A Big Data Architecture for Integration of Legacy Systems andData

by

Sanjay Jha

Thesis
Submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy
Central Queensland University

College of Information and Communications Technology

School of Engineering and Technology

February 2021

RHD Thesis Declaration

CANDIDATE'S STATEMENT*

By submitting this thesis for formal examination at CQUniversity Australia, I declare that it meets all requirements as outlined in the Research Higher Degree Theses Policy and Procedure.

STATEMENT AUTHORSHIP AND ORIGINALITY*

By submitting this thesis for formal examination at CQUniversity Australia, I declare that all of the research and discussion presented in this thesis is original work performed by the author. No content of this thesis has been submitted or considered either in whole or in part, at any tertiary institute or university for a degree or any other category of award. I also declare that any material presented in this thesis performed by another person or institute has been referenced and listed in the reference section.

COPYRIGHT STATEMENT*

ACKNOWLEDGEMENT OF SUPPORT PROVIDED BY THE AUSTRALIAN GOVERNMENT*

This RHD candidature was supported under the Commonwealth Government's Research Training Program/Research Training Scheme. I gratefully acknowledge the financial support provided by the Australian Government.

ACKNOWLEDGEMENT OF PROFESSIONAL SERVICES

Professional editor, Ms Sue Bond, provided copyediting and proof-reading services, according to the guidelines laid out in the University-endorsed Australian national guidelines, 'The editing of research theses by professional editors'.

DECLARATION OF CO-AUTHORSHIP AND CO-CONTRIBUTION

I. Jha, S., Jha, M., O'Brien, L., & Wells, M. (2014). Integrating legacy system into big data solutions: Time to make the change. *1st Asia-Pacific World Congress on Computer Science and Engineering 2014 (APWConCSE 2014).* DOI:10.1109/APWCCSE.2014.7053872

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL

In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, and Marilyn Wells, contributed to the paper by proof reading,suggesting section changes, and discussion on the analysis [10% of co-authors' contribution].

II. Jha, S., Jha, M., O'Brien, L., & Wells, M. (2015). Streaming big data into the enterprise architecture: Challenges and opportunities. *IEEE Conference 2nd Asia-Pacific World Congress on Computer Science & Engineering*, 2- 4 December,2015. Nadi, Fiji.

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, and Marilyn Wells, contributed to the paper by proof reading,suggesting section changes, and discussion on the analysis [10% of co-authors' contribution].

III. Jha, S., Jha, M., O'Brien, L., & Singh, P.K. (2016). Architecture for complex event processing using open -source technologies. *3rd Asia-Pacific World Congress on Computer Science and Engineering. (APWC on CSE)*, 2016, pp. 218-225, doi: 10.1109/APWC-on-CSE.2016.044.

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, and PK Singh, contributed to the paper by proof reading, suggestingsection changes, and discussion on the analysis [10% of co-authors' contribution].

IV. Jha, S., Jha, M., O'Brien, L., & Wells, M. (2017). Road map for data integration architecture for business intelligence. *19th International Conference on Advances in Information Systems* (*ICIES)*, 26-29 April, Porto, Portugal.

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, and Marilyn Wells, contributed to the paper by proof reading,suggesting section changes, and discussion on the analysis [10% of co-authors' contribution].

V. Jha, S., Jha, M., & O'Brien, L. (2018). A step towards big data architecture for higher education analytics. 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2018, pp. 178-183, doi: 10.1109/APWConCSE.2018.00036.

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, and Liam O'Brien, contributed to the paper by proof reading, suggesting sectionchanges, and discussion on the analysis [10% of co-authors' contribution].

VI. Jha, S., Jha, M., & O'Brien, L. (2019). Analysing computer science course using learning analytics techniques. *In 6th IEEE the Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2019, pp. 1-6, doi: 10.1109/CSDE48274.2019.9162369.

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 90% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, and Marilyn Wells, contributed to the paper by proof reading,suggesting section changes, and discussion on the analysis [10% of co-authors' contribution].

VII. Jha, S., Jha, M., O'Brien, L., Cowling, M., & Wells, M. (2020). Leveraging the organisational legacy: Understanding how businesses integrate legacy data into their big data plans. *Big Data andCognitive Computing 2020*, 4 (2), 15, **https://doi.org/10.3390/bdcc4020015**

NATURE OF CANDIDATE'S CONTRIBUTION, INCLUDING PERCENTAGE OF TOTAL
In conducting the study, I was responsible for forming the research question, collating literature, collecting data, analysing data, interpreting results, and drafting the paper.

This publication was written by me and my contribution was 95% contribution.

NATURE OF CO-AUTHORS' CONTRIBUTIONS, INCLUDING PERCENTAGE OF TOTAL
My co-authors, Meena Jha, Liam O'Brien, Michael Cowling and Marilyn Wells, contributed to the paper byproof reading, suggesting section changes, and discussion on the analysis [5% of co-authors' contribution].

ABSTRACT

Storing, analysing, and accessing data is a growing problem for organisations. Competitive pressures and new regulations are requiring organisations to efficiently handle increasing volumes and varieties of data, but this does not come cheap. Data sets grow rapidly in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. These kinds of data sets are referred to as big data and are too large or complex for traditional data-processing application software to adequately deal with. As the demands of big data exceed the constraints of traditional relational databases, evaluating legacy data and assessing new technology has become a necessity for most organisations, not only to gain competitive advantage, but also for compliance purposes. The challenge is how well an organisation's legacy data and processes can be integrated into the big data solutions. It is without a doubt that big data must be accommodated and the integration of legacy systems and processes into big data solutions must be dealt with. Legacy systems contain the significant and invaluable business logic of the organisation, with encoded 'business logic' that represents many years of coding, development, real-life experiences, enhancements, modifications, and debugging amongst other functions. Most legacy systems were developed without process or data models, which are now needed to support and be integrated into big data. To integrate legacy systems into a big data solution, re-engineering of the legacy processes is required depending on data used from the legacy system. Many approaches to re-engineer legacy systems have been developed; none are focused on integrating legacy systems with big data solutions (Vijaya & Venkataraman, 2018). Integrating legacy systems with big data solutions may change an organisation's Enterprise Architecture (EA), as EA demonstrates application, data, technology, and business architectures of an organisation. However, addressing the issues and scope related to incorporating legacy systems into big data allows mature legacy systems to become part of overall organisational changes so that big data solutions can be implemented in the organisation. This research addresses issues and concerns of existing legacy systems

within an organisation for decision making. This research further focuses on identifying current issues and concerns of integrating big data solutions with legacy systems in organisations and proposes a Big Data Architecture for Integration of Legacy Systems and Data.

This research is carried out using a combination of quantitative and qualitative studies. To understand the issues and concerns around integration of big data solutions with legacy systems, a survey was conducted on the practices of how organisations are using big data for different use cases and what practices are being employed to integrate big data solutions. The results from this survey were used to develop a Big Data Architecture for integration of legacy systems and data which addresses the identified issues and concerns of integrating big data solutions with legacy systems. A Big Data Architecture for integration of legacy systems and data was applied in industry to demonstrate the usefulness of the developed artifact of Big Data architecture and the architecture was evaluated using a higher institution. The key results and contributions emerging out of this thesis are listed below and divided into managerial implications, theoretical contributions, and future directions in this research:

- Identifying the requirements of organisations trying to achieve Big Data solutions for their organisations and what they are looking for while integrating legacy systems with Big Data solutions contributes to managerial implications.
- Identifying Analytics Value Chain as a part of the building blocks for streaming Big Data into EA is a theoretical contribution.
- Developing an e-business process model for a Big Data architecture to integrate data from different sources contributes towards the theoretical model which can be used by the organisations.
- Applying the Big Data architecture using open-source technologies contributes in verifying the theoretical model.
- Applying the Big Data architecture in Higher Education Institutions for Learning Analytics contributes towards validating the theoretical model.
- This Big Data architecture developed in this research should be applied to other case studies in different domains such as financials and the health sector to address real-time analytics. However, the possible obstacles the organisations can face are: insufficient understanding of Big Data technologies; complexity of big data technologies; complexity of managing data quality; and Big Data security issues.

ACKNOWLEDGEMENTS

I would like to acknowledge all the people who have assisted me throughout my studies at Central Queensland University. Firstly, I am extremely grateful to my supervisor Dr Marilyn Wells for her continuous guidance and encouragement. I also like to thank her for her invaluable insights and the time that she spent with me. I always came to her and got more than expected support when I had any research or general issues, Thanks for everything Marilyn.

I would like to thank my supervisor Dr Liam O'Brien for his encouragement and support. He has given me the expert advice while doing this research. I am very much grateful to him to show me the correct direction on my research. Thanks for your invaluable support and time Liam. Without you, this thesis would not have been possible. Without your support and encouragement, this thesis had no end.

I would also like extend my thanks to Dr Ergun Gide for his continuous support, and constructive suggestions during the planning and development of this research.

I would like to express my thanks to the Big Data Community, organisations and professionals who supported me by giving their time to finish the survey.

I would like to dedicate this thesis to my loving family....

Contents

LIST OF FIGURES

All figures listed above are referenced. Any self-drawn figures have not been referenced.

LIST OF TABLES

All tables listed above are referenced. Any self-drawn tables have not been referenced.

LIST OF ACRONYMS USED

API: Application Programming Interface

BDA: Big Data Analytics

BI: Business Intelligence

BIS: Business Information System

BLOB: Binary Large Object

BPI: Business Process Innovation BPI

BPM: Business Process Management

BPR: Business Process Re-engineering

CCTV: Closed-Circuit Television

CDC: Change Data Capture

CEP: Complex Event Processing

CI: Continuous Integration

CLOB: or Character Large Object

CRM: Customer Relationship Management

CSF: Critical Success Factors

DA: Data Analytics

DAG: Directed Acyclic Graph

DBMS: Data Base Management System

DBRE: Database Reverse Engineering

DSS: Decision Support Systems

EA: Enterprise Architecture

EAI: Enterprise Application Integration

EASI: Early Alert Student Indicator

ECDS: Electronic Coupon Distribution Service

ECM: Enterprise Content Management system

EFTPOS: Electronic Funds Transfer at Point of Sale

EIS: Executive Information Systems

ELT: Extract, Transform, Load

ERP: Enterprise Resource Planning

ESB: Enterprise Service Bus

ETL: Extract, Load, Transform

FEAPO: Federation of Enterprise Architecture Professional Organisations

GIA: Geospatial-Intelligence Agency

GIS: Geographic Information System

GPA: Grade Point Average

HDFS: Hadoop Distributed File System

HIPAA: Health Insurance Portability and Accountability Act

HRS: Human Resources System

HTTP: HyperText Transfer Protocol

IDC: International Data Corporation

IoT: Internet of Things

IT: Information Technology

JAD: Joint Application Development

JSON: Java Script Object Notation

KDD: Knowledge Discovery in Databases

KML: Keyhole Markup Language

LA: Learning Analytics

LMS: Learning Management Systems

MAV: Moodle Activity Viewer

MPP: Massively Parallel Processing MPP

MR4C: MapReduce for C

NoSQL: Not Only SQL

OLAP: Online Analytical Processing

OWL: Web Ontology Language

PDF: Portable Document Format

PRE: Procedure Reverse Engineering

RAD: Rapid Application Development

RDD: Resilient Distributed Dataset

RDF: Resource Description Framework

REPL: Read-Evaluate-Print Loop

RFID: Radio-Frequency Identification

SA: South Australia

SASE: Secure Access Service Edge

SCM: Supply Chain Management

SDLC: Software Development Life Cycle

SOA: Service Oriented Architectures

SIS: Student Information Systems

SQL: Structured Query Language

TPS: Transaction Processing Systems

WA: Western Australia

XML: Extensible Markup Language

YARN: Yet Another Resource Negotiator YARN

CHAPTER 1: INTRODUCTION

1.1 Introduction

Today, we live in a digital world in which information and technology are not only around us, but also play important roles in dictating the quality of our lives. Information and data are central components of our everyday activities. Industry studies have highlighted this significant development. Data is captured by organisations in our interactions across the various activities we do such as shop, work, play, study, and travel. An example is a supermarket checkout equipped with point-of-sale terminals. The transaction is primarily concerned with the sale to the customer but while the purchased items are being entered onto the bill it is usual for the machine to record, and thus capture, data that will allow calculation of stock movement, customer preferences and other information. Organisations are capturing information and data about us (or have done so in the past) where a lot of it is captured on legacy systems and is now regarded as legacy data (Pappas et al., 2018). By 2025, International Data Corporation (IDC) predicts there will be 163 zettabytes of data (Reinsel et al., 2017).

Traditionally, organisations mainly used structured data (meaning that it fits well within the rows and columns of a database) and internal data (meaning that it is created within the organisation) for decision making. Structured data fits well in rows and columns include mainly text and can be processed very easily using "Structured Query Language (SQL) (Groff & Welnberg, 2002). Structured data has various data types such as date, name, address, and character. Structured data are dependent on schema and can be analysed using schema based analyser. Internal data are data generated within the organisation and are structured. Internal data sources (manual and IT-based) are considered to be old and not fulfilling the requirements for a system, with continuous updates (Strand & Syberfeldt, 2020). However, because of the digital revolution data is now being generated and collected outside the organisation. This data can prove to be particularly useful for organisations while predicting and optimising the resources—such as human, machinery, and financial—with the needs of the organisation to achieve established goals. Big data is a concept which deals with data analytics; extracting information from datasets that are too large or complex to be processed by traditional data processing software applications. Big data describes data sets so large that they become difficult to manage with legacy systems.

Big data is characterised by 3Vs: Volume, Velocity and Variety (Laney, 2001). The 3Vs are three defining properties or dimensions of big data. Volume refers to the amount of data, Variety refers to the number of types of data, and Velocity refers to the speed of data processing. The 3Vs of big data was originally defined by Laney (2001) to describe the data management in 3-dimensions. However, other dimensions have also been added to big data characteristics such as veracity. According to Kitchin (2013), big data is characterised by high volume, velocity, variety, exhaustivity, resolution and indexicality, relationality and flexibility. A big data solution consists of big data sources, platforms and technologies which can collect, store, process, and analyse big data. Emerging academic research suggests that organisations that use big data for business analytics to guide decision making are more productive and experience higher returns on equity (Brynjolfsson et al., 2011). In this thesis, we explore the issues of integrating big data solutions with legacy systems and data.

There are many big data platforms developed by different companies' such as IBM, Talend, SAS to name but a few, which are used by organisations. However, they come with a proprietary cost associated with them. A big data platform does not have to be a cutting-edge technology. It can be any platform which supports the 4Vs (Volume, Velocity, Variety and Veracity) of big data. Legacy systems can process data very well in rows and column format. However, organisations have data in Binary Large Object (BLOB) or Character Large Object (CLOB) data format. These objects are only stored and cannot be analysed by legacy systems.

Organisations use a variety of systems to collect data, including enterprise resource planning systems, in house Information Systems, decision support systems and many others to support their day-to-day activities and operations. An organisation's operational capability is based on existing Information Systems. These existing Information Systems are called legacy systems (Bennett, 1995). Legacy systems contain significant and invaluable business logic of the organisation and is an Information System that may be based on outdated technologies but is critical to day-to-day operations. These legacy systems are old processes, technology, computer systems, or application programs that continue to be used, typically because they stillfunction for the users' needs, even though newer technology or more efficient processes of performing a task are now available. Organisations cannot afford to throw away or replace the business logic because of the many challenges associated with replacement, migration, and modernisation of legacy systems (Bisbal et al., 1999; Srinivas et al., 2016). Despite the availability of more cost-effective technology, about 80% of IT systems are running on legacy platforms and are assets of an organisation (Alkaseme, 2013). In 2002, IDC estimated that 200 billion lines of legacy code were in use on more than 10,000 large mainframe sites (Zoufaly, 2002).

Traditionally, organisations use only in-house data for decision making. In-house data is generated from these legacy systems and these legacy systems cannot be evolved with the changing requirements of the organisation (Malladi et al., 2016; Rahgosar, & Oroumchian, 2003).

Based on a survey of over 4,000 Information Technology (IT) professionals from 93 countries and 25 industries, the IBM Tech Trends Report identifies Business Intelligence (BI) and Data Analytics (DA) as major technology trends in the decade starting 2010 (IBM Report, 2011). Davenport has defined data analytics as: Analytics 1.0 is called an era of "Business Intelligence"; Analytics 2.0 is called an era of Big Data; and Analytics 3.0 an era of data-enriched offerings (Davenport, 2014). Organisations have shown great interest in Big Data analytics as it may help them better understand their business and market and give them information to make timely business decisions (Chen et al., 2012). 'Big Data' as a term is relatively new however concepts of 'Big Data' have been critical to all aspects of data storage, collection and retrieval since the early days of system implementations. Appropriate data processing and management could expose new knowledge, and facilitate in responding to emerging opportunities and challenges in a timely manner (Chen et al., 2013).

BI and DA based only on internal data from these legacy systems do not provide complete insight of the trends. BI and DA help business users to understand trends and drive insights to make sound business decisions. Decisions cannot be solely based on internal organisational data. For example, if an organisation wants to better manage its supply chain then it requires BI and DA to understand how the resources can be allocated to make the process of supply chain better and efficient, and it relies on transport data in the case of any delays happening. It is not only structured data but unstructured data, such as images, that is playing an important role in making decisions. Marketing is one of the examples; marketing brochures are designed based on extracting features for data analytics.

Collecting, storing and analysing unstructured data (for example, customer reviews of their Facebook pages or tweets), and external data, which is growing the fastest, is required to gain competitive advantage of BI and DA. Sources of external data are social media platforms such as Facebook, Twitter, and WhatsApp, but also search phrases in Google, data streams from smart devices, Internet of Technology (IoT), video streams from security cameras, or geographical information used by Uber or Lyft. All these sources, as well as many others, are adding to the enormous amount of unstructured data that is available to be used and analysed for BI and DA. BI and DA is not new to organisations. Information systems researchers and technologists have built and investigated BI and DA for more than 35 years. BI and DA have evolved. The stages of the evolution are: descriptive analytics, predictive analytics, and prescriptive analytics. Descriptive analytics analyses historical data and identifies patterns from samples for reporting of trends. Descriptive analytics helps us provide an understanding of the past. Predictive analytics uses data to find out what could happen in the future. The predictions are made by examining data about the past, detecting patterns, or relationships in these data, and then inferring these relationships forward in time. Prescriptive analytics analyses data to understand and prescribe what might happen in the future. It is based on descriptive and predictive analytics. Once the past is understood and predictions can be made about what might happen in the future, one needs to know what the best action will be, given the limited resources of the organisation. This is the area of prescriptive analytics. Nowadays, a tremendous amount of data is available in the form of structured data (from sensors) and unstructured data to perform analytics.

Because of the digital evolution it is now possible to add peoples' opinions and other information in the process of understanding customers. There is a large quantity of Big Data created every second awaiting to be processed and analysed but for Big Data to be more meaningful it should be integrated with legacy systems and data. The different data sources—such as sales, finance, marketing, product data with social data, sentiment data, demographic data, and competitor's data—when integrated, will be useful for the organisation to make decisions as it will combine data residing in different sources to provide users a unified view.

Big Data applications span a broad range of domains, including: (i) intelligent mobility systems and services; (ii) intelligent energy support systems; (iii) smart personnel healthcare systems and services; (iv) intelligent transportation and logistics services; (v) smart environmental systems and services; (vi) intelligent systems and software engineering; (vii) intelligent engineering and manufacturing (Acharjya & Ahmed, 2016; Akerkar, 2014; & Baker 2015).

Big Data is not only about massive amounts of data or the way data is being utilised. Today, organisations invest more in data manipulation and most of the time the stored data are unused, and they are not retrieved and utilised in a proper way (Arputhamary et al., 2015). It is also about the purpose of delivering added value to the organisation. Traditionally, this has been addressed by either storing this information in a database as Binary Large Object (BLOB) or Character Large Object (CLOB) data, or by using an Enterprise Content Management system (ECM). The drawback of storing a BLOB or CLOB in a database is that it can only be stored and retrieved. It cannot be searched or edited. With the tremendous growth of data everywhere, the decisions based only on in-house structured datasets and legacy systems will not provide a competitive edge to the organisation. There are a lot of legacy systems out there, with a lot of data that needs to relate to new data to make the best use of Big Data. 90% of the data getting generated today is unstructured and cannot be handled by traditional technologies (Linden, 2018).

Amazon, the Seattle-based e-commerce giant started in 1995 by Jeff Bezos, has always leveraged data. Every individual on Amazon.com undergoes a personalised experience, based on the actions they deal on the site. Amazon has been delivering products and services swiftly with a seamless user experience. In one of the latest business moves, Amazon has obtained a patent to ship goods to the required organisation before even the organisation decided to buy it, purely based on the predictive Big Data analytics. Amazon has combined the strengths in data analytics and the instinct for patenting key features to obtain a patent for what it calls Anticipatory Shipping.

LinkedIn, a social network service started in 2003, is used for professional networking. LinkedIn has created numerous data products, including People You May Know, Jobs You May Be Interested In, Groups You May Like, Companies You May Want to Follow, Network Updates, and Skills and Expertise.

Big Data has revolutionised the user experience and the way decisions are made. Today's organisations increasingly rely on technology that requires real time analytics. For this, organisations need to connect with their customers by tracking their purchasing histories, demographics, and how customers engage with them. Many organisations have not yetachieved this level of connection with their customers. They have multiple systems with multiple databases and multiple database vendors which do not include this connected data. Some of the systems, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and Supply Chain Management (SCM), have accumulated data over several years but this data often remains within its own silos and little connectedness has beenachieved to provide a real time view of their customers. For data to be useful to users, they must integrate customers with finance, and sales data, with product data, with marketing data,with social media data, with demographic data, with competitor's data, and more (van der Aalst, 2012).

Organisations need to take a holistic view to understand and recognise that success is built upon the integration of people, process, technology and data; this means being able to incorporate data from different sources into their business routines, their strategy and their daily operations. Organisations want to derive insights from information in order to make better, smarter, real time, fact-based decisions: it is this demand for depth of knowledge that has fuelled the growth of Big Data tools and platforms (Ernst Young, 2014).

Big Data analytics is evolving into a promising field for providing insight from exceptionally large data sets and improving outcomes while reducing costs. Making sense out of the vast data holding can help the organisation with more informed decision-making and provide competitive advantage (Raghupathi et al., 2014). Earlier, organisations used simple data analysis techniques like Structured Query Language (SQL) for their day-to-day operations that helped them in their decision making and planning. However, due to the increase in the size of data, especially the unstructured form of data, it has become almost impossible to process these data with the existing storage techniques and plain queries. Performance improvements in hardware and the emergence of new technologies have shown an explosive growth in the types of data and information processing capabilities.

Early adopters of Big Data solutions have gained a significant lead. Examining more than 400 large companies, Bain and Company found that those with the most advanced Big Data solution capabilities are outperforming competitors by wide margins (Pearson et al., 2013). The potential of Big Data is great; however, there remain several challenges to overcome (Bhadani et al., 2017). One of the challenges is data coming from multiple sources. In today's connected world, information often exists in multiple forms across multiple platforms. This can make it difficult for organisations to analyse all these formats and sources in a way that is accurate.

Big Data cannot be managed and processed by traditional software systems. Big Data are from different data sources, such as legacy data, internal data and external data, and integrating such data is especially important for decision making. A Big Data solution allows data to be integrated from different sources to provide users with a real-time view of business performance and operations. A Big Data solution produces a single, unified view of organisational data that can be used to provide actionable insights of the organisation.

The opportunities associated with Big Data analytics in different organisations have helped generate significant interest to help an enterprise better understand its business and market and make timely business decisions (Chen et al., 2012). BI and DA, that uses Big Data, include data processing, business-centric practices and methodologies that can be applied to various high impact applications such as e-commerce, market intelligence, e-government, healthcare, education, and security.

The goal of our research is to develop an architecture for integrating Big Data solutions with legacy systems and data. The architecture has been applied on an industrial organisation as well as a higher education to demonstrate the usefulness of the created Big Data architecture. In this chapter the background to the research and the research context is presented. The chapter describes the research questions, research design and the contributions of this research. The structure of the thesis is also presented later this chapter.

1.2 Background

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse. Big Data ranges from a few dozen terabytes tomultiple petabytes of data. Big Data is data whose scale, diversity, and complexity require new architectures, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big Data comes from a variety of sources including social media (generating data), scientific instruments (collecting data), mobile devices (tracking data), and sensor technology and networks (measurement data).

Every Big Data source has different characteristics, including the frequency, volume, velocity, type, and veracity of the data. Big Data can be acquired, stored, processed, and analysed in many ways. When Big Data is processed and stored, additional dimensions come into play, such as security, structure, and governance. Choosing an architecture and building an appropriate Big Data solution is challenging because so many factors must be considered. For decades, enterprises relied on relational databases, typical collections of rows and columns, for processing structured data. However, the volume, velocity and variety of data means that relational databases often cannot deliver the performance and latency required to handle large, complex data.

Big Data is being generated continuously at an exponential rate. Social media is one major contributor which is generating data explosively. Sensors, smartphones, and the internet are leading to huge data feeds. Eighty per cent of the world's information is unstructured and the amount of unstructured data is growing at 15 times the rate of structured information (Tanwar et al., 2015). The rise of unstructured data means that data capture had to move beyond merely rows and columns. Data integration provides uniform access to data from multiple sources (Doan et al., 2012), however the integration of heterogeneous data from different sources has been a critical challenge. Data quality plays an important role in Data Warehousing, and Data Warehousing plays a vital role in extracting the right information from the right place at the right time at the right cost, in order to make the right decision.

In Big Data solution, integration brings data gathered from legacy systems and other data sources together and makes it valuable for the organisation. It helps organisations achieve their target and objectives of return on investment. Integration of Big Data solutions with legacy systems can help organisations to identify missing opportunities; increase competitiveness; improve customer relationships; and supports better decision making (Gong et al., 2011). For Big Data integration new technologies are required which are capable of managing a vast variety of data and making it possible to run legacy applications on systems with thousands of nodes, potentially involving terabytes or petabytes of data, including legacy data. Integrating Big Data solutions with legacy systems is a challenging task. Some of these challenges include (Almeida & Calistru, 2012; Arputhamary & Arockiam, 2015; and Batlajery 2013):

- Legacy systems are resistant to change.
- Legacy systems have different data formatting.
- Legacy systems have incomplete data sets.
- Legacy systems have issues with data security.
- Legacy systems have issues with data governance.
- Legacy systems have silos and data is found at different locations.

However, legacy systems should be integrated with Big Data solutions. The reasons are as follows:

- According to IDC, the total volume of data will reach 163 zettabytes in 2025. It is expected that 80% of this will be unstructured data. Legacy systems are not capable of analysing datasets of unstructured data (van der Linden, 2018).
- The generation and storage of data can and will continue to grow exponentially and legacy systems are not capable of addressing the growth of unstructured data (Bloem et al., 2013).
- Legacy systems cannot deliver results of analytics at real time because of the cost and complexity. Many legacy systems were built to deliver data in batches, so they cannot furnish continuous flows of information for real-time decisions (Barton et al., 2012).

As far back as 1993, 3% of the world's information was stored on digital devices. In 2007, 94% of the world's information was stored on DVDs, CDs, memory cards and other digital devices. In 2014, 93% of all data were unstructured data, such as digital videos, audio files, photos, and graphics (Kościelniak et al., 2015). The research conducted among enterprises shows that the amount of data in enterprises is growing by 200% a year; moreover, it is predicted that the data holdings of enterprises will increase by 800% within five years, out of which 80% will be unstructured data. According to Jelonak et. al. (2014) managers are aware of the great significance of the integrated data coming from all departments—product, customer service and sales—and they indicate the necessity to share this data in real time.

Data integration in an organisation is a crucial and challenging problem. While business-critical information is often gathered in information systems such as ERP, CRM and SCM, the integration of these systems themselves, as well as the integration with the other information sources, is still a major challenge (Auer et al., 2014). Dong et al. (2013) have suggested data integration on schema mapping, record linkage and data fusion. The authors have identified the challenges of Big Data integration. The challenges identified were on volume and number of data sources, velocity, variety, and veracity. Following are the challenges identified while integrating Big Data solutions with legacy systems.

- The number of data sources, even for a single domain within an organisation, is estimated to be in the tens of thousands. The data sources have outgrown the capacity of existing database systems.
- Many data sources are continuously making data available on real-time.
- The data sources are heterogeneous in their structure, with considerable variety even for substantially similar entities.
- The data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. The same data element is of different size in different applications and may have different meaning in different business areas.
- Obsolete technology such as traditional databases and legacy applications make it difficult to integrate. Data handling system could be an out-dated database system.
- Lack of people with knowledge of the systems and data and what the data means or represents.
- Lack of documentation of the systems and data and what the data means or represents.
- Silos and Fifedoms where people do not want to give up access to their data.
- Change management and selling of the idea of why using the data as part of a Big Data solution is going to deliver benefit to all parts of the organisation.

From one perspective, legacy systems and data are time-tested, having their value proven by long use representing decades of effort and customisation, becoming reliable parts of an overall IT strategy along the way. These entrenched software systems often resist evolution because their ability to adapt has diminished through factors not exclusively related to their functionality. Software must be continually adapted, or it will become progressively less satisfactory in “real-world” environments (Lehman et al., 1998). This is due to the continuous change of user requirements and technical environments. Many legacy systems have been large investments for organisations, and they contain invaluable business logic and knowledge. An empirical study on software modernisation addresses that the feasibility of a legacy system being evolved, maintained, and integrated with other systems is improved due to modernisation (Mishra et al., 2009). The authors argue that the ending of technological support and expected system lifetime reflect mandatory system modernisation. The data handling system could be an outdated database system, such as home-grown database management system, and flat file system FILEMAN. Program-managed memory overlays were an innovative use of flat-file technology, using a table-driven approach to separate process from data as database technology has done. However, a fixed record length limits the number of fields available to accommodate the growing user requirements. The same field can be used for multiple uses with different meanings depending on the user groups. Allowing individual user discretion rather than enterprise standards permitted a restrictive physical data limit to serve more customers, but this approach is reaching its design and operation limit as this cannot be integrated to Big Data.

Data sources (even in the same domain) are heterogeneous both at the schema level regarding how they structure their data and at the instance level regarding how they describe the same real-world entity, exhibiting considerable variety even for substantially similar entities. One of the major issues raised in Big Data solutions is the input and output process. Data entry and storage cannot be handled with processes currently used for relational databases. Big Data can cause performance problems, especially when traditional databases are involved. Data analysis is also exemplified by applications with hard limits on the size of data they can handle. Another issue is large heterogeneous datasets; a major challenge is to figure out what data one has and how to analyse it (Jacobs, 2009).

Big Data requires IT leaders and technology specialists to acquire and apply tools, techniques and architectures for analysing, visualising, linking, and managing big, complex datasets. Data integration requires capturing and integrating both structured and unstructured data both internal and external to the enterprise. Big Data as a part of the overall decision support system or Business Intelligence (BI) system can only be realised by data integration from different sources (Chen et al., 2012). The different types of Big Data sources are:

- Media as a Big Datasource
- Cloud as a Big Datasource
- Web as a Big Data source
- IoT as a Big Data source
- Databases as a Big Data source.

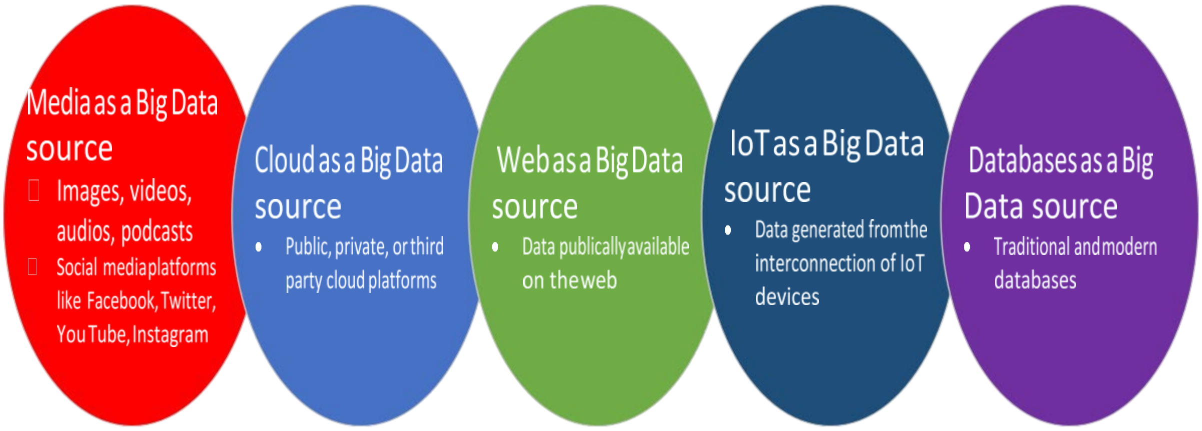

Media as a Big Data source
Images, videos, audios, podcasts
Social media platforms like Facebook, Twitter, YouTube, Instagram
Cloud as a Big Data source
Public, private, or third party cloud platforms
Web as a Big Data source
Data publically available on the web
IoT as a Big Data source
Data generated from the interconnection of IoT devices
Databases as a Big Data source
Traditional and modern databases

Figure 1.1: Big Data Sources

These Big Data sources which need to be integrated with a Big Data solution are based on business needs. For example, if the business need is to "find product information and knowledge to respond to customer query", this requires consolidated information around a customer or set of customers, their product purchase history, and past issues to correlate similar complex customer issues with most likely resolution paths. The outcome will be predicting and uncovering emerging customer issues following a product launch before they become endemic. According to a (Van der Linden et al., 2018):

> Digital customer experience is all about understanding the customer, and that means harnessing all sources – not just analysing all contacts with the organisation, but also linking to external data sources such as social media and commercially available data. For the digital supply chain, it is about collecting, analysing and interpreting the data from the myriad of connected devices.

Organisational silos and a dearth of data specialists are the main obstacles to putting Big Data to work effectively for decision making (Capgemini, 2012). Most of the legacy systems were developed without process models or data models which are now required to support data standardisation. The organisation requires that modernised systems use logical data models to represent data requirements. As organisations grow, whether organically or through mergers and acquisitions, they expand in many directions. Organisations are very much likely to gain duplicate technologies and business functions. To remove organisational silos business processes, need to be re-engineered so that Big Data analytics can be applied and work effectively for decision making.

Business Process Re-engineering (BPR) is the analysis and redesign of workflow within and between enterprises (Vidgen et al., 2017). BPR involves the radical redesign of core business processes to achieve dramatic improvements in productivity, cycle times and quality. Using BPR, companies start with a blank sheet of paper and rethink existing processes to deliver more value to the customer. They typically adopt a new value system that places increased emphasis on customer needs. Companies reduce organisational layers and eliminate unproductive activities in two key areas. First, they redesign functional organisations into cross-functional teams. Second, they use technology to improve data dissemination and decision making. BPR assumes the current process is irrelevant, does not work, or is broken and must be overhauled from scratch. Sucha clean slate enables business process designers to disassociate themselves from today's process and focus on a new process. Organisations should be re-engineering business processes now to fit Big Data analytics and not delay it any longer. It is like the designers projecting themselves into the future and asking: What should the process look like? What do customers want it to look like? What do other employees want it to look like? How do best-in-class companies do it? How can a new information system facilitate the process?

Business Process Management (BPM) involves an approach to combine business view from a technical perspective focusing on expert's views (Hepp et al., 2005). BPM should be implemented to have an impact on the business by delivering benefits. The core business processes for BPM are primary business activities and activities contributing towards the achievement of the strategic objectives of the organisation. BPM activities and BPM approach includes "Enterprise-wide business transformation program" including strategic business drivers that are heavily linked to business strategies (Jeston, & Nelis, 2008). In times of rapidly changing business environments and newly emerging technologies, organisations are not only forced to adapt their business models, their strategy and their organisational structures but also their information systems (Barth, & Koch, 2018). Organisations have implemented enterprise information systems called as enterprise resource planning (Davenport et al., 2004), however, evolution, maintenance and eventual replacement of such systems (Gable et al., 2001) has received considerably less attention.

1.3 Problem Outline

As organisations grow and expand in more than one direction through mergers, acquisitions, or through requirement generation, they are also likely to gain technologies that duplicate existing capabilities, workflows in need of significant overhaul, and legacy systems whose contributions to the business value are still very critical. Organisations in such situations must address the issue of integrating Big Data (data from internal sources, data from external sources), which can drive the alignment between business' operating needs and the processes, applications data and infrastructure required to support the ever-more dynamic requirements. As the demand for managing information increases, they need to focus the efforts on integrating business processes and data.

Because of globalisation an organisation may have several centres at different geographical locations. Most of the organisations maintain many billions of lines of code today associated with thousands of heterogeneous information systems at these different centres. Over the years, legacy systems have been continuously modified to implement changing needs, including functional requirements, business rules, and data architectures. There are many System

Development Life Cycle (SDLC) methodologies such as waterfall, Agile, Spiral, Joint Application Development (JAD), and Rapid Application Development (RAD). SDLC includes a detailed plan for how to develop, alter, maintain, and replace a software system. SDLC involves several distinct stages, including planning, design, building, testing, and deployment. SDLC is important because it breaks down the entire life cycle of systemdevelopment, thus making it easier to both evaluate each part of system development and for programmers to work concurrently on each phase. Continuous Integration (CI) within SDLC is a development practice where members of a team integrate their work frequently, each integration being verified by an automated build to detect integration errors as quickly as possible. Methodologies in SDLC are addressing dynamic development, risk assessment development, and traditional approach of development. However, methodologies used in

SDLC do not address the issues of integrating legacy systems with Big Data solutions.

There is also a need to recognise and to develop data migration plans which can incorporate Big Data. There is a need to address first what is required for the integration of legacy systems with Big Data and then we can focus on how this can be achieved with the identified scope. For this, we need to address the impact of legacy system complexity on Big Data, with an overview of the existing legacy modernisation approaches, describing data categories and how Big Data is different to it, what data administration strategies need to be addressed while modernising the legacy system, and the scope of integration of legacy systems with Big Data solutions.

Many approaches to modernising legacy systems have been developed. The current situation in legacy system modernisation can be summarised as follows:

- Database reverse engineering for migration of relational databases (Hainaut, 1998).
- Redevelopment approaches (Kang et al., 1998).
- Wrapping solutions (Rahgozar et al., 2003).
- Reverse engineering of procedural components of a large application (Deursen et al., 1999).
- Knowledge Based Software Reuse (KBSR) Process and Repository for systematic legacy system modernisation (Jha et al., 2013).

The current modernisation approaches do not address the issue of integrating legacy systems with Big Data solutions. Organisations such as IBM, SAS, Talend, Tableau, to name but a few, are using their own platforms for Big Data analytics. The reports generated from Big Data platforms are merged with legacy system reports. During our literature review, we identified that there is no architecture designed for integration of legacy systems with Big Data solutions (Seetharam et al., 2017). One of the reasons for not having an architecture for legacy systems with Big Data solutions, as an activity in any of the examined modernisation approaches, could be issues related to the complexity of integrating old systems due to their architecture, underlying technology, or design.

Big Data integration has several issues (Almeida et al., 2012) such as the uncertainty of data management, complexity behind the transmission, access, and delivery of data and information from a wide range of resources, and to ensure the right-time data availability to the data consumers.

While integrating legacy systems with Big Data solutions, the challenge is to create a target system which will incorporate Big Data with other data sources for rapid use and rapid data interpretation requirements. Logical data models are used to represent data requirements of an organisation. Data models must be developed to represent the policies, strategies, and Big Data issues. The architecture we are developing will include business functions, policies, rules, and data elements. This architecture will ensure that data structures, including Big Data, can be identified and linked to supported processes. In an organisation there are usually several legacy systems such as payment systems, health care systems, and human resource systems, as shown in Figure 1.2. There is a need to have data integration strategies based on data modelling, data standardisation, data migration and data architecting planning. Figure 1.2 shows that modernised systems should reuse functionality provided by legacy systems and data integration strategies. To employ the data integration strategies, we also need to understand a well-formed logical architecture for structured data. Structured data are collected from different source use integration techniques – such as Extract, Transform, Load (ELT)/ Extract, Load, Transform (ETL)/Change Data Capture – to transfer data into a Data Base Management System (DBMS) data warehouse or operational data store, and then offer a wide variety of analytical capabilities to reveal the data. Some of the analytical capabilities include dashboards, reporting, BI applications, summary and statistical query, semantic interpretations for textual data, and visualisation tools for high- density data.

Organisations are taking insights from multiple sources of Big Data including the Web, biological and industrial sensors, video, email, and social communications to support decision making in real time from fast-growing in-motion data from multiple sources (Broekema et al., 2012). Universities are looking at Google Reviews to see students' reviews on factors such as overall satisfaction; courses and teaching; support for students; and campus facilities. Organisations are using newer generated data on Twitter, Facebook, or other social media to support the decision-making process. Bulmer and DiMauro (2011) have surveyed 105 participants in 97 organisations in 20 countries and found that social network participation increasingly affects executive decision making at companies. Traditional decision-making processes are being disrupted by social media. Changes are taking place in organisations' internal and external use of data and there is a recognised need for external data and peer input in decision making. There is evidence of organisations looking for ways to harness the value of Big Data to improve the decision-making process. According to Olszak (2016), the ability to take advantage of all available information has become a critical ability for organisational success. Janssen, Estevez, and Janowski (2014), emphasis that creation of value from data requires combining large datasets originating from different and heterogeneous data sources.

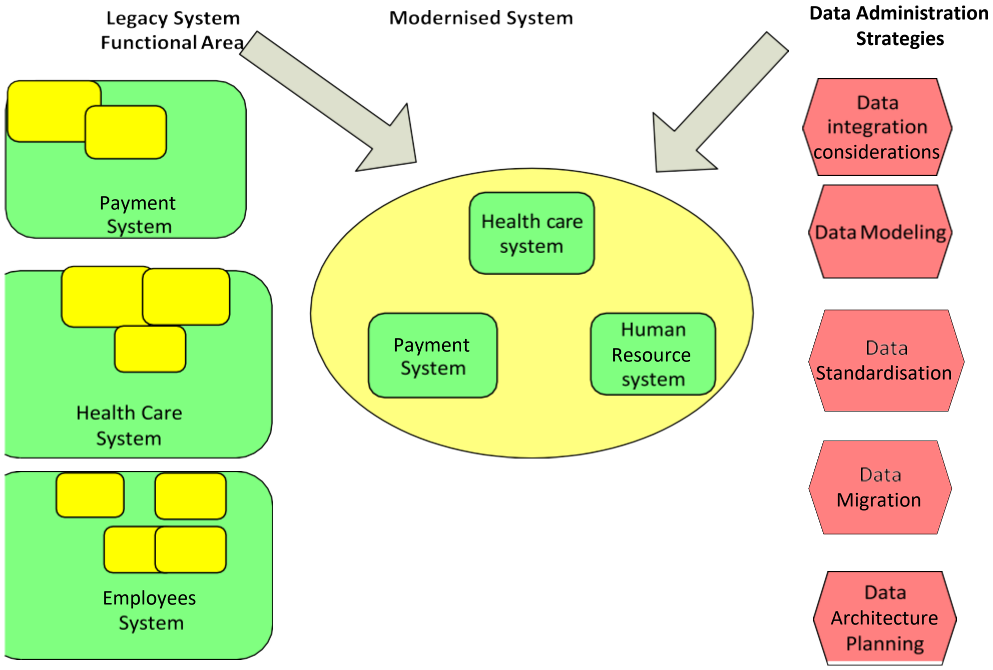


Figure 1.2: Modernised System Using Legacy System Functional Areas and Data Administration Strategies

However, in the research community we found no evidence of scholarly articles on integrating legacy systems with Big Data solutions. There are many scholarly articles (Janssen, Voort & Wahyudi, 2017; Lytras & Raghvan, 2017) supporting the need of, and adding value in, decision making by leveraging Big Data. However, there are no listed approaches as to how legacy systems and data are integrated with Big Data solutions. A search using the keywords “Legacy Systems” and “Integration with Big Data” results in Big Data technologies and how Big Data can be integrated with Hadoop sources but does not result in any approaches showing how legacy systems can be integrated with Big Data solutions.

1.4 Research Context and Research Questions

To explore our research, several research questions have been defined around integrating legacy systems and data with Big Data solutions. The research questions are:

Research Question 1: *What are the challenges in integrating Big Data solution(s) with legacy systems and data?*

Research Question 1 allows us to understand the challenges in integrating Big Data solution(s) with legacy systems and data, and the requirements to support BI and DA in an organisation. Big Data processing may require real time, near real time or batch processing. Using different characteristics of Big Data may require different technology infrastructure components and data architectures. The technology infrastructure components and the technology architectures as well as data architectures changes must be captured by an organisation's existing EA to enable conducting BI and DA, using a holistic approach at all times, for the successful development and execution of an organisation's strategy.

Research Question 2: *How does integration of legacy systems and data with Big Data solutions impact Enterprise Architecture (EA) in terms of the Business Architecture, Information Architecture, Technology Architecture and Data Architecture?*

Research Question 2 allows us to understand how integration of legacy systems and data with Big Data solutions impacts the Enterprise Architecture of an organisation. This question helps us understand the entire landscape of Enterprise Architecture and how Big Data solutions can be a proper fit in an organisation. Answering this question further enhances our understanding of an Enterprise Architecture approach to information management; that Big Data is an enterprise asset and needs to be managed from business alignment to governance as an integrated element of the existing legacy information management architecture. This helps us answer questions related to business context such as: How will we make use of Big Data? Which business processes can benefit from the use of Big Data?

Research Question 3: *How do we address the challenges of integrating legacy systems and data with Big Data solutions?*

Research question 3 allows us to generate an architecture to integrate legacy systems and data with Big Data solutions, which focuses on the challenges identified in the Research Question 1.

1.5 Research Goals

The goal of our research is to:

- Advance the state-of-the-art of integration of legacy systems and data with Big Data solutions through the development of an architecture for systematically integrating legacy systems and data with Big Data solutions.
- Identify the issues and challenges of integration of legacy systems and data with Big Data solutions and the state-of-the-art practice used in industrial organisations and higher education institutions.
- Construct the building blocks for Big Data into the enterprise architecture to address application, data, infrastructure, and technological issues related to Big Data integration with legacy systems.
- Propose and develop an architecture for integration of legacy systems with Big Data solutions.
- Show how the Big Data architecture is used in organisations and higher education institutions.

1.6 Research Methodology Design

This research has been carried out using a combination of quantitative and qualitative studies. This research has received ethical clearance from Central Queensland University's ethics committee. The ethical clearance numbers are H16/08-221/0000020267 and 0000020981.This research was conducted in four phases which are:

1. Phase 1
    - Study of Big Data Solutions

    - Study of Legacy System Modernisation Approaches
    - Survey of challenges from the issue of how legacy systems are integrated with Big Data solutions.
2. Phase 2
    - Study of Enterprise Architecture
    - Study of mapping Big Data Solutions to Enterprise Architecture
3. Phase 3
    - Study to identify business processes which will benefit using a Big Data solution.
    - Construct the building blocks of Big Data to address application, data, infrastructure, and technological issues related to Big Data integration with legacy systems.
4. Phase 4
    - Combining results from Phases 1, 2 and 3.
    - Develop an architecture to integrate legacy systems and data with Big Data solutions.
    - Apply case studies on the developed Big Data architecture to integrate legacy systems and data with Big Data solutions.
    - Evaluation of the Big Data architecture to integrate legacy systems and data with Big Data solutions using industrial and higher education case studies.

Phase 1 is dominated by a literature survey followed by an experimental survey to identify existing approaches towards integration of legacy systems and data with Big Data solutions. The literature survey on Big Data solutions helped us to explore the state-of-the-art of Big Data solutions research and practice. We carried out a survey to identify the issues and concerns organisations are facing in integrating Big Data solutions with legacy systems. This is a brand new survey based on literature review and supervisors' consultation. The survey questions were framed to answer our research question and contributed towards Phase 1 of our research methodology and design. This survey was sent out to only people working in the Big Data area. The survey questions were fine-tuned through literature review and supervisors' consultations. The educational qualifications and experiences of the participants suggests the reliability of this survey. We explored the problems and gaps in the legacy modernisation approaches to address integrating Big Data solutions with legacy systems. The experimental survey helped in understanding the existing approaches that are used in organisations for integrating legacy systems with Big Data solutions. This survey also helped in identifying issues and concerns related to the implementation of Big Data solutions in organisations. The results from this survey supported the creation of an effective process for integrating legacy systems and data with Big Data solutions which has not yet been completely answered by the research community and in practice.

Phase 2 is dominated by a literature survey of enterprise architecture and the role it plays in the integration of Big Data solutions with legacy systems. We explored the problems of mapping Big Data solutions to different layers of an EA. The layers used to map Big Data solutions were business architecture layer, application architecture layer, data architecture layer, and technology architecture layer.

Phase 3 outlines the study of identifying business processes, which benefit from the integration of Big Data solutions with legacy systems. This phase is dominated by a literature review of Big Data use within organisations and empirical studies on how Big Data solutions are used in organisations. This phase helps to gain an understanding of use cases for Big Data solutions and helped in the construction of the entire information landscape (EIL).

Phase 4 is dominated by quantitative and qualitative study. This Phase combines the results from Phases 1, 2 and 3 from which we developed an architecture to integrate Big Data solutions into legacy systems. In this Phase we demonstrated the usefulness of our developed Big Data architecture artifact on industry and evaluated the Big Data architecture artifact by applying on higher education.

1.7 Contributions

The contributions from our research are:

- Identifying current issues and concerns of integrating legacy systems and data with Big Data solutions in organisations: We conducted a survey and had responses from industries such as financial services, healthcare, aviation, higher education, the energy sector, and insurance. We have demonstrated that many organisations are implementing Big Data solutions and integrating their legacy systems and data with their solutions. A Big Data architecture for Big Data integration with legacy systems should be able to provide solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats. The integration should be able to maintain data accuracy and integrity, and this should be addressed by the Big Data architecture.

The contribution we made is by identifying the requirements of organisations trying to achieve Big Data solutions for their organisations and what they are looking for while integrating legacy systems with Big Data solutions. What are the obstacles the organisations would like to resolve while integrating legacy systems with Big Data solutions?

- Creating and developing a Big Data architecture to integrate legacy systems and data with Big Data solutions: We applied Six Sigma (Tennant, 2001) activities for business process re-engineering and developed organisational Big Data strategy for re-engineering. We identified that Big Data analytics and business processes can be combined using re-engineering and can deliver benefits to the organisations and customers.

1.8 Thesis Structure

The thesis is organised as follows:

Chapter 2 presents a literature review on the modernisation of legacy systems and Big Data solutions. It discusses from where the problem of integrating legacy systems and data with Big Data solutions comes. What is already known about modernisation approaches and Big Data solutions? It also discusses Big Data technologies used in a Big Data solution.

Chapter 3 presents the results of our survey conducted to identify the issues and the scope related to integrating legacy systems and data with Big Data solutions and identifies the gap of mature legacy systems and Big Data becoming part of decision making so that Big Data solutions can be implemented in the organisation.

Chapter 4 discusses requirements of Big Data solutions, what a Big Data Strategy is, what a BD Solution looks like and how it relates to EA, and what constitutes the Analytics Value Chain. This chapter also presents challenges and opportunities of Big Data and EA, different layers of EA, and how Big Data solution requirements are conveyed by an organisation's EA. This chapter outlines the building blocks of streaming data. This chapter also discusses technology and data architecture used to support the integration of legacy systems and data with Big Data solutions.

Chapter 5 presents an overview of the Big Data architecture that we have developed for the integration of legacy systems and data with Big Data solutions. This chapter focuses on combining Big Data analytics using business process re-engineering. It discusses how Business Process Re-engineering (BPR) involves the radical redesign of core business processes to achieve improvements in productivity, cycle times and quality by integrating legacy systems and data with Big Data solutions. This Chapter also focuses on selecting a business process for re-engineering and develops on tasks to combine Big Data analysis with business processes.

Chapter 6 is dedicated to demonstrating the usefulness of Big Data architecture on complex event processing for the Electronic Coupon Distribution System, where we have applied our Big Data architecture using open source technologies. It shows how Big Data architecture can be applied to integrate legacy systems and data with Big Data solutions to drive organisations' decisions on the distribution of an Electronic Coupon Distribution Service, using location information, past shopping/travel history, gender, likes/dislikes, and so forth. This Chapter also explores how different types of data such as static information (gender, age, etc.), previous history (where the person travelled to, what they bought, etc.), as well as real-time information about a customer (current location, current shopping habits, etc.) would all be utilised in Electronic Coupon Distribution Service.

Chapter 7 outlines the implementation and use of the Big Data architecture for higher education institutions and shows how Big Data architecture can be applied to integrate legacy systems and data with Big Data solutions to improve decision making within such institutions. This chapter outlines higher education business domains for Big Data analytics and evaluates our Big Data architecture. This chapter explores how Big Data can be leveraged to analyse students' online behaviour and how it relates to academic performance. It outlines how the architecture has been used in the development of a Big Data architecture that supports a model for academic success within an Australian University.

Chapter 8 outlines the implementation and use of our Big Data architecture integrating legacy systems and data with Big Data solutions selecting core business processes from higher education legacy systems and re-engineering these selected business processes to integrate with Big Data solutions.

Chapter 9 concludes and summarises this research work, revisits the research questions and outlines the contributions of the Big Data architecture and body of knowledge. This chapter also discusses the limitations of our research work and lists the future work.

1.9 Chapter Summary

Over the past few years there has been a great deal of interest in Big Data solutions in organisations trying to achieve competitive advantage. Data is growing at an enormous rate and is generated by other organisations, by users on social media or provided by devices of the Internet of Things (IoT). The creation of value from data requires combining large datasets originating from different and heterogeneous data sources. In practice there is often a whole chain of activities in which various actors play a role. Big Data needs to be analysed with organisational data. This requires integration of legacy systems and data with Big Data solutions. However, there remain many challenges to integrate legacy systems and data with Big Data solutions.

In this chapter we have discussed the need for the integration of legacy systems and data with Big Data solutions. We have also outlined the contributions of our research work with one of the main ones being the development of a Big Data architecture for integrating legacy systems and data with Big Data solutions that can be used for business intelligence and supporting decision making within the organisation. Big data is related to Big Data Analytics (BDA) which is needed to create value of the data. The next chapter outlines the literature review of Big Data definitions and challenges; legacy systems definitions and challenges; impact of legacy systems' complexity on Big Data solutions; and enterprise architecture and its role in integrating legacy systems and data with Big Data solutions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Many organisations own a few legacy systems and data and maintain them to fulfil their daily business operations. Legacy systems and data cannot always accommodate newly emerging business needs such as brand analysis, thus might negatively impact an organisation's decision-making capabilities (Sivarajah et al., 2017). Information is going to be our generation’s next natural resource, like steam was to the 19th century. Mobile is everywhere—more people have a cell phone than running water and 25% of the world will be on a social network—that is what has created big data: 2.5 billion gigabytes of data is created per day (Thau, 2014). This chapter discusses the state-of-the-art of Big Data solutions and what is missing while integrating legacy systems and data with Big Data solutions. This chapter also presents our literature review on the current state of research focusing on Big Data definitions and challenges. This chapter also discusses legacy systems definitions and challenges and the impact of the complexity of legacy systems and data while integrating with Big Data. It is important to capture the architectural requirements needed to generate Big Data architecture to integrate legacy systems and data with Big Data solutions. And this gets addressed while discussing the influence of Enterprise Architecture (EA) and its role on Big Data solutions. And finally, this chapter presents the influence of EA and the role it plays while integrating legacy systems and data with Big Data solutions.

2.2 Big Data Definitions and Challenges

There are many definitions of Big Data given by different researchers and data analytics practitioners. Big Data is extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. Every Big Data source has different characteristics, including the frequency, volume, velocity, type, and veracity of the data. We have examined the existing definitions within the literature and on the Web and the main definitions are:

1. According to Maurao et al. (2016), Big Data refers to large datasets. De Maurao et al. (2016) have defined Big Data as the information asset characterised by High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.
2. The initial definition of Big Data is characterised by 3Vs: Volume, Velocity, and Variety. The 3Vs are three defining properties or dimensions of Big Data. Volume refers to the amount of data, variety refers to the number of types of data, and velocity refers to the speed of data processing. The 3Vs of Big Data were originallydefined by Laney (2001) to describe the data management in 3-dimensions. Laney's3Vs underpinning the expected 3-dimensional increase in data Volume, Velocity and Variety, did not mention Big Data explicitly. His contribution was associated with Big Data several years later (Beyer & Laney, 2012; Zaslavsky et al., 2013; Zikopoulos & Eaton, 2011).
3. Kaisler et al. (2013) have defined Big Data as the volume of data in the range of Exabytes and argued that the existing technology is incapable and cannot effectively process and manage this volume of data.
4. Khan et al. (2014) have defined Big Data based on 7 dimensions. These 7 dimensions are: veracity; value; validity; variability; volatility; virtual; and visualisation. Landmark solutions (2016), added other dimensions to it and called it complexity.
5. Katal et al. (2013), have given different definitions of Big Data based on variety, volume, velocity, variability, complexity, and value.
6. Fisher et al. (2012) have defined Big Data as data that cannot be handled and processed in a straightforward manner.
7. Chen et al. (2012) have defined Big Data as the data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualisation technologies.
8. Ward and Barker (2013) and Microsoft Research Group (2013) have defined Big Data as the process of applying serious computing power, the latest in machine learning and artificial intelligence, to seriously massive and often highly complex sets of information.

9. Mayer-Schönberger and Cukier (2013) have defined Big Data as a phenomenon that brings three key shifts in the way we analyse information that transform how we understand and organise society: 1. More data 2. Messier (incomplete) data 3. Correlation overtakes causality.
10. Lee (2017) has defined Big Data with adding another dimension to it: decay. The author argues that the decay of data refers to the declining value of data over time. In a time of high velocity, the timely processing and acting on analysis is more important. IoT devices generate high volumes of streaming data, and immediate processing is often required for time-critical situations such as patient monitoring and environmental safety monitoring.

The definitions outlined above pertain to the characteristics of data, technologies, and techniques. There is an absence of a consensual definition of Big Data in the business and research communities. Also, the research community has not identified if data generated by legacy systems forms part of Big Data. According to De Maurao et al. (2016) researchers have adopted "implicit" definitions through anecdotes, success stories, characteristics, technological features, trends, or its impact on society, organisations, and business processes. The existing definitions for Big Data provide very different perspectives, denoting the chaotic state of the art. Big Data is considered in turn as a term describing a social phenomenon, information assets, data sets, storage technologies, analytical techniques, processes and infrastructures.

The handling of Big Data is extraordinarily complex. Big Data are collected from numerous data sources. Discovering value from Big Data requires Big Data challenges to be addressed. Following are the four Big Data challenges.

1. The first challenge is understanding business strategy, governance, and organisational key business processes for Big Data (EY, 2014).
2. The second challenge involves developing models, policy rules or standards that govern what data is collected and how it is stored, arranged, integrated, and used within an organisation and its various systems (Sivarajah et al., 2017).
3. The third challenge is defining the Big Data architecture of an organisation's application solutions against business requirements (Vidgen et al., 2017).

4. The fourth challenge is defining technologies which requires the hardware, operating systems, programming, and networking solutions a business employs and how those may be used to achieve its current and future objectives (Chen et al., 2014).

Based on the above identified four Big Data challenges we define Big Data as: An Information asset characterised by volume, velocity, variety, veracity, value, validity, visualisation and complexity which requires business strategy, data models, application solutions against business requirements, and technologies to support the characteristics of data to transform it into value.

2.3 Big Data and its Evolution

Big Data has evolved in several ways and the Big Data evolution can be divided into three activities:

1. Adding dimensions to its characteristics (Chen et al., 2012; Kwon et al., 2014; Laney, 2001; Lee, 2017).
2. Implementing technological changes (Bhadani & Jothimani, 2016).
3. Implementing analytical changes (Gonçalves et al., 2013; Kaplan & Haenlein, 2010; O'Reilly, 2007; ReportsnReports, 2016).

Big Data evolution started with the evolution of data collections. Figure 2.1 illustrates Big Data processing stack (Beyer et al., 2011; Driscoll, 2011). According to Driscoll (2011), Big Data processing stack comprises the following three major layers:

- Foundation layer: This layer provides the infrastructures for storage, access, and management of Big Data.
- Analytics layer: This layer provides machine learning algorithms to extract correlations and features from data and feeding classifications, predictions, and prescriptions.
- Applications layer: This layer provides the use of more generic lower-layer technologies and is exposed to end users for visualisations of reports and data.

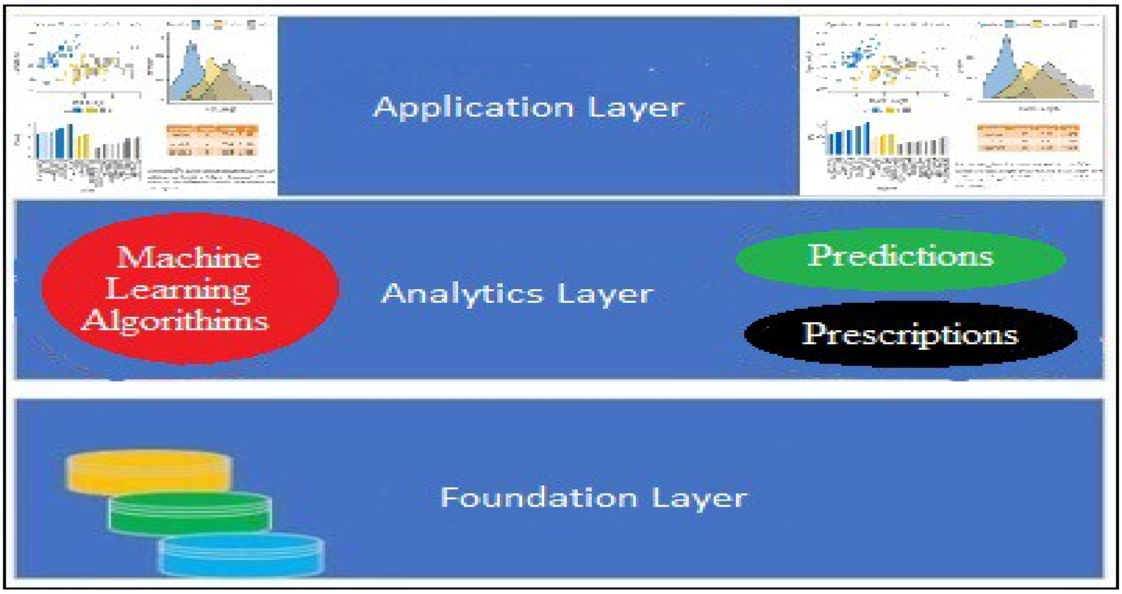


Figure 2.1: Big Data Processing Stack (Akerkar, 2014)

To assess and understand Big Data solutions, solutions that are provided to handle big and complex data for an organisation, we need to address and focus on what kind of data is required, what are the data sources, where and how data will be stored, what combination of technologies will be used, and what analytics is required from Big Data for an organisation.

According to Benes (2018), legacy systems are one of the biggest roadblocks preventing organisations from implementing Big Data solutions, as Big Data include textual to multimedia content on a multiplicity of platforms and legacy systems are incapable of handling this Big Data. Big Data solutions provide Big Data analytics services. Big Data analytics extract valuable information for the organisation. In a July 2017 survey of 560 marketing professionals worldwide conducted by Harvard Business Review Analytic Services, 36% of respondents said that legacy systems were one of the biggest roadblocks providing limitations of current data processing systems (Jin, Wah, Cheng, & Wang, 2015), and preventing organisations from implementing Big Data analytics using multiple information assets and some of these could be legacy. A third of respondents reported that legacy data silos stifled their progress. The results indicate that old technologies and organisational structures make life more difficult for marketers who want to utilise Big Data in novel ways (Benes, 2018).

Several Big Data initiatives have emerged in the public sector based on characteristics of Big Data. For example, in March of 2012, the Obama Administration put forward the Big Data

Research and Development Initiative. The objective of this initiative was to understand the technologies needed to manipulate and mine massive amounts of information and apply that knowledge to other scientific fields as well as address national goals in the areas of health, energy, defence, education and research (Mervis, 2012). Facebook, a company that did not exist a decade ago, gets more than 10 million new photos uploaded every hour. Facebook members click a "like" button or leave a comment nearly three billion times a day, creating a digital trail that the company can mine to learn about users' preferences (Mayer-Schonberger et al., 2013). There is a huge opportunity to work on the data and create value (Desouza et al., 2017). In a survey report by Bloomberg Businessweek (2011), 97% of companies with revenues exceeding $100 million were found to use business analytics.

2.4 Legacy Systems Definitions and Challenges

There are many definitions of legacy systems given by different researchers and software practitioners. We have examined the existing definitions within the literature and on the Web and these include:

1. A legacy system is defined as a system which was developed sometime in the past and is critical to the business in which the system operates. Typically, legacy systems were developed before the widespread use of modern software engineering methods and have been maintained to accommodate changing requirements. These two factors result in systems which are often difficult to understand and expensive to maintain. Many legacy systems thus present a dilemma – such systems are critical to the business process, but maintaining them incurs unjustifiable expense (Bennett, 1995).
2. A legacy system is an antiquated computer system or application program which continues to be used because the user (typically an organisation) does not want to replace or redesign it. Legacy systems are potentially problematic by many software engineers for several reasons. Legacy systems often run on obsolete (and usually slow) hardware, and sometimes spare parts for such computers become increasingly hard to obtain. These systems are often hard to maintain, improve, and expand because there is a general lack of understanding of the system; the designers of the

system have often left the organisation, so there is no one left to explain how it works. Such a lack of understanding can be exacerbated by inadequate documentation, or manuals getting lost over the years. Integration with newer systems may also be difficult because new software may use completely different technologies (McGee, 2005).

3. When software is new, it is very malleable; it can be formed to be whatever is wanted by the implementers. But as the software in each project grows larger and larger and develops a larger base of users with long experience with the software, it becomes less and less malleable and resists change. This kind of system is called a legacy system. Like a metal that has been work-hardened, the software becomes a legacy system, brittle and unable to be easily maintained without fracturing the entire system (Bisbal et al., 1999).
4. A legacy system is an asset to organisations and business requirements of organisations change over time which suggests that a legacy system must become part of the changing business requirements (Srinivas, Ramakrishna, Rao, & Babu, 2016).

A legacy system is a useful system which an organisation would like to keep but is difficult to understand, it lacks business functions that support the evolving organisation and cannot meet business requirements that change over time. Organisations often describe their legacy systems as business assets. To gain the competitive advantage of evolving data, technologies and analytics, legacy systems must become a part of a Big Data solution (Benes, 2018). However, according to Gauger et al. (2017), legacy solutions lack flexibility and carry a significant technology debt due to dated languages, databases, architectures, and a limited supply of aging baby-boomer programmers. This liability prevents many organisations from advancing and supporting analytics, real-time transactions, and a digital experience. However, these legacy systems cannot support Big Data storage, processing, and analysis. Integrating data from existing Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) systems and merging it with newer types of data stores is a necessity to support Big Data analytics and hence should be integrated with Big Data solutions. For the purpose of this research a legacy system is defined as a system which has dated languages, databases, architectures, manual processes of any kind such as data entry, scanner, and inhibits Big Data storage, processing and analysis. Examples of this type of legacy systems are: Learning Management System (LMS); Students Information System (SIS); and Conventional Shopping Coupon System (CSCS).

Jha and O'Brien (2013) have recognized the value of the legacy systems and identified the need to use modernisation of legacy systems to recover and reuse functional requirements, software artefacts, and system components to modernise legacy systems. There is also a need to recognise and to develop an information base for data migration planning which can incorporate Big Data.

2.5 Impact of Legacy System Complexity on Big Data

The impact of legacy systems complexity on Big Data integration can be categorised into Operational and Technical complexities. We briefly describe both in the following sections.

2.5.1 Operational Complexity

Organisations can have several information systems, such as personnel data systems, pay systems, health care systems, procurement systems, and so forth. Mergers and acquisitions of organisations have resulted in acquiring legacy systems which contain invaluable data. Organisations may have separate subordinate level organisation structures to maintain separate, functionally duplicative systems supporting operational requirements. To collect, analyse or process data which would be useful for decision making at higher organisational levels, management often collects the data from lower levels. The collection process depends on subordinate organisations feeding information upward – often manually. The format of the upward feed must be meticulously specified at each level. Outside of limited specific mission critical instances, the consistency and accuracy of data flows has been practically impossible to maintain and control. According to DalleMule and Davenport (2017), cross-industry studies show that on average, less than half of an organisation's structured data is actively used in making decisions – and less than 1% of its unstructured data is analysed or used at all. More than 70% of employees have access to data they should not, and 80% of analysts' time is spent simply discovering and preparing data. The biggest challenge identified for such a low number of unstructured data being analysed is legacy systems. The data from legacy systems is not a perfect fit to be integrated to Big Data. And it does hamper if managers or executives want to make some decision on or near real time. The impact of legacy systems on Big Data integration includes the following:

- The inventory of legacy system updates and documented evidence is difficult to collect and analyse. Even if documentation exists, it is often outdated or of poor quality. In addition, personnel with the required knowledge of the legacy systems and data are often no longer available.
- The existing interface data elements defined for transferring data instances do not represent completely identified, sharable data structure and semantic requirements among systems.

2.5.2 Technical Complexity

From one perspective, legacy systems are time-tested having value proven by long use representing decades of effort and customisation, becoming reliable parts of an overall IT strategy along the way. These entrenched software systems often resist evolution because their ability to adapt has diminished through factors not exclusively related to their functionality. According to Lehman's first Law (Lehman et al., 1985) software must be continually adapted or it will become progressively less satisfactory in "real-world" environments. This is due to the continuous change of user requirements and technical environments. Many legacy systems have been large investments for organisations, and they contain invaluable business logic and knowledge.

Koskinen et al. (2005) completed an empirical study on software modernisation decision criteria and found that the feasibility of a legacy system to be evolved, maintained, and integrated with other systems is improved due to modernisation. The ending of technological support and the expected system lifetime reflect mandatory system modernisation. A data handling system could be an outdated database system, such as a home-grown database management system, and flat file system FILEMAN. Program managed memory overlays was an innovative use of flat-file technology, using a table-driven approach to separate process from data as database technology has done. However, a fixed record length limits the number of fields available to accommodate the growing user requirements. The same field can be used for multiple uses with different meanings depending on the user groups. Allowing individual user discretion rather than enterprise standards, permitted a restrictive physical data limit to serve more customers, but this approach is reaching its design and operation limit as legacy systems cannot be integrated with Big Data. Legacy systems access persistent data in terms of I/O operations on flat files or through accesses to the tables of a relational database.

2.6 Overview of Approaches to Integrate Big Data Solutions with Legacy Systems

We identified several approaches to integrate legacy systems and data with Big Data solutions. The approaches to integrate legacy systems and data with Big Data solutions can be categorised into Data Integration, System Integration, and Business Process Integration.

2.6.1 Data Integration

Data integration is a big and challenging problem, since a lot of business and personal data is stored in large amounts of different computer systems. Also, information is stored in Web-based applications such as intranet portals, blogs, personal web pages, and so on. Bizer et al. (2011) have discussed ontology-based semantic data integration in their work. Ontology-based semantic data integration can be used to integrate structured data, semi-structured data, and unstructured data. Using Linked Data principles (Bizer et al., 2011) as the basis for integration gives a possibility to apply the same techniques for integrating any kind of data. Linked Data describes a method of publishing data on the Web so that it can be interlinked and become more useful. It is based on W3C standard technologies, such as HyperText Transfer Protocol (HTTP), Resource Description Framework (RDF), and Web Ontology Language (OWL), and, according to the authors, this is one of the best solutions for publishing interlinked data on the Web (Bizer et al., 2011). This data integration approach is used for publishing data on the Web. However, the shortcoming of this approach is that it cannot be used for integrating legacy systems with Big Data solutions.

Karma is data integration software developed by the University of Southern California, and is an information integration tool that enables users to quickly and easily integrate data from a large number of different data sources, including databases, spreadsheets, text file formats, Extensible Markup Language (XML), JavaScript Object Notation (JSON), Keyhole Markup Language (KML), and Web Application Programming Interface (API) (Xiao et al., 2017). The user integrates the information through modelling, which is based on the user's choice of the ontology using a graphical user interface. However, too much manual works such as transforming, modelling, and publishing data makes it time intensive.

Dong et al. (2013) focuses on Big Data integration using schema mapping, record linkage and data fusion to deal with the novel challenges faced by Big Data integration. According to Bellahsene, Bonifati and Rahm (2011), schema mapping in a data integration system refers to creating a mediated (global) schema and identifying the mappings between the mediated schema and the local schemas of the data sources to determine sets of attributes that contain the same information. Record linkage has traditionally focused on linking a static set of structured records that have the same schema. In Big Data integration, data sources tend to be heterogeneous in their structure, many sources such as tweets and blog posts provide unstructured text data, and data sources are dynamic and continuously evolving. Data fusion refers to resolving conflicts from different sources and finding the truth that reflects the real world (Dong & Naumann, 2009). The Web has made it easy to publish and spread false information across multiple sources. It is critical to address veracity of data.

2.6.2 System Integration

The principle of this approach is that a system contains the business logic of the enterprise, and the solution lies in preserving that business logic by extending the system's or application's interfaces to interoperate with other newer applications. System integration concerns interoperability of applications on heterogeneous platforms. Vernadat (1996) defines interoperability as the ability to communicate with pier systems and access the functionality of the pier systems. Establishing interoperability means to relate two systems together and remove any incompatibilities in between.

System integration addresses the problems related with the lack of systems' and applications' interoperability in organisations and proposes novel solutions for system integration. However, despite research efforts to date, the proper scientific foundations for system integration remain elusive (Jardim-Gonçalves et al., 2012).

Ulmer et al. (2013) have proposed a pivotal-based system integration approach. According to the authors an organisation must be able to describe and to react against any endogenous or exogenous event. Processes are supported by IT in an organisation. There is a requirement of alignment between business processes, strategy, and applications within the organisation. Organisations need flexibility to reach new opportunities. The processes supported by IT need to be flexible and agile. Authors have proposed a generic approach for modelling and implementing processes and to establish a platform that supports such an approach. This approach allows the control of business processes, from the modelling to their implementation within an information system. The work shows how an operational alignment can be achieved between business processes and IT models. Business processes are shown as business analysis models and do not contain technical logic as done in the implementation model. Analysis models may be inaccurate as they are informal models. Thus, when the IS evolves or is changing, analysis models are no longer updated or in sync with implementation models. Authors have identified that business process re-engineering is an essential approach toenhance the operational alignment and improve the agility of an organisation.

The data integration and system integration as discussed above does not provide the foundation to integrate Big Data to legacy systems.

2.7 Enterprise Architecture and its Role in Big Data Integration

Wamba et al. (2017) have done a literature review of Big Data integration with business processes to improve the understanding of the integration of business process management (BPM), business process re-engineering (BPR) and business process innovation (BPI) with Big Data. Their work focused on analysing and synthesising research published in the period 2006–2016 (182 research papers were analysed) to find and establish the facts about what the researchers in the Big Data community know and do not know about Big Data integration. The authors' work suggests that research in the Big Data community does not address Enterprise Architecture (EA) and its role in Big Data integration. An effective EA can produce many technical and business benefits, like more efficient IT operations for instance, reduced IT infrastructure complexity and faster procurement, to name a few.

To understand the term EA, it is firstly useful to define what is regarded as an enterprise and, furthermore, what is meant by architecture. An enterprise is any organisation or group of organisations which share common objectives that work together to achieve common goals. EA is concerned with the design of enterprises and according to the Open Group's TOGAF, an architecture is a description or detailed plan of an enterprise at component level that guides the implementation thereof (Zachman, 2011). An EA is a conceptual blueprint that defines the structure and operation of an organisation. EA provides documentation and information that supports planning and decision making at various levels of scope and detail, which means it should provide major benefits for management. The intent of an EA is to determine how an organisation can most effectively achieve its current and future objectives (Jha, et al. 2015). EA shows the alignment between IT and business concerns. The main purpose of EA is to guide the processof planning and designing the IT/IS capabilities of an enterprise to meet desired organisationalobjectives.

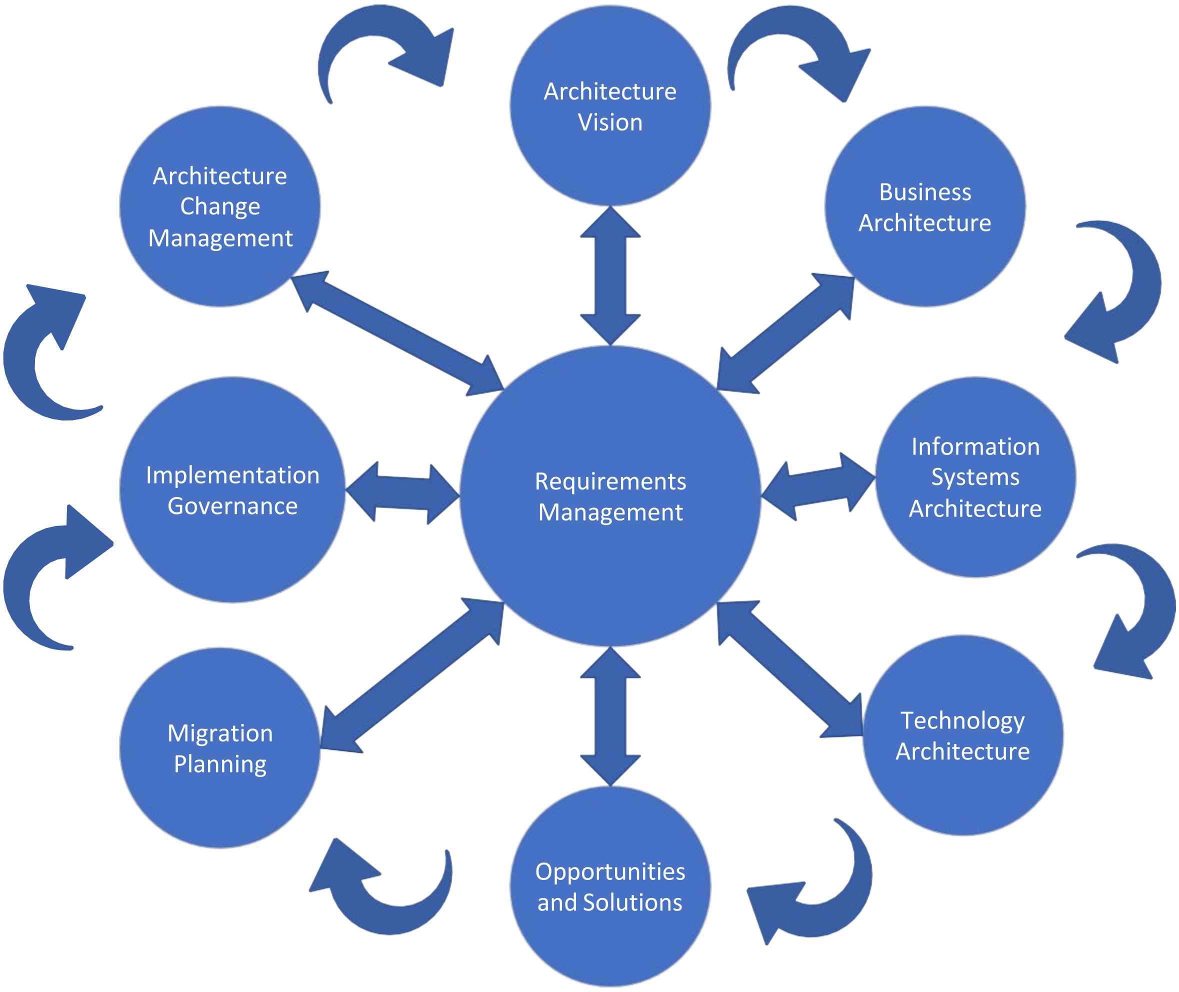

Architecture Vision
Business Architecture
Information Systems Architecture
Technology Architecture
Opportunities and Solutions
Migration Planning
Implementation Governance
Architecture Change Management
Requirements Management

Figure 2.2: Architecture Life Cycle

For organisations to prepare and integrate Big Data into their strategy and structure, the incorporation of data and its impact into the Enterprise Architecture (EA) of the organisation should be investigated (Waller et al., 2013). Enterprise Architecture of an organisation changes with business requirements. The basic structure of the Architecture lifecycle development as defined by TOGAF 9.2 is shown in Fig 2.2 (https://www.opengroup.org/togaf). If an organisation wants to implement a Big Data solution and wants to capitalise on its potential for the business, the first thing all organisations should really understand is where Big Data fits into their business. Like many new information technologies, Big Data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings and support internal business decisions (Davenport, 2014).

Deciding what the organisation wants from Big Data is a critical decision (Jeble et al., 2018) that has implications not only for the outcome and financial benefits from Big Data(Sivarajah et al., 2017) but also the process where it fits within the organisation and what is required to implement Big Data solutions in an organisation and the impacts on the EA (Lněnička et al., 2017). It is important to make sure that Big Data solutions support the business strategy (Watson, 2014). Embedding Big Data solutions in an organisation will have impacts on the Business, Application, Technology, and Data Architectures which are subsets of the architecture represented in the EA.

It is important to capture architectural requirements for Big Data. The Rational Unified Process (Kroll & Kruchten, 2003) gives the following definition of a requirement: A requirement describes a condition or capability to which a system must conform; either derived directly from user needs, or stated in a contract, standard, specification, or other formally imposed document. An architectural requirement, in turn, is any requirement that is architecturally significant. Impact is the action of one object coming forcibly into contact with another. Architectural requirements need to be addressed while capturing and storing unstructured data.

Application Architecture, Technology Architecture, Data Architecture and Business Architecture are the four subsets of EA to address architectural requirements on an organisation. These four subsets of EA are greatly impacted by bringing Big Data solutions into an organisation. An appropriate Big Data architecture design will play a fundamental role to meet the Big Data processing needs.

The Application Architecture focuses on applications within the organisations, variants, versions, and services. Even though Big Data is only possible through technology (Mann, 2012), applications and their versions are one of the important aspects within an organisation that needs to identify if and how an application can be integrated with Big Data solutions (Luna et al., 2014). Application Architecture governs the integration of applications within an organisation. It also suggests which applications can be integrated with Big Data solutions. There could be a possibility that the organisation does not need to integrate all applications with Big Data. It depends on the requirements of the use of Big Data. The first thing is to identify which applications should be integrated and then check if they can be integrated and then how they can be integrated and choose one of the integration options. A Big Data solution will impact many components in the organisation such as Order management system, and Customer Relationship Management (CRM) system. A Big Data solution will impact the way applications and services are configured. A Big Data solution will have an impact on business value propositions, solution/system capabilities, and system architecture requirements (Chen et al., 2017).

The Technology Architecture describes the hardware, software and infrastructure environment that is required to support the development, and host the deployment, of the application components described in the Application Architecture. The Technology Architecture focuses on components, platforms, deployments, technical infrastructures, and hardware requirements. A Big Data solution introduces new technologies in the organisation which may not have been in the organisation previously. To realise the value from Big Data, organisations should use multiple platforms. Which platforms are added depends on the platforms themselves, the applications that use the platforms, and the organisation's maturity in working with the various platforms (Watson, 2014). For example, an organisation might add an appliance to off-load some computationally intensive applications such as predictive, and prescriptive analytics from a data warehouse. Or there may be a need for new types of storage which can support the Big Data processing needs. There should be fast, seamless interaction and collaboration among the different platforms (LaValle, 2011). Applications with Service-oriented Enterprise Architectures in the cloud are emerging and will shape future trends in technology (Salim, 2014). The dynamic and experimental nature of Big Data introduces new technologies such as Hadoop and MapReduce. A Big Data solution is dependent on real-time, streaming, interactive, and machine-learning analytics. This requires changes in the tools and technology infrastructure. The organisation requires high performance computing, heterogeneous multi-provider services integration, new data centric service models, and new data centric security models for trusted infrastructure and data processing and storage. The high velocity data is required to be captured from a variety of sensors and data sources. The processed data is required to be delivered to different visualisation and actionable systems and consumers.

Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organisations (http://www.businessdictionary.com/definition/data-architecture.html). Data Architecture requires information about master and reference data management, data warehousing business intelligence, transaction data management, structured technical data management, unstructured data management, metadata, analytical data, documents and contents, historical data, temporarily data, and Big Data. Data Architecture focuses on data needs, data governance, data quality, ownership and accountability, business objects, data flow and interfaces. Data is critically important to BI and DA. When a strong data infrastructure is in place, applications can often be developed in days. Without a strong data infrastructure, applications may never be completed. IT understands the importance of the data infrastructure, but the business units sometimes assume it is a given and do not fully appreciate what is required to create and maintain it (Watson, 2014). When implementing Big Data solutions many of these areas of the Data Architecture will be impacted (Tekinerdogan et al., 2017). Characteristics of Big Data impacts the technologies associated with it (Oussous et al., 2018). Characteristics of Big Data type allows the organisation to understand its data categories (transaction data, historical data, or master data, for example), the frequency at which data will be made available, how the data needs to be processed (ad-hoc query on the data, for example), and weather the processing must take place in real time, near real time, or in batch mode.

Business Architecture is defined as a blueprint of the enterprise that provides a common understanding of the organisation and is used to align strategic objectives and tactical demands (Zachman, 2011). Business architecture represents holistic, multidimensional business views of capabilities, end- to-end value delivery, information, and organisational structure; and the relationships among these business views and strategies, products, policies, initiatives, and stakeholders (FEAPO) (2013).

According to the Federation of Enterprise Architecture Professional Organisations (FEAPO) (2013), an EA primarily had focused on the technological aspects of organisational change; the practice is quickly evolving to use a rigorous business architecture approach to address the organisational and motivational aspects of change as well. Business architecture provides a foundation for decision within organisations' business functions, organisations' business strategy, locations, processes, and products. Strategic objectives and tactical demands could be used to identify the proper fit of Big Data Solutions (Jha, et al. 2015).

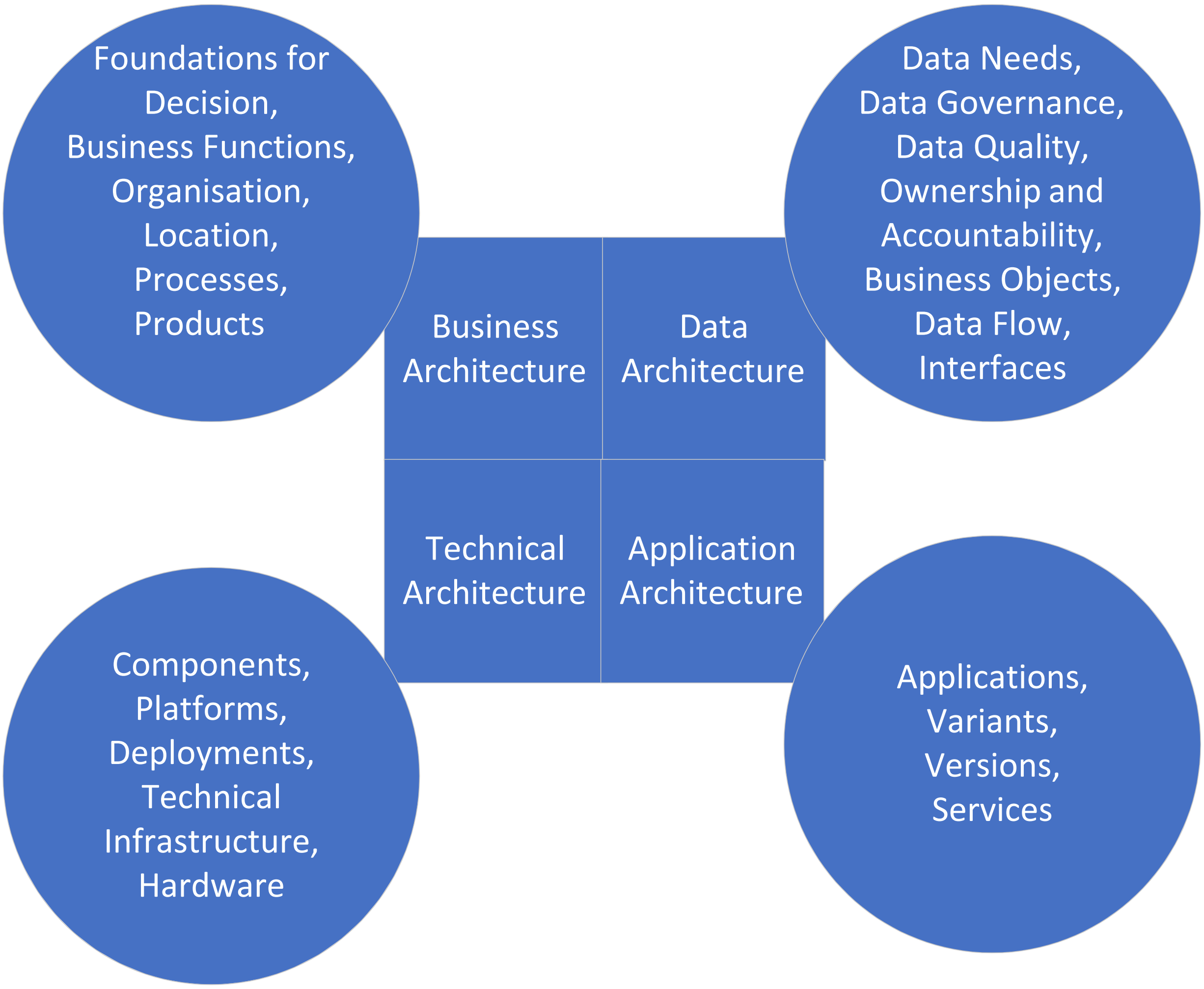

Foundations for Decision, Business Functions, Organisation, Location, Processes, Products
Data Needs, Data Governance, Data Quality, Ownership and Accountability, Business Objects, Data Flow, Interfaces
Business Architecture
Data Architecture
Technical Architecture
Application Architecture
Components, Platforms, Deployments, Technical Infrastructure, Hardware
Applications, Variants, Versions, Services

Figure 2.3: Requirements of Big Data on Enterprise Architecture

Each subset architecture within EA links to the others either directly or indirectly at some point. Business Capabilities and/or Business Processes (Business Architecture), links to the

Applications that enable the capability/process (Applications Architecture – Components Off the Shelf, Custom), links to the Information Assets managed/maintained by the Applications' links to the technology infrastructure upon which all this runs (technology architecture-integration, security, business intelligence/data warehouse, database infrastructure, deployment model). Figure 2.3 shows all the four subset architectures of EA and the requirements of Big Data integration on each architecture (Jha, et al. 2015).

EA plays a major role in ensuring that organisations maximise the business opportunities posed by Big Data. Big Data disrupts traditional information architectures of an organisation as requirements to handle unstructured, high volume, high velocity, and high variety data come into play. There is no architecture currently available that investigates how EA can be incorporated into Big Data for data-driven solutions. The main problem identified is that there exists little discussion within the research or practice literature with regards to the integration of data, Big Data and data science into the EA, specifically when considering data and Big Data as disruptors and the aim to be a data-driven enterprise (Kearny et al., 2016). There is a lack of understanding and any systematic development of a Big Data architecture that incorporates within EA the integration of BD solutions with legacy systems and data.

2.8 Chapter Summary

This chapter has discussed where the problem of Big Data solution comes from, what is already known about the problem of Big Data integration, what the different Big Data integration approaches are, and what is missing in the current approaches. It has also discussed the role of Enterprise Architecture in Big Data integration into legacy systems.

Many approaches to Big Data integration into legacy systems exist. These can be categorised into data integration, system integration, and business process integration. There is a lack of understanding and any systematic development of a Big Data architecture that incorporates within EA the integration of BD solutions with legacy systems and data. There is a need to develop a Big Data architecture for integrating Big Data solutions with legacy systems which uses EA. Integration of Big Data solutions with legacy systems should have foundations on EA as EA is the conceptual blueprint that defines the structure and operation of an organisation. EA provides a strategic context for the growth and change of the IT system in response to the shifting demand of the business environment. In short, Enterprise architecture aligns IT with business. An effective EA can produce many technical and business benefits like more efficient IT operations, for instance, reduced IT infrastructure complexity and faster procurement, only to name a few.

CHAPTER 3: A SURVEY ON BIG DATA: INTEGRATION OF LEGACY SYSTEMS AND DATA WITH BIG DATA SOLUTIONS

3.1 Introduction

This chapter presents review of the existing surveys on Big Data analytics and architectures, and we highlighted the carried out our work. More specifically, the highlights of this chapter are as follows:

- We present an overview of integration of Big Data solutions with legacy systems and data and categorise them according to some identified challenges.
- We highlight the practices related to the use of Big Data solutions with legacy systems.

The motivation and the rationale behind this survey is to find out the challenges organisations are facing arising out of legacy systems to embed real-time or near real-time analytics which can help improving the organisational goals. This survey contributes towards the phase 1 of our research methodology and design as discussed in chapter 1 and contributes towards identifying current issues and concerns of integrating legacy systems and data with Big Data solutions in organisations. We discuss the survey participants and why they are important in our study. Results from the survey are discussed in this chapter. BI and DA that uses Big Data include data storage and processing for business centric practices and methodologies that can be applied to data centric online applications such as e- commerce, market intelligence, e-government, healthcare, education and security. Extensive surveys have been conducted to discuss Big Data analytics, architectures, and its challenges. However, none of the survey focuses on integration of Big Data with legacy systems and how organisations are dealing with this integration (Gani et al., 2016; Singh et al., 2015; Zhang et al., 2018). In our work, we highlight and compare organisations which have integrated legacy systems with Big Data solutions that are not studied in existing surveys. At the end a summary of the chapter is presented.

3.2 Background

Our empirical survey differs from the existing ones by the fact that it considers legacy systems integration with Big Data solutions from a number of different aspects, such as the use of legacy systems and data, and the advantages, disadvantages, and factors influencing legacy systems and data in the business intelligence community. In this section, we highlight the existing eight recent surveys on Big Data solutions, and architectures, and we describe their main contributions. The surveys identified for the study were searched through different journals and scholarly articles and provided original research that further enhanced our understanding about Big Data solutions, architectures, and issues related to these. The keyword search was done on "Big Data survey", "Big Data architecture", and "issues implementing Big Data technology".

Zou et al. (2019) conducted a survey of Big Data analytics for smart forestry and have established the need of Big Data technology in forestry. In their work they have discussed that Big Data will bring greater opportunities for forestry development as the speed and accuracy of forestry data acquisition have been greatly improved with the development of technology. They have pointed out the challenges to organising the massive data reasonably and effectively and to analysing it fast. They have suggested a four-layer architecture model for data storage, query, analysis, and application. Forestry data analysis, based on stand-alone geographic information system (GIS) tools, cannot meet the requirements of speed and accuracy in massive heterogeneous forestry data analysis. The limitation of their work is integrating GIS with Big Data solutions they have proposed as SciDB only provides a C language interface. SciDB is an array-oriented storage model, and therefore is ideal for storing remote sensing data. However, SciDB only provides a C language interface. Hadoop is written in non-C languages, therefore, for distributed computing architectures it is difficult to call the SciDB interface. Data is the basis and prerequisite for the development of forestry Big Data. With the development of Big Data, the current forestry data system will not be able to support computing resources and cannot cope with the massive data. Therefore, an organisation needs to develop a Big Data strategy before they can integrate Big Data solutions with GIS.

Inoubli et al. (2018) conducted an experimental survey on existing Big Data architectures and have provided an experimental evaluation of the Big Data architectures with several representative batch and iterative workloads. They have compared Big Data architecture Hadoop, Spark, Storm, Flink, and Samza based on data formats, processing mode data sources, programming mode, supported programming language, cluster management, comments, iterative computation, interactive mode machine learning capability and fault tolerance, and have provided a list of the best practices with batch and stream processing.

Liu et al. (2014) conducted a survey of real-time processing systems for Big Data based on open source technologies. Their survey focused on system architectures and systems platforms used for real-time/near real-time processing. They have identified that "due to the nature of Big Data, it has become a challenge to achieve the real-time capability using the traditional technologies" such as Relational Database Management Systems and Information Systems.

Acharjya et al. (2016) conducted a survey on Big Data analytics to identify the potential impact of Big Data challenges, open research issues, and various tools associated with Big Data. The authors have categorised Big Data challenges into four broad categories, namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualisation of data; and information security. They have also discussed Big Data tools such as Apache Hadoop and MapReduce, Apache Mahout, Apache Spark, Dryad, Storm, Apache Drill, Jaspersoft, and Splunk. The limitation of their work is that they have not addressed how Big Data tools can be used to integrate Big Data solutions with legacy systems and data.

Oussous et al. (2018) conducted a survey on Big Data Technologies to facilitate the adoption and the right combination of different Big Data technologies according to their technological needs and specific applications' requirements. The authors have categorised the tools according to different layers such as data storage layer, data processing layer, data querying layer, data access layer, and management layer. They have identified which tools can be used in which layer. This brings an understanding of the tools and where they can be used in a Big Data solution.

Khan et al. (2014) conducted a survey on technologies, opportunities, and challenges associated with Big Data. The authors have identified that difficulties lie in data capture, storage, searching, sharing, analysis, and visualisation. According to the authors, the architecture of Big Data must be synchronised with the support infrastructure of the organisation. The authors claim that, to date, all the data used by organisations is stagnant. Data is increasingly sourced from various fields that are disorganised and messy, such as information from machines or sensors and large sources of public and private data.

Tsai et al. (2015) conducted a surveyon Big Data Analytics to develop a high-performance platform to efficiently analyse Big Data. The authors work is more aligned to selecting a data mining algorithm to handle Big Data analytics and relate to selection, pre-processing, transformation, data mining, and interpretation/evaluation. Their work is based on knowledge discovery in databases (KDD) and its operations. The authors discussed the three main processes. The identified processes are input, data analytics, and output. These processes are working on seven operators: gathering, selection, pre-processing, transformation, data mining, evaluation, and interpretation. Big Data analytics systems are designed to work on parallel computing, with other systems which can be hosted on the cloud, or on another search engine. The communication between the Big Data analytics and other systems will strongly impact the performance of the whole process of input, data analytics, and output. There is an infinite computing resource for Big Data analytics with an impracticable plan, an input and output ratio such as return on investment that need to be considered before an organisation constructs their Big Data analytics centre.

Praveena and Bharathi (2017) conducted a survey and gave an overview of Big Data analytics, issues, challenges, and various technologies related with Big Data. The authors provided a Big Data architecture with different layers and what technologies are used in different layers. Big Data challenges reported are with: storage; data representation; data life cycle management; analysis; reporting; redundancy reduction and data compression; energy management; data confidentiality; expendability; scalability; cooperation; and Big Data dimensional reduction.

The above literature review shows that there are existing surveys which discuss Big Data challenges, and Big Data tools. However, none of the identified surveys has listed how organisations are integrating Big Data solutions with their legacy systems and what challenges and issues they are facing in doing so.

In this chapter, we highlight that our survey differs from the above presented works and answers our Research Question 1: What are the challenges in integrating Big Data solution(s) with legacy systems and data?. Our work focuses on, and presents an overview of, the integration of Big Data solutions with legacy systems and data and reports on challenges and issues of doing so. Our survey also deals with what strategies an organisation needs to adapt before implementing Big Data solutions within their organisations. We also highlight the best practices used within the organisations related to the use of Big Data solutions with legacy systems.

3.3 Survey Participants

Our survey on identifying existing approaches towards Big Data integration with legacy systems and data was conducted between 2018 and 2019. We selected 210 participants from various industries to participate in our survey. The list was selected using internet browsing with key words "Big Data", "integration with legacy systems", and "Data Analyst". We also looked in journal and conference proceedings for people who are active in Big Data research. We had 97 respondents among 210 participants from different organisations, received responses from industries such as financial services, healthcare, aviation, higher education, energy sector and insurance.

Cluster sampling is used to accomplish the task and population respondents were divided into groups of financial services, healthcare, aviation, higher education, energy sector and insurance. The clusters were selected systematically to reflect on the sectors respondents were working in. Several reasons for choosing the participants are listed below.

Financial services institutions such as banks are data-intensive organisations. Banks are relying on information gained from data analytics for cyber security scams. Data analytics equip the banks with information on which they can make informed decisions about their financial assets (Kharote et al., 2014). Big Data is being used in security and fraud detection within the banking sectors (Wongchinsri et al., 2016). Researchers have been exploring advanced Big Data techniques for effectively identifying unusual fraud behaviour within the banking sector in order to maintain the high standards of security with the overwhelming flow of big banking data and the rapidly growing scale and complexity of cybercrimes.

Insurance fraud has been identified as one of the major crimes because of the digital revolution that is impacting society. Policyholders are hiding information from insurance companies. There are many cases identified of incorrect age, non-disclosure of pre-existing diseases, non-disclosure of medical history, hiding previous claims, providing false information regarding self or any other family member, frauds in physicians' prescriptions, false documents, false bills, and exaggerated claims (Dora & Sekharan, 2013). Paying off illegitimate claims are a huge cost to the insurance organisations.

The aviation industry in the recent years has undergone a drastic growth, resulting in increased passenger flow and air traffic. This sudden growth makes it difficult to manage operations and to ensure passenger and cargo safety. Most of the data collected (onboard sensors, ground stations, satellite sensors) goes unused unless an abnormality is found wherein the person responsible is notified. The data generated cannot be handled by legacy systems. An average trans-Atlantic flight generates around 1 Terabyte of data. Extracting useful information from this Big Data is an effective way of improving management and ensuring safety by increasing revenues and keeping expenses in check (Sumathi et al., 2017).

Energy efficiency is becoming one of the major concerns of a green and sustainable society. With climate change movements around the world, organisations are working towards generating green and clean energy and that has attracted increasing research and development efforts in recent years (Fan et al., 2015; Zhou et al., 2015). Data generated within a single household can be analysed to develop an energy efficient society. Understanding individuals' household energy consumption behaviour can be utilised as information to improve energy efficiency and promote energy conservation in the society making it energy sustainable (Cooper et al., 2013; Koseleva et al., 2017; Mashayekhy et al., 2017).

Higher education institutions are increasingly interested in measuring, demonstrating, and improving performance in education. According to Norris et al. (2009), analytics enables us to engage in a process of data assessment and measurement and is aimed at improving the performance of individuals or institutions. Big Data in higher education has an ability to transform the business processes within the higher education sector. Big Data analytics can be utilised to transform the academic rigor, administrative workload, students learning, and timely help required by students (Baer et al., 2011). Big Data can also contribute to policy and practice outcomes in helping institutions address contemporary challenges (Atif et al., 2013). Big Data analytics in the context of education is called Learning Analytics (LA), and refers to the process of collecting, evaluating, analysing, and reporting educational data for decision making. LA is a multidisciplinary approach that makes use of existing techniques such as data processing/science, artificial intelligence, technology-learning enhancement, business intelligence, educational data mining, text mining, statistics, mathematics, and visualisation (Scheffel et al., 2014). LA is focused specifically on students' learning processes and their learning behaviour (Greller & Drachsler, 2012). The data sources required for LA are Learning Management Systems (LMSs), Student Information Systems (SISs), and any other system with which students are interacting. The data analysed from these systems will develop knowledge construction, creativity, self-directed learning, sense of community, and assessing academic progress based on assessment and structured activities (Bienkowski et al., 2012; Dawson, 2011). LA addresses a student's behaviour and the issues of a student's progression and attrition using student's data.

Research organisations collect and use data. Data generated from online activities can portray customers' behaviour accurately. Customers leave a digital footprint while working online. This information can be exploited for research purposes. Data collected in a passive, unobtrusive, and non-reactive manner may overcome social desirability bias and other distortions that occur when researchers directly interact with research subjects (Webb et al., 1966). Detailed data generated as a non-reactive by-product of digital actions can complement data collected by researchers through traditional methods of surveys and interviews (Johnson et al., 2019).

In Australia, the supermarket industry is facinginternational pressure with Aldi and Costco opening their stores in Australia. Aldi entered the market in 2001 opening 470 stores across Australia and spreading its wings in the South Australia (SA) and Western Australia (WA) markets in 2016. Costco is another international retailer which entered the Australian market in 2009 opening eight warehouses in the country. Big Data has served retailers to better understand customer behaviour. According to Akter et al. (2016), the three main types of Big

Data utilised to understand consumers' behaviour in grocery purchasing are scanned items, credit card history and purchases, and electronic funds transfer at point of sale (EFTPOS) and purchase. Consumers shop at different retailers and the opportunity Big Data offers is to better understand the consumers and improve consumer loyalty. Over many years Coles and Woolworths have invested in Big Data to ensure they are able to make fact-based decisions (Akter et al., 2016).

Service providers are using Big Data to better gauge customers' needs. The new generation of customers are well informed. Service providers are challenged with what kind of services to offer to customers. The data streams are available for both providers and customers. Managing the new generation of informed customers is increasingly complex and challenging for the service providers as they need to balance the needs of their organisations and customers. The positive impact of data analytics on service quality, customer satisfaction, and firm performance has been well established (Devi et al. 2013; Sood et al., 2010).

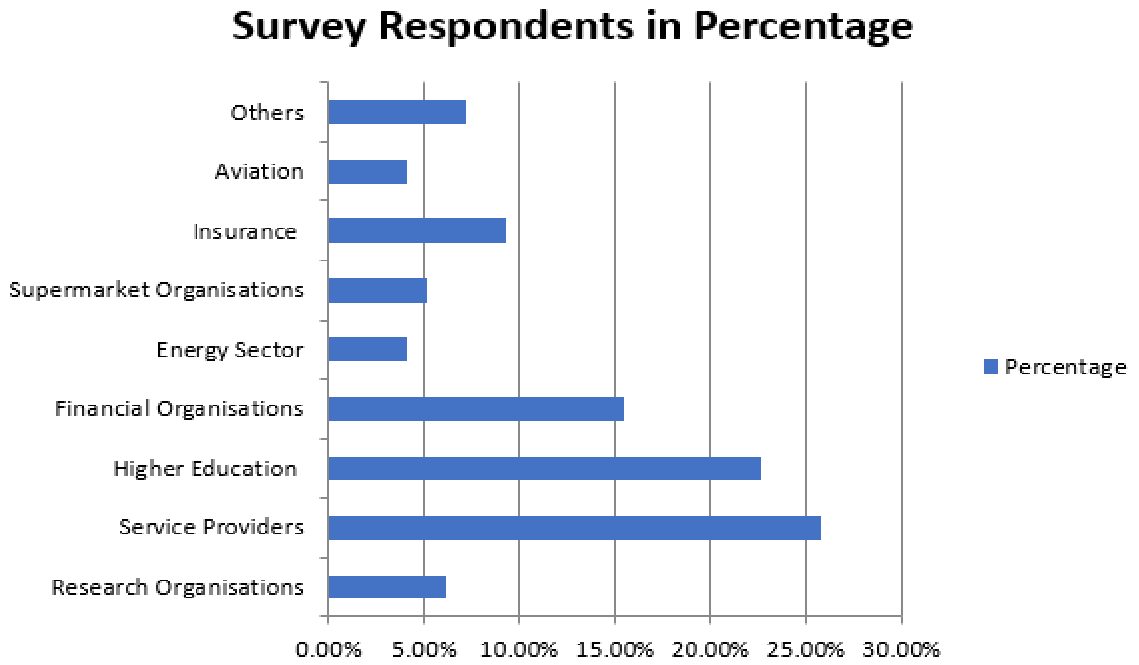


Figure 3.1: Survey Respondents in Percentage

Figure 3.1 shows graphical representation of survey respondents. The survey we conducted had the most respondents working in service provider organisations followed by higher education. We could not get many participants from the energy sector or aviation sector; however, these are still particularly important subjects for Big Data and how integration with their legacy systems can be achieved. We divided our survey Respondents into nine sectors:

Others, Aviation, Insurance, Supermarket Organisations, Energy Sector, Financial Organisations, Higher Education, Service Providers, and Research Organisations. Others constitute organisations which could not be classified into the rest, such as Legal firms, Astrological firms, and Central Link.

3.4 Survey Structure and Questions

The survey was divided into four subsections with 31 questions in total. The survey questionnaires and information sheet for the participants can be found in Appendix A. The survey questions were formulated based on the literature review and in consultations with supervisors so the data collected from this survey would help us understanding issues of Big Data solution, and help us designing a Big Data architecture to address the identified problems. The details of each sections are:

1. Section 1: Organisational Questions (1–3). This section captures answers to find out if Big Data projects are running within the organisation or not. This section also captured the name of the organisation, division, or department and if any Big Data projects are happening in the organisation. We also captured answers to what kind of analytics are being used in the organisation.
2. Section 2: General Questions (4–7). This section captures answers on the educational level, role, and experience level of the respondents.
3. Section 3: Legacy Systems Issues and Concerns Questions (8–13): This section captures answers about legacy systems and data, advantages, disadvantages, and factors influencing legacy systems and data in the business intelligence community.
4. Section 4: Big Data Initiatives and Implementation Questions (14–31). This section captures answers about finding the proper fit of a Big Data solution and technologies for an organisation. This section also gathered information regarding Big Data activities, the integration of Big Data solutions with legacy systems. The survey was targeted specifically where Big Data could be of use.

3.4.1 Section 1: Big Data Organisational Questions

This section captures the name of the organisation, division, or department and if any Big

Data projects are happening in the organisation.

Of the respondents, 27.83% believed that Big Data projects are running in their organisation. And the rest of 72.16% of the respondents believed that they are not running any Big Data projects in their organisation. Among the organisations contributing, 27.83% are financial organisations, research organisations, insurance companies, and some of the service providers. Figure 3.2 shows the organisations with Big Data projects currently running within their organisations.

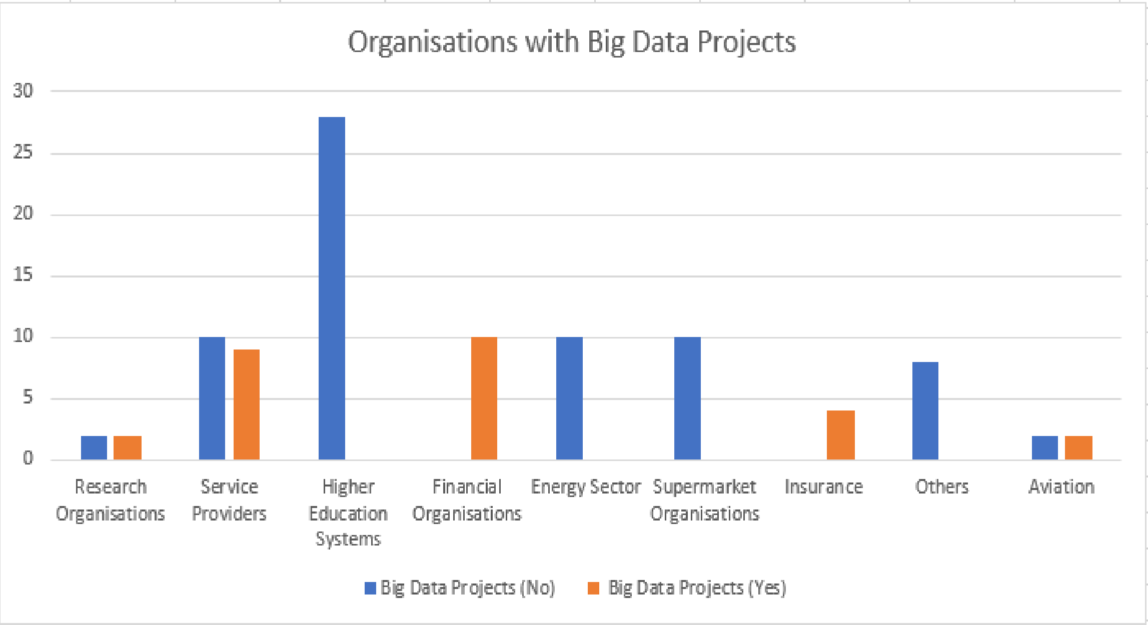


Figure 3.2: Organisations with Big Data Projects

3.4.2 Section 2: General Questions

This section captures answers to general questions about educational qualifications and years of experience our respondents have. To understand how our respondents perceived Big Data and its implementation we required to know the respondent's role they played in the organisation. The statistics shows that we had a good cohort of respondents playing important roles in the organisation around areas such as data analytics, decision making and business intelligence. Figure 3.3 shows the respondents role in percentage.

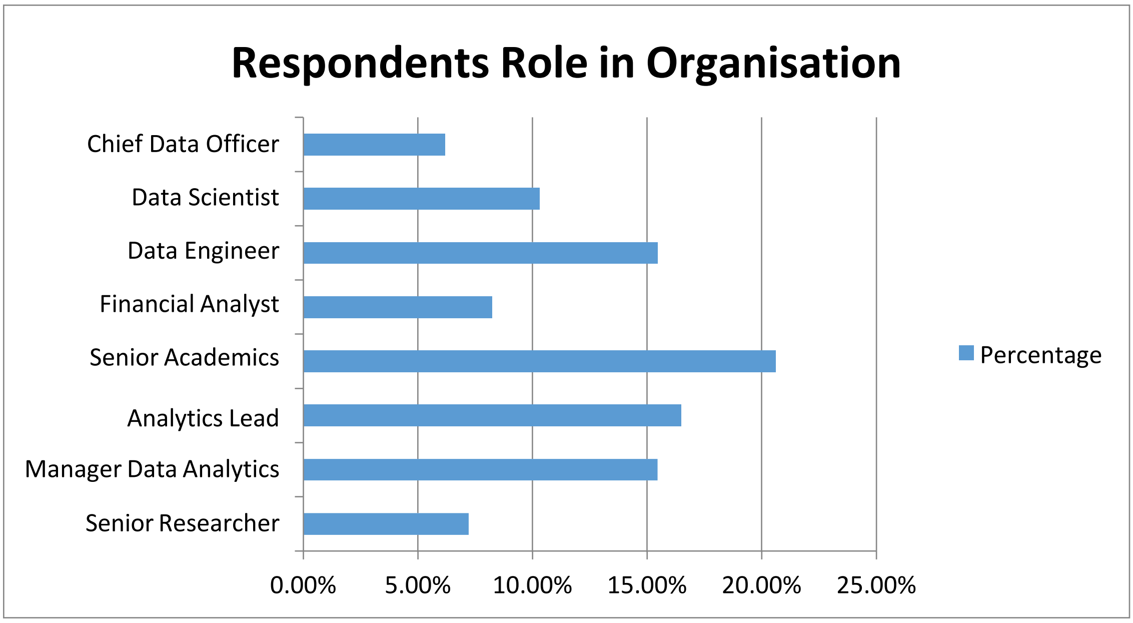


Figure 3.3: Respondents Role in Organisation

Of the survey respondents, 26.80% had an educational level of postgraduate in Computer Science or Information Technology. Even though our respondents have many years of experience, they did not receive any formal degree in Data Science. However, they have evolved with this new technology called Big Data.

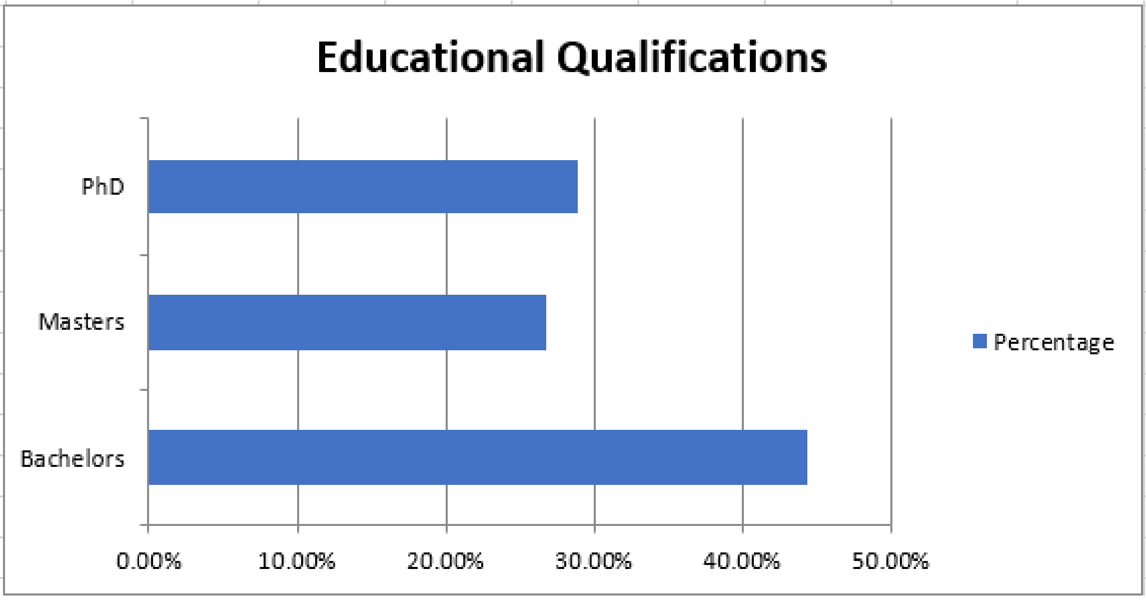


Figure 3.4: Educational Qualifications

The proportion who had a PhD in Information Technology or Computer Science was

28.86%. We did not specifically ask the age of the participants as this for some people may have been perceived as a violation of their privacy. However, we asked the years of experience they had in the Computing and Information Technology field. The survey result shows the years of experience of the respondents ranges from 10 to 20 years. The years of experience of our respondents show that we have a more mature population working in legacy systems. Figure 3.4 shows our respondents qualification level.

Areas of work: Figure 3.5 shows areas of work that our respondents are involved with.

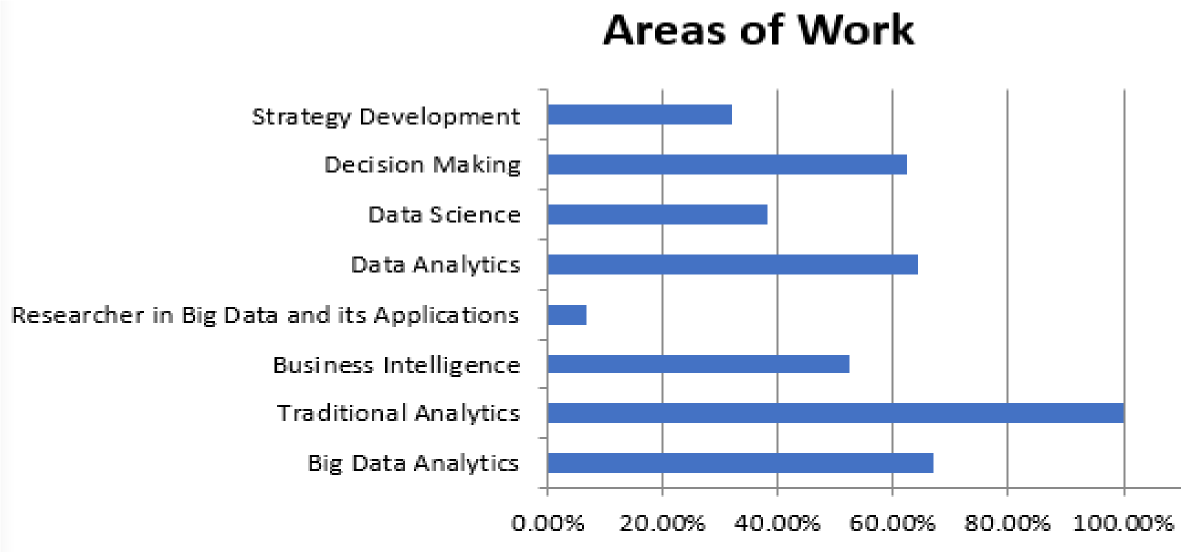


Figure 3.5: Areas of Work

Most of the respondents were involved in more than one kind of work area. One hundred per cent of the respondents are working in traditional analytics. This means that 100% of the organisations use traditional analytics for report generation or decision making. Of our respondents, 66.80% are from the area of Big Data Analytics followed by data analytics (62%) and business intelligence (56%). One hundred per cent of our respondents use legacy systems to generate reports for decision making or business intelligence.

3.4.3 Section 3: Legacy Systems Issues and Concerns Questions

This section captures answers about legacy systems and data, advantages, disadvantages, and factors influencing legacy systems and data in the business intelligence community. The following section provides a description of the survey results.

What kind of data does your organisation have?

Most of the organisations we surveyed are working on organisational transactional and historical data. In fact, none of the organisations admitted to integrating social media data to their existing business intelligence infrastructure. One hundred per cent of the respondents from financial organisations believed that they have good practices in place to manage data, processes, and infrastructures for detecting fraud. One hundred per cent of the respondents believed that data is the biggest asset of their organisation.

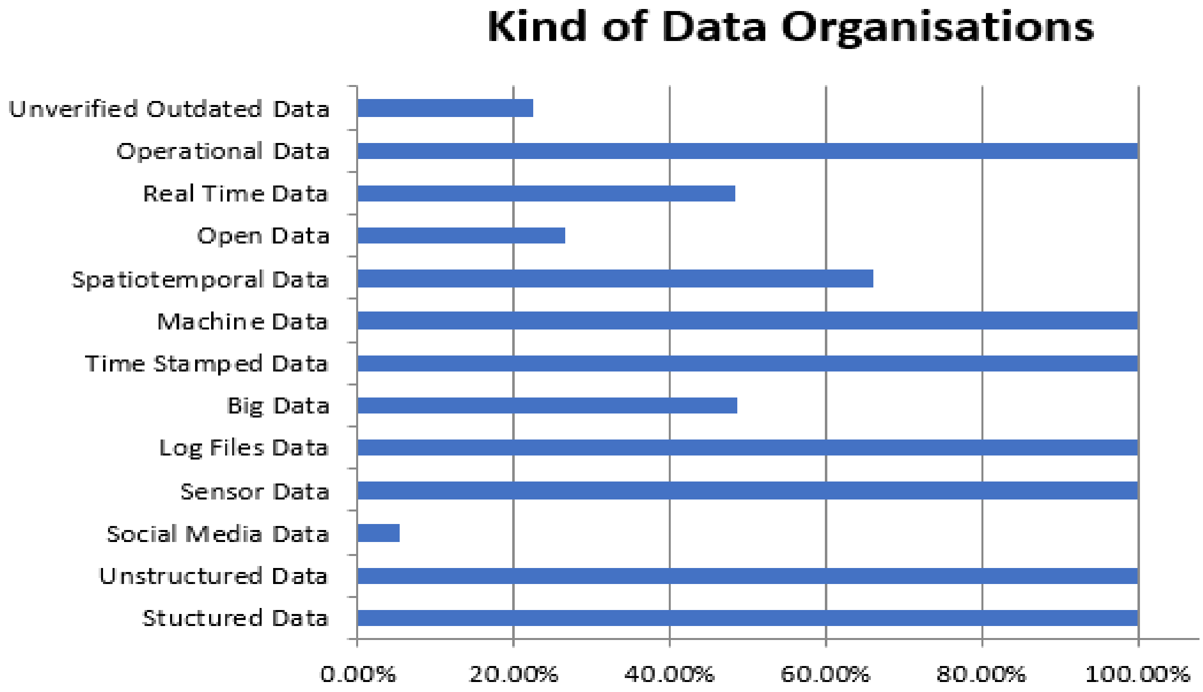


Figure 3.6: Kind of Data Organisations

Out of the respondents, 48% believe that the data generated within the organisation fits into the characteristics of Big Data and if analysed effectively can provide a competitive edge to the organisation. One hundred per cent of respondents said that the drawback is the existing technology that inhibits the organisation to exploit this volume of data. Our survey identified that the kinds of data used include structured data, unstructured data, sensor data, log files data, big data, time stamped data, machine data, spatiotemporal data, open data, real-time data, operational data, and unverified outdated data. Figure 3.6 shows the responses on the various kinds of data from our respondents.

Our survey identified that all organisations have structured, unstructured, sensor, log file, time stamped, machine, and operational data. However, only 5.21% have social media (such as Facebook, Twitter, and LinkedIn) data. Of respondents, 62.3% said that there are unverified outdated data sitting within the organisation. This data has been collected and stored within the organisation and people have no idea about what kind of data it is, its relevancy, or whether it can be put to any use.

What do you think are the benefits of legacy systems and data in your organisation?

Most of the respondents believe that their organisations are using legacy systems for dayto day operations. People are comfortable using these legacy systems as they are familiar with them. One hundred per cent of the respondents believed that business continuity is important, and a legacy system that works and keeps everyone on the same page is good for the organisation. Of the respondents, 64.94% believed that using a legacy system is easy to manage and within the control of the organisation. One hundred per cent of the respondents believed that using existing systems is less complex as over time people have gained the confidence in using them. Data is being retrieved from legacy systems for business intelligence and decision making. However, the insight of the existing data has become more meaningful with the advancement in tools and technologies. One hundred per cent of the respondents believed that there is a need for integrating Big Data with their legacy systems.

What do you think are the main disadvantages with reporting when using legacy systems and data?

Of the respondents, 63.91% believed that data silos are one of the identified biggest challenges to solve and deprives the power of Big Data. These respondents believed that data stored in different data sources, different data systems, and different organisational units that have nothing to do with one another result in no complete insights being generated from the available data because it isn’t integrated on the back end.

Of the respondents, 56% believed that existing systems and data cannot fulfill the requirements of the changing technology which should be treated as of prime importance if organisations need to survive the competitive edge in today’s world. One hundred per cent of respondents believed that business requirements change continuously, and organisations should be flexible to adapt to it. The disadvantages of legacy systems were identified as: costliness, being outdated, limited flexibility, cannot generate reports in real time, store only organisational data in a structured format, immobile systems, and data integrity problems. Of the respondents, 57.73% believed that their customer contact information is incorrect. One of the respondents mentioned that "if you've got a database full of inaccurate customer data, you might as well have no data at all". The proportion of the respondents who believed that data silos can be eliminated by integrating data was 57.73%. They also believed that in their organisation, accessing legacy system data takes a long time.

Where and when do you use legacy systems and data for decision making?

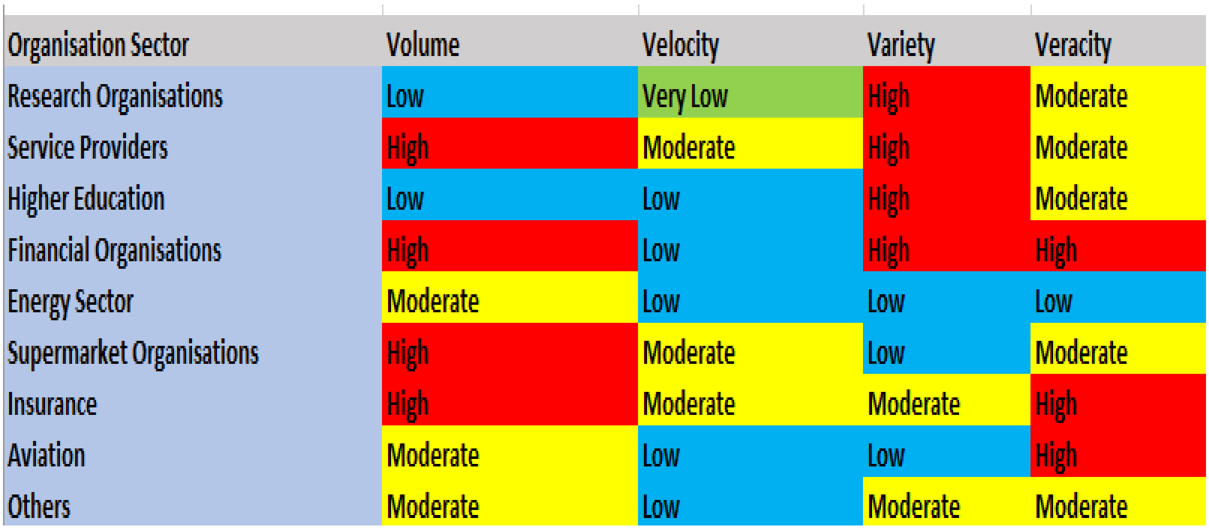
The data are used for historical analysis and operational analysis to understand how to improve the future based on historical evidence. Historical data is analysed to have better insights about processes or organisation. Data is analysed for descriptive and predictive analytics. Respondents want to have prescriptive analytics; however, they do not have the skills set, tools and technologies to run prescriptive analytics. Our respondents require the use of legacy systems for five key themes: information access; insight; foresight; business agility; and strategic alignment. All the respondents and their organisations are using data generated from systems such as enterprise resource planning systems, attendance tracking systems, and e-commerce systems. The data generated from these systems are stored in data warehouses, data marts and database management systems. These technologies have existed in various forms for years. These are large amounts of data and our respondents believe that this data fits into the meaning of Big Data. However, Big Data is not only about storing and retrieving semi-structured and unstructured data. One of the respondents believed that "extraordinary adaptive use of Big Data is triggered by major changes in the business processes and the way users perceive those".

3.4.3 Section 4: Big Data Initiatives and Implementation Questions

This section captures answers about finding the proper fit of a Big Data solution and technologies for an organisation. The following section discusses the results we found from the survey.

In your opinion what is the range (high; moderate; and low) of data processing in your organisation in relation to volume, velocity, variety, and veracity?

Of the respondents, 53.6% believe that their organisations are contributing towards one or the other stated characteristics of Big Data (Volume, Velocity, Variety and Veracity) in the range of High, Moderate, and Low. Figure 3.7 shows the responses from our respondents about data volume, velocity, variety, and veracity in their organisation.



Organisation Sector	Volume	Velocity	Variety	Veracity
Research Organisations	Low	Very Low	High	Moderate
Service Providers	High	Moderate	High	Moderate
Higher Education	Low	Low	High	Moderate
Financial Organisations	High	Low	High	High
Energy Sector	Moderate	Low	Low	Low
Supermarket Organisations	High	Moderate	Low	Moderate
Insurance	High	Moderate	Moderate	High
Aviation	Moderate	Low	Low	High
Others	Moderate	Low	Moderate	Moderate

Figure 3.7: Range of Data Processed in Organisations

Does your organisation have information management Big Data and analytics capabilities?

Of the respondents, 72.16% believed that their organisational data is measured in gigabytes and believed that applying new tools for business intelligence will help their organisation. However, they also believed applying new tools should not disrupt their existing running systems. Organisations do require different types of analytics for different purposes so the new technology should help them with these changing requirements. Organisations having gigabytes of data do not have any Big Data project in their organisation. However, data analytics is enhanced using machine learning to provide more insight into the data and make use of the data that the organisations are already collecting. Of the respondents, 27.83% believed that their organisations have Petabytes of data and Big Data projects are underwayto identify real time fraud and detection. These organisations include financial, insurance, and service providers. The aviation industry is establishing an architecture to use Big Data. Higher education systems are using Learning Analytics on already collected data. This fits into Big

Data as analysing discussion forums and emails are using unstructured data. The proportion of the respondents' who stated that Big Data has primarily been used to drive profits was 70.1%. Big Data analytics can provide deep insights into customer behaviour. Big Data help in gaining a 360° view of their customers, by analysing and integrating existing data. One of the respondents stated that "Big Data analytics is all about understanding the customer, and that means harnessing all resources not just analysing all contacts with the organisation, but also linking to external sources such as social media and commercially available data. For the digital supply chain, it is about collecting, analysing, and interpreting the data from the myriad of connected devices".

Do you think that integrating a Big Data solution will benefit your organisation?

One hundred per cent of the respondents believed that integrating Big Data solutions will bring benefit to their organisation in different forms. This question had interesting answers where higher education respondents highlighted that contact cheating, which is a form of fraud, can be detected using Big Data analytics. Fraud detection is common in financial, service providers, and insurance organisations. One of the respondents stated that "if you can obtain all the relevant data, analyse it quickly, surface actionable insights, and drive them back into operational systems, then you can affect events as they're still unfolding".

Has your organisation developed a Big Data Strategy?

Organisations with gigabytes of data have not developed any Big Data strategy within their organisation. They are using advanced analytics, pattern recognition and deep learning to identify the causes of problems. Organisations with Petabytes of data have a Big Data strategy in place as these organisations are already working on a Big Data architecture. However, their reporting system is not integrated to their legacy systems. According to our survey, 86.62% of the 27.83% of the respondents believed that Big Data architecture implementation requires organisational effort. These respondents believed that one of the disruptive facets of Big Data is the use of a wide range of Big Data tools and technologies for innovative data management to support different analytics.

What kind of analytics is used in your organisation?

One hundred per cent of the respondents believed that they are using descriptive analytics, 53% believed that they are using predictive analytics, and 27% believed that they are using prescriptive analytics. Organisations on the forefront of money management are using prescriptive analytics. The ecosystem of Big Data is very daunting and confusing; 10.3% of the respondents believed that there should be a guideline with requirements on how to useBig Data tools, technology, and architecture. One of the respondents stated that "the most practical use cases for Big Data involve the availability of data, augmenting existing storage of data, as well as allowing access to end-users employing business intelligence tools for the purpose of the discovery of data".

How do you generate reporting for business intelligence?

One hundred per cent of the respondents are using legacy systems such as Customer Relationship Management (CRM), Supply Chain Management (SCM), and Learning Management System (LMS) to generate reports for business intelligence. Of the respondents, 27% believed that they are using a Big Data architecture to generate reports. However, the architecture does not include data from legacy systems. One hundred per cent of the respondents believed that they use legacy systems and data for business intelligence and decision making.

What approaches does your organisation use to integrate legacy systems and data? What problems are associated with it?

One hundred per cent of the respondents believed that there is no architecture implemented to integrate Big Data solutions with legacy systems. However, they are generating reports from different information systems, and combining the reports for final analysis. Respondents cited a lack of experience slowing project progress (48%), struggling to keep up with new data sources (58%), and issues with constantly changing business requirements (44%) as their top challenges.

When making a business decision in your organisation what do you mostly rely on?

Our survey identified that business decisions are made at different levels and they can be classified as: functional level, business unit level, and corporate level. At all three levels people rely on data and reports generated by existing organisational systems. As identified by our respondents, 100% of the respondents believed that information is the key success factor influencing the decision-making process. Of the respondents, 19.58% believed that Big Data analytics is integrated into the decision-making process. The remaining 80.42% of respondents believed that Big Data analytics is not used in the decision-making process and 78.52% of the respondents believed that Big Data analytics should be used in the decision-making process.

Do you see value in integrating Big Data solutions into legacy systems and data in your organisation?

One hundred per cent of the respondents believed that integrating a Big Data solution with legacy systems and data in their organisation will bring benefit to their organisation. However, 87.62% of respondents also stated that their organisation must have a Big Data strategic plan for integrating data from multiple data sources. This is required for Big Data integration for receiving holistic information from different data sources including legacy systems and data. Integrating new datasets into existing pipelines (72% of respondents) were cited as the primary obstacles to Big Data projects. This was shown as the biggest concern that would hinder an organisation's progress towards a Big Data solution. These respondents believed that their organisation had hundreds of systems. This means that the data must be extracted from many different sources, and the volumes could be overwhelming. Besides the volume, the variety of sources also needs to be considered for integration purposes.

Does your organisation use any Big Data technology? Please specify the technology in the text box below.

Of the respondents, 12.38% said that their organisation is using Big Data technology and tools. The most common tools used are Hadoop, Apache Spark, Apache storm, Cassandra, RapidMiner, MongoDB, R programming, and Neo4J. The proportion of respondents who believed that Hadoop is not suitable for analysing social networking data was 3.09%. With large volumes and graph-related issues like social networking or demographic patterns, Neo4j, a graph database, may be a better choice. Neo4j is one of the Big Data tools that is widely used as a graph database in the Big Data industry.

What is the biggest challenge in your organisation for collecting, accessing, storing, processing, and analysing data?

While Big Data offers many benefits, implementation of Big Data has many challenges too.

The Big Data landscape is massive, making it even more challenging and complex for organisations to implement Big Data solutions. Business users do not have enough understanding and knowledge of how Big Data can be utilised within their organisations. Some of the commonly identified issues include inadequate knowledge about the technologies involved, data privacy, and inadequate analytical capabilities of organisations. Of the respondents, 85.56% believed that organisations lack the skills of Big Data implementation in the workforce. Employees are not trained enough to handle Big Data technologies with confidence. Not many people are trained to work with Big Data, which then becomes an even bigger problem as people are not trained to work with Big Data. Volumes, velocities, and varieties of data are growing continuously giving organisations a lot of opportunities to unfold the truth hiding behind the raw data available. Of the respondents surveyed, 87.62% are looking to increase their data team headcount to support a Big Data solution, but 85.56% also say it is difficult to find professionals with the right skills and experiences within Big Data. Organisations are struggling to satisfy the requirements of implementing Big Data.

What are your goals of adopting Big Data projects?

Our respondents believed that there are several goals for adopting Big Data projects within their organisation. Following are the listed goals for adopting Big Data projects according to our identified nine organisational categories.

- Research Organisation: Goals are to collect, process, and analyse a high variety of data to generate more insight from the data.
- Service Providers: Goals are to collect, process, and analyse a high volume and variety of data so that consumers' insights can be utilised, and recommendations can be built.
- Higher Education: Goals are to collect, process, and analyse a high variety of data so that it can be used to enhance good practice measures in learning and teaching. Educators and learners should be able to take control on informed decisions. Teacher's performance can be fine-tuned and measured against student numbers, subject matter, student demographics, student aspirations, and behavioural classification.
- Financial Organisations: Goals are to collect, process, and analyse a high volume and high variety of data for early fraud detection and mitigation and anti-money laundering.
- Energy Sectors: Goals are to collect, process, and analyse data from smart meters so

that energy consumption can be analysed for improved customer feedback and better control of utilities use.

- Supermarket Organisations: Goals are to collect, process, and analyse data to optimise staffing through data from shopping patterns, and local events.
- Insurance: Goals are to collect, process, and analyse data derived from social media, GPS-enabled devices, and closed-circuit television (CCTV) footage for fraudulent claims.
- Aviation: Goals are to collect, process, and analyse data to strengthen the customer value, relationship, and customer loyalty.
- Others: Goals are to collect, process, and analyse data for optimising resources within the organisation.

If you do not have Big Data Architecture in your organisation, how beneficial will it be for your organisation? Do you see any value in having Big Data Architecture in your organisation? Out of the respondents, 12.38% believe that they do have a Big Data Architecture within their organisation and 87.62% of the respondents believed that they do not have any Big Data Architecture within their organisation. However, 87.62% do believe that a Big Data Architecture will surely be beneficial to their organisation as an architecture provides structure to achieve long term success. A Big Data Architecture is about structure, technology, capabilities, and skilled people. The respondents believed that a Big Data Architecture can provide a structure for organisations that want to start with Big Data or aim to develop their Big Data capabilities further. Respondents also suggested that the Big Data Architecture should be vendor independent and can be applied to any organisation regardless of choice of technology, or tools. It should be able to provide a common reference model that can be used by any organisation depending on their requirements of Big Data analytics and solutions. One hundred per cent of the respondents believed that having a Big Data architecture will add value where organisations are struggling to embed a successful Big Data solution in their organisation. Of the respondents, 87.62% also believed that a Big Data Architecture will be useful in integrating a Big Data solution with legacy systems as it will dictate the architecture of Big Data and will help in developing a Big Data strategy.

3.5 Analysis of Survey Results

The analysis of our survey is divided into the following seven points:

- Big Data solution integration with legacy systems and dealing with silos and data from across different parts of an organisation (Variety): Organisations have different kinds of data generated from different systems on which decisions are made. Respondents believed that their organisations make decisions on structured (100%); unstructured (100%); social media (5.21%); sensor data (100%); log files data (100%); Big Data (48.51%); time stamped data (100%); machine data (100%); spatiotemporal data (66.14%); open data (26.60%); real time data (48.23%); operational data (100%) and unverified outdated data (22.30%). Of the respondents, 63.91% believed that data silos deprive their organisation of the power of Big Data. These respondents believed that data captured in different data sources cannot provide holistic views on which an informed decision can be made. Respondents believed that data silos with other data sources should be integrated so that an organisation can gather insights from it to make data-driven decision. Of the respondents, 57.73% believed that their customer contact information is incorrect or not up to date. Data silos are identified as one of the reasons for having different data stored in different systems giving rise to inaccurate data. One hundred per cent of these respondents believed that data silos can be eliminated by integrating data. Data integration will produce a single, unified view of an organisation's data. Business users for BI & DA applications can access this unified view to develop an actionable plan based on the organisation's data assets. Legacy systems are considered as the lifeblood of organisations running through different systems such as Supply Chain Management (SCM), Human Resources System (HRS), Customer Relationship Management (CRM), and Learning Management System (LMS). Legacy systems contain significant and invaluable business logic from the organisation. These business logics are particularly important for the running of the organisations and organisations cannot afford to throw them away. Legacy systems cannot be replaced as that has many other challenges associated with it. Legacy systems represent many years of change in organisational processes. These legacy systems are assets of the organisation. Redevelopment of these systems would be unaffordable in terms of time, costs, and the required human resources (Weiderman et al., 1997). Simple replacements of these systems may be infeasible or impractical because of the scope of previous investments (Jha & O'Brien, 2014).

Since they are vitally important for the organisation, they need to be evolved into new technology environments and run on modern platforms (Bisbal et al., 1999). Legacy systems cannot be replaced because of obvious reasons and hence Big Data solutions need to be integrated into legacy systems.

- **Data and processes are not scalable:** Data and processes are not scalable and cannot be cross referenced. Data redundancy exists along with an inability of systems to share data with each other. Vital information is lost as a result. Legacy systems support certain types of reporting and could supplement legacy system data with other data sources, which will improve the insights. Business decisions become totally dependent on accessing legacy data. Business users cannot access legacy systems directly, so the request to extract data from legacy systems must go through the IT department for the processing. This takes a long time as too many processes are involved. Organisations need to reduce these unwanted processes by integrating and eliminating data silos in today's world of BI.
- **Volume and Velocity of legacy data as opposed to Volume and Velocity of other types of Big Data:** One hundred per cent of the respondents believed that the data in legacy systems are typical relational data from enterprise applications such as CRM and SCM, which are very structured. Respondents also believed that the legacy data tends to be on-premises, behind firewalls in a bounded and constrained infrastructure so external security and data management are not an issue and were never considered while developing legacy systems. One hundred per cent of the respondents believed that legacy systems were not built to accommodate today's different variety of data as opposed to Big Data. All of the respondents believed that a Big Data source has been identified as having characteristics such as frequency, volume, velocity, type, and veracity of the data. All respondents unanimously believed that processing, accessing, visualising, and storing Big Data is complex. Organisations need to consider many dimensions such as: where the data is coming from, who is the owner of data, and how data can be shared. Policies, structure, procedures, and governance need to be in place before Big Data can be processed. One hundred per cent of the respondents believed building an appropriate architecture for Big Data solutions is challenging because so many factors are required to be considered. All of the respondents believed that review of legacy applications and data is needed to get a complete picture of integrating Big

Data solutions with legacy systems.

- **Skills shortage in Big Data and in integration of Big Data with legacy systems:** All of the respondents believed that Big Data solutions are beneficial to their organisation. However, 85.56% of the respondents believed that there are skill shortages within organisations to implement Big Data solutions. Employees are not trained to work with Big Data. Employees are not aware of the ethical concerns of processing Big Data. Volumes of data is growing continuously. Of the respondents, 87.62% are looking to increase their data team headcount to support their Big Data solutions, but the respondents also say it is difficult to find data professionals with the right skills and experience. The skills set required to integrate Big Data solutions with legacy systems is totally dependent on Big Data skills and understanding and handling legacy systems and data issues. Mainly there are three main types of data formats which are also called Geospatial-Intelligence Agency (GIS) Data formats. All these data formats are handled in a different way. They are being used for different purposes. The three data formats are: File-Based Data Format; Directory-Based Data Format; and Database Connections. Integration of different data sets requires an understanding of different data formats and how they can be integrated to provide actionable insights.
- **Benefits of legacy integration:** One hundred per cent of the respondents say that their organisations are using legacy systems and have shown the benefits around them. Legacy systems and the data are assets to the organisation for building informed decisions. BI and DA requires access to legacy systems and data. Organisations' payment strategies, sales strategies, marketing strategies, optimisations of human resources, satisfying customers need, are all required towards contributing to an organisation's competitive advantage. Legacy systems have inbuilt business processes and data, ranging from mainframes to organisational home-grown custom applications to proprietary applications. Data is being retrieved from legacy systems for business intelligence and decision making. Legacy data is a crucial resource for BI and DA but it is also very inefficient for data retrieval and expensive to maintain. Organisations are struggling to find ways in which these systems can be operated efficiently and in a cost-effective manner. There are many issues and challenges associated with legacy systems and data residing in legacy systems. Some of the issues and challenges are: the cost and

complexity of migrating to newer platforms; challenges in accessing data in legacy systems; data refreshes from legacy systems may be too slow for BI and DA purposes; and outdated obsolete technology. All of the respondents believed that legacy systems cannot be replaced as they are beneficial and in use. However, 100% of the respondents believed that their legacy systems must be integrated with Big Data technology to leverage competitive advantage for the organisation. Respondents believed that integration would help them in real-time decision making.

- **Lack of an architecture for legacy and Big Data integration:** Architectures provide structure. The core objective of the Big Data Architecture is to provide a structure for enterprise organisations that aim to benefit from the potential of Big Data. One hundred per cent of the respondents believed that integrating Big Data solutions with legacy systems and data in their organisation will bring benefits to their organisation. The deep understanding derived from integrating Big Data could help organisations to improve their business processes, optimisation of resources, fraud detection, and improve customer relationships and satisfaction. Organisations can use Big Data analytics as their primary source for reporting and analytics after integrating Big Data solutions with legacy systems. Most organisations do not have a strategic plan to execute Big Data integration with legacy systems. Of the respondents, 87.62% stated that their organisation needs to develop a Big Data strategy concerning Big Data integration. This strategy will help plan receiving information from multiple data sources. Integrating new datasets into existing pipelines (72% of the respondents) were cited as the primary obstacles to Big Data integration with legacy systems. This was shown as the biggest concern that would hinder organisations' progress towards a Big Data solution. All of the respondents believed that there is no architecture within their organisation to integrate Big Data solutions with legacy systems. They are generating reports from different legacy systems and combining the report for final analysis.
- **Architecture to integrate Big Data solutions with legacy systems:** The proportion of the respondents who believed that they do not have any Big Data Architecture within their organisation was 87.62%. However, all of them believed that a Big Data Architecture will help the organisations who are struggling with implementing Big Data solutions within their organisations. This will provide them with support and structure to start with their Big Data projects.

The survey helped us answering RQ1 and it has been identified through our survey that there is no architecture to implement Big Data solutions and integrate Big Data solutions with legacy systems in organisations. Big Data integration with legacy systems will provide solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats. However, the limitation of this survey is that Big Data issues and challenges cannot be generalised to all organisations. There is a need to conduct the survey for wider organisations wanting to implement Big Data technologies.

3.6 Chapter Summary

This chapter has discussed our survey and has demonstrated that not many organisations are implementing Big Data solutions, but that 100% of respondents believe that implementing Big Data solutions will bring benefits to organisations in many ways, such as using external data sources for making organisational decisions. It has been identified through our survey that there is no architecture to implement Big Data solutions and integrate Big Data solutions with legacy systems in organisations. Big Data integration with legacy systems will provide solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats.

Having a Big Data integration solution in place is as important as having good analytics tools for creating insights. However, from our survey we have identified there are many challenges to integrating existing legacy systems and data with Big Data solutions. These include skills shortages in Big Data, along with technical challenges, such as the lack of Big Data architectures, the lack of an organisational Big Data strategy, cloud computing, the lack of a Big Data architecture or an inability to define one, and lack of knowledge of software systems to support Big Data. We also identified the issue of a lack of people with the organisational knowledge of the legacy systems and the legacy data.

CHAPTER 4: BIG DATA STRATEGY AND THE ANALYTICS VALUE CHAIN

4.1 Introduction

This chapter discusses Big Data strategy and the analytics value chain. An organisation's Big Data strategy encompasses its approach to storage, analysis, data architectures, and decisions about data models. As discussed in chapter 2, Enterprise Architecture (EA) is a blueprint of an organisation. Research in Big Data community does not address EA and its role in Big Data integration. However, EA addresses the process by which organisations standardise and organise IT infrastructure to align with business goals. Big Data strategy and analytics value chain is dependent on an organisation's IT infrastructure. An effective EA will help organisations achieve business goals such as reduction of expenses, the implementation of a data-driven culture, innovation, the acceleration of the deployment of new capabilities and services, and the launch of new products and services with the help of organisations' IT infrastructure. Organisations around the world are slowly beginning to incorporate big data analytics into their business models and are using it for more educated decision-making (Attaran et al., 2018). The key purpose of this chapter is to propose Big Data strategy and for the successful implementation of a Big Data Architecture for analytics. Big Data strategy makes the benefits of Big Data actionable for the organisation. Big Data strategy makes the benefits of Big Data actionable for the organisation. Analytics using Big Data for Business Intelligence within an organisation are required to support critical operational processes of an Enterprise. Every data point is potentially valuable, from the publicly available data on the Internet's 60 trillion individual pages from consumer reviews, to internal data such as e-health records, smart cards, and financial transactions.

This chapter presents what constitutes the Analytics Value Chain, what a Big Data strategy is, what a Big Data solution looks like and how it is related to EA. At the end, a summary of the chapter is presented.

4.2 Background

In the age of the Internet, much of the focus is on faster delivery of more information, such as documents, medical images, movies, gene sequences, and sensor data streams to systems, PCs, mobile devices, and living rooms. The challenge for organisations in the next decade will be finding ways to better analyse, monetise, and capitalise on all these information channels and integrate them into their business. Most organisations have existing Enterprise Architectures. Defining how the streaming of Big Data into an existing Enterprise Architecture occurs is an issue.

Business and Technology architecture often reflect this flow, starting with transactions and operations. Organisations review and plan improvements. These improvements are done through projects that span months and years. Big Data solutions have changed this concept. Data, that is part of Big Data solutions, are in a constant state of change and flux, and those organisations that can recognise and react quickly and intelligently have the upper hand. The IT capabilities are today more focused on discovery and agility rather than stability. Data scientists should be able to work with Big Data tools and technologies to continuously mine new and existing data sources for patterns, events, and opportunities at an unprecedented scale and pace. This gives rise to stream processing.

The demand for stream processing is increasing. The reason is that often processing big volumes of data is not enough. Data must be processed fast, so that an organisation can react to changing business conditions in real time. This is required for trading, fraud detection, system monitoring, and many other examples. Many organisations have failed to effectively incorporate Big Data in their own decision-making processes. The Big Data analytics value chain can provide useful insights into the characteristics of the environments in which many organisations operate (Tabesh, 2019).

4.3 Analytics Value Chain

The idea of analysing data to make sense of what is happening in businesses has been around for a long time. Over the years the name and terminology for data analytics have changed because of the advent of technology and the way data is processed for use. The activity of making sense of data has been called Decision Support, Executive Support, Online Analytical

Processing, Analytics, and now Big Data Analytics (Davenport, 2014).

Analytics using Big Data for Business Intelligence within an organisation are required to support critical operational processes of an Enterprise. Every data point is potentially valuable, from the publicly available data on the Internet's 60 trillion individual pages from consumer reviews, to internal data such as e-health records, smart cards, and financial transactions.

Many organisations have multiple databases and multiple database vendors, with terabytes or even petabytes of data. Some of these systems accumulated data over several years. Many organisations build entire data warehouse and analytic platforms off this old data. For data to be useful to users they must integrate customer data with finance and sales data, with product data, with marketing data, with social media, with demographic data, with competitors' data,and more. After a decade of channelising data collection, structures, storage, access and retrieval, a value chain has emerged, as shown in Figure 4.1.

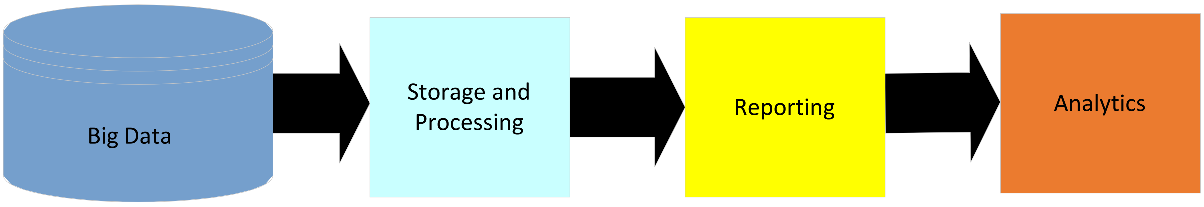


Figure 4.1: Analytics Value Chain (Akerkar, 2014)

The components of a typical analytics value chain are:

- **Big Data**: Sources and origin of heterogeneous data such as the Internet, sensors, machines, Web Logs, images/audio, and video. The formats of these data are structured and unstructured.
- **Storage and Processing**: Storage and Processing requires data to be indexed, organised, optimised and prepared for analysis. Technologies supporting data processing include Relational Databases (RDBMS), not only SQL (NoSQL), and distributed file systems. Big Data solutions require specialised technologies to efficiently process the large volume of data. Manyika et al. (2011) suggest suitable technologies such as crowdsourcing, data fusion and integration, generic algorithms, machine learning, time series and visualisation.

- **Reporting**: Reporting requires identification of relationships, that also identify and evaluate what happened in the past.
- **Analytics**: Business analytics functions, system supporting business analytics, and predictive analytics are the major intent of Big Data. Predictive analytics helps to find patterns in data to help decision makers.

Organisations must treat data and information as a strategic asset for any analytics. Every organisation needs to manage data generated from its automation systems such as: Enterprise Resource Planning, Customer Relationship Management, Time and Attendance, and e-commerce. Data warehouses, data mining, and database technologies have existed in various forms for years. Many IT professions have worked with large amounts of data in various industries for many years before the term Big Data was coined. Analysing semi-structured and unstructured data is new with the advent of Big Data and the tools that enable it to be done. A decade ago, email messages, PDF files, or videos were not analysed as there was not any value chain of collecting, storing, and accessing data (Akerkar, 2014).

Analytics can be divided into three categories: Descriptive analytics, Predictive analytics, and Prescriptive analytics. Descriptive analytics depicts what has already happened, that is, historical data. Descriptive analytics is the commonly used and well-understood type of analytics. It basically categorises, characterises, consolidates, and classifies data. Tools for descriptive analytics provide mechanisms for interfacing to Enterprise data sources. They contain report generation, distribution capabilities, and data visualisation facilities.

Until the advent of Big Data tools and solutions, descriptive analytics was all that was available. Spreadsheets, financial statements, and accounting reports are examples of descriptive analytics. Descriptive analytics enables us to look back and evaluate what has already transpired. Descriptive analytics can be classified into the following threetasks:

- *Usual reporting and dashboards:* What took place? How does it relate to the organisational objectives?
- *Ad hoc reporting:* How many? Where?
- *Analysis/query:* What exactly are the challenges? Why is this happening?

Real time or near real-time descriptive analytics depict what is happening now with the advent of tools and technology that enabled the processing of the underlying data in real or near real time. These are new variations of the analytics that came to be with the advent of Big Data capabilities. We can see the details of the events while the events are happening or soon after they occur and take immediate action to leverage good events and/or reduce the impact of bad events.

Predictive analytics is looking forward. Predictive analytics uses data to find out what could happen in the future. It is a more refined and higher-level usage of analytics. For example, banks issue loans based on predictive analytics. Predictive analytics can be classified into six tasks:

- *Data mining*: What data are correlated with other data?
- *Forecasting*: What will happen tomorrow?
- *Root cause analysis*: Why did it occur?
- *Pattern recognition*: When should a process be altered?
- *Monte-Carlo simulation*: What could emerge?
- *Predictive modelling*: What will happen then?

Predictive analytics existed in earlier forms as businesses have always sought to discern the future to prepare for it or to leverage a new insight. Today's predictive analytics are far more accurate and are generally available in mere minutes. With the advent of Big Data, decision making has become less reactive and more proactive (Baker, 2015).

Prescriptive analytics is the newest frontier. Once the past is understood and predictions can be made about what might happen in the future, one needs to know what the best action will be, given the limited resources of the organisation. In this class of analytics, results are accompanied with automated actions or a list of recommended actions with a likely outcome for each action. The user may pick one of the choices and implement it. Facebook uses Prescriptive analytics to determine who to recommend to us as a friend. Prescriptive analytics is based on the concept of optimisation, which can be divided into two tasks:

- *Optimisation*: How can the best results be achieved?
- *Stochastic optimisation*: How the best results can be achieved to tackle improbability in the data to make better decisions.

Prescriptive analytics is used in the maintenance of large systems such as road, water, and cable TV systems. Prescriptive analytics can determine when a road will need widening or a stoplight should be added at an intersection based upon actual and expected traffic flows and other factors (Akerker, 2014).

4.4 Big Data Strategy

A Big Data strategy defines and lays out a comprehensive vision across the organisation and sets a foundation for the organisation to employ data-related or data-dependent capabilities. Clearly stated business goals lie at the centre of any successful organisation. A well-defined and comprehensive Big Data strategy makes the benefits of Big Data actionable for the organisation. An organisation's Big Data strategy should able to build support for a big data initiative to ensure that the big data initiative is valued by, or of value to, the business stakeholders (Schmarzo, 2013). Figure 4.2 shows the Big Data strategy document.

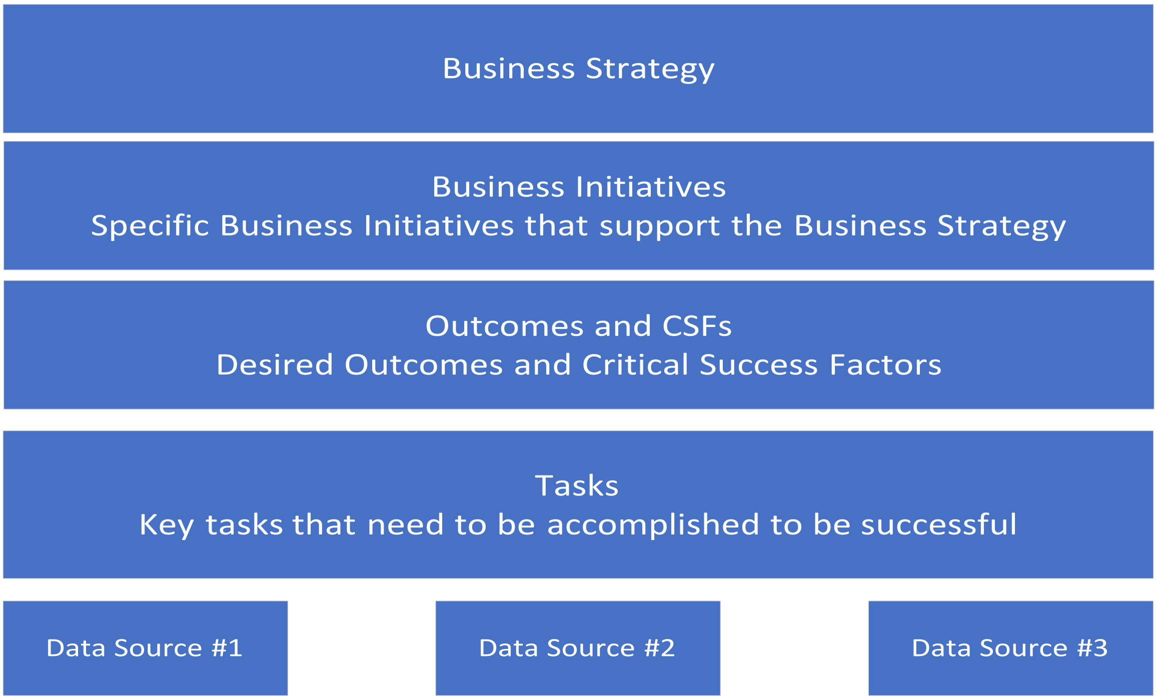

Business Strategy
Business Initiatives
Specific Business Initiatives that support the Business Strategy
Outcomes and CSFs
Desired Outcomes and Critical Success Factors
Tasks
Key tasks that need to be accomplished to be successful
Data Source #1
Data Source #2
Data Source #3

Figure 4.2: Big Data Strategy Document (Schmarzo, 2013)

Without a Big Data strategy, organisations will be forced to deal with a variety of data-related activities which may not be business-goal oriented. In the figure 4.2, targeted business strategy is captured as the title of the document and clearly defines the scope upon which the Big Data initiative will be focused. For example: “Improve customer intimacy” or “Reduce operational maintenance costs” or “Improve new product launch effectiveness”. Business strategy should be supported by business initiatives. A business initiative clearly states financial or business goals against which success will be measured. Outcome and critical success factors (CSF) capture the outcomes and critical success factors necessary to support the successful execution of the organisation’s key business initiatives. Outcomes define the desired or ideal end state. Critical success factors define “what needs to be done” for the business initiative to be successful. The next level of detail is provided by documenting the specific tasks that need to be executed to perfection to be successful in support of the targeted business initiatives. These are the key tasks around which the different parts of the organisation will need to collaborate to achieve the business initiatives. This is the “how to do it” section of the document, and it is at this level of detail where personal assignments and management objectives can be defined, assigned, and measured. Finally, the document highlights the key data sources required to support the business strategy and the supporting key business initiatives. From the definition of the tasks, one should have a strong understanding of the key metrics and measures, important business dimensions, level of granularity, and frequency of data access.

Big Data strategy is being used by the analytics value chain for managing efficiently value creation processes within organisations (Faroukhi et al., 2020). Big Data-driven organisations, that use analytics value chains, generate higher benefits than traditional organisations since, once data is generated, it can be mined multiple times and for different needs (McAfee & Brynjolfsson, 2012). Analytics value chains are dependent on Big Data strategy. Data sources identified in Big Data strategy are input to Big Data for analytics value chains for Business Intelligence reporting and analytics.

As Business Intelligence allows organisations to strategically plan, transform and consolidate their data into meaningful information delivered in a presentable format for effective decision making, user involvement and experience are the most important components of any Business Intelligence solution. Human talent is needed to analyse and extract insights from the data. The ability to use technology is crucial to being able to analyse and make sense of data. Using the right analytic functions, with the right presentation and visualisation of the results so they can be properly interpreted by humans, is a major challenge for Big Data solutions and how these are integrated into the EA and used within the organisation creates both a challenge and an opportunity for the organisation.

4.5 Challenges and opportunities for Big Data and Enterprise Architecture

The growth of data has been observed in every industry, though some industries have greater volumes of data than others from which to derive value. While small businesses have potential to benefit from data-driven innovation, businesses, government, and industries with high data intensity have even greater potential in the near term. Industries where the culture is more conducive to innovation and learning have a greater propensity to understand and capture the potential from data-driven insights. Having the right solutions that match the size and type of an organisation is a challenge and an opportunity for vendors and providers of Big Data solutions. Business Intelligence and Data Analytics require:

- Consolidation and merging of data from various systems.
- Data visualisation to best present the information.
- Dashboards and Scorecards for quick high-level report consumption.
- Ad-hoc reporting for quick self-service analysis.
- Mobile accessibility of information on smartphones and tablets.
- Implementing a Big Data architecture: Builds out the necessary architectural requirements addressing Extract, Transform, Load (ETL)/ Extract, Load, Transform (ELT), data staging area, data management platform, master data management capabilities, business intelligence, and advanced analyticsplatforms.

All of these provide challenges to Big Data solution providers and to organisations implementing Big Data solutions. So, the problem is how to overcome these impediments with Big Data solutions so that organisations can use Big Data to enable business transformation. Big Data enables organisations to transform from a "rear view mirror", hindsight view of the business using a subset of the data in batches to monitor business performance, into a predictive enterprise that leverages all available data in real-time to optimise business performance. Unfortunately, traditional data management and analytic technologies and approaches are hindering this business transformation and providing challenges because they are incapable of managing the tsunami of social, mobile, sensor, and telemetry data, and consequently, they are unable to make timely decisions (Schmarzo, 2013).

4.6 Building Blocks for Streaming Big Data into the Enterprise Architecture

According to Driscoll (2011), the Big Data processing stack is comprised of three layers: Foundation Layer, Analytics Layer, and Application Layer. Big Data Storage, Processing, and Reporting are handled in the Foundation layer, Analytics is handled in the Analytics Layer, and the result of the Analytics Layer is visualised in the Application Layer. The detailed purpose of these layers is described as follows:

- Foundational Layer: The foundational layer provides the infrastructure for storage and processing, access, and management of Big Data. Depending on the nature of data, stream processing solutions (Abadi et al., 2003; Salehi, 2010), distributed persistent storage (Shvachko et al., 2010), cloud infrastructures (Cusumano, 2010), or a reasonable combination of these (Sakr et al., 2011) is used for storing and accessing data in response to the upper-layer requests andrequirements.
- Analytics Layer: This is the middle layer and responsible for analytics. Here data warehousing technologies are currently exploited for extracting correlations and features from data, and feeding classification and prediction algorithms (Thusoo et al., 2010).
- Application Layer: This layer is also called focused application layer and is at the top of the stack. The functionality of the application layer is based on the use of more generic lower-layer technologies and exposed to end users as Big Data solutions.

All three layers are the building blocks for streaming Big Data into the EA incorporating the Analytics Value Chain. The functionality of all these layers is dependent on Technology Infrastructure and Data Architecture. Infrastructural technologies are the cornerstone of Big Data Solutions. They relate to data storage, data steam processing, data management, and query planning and execution. Possessing the right tools for storing, processing and analysing data is crucial in Big Data solutions.

4.6.1 Technology Infrastructure and Technology Architecture

Infrastructural technologies process, store, and often also analyse data. After the arrival and proliferation of IT in large enterprises, various approaches, techniques, and methods have been introduced to solve the data integration challenge. In the last decade, the prevalent data integration approaches were primarily based on Extensible Markup Language (XML), Web Services, and Service Oriented Architectures (SOA) (Halevy et al., 2006). XML defines a standard syntax for data representation, Web Services provide data exchange protocols, and SOA is a holistic approach for distributed systems architecture and communication based on the use of services. However, these technologies are not sufficient to solve the data integration challenge in large organisations. Overheads associated with SOA are still too high for rapid and flexible data integration, which are a prerequisite for the dynamic world analytics (Halevy et al., 2006).

One of the key infrastructural technologies is Hadoop (Apache Software Foundation, 2019). Hadoop is a whole ecosystem of technologies designed for the storing, processing, and analysing of data. Hadoop provides a software architecture for distributed storage and processing of Big Data using the MapReduce programming model. The core Hadoop technologies work on the principle of breaking up and distributing data into parts and analysing those parts concurrently, rather than tackling one monolithic block of data all in one go. It is more efficient to break up and distribute data into many parts, allowing processing and analysing of different parts concurrently. The main advantages of Hadoop are its cost and time effectiveness. Cost, because as it is open source, it is free and available for anyone to use, and can run off cheap commodity hardware. Time, because it processes multiple parts of the data set concurrently, making it a comparatively fast tool for in-depth analysis. The Apache Software Foundation are constantly updating and developing the Hadoop ecosystem, such as Hadoop on Premium. Hadoop on Premium services, such as Cloudera, Hortonworks and Splice, offer the Hadoop architecture with greater security and support, with added system and data management tools and enterprise capabilities. Some key components of Hadoop include:

- HDFS: Hadoop Distributed File System which is the default storage layer and Hadoop’s own rack-aware file system. This is designed to scale to tens of petabytes of storage and runs on top of the file systems of the underlying operating systems.
- MapReduce: This is composed of two components Map and Reduce. The Map job distributes a query to different nodes, and the Reduce gathers the results and resolves them into a single value.
- Yet Another Resource Negotiator (YARN): Responsible for cluster management and

scheduling user applications. YARN allows different data processing engines like graph processing, interactive processing, stream processing, as well as batch processing to run and process data stored in HDFS.

- Spark: Used on top of HDFS and promises speed up to 100 times faster than the two step MapReduce function in certain applications. This allows data to be loaded in memory and queried repeatedly, making it suitable for machine-learning algorithms.
- Not Only SQL (NoSQL): NoSQL is involved in processing large volumes of multi-structured data. Most NoSQL databases are most adept at handling discrete datastored among multi-structured data. Some NoSQL databases, like HBase, can work concurrently with Hadoop.

NoSQL is better suited for operational tasks, interactive workloads based on selective criteria where data can be processed in near real time. Hadoop is better suited to high-throughput, and in-depth analysis. Hadoop and NoSQL products are sometimes marketed concurrently. Some big names in NoSQL field include Apache Cassandra, MongoDB, and Oracle NoSQL. Many of the most widely used NoSQL technologies are open source, meaning security and troubleshooting may be an issue. It also places less focus on atomicity and consistency than on performance and scalability. Premium packages of NoSQL databases (such as Datastax for Cassandra) work to address these issues.

Massively Parallel Processing (MPP) Databases work by segmenting data across multiple nodes and processing these segments of data in parallel. Whereas Hadoop usually runs on cheaper clusters of commodity servers (allowing for inexpensive horizontal scale out), most MPP databases run on expensive specialised hardware (data warehouse appliances). MPP technologies process massive amounts of data in parallel. It may have hundreds (or potentially even thousands) of processors, each with their own operating system and memory, working on different parts of the same programme.

MPP uses SQL, and Hadoop uses Java as default (although the Apache Foundation developed Hive, a language used in Hadoop like SQL, to make using Hadoop slightly easier and less specialist). Many of the major players in the MPP market have been acquired by technology vendors. Netezza, for instance, is owned by IBM, Vertica is owned by HP, and Greenplum is owned by EMC.

MapReduce for C (MR4C) is an implementation framework which enables large-scale deployment of advanced data processing applications and allows native code to run on Hadoop. MR4C was originally developed at Skybox Imaging to facilitate large scale satellite image processing and geospatial data science. All external files are configured using JavaScript Object Notation (JSON). MR4C allows multiple algorithms to loop together and is used for geospatial data analysis. However, Zhao et al (2015), have suggested implementing Hadoop framework for applications written in C/C++ such as video read/write interface for Hadoop.

Another key technology is Cloud computing, which refers to a broad set of products that are sold as a service and delivered over a network. The cloud computing paradigm builds on the premise that computational capacity can be consumed on-demand, thus reducing the cost at the demand side while improving the utilisation of resources at the supply side. According to Biocic et al. (2011), virtualisation via cloud computing can increase the utilisation of computing infrastructures from an average of 15% up to 90% compared to traditional IT environments. The traditional infrastructure setting requires hardware and software for each person involved with the processing and analysis of data. In cloud computing access to only one application (a Web-based service) is required. The expected benefits draw from flexible management of capacities to perform computational tasks without investing in new infrastructure, training of personnel, or licensing of software. This is very much relevant around Big Data solutions where the increasing availability of heterogeneous data requires large amounts of storage capacities and computer intensive analytics to derive business value from cloud computing.

Case studies from market research and business literature suggest that Big Data storage systems based on the principle of cloud computing coincide with massive savings in IT infrastructure and operation. Walt Disney lowered the company's IT expense growth from 27% to 3%, while increasing its annual processing growth from 17% to 45% after setting up a Hadoop Cluster on the cloud (Gruman, 2010).

A Big Data–ready cloud-computing platform provides the following key capabilities:

- Agile Computing Platform: Agility is enabled through highly flexible and reconfigurable data and analytic resources and architectures. Analytic resources can be quickly reconfigured and redeployed to meet the ever-changing demands of the business, enabling new levels of analytics flexibility and agility.

- Linear Scalability: Access to massive amounts of computing power means that business problems can be attacked in a completely different manner. For example, the traditional Extract, Transform and Load (ETL) process can be transformed into a data enrichment process creating new composite metrics, such as frequency (how often?), recent in time (how recent?), and sequencing (the order?).
- On-Demand, Analysis-Intense Workloads: Previously organisations had to be content with performing "after the fact" analysis. The organisations lacked the computational power to dive deep into the analysis as events were occurring or to contemplate all the different variables that might be driving the business. With a cloud platform, these computationally intensive, short burst analytic needs can be exploited. Business users can analyse massive amounts of data in real time. This means uncovering the relevant and actionable facts across hundreds of dimensions and business metrics.

4.6.2 Data Architecture

Introducing Big Data solutions will have an impact on the Data Architecture which is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organisations. Data sources and data access requirements should include a detailed plan and roadmap for prioritising what Big Data to capture and where to store that data (both from a data access, as well as an analysis perspective). This plan will need to address both structured and unstructured data. It also needs to address external data sources, which means that the data plan will need to be updated every four to six months to accommodate the many new data sources that are becoming available (Liebowitz, 2013).

Information is called unstructured if it is not provided as a list of attribute value pairs for attributes with data types defined in a suitable schema. Typically for data types covering numbers, patterns of numbers, letters, and names, structured information can be provided by a table as an instance of a relation scheme fixing a set of attributes in a relational database. Examples of unstructured data are texts, sound, image, and multimedia data. The following conventions are used in a Data Architecture to model:

- **Class diagrams**: The key purpose of the class diagram is to depict the relationships among the critical data entities (or classes) within the enterprise. This diagram is developed to present the relationships and to help understand the lower level data models for the enterprise.

- **Data dissemination diagrams**: The purpose of the data dissemination diagram is to show the relationship between data entities, business services, and application components. The diagram shows how the logical entities are to be physically realised by application components. This allows effective sizing to be carried out and the IT footprint to be refined. Moreover, by assigning business value to data, an indication of the business criticality of application components can be gained. The diagram also shows data replication and system ownership of the master reference for data. This diagram can include services; that is, services encapsulate data and they reside in an application, or services that reside in an application and access data encapsulated within the application.
- **Data life cycle diagrams**: The purpose of the data lifecycle diagram is to capture changes of business processes. The data lifecycle diagram is an essential part of managing business data throughout its lifecycle, from conception through disposal, within the constraints of the business process. The data is considered as an entity, detached from business processes and activities. Each change in state is represented in the diagram, which may include the event or rules that trigger that change in state. The separation of data from process allows common data requirements to be identified, thereby enabling more effective resource sharing to be achieved. Defining the lifecycle of business entities enables better formalisation of these business entities, as well as the determination of the steps that are essential to their management. This state model will be a guide in the definition of business processes, since these processes will themselves have to respect the constraints defined for transitions between states: if a business entity has not passed through all its states within the business processes that handle it, then these are incomplete. If the business processes transgress the lifecycle of the business entities, then they are incorrect.
- **Data migration diagrams**: The purpose of the data migration diagram is to show the flow of data from the source to the target applications. The diagram will provide a visual representation of the spread of sources/targets and serve as a tool for data auditing and establishing traceability. This diagram can be elaborated or enhanced as detailed, as necessary. For example, the diagram can contain just an overall layout of migration landscape or could go into individual application metadata element

level of detail.

- **Data security diagrams**: Data is considered as an asset to the enterprise and data security simply means ensuring that enterprise data is not compromised and that access to it is suitably controlled. The purpose of the data security diagram is to depict which actor (person, organisation, or system) can access which enterprise data. This relationship can be shown in matrix form between two objects or can be shown as a mapping. The diagram can also be used to demonstrate compliance with data privacy laws and other applicable regulations such as Health Insurance Portability and Accountability Act (HIPAA) and Sarbanes-Oxley Act (SOX). This diagram should also consider any trust implications where an enterprise's partners or other parties may have access to the company's systems, such as an outsourced situation where information may be managed by other people and may even be hosted in a different country.

4.7 EA Data Management Architecture

Behind any information management lies the core doctrine of Data Quality, Data Governance and Metadata management along with considerations for privacy and legal concerns. Data governance helps an organisation to take a holistic view and to manage data in the context of business process, and to support application integration needs. Data Architecture requires information about master and reference data management, data warehousingbusiness intelligence, transaction data management, structured technical data management, unstructured data management, Metadata, Analytical Data, Documents and contents, Historical Data, Temporary Data, and Big Data analytics. All these data sources need to be organised in an information architecture which can support data governance and data asset planning.

The key element of organisational structure is the business functions. Organisations need to know the kind of data required, captured, acquired, stored, and analysed for business functions such as marketing, sales, supply chain, manufacturing, human resources, strategy making, finance, and information technology (Jha, et al. 2016). Data is an asset to the organisation and as such must be managed to ensure competitive advantage and to reduce the complexity of the asset management in the organisation.

Information Goals and Principles are top priority for any organisation, but this must be supported by Data Governance and Data Asset Planning. So, while constructing the roadmap of Information Management Architecture we require Data Architecture Information. The data sources are heterogeneous involving more diverse unstructured and semi-structured data sets. The data sources are likely to be found both outside the organisation and inside the organisation. For business intelligence the external and internal data sources are required. These data sources must be addressed by the EA.

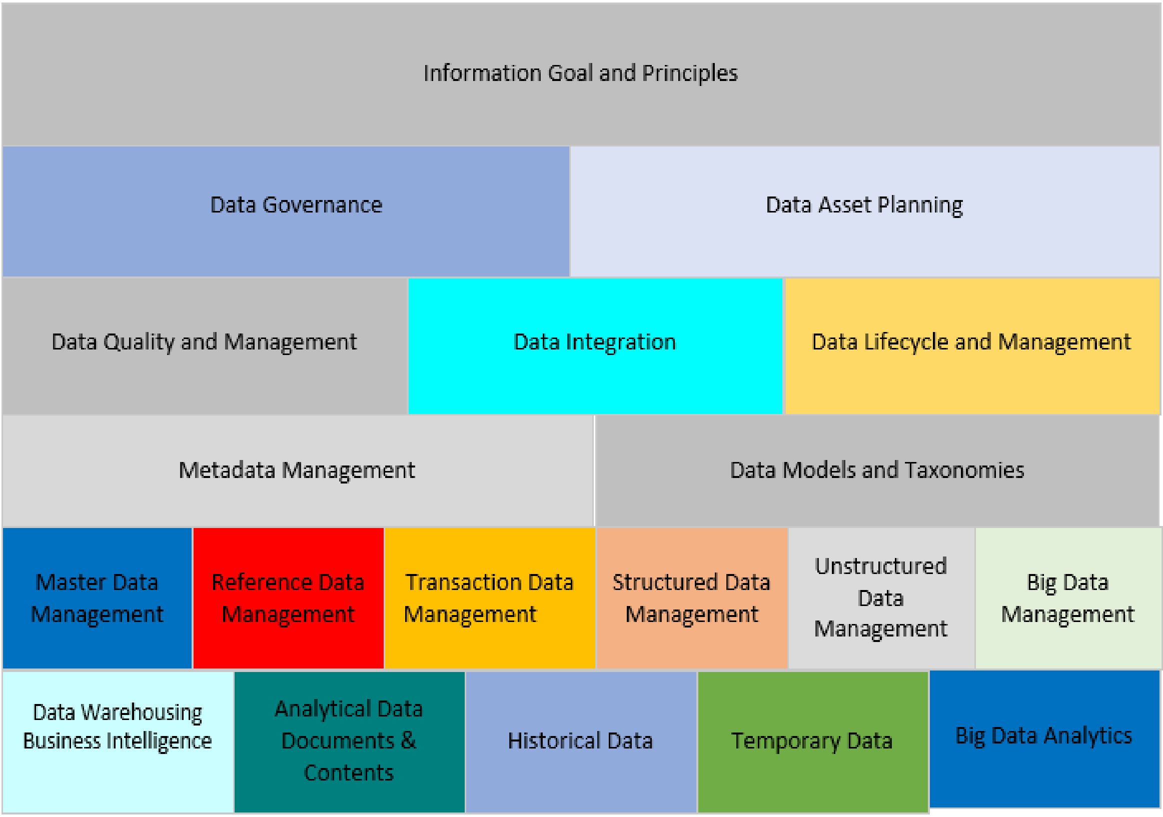

Information Goal and Principles
Data Governance
Data Asset Planning
Data Quality and Management
Data Integration
Data Lifecycle and Management
Metadata Management
Data Models and Taxonomies
Master Data Management
Reference Data Management
Transaction Data Management
Structured Data Management
Unstructured Data Management
Big Data Management
Data Warehousing Business Intelligence
Analytical Data Documents & Contents
Historical Data
Temporary Data
Big Data Analytics

Figure 4.3: EA Data Management Architecture

Figure 4.3 shows an EA Data Management Architecture (Jha, et al. 2016). In a high Information Technologyconsuming context such as Big Data solutions, data governance and data asset planning shouldbe shown in the enterprise architecture. Data governance and data asset planning remove the risk of bad operational/transactional/in-motion data which removes the risk of bad Intelligence,reporting and poor decision making.

4.8 Big Data Internal and External Sources of Data

The following are the internal and external sources of Big Data which need to be integrated to big data storage, processing and analysing platform, as shown in Figure 4.4.

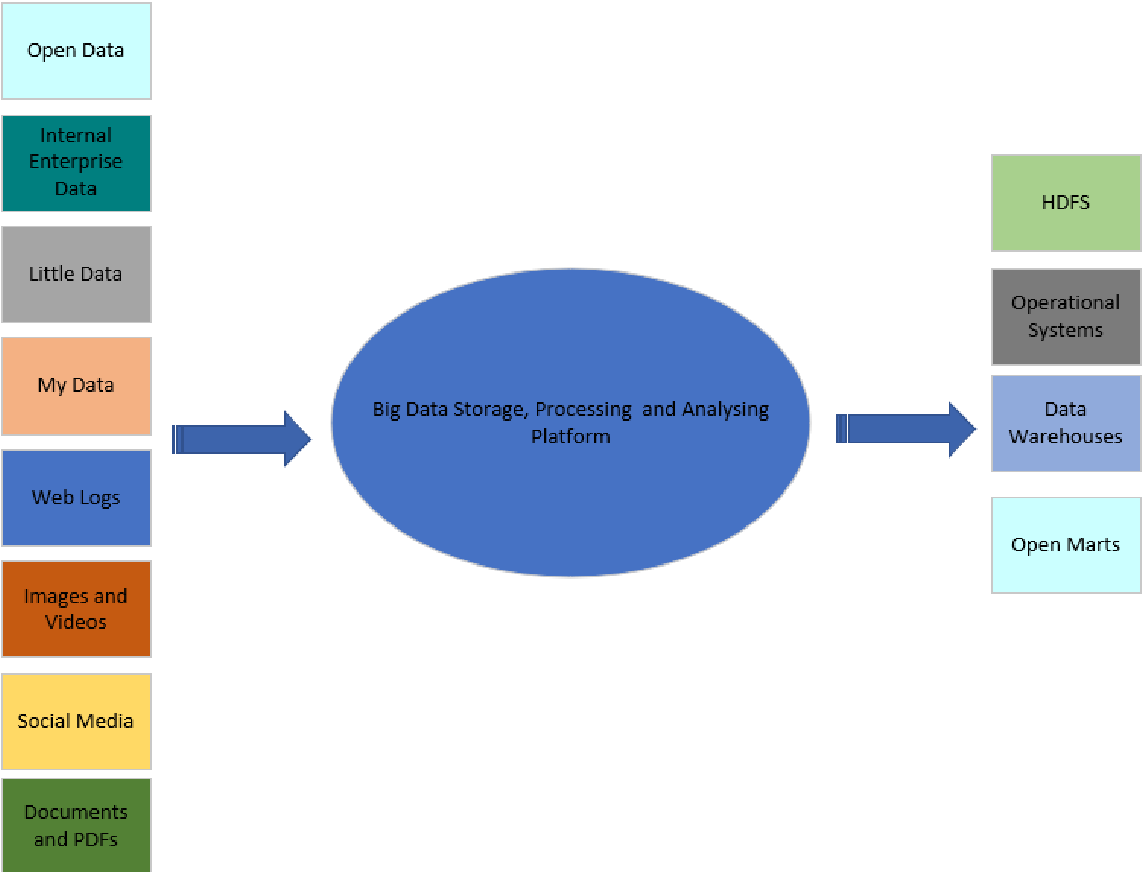


Figure 4.4: Integrating Big Data Sources for Big Data Solutions

- Open Data: A key source of data from governments and private institutions. Open relates to how accessible a data set is in terms of allowing others to use it without restriction (PWC, 2014).
- Internal enterprise data: Data that is collected by an organisation about its own systems and processes. This data may not be digital, can consist of both quantitative and qualitative information, and can also be anonymised. A bank using anonymised customer transaction records to predict and proactively refill its Automated Teller Machines (ATM) is an example of this type of data (PWC, 2014).

- Little data: Small businesses can also make use of data analytics across data that they have about their own business, like Big Data, but on a smaller scale (PWC, 2014).
- My Data: Internal data about a particular organisation or individual. This type of data is typically held securely with strict rules regarding access; for example, a hospital holding an individual's health records for their health care professionals, which would also allow them to diagnose the patient on the basis of other aggregate data on medical conditions (PWC, 2014).
- Web Logs: Web Logs are personal Web pages written in chronological order and maintained through specific software that helps their administration. A Web Log, sometimes written as Web Log or weblog, is a website that consists of a series of entries arranged in reverse chronological order, often updated frequently with new information about topics.
- Images and Videos: Today, images and image sequences (videos) make up about 80% of all corporate and public unstructured big data. As growth of unstructured data increases, analytical systems must assimilate and interpret images and videos as well as they interpret structured data such as text and numbers.
- Social Media: Websites and applications that enable users to create and share content or to participate in social networking.
- Documents and PDFs: PDF (Portable Document Format) is a file format that has captured all the elements of a printed document as an electronic image.

Once the Big Data which is unstructured data is being stored, processed and analysed, it can be stored in HDFS, operational systems, data warehouses and data marts from where data can be used for Business Intelligence, reporting, OLAP, ad-hoc reporting and modelling (Jha, et al. 2016).

4.9 Chapter Summary

This chapter discussed Big Data Strategy, Analytics Value Chain, EA, technology infrastructure, technology architecture, and data architecture of Big Data into EA for BI and DA. The components of Analytic Value Chain form the building blocks for Big Data. Technology infrastructure, technology architecture and data architecture will govern the Big Data architecture to integrate legacy systems and data with Big Data Solutions.

Working with legacy data sources like mainframes and integrating it with Big Data like stream-process transaction data requires integration of legacy systems and data with Big Data solutions. In this chapter we have explored the Analytics Value Chain which is a part of the building blocks for Big Data into the EA. Big Data solutions will impact the Data Architecture, Application Architecture, Technology Architecture and Business Architecture of EA. We have also explored the challenges and opportunities of Big Data in EA. Big data is about business transformation which requires Business Architecture to be changed. Traditional data warehouse and business intelligence technologies impede business growth.

New technologies should be leveraged, such as Hadoop, in memory computing, and in database analytics, to provide new data management and advanced analytics capabilities, and open new, more modern architectural options to leverage Big Data solutions. Organisations should be prepared to embrace open-source technologies and tools (Jha, et al. 2016). Organisations should document an Enterprise Architecture that is used to govern how Big Data can be streamed intothe BI environment of the organisation.

New ways of doing BI and DA impact on technology infrastructure components and data architectures. Since most data is directly generated in digital format today these impacts and changes must be captured by an organisation’s EA for conducting enterprise analysis, design, planning, and implementation, using a holistic approach at all times, for the successful development and execution of organisational strategy. Enterprise Architecture is a detailed overview of a business's processes and how they relate to the IT infrastructure. Business processes can foster an organisation’s flexibility, agility, and dynamism. However, most legacy systems’ business processes are saddled by fragmented analytics conditions characterised by business and data silos that limit the ability to operate with flexibility, agility, and dynamism.

Re-engineering business process will allow the integration of legacy systems and data with Big Data solutions to achieve flexibility, agility, and dynamism for BI and DA. In the next chapter we present an architecture for integrating legacy systems and data with Big Data.

CHAPTER 5: ARCHITECTURE FOR INTEGRATION OF LEGACY SYSTEMS AND DATA WITH BIG DATA SOLUTIONS

5.1 Introduction

Legacy systems and its business processes are complex. Most of the legacy systems were developed without process models or data models. For integrating legacy systems and data with Big Data, process and data models need to be standardised. Business processes within legacy systems are required to be identified, reviewed, and updated to fit Big Data solutions. Business processes are required to be re-designed through re-engineering. This chapter discusses architecture for the integration of legacy systems and data with Big Data solutions. This chapter also discusses how Big Data analytics and the business process can be combined using re-engineering and can deliver benefits to the organisations and customers. It describes Big Data analysis requirements by selecting a business process for re-engineering, the business and organisational impact of Big Data on business processes, and tasks to combine Big Data analytics with business processes to develop a Big Data architecture to integrate Big Data solutions with legacy systems and data. This chapter presents the architecture we have developed for integrating legacy systems and data with Big Data solutions.

5.2 Background

Organisations have various types of information systems to run basic business processes. Improving business processes is paramount to staying competitive in today's electronic marketplace. Organisations must improve their business processes because customers are demanding better products and services. If customers do not receive what they want from one supplier, they can simply click a mouse and have many other choices. Business process improvement attempts to understand and measure the current process and make performance improvements accordingly. Business Process Management (BPM) enables a company to successfully align business practices with strategic objectives and increase business performance. Functional silos or roadblocks disappear, and the organisation is better positioned to satisfy all stakeholders – customer, owners, andemployees.

Business Process Re-engineering (BPR) is the analysis and redesign of workflow within and between enterprises. BPR involves the radical redesign of core business processes to achieve dramatic improvements in productivity, cycle times, and quality. Using BPR, companies start with a blank sheet of paper and rethink existing processes to deliver more value to the customer. They typically adopt a new value system that places increased emphasis on customer needs. Companies reduce organisational layers and eliminate unproductive activities in two key areas. First, they redesign functional organisations into cross-functional teams. Second, they use technology to improve data dissemination and decision making.

BPR is a dramatic change initiative that contains four major steps (Tuomisaari et al, 2012). Companies should:

- Refocus company values on customer needs.
- Redesign core processes, often using information technology to enable improvements.
- Reorganise a business into cross-functional teams with end-to-end responsibility for a process.
- Rethink basic organisational and people issues.

Companies use BPR to improve performance substantially on key processes that impact customers (AbdEllatif et al., 2018). BPR can:

- Reduce costs and cycle time. BPR reduces costs and cycle times by eliminating unproductive activities and the employees who perform them. Reorganisation by teams decreases the need for management layers, accelerates information flows, and eliminates the errors and rework caused by multiple handoffs.
- Improve quality. BPR improves quality by reducing the fragmentation of work and establishing clear ownership of processes. Workers gain responsibility for their output and can measure their performance based on prompt feedback.

Big Data analytics are focused on transforming big quantities of data into usable information. Business processes generate information. Intelligent business processes need Big Data for analytics to provide competitive advantage. Organisations want to make good decisions. Good decisions require data from several different sources. Accessing data is the critical path in making good decisions.

Traditionally, three major classes of information systems are found in organisations: transaction processing systems (TPS); decision support systems (DSS); and executive information systems (EIS). All three systems support various kinds of decision making. Transactional data encompass all the raw facts contained within a single business process or unit of work, and their primary purpose is to support the performing of daily operational tasks. Examples of events where transactional data are captured include purchasing stocks, making an airline reservation, or withdrawing cash from an ATM. Analytical information encompasses all summarised or aggregated transactional data, and its primary purpose is to support the performing of analysis tasks. Analytical information also includes external information such as that obtained from outside market and industry sources. Examples of analytical information include trends, aggregated sales amounts by region, product statistics, and future growth projections. Examples of analytical information include the largest growing basket of stocks over the last quarter on the Toronto Stock Exchange (TSX) such as energy stocks, and technology stocks, the most popular destination of travel for British Columbia residents, and projections of cash withdrawals made from chequing accounts for the upcoming holiday weekend. Organisations use analytical information when making important ad hoc decisions such as whether the organisation should build a new manufacturing plant or hire additional sales personnel. DSS can be used on transactional data or analytical data depending on the level and depth of analysis required.

BPR assumes the current process is irrelevant, does not work, or is broken and must be overhauled from scratch. Such a clean slate enables business process designers to disassociate themselves from today's process and focus on a new process. Organisations should be doing re-engineering of business processes to fit Big Data analytics and not putting it off any longer. It is like the designers projecting themselves into the future and asking: What should the process look like? What do customers want it to look like? What do other employees want it to look like? How do best-in-class companies do it? How can a new information system facilitate the process?

5.3 Selecting a Business Process for Re-Engineering: Big Data Analysis Requirements

Organisations are finding it challenging to achieve their objectives and goals due to complicating factors arising from (Anand, 2013; Hanafizadeh & Moosakhani, 2009; Newman & Zhao, 2008):

- Making decisions at the right time
- Intense threat from competition
- Dependencies on suppliers to deliver the right partsat the right time
- Aging workforce with ineffective plans to transfer knowledge to the younger generation
- More stringent regulatory requirements
- Changing economic conditions
- Erratic public opinion
- Increasing environmental concerns.

The old way of doing business is simply no longer effective. For the selection of the process we need to understand the impact of Big Data on the existing processes. By taking a closer look at business processes, including the management of information contained in documents, businesses can more fully realise the benefits of Big Data analytics and data-driven decision-making—just one being increased profitability (Chen et al., 2012). Traditionally the information flow focus is through business processes such as expediting invoices in Accounts Payable, faster employee on-boarding from Human Resources, and more efficient contract management in Legal.

Organisations need to identify business processes which are using unstructured or semi-structured data that need to be analysed for decision making such as email, Power Points, pdf, XML, voice, video, and image. There is valuable business information contained in unstructured documents flowing through many existing business processes. This valuable business information needs to be identified, analysed, and utilised to integrate Big Data solutions with legacy systems and data. Business processes need to be re-engineered to fit unstructured and semi-structured data. The organisations need to determine the critical processes, such as product development, marketing, selling, and customer care, that need to be radically changed in order to realise quick wins in the areas of business performance improvement and/or cost reduction undertakings.

The outcome of the business processes that will enable the organisation to achieve dramatic improvements in performance measures are (Newman & Zhao, 2008):

- Improved customer satisfaction (improved user experience)
- Reduced business and IT costs
- Increased profitability
- Increased responsiveness
- Improved quality of execution and decision making (improved internal processes).

Six Sigma (Tennant, 2001) is a set of techniques and tools for process improvement. It can be used to define a set of best practices, efficiencies, and competitive advantages. Figure 5.1 illustrates the Six Sigma activities which can be employed for business process re-engineering.

The business process must have most of the following characteristics for re-engineering and introducing Big Data analytics:

- The business process must be major contributors to the organisation's core competencies.
- The business process must have the significant impact oncustomers.
- The business process must be ready and feasible for dramatic change due to high cost, and loss of market share.
- The business process must contribute to the organisation's vision and objectives.
- The level of risks for the process domains must beacceptable.
- The process domain change must impact one or more of the following: cycle time, cost, process value, keyissue, supplier performance, beats competition.

- The process domains must have inter-relationships with other functions/departments (i.e. they must be real process domains, not functional units) and the redesign will produce quick wins.

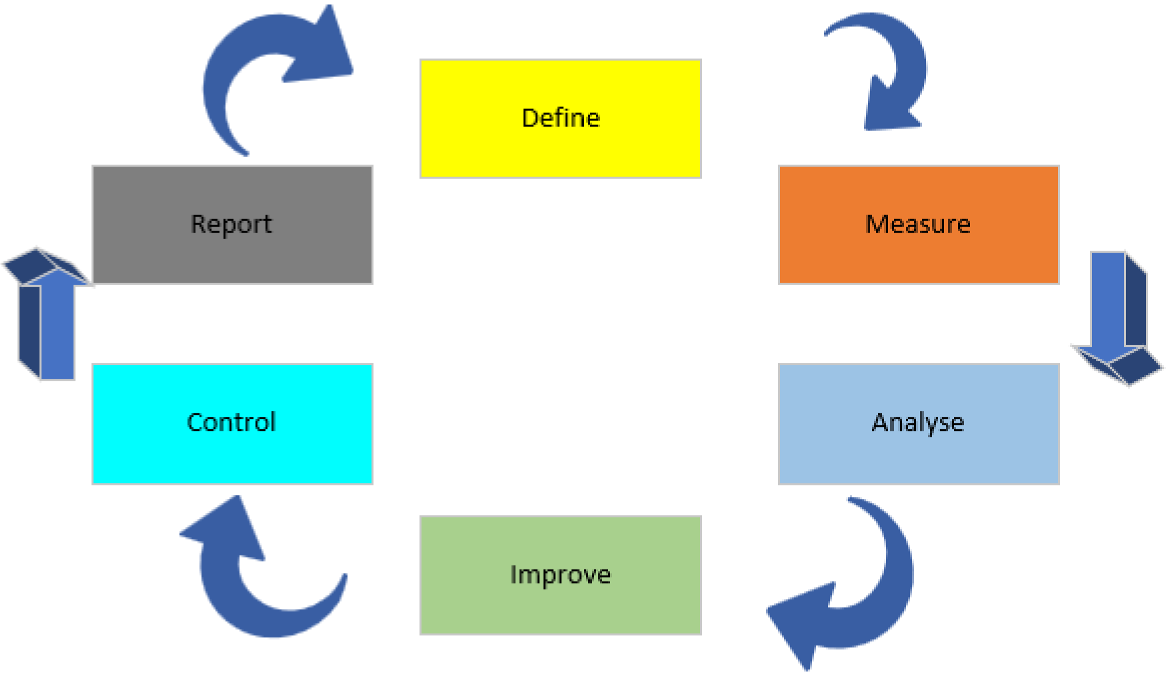


Figure 5.1: Six Sigma Activities for Business Process Re-engineering

5.4 Digitisation of Business Processes: Business and Organisational Impact of Big Data

Big Data solutions have allowed customers, when they log in to their online electricity account, to see a real-time report of their consumption. Organisations are starting to realise that Big Data is more about business transformation than IT transformation.

One of the more significant impacts of Big Data is the organisational change or transformation necessary to support and exploit the Big Data opportunity (Davenport, 2013). The process of capturing the roles, responsibilities, and expectations of the business users, identifying key performance indicators against which the performance of those business processes will be measured, and capturing, aggregating, aligning, cleansing, and making available the data (at the necessary levels of granularity and frequency) to support the monitoring of those business processes are required. Old roles will need to be redefined and new roles introduced, creating both opportunities and anxiety for individuals and organisations alike. Business Intelligence and Operational Intelligence traditionally have focused on understanding key business processes at a detailed enough level so that metrics, reports, dashboards, alerts, and some basic analytics (trending, comparisons) can be built that support those key business processes.

5.5 Architecture for Integration of Legacy Systems and Data with Big Data Solutions

Figure 5.2 shows the architecture for integrating legacy systems and data with Big Data solutions. We require reverse engineering and forward engineering to integrate legacy systems and data with Big Data solutions. The activities involved are:

- Reverse Engineering
- Forward Engineering
- System Integration and data administration.

Reverse Engineering: Reverse engineering supports the integrated analysis and redesign/development activities required to modernise the selected legacy system. Due to the massive and complex nature of software, modernisation must be conducted in multiple phases. Reverse engineering can be divided into the White-Box approach and Black-Box approach. In the White-Box approach, the business-logic is to be extracted through analysis of the legacy system. This approach can be separated into the two domains of Database Reverse Engineering (DBRE) and Procedure Reverse Engineering (PRE).

Database Reverse Engineering is the part of system maintenance work that produces a sufficient understanding of an existing database system and its application domain to allow appropriate changes to be made. Database Reverse Engineering deals with a subset of the problems addressed by software reverse engineering. DBRE recovers domain semantics of an existing database and represents them as a conceptual schema that corresponds to the possible (most likely) design specifications of the database. These design specifications are required when integrating legacy systems and data with Big Data solutions. While the first domain DBRE seems to be mature enough to be considered for the development of DBRE tools, the second Procedure Reverse Engineering is still an unsolved problem (Fyrbiak et al., 2017; Hainau,1998).

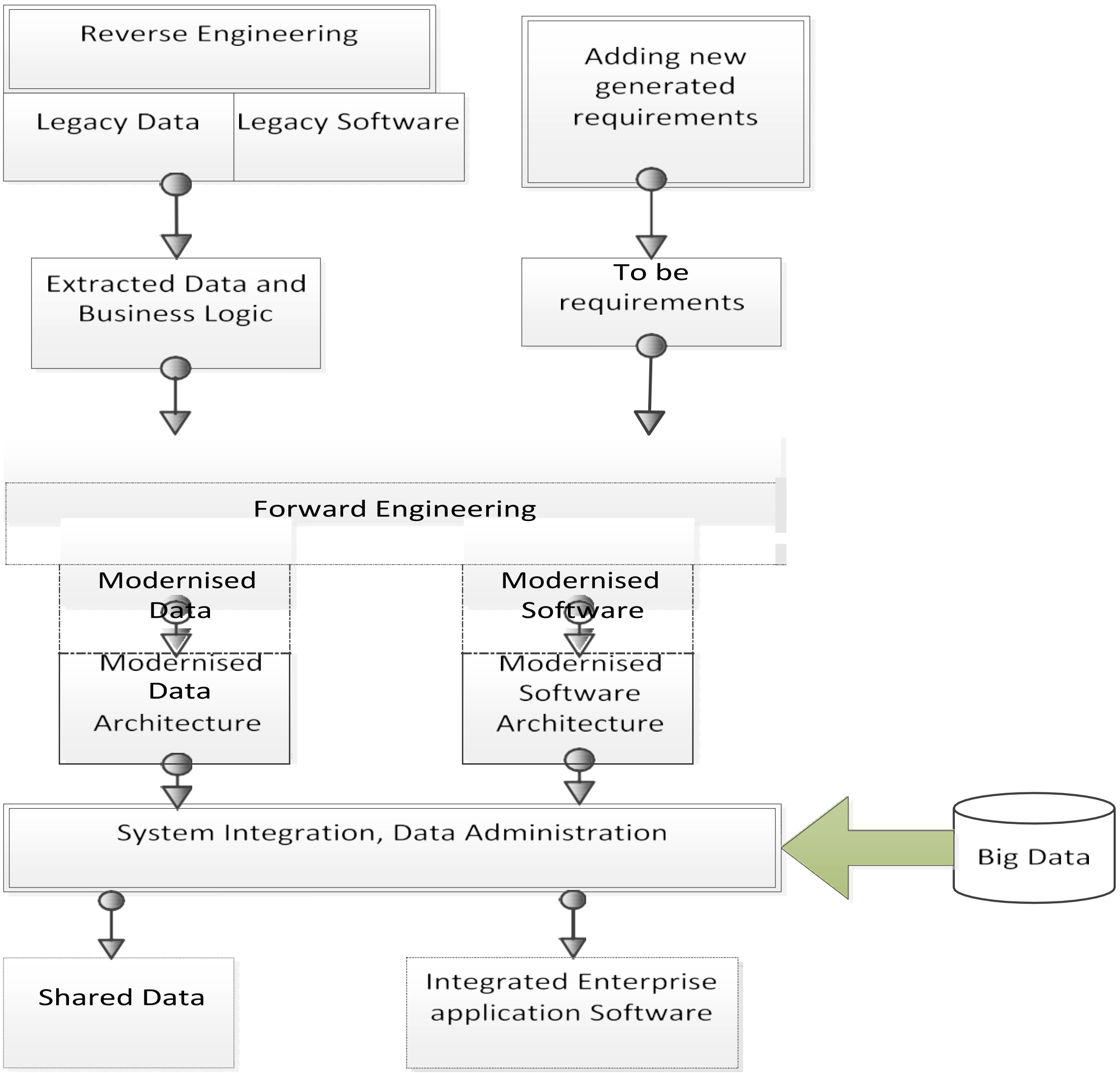

Reverse Engineering
Legacy Data
Legacy Software
Adding new generated requirements
Extracted Data and Business Logic
To be requirements
Forward Engineering
Modernised Data
Modernised Software
Modernised Data Architecture
Modernised Software Architecture
System Integration, Data Administration
Big Data
Shared Data
Integrated Enterprise application Software

Figure 5.2: Architecture for Integrating Legacy Systems and Data with Big Data

Procedure Reverse Engineering deals with analysing and understanding the old code. Analysing and understanding the old code is a difficult task. Some architecture reconstruction tools have been developed to aid in the understanding of such code, but these tools are human interactive and interpretive (Jha et al., 2004; Rahgozar et al., 2002). Jha and O'Brien (2004) have used a software architecture reconstruction tool to document the software architecture of a legacy system.

Forward Engineering: Forward engineering is the traditional process of moving from high-level abstractions and logical, implementation-independent designs to the physical implementation of a system. Forward engineering follows a sequence of going from requirements through designing its implementation. All Software Development Life Cycle is based on Forward Engineering.

The traditional software development life cycle has five phases: Analysis, Design, Coding, Implementation, and Maintenance. Forward Engineering is based on these five phases. One of the newer and more effective ones is called Agile Development. The basic philosophy of Agile Development is that neither team members nor the users completely understand the problems and complexities of a new system, so the project plan and the execution of the project must be responsive to unanticipated issues. It must be agile and flexible (Satzinger et al., 2011).

System Integration and Data Administration: System integration is defined as the process of bringing together the component subsystems into one system and ensuring that the subsystems function together as a system. It is also the process of linking together different computing systems and software applications physically or functionally to act as a coordinated whole. There are different methods of integration, such as Vertical integration, Horizontal Integration, Star integration, and common data format integration.

Vertical Integration is the process of integrating subsystems according to their functionality by creating functional entities also referred to as silos. The benefit of this method is that the integration is performed quickly and involves only the necessary vendors; therefore, this method is cheaper in the short term (Gold-Bernstein et al., 2005). On the other hand, cost-of-ownership can be substantially higher than seen in other methods, since in the case of new or enhanced functionality, the only possible way to implement (scale the system) would be by implementing another silo (Lau, 2005).

Horizontal Integration is also called Enterprise Service Bus (ESB) and is an integration method in which a specialised subsystem is dedicated to communication between other subsystems. This allows cutting the number of connections (interfaces) to only one per subsystem which will connect directly to the ESB. The ESB can translate the interface into another interface. This allows cutting the costs of integration and provides extreme flexibility. With systems integrated using this method, it is possible to completely replace one subsystem with another subsystem which provides similar functionality but exports different interfaces, all this completely transparent for the rest of the subsystems. The only action required is to implement the new interface between the ESB and the new subsystem. The horizontal scheme can be misleading, however, if it is thought that the cost of intermediate data transformation or the cost of shifting responsibility over business logic can be avoided.

Star Integration (Gold-Bernstein et al., 2005), also known as Spaghetti Integration, is a process of integration of the systems where each system is interconnected to each of the remaining subsystems. When observed from the perspective of the subsystem which is being integrated, the connections are reminiscent of a star, but when the overall diagram of the system is presented, the connections look like spaghetti, hence the name of this method. The cost varies because of the interfaces that subsystems are exporting.

In a case where the subsystems are exporting heterogeneous or proprietary interfaces, the integration cost can substantially rise. Time and costs needed to integrate the systems increase exponentially when adding additional subsystems. From the feature perspective, this method often seems preferable, due to the extreme flexibility of the reuse of functionality. A common data format is an integration method to avoid every adapter having to convert data to/from every other applications' formats, Enterprise Application Integration (EAI) systems usually stipulate an application-independent (or common) data format. The EAI system usually provides a data transformation service as well to help convert between application-specific and common formats.

5.6 Combining Big Data Analytics with Business Processes

To apply Six Sigma (Tennant, 2001) activities for business process re-engineering an organisation needs to develop its Big Data strategy for re-engineering. The Big Data strategy for re-engineering is shown in Figure 5.3.

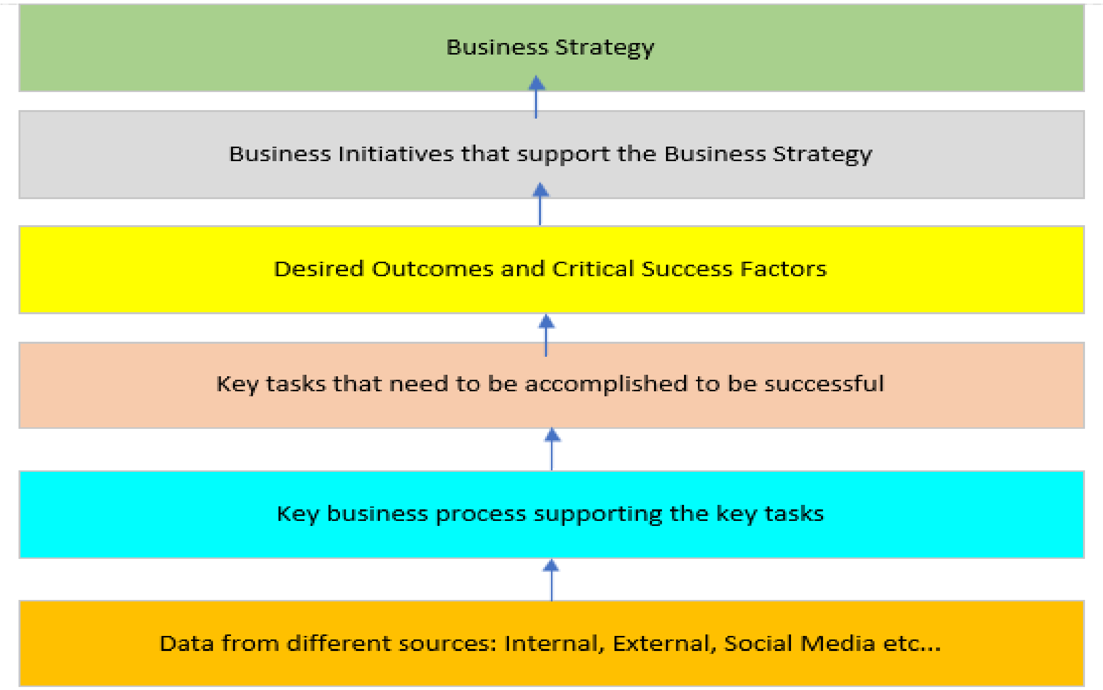

Business Strategy
Business Initiatives that support the Business Strategy
Desired Outcomes and Critical Success Factors
Key tasks that need to be accomplished to be successful
Key business process supporting the key tasks
Data from different sources: Internal, External, Social Media etc...

Figure 5.3: Big Data Strategy for Re-engineering

We have identified seven processes of an e-business process model as shown in Figure 5.4, which are:

1. Process of customer service
2. Process to ship product
3. Process to quality assurance
4. Process of credit card transaction
5. Process to extract customer information in/out of database
6. Data Processing
7. Process to extract information from external data sources.

The targeted business strategy clearly defines the scope upon which the Big Data initiative is focused. The title should provide enough detail to clearly identify the overall business objective; for example, “improve customer intimacy” or “reduce operational maintenance costs” or “improve new product launch effectiveness”. Business initiatives should support the business strategy. A business initiative is defined as a cross-functional project.

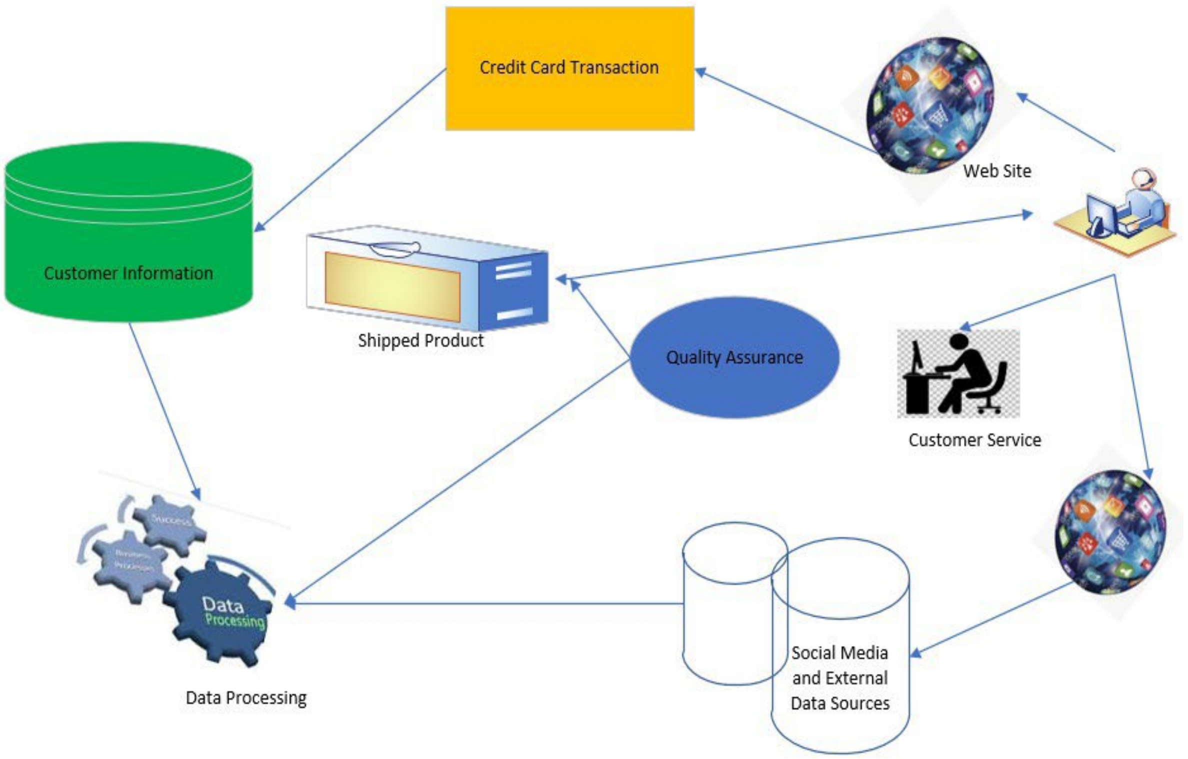


Figure 5.4: E-business Process Model

We worked on the process "Customer Service" to target the Business Strategy called "improve customer intimacy". To support the business strategy, the key business initiatives we focused on are:

- Acknowledge key customers.
- Acknowledge products.
- Acknowledge suppliers.
- Acknowledge market roles and responsibilities.

Desired Outcomes and Critical Success Factors (CSF) captured the outcomes and critical success factors necessary to support the successful execution of the organisation's key business initiatives. Outcomes define the desired or ideal end state. Critical success factors define "what needs to be done" for the business initiative to be successful. To support key business initiatives identified we focused on developing an intimate knowledge of the customer's life stage and their behaviour.

Key tasks that need to be accomplished to be successful provide the next level of detail by documenting the specific tasks that need to be executed to perfection to successfully support the targeted business initiatives. At this point different processes of the organisation need to collaborate to achieve the business initiatives already identified. This is “how to do it”. So how “to develop an intimate knowledge of customer’s life stage and their behaviour” is a challenging task. This is where we need to address different business processes which can collaborate to get the target result for the identified business task. The identified key tasks for “to develop an intimate knowledge of customer’s life stage and their behaviour” are:

- Collect information.
- Generate business intelligence.
- Contact relevant customers.
- Track operational intelligence for identified customers.

Key business processes support the identified key tasks by generating required output from the processes. A business process: has a goal; has specific inputs; has specific outputs; uses resources and many activities that are performed in some order; may affect more than one organisational unit; and creates value.

The required output is dependent on the input. For the identified key tasks, the business processes are:

- Process to collect data from different data sources.
- Process to format data.
- Process to clean dataset for analysis.
- Process to communicate visual report.

These processes are existing processes in the organisational structure. The only change with the advent of Big Data is that the existing processes are not capable to integrate external data sources.

Data from different data sources is required, but this also requires the “Right” data sources to be picked. The value is found in layering data from multiple sources, not in simply adding bulk to the final data count (Baker, 2015). Internal data sources are helpful in solving a number of business problems but by working in isolation and applying analysis only to internal data, it may improve existing processes and products that are on their way to becoming obsolete or unprofitable.

Data from different sources are required. However, hoarding data either from external or internal sources drives up the cost and potential liabilities. Transactional data reveals past information and may or may not have any bearing on future information. This can be resolved by analysing social media data. While social media analysis can provide many useful insights, these have largely proved unsatisfactory when used as the sole source for analysis because the data is often incomplete and mostlyreveals correlation rather than causation.

Choosing the "Right Data" source for decision making should enhance any existing analysis projects. It is not merely adding more data but understanding what variety of data are collected from different data sources required for the analysis. Data scientists need to first identify what additional variables could refine or improve the analytical result. By understanding what kind of data is required for analysis, it will help in locating and choosing the "Right" external data sources.

To collect information, we have identified the following data sources:

- Customers' reviews of data source.
- People looking for the similar product.
- Web Logs.
- Images and videos.

We integrated Big Data sources for processing and analysing so that business intelligence could be created. The following are the internal and external sources of Big Data which were required to be integrated to the Big Data storage, processing and analysing platform, as shown in Figure 5.5. As discussed in Chapter 4, Big Data sources were Web Logs, My Data, Open Data, and Internal Enterprise Data.

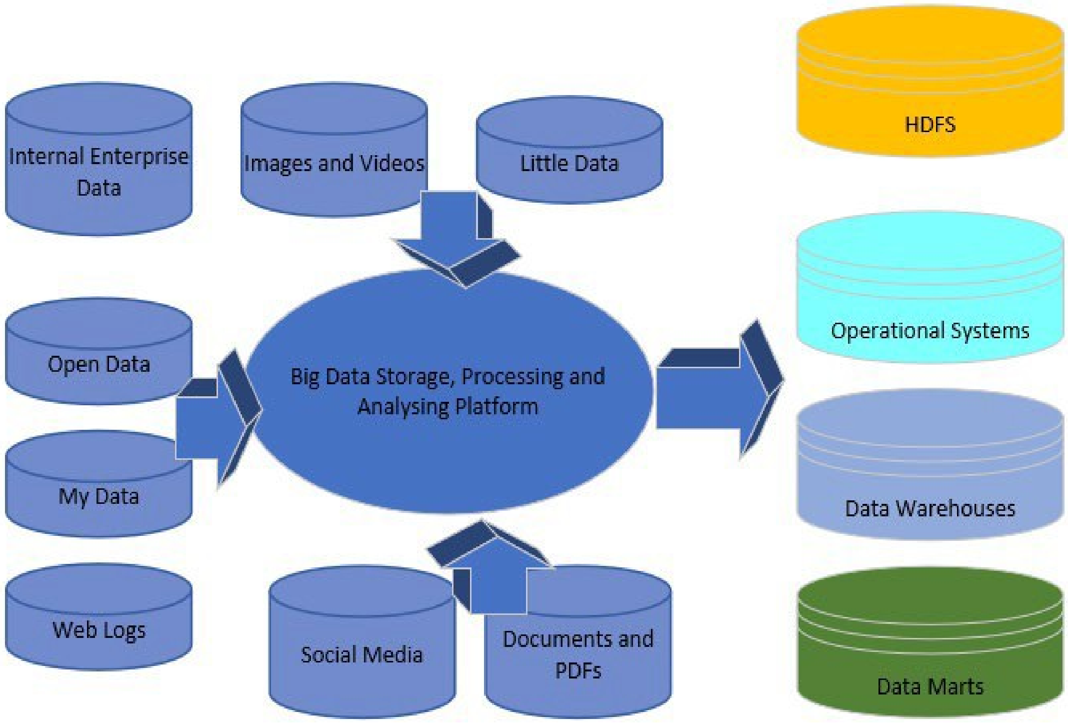


Figure 5.5: Integrating Big Data Sources

Once the Big Data is being collected, processed and analysed, it was stored in Hadoop Distributed File System (HDFS), from where data can be used for business intelligence, reporting, online analytical processing (OLAP), ad-hoc reporting and modelling.

5.7 Chapter Summary

This chapter discussed how the Big Data analytics can be combined with the business process using re-engineering to integrate legacy systems and data with Big Data Solutions using a Big Data architecture. It described Big Data analysis requirements by selecting a business process for re-engineering, business, and organisational impact of Big Data on business processes, and tasks to combine Big Data analytics with business processes.

Businesses and organisations from all sectors have gained critical insight from the structured data collected through various enterprise systems and analysed by commercial relational database management systems. Organisations should not let existing data warehouse and existing business intelligence processes, which are insufficient for today's deep, wide, and diverse data sources, hold the organisation back. Big Data analytics requires business processes to change and it must align with the organisation's IT infrastructure to support the business initiatives. New ways of doing data analytics and business intelligence impact on technology infrastructure components. Organisations need to focus on this now rather than later to gain competitive advantage in the marketplace.

CHAPTER 6: USE Of BIG DATA ARCHITECTURE IN COMPLEX EVENT PROCESSING

6.1 Introduction

This chapter discusses a Complex Event Processing (CEP) architecture developed to be used in the architecture for integrating legacy systems and data with Big Data solutions using open source technologies and shows how CEP architecture is applied to the use case of an Electronic Coupon Distribution Service (ECDS). In CEP architecture we have used legacy data such as member and store data, stored in a database and pulled into the CEP using location information, past shopping/travel history, gender, and likes/dislikes. This chapter further discusses how different types of data such as static information on gender, age, previous history (where the person travelled to, and what they bought), as well as real-time information about a customer's current location, and current shopping habits, would all be utilised in CEP and suggest how a Big Data architecture to integrate legacy systems and data with Big Data Solutions would work in this scenario. In this chapter we will discuss the architecture developed for CEP using open source technologies such as Hadoop, Apache Kafka, Spark, and Scala as a language and will discuss how CEP is applied to the use case of an ECDS.

Complex Event Processing (CEP) is a technique for tracking, analysing, and processing data as an event happens and is useful for Big Data because it is intended to manage data in motion. CEP helps to aggregate a lot of different information useful for identification and analysis for the cause-and-effect relationships among events in real time. CEP is a type of event processing that combines data from multiple sources to identify patterns and complex relationships across various events. Incoming events in the form of data is matched to provide a continuous pattern of the insights that is happening. CEP is the key part of multiple processes that run parallel to improve the performance. Data in motion is processed and communicated based on business rules and processes. For decisions to be better informed, data used for decision making must be timely, complete, accurate, trusted, valid, reliable, and relevant. CEP utilises data generated from moment-to- moment from different emerging sources such as sensor, sentiment, and geo-locational. There is a need to bridge the gap between traditional business intelligence and new Big Data technologies such as CEP. Bridging of this gap will enable organisations to become agile and data-driven so that business outcomes can be maximised by delivering better-informed decisions about a customer and delivering a better service to them.

6.2 Background

CEP has evolved into the paradigm of choice for the development of monitoring and reactive applications. It also has a strong impact on information systems and the way information is subscribed and consumed (Buchmann et al., 2019). CEP enables real-time analysis by utilising stream data generated from moment-to-moment to support better insight and decision making. With the recent explosion in data volume, variety, velocity, and diversity of data sources, this goal can be quite challenging for architects to achieve. Organisations like Macy's and Kohl's switched from sending shoppers blanket email promotions to sending targeted offers based on individual shopper purchases (Thau,2014).

CEP is a type of event processing that combines data from multiple sources to identify patterns and complex relationships across various events. The value of CEP is that it helps identify opportunities and threats across many data sources and provides real-time alerts to act on them. Today, CEP is used across many industries in a variety of use cases, including:

- Finance: Trade analysis and fraud detection
- Airlines: Operations monitoring
- Healthcare: Claims processing and patient monitoring
- Energy and Telecommunications: Outage detection.

CEP was developed at Stanford University in the mid-1990s by Professor David Luckham (2002). The goal of CEP is to enable information contained in the events flowing through all of the layers of the enterprise IT infrastructure to be discovered, understood in terms of its impact on high-level management goals and business processes, and acted upon in real-time to make well-informed decisions. CEP implementations are built around events such as a new purchase, a change of address, or an attempt to break into a network. Events can come from people, devices, applications, networks, or databases. Events can generate responses, or actions. For example, an "Attempted Fraud" event may trigger, in some cases, a "Put Account on Referral" action to make sure downstream account activity is legitimate. The need of CEP represents a paradigm shift in the approach to understanding and responding to business activity to make well-informed decisions through ITinfrastructure.

Coupons are certificates that entitle the bearer to stated savings on the purchase of a specific product or product bundle. Conventionally, manufacturers and merchants distribute coupons via newspaper inserts, in magazines, or by direct mail. To get the rebates using coupons usually require the customer to redeem the rebate certificate by mailing it to the manufacturer along with proof of purchase. Some of the popular uses for coupons include: promoting a new brand (manufacturer-issued coupons); persuade customers to switch to the promoted brand (manufacturer-issued coupons); increase sales of an existing product (both manufacturer- and merchant-issued coupons); and attract shoppers to a retail establishment (merchant-issued coupons).

The conventional approach to distribute coupons is to issue identical coupons regularly to all customers, as shown in figure 6.1. The conventional approach to distribute coupons faces many disadvantages, such as:

- Conventional distribution systems are slow and have long lead-times.
- Coupons do not get to the targeted customers.
- Redemption rates are low, 1% of the distributed coupons (Anand et. al.,1998).

The coupon concept has not been adopted widely on the Internet. Several websites offer printable versions of conventional coupons; these coupons cannot be redeemed online (Kotler et al., 1998). E-coupon issuers enjoy a high degree of flexibility in choosing which e-coupons are given to shoppers and when they are offered. For example, e-coupons could be offered to shoppers when they enter an online store, when they view a product description, or when they finalise their purchases. Similarly, e-coupons could be offered for a product for which a shopper has expressed interest, a product related to the product a shopper is buying, or a product the shopper never buys but the storekeeper is interested in promoting.

SASE (Secure Access Service Edge) (Wu et al., 2006), an event processing system that executes complex event queries over real-time streams of radio-frequency identification (RFID) readings, extending existing event languages to meet the needs of RFID-enabled monitoring applications, has been proposed by Wu, Diaoa and Risvi. Agrawal et al. (2008) present an evaluation model and query evaluation architecture for pattern matching over CEP that allows for optimisation and techniques to improve runtime efficiency.

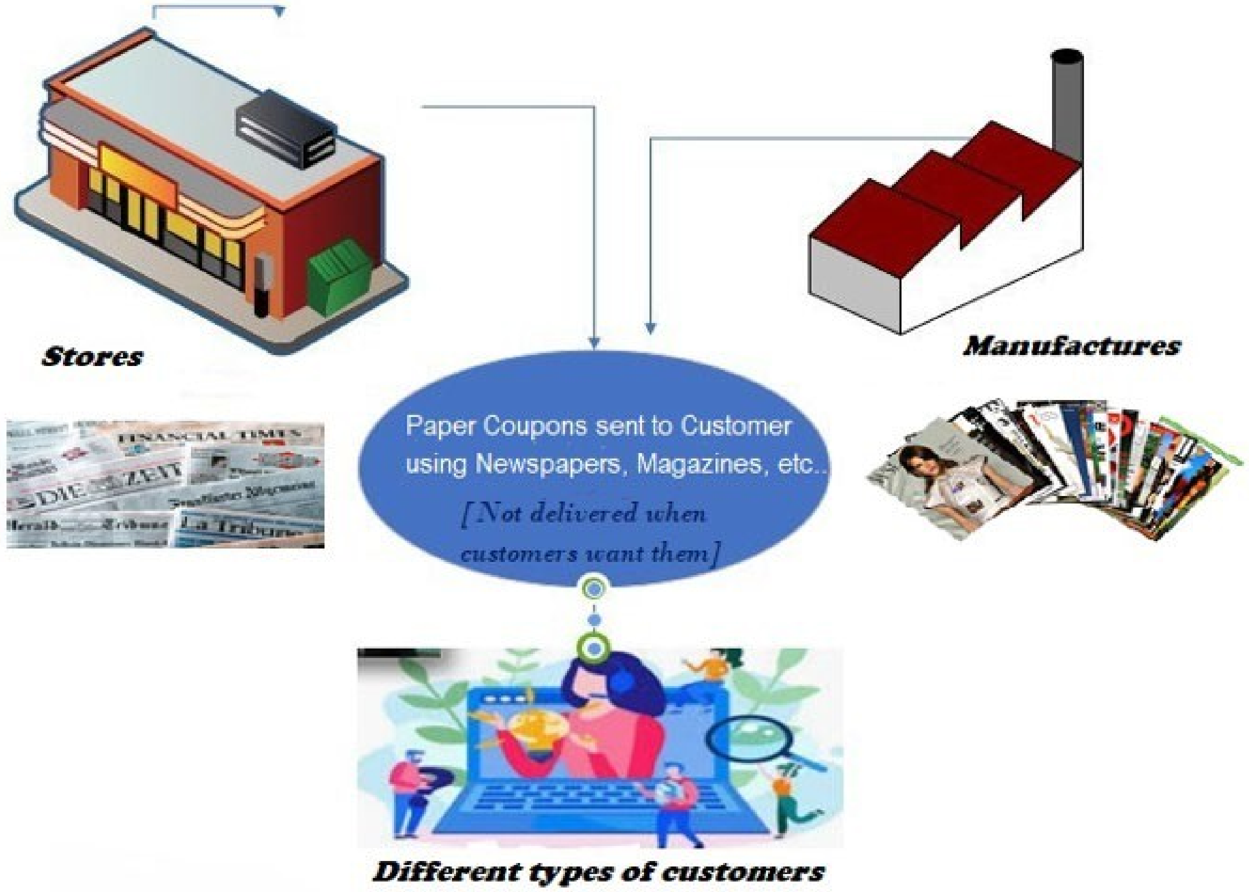

Stores
Manufactures
Paper Coupons sent to Customer using Newspapers, Magazines, etc..
[Not delivered when customers want them]
Different types of customers

Figure 6.1: Conventional Way of Coupon Distribution

The Cayuga System (Brenna et al., 2007) was built at Cornell University as a high-performance system for complex event processing. Cayuga is based on nondeterministic finite automata with buffers. Bry and Eckert (2007) claim that a sufficiently expressive language must be able to handle data (event object property) extraction, event composition (matching multiple events), temporal and causal relationships, and event accumulation such as aggregation of data over time, or checks for missing events, and order not filled in at a specific time.

Søberg et al. (2008) devise a CEP system that detects deviations from expected events. Their paper describes a query language to detect normal deviations and is open to sensors, distributes event processing to optimise resource utilisation, supports complex temporal and spatial events. The CEP provides possibility to query complex events and deviations from these complex events by generating regular patterns and identifying deviations from these general patterns.

As stated above several CEP solutions have been proposed and deployment strategies have been introduced by processing primitive events generated by sources, extracting new knowledge in the form of composite events, and delivering them to interested sinks (Cugola & Margara, 2013). The CEP can be internally built around several, distributed processors, connected together to form an overlay network, and cooperating to provide the processing and routing service. Unfortunately, none of the identified work utilises real-time analytics for decision making or uses open source software to build CEP. Real-time systems need to perform analytics on short time windows, that is correlating and predicting event streams generated over the last few minutes, which is different to batch processing systems. However, to make decisions these two types of systems, real-time and batch processing, need to be combined so that decisions are well informed. For instance, a credit card fraud prediction system could leverage a system using previous credit card transaction data over a defined period. This can be combined with a real-time system to find if there is any deviation in the real-time stream data. If a deviation is beyond a certain threshold, it can be tagged as an anomaly. There are several open source technologies available to process data in batch form and in real-time. Combining these two will create an architecture for CEP.

Predictive Analytics is established on a set of core values and guiding principles. It is not a rigid or prescriptive methodology; rather it is a style of building business intelligence applications, and analytics applications that focuses on the early and continuous delivery of business value (Collier, 2012). The Manifesto for Predictive Analytics is based on: individuals and interactions over process and tools; end-user and stakeholder collaboration over contract negotiation; and responding to change over following a plan.

6.3 Open Source Technologies for CEP

The following section presents the description of open-source technologies for CEP.

6.3.1 Hadoop

As discussed in chapter 4, the open source technologies for CEP used are Hadoop Distributed File System (HDFS) (https://hadoop.apache.org/docs/current/api/), MapReduce to distribute a query to different nodes, and to gather the results and resolves them into a single value using JobTracker, TaskTracker, and JobHistoryServer, and Not Only SQL (NoSQL) for processing large volumes of multi-structured data. NoSQL databases are most adept at handling discrete data stored among multi-structured data. However, some NoSQL databases, like HBase, can work concurrently with Hadoop.

NoSQL is better suited for operational tasks, interactive workloads based on selective criteria where data can be processed in near real-time. Hadoop is better suited to high-throughput and in-depth analysis. Hadoop and NoSQL products are sometimes marketed concurrently. Some big names in NoSQL field include Apache Cassandra, MongoDB, and Oracle NoSQL. Many of the most widely used NoSQL technologies are open source, meaning security and troubleshooting may be an issue. It also places less focus on

atomicity and consistency than on performance and scalability. Premium packages of NoSQL databases (such as Datastax for Cassandra) work to address these issues.

Massively Parallel Processing (MPP) Databases work by segmenting data across multiple nodes and processing these segments of data in parallel. Whereas Hadoop usually runs on cheaper clusters of commodity servers (allowing for inexpensive horizontal scale out), most MPP databases run on expensive specialised hardware (data warehouse appliances). MPP technologies process massive amounts of data in parallel. It may have hundreds (or potentially even thousands) of processors, each with their own operating system and memory, working on different parts of the same programme. MPP uses SQL, and Hadoop uses Java as default (although the Apache Foundation developed Hive, a language used in Hadoop like SQL, to make using Hadoop slightly easier and less specialist). Many of the major players in the MPP market have been acquired by technology vendors. Netessa, for instance, is owned by IBM, Vertica is owned by HP, and Greenplum is owned by EMC. Hadoop is a high-throughput system which can crunch a huge volume of data using a distributed parallel processing paradigm called MapReduce. But there are many use cases across various domains which require real-time/near real-time response on Big Data for faster decision making. Hadoop can be used for building a prediction model for sequence analysis with the help of the Machine Learning library.

6.3.2 Apache Spark Core Engine Spark APIs

Traditional data warehouses have focused on support for strategic Business Intelligence (BI). In operational data warehousing, the closer the warehouse is to real-time information, the more actionable it becomes for front-line users. CEP engines are utilised for rapid and large-scale data processing in real time.

One open-source CEP solution is the Apache Spark architecture. Apache Spark is used on top of HDFS and promises speeds up to 100 times faster than the two step MapReduce function. This allows data to be loaded in memory and queried repeatedly, making it suitable for machine-learning algorithms. An increase in performance is obtained by leveraging computations in-memory.

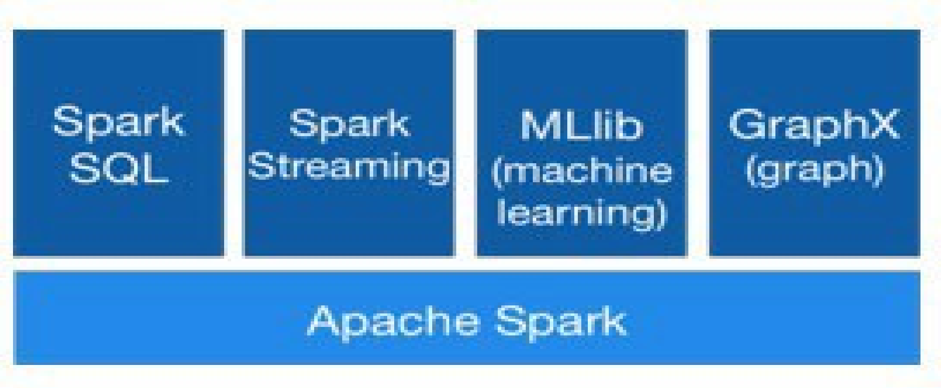


Figure 6.2: Apache Spark and Stack of Libraries

Image used from https://spark.apache.org/

Apache Spark is a fast and general engine for large- scale data processing. Apache Spark runs on Hadoop, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3. Apache Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming as shown in Figure 6.2. These libraries can be combined seamlessly in the same application. Apache Spark can be used interactively from the Scala, Python and R shells. Spark has an advanced Directed Acyclic Graph (DAG) execution engine that supports cyclic data flow and in-memory computing.

Spark has been adopted by several companies in the industry. To mention a few, Guavus (2014) has built its operational intelligence platform on Spark (Carr, 2014), Soomdata (Langseth, 2014) is using SparkSQL to do business intelligence-style analytics, and Graphflow (Pentreath, 2014) has used Spark to build a real-time recommendation and customer intelligence platform. The Spark engine is a multi-faceted tool that provides a suite of packages to build a variety of online streaming, batch processing, and machine-learning applications.

6.3.3 Spark APIs: Scala or Python for Spark

Spark provides an API for distributed data analysis and processing in four different languages: Scala, Java, Python, and R. Java is complex to learn and quite verbose and does not support Read-Evaluate-Print Loop (REPL) interactive shell. Python is a general-purpose programming language with excellent libraries for data analysis like Pandas and scikit-learn. But like R, it is still limited to working with an amount of data that can fit on one machine.

Scala is less complex than Java but more complex than Python. Scala presents a learning curve. But at least, any Java library can be used from within Scala. R offers a rich environment for statistical analysis and machine learning, but it has some rough edges when performing many of the data processing and cleanup tasks before the real analysis work. DataFrame makes Spark programs more concise and easier to understand, and at the same time exposes more application semantics to the engine. A comparative study of Scala and Python based on some attributes, such as performance, learning curve, Ease of Use, Libraries, and Porting R Code, is shown in Table 6.1.

Spark is built on Scala, thus being proficient in Scala helps in digging into the source code when something does not work as expected. Spark codes written in Scala running on a Java Virtual Machine (JVM) when called by Python wrapper might be the source of more bugs and issues while translation between two different environments (Scala and JVM) and languages. Since Spark is implemented in Scala, using Scala allows accessing the latest features.

Attributes	***Scala***	***Python***
Performance	+ (uses Java Virtual Machine)	- (slower than C)
Learning curve	- (Java alternative)	+ (large community support and tutorials)
Ease of use	- (complex to learn)	+ (grand library collection from community)
Libraries	- (small libraries for Machine Learning)	+ (better libraries in Machine Learning and Natural Language Processing)
Porting R code	- (complex way for calling R routines)	+ (easy ways to call R directly from Python)
Type of systems used in	Good for large scale systems	Good for simple to moderately complex analysis and for a quick demo
Typed language	Scala is statically typed. But looks like dynamic-typed language because it uses a sophisticated type inference mechanism.	Python is dynamically typed
Spark streaming data	Scala is preferred. Scala helps in digging into the source code	Python supports for Spark streaming but only for basic sources like text files and text data over sockets.
Kind of programming language	Scala is multi- paradigm programming language that promotes usage of functional principles.	Python is an interpreted language hence slow.

Table 6.1: Comparisons of Scala and Python for Spark

Most features are first available on Scala and then port to Python. Scala is designed for distributed systems. Hence performance is better than with traditional languages like Python and R. Scala is being integrated well with the big data ecosystem, which is mostly JVM-based. CEP challenges are complex. Hadoop-like systems handle the Volume and Variety aspects of Big Data. However, CEP systems also need to handle the Velocity aspect.

6.3.4 Apache Kafka

Kafka is a tool that is built to handle ingesting transaction logs and other real-time data feeds. Kafka works well as a replacement for a more traditional message broker. Message brokers are used for a variety of reasons such as to decouple processing from data producers, and to buffer unprocessed messages. In comparison to most messaging systems, Kafka has better throughput, built-in partitioning, replication, and fault-tolerance which makes it a good solution for large-scale message-processing applications.

The original use case for Kafka was to be able to rebuild a user activity tracking pipeline as a set of real-time publish-subscribe feeds. This means site activity (page views,

searches, or other actions users may take) is published to central topics with one topic per activity type. These feeds are available for subscription for a range of use cases including real-time processing, real-time monitoring, and loading into Hadoop or offline data warehousing systems for offline processing and reporting. Kafka is often used for operation monitoring data pipelines. This involves aggregating statistics from distributed applications to produce centralised feeds of operational data.

Many people use Kafka as a replacement for a log aggregation solution. Log aggregation typically collects physical log files from servers and puts them in a central place (a file server or HDFS perhaps) for processing. Kafka abstracts away the details of files and gives a cleaner abstraction of log or event data as a stream of messages. This allows for lower-latency processing and easier support for multiple data sources and distributed data consumption. In comparison to log-centric systems like Scribe or Flume, Kafka offers equally good performance, stronger durability guarantees due to replication, and much lower end-to-end latency.

6.4 Electronic Coupon Distribution Service Workflow Model

Electronic coupons are one of the ways to raise popularity for a service, product, or brand. Electronic coupons (or special offers, as most vendors prefer to call them) are generally embraced by customers, regardless of financial potential. One of the challenges of Electronic Coupon Distribution Service (ECDS) using locational information is distribution of individually tailored coupons and promotions, with real time analysis of customer information and present location. Customers are moving with mobile devices and looking for certain types of product or stores. If the ECDS delivers the coupon to the customers on their mobile device at right time, and at right location, it will improve the ability to attract customers into shops. The objectives of implementing ECDS are to have highest customer satisfaction rating and improvement of the ability to attract customers into shops to increase sales and hence profit for the organisation. Figure 6.3 depicts a workflow model of ECDS.

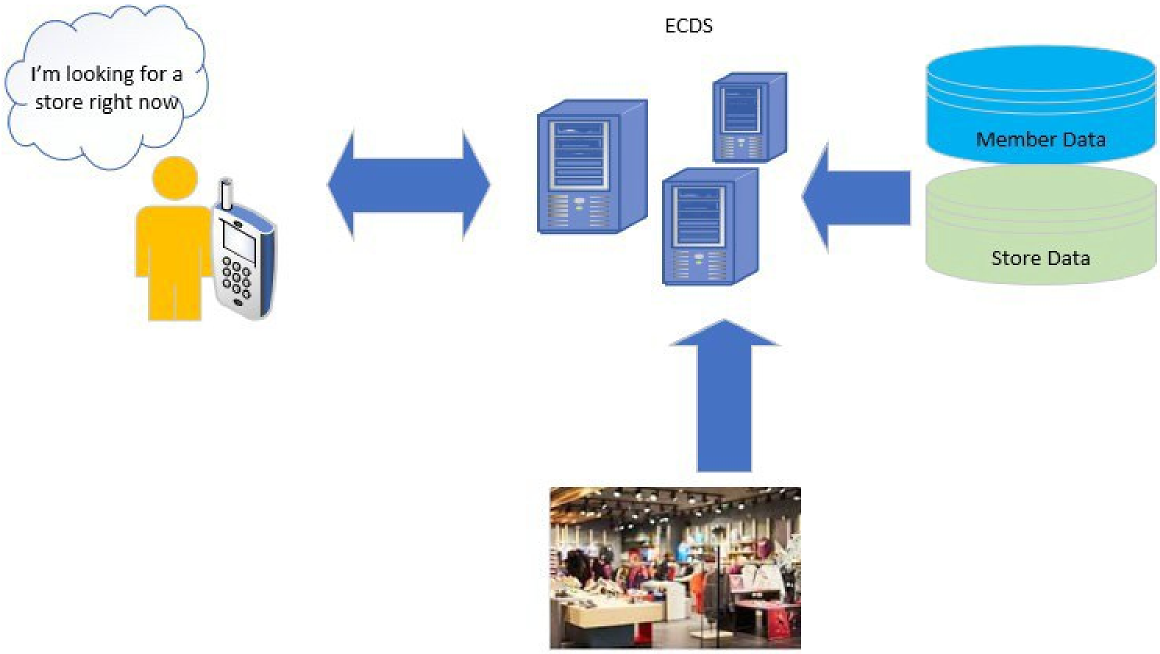


Figure 6.3: Workflow Model of ECDS

To accomplish the workflow model of ECDS the components required to construct the architecture to support ECDS are as follows:

- Sensing Devices such as mobile, Facebook, Twitter, creating structured and unstructured data.
- Data Platform service for storage of structured data, CEP, Hadoop.
- Navigation to internal or external Application hosted oncloud.
- End-customers and customers' systems where the coupons will be delivered.

The components of the architecture are shown in Figure 6.4. The sensors send the locational information to CEP and a high-performance filter is used for processing. CEP takes Member information and Store information from storage of structured data and Hadoop and processes the data with some additional rules to make decisions and then recommendations are being navigated to end users and customer systems.

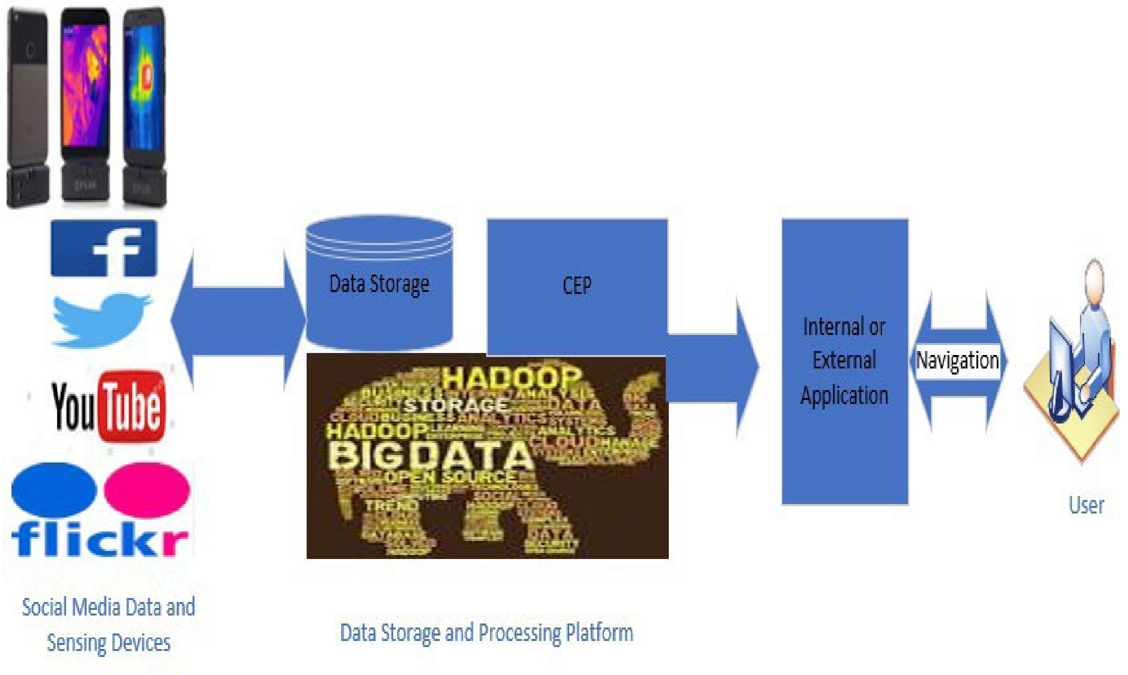

Data Storage
CEP
HADOOP
STORAGE
BIGDATA
OPEN SOURCE
Internal or External Application
Navigation
User
Social Media Data and Sensing Devices
Data Storage and Processing Platform

Figure 6.4: Architectural Components

Luckham (2002) gives the following steps for a design of a CEP system:

- Design a new process.
- Convey design to stakeholders, and form consensus.
- Simulate on expected data, and update design.
- Integrate into system, test, and update design.
- Monitor upgraded system.
- Modify system based on monitored results or business requirements.

6.5 Architecture for CEP for ECDS

Big Data Architecture for CEP requires capturing clickstream. A clickstream is a recording of the parts of the screen a computer user/mobile user clicks on while Web browsing (Zhou et al., 2015) as shown in Figure 6.5 .

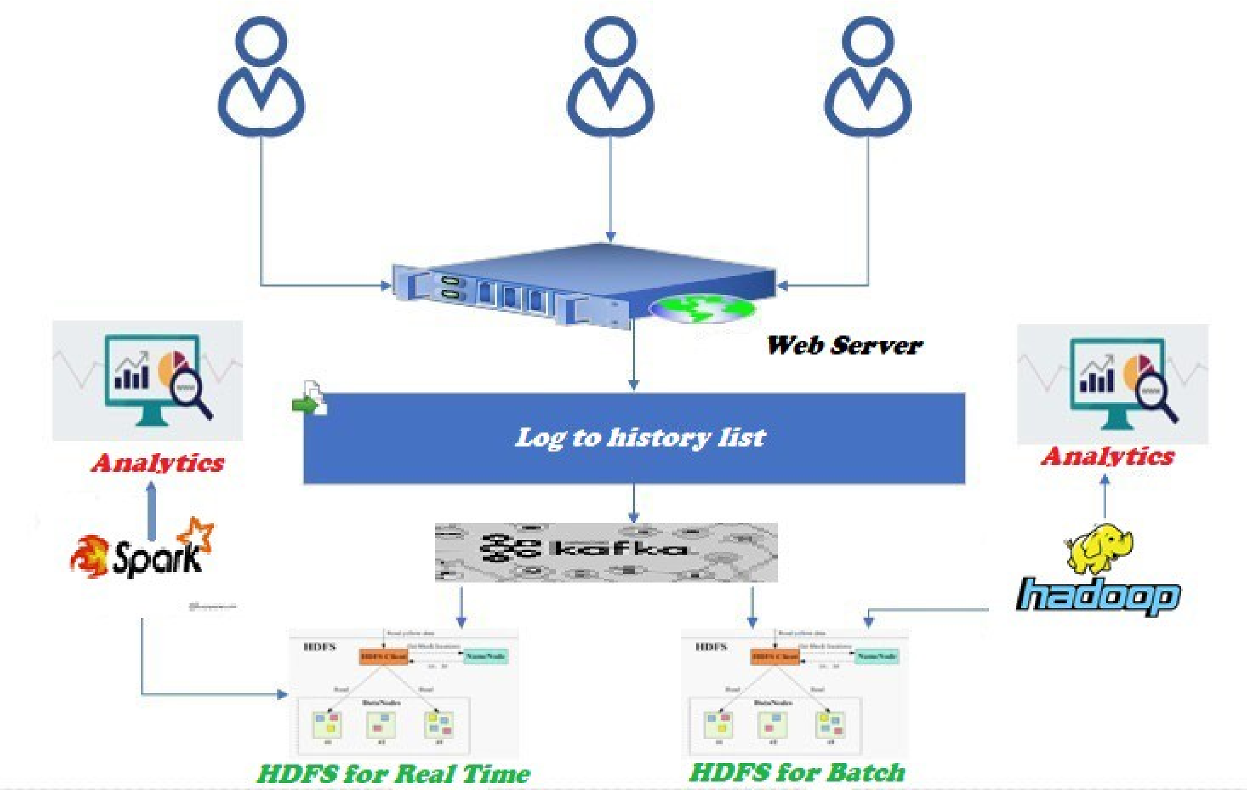


Figure 6.5: Big Data Architecture of CEP for Coupon Distribution

Hadoop was used for batch processing and Spark is used for real-time processing. Spark lacks its own distributed storage system, so it was installed on top of Apache Hadoop and used Hadoop Distributed File System for storing data. Coupons are sent based on location information, store status, previous history and profile of the customer, gender, age, previous history of what they bought, and real- time information about a customer. To send the coupon to interested customers requires processing of locational data, members data, and store information data. Members and store information requires filtering large volumes of location data which is being generated by log files. Locational information (real-time event data) is being sensed by sensing devices such as mobiles, and member and store information is stored in the database. These two sets of data need to be combined with various rules to develop recommendations and make decisions for the customers as shown in Figure 6.6.

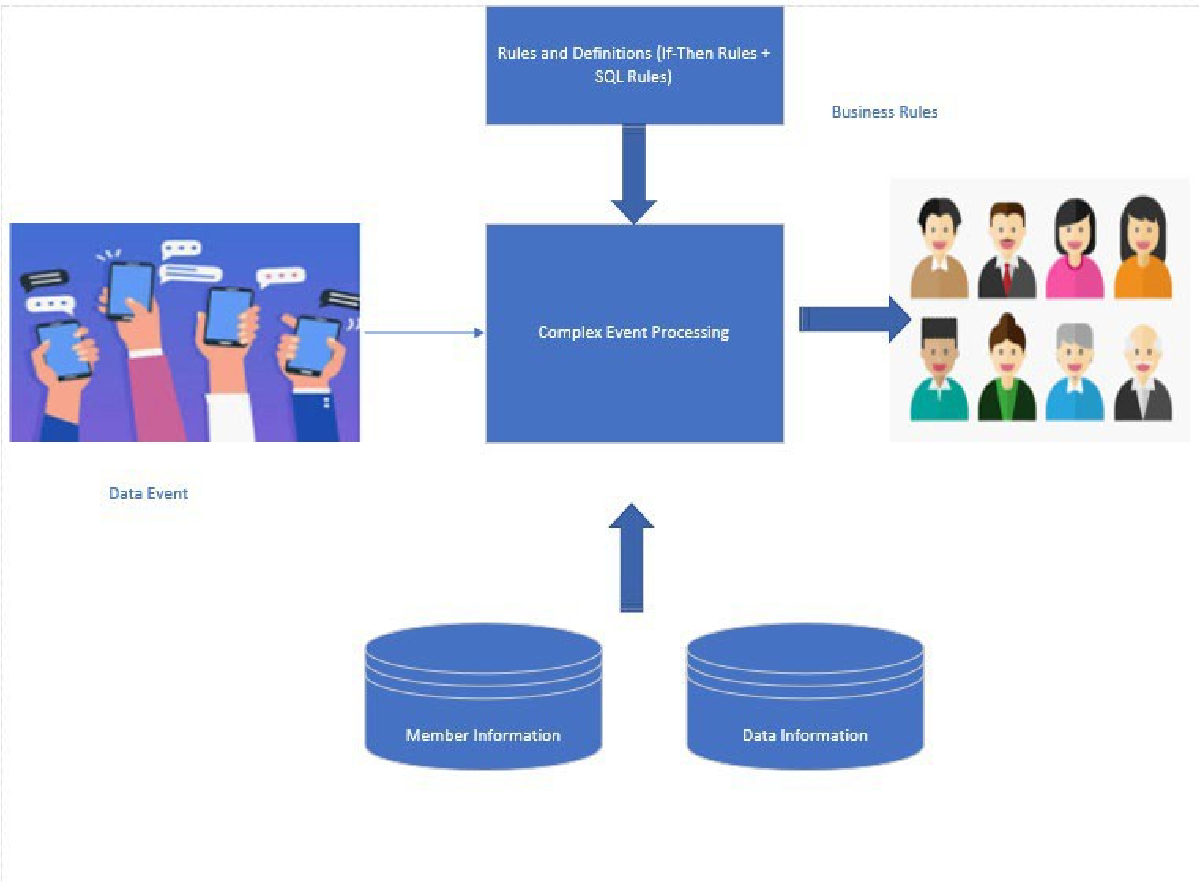


Figure 6.6: Information Processing using CEP

Business rules are frequently defined based on the occurrence of scenarios triggered by events (Cugola & Margara, 2012). Rules must be set in advance to execute complex event processing; for example, take an action if the customer is identified as a loyal customer based on his/her purchase habits. Business rules and event processing queries change frequently and require immediate response for the business to adapt itself to new market conditions, new regulations, and new enterprise policies. Business rules can be divided into two types: If-Then-Else and SQL rules. The rules are based on these three components:

- ***Event:*** defines the sources that can be considered as event generators, such as sensors.
- ***Condition:*** specifies when an event must be considered; for example, we can be interested in some data only if it exceeds a predefined limit.

- ***Action:*** identifies the set of tasks that should be executed as a response to an event detection; some systems only allow the modification of an internal database, while others allow the software application to be notified about the identifiedsituation.

To execute the rules at runtime, five phases have been identified (Thau, 2014). These are as follows:

- ***Signalling***: detection of an event
- ***Triggering***: association of an event with the set of rules defined for it
- ***Evaluation*****:** evaluation of the conditional part for each triggeredrule
- ***Scheduling***: definition of an execution order between selected rules
- ***Execution*****:** execution of all the actions associated with selected rules.

CEP puts great emphasis on the issue and ability to detect complex patterns of incoming data involving sequencing and ordering relationships. CEP relies on the ability to specify composite events through event patterns that match incoming event notifications based on their content and on some ordering relationships on them. CEP requires interaction with many distributed and heterogeneous information data sources and sinks which observe the external world and operate on it. This is typical of most CEP scenarios, such as environmental monitoring, business process automation, and control systems.

6.6 Results Achieved by using ECDS

A study was conducted during 2017–2018 for an organisation using ECDS.

The electronic coupons were developed for organisations wanting to enhance their marketing and sale. The problem was identified as the usages of conventional discount coupons were not used and organisation was spending money for printing, inserting and distributing conventional physical coupons. The organisation was looking to enhance distribution and utilisation of discount coupons to attract more customers and improve sale. The business process identified was distribution of electronic coupons which was benefitted using the Big Data solution. The building blocks of Big Data solution addressed application, data, infrastructure, and technological issues related to the conventional physical distribution of discount coupons and integrated Big Data with legacy systems. The legacy system adheres to our definition of legacy systems as defined in chapter 2, and has dated languages, databases and manual process for the distribution of discount coupons.

The electronic coupons for the organisation was developed. The downloadable electronic coupons were accessible to customers from the organisation's website and were made accessible from customer's mobile devices to conduct the study. Customers looking for discounted coupons went to the organisation's website to download the coupon to be used for the product purchase. An automatic notification was sent to the registered customer. The usage data of electronic coupons were fed into CEP for processing and a pattern was generated using member information and data information. Data collected through CEP on the usage of coupons were collected and analysed to understand the usage of electronic coupons vs physical conventional coupons. Sale record and member information of the organisational data revealed the past usage of the conventional coupons within the organisation. Member information data was linked to member sale data. The organisation using ECDS recorded their sale and customer satisfaction data and analysed the results based on attributes such as use of coupons, receiving discount coupon on time, customer satisfaction, and improvement of the ability to attract customers in shops/customer loyalty.

The results achieved by using ECDS over conventional coupon distribution system are summarised in Table 6.2.

Comparison Features	Results Achieved Using ECDS	Conventional Coupon Distribution system
Use of coupons	46% of coupons were used	5% of coupons were used.
Receiving discount coupon on time	Yes, 89% of customers received online coupon on time.	No, only 15% received coupons on time.
Customer satisfaction	14% increase in customer satisfaction from 50% to 64%.	2% increase in customers rated as satisfied. 50% to 52%.
Customer loyalty	Retention rate increased to 66% which is 10% higher.	56% customers are loyal to their stores.

Table 6.2: Result Summary of ECDS over Conventional Coupon Distribution System

The visualisation of the results is shown in Figure 6.7, which displays a positive response on the need of ECDS for organisations to improve customer satisfaction and to improve the ability to attract customers into shops. To send coupons at the right time to the right customers at the right location requires decisions to be made based on the historical data and the locational data of the customers. This was achieved using ECDS where Hadoop was used for batch processing and Spark is used for real-time processing.

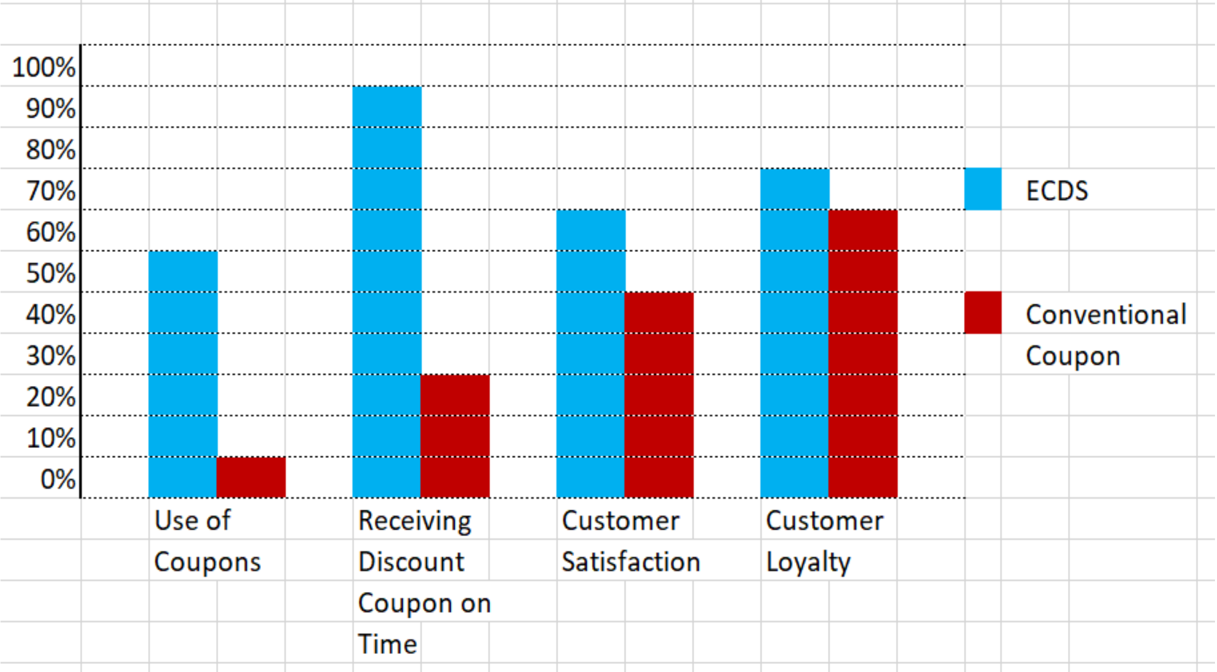


Figure 6.7: Visualisation of the Results

6.7 Chapter Summary

E-coupons are one of the ways to attract customers and increase customer satisfaction. Conventional ways of distributing coupons are using newspapers, magazines, and emails. The conventional process to distribute coupons targets all kinds of customers and does not differentiate between who will buy and who will not. Sending customers blanket promotions and coupons results in only 5% of the coupons being used and 95% of the coupons being wasted and not used.

In this chapter we have discussed the Big Data architecture developed for CEP using open-source technologies such as Hadoop, Apache Kafka, Spark and Scala as a language, and documented the results showing how CEP is applied to the use case of an ECDS, using location information, past shopping/travel history, gender, and likes/dislikes. We have explored different types of data such as static information on gender, and age, previous history (where the person travelled to, and what they bought), as well as real time information about a customer (current location, and current shopping habits) to send coupons to customers. The architecture of ECDS allows customers going online to get connected to Web servers and generates Web Logs. Apache Kafka is a tool that is built to handle ingesting transaction logs and other real-time data feeds. HDFS is used for batch and real-time processing. The batch processing is done using Hadoop and real-time processing is done using Spark with the results being sent to the analytics platform.

With the use of ECDS, it shows that 46% of the distributed coupons are used. Coupons and discounts do help sales, and if used and targeted well, can be effective in driving business forward. Poor coupon redemption is a poor KPI and Big Data can help in improving the redemption of discount coupons. Big Data often helps in discovering the history of the customer, the shopping habits, the product they looked for, and the product they would like to buy.

CHAPTER 7: USE Of BIG DATA ARCHITECTURE IN HIGHER EDUCATION INSTITUTION - 1

7.1 Introduction

This chapter discusses Big Data analytics (BDA) in the higher education sector. This was achieved by the integration of legacy systems and data with Big Data (BD) solutions. In the higher education sector, Big Data analytics is used relatively less than in other sectors. Also, higher education institutions rely on their legacy systems, such as Learning Management System (LMS), to make decisions. An LMS is a software application for the administration, documentation, tracking, reporting and delivery of educational courses, training programs, or learning and development programs. Learning Management Systems make up the largest segment of the higher education sector.

Big Data analytics in this sector needs to be combined with business processes within higher education institutions to improve institutional operations and support institutions in offering innovative services to students. The retention rate of students can be improved if an early alert system based on Big Data analysis is set up and intervention is appropriately deployed. Early alert systems extract data from LMS and other external sources such as forums and discussion boards. In this chapter, the functional capabilities of Big Data analytics in higher education and a step towards Big Data architecture to implement data analytics to benefit the higher education institutions and their stakeholders is also discussed. This chapter reports on an experimental study with 309 postgraduate students to explore how Big Data Architecture can be used for higher education analytics by integration of legacy systems and data with BD solutions.

7.2 Background

There has been growing interest in the higher education sector to take advantage of Big Data (Daniel, 2014; Liebowits, 2013; and Chen et al., 2012) to improve the learning performance of students, enhance working effectiveness of academic faculty and reduce administrative workload (Smolan, 2012). According to King (2014) and Crosling (2009), the overarching issue in institutions of higher education across the world is academic success and the retention of students. Liang et al. (2016) in their work have discussed that high-risk students are more likely to fail multiple classes in their final exam, leading to repeat or even drop-out of students. Administrators, faculty, and students are the main stakeholders of ensuring student success. Early in each semester, the discovery of high-risk students can remind counsellors, teachers, and instructors to intervene and help students in a timely manner, reducing the risk of student drop-out (Xiaogao et al., 2017). According to Godstein (2005), producing meaningful, accessible, and timely management information has long been the holy grail of higher education administrative technology. Data-driven decision making, popularised in the 1980s and 1990s, is evolving into a vastly more sophisticated concept known as Big Data that relies on software approaches generally referred to as analytics (Picciano, 2012). According to Mamčenko et al. (2006), student success rate and retention are pressing matters in the context of widening participation for under-represented student groups, increasing student diversity and educational quality assurance and accountability processes. The RP Group (2014), has discussed in the report that higher education institutions collect more and more data about their students, and as students' record databases have grown more complex, institutions are entering a new era of using data to improve student success, streamline processes, and more effectively utilise resources. While Big Data and analytics are not solutions for addressing all the issues and decisions faced by higher education institutions, they can become part of the solutions integrated into administrative and instructional functions (Picciano, 2012).

Higher education institutions rely heavily on student data for making critical and strategic decisions (Jha, et al. 2016). Higher education institutions are developing electronic learning modules, books, andquizzes, to enhance understanding of concepts amongst students. They also provide assessmentof students in systematic, real-time ways. Higher educational institutions are generating huge volumes of data, from grades or test scores to admissions or enrolment numbers while doing online evaluations and admissions, respectively. Higher education institutions have been collecting and tracking more student data than ever before, from student admission to student departure, and even after departure, such as application data, course registration data, attendance data, online learning data, performance data, extra-curricular data, internship and employability data.

Numerous e-learning platforms are available today and most popular platforms are the commercial systems such as Blackboard, Clix, and Desire2Learn, and the open-source platforms ILIAS, Moodle, OLAT, and Sakai (Cuomo et al., 2016). The strategic use and applications of Big Data analytics in higher education would lead to higher educational quality and better student and staff experience (Picciano, 2012). However, Big Data analytics are mostly employed to satisfy credentialing or reporting requirements rather than to address strategic issues, and much of the data collected are not used at all (Bichsel, 2012; Ong, 2015). Today the horizon of education is expanding electronically (Santos et al., 2014) and the old way of learning and teaching is simply no longer effective.

The prediction of which students would be high risk has always been the focus of educational science research. Xiaogao and Ruiqing (2017) have discussed an architecture for Big Data-driven high-risk student prediction which can help students identify problems as early as possible and improve students' learning methods and strategies. The authors have adopted Big Data technology and data from multiple data sources is utilised. However, their existing problems are manifested around keeping the data source relatively simple with a single dataset, the data being only from a system or a course, and the amount of data not being more than 1MB.

Tsao et al. (2017) proposed an at-risk students early alert system, and have compared different strategies and methods that deal with the participants and groups in different classes to identify the key attributes to improve the accuracy of the proposed early alert systems. The authors have shown that different data selection causes different predicting performance in at-risk student early alert systems. So, an early alert predictive model needs to consider different data volumes from different classes.

Cantebella et al. (2017) have designed and implemented an architecture based on Big Data technologies to search for behaviour patterns of LMS users. The authors have found that blended and online students are more similar with their behaviour when using an LMS than in the case of blended and on-campus students. However, they have not suggested how to improve success rate and student retention using these patterns.

According to Matsebula and Mnkandla (2016), the lack of interest in analytics and Big Data from higher education institutions has caused institutions to function with substantial delays in analysing readily available data. The authors have considered technological, organisational, and environmental factors that affect the adoption of Big Data in a South African context. They have identified that higher education institutions' data policies need to be compliant with federal, state, and local laws. The use of student data for research purposes needs to be governed by the state's data protection architecture. However, no architecture is being suggested by the authors to secure students' data.

The growth of data produced via the Internet of Things (IoT) has played a major role in defining the Big Data landscape. Immense opportunities are presented by the capability to analyse and utilise huge amounts of IoT data. The widespread popularity of IoT has made Big Data analytics challenging because of the processing and collection of data through different sensors in the IoT environment. With modern improvement of technology many higher education institutions are applying the IoT within their online learning platforms to collect, store and send the data to a central database system. The data collected is more complex and challenging due to an increase in the volume of data collected (Anthony, 2012; Selinger, 2013). Njerul et al. (2017) have demonstrated data collection from the IoT devices that can be analysed to improve decision making in higher education institutions. However, they have also identified that with the advancement of technologies, higher education institutions require taking advantageous opportunities to confront challenges, such as delivery methods, quality of contents, teachers' learning leadership, pedagogical theory, educational technology leadership, educational structures and ideology.

Kondo et al. (2017), have proposed an approach based on the log data of LMS to detect at-risk students by using machine-learning methods. According to their study the approach can detect about 45% of at-risk students at the end of the third week of first semester with only the

LMS log data. The potential of Big Data to enhance the higher education sector in Oman was the research focus of Riffai et al. (2016). Authors have concluded that the Big Data has the potential to provide student insight information that would help in addressing part of the challenges currently faced by higher education institutions in Oman. The authors have suggested that Big Data capabilities such as real-time feedback and recommendations, personalised learning and continuous improvement, sentiment and behaviour analytics, and improved student retention, can be considered as part of the solution to the challenges facing higher education in Oman. The authors have also suggested that while Big Data technologies provide opportunities to improve education, it could also pose implementation challenges such as development of technological infrastructure, capturing Big Data, changes to organisational policies and practices, competent human resources, and concerns over information privacy.

According to Marques et al. (2017), the analytics of student's behaviour in an LMS can be used as a predictor of learning success. Arnold and Pistilli (2017) have discussed an early intervention solution for collegiate faculty called Course Signals. Course Signals is designed on grades, demographic characteristics, past academic history, and students' effort as measured by interaction with Blackboard Vista, Purdue's learning management system to predict students' performance. Murmba and Micheni (2017) have done a review on Big Data analytics in higher education institutions and have identified the need for Big Data in academia. The authors have explored Big Data analytics and its relevance in higher education institutions with a view of helping educational institutions adopt Big Data analytics. The decreasing costs of big data storage, open-source software such as Apache Hadoop, NoSQL databases, network bandwidth and on-demand access to resources through cloud computing are bringing these complex technologies close to nearly everyone (Ohri, 2015; Segal et al., 2016). Underscore this by stating that while the cost of Big Data and analytic tools is coming down, they are becoming much easier to use. This is in turn opening opportunities for use of these tools by enterprises and educational institutions to achieve better outcomes and more efficient use of resources.

Many higher education institutions are in the process of using Big Data analytics; however, no Big Data architecture has been suggested which addresses all the domains of higher education for learning analytics (Jha, et al. 2016). The literature review suggests that Big Data is being applied in higher education institutions to help students in some areas. However, most of the literaturedoes not provide insights into the Big Data architecture for higher education analytics. As a result, we are demonstrating the use of Big Data architecture which can be used for higher education learning analytics. To evaluate our architecture, different data sets from a higher education institution LMS are used. We collected data sets for a postgraduate course in one semester (2018). We have collected data for a semester to find the correlation between studentsuccess rate and online activity/student behaviour.

Many higher education institutions are moving to cloud architectures and with the increased use of digital devices by users are leading to a situation where more data is being collected in these institutions than ever before, creating considerable opportunities for using Big Data to analyse and correlate information that enhances decision making. Big Data provides an opportunity for institutions to use their information technology resources strategically to improve educational quality and guide students to higher rates of completion, and to improve student persistence and outcomes (Daniel, 2014).

Once the data is analysed it promises better student placement processes; more accurate enrolment forecasts, and the results of the analysis can be fed into an early warning system to provide early and better intervention that identifies and assists students at risk of failing or dropping out. In recent years studies suggest (Daniel, 2014; Liang et al., 2016; Riffai et al., 2016) that leveraging Big Data technology and taking appropriate actions can enhance students' graduation rate and retention rate. Marsh et al. (2014) have observed that it is important for higher education institutions to use Big Data analytics to deliver the best of learning environments for the good of society.

Big Data describes data that is fundamentally too big and moves too fast, thus exceeding the processing capacity of conventional database systems (Manyika et al., 2012). A current tendency in higher education institutions consists of the analysis and processing of data related to the activities generated by the users using LMS. Higher education institutions have data contained in their databases, including grades, demographics, interaction with LMSs, interaction among peers through blogs, budget, and financial status. The unstructured and complex data extracted from these platforms provides fundamental information that can help both academics and students to improve their educational goals. One of the main problems at the present time is the analysis of this information, due to the different formats of these data, especially the management of unstructured data.

Lane et al. (2013) in their work have shown that higher education institutions still lag in the adoption of Big Data analytics for measuring student success rate and identifying students at risk. In parallel, advancements in information and communication technology are continually reshaping the teaching, learning and campus experience, resulting in many elements of the student journey being digitised, meaning that as students engage in the digital environment, institutions are deluged by data and such data is not put to maximum use. Big Data analytics is relevant in addressing a significant number of pressing issues for education systems (Marsh et al., 2014), key among them are increasing educator effectiveness, harnessing insights from learning experiences, delivering education for all that may also be tailored for individual learners needs, and equipping students with relevant skills for their future.

This chapter outlines an architecture for Big Data for integrating legacy systems and data with BD solutions that can be used in higher education institutions and details of a study in applying Big Data analytics which was guided by the following questions:

1. *How can Big Data Architecture for integrating legacy systems and data with BD solutions help in higher education analytics incorporating LMS datasets?*
2. *How, by leveraging unstructured behavioural data/sentiment data, can we predict the probability that any given student will pass or fail?*

7.3 Higher Education Business Domains for Big Data Analytics

Like other organisations, higher education institutions need to identify business processes, which are using unstructured or semi-structured data that need to be analysed for decision making such as email, PowerPoints, pdf, XML, voice, video, and image.

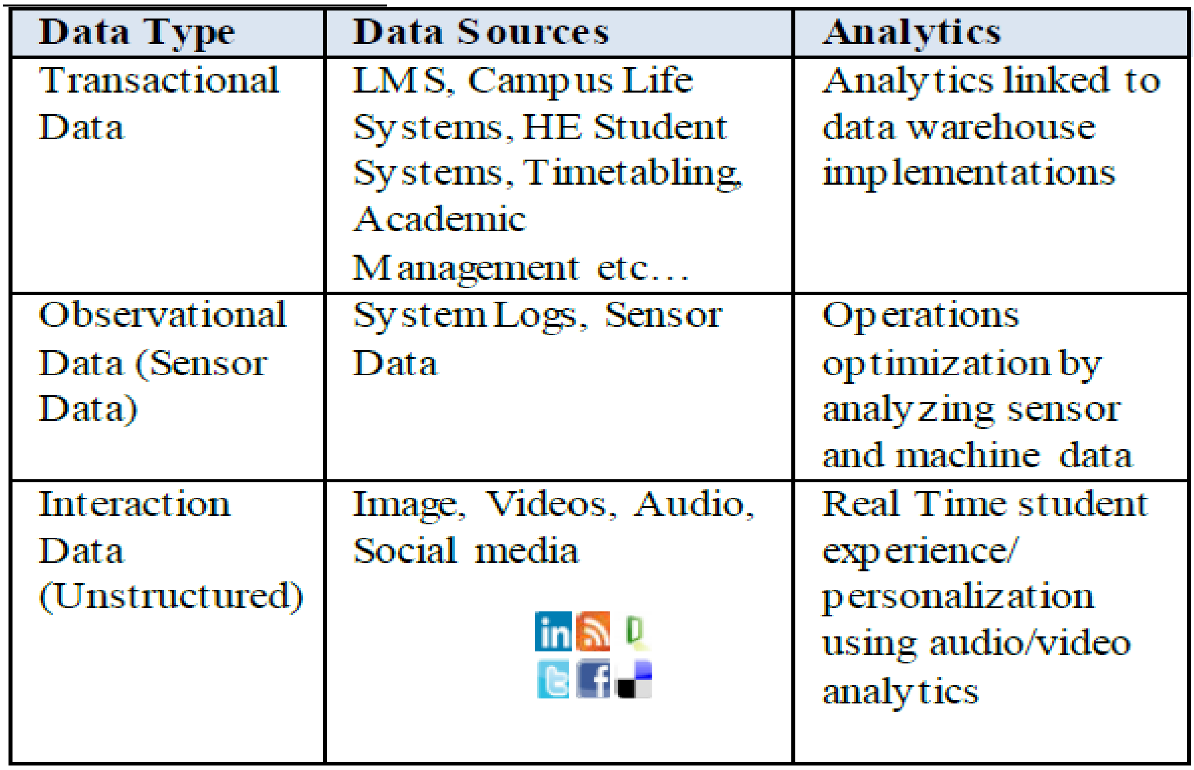
The major higher education domains identified are: websites and Portals, Academic Management, Timetabling, Business Intelligence (BI), Document and Process Management, Library and Research, Finance Systems, Facilities Management, Higher Education Student Systems, Campus Life Systems, and Learning Management Systems (LMSs). These identified domains contribute to the functional components and areas of higher education institutions. For our study we have selected Business Intelligence systems and LMS.

To answer the earlier questions, we identified higher education business domains to extract data sets. The data sets were comprised of: Student Id; Assignment marks; Exam marks; Total grade; Weekly Moodle activity; Units passed; Units attempted; Unit load, and Grade Point Average (GPA). These data are structured data sets. Students behavioural data/sentiment data were collected from "Have your Say" survey[1]. These data are unstructured data sets. For our study we have collected and processed unstructured data to enforce the need to analyse unstructured data sets within the context of higher education institutions. In the context of higher education, decision time is critical as student data, such as data from online learning activities, is generated in real-time. Reporting tends to be required on a weekly and/or monthly basis, such as a daily engagement report of online activity, weekly student admission report, monthly report on student external engagement and employability data (Daniel, 2014).

7.4 Big Data Architecture for Educational Institutions as a Decision Support

The types and sources of data available for analytics in higher education institutions have changed in recent years and are shown in Table 7.1. The services provided by HE Student Systems are management, financials, reporting, and integration. Data recorded in HE Student Systems are student biographics, student demographics, and student progression. The services provided by the Academic Management are to review and modify existing courses and program information. The data stored in the Academic Management systems are course and program information. Website and Portal Systems provide services of presentation and access to information and applications and data sources are consolidated higher education data. The growth of data produced via the Internet of Things (IoT) has played a major role in defining the Big Data landscape. Immense opportunities are presented by the capability to analyse and utilise huge amounts of IoT data. The widespread popularity of IoT has made Big Data analytics challenging because of the processing and collection of data through different sensors in the IoT environment.

[1] The responses to unit evaluations have been kept anonymous. The data collected and used has been approved by the Central Queensland University's Human Research Ethics Committee involving students in a unit. All participants enrolled in the unit were asked for their consent.



Data Type	Data Sources	Analytics
Transactional Data	LMS, Campus Life Systems, HE Student Systems, Timetabling, Academic Management etc…	Analytics linked to data warehouse implementations
Observational Data (Sensor Data)	System Logs, Sensor Data	Operations optimization by analyzing sensor and machine data
Interaction Data (Unstructured)	Image, Videos, Audio, Social media	Real Time student experience/ personalization using audio/video analytics

Table 7.1: Data Availability for Learning Analytics in Higher Education Institutions

The layers of the proposed Big Data architecture, as shown in Figure 7.2, have been selected to address different formats of data types, provision for data integration, and the tools employed for the processing of data.

Provision Data Layer contains all information required for processing. It includes everything from higher education institutions such as: institutional data; records of a particular student; Web Logs in chronological order and maintained through specific software; images; videos; documents; PDFs (Portable Document Format); and social media data. Apache SQOOP is used in this layer for transferring data between relational databases and Hadoop. This is one of the connectors between legacy and BD Architecture we used where the connection to the integration of legacy systems and data occurs with BD solutions.

Store Data Layer contains all collected data in a single location. This is where our Big Data lives, once it is gathered from all the sources. This layer serves as a system for storing data as well as a system for organising and categorising it in a way that is understandable—the database. At this layer Hadoop, HBase and its files system, HDFS is used.

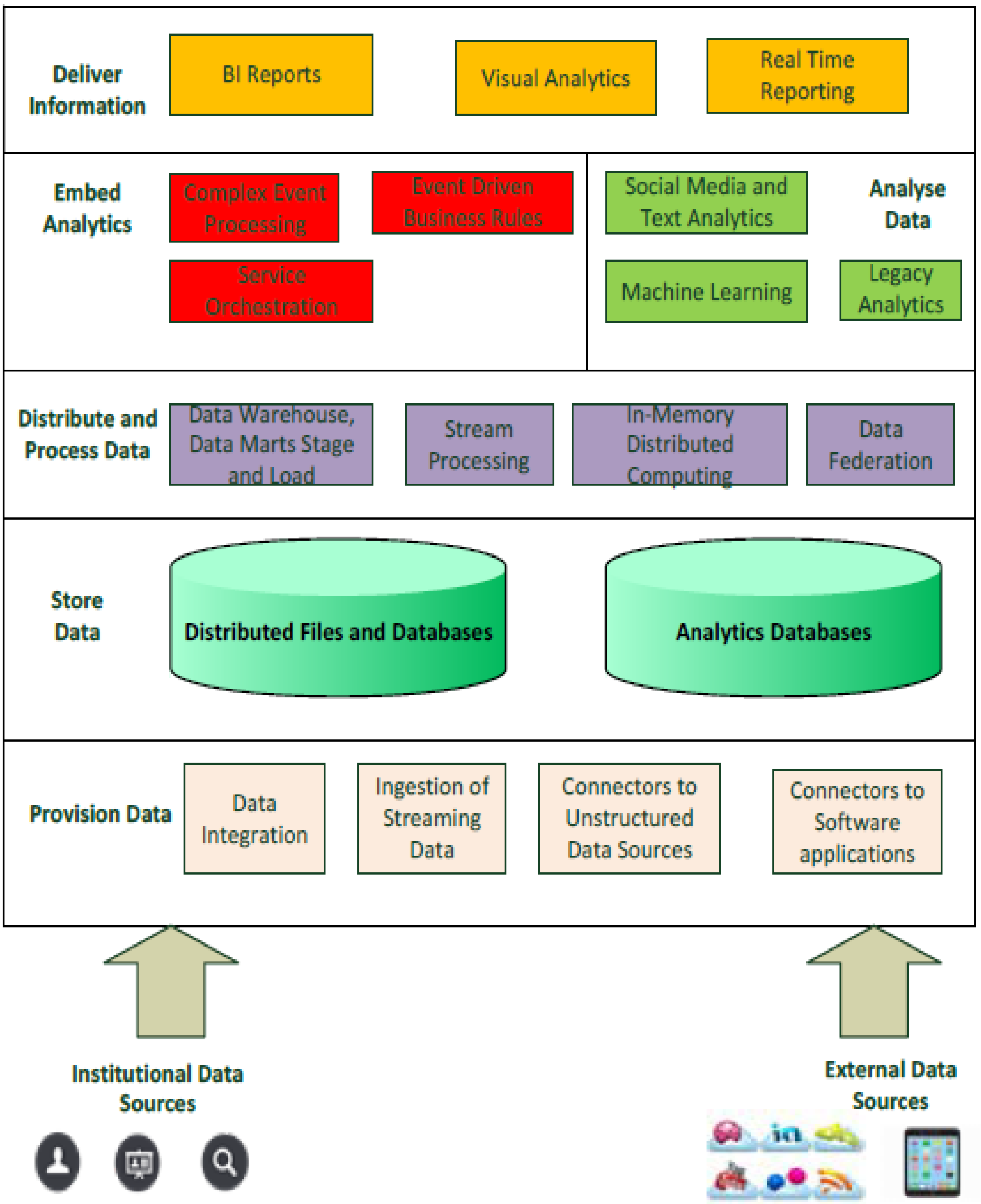

Deliver Information
BI Reports
Visual Analytics
Real Time Reporting
Embed Analytics
Complex Event Processing
Event Driven Business Rules
Service Orchestration
Social Media and Text Analytics
Analyse Data
Machine Learning
Legacy Analytics
Distribute and Process Data
Data Warehouse, Data Marts Stage and Load
Stream Processing
In-Memory Distributed Computing
Data Federation
Store Data
Distributed Files and Databases
Analytics Databases
Provision Data
Data Integration
Ingestion of Streaming Data
Connectors to Unstructured Data Sources
Connectors to Software applications
Institutional Data Sources
External Data Sources

Figure 7.1: Big Data Architecture for Higher Education Analytics

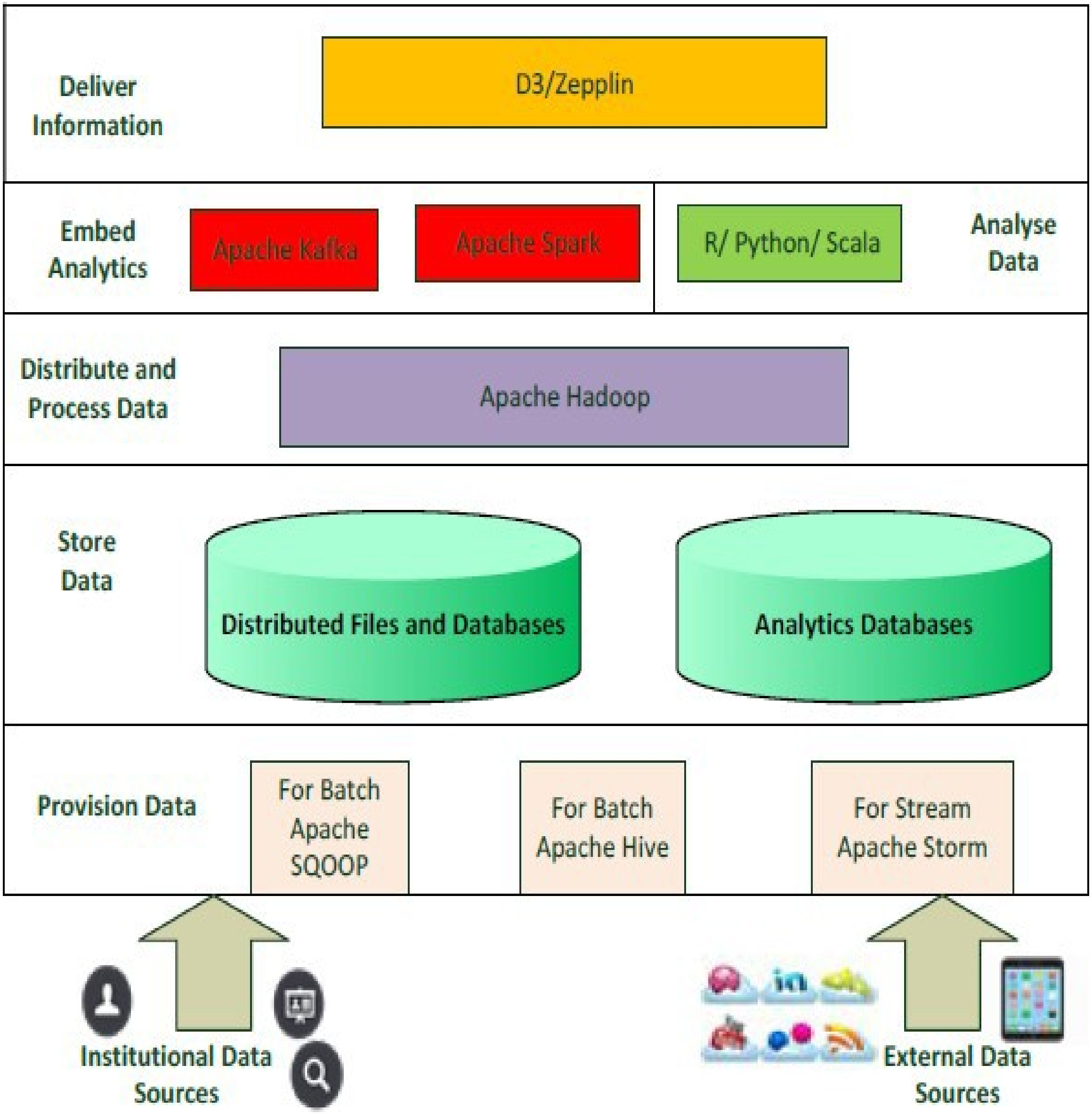


Figure 7.2: Tools Used in Big Data Architecture

Distribute and Process Data Layer uses Spark. All the data is transferred in the form of Resilient Distributed Data (RDD) when Spark is used. The RDDs were created byreferencing data stored in HBase which is used as external storage.

In the Embed Analytics Layer learning algorithms are used. We have used Python for Logistic Regression to predict the probabilities of pass and fail.

Apache Zepplin is used in the Deliver Information layer. Apache Zepplin is integrated with distributed, general-purpose data processing systems such as Apache Spark which is large-scale data processing.

We processed and analysed 309 postgraduate students' data collected from Moodle and different modules of Moodle, namely Early Connect Student Alert Indicators Systems (EASI Connect) and "Have your Say" unit evaluation. The purpose of this study is to show that using the proposed Big Data Architecture is more efficient than a traditional architecture, especially if the data is varied and has unique characteristics.

To conduct our study, we have used two architectures, namely Traditional architecture and Big Data architecture. To show that Big Data architecture is more efficient than traditional architecture we have processed structured and unstructured data sets. Traditional architecture only processes structured data such as Assessment marks, Final exam marks, Student Id, Unit attempted, Unit passed, Unit load, GPA, Tutorial, Absence rate, Total grade. For traditional architecture, the dependent target variable was total final marks based on the input variables of assessment marks, tutorial completed, and absence rate. To facilitate analysis, the data were transformed and prepared to be fed to HBase in preparation for processing by Apache Spark to predict the probability of students passing and failing the course based on the input variable. These input variables can be changed. We used supervised regression technique to develop a predictive model based on both input and output data.

Unstructured data was collected from a survey where students can write about their opinion of the unit. The student's opinions were collected using "Have your Say" evaluation about the unit. Only 173 students completed the survey. The students' behavioural engagement data was captured via their interaction with the online Moodle activity consisting of how many times they visited, how long they were on the site, the number of navigation steps and navigation patterns. Students are unlikely to be involved in academic misconduct if they are interacting and engaging themselves with online Moodle activity.

We have used the above Big Data architecture (figure 7.2) for higher education analytics for analysing LMS Moodle data and Early Connect Student Alert Indicators Systems (EASI Connect) which collects sets of data from students, teachers, instructors, and other stakeholders in near real time. EASI relies on data from upstream systems holding students' related data and student activities data. Thus, in this experiment we used multiple data types to examine students' performance. Data is refreshed every day based on students' online activities. A course activity report, showing the number of views for each activity and resource (and any related blog entries), can be viewed by teachers using Moodle Activity Viewer (MAV). It does this using a heatmap, colouring links lighter or darker according to how often records are accessed. Moodle data when analysed can serve as a tool for teachers to follow students' behaviour to identify critical situations. Typical data relate to the competency breakdown, logs, live logs, activity reports, activity completion, number of accesses, duration of accesses, paths traversed in the platform, tools used, resources used or downloaded, participations in the forum, and other activities (Murrumba et al., 2017).

For Big Data Architecture we have processed and analysed data generated by the MAV for students' online activity to courses they have passed to see the correlation between students' engagement to students' success rate. We read the text data collected from "Have your Say" evaluation from HBase and stored it in Resilient Distributed Dataset (RDD). At the Deliver Information layer, we have used Apache Zeppelin Web-based notebook which brings data ingestion, data exploration, visualisation, sharing, and collaboration features to Hadoop and Spark. For the Embed analytics and analyse data layers we have used Apache Spark. Apache Spark has as its architectural foundation RDD and data items distributed over a cluster of machines. Apache Spark supports and can interface with Hadoop which we have used at layer Store distribute and process data. For layer Embed Analytics and Analyse Data we have used Python for Logistic Regression to predict the probabilities of pass and fail.

7.5 Results from the Study

To understand the relationship between Moodle activities and the grades obtained by the students' we have used statistical method correlation analyses. Correlation analysis is used to evaluate the strength of relationship between two quantitative variables. This method is used for our analysis for the identified variables: students online Moodle activities, and the grades obtained. The correlation method allowed us to understand the extent that a change in one variable online Moodle activity creates some change in the other variable grades obtained. Extracting data from Moodle has reflected upon online activities students are involved in during the term. The grades obtained by students can be discussed and explored after analysingthe correlation between their online activities and grades obtained.

The quantitative data analysed in this study were retrieved and collected from the Learning Management System, Moodle. The course coordinator of the unit provided the data for selected unit of study. Specific information required for this study included student online activity, access to documents, assignment submission details, forum participation, academic misconduct, absence rate, behavioural engagement data, and students' final grades of officially enrolled students in the unit. The correlation design for the research was selected to enable the study to determine if a "relationship between variables" exists (Gay & Airasian, 2003). Using Microsoft Excel spreadsheet and Statistical Package for the Social Sciences SSPS software, the data of student online activity, access to documents, assignment submission details, forum participation, academic misconduct, absence rate, behavioural engagement data, and students' final grades was tabulated for each student. Using the statistical and data analysis tools available through Microsoft Excel and the SPSS correlation analyses were conducted between the variables of interest. We used Pearson's r to represent the strength and direction of the relationship between two variables. The numeric value of the Pearson's r indicates the strength of the linear relation between two variables. It can range from –1 to 1 and the closer the value is to the absolute value of 1, the stronger the linear relation between two variables (Odom & Morrow, 2006). Prior research (Basaran, 2013) in education supports this correlation testing method.

Table 7.2 shows several correlation scores calculated using access and participation data from Moodle activities and the grades obtained by the students. Students' online participation data in the unit and passing unit had a positive correlation. More online participation reflects higher grades in the unit.

We can observe that there are many correlations to be found regarding students' success. The traditional architecture has shown correlation between accesses to documents, assignment submission, participating in forum, academic misconduct, and absence rate. The traditional architecture was improved by adding the behavioural engagement score using Big Data architecture. The Big Data architecture has shown correlation between the above-mentioned variables and behavioural engagement/sentiment data. Behavioural engagement data wasfound to be more efficient in predicting success rate. Our study suggests that the more the behavioural engagement then the lower the failure rate. This indicates that there are genuine indicators within textual student feedback that can be extracted and used as important predictors of student success rate within the term.

Variables	Type of Data	Correlation
Access to documents/Final grades	Structured	Correlation final = 0.5593
Assignment submission	Structured	Correlation final = 0.25896
Participating in forums	Structured/Semi-structured	Correlation final = 0.12589
Academic misconduct	Structured/Semi-structured	Correlation final = 0.1152
Absence rate	Structured	Correlation final = 0.1238
Behavioural engagement data/Sentiment data	Text data unstructured	Correlation final = 0.0135

Table 7.2: Correlation Scores Calculated Using Access and Participation data

Our Big Data architecture has given a picture of online activity driving the student success rate incorporating LMS datasets. Online activity data can be used to keep students on track all the way to graduation and help students who have struggled to stay in the course as well as higher education institutions struggling to understand how to lower dropout rates and keep students on track during their study program. The results of this experiment provide evidence of the importance of unstructured data.

As in this experiment, several other studies utilised Apache Spark as a platform for processing unstructured data. All these studies (Assiri et al., 2016; Solaimani et al., 2016) acknowledge the high and rapid performance of Spark, which can be used to manage massive amounts of data and provide real-time analytical power.

7.6 Chapter Summary

In this chapter we have implemented and demonstrated the usefulness of Big Data architecture integrating legacy systems and data with Big Data solutions. The investigation for Higher Education analytics that has been used to gain some critical insight from the data collected through the Learning Management System (LMS) and Early Alert Student Indicator (EASI). Our research in Big Data analytics for higher education provides the core frame and tools of Big Data applications that are needed by Big Data analytics and enforces the need to analyse unstructured student behavioural data. All these data-driven initiatives aim to increase student learning, enhance the students' experience, and build an environment that encourages retention at higher education institutions.

This proposed and implemented Big Data architecture integrating legacy systems and data with BD solutions for higher education institutions for data collection, storage and analysis offers potential benefits for future development activities in the learning analytics field.

CHAPTER 8: USE OF BIG DATA ARCHITECTURE IN HIGHER EDUCATION INSTITUTION - 2

8.1 Introduction

In this chapter we discuss how Big Data analytics can be combined with higher education business processes using re-engineering for structured data, unstructured data, and external data for integrating legacy systems and data with BD solutions. To achieve this objective, we investigate the core business processes of learning and teaching and define a re-engineered higher education business process model.

From a review of the literature in the Business Information System (BIS) research, most of the research is concerned with the use of BIS in industry. The education sector is rarely addressed in the research (Chiasson & Davidson, 2005). We believe there is an opportunity for the BIS community to address the changing requirements of educational business processes. One of the key business applications used in the education sector is Learning Management Systems (LMS).

8.2 Background

Data within LMSs and other sources within the University should be utilised through the introduction of business analytics to provide additional information to business processes within the academic institutions. The data within these systems individually would not be considered Big Data; however, combining the structured and unstructured data from all of the various systems, as well as external data, turns it into a Big Data problem that needs to be addressed by a Big Data solution.

This business process model can shape the idea of correlating students' requirements regarding the developed course structure and delivery used by higher education institutions. This business process model addresses the importance of understanding students' requirements and their needs from course content and course delivery.

A large shift towards Big Data from traditional approaches has been observed to handle different business processes and to develop better predictive models for the organisation.

Business intelligence and analytics is helping many companies to improve their efficiency in customer satisfaction. Business analytics is becoming standard to communicate data-driven business decision making. Big Data analytics have been one of the major reasons for drastically changing the products and services provided by companies in recent years (Baesens et al., 2016). There is an opportunity to apply the same within the education sector.

Analytics goes beyond business intelligence in that it is not simply more advanced reporting or visualisation of existing data to gain better insights. Instead, analytics encompasses the notion of going behind the surface of the data to link a set of explanatory variables to a business response and outcome.

The widespread introduction of LMS (Paule-Ruis et al., 2015), such as Moodle and Blackboard, resulted in increasingly large datasets. Each day, LMS accumulate increasing amount of students' interaction data, personal data, systems information, and academic information (Romero et al., 2008). However, LMS does not collect data from interaction among peers through blogs, social media, external review sites, students' feedback surveys, tweets directed at higher education or on related hashtags, reviews on Facebook, to name a few external data sources. Higher education should not overlook what students are saying about them on social media. Students posting comments on a Facebook wall is also a review feature where students can leave a detailed comment about higher education institutions and its offered courses. The current LMS does not support external and unstructured data sources.

To combine Big Data analytics with higher education business processes, we need to re-engineer existing higher education business processes by identifying input and output from the core process of learning and teaching and identifying how unstructured and external data sources can be helpful, so that Big Data analytics can be applied and work effectively for decision making (Capgemini, 2012). The obvious reason behind it is higher education business processes are not capable of handling different varieties of data. By definition, variety is one of the characteristics of Big Data. The technical and managerial issues resulting from the adoption and application of Big Data analytics is worth exploring (Riffai et al., 2016). Analysing and understanding the online behaviour of students results in predictive capability that is important for intervention and is a cornerstone of effective personalised learning (Daniel, 2014).

The biggest benefits of Big Data analytics in higher education institutions are: the ability to analyse, track and predict student performance; improve graduation and retention rate; and adjust teaching strategies for just-in-time intervention (Attaran et al., 2018). Traditionally, it has been difficult for higher education institutions to provide critical and timely intervention simply because they did not know a student was struggling until it was too late. Organisational silos and a dearth of data specialists are the main obstacles to putting Big Data to work effectively for decision making (Reid-Martines et al., 2015). Most of the legacy systems were developed without process models or data models which are now required to support data standardisation. Higher education institutions need to improve processes and technological advances that can be integrated in the development of efficient processes through business process re-engineering and business process innovation (Anand et al., 2017). A successful integration of Big Data analytics and business processes will create a "new class of higher education economic asset" and help higher education institutions redefine their business and outperform theircompetitors.

8.3 Selecting Core Business Processes of Learning and Teaching

Re-engineering is a systematic process of analysis, design, and implementation. Hammer and Champy (1993, p.32) have defined Business Process Re-engineering as "*the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service and speed*". Hammer and Champy (1993, p.35) also define a process as "*a collection of activities that takes one or more kinds of input and creates an output that is of value to the customer*". In higher education the output from each process of learning and teaching should be used for Big Data analytics so that the output generated by each process is of value to students, educators, and other stakeholders. Higher education institutions are finding it challenging to achieve their objectives and goals due to complicating factors arising from making decisions at the right time. Khalid et al. (2011) have coined the term "*Educational Process Re-engineering*" based on the established concept of Business Process Re-engineering for process improvement of learning and teaching activities, academic administration and evaluation and assessment. However, their work lacks combining Big Data analytics within the business processes. The dramatic transformation of society requires that the educational system adapt and change, or face obsolescence. The stimulus for re-engineering higher education business processes is the combination of four ongoing transitions in the society (Jha et al. 2016). They are enhancing learning and implementation of learning activities; changing educational needs; requirements for alternative learning opportunities; and extensive availability of digital technology.

A re-engineered educational activity is a significant enhancement to the traditional decision-making process. Re-engineering higher education learning and teaching processes to combine Big Data analytics will exploit significant amounts of student-level data (Jha, et al. 2019).

Big Data analytics should be used to guide the re-engineering educational process to address the following (Wagner & Ice, 2012):

- Educational needs are to be met by a proposed educational activity. This can range from the institutional to the individual lesson/module level.
- Redesign a curriculum to define the scope, general content, and structure of the proposed activity (program or area of study).
- For each area of study define the elements or units, such as class topics, modules, and exercises, that are appropriate for concentrated study.
- Develop well-defined learning objectives and expected outcomes. These are derived from the documented or perceived needs for the activity.
- Develop an analysis of the proposed learners (students) with respect to educational background, location (geographical distribution), and times available for learning activities.
- Apply established principles of learning to define the materials, media, and methods to design each learning unit.
- Develop or acquire the learning materials and media as required.
- Identify and make available additional references and resources to support the learning activities.
- Develop or acquire access to facilities with sufficient technology infrastructure to support the learning activities.

- Develop an appropriate management and administration system for the educational activities.
- Develop a faculty and collaborators with "real world" experience and the ability to guide and stimulate the learning activities.
- Provide faculty with learning and development opportunities to enhance their effectiveness as learning facilitators.
- Provide faculty with incentives to develop and use technology to enhance their teaching and academic activities.

Core Processes of Learning and Teaching	**Prepare Learning and Teaching Resources**	**Implement Learning and Teaching Resources**	**Outcomes of Learning and Teaching Resources**	**Review of Learning and Teaching Resources**
Input artefacts	Teaching strategies; Learning materials	Develop course contents Coordinate courses	Graduates	Review course Review subjects Review industry feedback
Output artefacts	Attract students Select students	Graduation Retention rate Success rate	Scholarship of learning and teaching	Teachers and student prospective

Table 8.1: Core Business Processes of Learning and Teaching

Table 8.1 shows the identified core business processes of learning and teaching with their input and output. The core processes of learning and teaching to address the above points are identified as: prepare learning and teaching resources; implement learning and teaching resources; outcomes of learning and teaching resources; and review of learning and teaching resources. These business processes are required to be re-engineered to fit data-driven outcomes in the higher education environment. Each business process has inputs and outputs. Outputs from these business processes will help higher education to make near real-time evidence-based decisions if processes are combined with Big Data analytics. Inputs to these business processes should not be only data from students' success rates and graduation. Nowadays a large amount of log data is generated in higher education institutions' learning management systems. The benefits of analysing these log data are to be found when processing them *en masse*.

8.4 Re-engineering Learning and Teaching Business Processes

We have re-engineered and generated re-engineered a higher education business process model combining with Big Data architecture using integration of legacy systems and data with BD solutions (Jha, et al. 2019). Some of the enhanced characteristics of combining Big Data architecture with business processes of higher education are:

- Attracting students to a program or a course (which is an outward and visible sign of the popularity of a program or a course)
- Observing student performance (which is an outward and visible sign of student learning)
- Analysing the factors that improve student achievements (such as teaching practices and student motivation)
- Analysing the factors that affect success and retention rate (such as gender and socioeconomic status)
- Reviewing of educational policies (such as curriculum revision and testing regime).

The key components of a re-engineered higher education business process model to combine Big Data analytics in higher education are illustrated in Figure 8.1. Data (external and internal) is acquired and organised as appropriate and then analysed to make meaningful decisions. LMSs collect data from students that, properly analysed, can provide educators and students with the necessary information to support and constantly improve the learning process (Evans, & Linder, 2012; Garcia et al., 2009). Unfortunately, these platforms do not provide specific tools to allow educators to thoroughly track and assess all students' learning processes. To test our e-business process model, we applied it to log file data from a Moodle course offered in mixed mode (face-to-face and distance education).

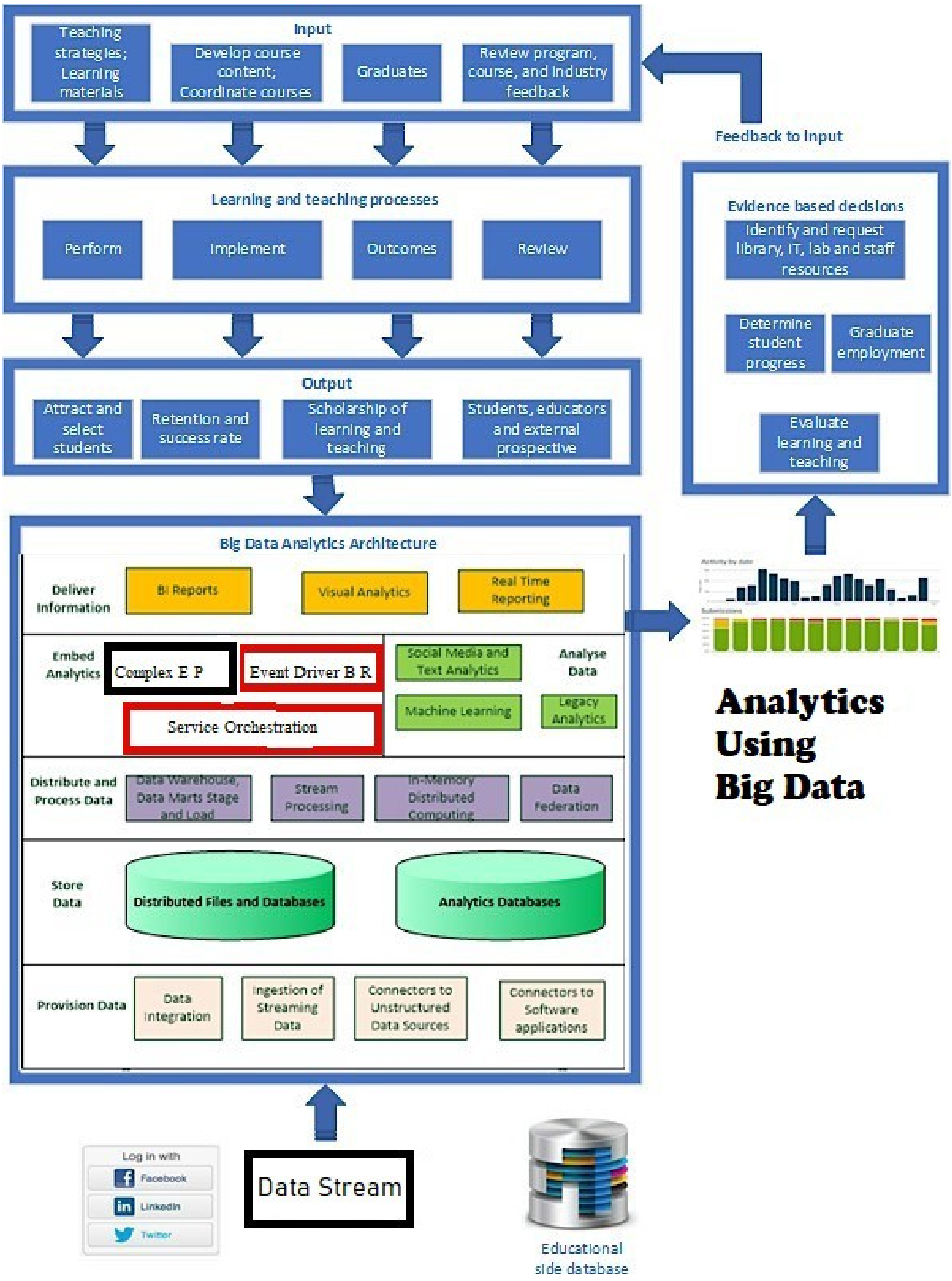

Input
Teaching strategies; Learning materials
Develop course content; Coordinate courses
Graduates
Review program, course, and industry feedback
Feedback to Input
Learning and teaching processes
Perform
Implement
Outcomes
Review
Output
Attract and select students
Retention and success rate
Scholarship of learning and teaching
Students, educators and external prospective
Evidence based decisions
Identify and request library, IT, lab and staff resources
Determine student progress
Graduate employment
Evaluate learning and teaching
Big Data Analytics Architecture
Deliver Information
BI Reports
Visual Analytics
Real Time Reporting
Embed Analytics
Complex E P
Event Driver B R
Service Orchestration
Social Media and Text Analytics
Machine Learning
Analyse Data
Legacy Analytics
Distribute and Process Data
Data Warehouse, Data Marts Stage and Load
Stream Processing
In-Memory Distributed Computing
Data Federation
Store Data
Distributed Files and Databases
Analytics Databases
Provision Data
Data Integration
Ingestion of Streaming Data
Connectors to Unstructured Data Sources
Connectors to Software applications
Analytics Using Big Data
Log in with
Facebook
LinkedIn
Twitter
Data Stream
Educational side database

Figure 8.1: Re-engineered Higher Education Business Process Model

We have used the re-engineered higher education business process model for analysing LMS Moodle data, which collects a set of data from students, educators, and other stakeholders (Jha, et al. 2019). Wehave also analysed Facebook reviews and University Google review data to see the satisfaction rate of course and content delivery. The LMS course activity report, showing the number of views for each activity and resource (and any related blog entries), can be viewed by teachers using Moodle Activity Viewer (MAV).

The LMS course activity report is shown using a heatmap—colouring links lighter or darker according to how often records are accessed. Moodle data when analysed can serve as a tool for teachers to follow their students' behaviour to identify critical situations. Typical data relate to the competency breakdown, logs, live logs, activity reports, activity completion, number of accesses, duration of accesses, paths traversed in the platform, tools used, resources used or downloaded, participations in the forum and other activities (Arnold & Pistille, 2017; Rice, 2006). Data was collected for two semesters (six months) for three courses. The population consisted of 620 participants. We have analysed data generated by MAV for students' online activity to courses they have passed to see the correlation between these two variables that includes students' academic success and students' engagement with online learning activities.

At the Deliver Information layer, as shown in Big Data analytics architecture of re-engineered higher education business process model in Figure 8.1, we have used Apache Zeppelin Web-based notebook which brings data exploration, visualisation, sharing and collaboration features to Hadoop and Spark.

For the Embed Analytics layer and Analyse Data we have used Apache Spark. Apache Spark has, as its architectural foundation, the Resilient Distributed Dataset (RDD) and data items distributed over a cluster of machines. Apache Spark supports, and can interface with, Hadoop which we have used at layer Store Distribute and Process Data. For layer Provision Data, we have used Hive to import datastores into Hadoop.

We have uploaded data into Hadoop Distributed File System (HDFS) Files View and created Hive queries to manipulate data. We created a table called temp_students to store data. We created another table student, so we can overwrite that table with extracted data from the temp_students table we created earlier. Then we did the same for temp_online_activity and online_activity. Finally, we created queries to filter the data to have the result show the sum of online activities and courses passed by each student. Table 8.2 shows the data that were integrated into each of the two tables.

These data sources reflect students' engagement with learning activities and review on how the course is delivered and structured. The main objective in selecting these sources is to understand the correlation between students' academic success and students' engagement with learning activities. Facebook reviews and University Google reviews were used to improve on lecture delivery, course content, and resources developed by educators.

Data Sources	**Indicator Count from Legacy Systems and External Data Sources**
Online access to Assignment 1	220914
Online access to Assignment 2 marks	90722
News forum	7653
General discussion	6858
Questions and answers	2645
Facebook review	556
University Google reviews	76

Table 8.2: Contents of Legacy Data and External Data Used for Big Data Architecture

For analytics we used Apache Spark with language R. For data sets that are not too big, calculating rules with association rules (arules) in R is not a problem. Since association mining deals with students, the data had to be converted to one of class students, made available in R through the arules pkg. This is a necessary step because the apriori() function accepts students' data of class students only. We included library (arules). We used several correlation scores calculated using access and participation data from Moodle activities and the grades obtained by the students. Students' online participation data in the unit and passing unit had a positive correlation. More online participation reflects higher grades in the unit. We used association-based mining techniques which include Apriori. We selected Apriori algorithm. In R there is a package arules to calculate association rules, it makes use of the Apriori algorithm. We first connected to Apache Spark and uploaded the data to Spark. Association rule mining is a rule-based machine learning method for discovering interesting relationships between variables in large datasets. It is intended to identify strong rules discovered in data using some measures of interestingness (Anand et al., 2013). The correlation between online activities and students' grades were analysed as shown in Table 8.3.

Variables	Type of Data	Correlation
Access to summative assessment documents/Final grades	Structured	Correlation final = 0.6593
Assignment submission	Structured	Correlation final = 0.55896
Participating in forums	Structured/ Semi-structured	Correlation final = 0.22589
Academic misconduct	Structured/ Semi-structured	Correlation final = 0.1052
Absence rate	Structured	Correlation final = 0.1238
Behavioural engagement data/ Sentiment data	Text data unstructured	Correlation final = 0.1135

Table 8.3: Correlation Scores Calculated

The correlation was used to gain the understanding of students' online activity to their academic success and how much these online activities are used by students. The regression algorithm shows the correlation between access to documents/final grades; assignment submission; participating in forums; academic misconduct; absence rate and behavioural engagement data with online activities. The result shows strong correlation between online activity and courses passed; however, students are often not active on forums and not working towards tutorial exercises. We identified that the student's online activity decreased towards the end of the term. This resulted in a poorer exam outcome. Interestingly, their grades remain good. The reason behind this was to pass the course it was not necessary to pass the exam. Students are active on summative tasks. This requires the course to be redesigned to address the issues of students not using all the required learning and teaching re- sources. This gives an input to re-designing and re-developing course content. Students were posting comments on University Google reviews about the quality of the course contents and course delivery style. The decision tree algorithm is used to predict potential performance using online activities, courses attempted, and estimation of success factors. Our research findings are that students' academic success is not only affected by the extent of online activities, and experience with LMS, it is also affected by the design and structure of the formative and summative assessments. Students' online activities increase and decrease during the term according to the overall contribution of formative and summative assessment to pass the course. With our results we were able to re-design the course and delivery style which resulted in a higher student success rate and attracted more students to our course. We witnessed the in- crease of 12% success rate with the new course design and 20% more students enrolled in the course.

8.5 Chapter Summary

This chapter described our approach of using the Big Data architecture for higher education. To combine Big Data analytics with higher education business processes, we re-engineered existing higher education business processes by identifying input and output from the core process of learning and teaching and identified how unstructured and external data sources can be helpful so that Big Data analytics can be applied and work effectively for decision making for higher education. In this chapter we have reported that it was a positive first step in allowing our students and teachers to quickly gain an understanding of the integrated data and to visually extract interesting patterns. These patterns suggest the improvement on redesigning a curriculum to define the scope, general content, and structure of the proposed activity (program or area of study) based on observing student performance (which is an outward and visible sign of student learning) and analysing the factors that improve student achievements (such as teaching practices, student motivation, and nudging students when they are not engaged with online resources).

CHAPTER 9: CONCLUSION

9.1 Introduction

Many organisations own some legacy systems and data and maintain them to fulfil their daily business operations. Legacy systems and data cannot always accommodate newly emerging business needs, thus might negatively impact an organisation's decision-making capabilities.

The purpose of this thesis is to build an architecture to integrate legacy systems and data with Big Data solutions. To build the architecture, we needed to understand challenges in integrating Big Data solution(s) with legacy systems. Due to the complexity of legacy systems and the landscape of Big Data the task is complex. We firstly developed an architecture for integrating Big Data into the Enterprise. Big Data can be in many forms, such as Web data, audio, and video data. Big Data can be found internal to the organisation and/or external to the organisation. We identified building blocks for integrating Big Data with the Enterprise Architecture. Secondly, we re-engineered a higher education business process to fit into an architecture of Big Data solutions. Big Data analytics need to be combined with business processes to improve operations and offer innovative services to customers. Business processes need to be re-engineered for Big Data analytics. Re-engineering of business processes in a legacy system was done to incorporate legacy systems into Big Data solutions where Business Architecture, Information Architecture, Technology Architecture, and Data Architecture are used. To do this, these re-engineered processes were used within Big Data architecture to address the challenges in integrating Big Data solution(s) to fit into the legacy systems.

This thesis presents original work in four areas:

1. Identifying the challenges in integrating Big Data solution(s) to integrate with legacy systems and data.

2. Impact on Business Architecture, Information Architecture, Technology Architecture and Data Architecture while integrating legacy systems with Big Data solutions.
3. Addressing the challenges of integrating Big Data solutions with legacy systems and data.
4. Combining Big Data analytics with business processes using re-engineering.

This concluding chapter comprises of the following sections. Section 9.2 summarises the research questions, section 9.3 summarises the contributions of our research, while section 9.4 suggests future research directions.

9.2 Research Questions

We investigated the challenges in integrating Big Data solution(s) to integrate with legacy systems and data and addressed those challenges with developing a Big Data Architecture to improve integration of Big Data solutions with legacy systems and to build an architecture to integrate legacy systems and data with Big Data solutions. Following research questions are explored in this thesis to develop and demonstrate the usefulness of the Big Data Architecture:

Research Question 1: *What are the challenges in integrating Big Data solution(s) to integrate with legacy systems and data?*

Research Question 1 allowed us to identify the challenges, understand them and address them in the development of the Big Data Architecture and architecture for integration of legacy systems and data with Big Data solutions. In chapter 3 we have discussed a survey on Big Data and its integration with legacy systems and data. We have identified that 100% of respondents believe that implementing Big Data solutions will bring benefits to organisations in many ways, such as using external data sources for making organisational decisions. It has been identified through our survey that there is no architecture to implement Big Data solutions and integrate Big Data solutions with legacy systems in organisations. An architecture was developed based on the issues and challenges identified by this survey to integrate Big Data with legacy systems provided solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats.

The key challenges identified and addressed include: Big Data processing requires real-time, near real-time, or batch processing; using different characteristics of Big Data requires different technology infrastructure components and data architectures; the technology infrastructure components and the technology architectures, as well as data architecture changes, must be captured by an organisation's existing EA to enable conducting BI and DA, always using a holistic approach, for the successful development and execution of an organisation's strategy.

Research Question 2: *How does Big Data solution integration impact Enterprise Architecture (EA) in terms of the Business Architecture, Information Architecture, Technology Architecture and Data Architecture?*
Addressing Research Question 2 gave us an understanding of how Big Data solution integration impacts the EA of an organisation. This question helped us understand the entire landscape of Enterprise Architecture and how Big Data solutions proper fit within an organisation. Answering this question further enhanced our understanding of an Enterprise Architecture approach to information management; that Big Data is an enterprise asset and needs to be managed from business alignment to governance as an integrated element of the existing legacy information management architecture. This helped us answer questions related to business context, such as: How will we make use of Big Data? Which business processes can benefit from the use of Big Data?

Research Question 3: *How do we address the challenges of integrating Big Data solutions with legacy systems and data?*
Research Question 3 allowed us to generate a Big Data architecture to integrate Big Data solutions to legacy systems and data, which focuses on the challenges identified in the Research Question 1. We identified business processes which are using structured and semi-structured data that needs to be integrated for the purpose of analytics and decision making. Business processes were re-engineered to fit unstructured and semi-structured data. We addressed the challenges of integrating Big Data solutions with legacy systems and data involving the activities of re-engineering and system integration and data administration.

Big Data Strategy, Analytics Value Chain, EA, technology infrastructure, technology architecture, and data architecture of Big Data as discussed in chapter 4 built the components of our Big Data Architecture for BI and DA. The components of Analytic Value Chain form the building blocks for Big Data. Technology infrastructure, technology architecture and data architecture governed the technology choice of Big Data architecture to integrate legacy systems and data with Big Data Solutions. In our survey we have identified that technology choice and technology skills are one of the deterrents to implement Big Data solutions.

Research Question 4: *How can Big Data Analytics be combined with the Business Process using re-engineering?*

Research Question 4 allowed us to investigate how organisations, that have been mainly using customer and business information contained within in-house systems, are making use of and analysing external data and how these organisations are combining Big Data analytics, with business process workflows, to deliver benefits to the organisations and their customers. To combine Big Data analytics with higher education business processes, we re-engineered existing higher education business processes by identifying input and output from the core process of learning and teaching and identified how unstructured and external data sources can be helpful so that Big Data analytics can be applied to, and work effectively for, decision making for higher education.

In chapter 5 and 6, we demonstrated how Big Data analytics was combined with the business process using re-engineering to integrate legacy systems and data with Big Data Solutions using the Big Data architecture for E-coupons. Sending customers blanket promotions and traditional physical coupons results in only 5% of the coupons being used and 95% of the coupons being wasted and not used. With the use of Electronic Coupon Distributed System (ECDS), we have shown that 46% of the distributed coupons were used. Coupons and discounts do help sales, and if used and targeted well, can be effective in driving business forward.

In chapter 7 and 8, we have demonstrated the usefulness of our Big Data Architecture for Higher Education analytics to gain some critical insight from the data collected through the Learning Management System (LMS) and Early Alert Student Indicator (EASI). We have found many correlations regarding students' success. We found that access to online documents with students' final grade has positive linear correlation of 0.5593. This means that an increase in the online access to documents will correspond to an increase in the students' final grades.

9.3 Contributions and Significance of the Research

This section presents the contributions and the significance we made to the research community. The contributions and significance we made can be grouped into five areas to address the research questions as stated in the previous section. They are discussed in the following sections.

9.3.1 Identifying what are the issues and concerns of existing legacy systems within an organisation for supporting decision making

Chapter 3 reports on the survey we conducted where we had 97 respondents from different organisations. We had responses from industries such as financial services, healthcare, aviation, higher education, the energy sector, and insurance. The analysis of our survey was divided into six categories. We have demonstrated that many organisations are implementing Big Data solutions and integrating their legacy systems and data with their solutions. A Big Data architecture for Big Data integration with legacy systems should be able to provide solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats. The integration should be able to maintain data accuracy and integrity, and this should be addressed by the Big Data architecture.

The contribution we made is by identifying the requirements of organisations trying to achieve Big Data solutions for their organisations and what they are looking for while integrating legacy systems with Big Data solutions. What are the obstacles the organisations would like to resolve while integrating legacy systems with Big Data solutions?

9.3.2 Identifying current issues and concerns with the development of architecture to integrate Big Data solutions with legacy systems in organizations'

Based on the results of the survey we identified the building blocks of Big Data strategy and the analytics value chain and this is discussed in Chapter 4. We developed three layers for different functionality consisting of a Foundation Layer, Analytics Layer, and Application Layer. Big Data storage, processing, and reporting are handled in the Foundation Layer, analytics is handled in the Analytics Layer, and the result of analytics layer is visualised in the Application Layer. All three layers are the building blocks for integrating Big Data into the EA incorporating the Analytics Value Chain. Technology infrastructure, technology architecture, and data architecture govern the Big Data architecture to integrate legacy systems and data with Big Data solutions. The contribution we have made is towards identifying Analytics Value Chain as a part of the building blocks for streaming Big Data into EA.

9.3.3 Creating and developing a Big Data architecture to integrate Big Data Solutions with legacy systems and data

We applied Six Sigma (Tennant, 2001) activities for business process re-engineering and developed organisational Big Data strategy for re-engineering. We identified that Big Data analytics and business processes can be combined using re-engineering and can deliver benefits to the organisations and customers. In Chapter 5 we reported on combining Big Data analytics with business processes using re-engineering. We have focused on Business Process Re-engineering. We have used four major steps for re-engineering, and they are:

- Refocus company values on customer needs.
- Redesign core processes, often using information technology to enable improvements.
- Reorganise a business into cross-functional teams with end-to-end responsibility for a process.
- Rethink basic organisational and people issues.

We have identified seven processes of an e-business process model and they are:

1. Process of customer service
2. Process to ship product
3. Process to quality assurance
4. Process of credit card transaction
5. Process to extract customer information in/out of database
6. Data Processing
7. Process to extract information from external data sources.

We worked on the process “Customer Service” to target the Business Strategy called “improve customer intimacy”. To support the business strategy the key business initiatives we focused on are:

- Acknowledge key customers
- Acknowledge products
- Acknowledge suppliers
- Acknowledge market roles and responsibilities.

The contribution we made was by creating an e-business process model for a Big Data architecture to integrate data from different sources. We integrated Big Data sources in a Big Data architecture for processing and analysing so that business intelligence could be created and developed an e-business model.

9.3.4 Applying the Big Data architecture using open-source technologies

In Chapter 6 we report on applying the Big Data architecture using open-source technologies. We reported on how different types of data such as static information on gender, age, previous history (where the person travelled to, and what they bought), as well as real-time information about a customer’s current location and current shopping habits, was utilised in our Big Data architecture. The open source technologies we used are Apache Hadoop; Apache Spark Core Engine Spark APIs; Scala; and Apache Kafka. We developed architecture for coupon distribution, executing five phases:

- Signalling: detection of an event
- Triggering: association of an event with the set of rules defined for it
- Evaluation: evaluation of the conditional part for each triggered rule

- Scheduling: definition of an execution order between selected rules
- Execution: execution of all the actions associated with selected rules.

The contribution we made is the development of an architecture for Complex Event Processing (CEP) using open-source technologies and we documented the results showing how CEP is applied to the use case of an Electronic Coupon Distribution System.

9.3.5 Applying the Big Data architecture in Higher Education Institutions for Learning Analytics

In Chapter 7 we have discussed the use of Big Data analytics in the higher education sector and applied our BD architecture to the problem of student success rate and online engagement. We provided details of a study in applying Big Data architecture for analytics which was guided by how Big Data architecture can help higher education analytics incorporating Learning Management System datasets, and how by leveraging unstructured behavioural data/sentiment data we can predict the probability that any given student will pass or fail.

The method that we used to apply our Big Data architecture to an organisational problem involved the following steps:

- Identify business domains with the organisation and select the domain to work on.
- Identify the types of sources of data for the domain.
- Identify tools to be used within the Big Data architecture (could be different depending on data available).
- Apply the Big Data solution and generate the results.
- Apply the results to improve the organisations (which could include re-engineering business processes).

This helped us to identify the types and sources of data available for analytics in higher educational institutions. We applied our Big Data architecture based on open-source technologies for higher education institutions. We divided different tools for different layers to be used in our Big Data architecture. For Big Data Architecture we have processed, and analysed data generated by the Moodle Activity Viewer for students' online activity to courses they have passed to see the correlation between students' engagement to students' success rate.

In Chapter 8 we have evaluated and discussed how Big Data analytics can be combined with higher education business processes using re-engineering for structured data, unstructured data, and external data.

The contribution we made is the application of the Big Data architecture, applying it using a combination of tools on data sets and producing results that we applied to make organisational improvements.

9.4 Limitations and Future Work

While developing this architecture we did not pay much attention to addressing real-time analytics. This Big Data architecture should be applied to other case studies in different domains such as financials and the health sector to address real-time analytics. However, the possible obstacles the organisations can face are: insufficient understanding of Big Data technologies; complexity of big data technologies; complexity of managing data quality; and Big Data security issues.

There are a number of research issues related to the integration of legacy systems and data with Big Data solutions which should be addressed in the use of this architecture. Real-time analytics is the need of today's era. The architecture can be evolved to address real-time analytics using legacy systems and Big Data solutions.

Our approach can also be used by the data integration community on the topics of legacy data integration. There is evidence of solutions for achieving schema mapping using Hadoop. Clustering techniques can also be used on top of our architecture to address system integration concerns, such as interoperability of applications on heterogeneous platforms. System integration is a well-established area of applied research that addresses the problems related to the lack of systems and applications' interoperability in organisations and proposes novel solutions for system integration. However, despite research efforts to date, the proper scientific foundations for system integration remain elusive.

If an organisation wants to implement a Big Data solution and wants to capitalise on its potential for the business, the first thing all organisations should really understand is where Big Data fits into their business, and what kind of legacy systems and data organisations possess. Big Data architecture for integration of legacy systems and data would require selecting a business process for re-engineering. Implementing a Big Data architecture would suggest business and organisational impact of Big Data on business processes, and tasks to combine Big Data analytics with business processes.

REFERENCES

Abadi, D.J., Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., & Zdonik, S. (2003). Aurora: A new model and architecture for data stream management. *Very Large Data Base Journal*, *12*(2), 120–139.

AbdEllatif, M., Farhan, S.M., & Shehata, N.S. (2018). Overcoming business process reengineering obstacles using ontology-based knowledge map methodology. *Future Computing and Information Journal*, 3(1), 7-28.

Acharjya, D.P., & Ahmed, P. K. (2016). A survey on Big Data analytics: Challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(2), 511–518.

Aebi, D. (1997). Data re-engineering-A case study. *East-European Symposium on Advances in Database and Information Systems (ADBIS'97)*, Springer-Verlag Electronic Workshops in Computing, Ed.: C.J. van Rijsbergen.

Agrawal, J., Diao, Y., Gyllstrom, D., & Immerman, N. (2008). Efficient pattern matching over event streams. *ACM International Conference on Management of Data (SIGMOD)*, 147–160.

Akerkar, R. (Ed.). (2014). *Big Data computing*. Chapman and Hall/CRC Press.

Akter, S., & Wamba, S.F. (2016). Big data analytics in e-commerce: A systematic review and agenda for future research. *Journal of Electronic Markets*, *26*(2), 173–194.

Alkaseme, B.Y., Nour, M.K., & Meelud, A.Q. (2013). Towards a Framework to Assess Legacy Systems. *IEEE International Conference on Systems, Man, and Cybernetics*, 924-928.

Almeida, F., and Calistru, C. (2012). The main challenges and issues of Big Data management. *International Journal of Research Studies in Computing*, *2*(1), 11–20.

Altman, R., Natis, Y., Hill, J., Klein, J., Lheureux, B., Pessini, M., Schulte, R., & Varma, S. (1999). *Middleware: The glue for modern application*. Gartner Group, Strategic Analysis Report.

Anand, R., Kumar, M., Jhingran, A., & Mohan, R. (1998). Sales promotions on the Internet. *Second Usenix Conference on E-Commerce*, IBM T.J. Watson Research Center, 1–11.

Anand, A., Wamba, S.F., & Gnansou, D. (2013). A literature review on business process management, business process re-engineering, and business process innovation. *Workshop on Enterprise and Organisational Modeling and Simulation*, 1–23.

Apache Software Foundation. (2019). Hadoop Releases. Retrieved April 1 2020, from: https://archive.apache.org/dist/hadoop/common/

Arnold, K.E., & Pistilli, M.D. (2017). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of IEEE Frontiers in Education Conference (FIE)*, 1–8.

Arputhamary, B., & Arockiam, L. (2015). Data integration in Big Data environment. *International Journal of Data Mining*, *5*(1), 1–5.

Assiri, A., Emam, A., & Al-Dossari, H. (2016). Real-time sentiment analysis of Saudi dialect tweets using SPARK. *IEEE International Conference on Big Data*, 3947–3950.

Atif, A., Richards, D., Bilgin, A., & Marrone, M. (2013). Learning analytics in higher education: A summary of tools and approaches. *30th ASCILITE Conference*, 68–72.

Attaran, M., Stark, J.B., & Stotler, D. (2018). Opportunities and Challenges for Big Data Analytics in American Higher Education- A Conceptual Model for Implementation. *Industry and Higher Education*, 32(1), 1-14.

Auer, S., Ngonga Ngomo, A.C., Frischmuth, P., & Klimek, J. (2014). Linked data in enterprise integration. In R. Akerkar (Ed.), *Big Data Computing*. Chapman and Hall/CRC Press.

Baer, S., Boguss, E., & Green, D. (2011). Stuck on screens: Patterns of computer and gaming station use in youth seen in a psychiatric clinic. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *20*(2), 86–94.

Baker, P. (2015). *Data divination Big Data strategies*. Cengage Learning.

Barton, D., & Court, D. (2012). Making advanced analytics work for you. *Harvard Business Review Press*. Retrieved February 2 2019, from:

https://hbr.org/2012/10/making-advanced-analytics-work-for-you

Barth, C., & Koch, S. (2018). Critical success factors in ERP upgrade projects. *Industrial Management & Data Systems*. 119(3), 656-675.

Basaran, M. (2013). Reading Fluency as an Indicator of Reading Comprehension. Educational Sciences: Theory & Practice, 13(4), 2287-2290.

Barwick, H. (2014). Lack of in-house skills a barrier to Big Data adoption in A/NS report on CIO magazine. Retrieved November 28 2013, from: http://bit.ly/1dS41yD

Batlajery, B.V. (2013). *Revisiting legacy systems and legacy modernisation from the industrial perspective*. Master's Thesis Business Informatics. University of Utrecht, Utrecht, the Netherlands. Unpublished or available via database? See p. 49 of CQUni guide

Bennett, K. (1995). Legacy systems. *IEEE Software*, *12*(1), 19–23.

Benes, R. (2018). How legacy systems stifle marketing analytics: Outdated technology stands in the way of marketers' goals. Retrieved on 20th February 2019 from: https://www.emarketer.com/content/how-legacy-systems-stifle-analytics-progress

Beyer, M.A., and Laney, D. (2012). *The importance of "Big Data": A definition. Stamford, CT*. Gartner Research Report.

Bellahsene, S., Bonifati, A., & Rahm, E. (2011). *Comprehensive survey of current and past research on schema matching and mapping*. Springer.

Bhadani, A.K., & Jothimani, D. (2016). book chapter - *Big Data: Challenges and Opportunities, and Realities in Effective Big Data Management and Opportunities for Implementation*, IGI Global, 1-24.

Bichsel, J. (2012). *Analytics in higher education-Benefits, barriers, progress, and recommendations*. Research Report. *EDUCAUSE Center for Applied Research*.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. A report from US Department of Education, Washington, D.C..

Biocic, B., Tomic, D., & Ogrisovic, D. (2011). Economies of cloud computing. *International Convention on Information and Communication Technology,* *Electronics and Microelectronics (MIPRO)*, 1438–1442.

Bisbal, J., Lawless, D., Wu, B., & Grimson, J. (1999). Legacy information systems: Issues and directions. *IEEE Software*, *16*(5), 103–111.

Bisbal, J., Lawless, D., Wu, B., & Grimson, J. (1999). Legacy information system migration: A brief review of problems, solutions and research issues. *Computer Science*, 16(5),1–18.

Biser, C., & Heath, T. (2011). *Linked data: Evolving the web into a global data space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1–136. Morgan & Claypool.

Bloem, J., van Doorn, M., Duivestein, S., van Manen, T., van Ommeren, E., & Sachdeva, S. (2013). *No more secrets with Big Data analytics*. Book Production, The Sogeti Trend Lab VINT.

Bloomberg Businessweek. (2011). The current state of business analytics: Where do we go from here? Bloomberg Businessweek Research Services. Retrieved February 19 2015, from: https://www.sas.com/content/dam/SAS/bp_de/doc/studie/ba-st-the-current-state-of-business-analytics-2317022.pdf

Brenna, L., Demers, A., Gehrke, J., Hong, M., Ossher, J., Panda, B., Riedewald, M., Thatte, M., & White, W. (2007). Cayuga: A high performance event processing engine. *International Conference on Management of Data (SIGMOD '07),* 1100–1102.

Brodie, M.L., & Stonebraker, M. (2007). *Migrating legacy systems: Gateways, interfaces & the incremental approach*. Morgan Kaufmann.

Broekema, C.P., Boonstra, A.J., Cabesas, V.C., Engbersen, T., Holities, H., et al., (2012). DOME: Towards the ASTRON and IBM Center for Exascale Technology. In *Proceedings of the 2012 Workshop on High-Performance Computing for Astronomy Data*, 1–4.

Bry, F., & Eckert, M. (2007). Temporal order optimisations of incremental joins for composite event detection. *International Conference on Distributed Event-Based Systems (DEBS '07)*, 85–90.

Brynjolfsson, E., Hitt, L.M., & Kim, H.H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *Social Science Research Network.*

Retrieved February 26, 2018, from: http://dx.doi.org/10.2139/ssrn.1819486

Buchmann, A., & Koldehofe, B. (2009). Complex event processing. *IT-Information Technology*, *51*(5), 241–249.

Bulmer, D. & DiMauro, V. (2011). New Symbiosis of Professional Networks Study Benchmarks the Impact of Social Media on Enterprise Decision-Making. *2nd Annual New Symbiosis of Professional Networks study*.

Cantabella, M., Domingues de la Fuente, E., Martines-Espana, R., Ayuso, B., & Munos, A. (2017). Searching for behavior patterns of students in different training modalities through Learning Management Systems. In the *Proceedings of 13th International Conference on Intelligent Environments*, 44–51.

Capgemini. (2012). *The deciding factor: Big Data and decision making*. Economist Intelligence Unit, White Paper.

Capgemini. (2017). *Driven by the conviction that the business value of technology comes from and through people*. Annual Report. Retrieved February 26 2019, from: https://reports.capgemini.com/2017/wp-content/uploads/2018/03/CapG_RA17_UK-2.pdf

Carr, E. (2014). Building Big Data operational intelligence platform with Apache Spark, Guavus, Spark Summit. A Report by Apache Spark.

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From Big Data to big impact. *MIS Quarterly*, *36* (4), 1165–1188. doi: 10.2307/41703503.

Chen, H., Schüts, R., Kasman, R., & Matthes, F. (2017). How Lufthansa capitalised on Big Data for business model renovation. *MIS Quarterly Executive*, 16(1), 19–34.

Chen, C.P., & Shang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Science*, *275* (2014), 314–347.

Chen, Z., Xiao, N., & Liu, F. (2013). An SSD-based accelerator for directory parsing in storage system containing massive files. *Peer-to-Peer Networking and Applications*, 5(4), 31-45.

Chiasson, M.W., & Davidson, E. (2005). Taking industry seriously in Information Systems research. *MIS Quarterly*, 29(4), 591-605.

Collier, K. (2012). *Agile analytics: A value driven approach to business intelligence and data warehousing*. Pearson Education.

Comella-Dorda, S., Wallnau, K., Seacord, R., & Robert, J. (2000). *A survey of legacy system modernisation approaches*. A Report by Software Engineering Institute, Carnegie Mellon University, Pittsburgh.

Cooper, J., Noon, M., Jones, C., Khan, E., & Arbuckle, P. (2013). Big Data in life cycle assessment. *Journal of Industrial Ecology*, *17*(6), 796–799.

Crosling, G., Heagney, M., & Thomas, L. (2009). Improving student retention in higher education. *Australian Universities Review*, *51*(2), 9–18.

Cugola, G., & Margara, A. (2012). Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR) Surveys Homepage Archive*, *44*(3), 458–497.

Cugola, G., & Margara, A. (2013). Deployment strategies for distributed complex event processing. *Journal of Computing*, 2013(95), 129-156.

Cuomo, S., De Michele, P., Galletti, A., & Piccialli, F. (2016). A cultural heritage case study of visitor experiences shared on a social network. In *Proceedings of 10th International Conference on P2P, Parallel, Grid, Cloud Internet Computing*, 3PGCIC, 539–544.

Cusumano, M. (2010). Cloud computing and SaaS as new computing platforms. *Communications of the ACM*, *53*(4), 27–29.

DalleMule, L., & Davenport, T.H. (2017). What's your data strategy? *Harvard Business Review*. https://hbr.org/2017/05/whats-your-data-strategy

Daniel, B. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 1–17.

Davenport, T.H. (2014). *Big Data at work: Dispelling the myths, uncovering the opportunities*. Harvard Business Review Press.

Davenport, T.H., Cantrell, S., & Harris, J.G. (2004), Enterprise systems and ongoing process change, *Business Process Management Journal*, 10(1), 16-26.

Dawson, S. (2011). *Analytics to literacies: Emergent learning analytics to evaluate new literacies*. Workshop on New Media, New Literacies, and New Forms of Learning, London.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Published on Library Review*, *65*(3),122–135, doi: 10.1108/LR-06-2015-0061.

Desousa, K.C., & Jacob, B. (2017). Big Data in the public sector: Lessons for practitioners and scholars. *Administration & Society*, *49*(7),1043–1064.

Deursen, A.V., Klint, P., & Verhoef, C. (1999). Research issues in software renovation. *Fundamental Approaches to Software Engineering*, 99–105.

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier/Morgan Kaufmann.

Dong, X.L., & Srivastava, D. (2013). Big Data integration. *29th IEEE International Conference on Data Engineering*, 1–4.

Dong, X.L., & Naumann, F. (2009). Data fusion – Resolving data conflicts for integration. *Proceedings of Very Large Database Endowment*, 1–2.

Dora, P., & Sekharan, G.H. (2013). Healthcare insurance fraud detection leveraging Big Data analytics. *International Journal of Science and Research (IJSR)*, *4*(4), 2073–2076.

Eichmann, D. (1995). Application architectures for Web based data access. Workshop Web Access to Legacy Data. *Fourth International World Wide Web Conference*.

Ernst Young Business Report. (2014). *Big Data changing the way businesses compete and operate. Report on building a better working world*. Retrieved April 11 2016, from:
http://www.ey.com/Publication

Erricson-Connor, B. (2003). Truth and consequences. *Science Journal*, *48*(2), 38–43.

Evans, J.R., & Lindner, C.H. (2012). Business analytics: The next frontier for decision sciences. *Decision Science Institute Line*, *43*(28), 4–6.

Fan, C., Xiao, F., Madsen, H., & Wang, D. (2015). Temporal knowledge discovery in

Big BAS Data for building. *Energy Management, Energy and Buildings*, *190*(2015),75–89.

Faroukhi, A. Z., El Alaoui, I., Gahi, Y., & Amine, A. (2020). Big Data monetization throughout Big Data value chain: A comprehensive review. *Journal of Big Data*, *7*, 3 (2020). https://doi.org/10.1186/s40537-019-0281-5

Federation of Enterprise Architecture Professional Organisations. (2013). A common perspective on enterprise architecture. *Architecture and Governance Magazine*, *11*, 1–2.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with Big Data analytics. *Data Mining and Knowledge Discovery*, *18*(1) (2008), 140–181.

Franke, J., Charoy, F., & El Khoury, P. (2013). Architecture for co-ordination ofactivities in dynamic situations. *Enterprise Information Systems*, *7*(1), 33–60.

Fuhr, A., Horn, T., Riediger, V., & Winter, A. (2013). Model-driven software migration into service-oriented architectures. *Journal of Computer Science - Research and Development*, *28*(1), 65–84.

Fyrbiak, M., Strauss, S., Kison, C., Wallat, S., Elson, M., Rummel, N., & Paar, C. (2017). Hardware reverse engineering: Overview and open challenges. *IEEE 2nd International Verification and Security Workshop (IVSW)*. DOI: 10.1109/IVSW.2017.8031550.

Gable, G.G., Chan, T., & Tan, W.G. (2001). Large packaged application software maintenance: a research framework. *Journal of Software Maintenance and Evolution: Research and Practice*, 13(6), 351-371.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management*, *35*, 137–144.

Ganti, N., & Brayman, W. (1995). *Transition of legacy systems to a distributed architecture*. John Wiley & Sons.

García, E., Romero, C., Ventura, S., & de Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modelling and User-Adapted Interaction*, *19*, 99–132.

Gauger, J., Art, M.M., & Sondergeld, E. (2017). *Legacy systems and modernisation:*

Core systems strategy for policy administration systems. A white paper by Deloitte and LIMRA.

Gay, L. R., & Airasian, P. (2003). Educational research: Competencies for analysis and applications (7th ed.). Upper Saddle River, NJ: Pearson.

Ghemawat, S., Gobioff, H., & Leung, S.T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, *37*(5), 29–43.

Goldstein, P. J., & Katz, R. (2005). *Academic analytics: The use of management information and technology in higher education*. Educause, Centre for Applied Research.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. *ACM Conference on Online Social Networks*, 27–37.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic architecture for LA. *Educational Technology and Society*, *15*(3), 42–57.

Groff, J.R., & Welnberg, P.N. (2002). *SQL: The complete reference*, second edition, Mc-Graw Hills Companies.

Gruman, G. (2013). *Tapping into the power of Big Data. Price Warehouse Coopers Technology Forecast*. Making Sense of Big Data, Accessed on 14th January 2010 http://www.pwc.com/us/en/technology-forcast/2010/issue3/index.html

Hainaut, J.L. (1998). Database reverse engineering. [Doctoral Dissertation]. University of Namur- Institute d'Informatique, 211B-5000 Namur, Belgium.

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. *32nd International Conference on Very Large Data Bases*, 9–16.

Hammer, M., & Champy, J. (1993). *Re-engineering the corporation: A manifesto for business revolution*. Nicholas Brealey Publishing.

Hanafizadeh, P., & Moosakhani, M. (2009). Selecting the best strategic practices for business process redesign. *Business Process Management Journal*, 15(4), 609-627.

Hepp, M., Leymann, F., Domingue, J., Wahler, A., and Fensel, D. (2005). Semantic business process management: a vision towards using semantic Web services for business process management. *IEEE International Conference on e-Business*

Engineering (ICEBE'05), 535-540.

Hong, J.H., & Huang, M.L. (2011). A quality-aware approach towards the integration of feature-based geospatial data. *7th International Conference on Digital Content, Multimedia Technology and its Applications*, 33–38.

IBM Report. (2011). *The 2011 IBM Tech Trends Report: Tech Trends of Today. Skills for Tomorrow*. Retrieved February 1 2019, from: https://ai.arisona.edu/sites/ai/files/MIS510/2011ibmtechtrendsreport.pdf

Inoubli, W., Aridhi, S., Mesni, H., & Maddouri, M. (2018). An experimental survey on Big Data architectures. *Future Generations Computer Systems*, *86*, 546–564.

Ives, S.G., Florescu, D., Friedman, M., Levy, A., & Weld, D.S. (1999). An adaptive query execution system for data integration. *The ACM Special Interest Group on Management of Data*, 28(2), 299–310.

Jacobs, A. (2009). Pathologies of Big Data. *Communications of the Association for Computing Machinery*, 52(8), 36–44.

Jardim-Gonçalves, R., & Grilo, A. (2013). Systematisation of interoperability body of knowledge: The foundation for EI as a science. *Enterprise Information Systems*, *7*(1), 7–32.

Janssem, M., Voort, H., & Wahyudi, A. (2017). Factors influencing Big Data decision-making quality. *Journal of Business Research*, 70(2017), 338–345.

Janssen, M., Esteves, E., & Janowski, T. (2014). Interoperability in big, open, and linked data-organisational maturity, capabilities, and data portfolios. *IEEE Computer Society*, 47(2014), 44–49. DOI: 10.1109/MC.2014.290.

Jeble, S., Kumari, S., & Patil, Y. (2018). Role of Big Data in decision making. *Operations and Supply Chain Management*, 11(1), 36–44.

Jelonek, D., & Stępniak, C. (2014). Identification of e-customers’ activities in the context of creating strategy, management and managers facing challenges of the 21st century. In F. Bylok, I. Ubresiova, L. Cichobłasiński, & S. Istvan (Eds.), *Theoretical background and practical applications* 335–345. Egyetemi Kiado Nonprofit Kf, Godollo.

Jeston, J., & Nelis, J. (2008). *Business process management*, *Practical guidelines to successful implementations*. Third edition, Routledge, Taylor & Francis, London and New York.

Jha, M., & O'Brien, L. (2013). Re-engineering legacy systems for modernisation: The role of software reuse. *International Conference on Advances in Computer Sciences and Electronics Engineering*, 49–59.

Jha, M., & O'Brien, L. (2013). Comparison of modernisation approaches: With and without the knowledge-based software reuse process. *International Conference on Advances in Computer Science and Engineering Computer Science*. DOI:10.2991/cse.2013.17 Corpus id: 55569711.

Jha, M., & Maheshwari, P. (2005). Reusing code for modernisation of legacy systems. *IEEE Conference on Software Technology and Engineering Practice (STEP)*, 102–114.

Jha, M., Jha, S., & O'Brien, L. (2015). Integrating big data solutions into enterprise architecture: Constructing the entire information landscape. *The International Conference on Big Data, Internet of Things, and Zero-Size Intelligence (BIZ2015)*, 8-10 September 2015, Kuala Lumpur, Malaysia. Held at the Fourth World Congress on Computing, Engineering and Technology (WCET 2015), ISBN: 978-1-941968-18-5 ©2015 SDIW.

Jha, M., Jha, S., & O'Brien, L. (2016). Big data and its impact on enterprise architecture. In S.G. Tomar, N.S. Chaudhari, R.S. Bhadoria, & G.C. Deka (Eds.), *The Human Elements of Big Data: Issues, Analytics and Performance*, Chapman and Hall/CRC, 119–136.

Jha, M., Jha, S., & O'Brien, L. (2016). Social media and big data: A conceptual foundation for organisations. In R. Chugh (Ed.) *Harnessing Social Media as a Knowledge Management Tool*, IGI Global Publication, 315–332.

Jha, M., Jha, S., & O'Brien, L. (2016). Combining big data analytics with business process using re-engineering. *IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, 1-6, doi: 10.1109/RCIS.2016.7549307

Jha, M., Jha, S., & O'Brien, L. (2019). Re-engineering higher education learning and teaching business processes for big data analytics. In W. Abramowicz, & R. Corchuelo (Eds.), *Business Information Systems: 22nd International Conference, BIS 2019*, Seville, Spain, June 26–28, 2019 Proceedings, Part II Vol. 354 (pp. 233–244). Cham, Switzerland: Springer. doi:10.1007/978-3-030-20482-2_19.

Jin, X., Wah, B. W., Cheng, X, & Wang, Y. (2015). Significance and challenges of Big Data research. *Big Data Research*, *2*(2), 59–64.

Johnson, S. L., Gray, P., & Sarker, S. (2019). Revisiting IS research practice in the era of big data. *Elsevier, Information and Organisation*, *29*(1), 41–56.

Kaisler, S., Armour, F., Espinosa, J.A., & Money, W. (2013). Big data: Issues and challenges moving forward. *IEEE 46th Hawaii International Conference on System Sciences (HICSS)*, 995–1004.

Kang, B.K., & Bieman, J. (1998). Using design abstractions to visualise, quantify, and restructure software. *The Journal of Systems and Software*, *42*(2), 172–187.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68.

Katal, A., Wasid, M., & Goudar, R. (2013). Big data: Issues, challenges, tools and good practices. *IEEE 6th International Conference on Contemporary Computing (IC3)*, 404–409.

Kearny, C., Gerber, A., & van der Merwe, A. (2016). Data-driven enterprise architecture and the TOGAF ADM phases. *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016*, 004603–004608.

Khan, M. A., Uddin, M.F., & Gupta, N. (2014). Seven V's of Big Data: Understanding Big Data to extract value. *Conference of the American Society for Engineering Education (ASEE Sone 1)*, 1–5.

Khan, N., Yaqoob, I., Hashem, I., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M., & Gani, A. (2014). Big Data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal 2014*, ID: 71286. http://dx.doi.org/10.1155/2014/712826.

Kharote, M., & Kshirsagar, V. P. (2014). Data mining model for money laundering detection in financial domain. *International Journal of Computer Application*, *85*(16), 61–64.

King, I. (2014). Big education in the era of Big Data. *Federated Conference on Computer Science and Information Systems*.

Kitchin, R. (2013). Big Data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. Doi: 10.1177/2043820613513388.

Kościelniaka, H., & Putoa, A. (2015). Big Data in decision making processes of enterprises. *International Conference on Communication, Management, and Information Technology*, 65, 1052–1058.

Koseleva, N., & Ropaite, G. (2017). Big data in building energy efficiency: Understanding of big data and main challenges. *Procedia Engineering, 172*, 544–549.

Koskinen, J., Ahonen, J. J., & Sivula, H. (2005). Software modernisation decision criteria: An empirical study. *Ninth European Conference on Software Maintenance and Re-engineering (CSMR'05)*, 324–331.

Kotonya, G., & Hutchinson, J. (2007). A COTS-based approach for evolving legacy systems. *Sixth International Conference on Commercial-Off-the-Shelf(COTS)-Based Software Systems (ICCBSS'07),* 205–214.

Kroll, P., & Kruchten, P. (2003). *The rational unified process made easy: A practitioner's guide to the RUP.* Addison-Wesley Longman Publishing.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience, and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387–394.

Lane, J.E., & Johnstone, D.B. (2013). *Higher education systems 3.0: Harnessing systemness, delivering performance*. State University of New York Press.

Landmark Solutions. (2016). *The 7 pillars of Big Data. A White Paper of Landmark Solutions.* Retrieved June 10 2016, from: https://www.landmark.solutions/Portals/0/LMSDocs/Whitepapers/The_7_pillars_of_Big_Data_Whitepaper.pdf

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. Gartner, file No. 949. Retrieved February 6 2016, from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-ControllingData-Volume-Velocity-and-Variety.pdf

Langseth, J. (2014). BI-style analytics on Spark (without Shark) using SparkSQL and SchemaRDD. Soomdata, Spark Summit. A Report by Apache Spark.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., & Kruschwits, N. (2011). Big Data, analytics and the path from insights to value. *Sloan Management Review*, *52*(2), 21–31.

Lee, I. (2017). Big Data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*, 293–303.

Lehman, M.M., & Belady, L. (1985). *Program evolution: Process of software change*. Academic Press.

Lehman, M., Perry, D., & Ramil, J. (1998). Implications of evolution metrics on software maintenance. *International Conference on Software Maintenance,* 208–217.

Li, S., Anming, X., Naiyue, S., Jianbin, H., & Shong, C. (2009). A SOA modernisation method based on Tollgate model. *International Symposium on Information Engineering and Electronic Commerce*. doi: 10.1109/IEEC.2009.65

Liang, J., Yang, J., Wu, Y., Li, C., & Sheng, L. (2016). Big Data application in education: Dropout prediction in Edx MOOCs. In the *Proceedings of IEEE Second International Conference on Multimedia Big Data*, 440–443.

Liebowits, J., & Boca, R. (2013). *Big Data and business analytics.* Auerbach Publications/CRC Press.

Liu, X., Iftikhar, N., & Xie, X. (2014). Survey of real time processing systems for Big Data. *Proceedings of the 18th International Database Engineering & Application Symposium*, 356–361. Doi: 10.1145/2628194.2628251.

Long, P., & Seimen, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, *46*(5), 30–40.

Lněnička, M., Máchová, R., Komárková, J., & Čermáková, I. (2017). Components of Big Data analytics for strategic management of enterprise architecture. *12thInternational Conference on Strategic Management and its Support by Information Systems*.

Luckham, D. (2002). *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Pearson Education.

Luna, D.R., Mayan, J.C., Garcia, M.J., Almerares, A.A., & Househ, M. (2014). Challenges and potential solutions for Big Data implementations in developing countries. *Yearbook of Medical Informatics*, *9*(1), 36–41.

Lytras, M.D., & Raghavan, V. (2017). Big Data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *13*(1). DOI: 10.4018/IJSWIS.2017010101.

Malcolm, R., Morrison, C., Grandison, T., Thorpe, S., Christie, K., Wallace, A., Green, D., Jarrett, J., & Campbell, A. (2014). Increasing the accessibility to Big Data systems via a common services API. *IEEE International Conference on Big Data*, 883–892.

Malladi, S., Ramakrishna, G., Rao, K.K., & Babu, E.S. (2016). Analysis of Legacy System in Software Application Development: A Comparative Survey. *International Journal of Electrical and Computer Engineering*, 6(1), 292-297

Mamčenko, J., & Šileikienė, I. (2006). Intelligent data analysis of e-learning system based on data warehouse, OLAP and data mining technologies. *Proceedings of the 5th WSEAS International Conference on Education and Educational Technology*, 171–175.

Mann, S. (2012). 'Big data' and the changing role of enterprise architects. Essential Guide Tech Target. Retrieved August 24 2019, from: https://searchapparchitecture.techtarget.com/tip/Big-data-and-the-changing-role-of-enterprise-architects

Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. Retrieved January 20 2015, from: http://www.mckinsey.com/insights/mgi

Marr, B. (2015). *Big Data: Using SMART Big Data, analytics and metrics to make better decisions and improve performance*. Wiley Publications.

Marsh, O., Maurovich-Horvat, L., & Stevenson, O. (2014). Big Data and education: What's the Big Idea? *Big Data and Education Conference*, 14–20.

Mashayekhy, L., Nejad, M. M., Grosu, D., Shang, Q., & Shi, W. (2015). Energy-aware scheduling of MapReduce jobs for Big Data applications. *IEEE Transactions on Parallel and Distributed Systems*, *26*(10), 2720–2733.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work and think*. John Murray.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The management revolution. *Harvard Business Review*, *90*(10), 61–68.

McGee. J. (2005). Legacy systems: Why history matters. *Enterprise Systems Journal*, 10th November 2005. https://esj.com/articles/2005/10/11/legacy-systems-why-history-matters.aspx

Mervis, J. (2012). Agencies rally to tackle Big Data. *Science*, *36*(6077), 22. doi: 10.1126/science.336.6077.22.

Microsoft Research Group. (2013). *The Big Bang: How the Big Data explosion is changing the world*. Retrieved April 14 2016, from: http://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx

Mishra, S.K., Kushwaha, D.S., & Misra, A.K. (2009). Creating reusable software component from object-oriented legacy system through reverse engineering. *Journal of Object Technology*, *8*(5), 133–152.

Morgan, T.P. (2006). Legacy application modernisation strategies hinge on SOA. *Four Hundred iSeries and AS/400 Insight*, *15*(40), 14–28.

Murumba, J., & Micheni, E. (2017). Big Data analytics in higher education: A review. *International Journal of Engineering and Science (IJES), 6*(6), 14–21.

Nelson, G.S. (2017). Developing your data strategy. *Pharma SUG 2017- A White Paper HA01*, 1–15. Retrieved April 15 2016, from: https://support.sas.com/resources/papers/proceedings17/0830-2017.pdf

Newcomb, P. (2005). Architecture-Driven Modernisation (ADM). *12th Working Conference on Reverse Engineering (WCRE'05)*, 237–247.

Newman, M., & Zhao, Y. (2008). The process of enterprise resource planning implementation and business process re-engineering: tales from two Chinese small and medium-sized enterprises. *Information Systems Journal*, 18(4), 405-426.

Njerul, A.M., Omar, M.S., Yi, S., Paracha, S., & Wannous, M. (2017). Using IoT technology to improve online education through data mining. In the *Proceedings of the IEEE International Conference on Applied System Innovation*, IEEE-ICASI - Meen, Prior & Lam (Eds), 515–518.

Norris, F. H., Tracey, M., & Galea, S. (2009). Looking for resilience: Understanding the longitudinal trajectories of responses to stress. *Social Science and Medicine*, *68*(2009), 2190–2198.

O'Callaghan, A.J. (1996). *Practical experiences of object technology*. Stanley Thornes in association with UNICOM.

Odom, L.R., Morrow, J.R. (2006). What's this R? A correlational approach to explain validity, reliability and objectivity coefficients. *Measurements in Physical Education and Exercise Science*, 10(2), 137-145.

Olssak, C.M. (2016). Toward better understanding and use of business intelligence in organisations. *Information Systems Management*, *33*(2), 105–123.

Ong, V.K. (2015). Big Data and its research implications for higher education: Cases from UK higher education institutions. *IIAI 4th International Congress on Advanced Applied Informatics*, 487–491.

O'Reilly, T. (2007). What is Web 2.0? Design patterns and business models for the next generation of software. *Communications and Strategies*, *65*(1), 17–37.

Oladipo, F.O., & Raiyetumbi, J.O. (2015). Re-engineering legacy data migration methodologies in critical sensitive systems. *Journal of Global Research in Computer Science*, *6*(11), 1–6.

Oussous, A., Benjelloun, F., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of Computer and Information Sciences*, *30*(4), 431–448.

Pappas, I.O., Mikalef, P., Giannakos, M.N., Krogstie, J., & Lekakos, G. (2018). Big data and business analytics ecosystems: paving the way towards digital transformation and sustainable societies. *Information Systems and e-business Management*, 16 (2018), 479-491.

Paule-Ruis, M.P., Riestra-Gonsales, M., Sánches-Santillan, M., & Péres-Péres, J.R. (2015). The procrastination related indicators in e-learning platforms. *Journal of Computer Science*, *21*(1), 7–22.

Pearson, T., & Wegener, R. (2013). *Big Data: The organisational challenge*. Bain and Company. Retrieved on 1st February 2019 https://www.bain.com/contentassets/25c167a5149c42168994338f9dc99ffe/bain_brief_big_data_the_organizational_challenge2.pdf

Pentreath, N. (2014). Using Spark and Shark to power a real-time recommendation and customer intelligence platform. *Graphflow, Spark Summit*.

Peres-Castillo, R., de Gusman, I. G., Piattini, M., & Avila-Garcia, O. (2008). On the use of ADM to contextualisation data on legacy source code for software modernisation. *Working Conference on Reverse Engineering*, 128–132.

Picciano, A.G. (2012). The evolution of Big Data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, *16*(3), 9–20.

Praveena Anto, M.D., & Bharathi, B. (2017). A survey paper on Big Data analytics. *International Conference of Information, Communication & Embedded Systems* (ICICES). doi:10.1109/ICICES.2017.8070723

Price Waterhouse Coopers. (2014). *Deciding with data: How data-driven innovation is fuelling Australia's economic growth*. Retrieved January 19 2015, from: https://www.pwc.com.au/consulting/assets/publications/data-drive-innovation-sep14.pdf

Raghupathi, W., & Raghupathi, V. (2014). Big Data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*(3). doi: 10.1186/2047-2501-2-3

Rahgosar, M., & Oroumchian, F. (2003). An effective strategy for legacy system evolution. *Journal of Software Maintenance and Evolution: Research and Practice*, *15*(5), 325–344.

Rajavat, A., & Tokekar, V. (2014). Effect of managerial dimensions on re-engineering process of legacy software systems. *The Conference on IT in Business, Industry and Government (CSIBIG)*, 1–6.

Reinsel, D., Gants, J., & Rydning, J. (2017). Data age 2025: The evolution of data to life-critical. seagate.com. Framingham, MA, US: International Data Corporation.

Reports n Reports. (2016). Social media analytics market to rise at 27.6% CAGR to 2020. *PRNewswire*. Retrieved February 16 2019, from: https://www.prnewswire.com/news-releases/social-media-analytics-market-to-rise-at-276-cagr-to-2020-568584751.html

Rice, W.H. (2006). *Moodle e-learning course development: A complete guide to successful learning using Moodle*. Packet Publishing.

Riffai, M.M.M.A., Edgar, D., Duncan, P., & Al-Bulushi, A.H. (2016). The potential for Big Data to enhance the higher education sector in Oman. In the *Proceedings of 3rd*

MEC International Conference on Big Data and Smart City,1–6.

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computer and Education*, *51*(4), 368–384.

Sakr, S., Liu, A., Batista, D.M., & Alomari, M. (2011). A survey of large scale data management approaches in cloud environments. *IEEE Communications Society Surveys and Tutorials, 13*(3), 311–336.

Salehi, A. (2010). *Low latency, high performance data stream processing: Systems architecture, algorithms and implementations*. VDM Verlag.

Salim, S.E. (2014). Service oriented enterprise architecture for processing Big Data applications in cloud. *International Journal of Engineering Sciences & Research Technology*. *3*(6), 647–655.

Santos, I., Escudeiro, P., & Vas de Carvalho, C. (2014). ICT Ways Network: ICT in Science Classrooms. In the *Proceedings of 9th Conferencia Ibérica de Sistemas Tecnologías de Informacion*.

Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Educational Technology and Society*, *17*(4), 117–132.

Schmarso, B. (2013). *Big Data: Understanding how data powers big business*. John Wiley & Sons.

Seacord, R.C., Plakosh, D., & Lewis, G.A. (2003). *Modernising legacy systems: Software technologies, engineering processes, and business practices*. Addison- Wesley.

Seetharam, T.H., Murlidhar, N.N., & Chandrasekaran, K. (2017). Implication of legacy software system modernisation – A survey in a changed scenario. *International Journal of Advanced Research in Computer Science*, *8*(7), 19–29.

Shvachko, K., Hairong, K., Radia, S., and Chansler, R. (2010). The Hadoop Distributed File System. *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1-10.

Siemens, G., Dawson, S., & Lynch, G. (2013). *Improving the quality and productivity of the higher education sector: Policy and strategy for systems-level deployment of learning analytics.* Office of Learning and Teaching. Australian Government.

Simon, P. (2014). Why Big Data in the enterprise is mostly lip service. *Information Week*. Retrieved February 19 2014, from http://ubm.io/MeGug1

Singh, D., & Reddy, C. (2015). A survey on platforms for Big Data analytics. *Journal of Big Data*, *2*(1), 1–20.

Sivarajah, U., Kamal, M.M., Irani, S., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286.

Solaimani, M., Gopalan, R., Khan, L., Brandt, P.T., & Thuraisingham, B. (2016). Spark-based political event coding. *IEEE Second International Conference on Big Data Computing Service and Applications (Big Data Service)*, 14–23.

Sood, A., & Lings, I. (2010). Empowerment and role stress in the human interface between the firm and its markets. *International Journal of Services, Technology and Management*, *14*(2–3), 233–249.

Smolan, R., & Erwitt, J. (2012). *The human face of Big Data*. Sterling Publishing Co.

Sneed, H.M. (1996). Encapsulating legacy software for use in client/server systems. *Working Conference on Reverse Engineering*, 104–119.

Søberg, J., Goebel, V., & Plagemann, T. (2008). To happen or not to happen: Towards an open distributed complex event processing system. *Proceedings of the 5th ACM Middleware doctoral symposium*, 25–30.

Srinivas, M., Ramakrishna, G., Rao, K.R., & Babu, E.S. (2016). Analysis of legacy system in software application development: A comparative survey. *International Journal of Electrical and Computer Engineering*, *6*(1), 292–297.

Strand, M., & Syberfeldt, A. (2020). Using external data in a BI solution to optimise waste management. Journal of Decision Systems, 29(1), 53-68.

Sumathi, N., Gokulakrishnana, S., Kaushik Ramana, S., Muralitharan, R., & Kamal, V. S. (2017). Application of Big Data systems to airline management. *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS)*, *6*(12), 129–132.

Tabesh, P. (2019). Implementing Big Data strategies. *Business Horizons*, 1–31.

Tanwar, M., Duggal, R., & Khatri, S.K. (2015). Unravelling unstructured data: A wealth of information in Big Data. *International Conference on Reliability, Infocom Technologies and Optimisation: Trends and Future Directions*, 1–6.

Tekinerdogan, B., & Oral. A. (2017). Performance isolation in cloud-based Big Data architectures. *Journal of Software Architecture for Big Data and the Cloud*, 2017, 127-145.

Tennant, G. (2001). *Six Sigma: SPC and TQM in manufacturing and services*. Gower Publishing, Ltd.

Thau, B. (2014, January 24). How Big Data helps stores like Macy's and Kohl's track you like never before. *Forbes*. Retrieved June 24 2015, from https://www.forbes.com/sites/barbarathau/2014/01/24/why-the-smart-use-of-big-data-will-transform-the-retail-industry/#1e655b46de8a

The RP Group. (2010). Possibilities for improving student success using predictive analytics: A report. *The Research Planning group for California Community Colleges*.

Thusoo, A., Shao, S., Anthony, S., et al. (2010). Data warehousing and analytics infrastructure at Facebook. *ACM SIGMOD International Conference on Management of Data*, 1013–1020.

Tomar, G.S., Chaudhari, N.S., & Bhadorai, R.S. (2016). *The human elements of Big Data: Issues, analytics, and performance*. CRC Publications.

Tsai, C.W., Lai, C.F., Chao, H.C., & Vasilakos, A.V. (2015). Big Data analytics: A survey. *Journal of Big Data*, *2*(21). Doi: 10.1186/s40537-015-0030-3

Tsao, N.L., Kuo, C.H., Guo, T.L., & Sun, T.J. (2017). Data consideration for at-risk students early alert. *IIAI International Congress on Advanced Applied Informatics*, 208–211.

Tulasi, B. (2013). Significance of Big Data and analytics in higher education. *International Journal of Computer Applications*, *68*(14), 23–25.

Tuomisaari, H., Nyberg, T., Karjalainen, J., Makelin, M., & Xiong, G. (2012). Business model transformation as an explanation of dramatic demand supply change events. *IEEE International Conference on Service Operations and Logistics, and Informatics*, 345–349.

Ulmer, J.S., Belaud, J.P., & Le Lann, J.M. (2013). A pivotal-based approach for enterprise business process and IS integration. *Enterprise Information Systems*, *7*(1), 61–78.

Van der Aalst, W. (2012). Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems*, *7*(1), 7–17.

Van der Linden, P. (2018). *Organisations need to give unstructured data its rightful place if they want to get value out of data*. Capgemini Report. Retrieved February 4 2019, from https://www.capgemini.com/2018/08/reorganising-unstructured-data/

Vernadat, F.B. (1996). *Enterprise modelling and integration: Principles and applications.* Chapman and Hall.

Vidgen, R., Shaw, S., & Grant, D.B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, *261*(2), 626–639.

Vijaya, A., & Venkataraman, N. (2018). Modernising legacy systems: A re-engineering approach. *International Journal of Web Portals,* 10(2), 50-60.

Voas, J.M. (1998). The challenges of using COTS software in component-based development. *Journal of Computer Science*, *44*(3), 31–37.

Wagner, E., & Ice, P. (2012). Data changes everything: Delivering on the promise of learning analytics in higher education. *Educause Review*, *47*(32), 1–18.

Waller, M.A., & Fawcett, S.E. (2013). Data science, predictive analytics, and Big Data: A revolution that will transform supply chain design and management, *Journal of Business and Logistics*, *34*(2), 77–84.

Wamba, S.F., & Mishra, D. (2017). Big data integration with business processes: A literature review. *Journal on Business Process Management*. *23*(3), 477–492.

Ward, J.S., & Barker, A. (2013). Undefined by data: A survey of Big Data definitions. *Journal of Cloud Computing*, *3*(1), 24–38.

Watson, H.J. (2014). Tutorial: Big Data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*. *34*(65), 1248–1268.

Webb, E.J., Campbell, D.T., Schwartzs, R.D., and & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally. *Social Forces*, *45*(2), 290–291.

Weiderman, N., Northrop, L., Smith, D., Tilley, S., & Wallnau, K. (1997). *Implications of distributed object technology for re-engineering. Technical Report CMU/SEI-97-TR-005*. Carnegie Mellon University.

Winsberg, P. (1995). Legacy code: Don't bag it, wrap it. *Journal on Datamation*. *41*(9), 36–41.

Wongchinsri, P., & Kuratach, W. (2016). A survey of data mining architectures in credit card processing. *13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1–6.

Wu, E., Diao, Y., & Risvi, S. (2006). High-performance complex event processing over streams. *ACM International Conference on Special Interest Group on Management of Data (SIGMOD)*, 407–418.

Xiao, W., Guoqi, L., & Bin, L. (2017). Research on Big Data integration based on karma modeling. *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 245–248.

Xiaogao, Y., & Ruiqing, P. (2017). Research on Big Data-driven high-risk students prediction. *IEEE International Conference on Cloud Computing and Big Data Analysis*, *28*(30), 145–149.

Youseff, L., Butrico, M., & Da Silva, D. (2008). Toward a unified ontology of cloud computing. *IEEE Grid Computing Environments Workshop*. doi: 10.1109/GCE.2008.4738443.Zachman, J. (2011). Zachman Enterprise Architecture Framework. *Zachman International*.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. *Proceedings of the International Conference on Advances in Cloud Computing (ACC)*, arXiv:1301.0159.

Zhao, D., Zhang, S., Zhou, X., Li, T., Wang, K., Kimpe, D., Carns, P., Ross, R., & Raicu, I. (2014). FusionFS: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems. *IEEE International Conference on Big Data*, 61–70.

Zhao, X., Ma, H., Zhang, H., Tang, Y, $ Kou, Y. (2015). HVPI: Extending Hadoop to support video analytic applications. *IEEE 8th International Conference on Cloud Computing*, 789-796.

Zhao, W.X., Li, S., He, Y., Chang, E.Y., Wen, J., & Li, X. (2016). Connecting social media to e-commerce: Cold-start product recommendation using micro-blogging information. *IEEE Transactions on Knowledge and Data Engineering*, *28*(5), 1147–1159.

Zhou, K., Fu, C., & Yang, S. (2015). Big Data driven smart energy management: From Big Data to big insights. *Renewable and Sustainable Energy Reviews*, *56*(2015), 215–225.

Zhou, J., Hu, L., Wang, F., Lu, H., & Zhao, K. (2013). An efficient multidimensional fusion algorithm for IoT data based on partitioning. *Tsinghua Science and Technology*, *18*(4), 369–378.

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill Osborne Media.

Zou, Y., & Kontogiannis, K. (2000). Web-based specifications and integration oflegacy services. *Conference of the Centre for Advanced Studies on Collaborative Research*, 7–17.

Zoufaly, F. (2002). *Issues and challenges facing legacy systems*. MEX Maintenance Software. Retrieved February 26 2019, from: https://www.developer.com/mgmt/article.php/1492531/Issues-and-Challenges-Facing-Legacy-Systems.htm

APPENDIX A

Information Sheet

A survey to identify existing approaches towards Legacy Systems and Data Integrationto Big Data solutions.

This survey will help to get a better understanding of existing approaches that are used in organisations for integrating legacy systems to Big Data solutions. This survey will also help us in identifying issues and concerns related to the implementation of Big Data solutions in organisations. The results from this survey will help in our thinking towards creating an effective process for integrating legacy systems and data into Big Data solutions which has notyet been completely answered by the research community. This survey will benefit data scientists, data analysts, decision makers, managers, educators, and others working in big datacommunity and research community about their beliefs, and practice in integrating legacy systems and data to Big Data solutions.

The survey should not take more than 20 minutes. We appreciate your valued time and your response to this survey. The survey questions will be divided into four categories: organisational information, general questions, legacy systems and data issues and concerns, and Big Data initiatives and implementation concerns. Please provide a more detailed answer than YES or NO where appropriate. The survey responses will be anonymous. In addition to that survey participant will be given an opportunity to send separate email request to receive acopy of the survey results.

Your participation in this research survey is voluntary. You may withdraw at any time prior tocompleting the survey simply by closing the survey window. If you wish to withdraw prior tocompleting the survey, the information you have already provided can be deleted. Online information will only be saved once you press submit button. In other modes of data collection,you can ask to destroy the copy of data collected.

If you feel upset at any time during the survey please contact *CQUniversity's Office ofResearch(Tel: 07 4923 2603;*

E-mail: ethics@cqu.edu.au; Mailing address: Building 32,

CQUniversity, Rockhampton QLD 4702) should there be any concerns about the nature and/orconduct of this research project.

This project has received ethical clearance from the CQUniversity Human Research Ethics Committee and is being undertaken by Mr Sanjay Jha. If you have any questions about this project you can contact:
Sanjay Jha s.jha@cqu.edu.au (Researcher Details)

Concerns / Complaints

Please contact CQUniversity's Office of Research (Tel: 07 4923 2603; E-mail: ethics@cqu.edu.au; Mailing address: Building 32, CQUniversity, Rockhampton QLD 4702) should there be any concerns about the nature and/or conduct of this research project.

ELECTRONIC CONSENT:

Clicking on the "agree" button below indicates that: (include relevant statements as indicated on the consent form)

□ You have read the above information

□ You voluntarily agree to participate; and

□ You give your consent for the data you provide in the following survey to be used for the research purpose described above.

Can we share the detailed information you submit?

Please be as open as possible. We are collecting this information primarily to help the Big Data community to make informed decisions. If there are a few details you would like to keep confidential you may choose to submit them by email instead of including them in this survey.

- Yes, share it publicly
- Yes, share it with organisations participating in the survey
- No, do not share this information except in aggregate/ anonymous form
- Any other comments:

Section 1: Big Data Organisational Questions (1–3)

1. Name of the organisation

2. Name of division or department.

3. If applicable name Big Data projects in the organisation

Section 2: General Questions (4–7)

4. What is your position in the organisation?
   - Manager
   - Decision Maker
   - Data Scientist
   - Other please specify

5. What is your educational level?
   - Big Data related Qualification (Please specify)
   - Research in IT
   - Postgraduate in IT
   - Bachelors in IT
   - Other Professional Qualifications

6. What area do you consider yourself working in?
   - Big data Analytics
   - Traditional Analytics
   - Business Intelligence
   - A researcher
   - Other please specify

7. How many years of experience do you have in this area?
   - Less than 5 years
   - 5-10 years
   - 10-20 tears
   - More than 20 years

Section 3: Legacy Issues and Concerns Questions (8–13): This section captures answers about legacy systems and data, advantages, disadvantages, and factors influencing legacysystems and data in the business intelligence community.

8. Do you use legacy systems for generating reports and decision making?
   - Yes
   - No
   - Any other system for reporting please specify

9. What kind of data does your organisation have?
   - Proprietary data
   - Open data
   - Acquired data
   - Social media data
   - Other please specify

10. What do you think are the benefits of legacy systems and data in your organisation?

11. What do you think are the main disadvantages with reporting when using legacy system and data?

12. Where and when do you use legacy systems and data for decision making?

13. In your organisation do you mainly use your data to look back and assess or think ahead?

- Look back
- Think ahead
- Any other reason, please specify

Section 4: Big Data Initiatives and Implementation Questions (14–31). This section captures answers about finding the proper fit of a Big Data solution and technologies for an organisation.

14. What is the volume of data generated in your organisation every year?

- Gigabytes
- Petabytes
- Zettabytes
- Please Comment if any other answer

15. Does your organisation have information management Big Data and analytics capabilities?

- Yes (Please answer question 17)
- No (Please answer question 16)
- Please comment if any other answer

16. Do you think that integrating a Big Data solution will benefit your organisation?

- Yes
- No
- Please comment if any other answer

17. Has your organisation developed a Big Data Strategy?

- Yes
- No
- Please comment if any other answer

18. What kind of analytics is used in your organisation?

- Descriptive
- Predictive
- Prescriptive
- Please comment if any other answer

19. How do you generate reporting for business intelligence?

20. Do you use legacy systems and data for business intelligence and decision making?

- Yes
- No
- Please comment if any other answer

21. What approaches does your organisation use to integrate legacy systems and data?

22. Is there an architecture to integrate legacy systems and data to Big Data solutions in your organisation?
    - Yes
    - No
    - Please comment if any other answer

23. When making a business decision in your organisation what do you mostly rely on?

24. Do you see value in integrating legacy systems to Big Data solutions in your organisation?
    - Yes
    - No
    - Please comment if any other answer

25. What’s hindering your organisation’s progress towards Big Data solution?
    - Organisational resistance: “it’s just too hard”
    - Lack of access to technology or investment
    - Lack of the right skills or access to training
    - Lack of data (whether proprietary, open data or acquired data)
    - Data security and privacy issues
    - Please comment if any other answer

26. Does your organisation use any Big Data technology? Please specify the technology in the text box below.

- Yes
- No

27. What is the biggest challenge in your organisation for collecting data?

28. What are the biggest challenges in your organisation for accessing data?

29. What are the challenges in your organisation for storing data?

30. What are the challenges in your organisation for processing data?

31. What are the challenges in your organisation for analysing data?