

A STUDY OF CONCEPTUAL DATA MODELLING IN THE ERA OF BIG DATA: A CASE OF ZAMBIA

By

Listone Kaputula

A Dissertation Submitted to the University of Zambia in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Information and
Communication Technology

School of Engineering

Department of Electrical and Electronics Engineering

University of Zambia

LUSAKA

2019

COPYRIGHT

All rights reserved. No part of this dissertation may be reproduced without prior consent of author and all information derived should be fully acknowledged.

©2019 Listone Kaputula

DECLARATION

I hereby truthfully declare that I am the sole author of this report and that all content is my original work that has not been presented before for an award at any university or learning institution:

Signed_____ **Date:** _____

Listone Kaputula

Author

Signed_____ **Date:** _____

Dr. Dani Eliya Banda

Supervisor

APPROVAL

This dissertation of Listone Kaputula has been approved as fulfilling the requirements for the award of Master of Engineering in Information and Communication Technology by the University of Zambia.

Name	Signature	Date
.....		
Internal Examiner 1		
.....		
Internal Examiner 2		
.....		
Internal Examiner 3		
.....		
Chairperson Board of Examiners		
.....		
Supervisor		

ACKNOWLEDGEMENT

I would wish to recognize and acknowledge the contribution made by various organisations and individuals towards the successful completion of this study. However, the special acknowledgement goes to, my supervisor, Dr. Dani Eliya Banda, for his professional guidance in this study.

Further, I would want to thank my wife Hope and our two sons (Lukumo and Muteende) for being by my side and for the motivation to pursue further studies.

Lastly but not the least, I would like to express my heartfelt gratitude to all the lecturers and management of University of Zambia, School of Engineering who supported me throughout the study.

ABSTRACT

The way data is collected and stored has evolved over time due to various developments which have led to advancements in database technologies. Few decades ago, only structured data were stored, however, with the increase in the business demand for data in this competitive global market, developments in new technologies for storing data has exponentially emerged. This has resulted in unstructured data also being accommodated in the databases. These new technologies have, however, come with challenges on how unstructured data can be modelled before it is stored. This research project dealt with the study of the conceptual data modelling applied in the era of digital world especially when storing data in NoSQL databases. The researcher, further explores the use of non-relational databases (type of databases which is developed to handle unstructured data) in the selected industries from public, private and non-governmental organisations in Zambia. The study was conducted to have feasible approach for conceptual data modelling that can help to mitigate the challenges of conceptual data modelling for non-relational databases. From the study, indication shows that most research works are on transforming conceptual data models for relational database to a particular logical or physical model of a non-relational database. Further, the study shows that most local organisations use relational databases and have limited or no knowledge on NoSQL databases. The results show lack of skills to create conceptual schema for non-relational database was a hindrance to Big Data implementation. This is attributed to insufficient knowledge about Big Data. The findings in this study reveals that local firms and organisations are yet to start implementing non-relational databases for data storage. It also showed that the use of Big Data is still at infancy stage. The results obtain in this study were used to create a proposed approach for conceptual data modelling that can be used in NoSQL databases.

Key words: Big Data, Conceptual data modelling, Database, Dataset, NoSQL.

TABLE OF CONTENTS

COPYRIGHT	i
APPROVAL.....	iii
ACKNOWLEDGEMENT.....	iv
ABSTRACT	v
LIST OF FIGURES	ix
LIST OF TABLES.....	x
LIST OF APPENDIXES	xi
ABBREVIATION AND ACRONYMS	xii
CHAPTER 1: INTRODUCTION	1
1.1 Background to the Study	1
1.2 Statement of the problem.....	8
1.3 Objectives	9
1.4 Research question	9
1.5 Purpose of the study	9
1.6 Thesis disposition.....	9
CHAPTER 2: LITERATURE REVIEW	11
2.1 Dataset	11
2.1.1 Small Data Vs Big Data.....	11
2.1.2 Big Data	13
2.1.3 Types of Dataset in Big Data	16
2.1.4 Sources of Big Data	17
2.2 Evolution of Databases.....	21
2.2.1 Database Types Overview	22
2.2.2 Relational verses Non-Relational Database	25
2.2.3 Relational Database	28
2.2.4 NoSQL database	31
2.3 Data Modelling	35

2.3.1	Features of Data Modelling.....	35
2.3.2	Conceptual Data Modelling	39
2.3.3	Roles of Conceptual Data Modelling	42
2.3.4	Benefit of Conceptual Data Modelling.....	43
2.3.5	Purpose of Conceptual Data Modelling.....	45
2.3.6	Data Modelling in Non-Relational Databases.....	45
2.3.7	Challenges of Conceptual Data Modelling in Non-Relational Databases	47
2.4	Analysis on the Peer-Reviewed Papers on Conceptual Data Modelling for Non-Relational Databases.....	48
2.4.1	Discussions on the Findings from Peer-Reviewed Papers on Conceptual Data Modelling for Non-Relational Databases	50
CHAPTER 3: METHODOLOGY		54
3.1	Introduction.....	54
3.2	Research Topic Selection	54
3.3	Study Area	54
3.4	Research Approach	55
3.5	Research Design.....	56
3.5.1	Research Instruments.....	57
3.6	Data Collection	57
3.6.1	Questionnaire Design.....	57
3.6.2	Primary Data Collection	58
3.6.3	Secondary Data Collection.....	59
3.7	Data Analysis.....	59
3.7.1	Manual Data Analysis.....	60
3.7.2	Computer Based Data Analysis.....	60
3.8	Research Findings and Discussions	60
3.9	Limitation	61
CHAPTER 4: DATA COLLECTION AND ANALYSIS		62
4.1	Research Findings	62
4.1.1	Bio Data	62
4.1.2	Dataset in the Digital World	63

4.1.3	Databases	66
4.1.4	Data Modelling Overview.....	68
4.1.5	Conceptual Data Modelling	70
CHAPTER 5: DISCUSSIONS AND CONCLUSION		72
5.1	Discussions	72
5.1.1	Demographic Data.....	72
5.1.2	Dataset in the Digital World	72
5.1.3	Database Deployments in Zambia.....	73
5.1.4	Data Modelling.....	74
5.1.5	Proposed Conceptual Modelling Approach for Non-Relational Database ...	76
5.2	Conclusion	80
5.3	Recommendations	82
REFERENCES		83
APPENDICES		92

LIST OF FIGURES

Figure 2. 1 The 5Vs of Big Data	13
Figure 2. 2 Big Data Sources	18
Figure 2. 3 Object-Oriented Data Model Vs Relational Data Model.....	24
Figure 2. 4 CAP Theorem with databases that chooses AP, CA and CP	27
Figure 2. 5 Relational Database Table	29
Figure 2. 6 Example of a Conceptual Data Model	36
Figure 2. 7 Example of Logical Data Model Design.....	37
Figure 2. 8 Example of Physical Data Model Design	38
Figure 2. 9 Artifacts of conceptual modelling	41
Figure 2. 10: Target meta-model.....	51
Figure 3. 1 Research Design	56
Figure 4. 1 Respondents' major roles in their organisation	63
Figure 4. 2 Dataset Types Stored in Organisations	64
Figure 4. 3 Big Data Rating locally	65
Figure 4. 4 Application of Big Data	66
Figure 4. 5 The types of databases in use in local organisations	67
Figure 4. 6 Necessity of Data Modelling.....	68
Figure 4. 7 Data Modelling Tools	69
Figure 4. 8 Areas of Application for Conceptual Modelling	70
Figure 5. 1 Entity Identification of the collected Data	78
Figure 5. 2 Proposed NoSQL Conceptual data modelling.....	80

LIST OF TABLES

Table 2. 1: Small Data Verses Big Data in terms of 5V	12
Table 2. 2 Difference between relational and non-relational databases	28
Table 2. 3 Data Modelling in Relational Database.....	35
Table 2. 4 Peer Reviewed Paper Analysed	49
Table 4. 1 Respondents' Industries	62
Table 4. 2 Data Modeler Skills.....	69
Table 4. 3 Areas of Application for EDM	71
Table 5. 1 Entity Identification and Documentation	79
Table 5. 2 Identifying relationship types	79

LIST OF APPENDIXES

Appendix 1: Questionnaire	92
Appendix 2: Bio Data Questions coding table	101
Appendix 3: Technical Data Questions coding table	102
Appendix 4: Journal Published.....	106

ABBREVIATION AND ACRONYMS

ACID	Atomicity, Consistency, Isolation and Durability
ATM	Automated Teller Machine
BASE	Basically Available, Soft-state, Eventual consistency
CAP	Consistency, Availability and Partition-tolerance
CPU	Control Processing Unit
CRUD	Create, Read, Update and Delete
DBMS	Database Management System
EER	Enhanced entity–relationship
ERD	Entity Relational Diagram
FFS	Format File System
GB	Gigabyte
GDBS	Graph Database Schema
I/O	Input/output
IBM	International Business Machine
ICT	Information and Communication Technology
ID	Identity
IDMS	Integrated Data Management System
IDS	Integrated Data Store
IEEE	Institute of Electrical and Electronics Engineers
IMS	Information Management System
IoT	Internet of Things

ISAM	Indexed Sequential Access Method
ISP	Internet Service Provider
IT	Information Technology
MDA	Model Driven Architecture
MySQL	My Structured Query Language
NoSQL	Not Only Structured Query Language
ORDBMS	Object-Relational Database Management System
PB	Petabyte
PC	Personal Computer
SEQUEL	Structured English Query Language
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
TTL	Time To Live
TV	Television
UML	Unified Modelling Language
UNZA	University of Zambia

CHAPTER 1: INTRODUCTION

1.1 Background to the Study

Data is generated every second. This data is generated from various sources such as social media, sensor data and transactions (business system). Due to data's enormous in size, its generation speed and its varieties, the term called Big Data suits its description. This Big data has come with challenges on how to model them using the relational approach for the purpose of storing and processing. Due to its large volume, high speed and different formats in Big Data, it is becoming difficult to manage it using relational databases. Thus, non-relational database is a better choice for storing and processing of big data. Before data is stored in a database, they have to undergo data modelling. Data modelling is the key to help store and process data in an efficient manner. Thus, this study is focusing on analysing the existing conceptual data models in use and suggest a better approach for the unstructured data.

Data is something humans of different fraternity handle every day either knowingly or unknowingly. In computing, data is defined as known facts that can be recorded and that have implicit meaning (Ramez & Shamkant, 2011). For example, names, telephone numbers, and addresses of the people can be recorded as data in a note book or stored on a hard drive of a personal computer using spreadsheet such as Excel. The data collected and analysed can find its usefulness from home to business, from private to public institutions for different purposes such as marketing, sharing with government agencies or selling to the third parties. Some data are important to our daily lives as they enhance knowledge.

In the business world, data is basically the plain facts and statistics collected when running the affairs of a particular business (Legay, 2016). This data can be used to measure performance in the Organisation's business activities such as customer data as the metrics that relate to customer interaction.

Data can be classified into four different types which includes personal data, transactional data, web data and sensor data (Import.io, 2018). Personal data is

anything that is specific to a particular person. Examples of personal data includes but not limited to your home address, family matter, your email address and other identifying factors on social or business media. Those things that involves action, generate transactional data, swiping on a door using an authorized card or biometric identity in order to have access into the building is an example of transactional data type. While data that is pulled from the internet whether for research study or any other purpose is classified as web data. Sensor data is the type of data which is generated from objects with inbuilt sensors. This type of data is also called internet of things.

With more and more developments in technologies, dataset has also evolved from structured to semi-structured and finally to the unstructured data (Hurwitz, et al., 2018). By structured data, it means data is well defined, can relatively be searched, grouped or sorted in accordance to certain criteria (K.V & M, 2015). In semi-structured data, some data are predefined while in unstructured ones, predefined format is not obvious.

The varieties of data (video, images, sounds and text) are stored in computer as binary values using patterns of two digits of 1 and 0. A bit is the smallest unit of data, and represents just a single value. A byte is eight binary digits called bits. Common computer storage and memory is measured in megabytes and gigabytes which is equivalent to 1,000,000 bytes and 1,000,000,000 bytes respectively. Margaret Rouse stated that, “the units of data measurement continue to grow as the amount of data collected and stored grows. The relatively new term "brontobyte," for example, is data measurement size that is equal to 10 to the 27th power of bytes”, (Margaret, 2017).

Generally, all categories of dataset (structured, unstructured and semi-structured) contributes to Big Data. The definition of Big Data differs from one community to another. What may be considered to be a Big Data at a mine company may not be a Big Data in banking industries. In the recent past, a gigabyte of memory on mobile

handset were impossible and un-heard of experience but in today's world, people are found with gadgets with tens of gigabytes of memory. In this view, Big Data can be considered to be volume, velocity and variety of data which is beyond today's norm (Cameron, 2012). Due to advancement in technology, data has become too big, moving at a faster rate which cannot easily fit in the traditional storage structure. Initially, Big Data was measured by three characteristics of dataset (Volume, Velocity and Variety) popularly known as 3V, however, two other factors such as veracity and value have been added. The Vs of Big Data are lenses to help us understand and view the nature of data and how it can possibly be stored (Edd, 2012).

Volume in Big Data is referred to the amount of data which is generated every fraction of a second which is very large (Jyothi & S, 2015). This large amount data is generated from different sources in different formats, thus, introducing Variety in data. The third "V" depicts the speed at which data generated. The collected data from different sources has some possibility of errors in it. The level of accuracy in the collected data is what is called the Veracity in data. The Value in data is determined by its usefulness in decision making.

The volume of data is exponentially increasing daily as many marketing companies are trying to reach their potential customers. This data which is not only generated from business organisations but also from many other sources such as social media, health sectors, agriculture and others come in different formats are generated at high speed. The use of information and communication technology through web and mobile internet, consumer products and services are being delivered in time and generating Big Data.

Big Data had been used for some years without a name. It was first called Big Data by Roger Mougallas in 2005, which he referred to as a large set of data that was impossible to manage and process using the traditional business intelligence tools available (Keith, 2017).

Sources of data, particularly for Big Data, is on the increase. For example, Big Data is made available through data generated in sensors, social media, business transactions and internet of things device. These sensors are used almost in every aspect of our lives. In agriculture, the use of integrated filed system (IFS) provides us with another source of Big Data as sensors are used to collect data about weed, soil moisture, weather and animal movement (Kelly & Irena, 2016). This helps farmers minimise risk and provide better platform for decision making.

Facebook and twitter are some examples of social media where data is generated in big quantity at a very fast rate. Every 60 seconds, there are over 3,000,000 likes and over 500,000 comments on [Facebook](#) and over 8000 [tweets](#) every second (Admin, 2018) (LiveStats, 2018).

Data is always generated in business transactions. Online financial transitions, goods and services purchased using debit cards or mobile money are some of the examples of source of data. In Zambia, we are seeing the introduction Toll Plazas and speed cameras on the trunk road. These devices are other areas where data is generated, thus, contributing to Big Data.

In the internet world, we are seeing more and more personalized data through search engines and social network sites through specific advertisements to individuals or a micro-segment (Abdul & Tariq, 2015). The increase in the use of internet cookies which tracks user behavior and tailor advertisement content on individual behavior is another source of Big Data.

Health sector is another major source of Big Data though the application of analytic tools and electronic communication systems to retrieve patients' past records when they seek medical attention. Historical data for patients obtained from various diagnosis at different time can be helpful to understand the symptom pattern and prescription trends (Remya, 2013).

Information is very important to the business of an organisation, thus, data has to be managed properly. The collection of data is referred to as database while Database Management System (DBMS) is the collection of interrelated data and a set of programs to access that data (Abraham, et al., 2011). The purpose of DBMS is to allow the organisation's storage and retrieval of data from their computers. In the 1950s, magnetic tapes were developed for data storage and was replaced in the late 1960s with hard disks which changed the way data was being processed (Abraham, et al., 2011). Data processing in the magnetic tapes involved reading of data from one or more tapes and then write data to a new tape.

The true evolution on data storage started in 1970 when a mathematician by the name of Edgar F. Codd published "a paper showing how information stored in large databases could be accessed without knowing how the information was structured or where it resided in the database" (IBM100, 2016). Codd called his paper, "A Relational Model of Data for Large Shared Data Banks." This relational database was introduced to separate conceptual representation of data from its physical data storage in the computer. To access data in a database, one does not have to know or understand how information is structured and where it is stored.

In relational databases, data and their relation are organised in tables. A table is collection of records and each record contains some fields (Suri & Sharma, 2011). The most common types of relational databases include Oracle, Microsoft SQL server and an open source MySQL. Schema in relational databases are rigid, they cannot easily be changed once the database is implemented, thus, schemas have to be predefined. With the development of web applications which handles millions of users at the same time, relational databases are facing difficulties on how to handle the huge amount data (Cornelia, et al., 2015). The limitations in relational databases has led to the development of non-relational databases. The non-relational is commonly known as NoSQL technology and it is put into four categories such as key-value stores, graph stores, column family stores and document stores NoSQL Databases (Robert, 2015).

The way data is stored in relational database is completely different from the way non-relational databases store it. In non-relational databases, defining schema before inserting data is not a requirement, instead, they use identification keys (Robert, 2015).

In order to handle dataset in Big Data, there was need to abandon ACID (Atomicity Consistency, Isolation and Durability) which is the key principal in relational databases (Abert & Sherif, 2017). Non-relational database follows the principal of BASE. BASE stands for Basically Available, Soft State, Eventual consistency. **Basically Available** implies that the system guarantee availability, **Soft State** in NoSQL databases means that information may expire if not refreshed and **Eventual consistency** indicates that when nodes are added data is replicated (Ganesh, 2015). Scalability and complexity are the main challenges in most relational databases. To attain scalability in relational database, it has to be distributed on multiple servers and powerful servers are very expensive unlike non-relational databases which are designed to expand horizontally by adding machines into your pool of resources (Couchbase, 2017). In relational database, data has to fit into tables, otherwise, a complex database structure has to be designed.

For data to be stored in a database, data models have to be created. The process of creating data model is what is termed as data modelling. The data model gives guidance in the design and implementation of a database. The data model is said to be the conceptual representation of data objects and how they relate to each other. Data models are employed in the design and implementation of database so that a clear picture of the database is provided (Watt, n.d.). This helps to eliminate errors. It is during data modelling where all required data objects are captured. If data is wrongly captured or important data objects are left out, the output cannot give the desired results.

There are at least three types of data model that have been employed in database design. These levels in data modelling approach includes conceptual, logical and physical data modelling. During conceptual modelling, which is the high level usually created by business shareholders and data architects, the business concepts and rules are defined and organised. Conceptual models define the real world data objects and it is independent of the database technology (Bansari & Hetal, 2017). Logical data model is the refinement of the conceptual model where entities' attributes, foreign keys and unique identifier are identified (André, et al., 2015). Unlike conceptual and logical models, physical data models are dependent on a specific database type. During physical modelling, data types are defined in readiness for database implementation.

The purpose of conceptual data models is to identify and define concepts, and how they relate to each other in a particular business circle. By concept we mean key ideas which can be either be basic or critical concepts (Steve, 2015). Basic concepts are those terms which are used many times in a business while critical concepts are terms which could make a business different or non-existence without them.

The defined concepts in conceptual data models are analysed to provide detailed business solution in logical data models. The attributes of the concepts are identified during logical data modelling. It is in logical data model where the attributes and relationships between business objects are identified. Further, the purpose of logical data model is to create business rules required to deliver the application.

The business solution defined on a logical data model is transferred to technical solution through physical data modelling process (Smith, 2016). Unlike at logical data model stage, the technical business solution on a physical data model are dependent on database technology. Thus, the purpose of physical data model is to create technical solution from business solution on a logical data model.

In this research, the focus was on conceptual data modelling and how it can be employed in big data. Stewart Robinson defines conceptual modelling as a “concise

and precise consolidation of all goal-relevant structural and behavioral features of the system under investigation presented in a predefined format” (Stewart, et al., 2015). Conceptual data model describes the physical and social aspect of the World around us. The developed conceptual model facilitates the understanding of the shape of the system to be modeled and provides a platform for acceptable solution to the problem. Conceptual modelling provides a broader understanding of the business process, application requirements and even the whole industry in a single piece of paper (Steve, 2015).

Analysing the three levels of data models explained above, a gap between business stakeholders and IT leadership is seen, in that, business stakeholders view things at conceptual data model while the technical personnel tend to focus on physical data model. This gap which has been ignored in relational database is a critical challenge in today’s Big Data environment. In this Big Data environment, immense data is generated at faster rates in different varieties. With the right database technology known as NoSQL, a data model where both business solution and technical solution can be merged. The solution data model can be referred to as enterprise or business data model.

1.2 Statement of the problem

There is a tremendous need to look at the technique of conceptual data modelling in modelling datasets which is too large, generated at high speed and of diversity in nature. Relational database approach in data modelling is more appropriate to datasets which is well organised, structured and does not often change. The advent of Big Data, the traditional approach in data modelling cannot cope, as data is generated in large volumes at high speed from different sources. The rate at which data is being generated from different sources, the current databases have limited capability to handle big volume of data in this era of digital world.

As creativity in database technology continues to advance and the need to store limitless volume, variety and velocity of data remains in demand. Further, there is need for cost effective and technology-based solution platform which is able to handle

data in the Big Data environment. Therefore, a study on conceptual data modelling and how it can be applied to Big Data is essential.

1.3 Objectives

- i. The main objective was to determine conceptual data modelling approach for non-relational database in the Big Data environment.
- ii. To investigate dataset types in the era of digital world (era of Big Data).
- iii. To explore the deployment of relational and non-relational databases in Zambia.

1.4 Research question

- i) What are the types of datasets handled in the local organisations of Zambia?
- ii) How do data structures in small data lead to the development of Big Data and describe their characteristics?
- iii) What are the types of databases deployed in Zambia?
- iv) What roles does conceptual data models play in building a system model?
- v) What are some of the developments in conceptual data modelling to support NoSQL technology and establish a possible approach for its conceptual modelling?

1.5 Purpose of the study

The purpose of this study is to evaluate the conceptual data modelling used in the implementation of relational databases so as to propose an approach to be used in designing non-relational database.

1.6 Thesis disposition

The work done in this thesis is organised into five chapters. Chapter 1 is the introduction to the research where a brief overview of the work in this thesis is given. The problem statement, objectives and the purpose of the study of this thesis, are also given. This chapter concludes by the giving an outline of the thesis.

Chapter 2 provides an overview of literature review about this dissertation. Literature review covers conceptual data modelling for relational and non-relational databases, datasets types leading Big Data and Databases evolution.

Chapter 3 presents the methodology of the study which is a mixed approach on how the survey was carried out. Methodology shows how data was collected and analysed. Chapter 4 presents the findings of the study. The findings as guided by the objectives of the study. The findings were further discussed in the next chapter accordingly. The final chapter is Chapter 5 where the discussions on the findings, conclusions and recommendations, and summary of the study were presented. It is in this chapter that the proposed NoSQL conceptual model approach is discussed.

CHAPTER 2: LITERATURE REVIEW

2.1 Dataset

Data are facts representing an object or event that can be recorded and stored in some form, which possess significance and are useful to the users. Conclusions and concepts can be drawn from these facts. Dataset is the collection of values either as numbers or strings (Wickham, 2014). These numbers and strings are known as data. When the collected data is arranged, presented and interpreted it is referred to as dataset (Renear, et al., 2018). As with clothes, cotton wools are the raw materials from which they are produced so is with dataset, data is its raw materials. When datasets are processed, interpreted, organised, structured or presented so as to make them meaningful or useful, they are called information. The general sources of data include primary and secondary source. Data collected direct by a researcher through questionnaire, interview or observation is referred to as a primary source whereas secondary source, is data obtained from the third part who collected from other sources such as data in a scientific journal or books.

Dataset comes in different forms which may be classified as small data and big data. It was not common to refer to data as either “small” or “big” before 2008 (Kitchin & Lauriault, 2015) (Kitchin & Lauriault, 2015) instead data was just called data. The data which may be termed as “small data” today were all called data yesterday, regardless of their volume. In the past, classifying data as big or small was not there. This was because data was usually measured in terms of volume. What is now is called “big data” is beyond the factor of volume.

2.1.1 Small Data Vs Big Data

Small data is defined as “records of our activities that are stored as we interact with the world around us” (Kitchin & Lauriault, 2015). Example of such data are records of tools at home, lecture notes stored on the laptop by the professor, a farmer’s records for the farm activities, etc. Such dataset which has few records in it are called small data (Rurael, et al., 2013). Big data is data whose volume, variety and velocity are bigger than today’s norm. Big data is not just small data which has grown, but there

are many other aspects that can help to differentiate one from the other. The voluminous size of data does not make it big data, in the sense that the data still lacks velocity, variety and other characteristics of Big Data.

Small data is focused in specific cases and tailored to achieved a particular objective. This type of data can be such data from non-governmental organisations, public institution, academia and private entities where access to it is limited to the public (Rob & Lauriault, 2014). Access to this small data may only be available for a fee. There are many ways of comparing big data to small data. Below is their comparison in relation to the characteristics of Big Data, commonly known as 5V:

Table 2. 1: Small Data Verses Big Data in terms of 5V

SMALL DATA	BIG DATA
VOLUME	
Range up to tens or hundreds of GB. Some case few TBs (1 TB=1000 GB)	More than few Terabytes (TB)
VELOCITY	
Data accumulation is slow, it is controlled and the data flow is steady	Enormous data can accumulate within very short periods of time and it arrive at a very fast speed
VARIETY	
Structured data in tabular format with fixed schema and sometimes in semi-structured format	High variety data sets which include tabular data, text files, images, video, audio, etc.
VERACITY	
Contains less uncertainty as data collection is controlled	Quality of data not guaranteed, thus, data validation is required before processing
VALUE	
Data is worth for business intelligence, analysis and reporting	Complex to determine its worth, thus, good for prediction and requires brainstorming to identify its value

The volume and velocity of data in small data is limited. In small data, volume goes up to a few terabytes while in Big Data its volume is above a terabyte. Data is generated, processed and analysed at a very high speed in Big Data as compared to small data where its speed is controlled.

The structure of data in small data is well structured with fixed schema and in some cases presented in tabular form. In Big Data, data is high varieties which includes tabular data, text, images, videos, audios, etc.

The quality and worthiness of data is not guaranteed in Big Data as it requires to be validated and brainstorming to identify its value. In small data, data is well structured, thus, veracity and value is guaranteed. This is because data collection is controlled and it is worth for business intelligence, analysis and reporting.

2.1.2 Big Data

In the last decade or so, a gigabyte of memory on our handset was not possible; unheard-of experience but in this era of Big Data we walk around with tens of gigabytes of data on our memory sticks, smart phones and many other portable electronic gadgets. Thus, what was big yesterday may not necessarily be big today.

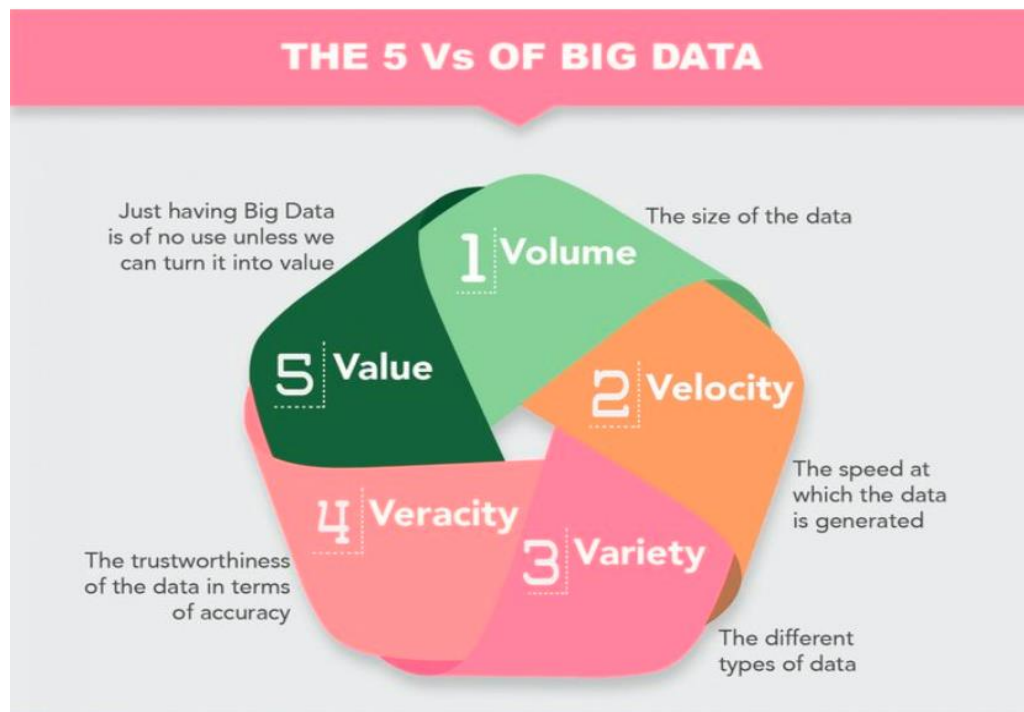


Figure 2. 1 The 5Vs of Big Data (*BigData*, 2016)

In this view, Big Data is data whose volume, velocity and variety is bigger than today's norm (Cameron, 2012). This definition of Big Data is adopted from data

characteristics which was initially identified by Dong Laney in 2001 (Erl, et al., 2016). Laney considered dataset to be Big Data when it possesses the three data characteristics commonly known as 3V such as Volume, Velocity and Variety. However, today at least two other characteristics of Big Data have been identified and these are Veracity and Value.

Big Data is steadily changing the way we think about data. Big Data is a big thing and to understand it, it is now described in terms of volume, velocity, variety, veracity and value as shown on the Figure 2.1 below. Further, Big Data is defined as: “data that exceeds the processing capacity of traditional database system as it is too big, moves too fast and does not fit the structure of relational database management system architectures” (MacCreary & Kelly, 2014).

i. Volume

This refers to the quantity of data that is generated, analysed and manipulated (Erl, et al., 2016). It is also referred to dataset whose size is beyond which the traditional database software tools can capture, store, manage and analyse (Jackson, et al., 2016). To determine the value of the data, the size of data plays very crucial role. Thus, volume of data plays an important role for a particular dataset to be considered as a Big Data or not. In this line, '**Volume**' is one of the characteristics which needs to be considered while dealing with 'Big Data'. The current amount of data can be staggering. For instance, every second there are 8,288 tweets sent, 883 Instagram photos are uploaded, 1,460 Tumblr posts, 3,502 Skype calls, 65,114GB of Internet traffic, 71,068 Google searches, 76, 545 YouTube videos viewed and 2,750,992 Emails sent (Anon., 2018).

ii. Velocity

Velocity refers to the speed at which data is being generated or collected (Hammad, et al., 2015). This characteristic of Big Data is not only about the speed at which data

is been generated and collected but also about the rate of change of the generated data. Velocity as a characteristic of Big Data is concerned with the speed at which data flows in from sources such as business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. This flow of data is usually massive and continuous. Important activities require instant action, thus, fast processing is needed to maximize efficiency. An example of fraud detection, which is time-sensitive process, Big Data flows must be analysed and used as they stream into the organisations in order to maximize the value of the information (Hadi, et al., 2015).

iii. Variety

Typical databases store datasets in structured format. This is not so when it comes to Big Data storage platforms. In Big Data technologies, dataset comes in different formats such as structured, semi-structured and unstructured data. In past data collected was well structured and of the same format, today it comes in array of different forms such as text, audio, video, social media posts and many more other unstructured or semi-structured data format (Hammad, et al., 2015). For example, email messages may contain typed text and attached messages in the form of voice, video or text. An organisation might require to sift through thousands of email messages from different sources. No one single message is going to be exact same as another, though they all may have sender address and human written message and possibly an attachment.

iv. Veracity

In this world of fake news, another important characteristic of Big Data is Veracity of the information collected before it can stored. Understanding and eliminating flaws in dataset, especially about their reliability, accuracy and representativeness is very important in conceptual data modelling of Big Data (Tole, 2013) (Hammad, et al., 2015). This characteristic of Big Data is concerned with the reliability of the source and content of data and how meaningful it is to the user. It is one of the characteristics of Big Data which inversely to Volume, Velocity and Variety in that the higher is the value of the 3V the less is the veracity of the data. For example,

consider the dataset about the cause of a particular pandemic outbreak, many questions have to be asked. What is the source? Who reported? Was the report edited or modified by someone else? To address these and many other questions it helps to determine the veracity of the information in this era of Big Data. Thus, working processes have to be done to keep your data clean and keep ‘dirty data’ away from accumulating in the system.

v. Value

This is equally an important characteristic in the Big Data application. Value allude to obtaining, analysing valuable information from large dataset (Koseleva & Ropaite, 2017). This is about the quality of data to be stored, otherwise there is no need of storing data which is of no value and that cannot be managed properly (Tole, 2013). If business value is not derived from data, other characteristics of Big Data becomes useless. When the value in dataset is properly gotten, there a lot of benefits an organisation would enjoy. The customers will be understood better and served them appropriately, thus, retaining them. The retained customers are the best marketer for your goods or services. Other benefits include improved business performance and targeting potential clients or customers accordingly. The value of dataset in an organisation determines the value of their business outcomes.

2.1.3 Types of Dataset in Big Data

Data generated from various sources comes in different formats with differing sizes at high rate. This complexity in the nature of data, makes it difficult for traditional conceptual data modelling to handle. Thus, Ahmed and his colleagues opted to define big data in terms of complexity. Ahmed refers Big Data to a large growing datasets which appears in heterogeneous formats such as structured, semi-structured and unstructured data (Oussous, et al., 2017). The different formats are (Cameron, 2012) classified into three types as:

- i. **Structured data.** Type of data usually used in a traditional database where fields of records are well defined, that can relatively be searched, grouped or

sorted in accordance to some certain criteria (K.V & M, 2015). Data is well organised that come either as in tables or some other ways and can be worked on without much difficulties. Data stored on a spreadsheet and in relational databases are some of the examples of structured data.

- ii. **Unstructured data.** It is the dataset type where predefined format is not obvious (Frappolli, 2018). The form for unstructured data is not known. This gives a system designer freedom to store any kind of data without defining or creating schema. The challenge with unstructured data is that, it is difficult to query, thus, organisations finds it difficult to get value out of it.
- iii. **Semi-structured data.** This is referred to as the combination of structured and unstructured data type where some aspect of data may be defined while others are not in a predefined format. This type of data lies just in between structured and unstructured data.

2.1.4 Sources of Big Data

Sources of data is categorized into three basic dimensions which includes machine, human interaction and data processing. Typically, the first source is the widest source of data. This data in most cases are generated by sensors that are connected to electronic devices (Joshi, 2017). These connected electronic devices are the basis of internet of things (IoT), which is one of the major contributor to Big Data. These devices ranges from mobile phones collecting traffic status using application wares to various mounted sensor on roads or entry points to satellite monitoring machines. Data is generated from interconnected devises such as speed cameras, home appliances, satellites and many other devices is also referred to as internet of things (Meola, 2018).

In today's technology, data generation through human interaction is increasing exponentially. This is as a result of the increase in number of social media sites such as Facebook, Instagram, YouTube, twitter and many others. Social media, being a self-broadcast and is not limited by physical border lines it is providing business

analysts with valuable insights on consumer preferences and changing trends. It is proven to be the fastest way for businesses to get their services and products to their targeted audience. On the other hand, Data Analyst/scientist are taking advantage of social media users' posts, tags, comments, upload and download to penetrate their marketing advancement (Nidhi, 2018).

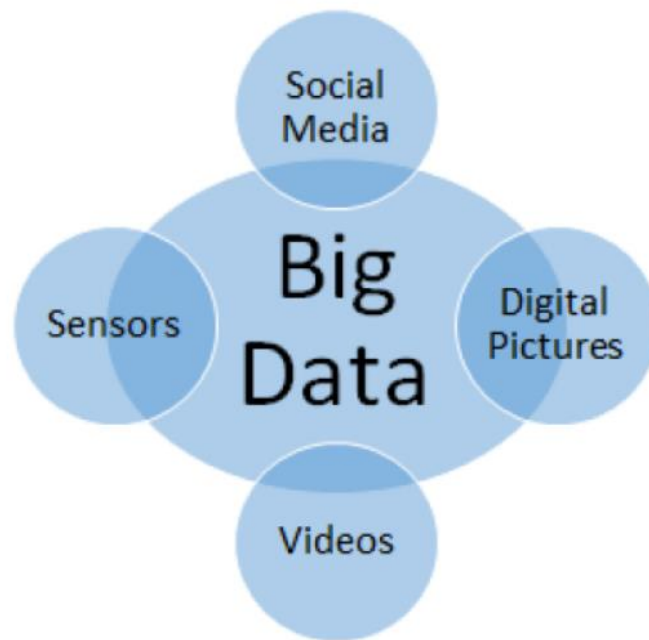


Figure 2. 2 Big Data Sources (*Hammad, et al., 2015*)

Further, another source of Big Data is through data processing. More people are using point of sales electronic gadgets when doing financial transaction. This generates data which is transmitted in real-time, creating another source of Big Data. Further, data is generated when one withdraws or deposit money using the ATM, this data is processed in real-time and information is sent to the register for account status update. Big Data has gained much attention in the past few years which has led to this generation of Big Data environment. In this era of technological advancement, almost all aspect of our lives are dependent on technological gadgets. These aspects of human life include but not limited to personal, social and professional (Abdul & Tariq, 2015).

Each of the dimension of human dimension deals with one form of data or another. This has led to the coming of the Big Data concept as data is being generated every second.

i. Social Medium as a source of Big Data

The increase and availability of internet on various portable devices has made social media to become an integral part of human lifestyle. Social media not only that it connects people, but also a platform for business houses to meet their targeted audience. With Big Data, business information is personalized so that the products and services arrives at their audience in a personal way in accordance to the customer's choices and likes (Nidhi, 2018). For instance, google search tends to suggest some topics based on the previous frequently searched.

Marketing analyst are able to predict their consumer's future requirements based on their behavior on social media. This is called predictive analysis; it is gaining popularity with the marketers and big data has a huge role to play (Warner, 2018). The use of social media does not only provide platform for marketing analyst to reach their audience, but also provide consumers with an opportunity to make an informed decision about products and services. Social media is constantly producing data for consumers, thus, enable them to make decision about what and where to buy their products and services (Mallon, 2018).

ii. Sensors as a source of Big Data

The emerging of Internet of Things (IoT) through the interconnection of objects such as PCs, tablets, sensors, smartphones, wearable devices and household appliances has proven to be the main source of Big Data (Ahmed, et al., 2017). Internet of Things is taking shape; we are seeing intelligent devices, wearables, and connected items that aim to make our lives easier (Ebner, 2015). The most common wearable devices with inbuilt sensors are watches such as Apple watch and Samsung Simband. Most of these watches and some phones have emotional sensors to measure your heartbeat, body

temperature and many other things. The measured parameter can be transmitted and saved on a database for future reference.

Robots and drones require good data to navigate and operate; the use of sensors makes them safe to work with humans (Demaitre, 2018). This data generated from sensors produces Big Data.

iii. Digital picture as a source of Big Data

Unlike film pictures, digital pictures are captured and stored using bits of 0's and 1's. This makes digital photos being easily transferred and stored. These types of photos which can be captured using various handheld devices which includes the phone are uploaded on social media such as Facebook, Twitter or Instagram (Woodford, 2018). This adds to Big Data.

iv. Videos as a source of Big Data

YouTube is the most popular video online sharing platform and is one of the major contributor of Big Data (Cheng, et al., 2014). Another area where videos are having impact on Big Data is through video advertising which comes with text, audio and video (Kitoyi, 2018). This may occur before, during or after the streaming of the video on the internet. This digital advertising is adding massive volume of data to the internet. Digital videos are slowly overtaking TV adverts.

In this era of digital age, location of the patient is not a factor anymore. Through the use of Big Data analytic tools and employing of electronic communication systems (Chinchmalatpure A & Dhore, 2016), information about individual patients can be retrieved at the click of the button. This use of technology to generate, analyse, store and manipulate the patient's information is another source of Big Data and data in general. In Big Data, historical data for patients which would have been obtained from various diagnosis at different occasion can be used to find out symptom pattern and prescription trends (Remya, 2013).

The use of integrated field system (IFS) in agriculture provides us with another source of Big Data. IFS tools collect information about soil condition, weather and variety of weeds which provides farmers with information to help them minimise risk and enhance better decision making (Kelly & Irena, 2016).

2.2 Evolution of Databases

Database is the collection of data with activities of one or more related to an organisation. A software designed to manage and utilize the collection of data is called Database Management System (DBMS).

To get to where it is today, databases have gone through various evolution. The year 1958 is an important year for data storage. In this year IBM Corporations of United States, based in New York produced a system file that could store and retrieve data (Olofson, 2011). IBM stands for International Business Machine. This system was called Format File System (FFS) and was very instrumental to the introduction of automated database management. FFS is known to be the first generation of database management system.

Four years later, another team of IBM developers came up with a system which improved the facilities of FFS by enabling users to collect data and easily code reports on a recurring basis (Olofson, 2011). This system was called Generalized Information Retrieval and Listing System (GIRLS) for IBM 7090, capable of handling a collection of data with many different records unlike FFS.

Charles W. Bachman (Berg, et al., 2013) (Olofson, 2011), General Electric consultant using network orientation designed a DBMS called Integrates Data Store (IDS) which was able to share complex data under schematic control. The IDS was developed in 1964.

A dawn of relational database from navigational was based on mathematical theory by Edwin Frank Codd. He authored a paper entitled “A Relational Model of Data for

Large Shared Data Banks”. It was his concept of relational that tuples, attributes and relations were translated into rows, columns and tables respectively (Olofson, 2011). Another IBM engineer named Donald Chamberlin led a team that developed Structured English Query Language (SEQUEL) in 1974 (Olofson, 2011). Three years later, Chamberlin with his colleague Robert Boyce changed SEQUEL to SQL (Structured Query Language). This relational database dominated the field of database management system for a period of time. Most of the relational database systems are not only complex and use structured query language but also rigid with limited ability to handle unstructured data (Foote, 2018). This led to the development of a non-relational database called NoSQL. NoSQL means Not Only SQL.

The development of NoSQL was partially due to the need to process unstructured data and requirement to handle web data. NoSQL model is a quicker approach to organise and process large and variety amount of data. It is not only able to handle structured but also unstructured data, making a better choice in this digital world.

2.2.1 Database Types Overview

Database is the collection of organised data. Despite that the term “Database” came into use in the late 1960s, data collection and its organisation has been part of human development (Guy, 2016). As human civilization develops alongside technology, they have adopted different means of collecting and organizing data to suit their data utilization.

From the time databases were introduced, we can identify five types of databases which will be discussed under this section. The types of databases to be discussed briefly, includes Hierarchical, Network, Object-oriented, Relational and Non-Relational (NoSQL) databases.

i) Hierarchical Database

In Hierarchical Database, data models are in a tree-like structure where records contain information about their groups of parent/child relationships (Team, 2017).

This type of database is usually used on mainframe computers. Hierarchical Database (DBMS) is the oldest technical method of organizing and storing data (Singh, 2015). Its structure is similar to family tree of parent/children relationship of 1: n. just like a normal family, a parent can have one or more children and a child can only have one father/mother, so it is in hierarchical database a child record only has a link to one parent record. The IBM Information Management System (IMS) and the RDM Mobile are examples of a hierarchical database model.

ii) Network Database

Network database is similar to hierarchical database in that both contain hierarchical structure but the Network database is more like a cobweb or interconnected network of records. Network databases not only do they look like hierarchical databases but they were also commonly used in the mainframe computing (Guy, 2016). Unlike in hierarchical database, children records are called members while parents are owners in a network database model (Singh, 2015). This type database model is more flexible compared to hierarchical database mode as the member (or child) can have more than one parent (owner). Some of the common databases that use network database model includes Integrated Data Store (IDS) and Integrated Database Management System (IDMS).

iii) Object-Oriented Database

Object-Oriented Database, also called Object Database Management System was designed to handle programming language such as java and C+. The object-oriented structure is a type which is common in multimedia web-based applications and they are capable of handling voice, text and pictures types of data (Singh, 2015). Object database stores object rather than storing data. The objects consist of attributes and methods. By attribute, we mean data which defines an object whereas the method is referred to the behavior or function of an object. Examples of this type of database are ConceptBase, ObjectDB and WakandaDB.

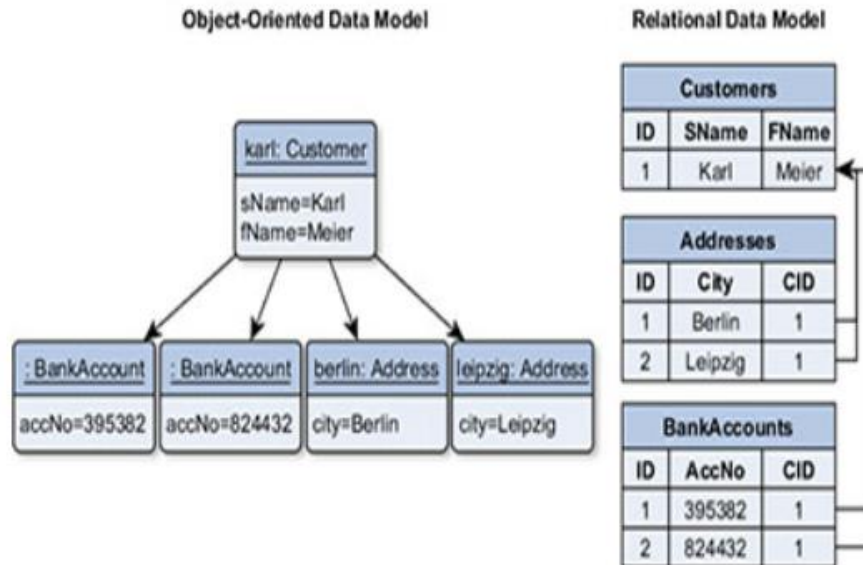


Figure 2. 3 Object-Oriented Data Model Vs Relational Data Model

iv) Relational Database

In relational databases, data is stored according to a schema that shows data as tables with rows and columns (Hammink, 2018). Basically, relational database is a collection of tables and each table has a schema representing attributes and data type. The RDBMS offers the CRUD functionalities on data by means of Structured Query Language (SQL). CRUD stand for Create, Read, Update and Delete.

Relational databases work on the principle that each table has a key field that uniquely identifies each row, and that these key fields can be used to connect one table of data to another (Singh, 2015). For the data contained in tables to be reliable and accurate, relational database management system uses a number of constraints.

They work well on structured data, however, on semi or unstructured data they fail to handle due to the schema and constraints employed in relational databases.

While there are many relational databases, over time these have become the most popular (Hammink, 2018):

- Oracle Database, also called Oracle RDBMS or simply as Oracle is a multi-model database management system produced and marketed by Oracle Corporation.
- MySQL is an open source RDBMS based on Structured Query Language (SQL). MySQL runs on virtually all platforms, including Linux, UNIX, and Windows.
- Microsoft SQL Server is an RDBMS, that supports a wide variety of transaction processing, business intelligence and analytics applications in corporate IT environments.
- PostgreSQL, often simply Postgres, is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards compliance.
- DB2 is an RDBMS designed to store, analyse and retrieve data efficiently.

v) Non-Relational Database

Non-relational database also known as NoSQL database is a type of database which does not use relational models of traditional relational database management system. NoSQL also known as Not Only SQL are schema free type of database.

NoSQL database was promoted to handle web-based applications which is increasing exponentially and other semi and unstructured data. As indicated above, the relational databases are unable to handle large amount of data in this era of digital world.

There are different types of non-relational database which may fall under of any of the following classified NoSQL database; key-value stores, graph stores, column family stores and document stores.

2.2.2 Relational verses Non-Relational Database

Relational and non-relational are the two different types of databases; relational database is based on the concept of table which was laid by Edger Codd (Pore & Pawar, 2015). The advent of web 2.0 in the 21st century such as Facebook, google, etc. has triggered the need for NoSQL database application (Sareen & Kumar, 2017).

To compare relational to non-relational databases, we will start by looking at the basics from both databases.

Relational Databases. This type of database was developed in 1970s and possesses features with support ACID properties to ensure that data is reliably processed. ACID is an acronym for Atomicity, Consistency, Isolation and Durability. Below is brief description of ACID property of relational databases (Ramachandran, 2014):

Atomicity: Each transaction is unique and make sure that if one logical part of a transaction fails everything is rolled backed so that data are unchanged.

Consistency: All data written to the database are subject to the rules defined (constraints, triggers, etc.)

Isolation: Changes made in a transaction are not visible to other transactions until they are committed.

Durability: Changes committed in a transaction are stored and available in the database even if there is power failure or the database goes offline suddenly.

Non-relational databases. The rapid and unpresented growth of unstructured data led to non-relational database concept. Unlike relational databases, non-relational is a schema free and does not require JOINS between objects (Pore & Pawar, 2015). As seen above, relational database supports ACID property. However, non-relational database supports CAP (Consistency, Availability and Partition-tolerance).

According to Brewer's theorem, also known as CAP theorem, it is not possible for the distributed computer system to simultaneously guarantee all the three component of CAP theorem.

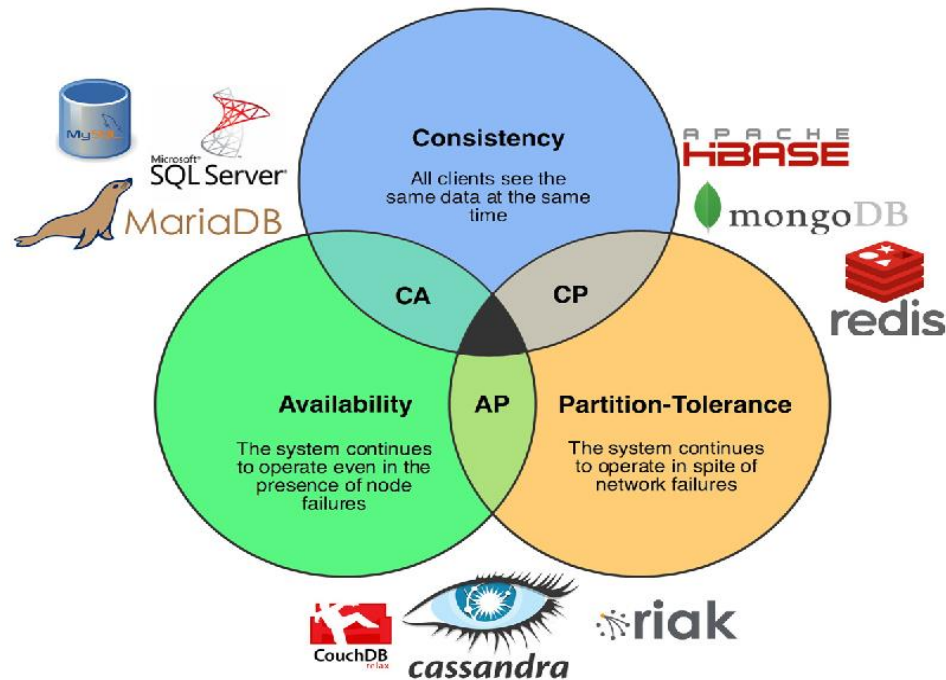


Figure 2. 4 CAP Theorem with databases that chooses AP, CA and CP

The CAP Theorem states that in the distributed computer system it is impossible to provide the three properties of Consistency, Availability and Partition-tolerance at a time (João, et al., 2015). The Cap Theorem is explained as follows:

- **Consistency:** this condition states that when performing an operation all nodes have the same information at the same time (Nazrul, 2018). For example, a 'read' operation returns the value of the most recent 'write' operation causing all nodes to return the same data. This means that regardless of which node in the Database system have an operation, every other node will have the same updates immediate after committed.
- **Availability:** the system provides answers for all requests it receives, even if one or more nodes are down.
- **Partition-Tolerance:** the system still Works even though it has been divided by a network failure.

The differences between relational databases and non-relational databases is summarised in the table below:

Table 2. 2 Difference between relational and non-relational databases
(Kumawat & Pavate, 2016)

	SQL	NoSQL
1	Relational Databases(RDBMS)	Non-relational or distributed database
2	Vertically scalable (one node adding multiple cores or memory)	Horizontally scalable (partition of data i.e. each node contains only part of data)
3	Table based databases	Document based, key-value pairs, graph databases or wide-column stores
4	Supports predefined schema	Supports dynamic schema
5	SQL (structured query language) for defining and manipulating the data	Uses unstructured Query Language
6	SQL databases maintains on ACID properties (Atomicity, Consistency, Isolation and Durability)	NoSQL database follows the Brewers CAP theorem/BASE properties

2.2.3 Relational Database

A relational database stores information in tables. Each entity, a ‘real’ thing you are trying to keep track of, is represented by a table (Shahid, 2018). For example, in an Agribusiness database, there would be a table for products and a table for customers (among many others). Each row of a table represents one instance of that entity- a value of customer ID, or his name. Each column of the table is an attribute — type, name.

The database is called “relational” because the tables are related to each other. For example, each worker is related to an animal because a worker’s responsibility may be to take care of animals. Below is an example of a table in a relational database:

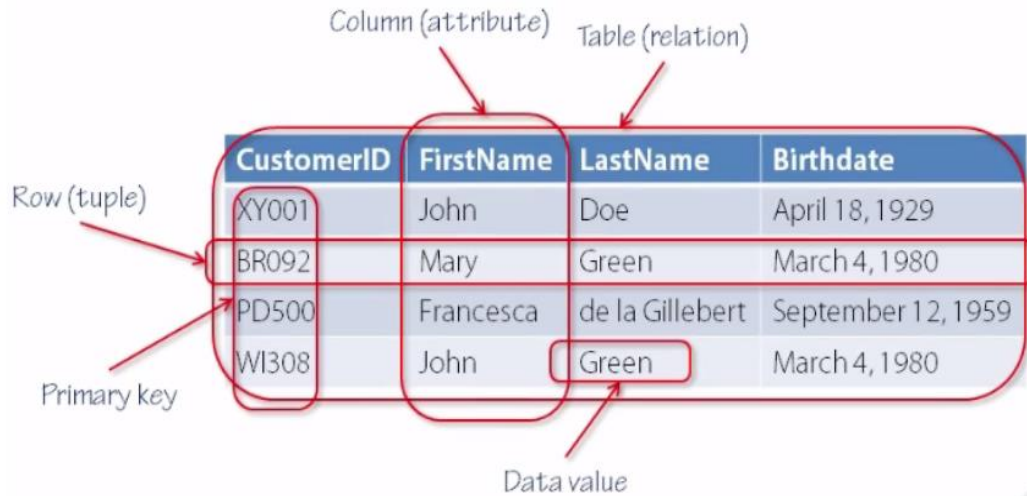


Figure 2. 5 Relational Database Table (Shahid, 2018)

Relational database uses Structured Query Language (SQL) to manipulate database tables and retrieving information from them. The type of SQL used in most relational database is called Declarative Language; that means that you explain to the database what you want. All Relational Databases use Common Sub Language Keywords:

- Data Manipulation Language (SELECT, INSERT, UPDATE, DELETE)
- Data Definition Language (CREATE, ALTER, DROP)
- Data Control Language (GRANT, DENY, REVOKE)

i. Data Modelling in Relational Databases

To create database in relational database; various model abstracts are made. These models include conceptual, logical and physical data models.

Conceptual Data Model (Conceptual Schema): A conceptual data model is used as a starting point to implement a database and uses the terminology that is common in everyday language of the people in the business. The main purpose for conceptual data model is to capture all information that is relevant for the business.

Logical Data Model: This defines how the business information should be stored in a database. During logical data modelling, schemas are designed for a specific database technology i.e. hierarchical, relational or object-oriented.

Physical Data Model: This model is independent of a particular database technology; it defines the exact implementation of the information in the system. Physical data models are derived from the logical data model.

ii. MySQL Database

The database was initially called mSQL database system connecting tables using fast low-level (ISAM- Indexed Sequential Access Method) routines, however, after testing it was found that mSQL was not fast as expected (Editorial, 2018). This resulted in new SQL interface, API (application programming interface) interface, which was designed to enable third party code. The name MySQL was adopted after co-founder Monty Widenius's daughter, My, while its logo is called Sakila. The winning name was submitted by Ambrose Twebaze, an Open Source software developer from Swaziland, Africa. According to Ambrose, the feminine name Sakila has its roots in SiSwati, the local language of Swaziland (Dybka, 2014).

MySQL is the most common open source relational database management system (RDBMS) based on Structured Query Language (SQL) (Editor, 2013). MySQL runs on different platforms, including Linux, UNIX and Windows; is most often used in web applications and online publishing despite it being used in a wide range of applications (Nishtha, et al., 2012). Client/server model is basis of MySQL database; MySQL server is the core, which handles all of the database instructions (or commands) and is available as a separate program for use in a client-server networked environment. MySQL is secured through the use of access privilege and encrypted password system that enables host-based verification (Nishtha, et al., 2012). MySQL server can be accessed by MySQL clients using several protocols, including TCP/IP sockets on any platform and it also supports a number of client and utility programs, command-line programs and administration tools such as MySQL Workbench (Rouse, 2018).

Relational database systems operate by reading data from disk, bringing it across an I/O (input/output) interconnect, and loading data into memory and into a CPU cache for data processing (Pal, 2016). This type database uses shared memory and it encounters scaling issues when data reaches terabytes (Pal, 2016). Thus, relational databases are not suitable for Big Data. Further, schema in relational databases are not flexible; this adds to its challenges to handle Big Data as data here are comes in different formats and a high speed.

2.2.4 NoSQL database

NoSQL database is non-relational database, originally, was referred to as “non SQL” database (Kirti & Maan, 2015). This new technology in the field of database is now commonly known as “Not Only SQL” and their use in big data and real-time web application has increased (Andlinger, 2013). NoSQL system is put into four categories: graph stores, document stores, wide-column stores and key-value store.

In NoSQL database, storing of data does not follow traditional Relational Database Management System (RDMS) mode (Vaish, 2013). The writer further clarified that NoSQL is neither a database nor a type of database but a generic term for filtering out a group of databases (Vaish, 2013). This set of databases have emerged to tackle the challenges Big Data has come with (Abert & Sherif, 2017). These challenges cannot be fully handled by Relational Database Management System (RDMS) such as MySQL, PostgreSQL, Oracle and SQL Server which is built on ACID property. ACID stands for Atomicity, Consistency, Isolation and Durability of the database.

NoSQL databases have abandoned ACID, which is the key principle in traditional databases, to provide horizontal scalability in order to handle large dataset in this era of Big Data (Abert & Sherif, 2017). This new type of databases come in different architectural patterns or technologies, posing challenges for NoSQL databases users when choosing a particular technology for a certain application (MacCreary & Kelly,

2014). Here, we will discuss the four main and mature types of NoSQL databases technologies. For the purpose of order, the key-value store was the first to be discussed, and then we looked at graph stores, column family stores and finally document stores.

i) Key-value Stores NoSQL Databases

Key-value store NoSQL databases use key/value method to store data. This pair of key/value is used to store data in database as a collection of key/value pairs. It is the most basic type of NoSQL model, but one which is quite efficient and powerful (Manoj, 2014). Data in Key-value store databases are stored in a schema free manner and data is only retrieved through the use of keys.

Manoj said that data in a key-value stores consist of two parts, “a string which represents the keys and the actual data which is to be referred as value, thus, creating a key-value pair (Manoj, 2014). Examples of key-value includes Redis, Oracle NoSQL database, etc.

Redis database: Redis, also referred to as Remote Distribution Server is an open source key-value NoSQL database that store any type of values from text to image and videos. It is a NoSQL database that trades durability to speed, thus, one of the fastest non-relational database.

Operations to create and update data are made using the SET and MSET keywords (Baron, 2015). The syntax follows this patterns: SET <key> <value> or MSET <key1> <value1> <key2> <value2>. MSET keyword is used to specify a multiple set operation provided by Redis for reducing the traffic. In the case of successful adding or updating the data, the Redis server will respond with an “OK” message. Example: “SET bmw 320i” and “MSET bmw 320i audi A4” For data retrieval we have the counterpart keywords GET and MGET using the following syntax: “GET <key>” and “MGET <key1> <key2>”

ii) Graph- Stores NoSQL Databases

According to (Connolly & Begg, 2015) Graph stores is defined as “a structure of nodes, edges, properties to represent and store data in form of graph”. With Graph stores, data is stored in the form of a graph. In Graph Stores NoSQL databases (Nayak, et al., 2013) graphs are made of nodes and edges where nodes act as the object while edges as relationship between objects. This type of NoSQL database is the object-oriented conceptual where every object is directly connected to the adjacent ones (Saxena, et al., 2014). ArangoDB (a multi-model), Neo4j and OrientDB (a multi-model) are some of the common graph store NoSQL databases.

Neo4J database: Neo4J is a graph database which was developed by Neo Technology and was first released in 2007 using Java. It is a high performance graph database which provides object oriented, flexible network structure (Nayak, et al., 2013). Neo4J is not only reliable but also ACID compliant, highly available and scalable. The query language it uses is called CYPHER and Neo4J is used in complex relationships requirement such as social networking.

Column Family Stores NoSQL Databases

Column family stores is also called as column-oriented, wide columnar stores or extensible record stores. Despite having the concept of column by column storage (Manoj, 2014), Column-family stores databases do not store data in tables, instead data is distributed across. In order to offer high scalability in data storage, data in wide columnar stores databases is stored in such a way that there is rapid aggregation with less I/O activities.

In order to decrease the disk access, data is stored in form of whole column compared to relational table which have column and rows with uniform sized field for each record (Pore & Pawar, 2015). For a relational database to search on a particular table all entities in columns should have indexes but is not so with the column family stores

NoSQL database (Connolly & Begg, 2015). Cassandra is one of the most popular column family stores databases.

Apache Cassandra: Apache is an open source NoSQL dataset belonging to the family of column-family stores database. It was developed using Java by Apache Software Foundation. Among others, Cassandra offers feature like high availability, partition tolerance, persistence and high scalability (Vladimir, 2018). It has a dynamic schema, thus, it can be used for a variety of applications which includes but not limited to social networking websites, banking and finance, real time data analytics and online retail.sd

iii) Document Stores NoSQL Databases

Documents stores NoSQL database was primarily designed to store and manage documents (Moniruzzaman & Hossain, 2013). A unique key is used to represent each document in document store NoSQL databases. Robert T. Mason narrated at the conference of (Robert, 2015) that information is not only embedded within a particular document but can also be referenced to another collection of documents. It is this reference concept which is applied in Relational Database Management System (RDBMS) as a foreign key.

In comparing document stores NoSQL database to key-value store database, Ameya Nayak (Nayak, et al., 2013) said that document stores are a bit more complex as they allow the encasing of the key-value pairs in a document. Unlike in RDBMS where every column needs to be defined, document database allows different representation of data (Sadalage & Martin, 2013). Not only that document database accepts empty attributes but also allows for new attributes to be created without the need to be predefined or change the existing document (Sadalage & Martin, 2013). Among many other types of document store database, MongoDB is the most used Database.

MongoDB: MongoDB is also an open source NoSQL document stores written in C++ and uses JavaScript as a query language (Sareen & Kumar, 2017). MongoDB stores data in form of collections and the collection consists of documents. These documents

are in BSON format which is binary form of JSON enabling it support different data types such as Boolean, date, float, etc.

To distribute the collections across multiple nodes MongoDB offers sharding technique (João, et al., 2015). To attain high availability, scalability and performance; Mongo DB was developed. It also leverages on an in-memory computing to provide high performance for both reads and writes operations.

MongoDB, provide a rich set of indexing options to optimize a wide variety of queries, including text indexes, geospatial indexes, compound indexes, sparse indexes, time to live (TTL) indexes, unique indexes, and others (MongoDB, 2018).

2.3 Data Modelling

Data modelling is a process that involves the categorising and defining data so that they become consumable to the organisation's information systems. In the way, data modelling helps to make sense out of the data. Data objects and their relationship to one another are analysed during data modelling. In relational data modelling, there are three levels of data modelling and these are conceptual, logical and physical data models. It is a development process from conceptual to logical to physical schema.

2.3.1 Features of Data Modelling

Table 2. 3 Data Modelling in Relational Database

Features	Conceptual Data Model	Logical Data Model	Physical Data Model
Entity Sets	√	√	X
Entity Relationship	√	√	X
Attributes	√	√	X
Primary Key	X	√	√
Foreign Key	X	√	√
Table Names	X	X	√
Column Names	X	X	√
Column Data Types	X	X	√

Features of Conceptual Data Model:

An Entity is an object or a thing in the real world and is unique from another object or things (Connolly & Begg, 2015). They are identified by enterprise having an independent existence. For example, each animal or crop at the farm is an entity. Entities have properties and values which can uniquely identify them. It is an important feature of conceptual data modelling. It is during conceptual data modelling that entities are identified and grouped according. There are times when a group of entities of same type share the same properties. Such entities are called entity set.

Another feature of conceptual data modelling is entity relationship. Entity relationship is an association between two or more participating entities (Abraham, et al., 2011). The relationship between entities is given a name which describes the purpose of the association. Besides entity and relationship, conceptual data modelling is characterized by attributes.

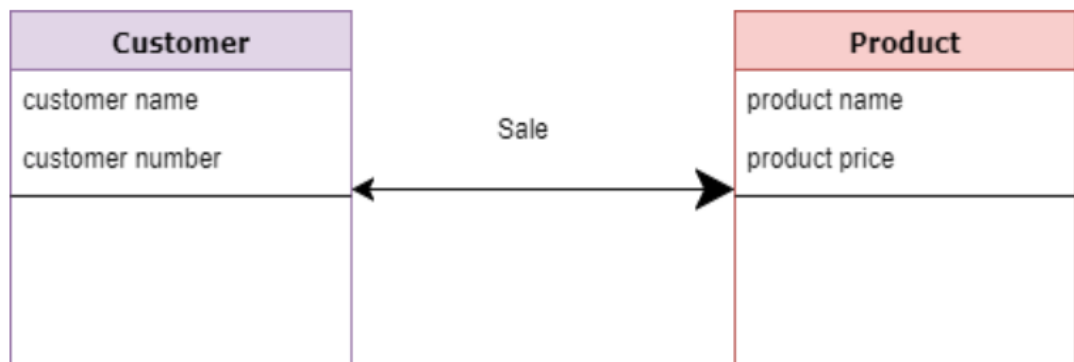


Figure 2. 6 Example of a Conceptual Data Model

Both entities and entity relationship have one or more values and properties. The properties that describe the entity is called attributes. Attributes provide the description of an entity.

Features of Logical Data Model:

A logical data model describes the data in as much detail as possible, without considering how they will be physical implemented in the database. Logical data

model possesses all features of conceptual data modelling mentioned above. As shown on the table 2.1 above, besides entity, relationship and attribute, logical data modelling has primary and foreign keys. These two features of logical data modelling are also found in physical data modelling.

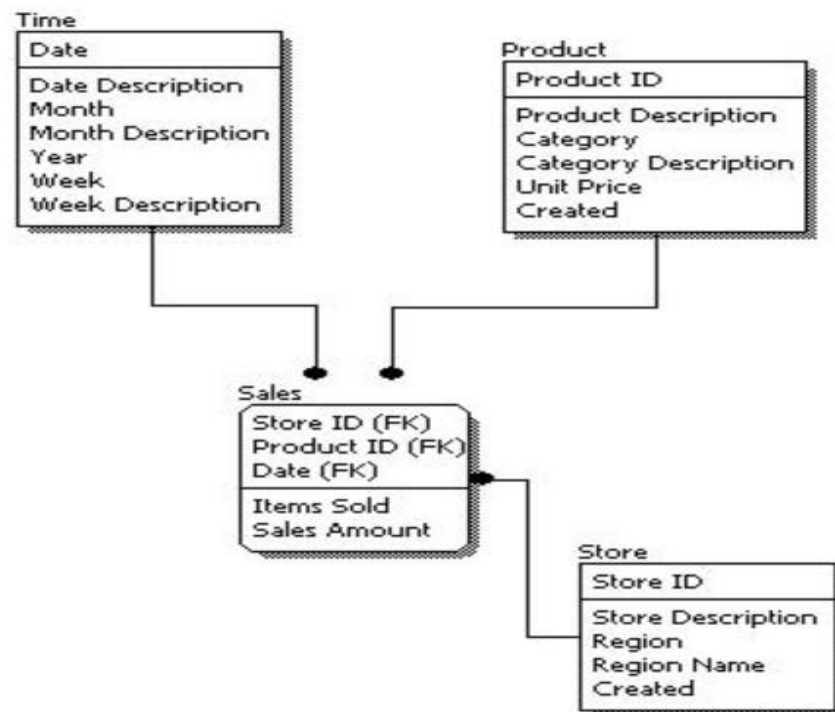


Figure 2. 7 Example of Logical Data Model Design

A primary key is a set of attributes (column) that uniquely identifies the tuples (rows) in a particular table. Each row can only have one primary key at any time. Records in a table are identified independently using primary key. Other functions of the primary key are to compare, sort and store records (Chapple, 2018). It can as well be used to create relationship between records.

A foreign key is column in table which is a primary key of another table which act as a cross reference between tables. It is not always the primary key of another table but foreign key can be a candidate or super key of another table. A super a set of one or more attributes which can uniquely identify a row in a table. The sole use of foreign

keys is to create relationship between tables. Beside the specifying of primary keys for all entities and identifying the relationships between different entities, it is during logical data modelling that all attributes for each entity are found, many-to-many relationships are resolved and normalization.

Features of Physical Data Model:

Just like logical data model, physical data model has primary and foreign keys in addition to table name, column name and column data type.

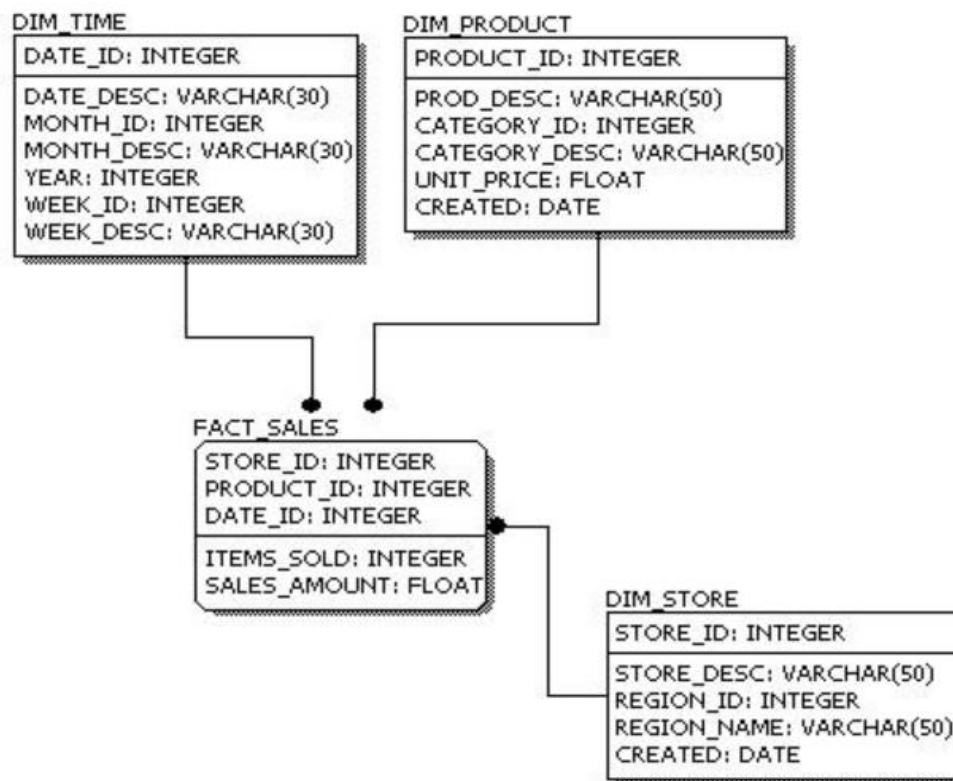


Figure 2. 8 Example of Physical Data Model Design

In physical data model, models are represented the way they will be built in the database. All table structures such as column name, column data type, column constraints, primary key, foreign key and relationships between tables are shown in physical database model. Activities that takes place during physical data modelling includes converting entities into tables, relationships into foreign keys and attributes into columns. Each database type has a unique physical data model. For example, data type for a column may be different between MySQL and SQL Server.

2.3.2 Conceptual Data Modelling

Conceptual data models are high-level description of information based on the business needs for a particular organisation. It is usually developed for business audience. Conceptual data models are not solutions models and they are independent of the application and technology. In conceptual modelling some aspects of the social and physical world around us are described for the purpose of understanding and communications. These social and physical aspects description typically include only the key concepts and their relationships among them. The entities are identified, in most case, without their detailed attributes. During conceptual data modelling, relationships between entities are recognised. It is the conceptual data models which help to establish relationships between entities or business concepts.

The physical world is described in detail and concisely once the requirements have been collected and analysed (Ramez & Shamkant, 2011). The data requirements contain the detail description of objects an organisation handles. The relationship among these objects are easily identified when entity types are known. The inputs are gathered from business shareholders in a context of business process. All this happens during the process of conceptual data modelling.

With defined entities, relationship and their attributes; conceptual data model helps both modeler and non-technical people to understand and communicate among themselves. These concepts in the process of data modelling is independent of the system model.

To describe the physical or social aspect of the World around us is what is called conceptual modelling. Stewart Robinson (Stewart, et al., 2015) defines conceptual modelling as a “concise and precise consolidation of all goal-relevant structural and behavioral features of the system under investigation presented in a predefined format”. Conceptual modelling aims to represent abstract concepts of the connections among objects in the system model. Adding to the above definition of conceptual modelling, the researcher further defined it as presenting project goals from problem statement in a concise and precise predefined format in order to provide the prerequisite guidance for the system model development.

In conceptual modelling process, decisions need to be arrived at in regards to the level of the model and its scope. This decision should be agreed upon by both the problem owners and the modeler. The problem owners, here is referred to as the business stakeholders, who needs the model to help in decision making. The process of conceptual modelling is joint venture task.

Conceptual modelling displays data independent of the database technology (Chen, et al., 2016) for the system. Different directions in conceptual modelling research have been done by various researchers, but no one standard has been agreed upon. This has led to different kinds of conceptual modelling for NoSQL databases which sometimes lead to confusion (Stewart, et al., 2015).

With all these different views about conceptual modelling on various researches and models, it can be concluded by defining conceptual modelling as a model that captures information used in a particular of the business.

The figure below, shows the key artifacts of conceptual modelling as presented by Kotiadis and Robinson.

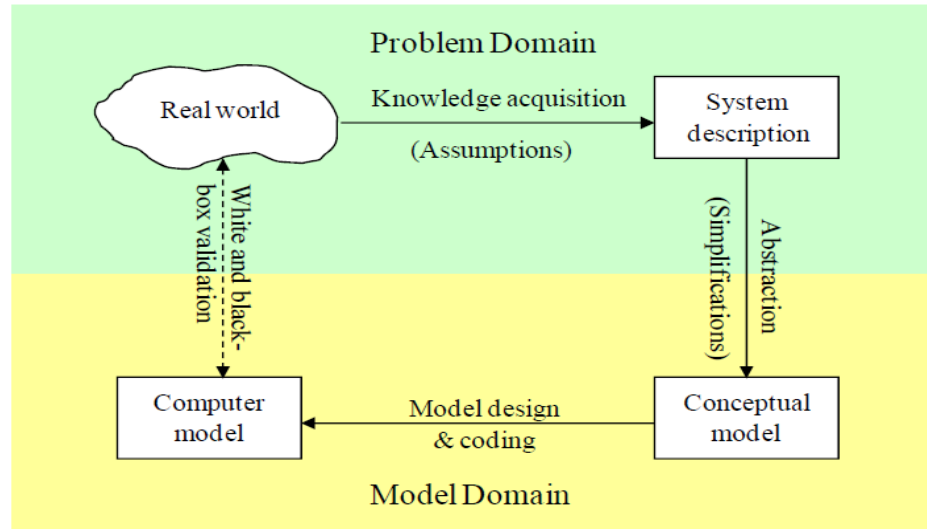


Figure 2. 9 Artifacts of conceptual modelling (Kotiadis & Robinson, 2008)

The flow of information and the real world is represented by the arrow and cloud in the diagram respectively while the three rectangles represent the conceptual modelling process. In the above figure, the flows of information from the real world to system description is an assumption stage.

The conceptual models are as follows:

System description: a description of the problem situation and the system in which the problem situation resides.

Conceptual model: ‘the conceptual model is a non-software specific description of the *computer* simulation model (that will be, is or has been developed), describing the objectives, inputs, outputs, content, assumptions and simplifications of the model.’ (Stewart, et al., 2015)

Computer model: a software specific design and software representation of the conceptual model.

Conceptual data modelling is an art and not a science, thus, there is still room for improving *through* more research. It is also involves understanding the needs of an organisation and how to meet those needs.

Andrew Powell-Morse emphasised that “Conceptual data model should capture the key business entities (a person, place, concept, event, or thing about which the organisation wants to collect data), as well as the relationships between these entities.” (Andrew, 2017)

2.3.3 Roles of Conceptual Data Modelling

Conceptual data modelling plays an important role in obtaining an acceptable solution to the problem. Solution to any problem begins as an idea. This idea, if properly nurtured and modeled becomes an acceptable solution to the problem. A good conceptual data model does not require the physical existence of the system (Varga, et al., 2016); this is because conceptual data models are not technology or system specific. In developing conceptual data modelling, it is important to know that problem description and project goals are the two main prerequisite artefact. In problem description, the problem is defined and the expected solution is previewed.

Robinson (Chen, et al., 2016) argued that conceptual data model is not for data storage structure but for content that made part of information. Conceptual data models are a prerequisite to logical and physical data model. Conceptual modelling help modeler to understand a specific real-world domain and enhance better communication among modelers.

Conceptual modelling is an important base of interoperability and application engineering. Thus, a good design of it enables effective communication of information in the problem description. Conceptual modelling is the first and an important step in application engineering. Malcolm Chisholm stated that conceptual data modelling requires some levels of artistic imagination as they are a product of analysis and not design (Chisholm, 2012).

Mathew West stated that conceptual modelling helps in identifying the main or key objects in the business and how these objects relate to each other. Matthew also added that conceptual modelling provides the specifications required for information in the

business (West, et al., 2016). In the presentation at the 2010 International Conference on E-business Intelligence, it was argued that enterprise data models motivate users and researcher to focus on the identified area of interest (Miao, et al., 2016). It was discovered in the literature of (Steve, 2015) that conceptual modelling offers directions as it reveals business rules which is identified as a subset of model analysis.

The role of conceptual data modelling in design can be summarised as follow:

- A conceptual model is a model of concepts or ideas (abstract) that exist in the mind.
- Conceptual models are used to help us know and understand, design thinking, ideas, casual relationships, principles, data, systems, algorithms or processes.
- Can be used to illustrate relationships that is in the designer's mind to others.
- Concept models can help to explain the thinking behind new ideas.
- Concept models help us to communicate with other members of design team.
- Concept models can help us visualize ideas through graphic, physical and virtual models.

2.3.4 Benefit of Conceptual Data Modelling

Conceptual data modelling provides a broader understanding of the business process, application requirements and even the whole industry in a single piece of paper (Steve, 2015). It is re-echoed that conceptual data modelling offers proactive analysis on the high-level understanding of the system model. More time and resources is saved, as important concerns are identified and discussed at the high level analysis of the business requirements through the process of conceptual data modelling.

Value added information which is obtained through data sharing is another benefit of the conceptual models (Enofe, 2017). This improves the performance in the operations of the organisation. Conceptual modelling allows defining the objects for what they are and what type of attributes they possess. Businesses benefit from enterprise data modelling as it integrates information. Information which is used in

organisation may be kept in form of data in different systems which do not communicate to each other. When the objects of the business in an organisation is properly captured during enterprise modelling, its information is integrated. Thus, conceptual modelling helps to integrate the existing information system.

It was also noted in an online article (Parms, 2018) that conceptual data modelling enables organisation's database design understand their business needs are in terms of data storage and service. It also helps them make an informed decision on the type of database to use as conceptual modelling gives a clear view of the data. However, it should be noted that conceptual data models are not software or technology specific. Conceptual modelling is independent of the software in which it will be used (Robinson, 2013). This makes it to be re-used. It is during conceptual data modelling that modelers get to understand the business of the particular organisation.

For someone to design and build a database, he needs to understand his client's requirements by gathering all important data and how they use it the organisation. A modeler takes time to ask relevant questions patterning to the organisation business. Where definition and clarification of the terms employed in their business is required, the modeler does not hesitate to ask. Thus, every data collected should be fully understood and well gathered if it is to give a full picture of the business of a particular organisation.

Benefits of conceptual data model as adopted from (CareerRide, 2018):

- Helps to create a user interface which is easy to understand and interpret.
- Gives a clear idea of what the product can do, intended use.
- Considers the users points of view hence actual results are easier to achieve
- Describes the roles of different users and their requirements
- Reduces repetitive tasks

2.3.5 Purpose of Conceptual Data Modelling

Conceptual data modelling (Stewart, et al., 2015) help to reveal the completed part of the real world problem that the modeler is trying to solve. With a good conceptual model, the road map to solve the existing problem is halfway done. Thus, the main purpose of developing conceptual models is to give a clear picture of the part of the real world problem. In the same conference paper, Tolk added that conceptual modelling capture and communicate the concept for the unforeseen modeler. It can be seen from above that conceptual data modelling gives an abstract of the world around us, identifying what needs to be model and how to go about modelling (Furian, et al., 2015).

The basis of developing designing conceptual models is that it facilitates the understanding of the shape of the system to be modeled. With the conceptual modelling in place, it can be argued that the future system is given a big picture which enables communication among the participating parties (Kadar, 2017). As the future system model is highlighted by means of enterprise data modelling (i.e. conceptual modelling as other call it) to facilitate discussion. By engaging in detailed discussion by the concerned individuals, mistakes and misunderstanding is minimised on the system model under design. It should be noted that enterprise data models form an important system model documentation.

2.3.6 Data Modelling in Non-Relational Databases

Data modelling is always there in database, whether you are dealing with schema-less non-relational database or the traditional relational approach. That process of understanding the needs or requirements of the business, one may consider creating some drawings. It is this “understanding” which is the aspect of data modelling.

For data modelling to remain relevant in this era of NoSQL technology, it needs to be re-inverted from the relational sequence of Conceptual>Logical>Physical. It should

be noted that designing a NoSQL database should be application-specific in order to leverage the benefits of the technology (Pascal, 2018). Pascal suggested (Pascal, 2018) that logical modelling exercise should be eliminated in NoSQL data modelling because NoSQL schema design is de-normalized and logical modelling is supposed to be normalized. Conceptual modelling of some kind should be applied in NoSQL data modelling to document the understanding and the blueprint of the business.

Traditional conceptual models such as entity-relationship model has been in use since (Connolly & Begg, 2015) 1979 when Peter Chen developed. This conceptual modelling has been a choice to use by the majority database designers since its inception. In conceptual modelling, the first step is to determine the business entities and how they relate to each other. To fully understand how these entities, how they relate to one another, their properties have to be defined (ArangoDB, 2018).

In contrary to Pascal's suggestion, Kwangchul Shin and his colleagues (Claudino de Freitas Myller & Sagado, 2016) proposed the conceptual data model for NoSQL database design based on Peter Chen's frame work. Their proposed of NoSQL database design consists of three phases just like the traditional relational database. The first phase is the developing of conceptual schema which is independent of the NoSQL database. The conceptual schema can be used to develop logical schema which is dependent on data model used in NoSQL database but independent from NoSQL database software. In physical design, physical schema is developed which describes a data structure for implementation in a certain NoSQL database software (Claudino de Freitas Myller & Sagado, 2016).

A big data system requires high-quality data modelling methods for organising and storing data, allowing us to reach the optimal balance of performance, cost, efficiency, and quality. Appropriate models and storage environments offer the following benefits to big data:

Performance: Good data models can help us quickly query the required data and reduce I/O throughput.

Cost: Good data models can significantly reduce unnecessary data redundancy, reuse computing results, and reduce the storage and computing costs for the big data system.

Efficiency: Good data models can greatly improve user experience and increase the efficiency of data utilization.

Quality: Good data models make data statistics more consistent and reduce the possibility of computing errors.

2.3.7 Challenges of Conceptual Data Modelling in Non-Relational Databases

In non-relational database, database does not follow the relational model used in traditional database management system. Data in this big data environment comes in different formats such structured, unstructured and the combination of the two. This data comes in huge volume and are generated at a high speed, thus, creating uncertainty in the minds of users on its value.

The fact that data which possesses 5V characteristics cannot properly be stored and managed using traditional relational databases poses a challenge to be modeled using relational approach. Relational data model is not flexible, making difficult to accommodate data which comes in different formats. The volume of data keeps on increasing as its velocity, this brings a challenge on how to model data using state-of-art relational data modelling. In the white paper of MongoDB, Non-relational databases have evolved from relational technology in these ways:

Data models: Relational models require predefined schema, while non-relational databases offer flexible schema design that make it much easier to update the database to handle changing application requirements.

Data structure: Unlike relational databases, non-relational databases are designed to handle unstructured data that doesn't fit neatly into rows and columns to suit most of the data generated today which is unstructured.

Scaling: with non-relational databases, scale your system horizontally easier by taking advantage of cheap, commodity servers.

Development model: Most of NoSQL (non-relational) databases are open sources. This eliminates the cost of paying for any software licensing fees upfront.

However, these NoSQL technologies vary greatly in the way data is stored despite that these databases are all more scalable and flexible than their relational counterparts. The popular ways in which data is stored among the NoSQL technologies include key-value, graph, column family and documents stores. Thus, a need to look at a general approach for conceptual data modelling to cater for the most common available NoSQL technologies.

2.4 Analysis on the Peer-Reviewed Papers on Conceptual Data Modelling for Non-Relational Databases

In this study, further survey from peer reviewed papers on conceptual modelling for non-relational databases was conducted so that an approach for conceptual data model for non-relational database can be drawn from the information obtained in both the local survey and international literature publications. As indicated in chapter three, very few papers attempted to discuss this topic of conceptual data modelling in the era of big data. Table 2.4 gives an overview on non-relational database as discussed in the peer reviewed papers.

Table 2. 4 Peer Reviewed Paper Analysed

No	NoSQL Database Area applied	Problem Description	Proposed Solution	Type(s) of Conceptual Model Used	Source (Reference)
1	Key-Value stores	Traditional techniques for data modelling not sufficient for non-relational databases	Methodology that can support the design of NoSQL Key-value stores	EER and UML	A modelling methodology for NoSQL key-value database (Rossel & Manna, 2017)
2	Column-family stores	The need for conceptual model to define how data can be structured in databases	Proposed UML conceptual model to describe Big Data in column-family stores	UML	Big Data and knowledge management: How to implement conceptual models in NoSQL systems (Abdelhedi, et al., 2016)
3	All four NoSQL databases	In the era of fast accumulation of data than it can be stored or processed	Proposed the use of ontological approach		Big Data- Conceptual modelling to the rescue (David, et al., 2013)
4	All four NoSQL databases	Lack of indicator models to support Big Data implementation	Proposed a holistic model of six (6) indicators for Big Data implementation		Conceptual model for successful implementation of Big Data in organisations (Mohanad & Ahmed, 2015)
5	Column-family, document and key-value stores	Need for conceptual model to define data structure in database	Proposed a Model Driven Architecture (MDA)-based approach to implement UML conceptual model in Big Data	UML and MDA-approach	Logical unified modelling for NoSQL databases (Fatma, et al., 2017)

6	Graph stores	Requirements for database methodology for NoSQL to meet the needs of Big Data	Proposed a graph database schema (GDBS) which is similar to UML except it didn't have "operation"	EER and UML	Modelling graph database schema (Noa, et al., 2017)
7	Document stores	The coexisting of relational and NoSQL databases in polyglot persistence environment makes the consistent data management difficult	Proposed the use of UML class diagram to represent conceptual schema which was transformed document data model	UML	NoSQL database design using UML conceptual data model based on Peter Chen (Kwangchul, et al., 2017)

2.4.1 Discussions on the Findings from Peer-Reviewed Papers on Conceptual Data Modelling for Non-Relational Databases

Table 2.4 above shows the findings from the analysis of peer reviewed papers

Overview of Peer Reviewed Papers

The paper "A modelling methodology for NoSQL key-value databases" by Gerardo Rossel and Andrea Manna explained how to present entities of conceptual by using the concepts of datasets. In their paper, the authors claimed to have presented an original methodology that can support NoSQL key-value databases. The approach involved establishing of the datasets then define the relationships among them. The datasets were group into some forms to represent the real world entity type, relationship and events. The steps for the methodology, Gerardo and Andrea proposed had three steps. The first step was requirement analysis, at this stage, data requirement was transferred to conceptual modelling. By applying conceptual models, logical design was established. This logical design had defined datasets, relationships and value specializations. The graphic representation was similar to the unified modelling

language (UML) diagram except their proposed graph indicated “key” instead of UML’s entity and attributes. For the “operation” in UML, they called it “value”. The proposed conceptual data modelling in this paper was the use of UML class diagram as a conceptual model for dataset, but it is specific to a specific class of NoSQL database, Key-Value databases. Thus, it is not generic and does not adapt to other classes of NoSQL databases.

The second paper which was analysed was authored by four researchers namely Fatma Abdelhedi, Amal A. Brahim, Faten Atigui and Gilles Zurfluh. In their paper entitled “Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems” suggested model driven architecture (MDA) approach using UML class diagram to implement a conceptual model for column-oriented NoSQL databases. All attributes of a particular class are grouped into the same column-family with each row describing as an instance. In their study, the proposed conceptual data modelling was for column-families NoSQL database by the use of UML class diagram. Below is the target meta-model for their proposed conceptual data modelling.

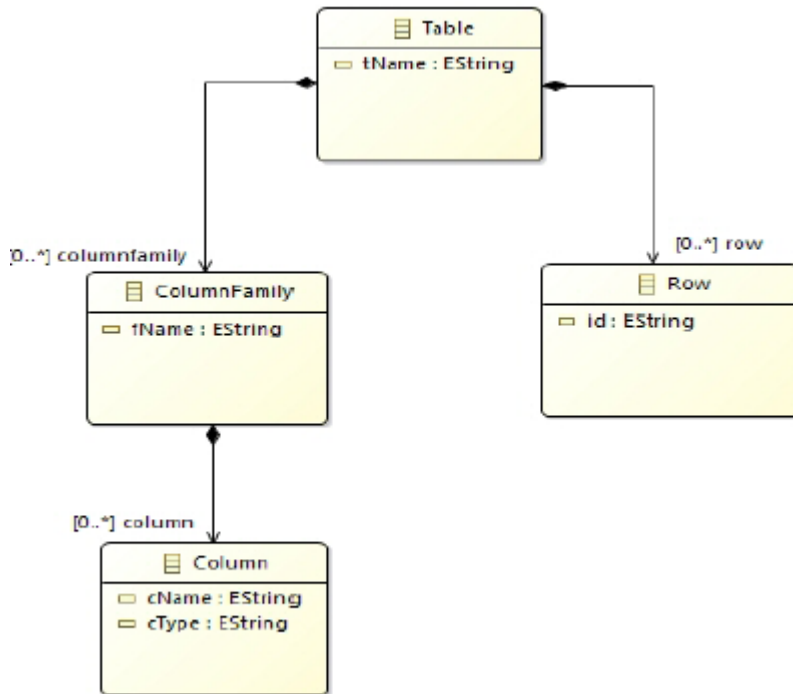


Figure 2. 10: Target meta-model

Another paper which the researcher looked at was an extended abstract from the paper “Big Data- Conceptual Modelling to the Rescue”. This article was authored by David W. Embley and Stephen W. Liddle. These two authors looked at many different approaches to address the challenge of Big Data. One of the approach used to address the issue of 3V (volume, velocity and variety) of Big Data suggested conceptual modelling by information extract which involved grouping conceptual models by associating regular patterns in objects and relationships.

Mohanad Halaweh and Ahmed E. Massry proposed a six indicator conceptual model to successfully implement Big Data in an organisation. The indicators included Top Management support, Organisation Change, Data availability and quality, Infrastructure, Require skills, and Privacy and Security.

Noa Roy-Hubara, Lior Rokach and Bracha Shapira authored a paper entitled “Modelling Graph Database Schema”. In their paper, they proposed a graph database schema (GDBS) by transforming entity-relational diagram (ERD). The schema for NoSQL graph database was created from the entity-relationship diagram (ERD) conceptual schema. Their approach involved the process of two steps for mapping ERD to a NoSQL graph database. The proposed NoSQL graph database is made up of nodes and edges. The nodes are subdivided into two layers, the top and lower, referred to as the “label or name” and “properties” respectively. The nodes are what represented objects or events of the real world. The edge provides an interconnection between nodes and have properties and cardinality constraints. This conceptual schema for NoSQL graph database is similar to UML class diagram for relational database except it does not have operations. The authors of this paper went further to suggest the graph database schema by transforming the conceptual schema for NoSQL graph database.

The last paper which was analysed was written by Kwangchu Shin, Chulhyun Hwang and Hoeyung Jung and was called “NoSQL database Design Using UML Conceptual

Data Model Based on Peter Chen's Framework". Similar to the authors above, Shin and the colleagues used UML diagram to represent conceptual data model for NoSQL database. This UML was used as conceptual data model because it is independent of all the popular logical data model for NoSQL databases (key-value, graph, document and column-family stores). The UML applied by these authors omitted operation layer which is applicable in the relational database. These authors appear to be suggesting that their proposed UML diagram for NoSQL database can be applied to any logical data model for NoSQL database. In their study, the focus was on converting the UML diagram to document stores data model. The conversion was achieved by transforming "class" into "collection", attribute to column in document and association to references or embedded.

CHAPTER 3: METHODOLOGY

3.1 Introduction

The study was carried out both as a field assessment and desk research study. The field sampling study and literature review was designed as a descriptive study to provide baseline information on the existing conceptual modelling approach for the area under investigation.

3.2 Research Topic Selection

The process of selecting a “topic” also referred to as a subject, issue or area under discussion is important in the success of the research project. This is because a well selected topic motivates the researcher to collect the required data for analysis (Kombo & Tromp, 2014). A broad area of study in Information and Communication Technology was identified which gave the researcher plenty of options. The researcher reviewed some papers on the selected area of study. This helped the researcher to come up with a proposed number of topics.

In order to make the topic manageable, the researcher worked on word selection and combination so that it is not too broad or narrow. The keywords which were identified includes but not limited to “database design”, “system implementation”, “digital era”, “data models” and “data analysis”. These combination of keywords help the researcher to develop clarity over the topic selected. A lot of care was taken to ensure that the research topic is not only researchable, interesting and can make contribution to the knowledge bank but also provocative, clear and focused (Kombo & Tromp, 2014). After reflecting on these keywords, this topic of study was formulated. The topic was submitted to the School of Engineering, University of Zambia for approval. The researcher had also put into consideration the available financial resources and time during the selection of this study topic.

3.3 Study Area

The study area for this research was Lusaka which is the hub of Centre for Zambia’s industrialization. Lusaka host most of the headquarters to the major organisations and

institutions. Due to its rich in industrialization, Lusaka was made to be the focal point for this research. Besides highest learning of institutions, Lusaka harbours headquarters for banking institutions, non-governmental organisations, government agencies/authorities, telecommunications/ICT/ISP companies, energy/utility companies, health cares and many other firms which consumes ICT services, particularly in the area database technology. The research intended to extend its coverage for data collection to other parts of the country, however, due to budget constraints and time, data collected by the use of questionnaires came from Lusaka. In order to get deep understanding of conceptual data modelling for non-relational databases, a new database technology for Big Data, the study area was extended to peer-reviewed papers which were published within the period of eight (8) years. The peer-reviewed papers were discussed under chapter two above.

3.4 Research Approach

The study employed the combination of quantitative and qualitative approaches in the collection and analysis of data. A mixed approach was applied to minimise biasness. This is because any used approach in research has its own limitations. The combined approach was used to collect data through questionnaires and peer-reviewed papers analysis. The main approach used was qualitative in the data collection while quantitative was also applied especially in analysing the collected data. The qualitative approach was applied on both peer-reviewed papers and questionnaires, particularly on part B which dealt with technical data while quantitative was applied part A of the questionnaire on the respondent's line of specialty, professional certification, organisation's primary industry and their main role their respective organisations.

The questionnaires were distributed to forty (40) database users and experts of different organisations in Lusaka, Zambia. The questionnaires consisted of both structured and closed questions, and utmost eight answer options in some few instances which allowed the respondents to provide more than one (1) answer in certain questions. Generally, the questionnaire was designed in such a way that some questions required only one answer, while other questions provided the respondents

with an option of selecting more than one answer. Prior to giving out the questionnaires, respondents were requested to participate and only those who consented participated.

The other source of data collection was through peer-reviewed papers where more than 100 articles were retrieved. In this approach, care was taken in selecting phrases for searching on the internet in order to get the best from the search engines.

3.5 Research Design

Research design is a procedural plan adopted by the researcher on how data collection and analysis will be done to suit the research purpose (Kumar, 2011). It is through the research design that the researcher chooses how he is going to collect data from respondents, how he is going to choose respondents, and the collected information will be analysed.

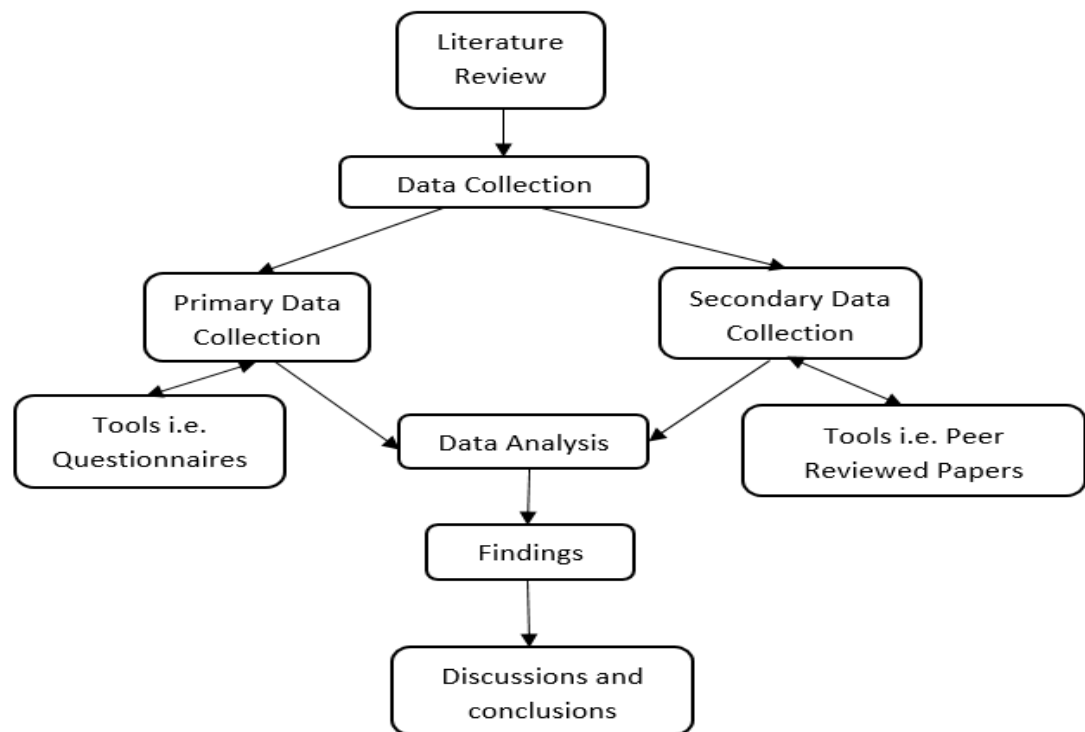


Figure 3. 1 Research Design

Figure 3.1 above summarises the research methodology employed in this study. The researcher engaged both primary data collection and secondary (desk research) which made use of both qualitative and quantitative approaches.

3.5.1 Research Instruments

To gather as much relevant information for the study, the researcher used the methods of questionnaire and peer-reviewed papers. A well-structured and closed questionnaire was designed. The questionnaire had two main parts, the bio data and technical. The bio data covered the respondent's specialty, professional certification and their primary role in the organisation. The other information in the part of bio data is organisation's primary industry. The technical part was used to collect data about subject of study by covering the objectives.

In order to deepen the understanding for the area under discussion, another instrument was used besides the questionnaire, the peer-reviewed papers, where only relevant papers to the study were analysed. These were the two main research instruments used in this research study

3.6 Data Collection

Data was gathered from two main sources: questionnaires and peer-reviewed papers retrieved from online journals, conferences and other online database storages. Two sources of data collection were adopted so that information gap in one source could be met by the other source.

3.6.1 Questionnaire Design

A questionnaire is a research instrument made up of related set of questions for the purpose of gathering information from respondents on a particular subject of study. The questionnaire design is dependent on the type of information that is required to be collected.

In order to achieve the study objectives, quantitative questionnaire was used. The questions in the questionnaire were planned and designed to collect specific information through structured questions. The questions were dichotomous and multiple-choice in nature i.e. "yes/no" and single select or multi-select respectively. The researcher wanted to distinct choices between open and closed-ended questions.

In a situation where the researcher did not want to restrict the respondents' thoughts, open ended choices of "others" were provided so as to allow them use their opinions.

3.6.2 Primary Data Collection

During questionnaire distribution and collection, the researcher keenly observed the presence of skilled manpower and ICT infrastructure for database technology of the target institutions. Insightful information was collected and was useful in the research. A total of 40 questionnaires were distributed to randomly selected database users and expert from different organisations. The targeted respondents for this research were considered to be experts in database technology for their respective institutions as they were directly or indirectly associated to data management.

The researcher designed the questionnaires in such a way that the respondents had option of providing more than one (1) answer to some questions and only one to other questions. The respondents were required to pick the applicable answer(s) by ticking in a box next to the chosen answer or answers. In few instances, the respondents had options of provided extra information if their choice on a particular question was "Other". This was done by physically writing on the provided space by the respondents. The researcher waited for a period of not more than two (2) weeks for the participants to complete the task of filling in and then the researcher made a follow up in person to collect them.

Prior to attending to the questionnaires, participants were requested to participate and only those who consented took part. Ethical issues such as anonymity, confidentiality and non-disclosure of respondents' identities and their organisations was assured to them and the researcher faithfully complied. The researcher ensured that the collected information was verified and correctly recorded in a spread sheet.

The questionnaire was divided in two parts: "Part A" for Bio Data and "Part B" for technical data. Question codes were designed as shown in Appendices II and III respectively. The codes were useful during data analysis using spreadsheet.

3.6.3 Secondary Data Collection

In this technique, data collection was from already collected data by other researchers whose findings were reviewed. This is referred to as the secondary source of data collection. To further address the objectives of this study, the researcher surveyed literature between 2010 and mid-2018. The criterion of the reviewed period was arrived at, after taking into consideration that non-relational database is rather a new phenomenon; it was not expected to find reviewed study on conceptual data modelling for non-relational databases prior to 2010. Beside the choice of the period of publication, the other two inclusion criteria used were: 1) English published articles; and 2) relevance to the research questions. This involved three steps: 1) collection of related works, 2) filtering of relevant works and 3) close analysis of the relevant works. In the first step, the researcher search used keyword-based search for journals, conference papers and other peer-reviewed articles was carried out from ResearchGate and ScienceDirect, as well as from Google Scholar. The two databases were considered due to their broad coverage of relevant literature and advanced features such as suggesting related literature. From the two databases and Google Scholar over 100 papers were retrieved. After filtering on the relevance of the papers, 84 papers were selected of which only seven (7) papers were used after detailed review and analysis of the state of the art related works. The keywords for searching on the internet were grouped into two, the first group addressed conceptual modelling (i.e. conceptual schema, business modelling, enterprise data modelling) while the second group focused on the Big Data (i.e. big data, digital era, non-relational, NoSQL database).

3.7 Data Analysis

Data analysis is the most crucial part of any research. Data analysis summarises collected data. It involves the interpretation of data gathered through the use of analytical and logical reasoning to determine patterns, relationships or trends. In this study, both manual data analysis and the use of computer was applied.

3.7.1 Manual Data Analysis

Data analysis involves the editing of raw data. This helps to detect and eliminate errors; and proactively make corrections. Data collected through questionnaire was analysed before it was subjected to spreadsheet on a computer by checking on the correct entry and consistent. Checking on consistence was verified in a situation, as an example, where a respondent is qualified to attend to a particular question on a condition that his or her response to a previous question was a “yes”. The other reason to why manual analysis was applied is to check content completeness and formulate aid coding for computer based analysis.

To analyse the data which was collected from peer-reviewed (secondary data collection) papers, content analysis was applied. The content analysis was adopted to help structuring the qualitative data collected in a way that satisfies the succeeding of the research project. It also assisted in eliminating papers which were not relevant to the study

3.7.2 Computer Based Data Analysis

The qualitative data collected particularly from questionnaire was analysed using Microsoft excel, a computer based analysis using the spread sheet. Microsoft-excel is a statistical in nature and has better features with capability in generating graphs. This research is an academic fulfillment, thus, the researcher was the sole editor, however, was under the guidance of the supervisor so that data consistency is maintained.

3.8 Research Findings and Discussions

The findings relate to the outcomes from the research questions that guided the study. Findings from the analysis of the primary and secondary data were presented in chapter four while discussions are in chapter five. Data were analysed to identify the use of conceptual data modelling in unstructured dataset. Further, the data were analysed to explore the use of non-relational database in local organisations. Data were obtained from self-administered questionnaires and peer-reviewed papers.

3.9 Limitation

This research is not without its limitations like many other studies. One of the limitation was the sample which consisted of organisations and business firms from Lusaka province of Zambia which may limit the generalisability of the results despite the regions being the industrial hub of the country. The other major limitation is limited number of open database for peer-reviewed papers. Most reliable (e.g. IEEE Xplore) online databases required subscription and purchase of their papers.

CHAPTER 4: DATA COLLECTION AND ANALYSIS

4.1 Research Findings

This chapter presents the results in regards to the types of datasets in the era of digital world, deployment of various databases in Zambia and conceptual data modelling approaches in Big Data environment. Graphs were used in presenting some data that was collected. Data in graphs was also presented in percentages and frequencies. Questionnaires were used to collect data in this research. A total of 32 completed questionnaires were retrieved representing 80% response rate.

4.1.1 Bio Data

The researcher distributed copies of questionnaires for this research to forty (40) database experts and users from various organisations; thirty-two (32) responded. From the respondents, 43.75% studied Computer science/engineering while 37.5% and 18.75% specialised in IT and Telecommunications respectively.

The majority of those with certifications were certified by Oracle with an overall percentage of 44%. On the other hand, about 28% of the respondents did not have any certification. Others were certified in Microsoft and IBM mounting to 28% and 18.75% respectively.

Table 4. 1 Respondents' Industries

Industry	Responses	
	Frequency	Percentage
Government	3	9%
Health	8	25%
Telecom/ICT	5	16%
Banking	5	16%
Utility	3	9%
Transport	1	3%
Academia	6	19%
Others	1	3%
Total	32	100%

Table 4.1 above, show that respondents were selected from different organisations which included non-governmental organisation in health sector (25%), Telecom/ICT

companies (16%), Banking (16%), Academia (19%), government agency/authority (9%), Energy/Utility (9%), Transport (3%) and Others (3%).

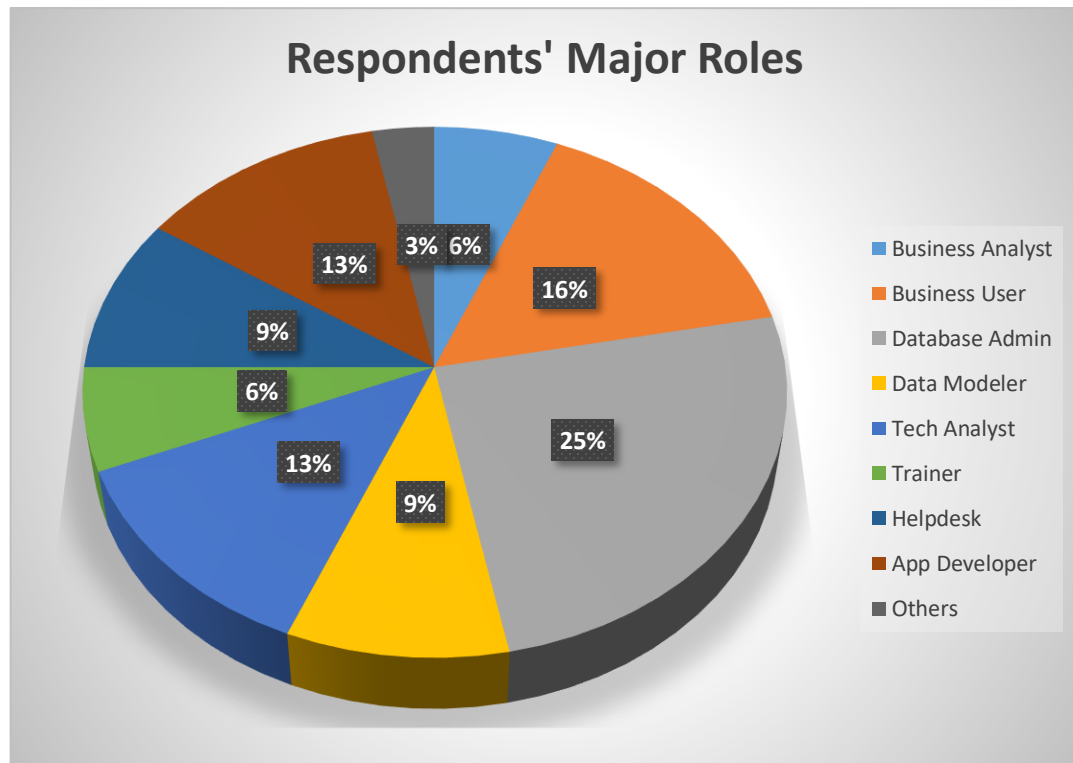


Figure 4. 1 Respondents' major roles in their organisation

The respondents were asked the major role they play in their respective organisations; the majority of the respondents were Database Administrators and this was followed by business users as shown in figure 4.1 above.

4.1.2 Dataset in the Digital World

Technology is developing in all aspect of business enterprise across all industries man is engaged in. With this eruption in technologies come with not opportunities but also challenges. The usefulness of data has evolved over time from valueless commodity to highly valued product. In this subsection, we will look at how dataset ranging from small data to big data is considered in different local industries.

Dataset

Respondents were asked about data collection; and the majority said that data is always collected in one form or another for future use while a few, just about 15% expressed ignorance on the collection of data. The other 9% claimed that their organisations does not collect data.

The researcher discovered that dataset about employees or clients, whose volume and format is accessible, informative and actionable was considered 78% of the respondents as small data, while others, just about 16%, thinks that this type data can either be small data or big data; and only 6% said this is Big Data.

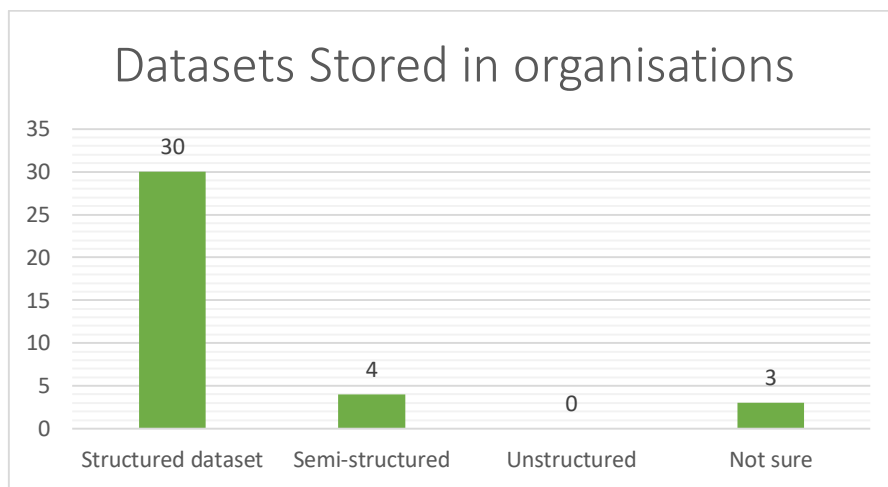


Figure 4. 2 Dataset Types Stored in Organisations

It was observed that the surveyed organisations have not started the storing of unstructured data as no one was found to have worked or is working on this dataset type. Almost every institution survey indicated that they collect and store structured dataset with only about 12.5% collect semi-structured ones while 9% were not sure. The figure 4.2 above shows the respondents' frequency on the dataset types stored in their organisations.

Big Data

From the survey, volume was the most known characteristic of Big Data and was followed by value of the dataset. Figure 4.3 below shows that other characteristics of Big Data, which is recognised and defined by Data Scientists are not well known by the local organisations. The less known characteristics of Big Data includes velocity, veracity and variety which were rated at 25%, 22% and 16% respectively.

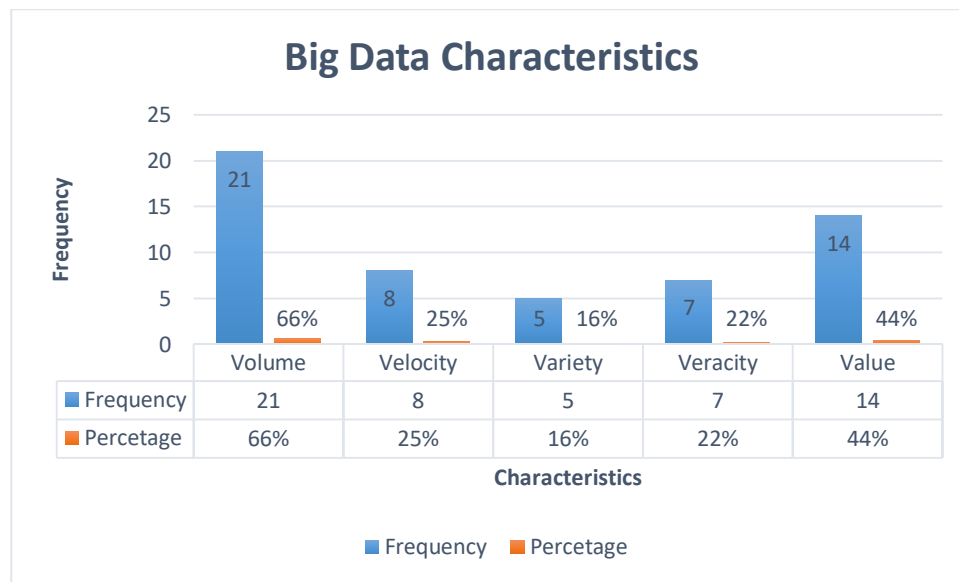


Figure 4. 3 Big Data Rating locally

The survey showed that data collected from sensors such as biometric identity and business transaction are the major known source of Big Data as confirmed by the majority respondents. Other noted sources of Big Data locally included information about employees and patients, social media and public data. With Toll Plaza and speed cameras mounted on roads, the majority (about 81%) agreed that Big Data sources in Zambia is on the rise.

The survey shows that more than half of the respondents said dataset whose volume is more than one (1) Petabytes is classified as Big Data, while about 25% indicated that any data stored in a database qualifies to be called Big Data and 9% of respondents did not respond.

The few organisations who claimed to have started working on Big Data, mainly used it in marketing, business intelligence, customer feedback analysis and information about the health of their clients.

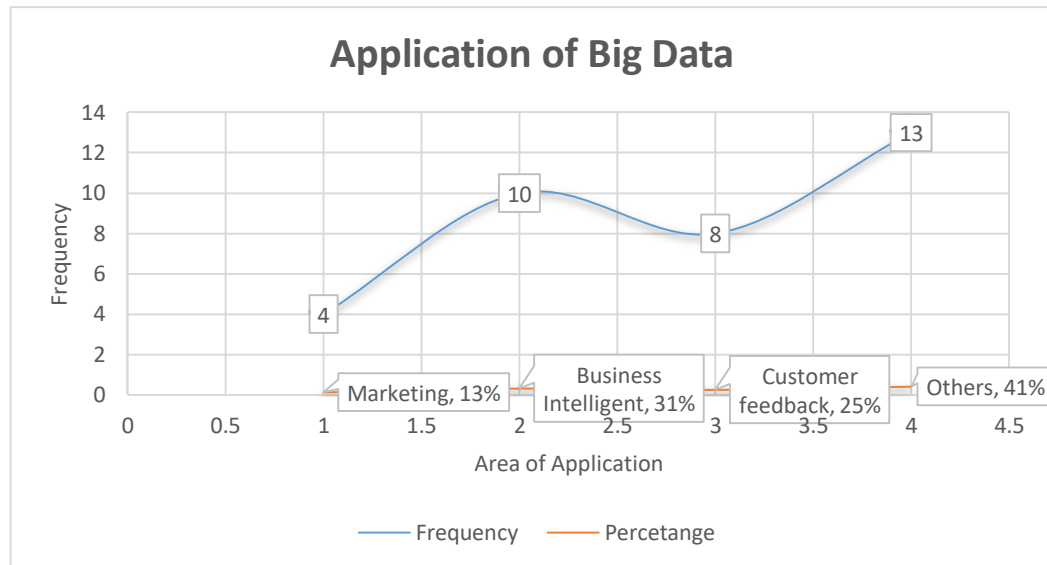


Figure 4. 4 Application of Big Data

4.1.3 Databases

The result of the survey shows that majority of the people and organisations are still using relational databases such as MySQL (45%), Oracle (33%) and Microsoft SQL server (22%). Figure 4.5a shows the comparative percentage of the deployed databases. This confirms to the responses obtained on the categories of database deployed; this revealed just about a quarter use the licensed while the rest uses either open source database applications or both.

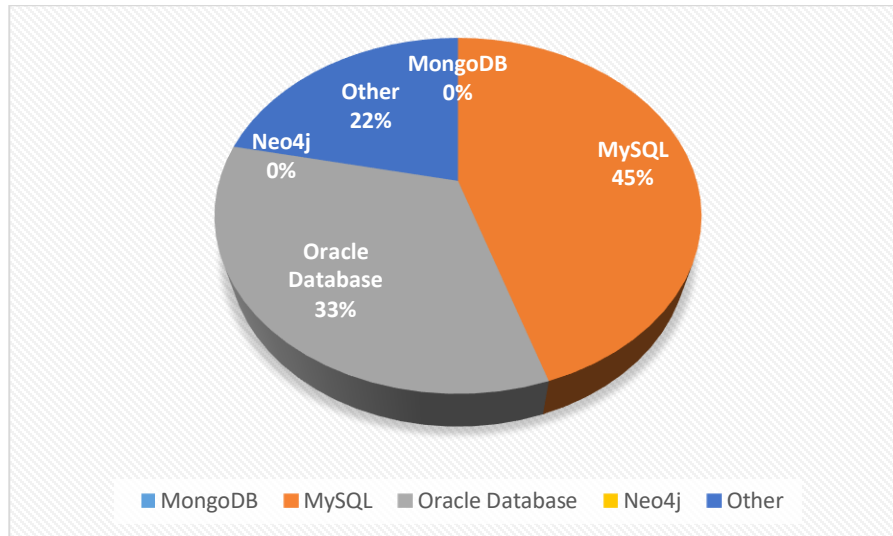


Figure 4.5a. Comparative Percentage on the Deployed Databases

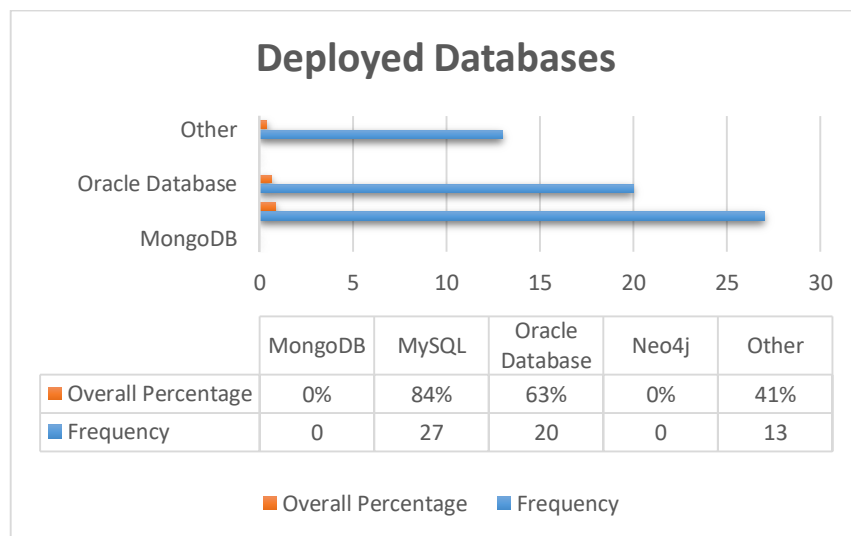


Figure 4.5b. Overall Percentage of the Deployed Databases

Figure 4. 5 The types of databases in use in local organisations

On the evolution of database, the survey shows that all respondents and their organisations have deployed relational databases while only about 22% and 9% of the respondents indicated to have worked on network and object-oriented types of database respectively. It was discovered that none has deployed hierarchical and non-relational databases.

4.1.4 Data Modelling Overview

In the survey, it was noted that data modelling is a necessity when designing database. All respondent confirmed on the importance and for data modelling before loading data in the database application.

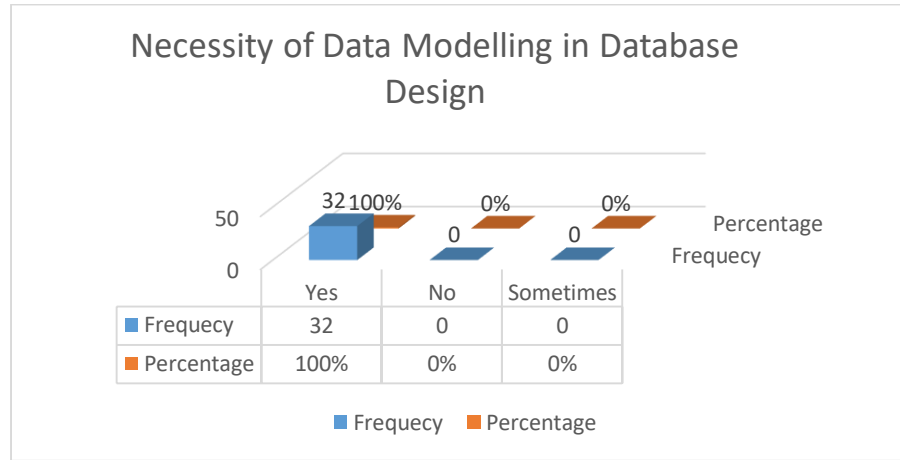


Figure 4. 6 Necessity of Data Modelling

A question was asked to find out if the process of data modelling act as a hindrance in database design process. The responses show that about 47% said that data modelling does not hinder the process of database design while 37.5% shows that it does hinder and the other 15.5% said, sometimes it stands in the way of database design. The figure below revealed MySQL workbench as the most popular modelling tool and is followed by Oracle data modeler while others are available but not so much used.

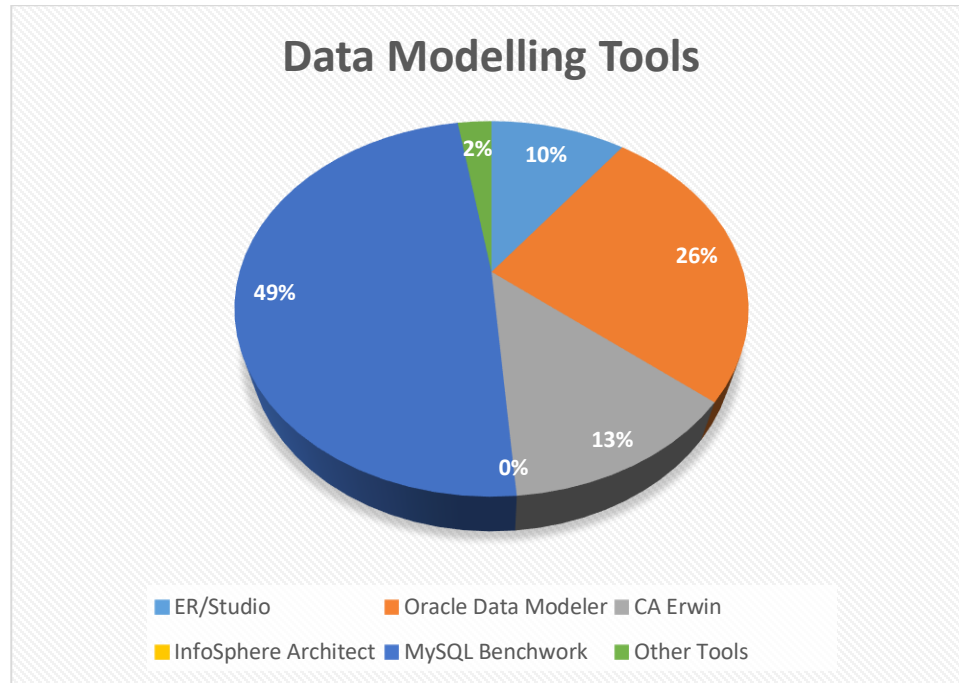


Figure 4. 7 Data Modelling Tools

In the survey, all respondents indicated that the knowledge of database management system is an important skill. Other skills with more than 90% responses include the knowledge of both business and modelling pattern. Further, the ability to differentiate between what is ideal from what is practical was considered to be an important skill. It was noted also that the skills of knowing when to stop and verbal communication are essential.

Table 4. 2 Data Modeler Skills

Modeler Skills	Responses		Percent of Case (%)
	Frequency	Percent (%)	
Business knowledge	28	17%	88%
Modelling pattern knowledge	30	18%	94%
Differentiate idea from practical	26	16%	81%
Knowing when to stop	22	13%	69%
DBMs Knowledge	30	18%	94%
Communication skills	27	17%	84%
Normalization knowledge	0	0%	0%
Total	163	100%	

4.1.5 Conceptual Data Modelling

Respondents were asked about the areas where their organisations use conceptual data modelling. From the survey, it was revealed that defining the key issues of problems which needs to be addressed by the system, addressing digital and non-digital concepts, and closing the gap between a solution model and document requirements are the many areas where conceptual modelling used.

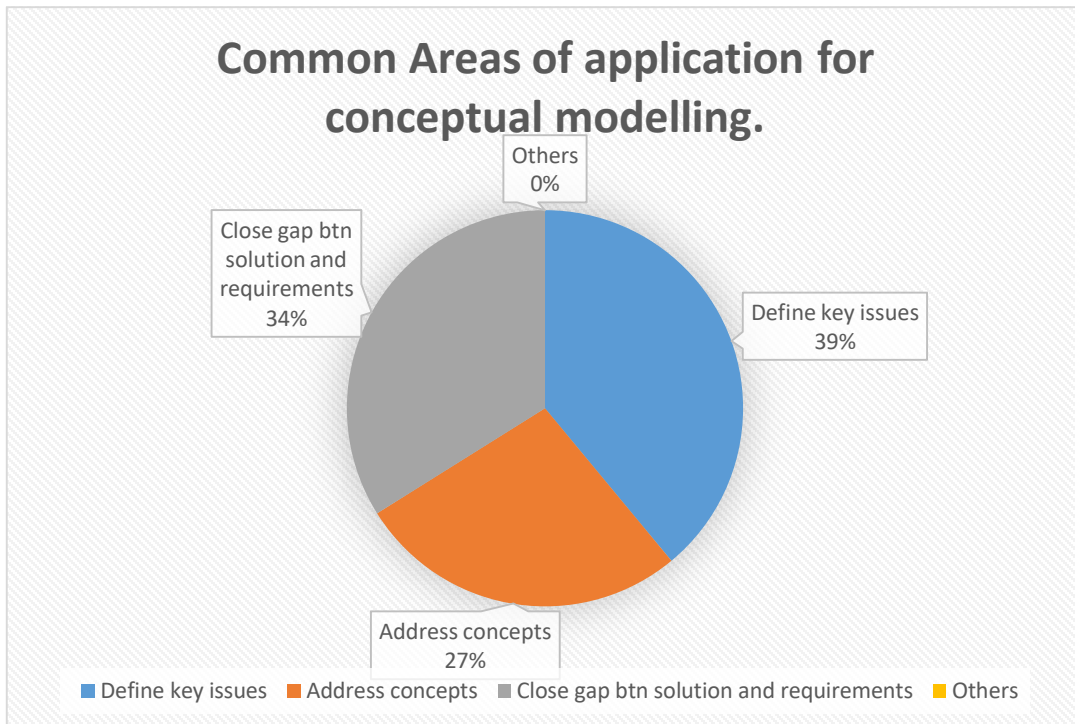


Figure 4. 8 Areas of Application for Conceptual Modelling

Majority of respondents confirmed that during conceptual data modelling, modeler needs to understand and capture the business knowledge about a particular enterprise, however, few respondents thought that it can also be understood at either logical or physical level.

In the research, it was observed that conceptual data modelling provides a big picture for better communication between the modeler and the business shareholders. This confirms the 91% of the respondents who stated that it is important to engage business owners or shareholders during the process of conceptual data modelling. Further, all

respondents agreed that conceptual modelling displays data independent of the technology or software application.

Enterprise Data Modelling is similar to Conceptual Data Modelling except that Enterprise Data Modelling is the practice of creating a graphical model of the data used by an enterprise or company. It was discovered that those organisations that has Enterprise Data Models for their activities, they use it to support applications integration (21%), as starting point for application modelling (22%), information system planning (28%) and business intelligence (29%). The survey shows that Enterprise Data Models is slightly more used for business intelligence and starting point for modelling than it is applied for system planning and applications integration.

Table 4. 3 Areas of Application for EDM

Area	Frequency	Comparative (%)	Overall (%)
Application integration	14	21%	44%
Application modelling	15	22%	47%
Inform sys planning	19	28%	59%
Business intelligence	20	29%	63%
Total:	68	100%	213%

CHAPTER 5: DISCUSSIONS AND CONCLUSION

5.1 DISCUSSIONS

This research can serve as a starting point for further research in data modelling for non-relational databases, while encouraging further exploration on the applications of Big Data in this era of digital business. All the research questions have been addressed in this chapter. The results of the study collected were interpreted in reference to the relevant literature and principal data sources such as questionnaires.

5.1.1 Demographic Data

The results showed that majority of the respondents, who were drawn from different organisations, have qualifications in Computer Science/Engineering while others are specialised in Information Technology and Telecommunications Engineering. From these findings, it was revealed that most institutions employ computer scientist and IT specialists to handle data in their organisations and in few cases Telecommunications engineers. However, it was discovered that few of them have professional certifications in database; from the few who are certified they have database certification in oracle, Microsoft and IBM.

The major roles of the respondents were Database Administrator and followed by business users. Other roles found among respondents included technical analyst, application developer and data modeler as shown in Figure 4.1 above.

5.1.2 Dataset in the Digital World

Technology is developing in all dimension of business enterprises across all industries man is working on. With this eruption in technologies come not only with opportunities but also challenges. The usefulness of data has evolved over time from valueless commodity to highly valued product. In this subsection, we looked at how dataset ranging from small data to big data is considered in different local industries. There is a good indication on data collection as revealed in this survey, though it is mainly in the area of structured datasets and a bit of semi-structured. With no one engaged in the collection of unstructured dataset shows that organisations are losing

out on the business potential which can be harvested in unstructured data. To stand a better chance in competing with the developed countries, our local institution need to start the collection of unstructured data.

The well-known and internationally recognised characteristics of Big Data are volume, velocity, variety, value and veracity also known as 5V. In the results obtained from this survey revealed that most of the local data scientists can only identify “volume” as the only characteristic of Big Data and a few other added “value”. This shows that most people who handles data in Zambia defines Big Data in terms of volume. The other three important characteristics of datasets are barely associated with Big Data. This is because organisations have not started storing unstructured datasets.

Besides the biometric identity and business transactions the researcher discovered that Toll Plaza and speed cameras mounted on the trunk roads are the major sources of Big data in Zambia. Despites all these known sources of Big Data, the potential business in these datasets remain untapped due lack of deployment in non-relational databases which is capable of managing unstructured data. With a proper conceptual data modelling approach which can lead to the implementation of non-relational databases, there is high chance that Big Data can be a source revenue in our local enterprises as it will provide a good understanding of each component of the 5V of Big Data. Though, few organisations have started to work on Big Data, which is mainly used in marketing, business intelligence, customer feedback analysis and information about health of their patients; the researcher strongly feels that a clear and simple approach in conceptual modelling can be the good starting point for many firms to actively engage in Big Data analysis.

5.1.3 Database Deployments in Zambia

Zambia has not been left out in the deployment of databases especially, relational databases. It was clearly shown in the survey that all respondents’ organisations are actively using relational databases. MySQL database is most popular database; this is because it an open source and easy to implement. Other deployed databases are oracle

and Microsoft SQL Server. It was observed from the survey that neither an individual nor organisation are using or have worked on non-relational databases. The fact that most organisations are comfortable to use open source databases is a good indicator that the deployment of NoSQL databases will be a success once a clear and simple conceptual data modelling approach for non-relational databases is developed.

Among the five major developments in database evolution that have been seen on the market (hierarchical, network, object-oriented, relational and non-relational databases), the survey shows that all respondents and their organisations are using relational databases while only a few of the respondents indicated to have used network and object-oriented databases respectively, at one time. This is because at the time some of these databases were introduced very few organisations in Zambia, if any, were using database. As for NoSQL database, which is new on the market, it is an indication that, more sensitization on non-relational databases need to be done so that organisation can move with technology as it unfold. Further, the researcher thinks that NoSQL databases are not deployed because Zambians takes time to adopt to new change. The researcher concluded that the fact that there are few publications and research works on conceptual data model for non-relational databases, this has resulted on organisations not implementing non-relational databases.

5.1.4 Data Modelling

All respondents said data modelling is necessary in database development. The responses, in relation to the three steps of data modelling in relational database, show that modeler faces less or no challenges during data modelling, however, few indicated that it hinders and sometimes slows down the process of developing database system. This is because all institutions surveyed are still using relational databases. From this, it can be seen that there an element of hindrance in the traditional data modelling process. Thus, if this process is applied in non-relational database, more challenges can be encountered as dataset comes in enormous sizes at high speed in different formats and are stored differently from the way relational databases stores them. Further, it can be seen that traditional approach in data

modelling is still appreciated by the majority because the modelers and their organisations have not started using non-relational databases.

All the modelling tools found are associated with relational databases. This is just another good reason why they still using relational database like MySQL, Oracle and Microsoft Server. It was observed that all these leading data modelling tools are typically used for relational database, this show that there is less or no engagement in non-relational data modelling.

In the survey, almost all respondents indicated that the knowledge of database management system is an important skill. Other skills which attracted more than 90% response includes the knowledge of both business and modelling pattern. The ability to differentiate between what is ideal from what is practical was considered to be an important skill as well. It was noted also that the skills of knowing when to stop and verbal communication are essential. Thus, with all these skills in data modelling, adopting to NoSQL databases will not be a challenge. It is also observed that the only lacking component in the deployment of non-relational database is the right approach in conceptual data modelling for this type of database.

In this study, it was revealed that organisations that use conceptual modelling is mostly applied when defining the key issues of problems which needs to be addressed by the system, addressing digital and non-digital concepts, and closing the gap between a solution model and document requirements. This is an indication that most organisation can easily handle Big Data when an approach for conceptual data modelling for it is developed. Thus, with an approach in place, engaging the business owners during conceptual data modelling on any type of dataset can provide the data modelers with a big picture for better communication.

Enterprise Data Modelling is a key to gain competitive advantage in this digital world business. It was discovered that less than 20% of the respondents' organisations have Enterprise Data Models for their business activities. The few institutions that uses

them, they are mainly used to support applications integration, as starting point for application modelling, information system planning and business intelligence. It is our hope that the proposed conceptual data modelling approach for non-relational database can help organisations to benefit from the increasing data in this era of Big Data.

5.1.5 Proposed Conceptual Modelling Approach for Non-Relational Database

The survey was conducted on various organisations and it revealed that, Big Data is not fully understood in organisations surveyed in Zambia. Further study on peer reviewed papers showed that no research work was found with a proposed conceptual data modelling for Non-Relational databases which can accommodate all the four developed classes of NoSQL databases. Below is the discussion of findings on the selected peer review papers which helped to develop the proposed conceptual data modelling for all the four well developed classes of NoSQL databases.

In their paper, the authors said to have presented an original methodology that can support NoSQL key-value databases (Rossel & Manna, 2017). The approach involved establishing of the datasets then define the relationships among them. The datasets were group into some patterns to represent the real world entity type, relationship and events. The steps for the methodology proposed by Gerardo and Andrea had three steps. The first step was requirement analysis, at this stage, data requirement was transferred to conceptual modelling. By applying conceptual models, logical design was established. This logical design had defined datasets, relationships and value specializations. The graphic representation was similar to the unified modelling language (UML) diagram except their proposed graph indicated “key” instead of UML’s entity and attributes. For the “operation” in UML, they called it “value”. The proposed conceptual data modelling in this paper was the use of UML class diagram as a conceptual model for dataset, but it is specific to a specific class of NoSQL database, Key-Value databases. Thus, it is not generic and does not adapt to other classes of NoSQL databases.

In this paper, “Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems”, the authors suggested model driven architecture (MDA) approach using UML class diagram to implement a conceptual model for column-oriented NoSQL databases (Abdelhedi, et al., 2016). All attributes of a particular class are grouped into the same column-family with each row describing as an instance. In their study, the proposed conceptual data modelling was for column-families NoSQL database by the use of UML class diagram. Below is the target meta-model for their proposed conceptual data modelling.

Another paper which the researcher looked at and was helpful to come with the proposed conceptual data modelling was an extended abstract from the paper “Big Data- Conceptual Modelling to the Rescue” (David, et al., 2013). The authors looked at many different approaches to address the challenge of Big Data. One of the approaches which was used to address the issue of 3V (volume, velocity and variety) of Big Data suggested conceptual data modelling by information extract which involved grouping conceptual models by associating regular patterns in objects and relationships. Objects with similar patterns were grouped together, from which relationships to other groups were generated.

Noa, et al., proposed a graph database schema (GDBS) by transforming entity-relational diagram (ERD) (Noa, et al., 2017). The schema for NoSQL graph database was created from the entity-relationship diagram (ERD) conceptual schema. Their approach involved the process of two steps for mapping ERD to a NoSQL graph database. The proposed NoSQL graph database is made up of nodes and edges. The nodes are subdivided into two layers, the top and lower, referred to as the “label or name” and “properties” respectively. The nodes are what represented objects or events of the real world. The edge provides an interconnection between nodes and have properties and cardinality constraints. This conceptual schema for NoSQL graph database is similar to UML class diagram for relational database except it does not have operations. The authors went further to suggest the graph database schema by transforming the conceptual schema for NoSQL graph database.

Similar to the authors above, Shin and the colleagues used UML diagram to represent conceptual data model for NoSQL database (Kwangchul, et al., 2017). This UML was used as conceptual data model because it is independent of all the popular logical data model for NoSQL databases (key-value, graph, document and column-family stores). The UML applied by these authors omitted operation layer which is applicable in the relational database. These authors appear to be suggesting that their proposed UML diagram for NoSQL database can be applied to any logical data model for NoSQL database. In their study, the focus was on converting the UML diagram to document stores data model. The conversion was achieved by transforming “class” into “collection”, attribute to column in document and association to references or embedded.

The findings from peer reviewed papers, discussed above, helped to provide the researcher with gap for the proposed conceptual data modelling approach for non-relational databases. The proposed approach involves three steps as follows:

Step 1: Entity Identification and Documentation

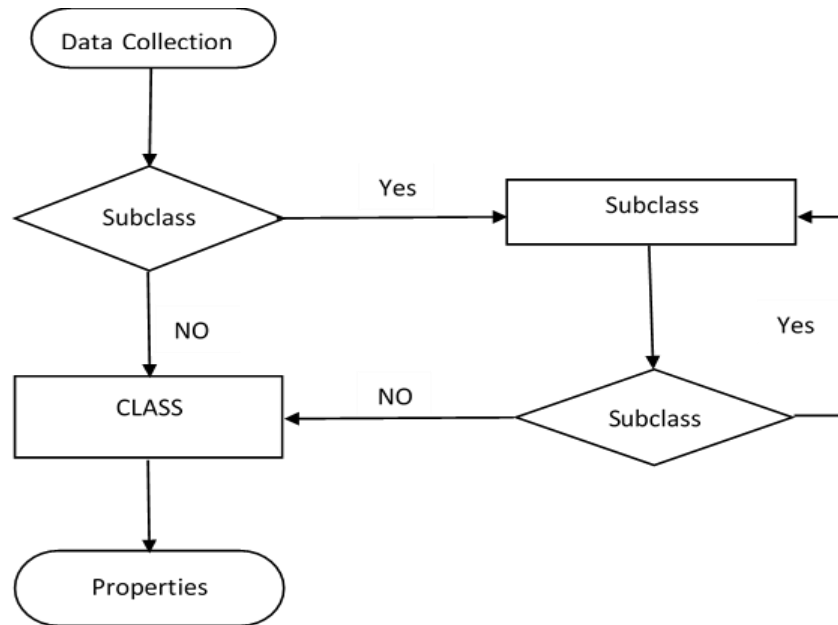


Figure 5. 1 Entity Identification of the collected Data

From figure 5.1 above, data collected is grouped according to their families which is referred to as a “Class” in this study. A class may have subclass, for example, the collected data in agriculture can be referred to as “Farming”. In this case “Farming” is represented by the term, Collection. The ‘Collection’ can be grouped in several classes e.g. in example of agriculture, classes may include arable, pastoral, workers, etc. In our example arable (crop) is considered to be a “class” which can have subclasses such as Maize, Cassava, Sorghum, etc. Each class or subclass has properties.

Table 5. 1 Entity Identification and Documentation

Collection	Class	Subclass	Properties
Farming	Crop	Maize	Family, Variety, Maturity, Soil, Season, etc.
		Sorghum	Family, Variety, Maturity, Soil, Season, etc.
	Animal	Chicken	Family, Types, Breed, Sex, colour, etc.
		Goat	Family, Types, Breed, Sex, colour, etc.
	Workers	Skilled	Names, Sex, Duty, Sex, Salary, etc.
		General	Names, Sex, Duty, Sex, Salary, etc.

Step 2: Identifying Relationship Types

The table below shows the proposed approach in identifying relationship types for conceptual modelling for non-relational database.

Table 5. 2 Identifying relationship types

Class (Entity Name)	Multiplicity	Relationship	Multiplicity	Class (Entity Name)
Workers	0: 1	Manages	0: n	Workers
Workers	1: n	Feeds	0: n	Animals
Animals	0: n	Eats	0: n	Crops
Crops	1: n	planted	1: n	Soil-type

Step 3: Representing Data in diagram or Graphs

The final step in the proposed conceptual data modelling is representing of collected data in a diagram which independent of any NoSQL database technology (key-value, graph, column-based and document stores).

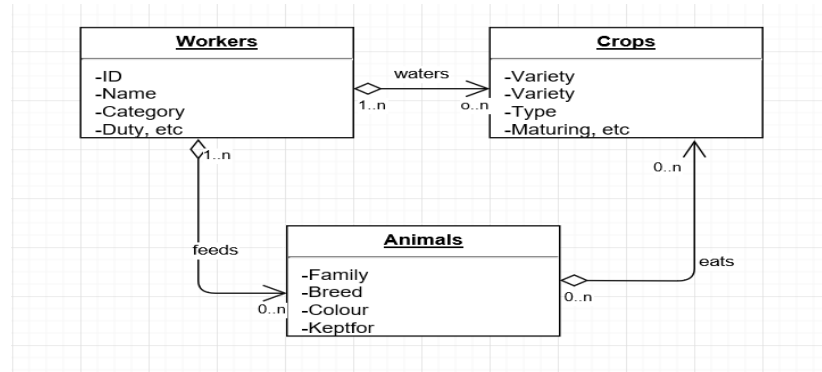


Figure 5. 2 Proposed NoSQL Conceptual data modelling

5.2 CONCLUSION

This chapter concludes the study and also makes recommendations based on the major findings of the research. The research findings show that organisations are engaged in the collection and storing of data. The datasets most stored are structured and a bit of semi-structured. This types of datasets organisations collects does not fully qualify to called Big Data. This is due to the fact that, the data collected are mainly defined by the volume while the other aspect of big data such as velocity, variety, veracity and value are not properly.

Further, the finding in terms of dataset, shows the country is experiencing an increase in the source of Big Data. For instance, data are daily generated through internet of things from the Toll Plaza and speed cameras mounted on major roads. Despite of these sources of Big Data; its value is not harvested. Thus, organisations need to be smart enough when embracing this information in order to mitigate risks, control costs, and more accurately project future revenue and minimise losses (Frappolli,

2018). In the study, no institution was found to be fully engaged in the business of Big Data and the few who claim to have started, it is mainly to do with volume and value of datasets which is applied in analysing their prospects and clients.

It was found that all organisations surveyed are using relational databases to store data. The type of relational database varied from one institution to another, with MySQL found to be the most common database deployed. Other notable types of databases in use were Oracle and SQL Server databases. MySQL database is the most popular in our local institution, this could be attributed to it being an open source software. In this study, no individual respondent or company they represented was found using any type of non-relational databases. This shows that as country, we are still lagging behind in terms of adapting to new change on the market.

In the study, it was revealed that data modelling process designed for traditional (relational) databases slowdown the pace for database designing. This is confirmed through the respondents as less than half said that the three (conceptual, logical and physical) steps in data modelling does not hinder or slow down the process of database implementation. There are various modelling tools found in use in our local firms. MySQL workbench is the most common and it was followed by Oracle Data modeler. The most required skills for data modeler was found to be the knowledge of business, data modelling pattern and database management systems. Other important skills pointed out were knowing the line between ideal and practical, knowing when to stop and having the skill in documentation and verbal communication. In the study, it showed that conceptual data modelling is used to define the key issues of problem which need to be address by the system, to address both digital and non-digital concepts, and also in closing the gaps between a solution model and requirements document. Further in the study, it was discovered that organisations that uses enterprise data models is mainly used for business intelligence, system planning and application integration; also as integrated view of the data and processes (Smith, 2011). In this study, only about 20% of respondents indicated that their organisations have fully implemented enterprise data models.

5.3 RECOMMENDATIONS

As revealed in this study, there is need for awareness activities, which may be in the form of researches, trainings or conferences, to inform local organisations on the importance of Big Data. The importance of Big Data does not rely on how much data a company is in position with, instead it is more on its utilisation of the collected data. The more efficient an organisation uses its data, the greater is its potential to grow its business. The proper use of Big Data can result in many benefits for the company. Some of the benefits of Big Data includes cost reduction, improved business efficiency, focus on local preferences, improved product pricing, hiring of the right personnel and increases sales and customer loyalty.

Enterprise Data Modelling is an important aspect of conceptual data modelling that gives more benefits to data users, thus, educating data users in industries on the importance of enterprise data models is a necessity. Data users benefit from enterprise data model through increased efficiency of the company's operation, re-engineering of business processes to take advantage of improved data understandability, availability, and quality. Other advantage of enterprise data model is, its helpful to meet the enterprise's business goals, including adaptation to changing market, regulatory, and other environments.

There is also need to sensitizing local firms to adopt new database for storing data. This can greatly help them to survive and grow their business in this global market. In the research, it was found that all organisations surveyed deployed relational database. The findings in this study reveals that local firms and organisations are yet to start implementing non-relational database for data storage. It also showed that the use of Big Data is still at infancy stage. There are two areas which are closely related to this study were not looked at. The first was the ways of converting the proposed NoSQL conceptual modelling to the particular type of NoSQL database. The other research work area which was not handled is the benefits of using Big Data in digital business, a case of a Zambian firm.

REFERENCES

- Abdelhedi, F., Amal, A. B., Atigui, F. & Gilles, Z., 2016. *Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems?..* Paris, SCITEPRESS – Science and Technology Publications,, pp. 235-240.
- Abdul, S. G. & Tariq, S. R., 2015. Big Data Analysis: Ap Spark Perspective. *Global Journal of Computer Science and Technology*, 15(1), pp. 6-14.
- Abert, Z. Y. & Sherif, S., 2017. *Handbook of Big Data Technologies*. 1st ed. Sydney: Springer International Publishing.
- Abraham, S., F.Henry, K. & S, S., 2011. *Database Sytem Concepts*. 6th ed. New York: McGraw-Hill Companies, Inc..
- Admin, 2018. *Facebook -Every Second*. [Online]
Available at: www.everysecond.io/facebook
[Accessed 26 August 2018].
- Ahmed, E. et al., 2017. The Role of Big Data Analytics in Internet of Things. *Elsevier*, 129(2), pp. 459-471.
- Andlinger, P., 2013. *RDBMS dominate the database market, but NoSQL systems are catching up*. [Online]
Available at: https://db-engines.com/en/blog_post/23
[Accessed 1 September 2018].
- André, R., Afonso, S. & Alberto, R. d. S., 2015. Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *Journal of Software Engineering and Applications*, I(1), pp. 617-634.
- Andrew, P.-M., 2017. *Conceptual Models- What are they and How Can You Use Them*. [Online]
Available at: <https://airbrake.io/blog/sdlc/conceptual-model>
[Accessed 1 September 2018].
- Anon., 2018. *Internet Live Stats*. [Online]
Available at: <http://www.internetlivestats.com/one-second/>
[Accessed 4 December 2018].
- ArangoDB, 2018. *Switching from Relational Databases to ArangoDB-Comparing RDBMS to a Native Multi-Model Database System*, Cologne: ArangoDB GmbH.
- Bansari, K. H. & Hetal, J., 2017. A Survey of Non -Relational Databases with Big Data. *International Journal on Recent and Innovation Trends in Computing and Communication*, V(11), pp. 143-148.
- Baron, C. A., 2015. NoSQL Key-Value DBs Riak and Redis. *Database Systems Journal*, VI(4), pp. 3-10.

Berg, L. K., Seymour, T. & Goel, R., 2013. History Of Databases. *International Journal of Management & Information Systems*, 17(1), pp. 29-36.

BigData, 2016. *What is Big Data*. [Online]
Available at: bigdata.black/featured/what-is-big-data/
[Accessed 30 August 2018].

Cameron, M., 2012. *Big Data The definitive guide to the revolution in business analytics*, London: Fujitsu Services Ltd.

CareerRide, 2018. *What are the benefits of creating a conceptual design?* *Business analyst*. [Online]
Available at: <https://www.careerride.com/ba-benefits-of-conceptual-design.aspx>
[Accessed 24 June 2018].

Chapple, M., 2018. *What is a Primary Key*. [Online]
Available at: <https://www.lifewire.com/primary-key-definition-1019179>
[Accessed 12 December 2018].

Cheng, X. et al., 2014. Understanding the YouTube partners and their data: Measurement and analysis. *IEEE Explore*, 11(12), pp. 26-34.

Chen, P., Bernhard, T. & Leah, W. Y., 2016. *Future directions of Conceptual Modeling*, Louisiana: Louisiana State University.

Chinchmalatpure A, M. & Dhore, P. M., 2016. Review of Big data Challenges in Healthcare Application. *IOSR Journal of Computer Engineering (IOSR-JCE)*, I(1), pp. 6-9.

Chisholm, M., 2012. *Big Data and the Coming Conceptual Model Revolution*. [Online]
Available at: <https://www.information-management.com/news/big-data-and-the-coming-conceptual-model-revolution>
[Accessed 18 May 2018].

Claudino de Freitas Myller, S. D. Y. & Sagado, A. C., 2016. *Conceptual Mappings to Convert Relational into NoSQL Databases*. Recife, SCITEPRESS – Science and Technology Publications.

Connolly, M. T. & Begg, E. C., 2015. *Database Systems- A Practical Approach to Design, Implementation, and Management*. 6th ed. Harlow: Pearson Education Limited.

Cornelia, G., Robert, R. & Roxana, S., 2015. A Comparative Study of Relational and Non-Relational Database Models in a Web- Based Applications. *International Journal of Advanced Computer Science and Applications*, VI(11), pp. 78-83.

Couchbase, 2017. *Why NoSQL Database?*. [Online]
Available at: <https://www.couchbase.com/resources/why-nosql>
[Accessed 3 September 2018].

David, E. W., Stephen & W, L., 2013. *Big Data—Conceptual Modeling to the Rescue*. Utah: Brigham Young University.

Demaitre, E., 2018. *Sensors Lead to Big Data, But Don't Expect Shortcuts, Says Sensors Expo Host*. [Online]
Available at: <https://www.roboticsbusinessreview.com/events/sensors-lead-big-data-dont-expect-shortcuts-says-sensors-expo-host/>
[Accessed 20 November 2018].

Dybka, P., 2014. *MySQL: Why "My"?*. [Online]
Available at: <https://www.vertabelo.com/blog/notes-from-the-lab/mysql-history>
[Accessed 27 January 2019].

Ebner, J., 2015. *How Sensors Will Shape Big Data and the Changing Economy*. [Online]
Available at: <https://dataconomy.com/2015/01/how-sensors-will-shape-big-data-and-the-changing-economy/>
[Accessed 15 June 2018].

Edd, D., 2012. *Big Data Now*. 1st ed. Beijing: O'Reilly Media.

Editor, I., 2013. *What Can MySQL Do For You?*. [Online]
Available at: <https://www.inap.com/blog/what-can-mysql-do-for-you/>
[Accessed 27 January 2019].

Editorial, 2018. *History of MySQL*. [Online]
Available at: <https://dev.mysql.com/doc/refman/8.0/en/history.html>
[Accessed 27 January 2019].

Enofe, M., 2017. *Data Management in an Operational Context: A study at Volvo Group Trucks Operations*. 1st ed. Lund: Dept. of Informatics, Lund University School of Economics and Management.

Erl, T., Khattak, W. & Buhler, P., 2016. *Big Data Fundamentals- Concepts, Drivers & Techniques*. 2nd ed. Boston: Arcitura Education Inc.

Fatma, A., AIT BRAHIM, A. F. A. & Gilles, Z., 2017. *Logical Unified Modeling For NoSQL DataBases*. Porto, s.n.

Foote, D. K., 2018. *A brief History of Non-Relational Databases*. [Online]
Available at: www.dataversity.net/a-brief-history-of-non-relational-databases/
[Accessed 2 September 2018].

Frappolli, M., 2018. *Data Analytics and Big Data Management*. [Online]
Available at: <https://www.ewsolutions.com/data-analytics-and-big-data-management/>
[Accessed 25 November 2018].

Furian, N. et al., 2015. A conceptual modeling framework for discrete event simulation using hierarchical control structures. *Elsevier*, I(1), pp. 82-96.

- Ganesh, C. D., 2015. Base Analysis of NoSQL database. *Elsevier*, I(1), pp. 13-21.
- Guy, H., 2016. *Next Generation Databases- NoSQL, NewSQL and Big Data*. 1st ed. New York: Apress Media.
- Hadi, H. J., Shnain, A. H., Hadishaheed, S. & Ahmad, A. H., 2015. Big Data and 5V Characteristics. *International Journal of Advances in Electronics and Computer Science*, January, 2(1), pp. 16-23.
- Hammad, K. A. I., Fakharaldien, M. A. I., Zain, J. M. & Majid, M. A., 2015. *Big Data Analysis and Storage*. Orlando, IEOM Society.
- Hammink, J., 2018. *The Types of Modern Databases*. [Online]
Available at: <https://alooma.com/blog/types-of-modern-databases>
[Accessed 3 September 2018].
- Hurwitz, J., Nugent, A., Halper, F. & Marcia, K., 2018. *Unstructured Data in a Big Data Environment*. [Online]
Available at: <https://www.dummies.com/programming/big-data/engineering/unstructured-data-in-a-big-data-environment/>
[Accessed 17 12 2018].
- IBM100, 2016. *Relational Database*. [Online]
Available at: www.03.ibm.com/ibm/history/ibm100/us/en/icons/reldb/
[Accessed 3 September 2018].
- Import.io, 2018. *What is data, and why it it important*. [Online]
Available at: <https://www.import.io/post/what-is-data-and-why-is-it-important/>
[Accessed 3 September 2018].
- Jackson, E. et al., 2016. *The Value of Big Data in Agriculture Inputs Farming and Processing: Big Data characteristics*, Illinois: International Food and Agribusiness Management Review.
- João, R. L. et al., 2015. Choosing the right NoSQL database for the job: a quality attribute evaluation. *The Journal of Big Data*, 2(18), pp. 1-26.
- Joshi, N., 2017. *Top 5 sources of big data*. [Online]
Available at: <https://www.allerin.com/blog/top-5-sources-of-big-data>
[Accessed 6 December 2018].
- Jyothi, S. B. & S, J., 2015. A Study On Big Data Modeling Techniques. *International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC)*, 5(6), pp. 19-26.
- K.V, K. & M, V., 2015. Unstructured Data Analysis-A Survey. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(3), pp. 223-225.

- Kadar, M., 2017. *Data Modeling and Relational Database Design in Archaeology*, Romania: University of Alba Iulia.
- Keith, F. D., 2017. *A Brief History of Big Data*. [Online]
Available at: www.dataversity.net/brief-history-big-data/
[Accessed 3 September 2018].
- Kelly, B. & Irena, K., 2016. Big Data in food and agriculture. *Big Data and Society*, 1(1), pp. 1-5.
- Kirti & Maan, P., 2015. Database for Unstructured, Semistructured data-NoSQL. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, IV(2), pp. 466-469.
- Kitchin, R. & Lauriault, P. T., 2015. *Small data, data infrastructures and big data*, County Kildare: The Programmable City.
- Kitchin, R. & Lauriault, T., 2015. Small data in the era of big data. *GeoJournal*, II(3), pp. 463-475.
- Kitoyi, N., 2018. *The impact of Big Data on Video Streaming*. [Online]
Available at: <https://channels.theinnovationenterprise.com/articles/the-impact-of-big-data-on-video-advertising>
[Accessed 20 December 2018].
- Kombo, D. K. & Tromp, D. L. A., 2014. *Proposal And Thesis Writing: An Introduction*. 14th ed. Nairobi: Paulines Publications Africa.
- Koseleva, N. & Ropaite, G., 2017. Big data in building energy efficiency: understanding of big data and main challenges. *Elsevier*, I(1), pp. 544-549.
- Kotiadis, K. & Robinson, S., 2008. *Conceptual Modeling: Knowledge Acquisition and Model Abstraction*. Coventry, IEEE.
- Kumar, R., 2011. *Research Methodology: A step-by-step guide for beginners*. 3rd ed. London: SAGE Publications Ltd.
- Kumawat, D. & Pavate, A., 2016. Correlation of NOSQL & SQL Database. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 18(5), pp. 70-74.
- Kwangchul, S., Chulhyun, H. & Hoekyung, J., 2017. *NoSQL database design using UML conceptual data model based on Peter Chen*. s.l., Research India Publications., pp. 632-636.
- Legay, L., 2016. *What is Data?*. [Online]
Available at: <https://www.doingdata.org/blog/what-is-data>
[Accessed 31 December 2018].

- LiveStats, I., 2018. *Twitter Usage Statistics*. [Online]
Available at: www.internetlivestats.com/twitter-statistics
[Accessed 26 August 2018].
- MacCreary, D. & Kelly, A., 2014. *Making Sense of NoSQL*. 1st ed. Shelter Island,:
Manning Publications Co.
- Mallon, S., 2018. *What To Know About How Big Data Is Affecting Social Media*.
[Online]
Available at: <https://www.smartdatacollective.com/what-to-know-about-how-big-data-affecting-social-media/>
[Accessed 20 December 2018].
- Manoj, K. V., 2014. Comparative Study of NoSQL Document, Column Store Databases and Evaluation of Cassandra. *International Journal of Database Management Systems (IJDMS)*, 6(4), pp. 11-26.
- Margaret, R., 2017. *Data*. [Online]
Available at: <https://searchdatamanagement.techtarget.com/definition/data>
[Accessed 3 September 2018].
- Meola, A., 2018. *What is the Internet of Things (IoT)? Meaning & Definition*. [Online]
Available at: <https://www.businessinsider.com/internet-of-things-definition?IR=T>
[Accessed 28 November 2018].
- Miao, H. S. J., Ge, S. & Wang, N., 2016. *Research on Metadata-driven Enterprise Data Modeling System*, Nanjing: Atlantis Press.
- Mohanad, H. & Ahmed, E. M., 2015. Conceptual Model for Successful Implementation of Big Data in Organizations. *Journal of International Technology and Information Management*, XXIV(2), pp. 21-34.
- MongoDB, 2018. *Top 5 Considerations When Evaluating NoSQL Databases*, New York: MongoDB.
- Moniruzzaman, A. B. M. & Hossain, A. H., 2013. NoSQL Database: New Era of Databases for Big data Analytics-Classification, Characteristics and Comparison. *International Journal of Database Theory and Application*, VI(4), pp. 1-13.
- M, V. & K.M, K., 2015. Unstructured Data Analysis-A Survey. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(3), pp. 223-225.
- Nayak, A., Poriya, A. & Poojary, D., 2013. Type of NOSQL Databases and its Comparison with Relational Databases. *International Journal of Applied Information Systems (IJAIS)*, V(4), pp. 16-19.
- Nazrul, S. S., 2018. *CAP Theorem and Distributed Database Management Systems*. [Online]
Available at: <https://towardsdatascience.com/cap-theorem-and-distributed-database->

[management-systems-5c2be977950e](#)

[Accessed 25 January 2019].

Nidhi, D., 2018. *4 Major Ways in Which Big Data is Impacting Social Media Marketing*. [Online]

Available at: <https://insidebigdata.com/2018/10/06/4-major-ways-big-data-impacting-social-media-marketing/>

[Accessed 18 December 2018].

Nishtha, J. et al., 2012. A Survey and Comparison of Relational and Non-Relational Database. *International Journal of Engineering Research & Technology (IJERT)*, I(6), pp. 1-5.

Noa, R.-H., Lior, R., Bracha, S. & Shoval, P., 2017. *Modeling Grap Database Schema*. s.l.:IEEE C omputer Society.

Olofson, C., 2011. *The Database Revolution*, New York: IBM Corporation.

Oussous, A., Benjelloun, F.-Z., Lahce, A. A. & Belfkih, S., 2017. Big Data technologies: A survey. *Journal of King Saud University- Computer and Information Sciences*, XXX(XXX), pp. 1-18.

Pal, S., 2016. *SQL for Big Data- Technology, Architecture and Innovation*. 1st ed. Wilmington: Apress.

Parms, J., 2018. *6 Benefits of Data Modeling in the Age of Big Data*. [Online]

Available at: <https://dataflog.com/6-benefits-data-modeling-in-the-age-of-big-data/1479>

[Accessed 12 July 2018].

Pascal, D., 2018. *Data Modeling is Dead...Long Live Schema Design!*. [Online]

Available at: www.dataversity.net/data-modeling-dead-long-live-schema-design/

[Accessed 1 September 2018].

Pore, S. S. & Pawar, B. S., 2015. Comparative Study of SQL & NoSQL Databases. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, IV(5), pp. 1747-1753.

Ramachandran, M., 2014. *Relational Vs Non-Relational Databases-Part 1*. [Online]

Available at: bigdata-madesimple.com/relational-vs-non-relational-databases-part-1/

[Accessed 2 September 2018].

Ramez, E. & Shamkant, N. B., 2011. *Fundamentals of Databases Systems*. 6th ed. Boston: Addison-Wesley.

Recker, J., 2015. *Research on Conceptual Modelling: Less Known Knowns and More Unknown Unknowns, Please*. Sydney, Australian Computer Society, Inc, pp. 3-7.

Remya, P., 2013. *Adoption of Big Data Technology for the Development of Developing Countries*. Nesik, MET's Institute of Engineering BKC.

- Renear, A. H., Sacchi, S. & Wickett, K. M., 2018. *Definitions of Dataset in the Scientific and Technical Literature*. Pittsburgh, ASIST 2010.
- Robert, M. T., 2015. *NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database*. Denver, College of Computer & Information Sciences, Regis University, pp. 259-268.
- Robinson, S., 2013. *Conceptual Modeling for Simulation*. Loughborough, IEEE.
- Rob, K. & Lauriault, T., 2014. Small Data, Data Infrastructures and Big Data. *GeoJournal*, 80(4), pp. 1-27.
- Rossel, G. & Manna, A., 2017. A Modeling methodology for NoSQL Key-Value databases. *Database Systems Journal*, VIII(2), pp. 12-18.
- Rouse, M., 2018. *MySQL*. [Online]
Available at: <https://searchoracle.techtarget.com/definition/MySQL>
[Accessed 27 January 2019].
- Rurael, N. H., Shahane, N. M. & Bhamare, D. P., 2013. Learning from Small Data Set to Build Classification Model: A survey. *International Journal of Computer Applications* (0975 – 8887), I(1), pp. 23-26.
- Sadalage, J. P. & Martin, F., 2013. *NoSQL Distilled- A Brief Guide to the Emerging World of Polyglot Persistence*. 2nd ed. New Jersey: Pearson Education, Inc.
- Sareen, P. & Kumar, P., 2017. NoSQL Database and its Comparison With SQL Databases. *International Journal of Computer Science & Communication Networks*, V(5), pp. 293-298.
- Saxena, M., Ali, Z. & Singh, K. V., 2014. NoSQL Databases- Analysis, Techniques, and Classification. *Journal of Advanced Database Management & Systems*, I(2), pp. 1-12.
- Shahid, U., 2018. *SQL Server And Relational Database - Part One*. [Online]
Available at: <https://www.c-sharpcorner.com/article/sql-server-and-relational-database-part-one/>
[Accessed 25 January 2019].
- Singh, S., 2015. Database Management System. *Journal of Management Research and Analysis*, II(1), pp. 72-80.
- Smith, A. M., 2016. *Logical and Physical Data Modeling Overview*. [Online]
Available at: <https://www.ewsolutions.com/logical-physical-data-modeling-overview/>
[Accessed 21 December 2018].
- Smith, M. A., 2011. *Enterprise Data Modeling*. [Online]
Available at: <https://www.ewsolutions.com/enterprise-data-modeling/>
[Accessed 25 November 2018].

- Steve, H., 2015. *Data Modeling Made Simple*. 2nd ed. New York: Technics Publications, LLC.
- Stewart, R. et al., 2015. *Conceptual modeling: Definition, purpose and benefits*. Huntington, IEEE.
- Suri, P. & Sharma, M., 2011. A Comparative Study Between the Performance of Relational & Object Oriented Database in Data Warehousing. *International Journal of Database Management Systems (IJDBMS)*, III(2), pp. 116-127.
- Team, E., 2017. *Understanding the Hierarchical Database Model*. [Online] Available at: <https://mariadb.com/kb/en/library/understanding-the-hierarchical-database-model/> [Accessed 22 October 2018].
- Tole, A. A., 2013. Big Data Challenges. *Database Systems Journal*, IV(3), pp. 31-41.
- Vaish, G., 2013. *Getting Started with NoSQL*. 1st ed. Birmingham: Packt Publishing.
- Varga, V., János-Rancz, K. T. & Kálmán, B., 2016. Conceptual Design of Document NoSQL Database with Formal Concept Analysis. *Acta Polytechnica Hungarica*, XII(2), pp. 229-248.
- Vladimir, F., 2018. *Top 5 reasons to use the Apache Cassandra Database*. [Online] Available at: <https://towardsdatascience.com/top-5-reasons-to-use-the-apache-cassandra-database-d541c6448557> [Accessed 27 January 2019].
- Vozábal, M., 2016. *Master Thesis-Tools and Methods for Big Data Analysis*. 1 ed. Pilsen: University of West Bohemia.
- Warner, J., 2018. *Big Data, Big Impact: How Data Improves Your Social Media Marketing*. [Online] Available at: <https://www.business.com/articles/big-data-social-media-strategies/> [Accessed 24 October 2018].
- Watt, A., n.d. *Chapter 5 Data Modelling*. [Online] Available at: <https://opentextbc.ca/dbdesign01/chapter/chapter-5-data-modelling/> [Accessed 31 December 2018].
- West, M., Partridge, C. & Lycett, M., 2016. *Enterprise Data Modelling: Developing an Ontology-Based Framework for the Shell Downstream Business*, London: Research Gate.
- Wickham, H., 2014. Tidy Data. *Journal of Statistical Software*, 59(10), pp. 1-23.
- Woodford, C., 2018. *Digital Cameras*. [Online] Available at: <https://www.explainthatstuff.com/digitalcameras.html> [Accessed 20 December 2018].

APPENDICES

Appendix 1: Questionnaire

A Study of Conceptual Data Modelling in the Era of Big Data- A Case of Zambia

By Listone Kaputula (2016145535)

Master of Engineering in Information and Communication Technology (MEng-ICT)

For further details or any queries, kindly get in touch on 0950008002 or 0966241670

Dear Respondent,

I am a student at the University of Zambia in my final year pursuing a Master of Engineering in Information and Communication Technology (ICT). As partial fulfillment for the award of a Master's degree, I am conducting a baseline study on: **“A Study of Conceptual Data Modelling in the Era of Big Data -A Case of Zambia”**.

You have been purposively sampled to provide information for the topic indicated above. The information being collected is purely for academic purposes as such, it will be treated with maximum confidentiality. Subsequently, you are not supposed to indicate your name or any personal information that can lead to revealing of your identity.

Your co-operation will be greatly appreciated.

For more information or queries in relation to the survey, you may wish to contact the following:

Research Supervisor: Dr. Dani Banda (Brigadier General- ZAF) on 0977 882297 or

Assistant Dean: Dr. Mwanaumo on 0969 561 353

.....

Part A: Bio Data - Please tick in the box as appropriate

1) **Line of specialty**, please tick only on one box: (Tick one)

- | | |
|---------------------------------|--------------------------|
| a) Computer Science/Engineering | ☐ |
| b) IT | ☐ |
| c) Telecommunications | ☐ |
| d) Other | ☐ |

If others, indicate.....

2) **Professional certifications**, please tick only on one box: (Tick all that applies)

- | | |
|---|--------------------------|
| a) Oracle Certified Associate/Professional | ☐ |
| b) Microsoft Office Specialist Access Track | ☐ |
| c) IBM Database certified | ☐ |
| d) Others | ☐ |

If others, indicate.....

3) **What is your organisation's Primary industry? (Tick one)**

- | | | | |
|---------------------------------|--------------------------|-----------------|--------------------------|
| (a) Government Agency/Authority | ☐ | (b) Health Care | ☐ |
| (c) Telecommunications/ICT/ISP | ☐ | (d) Banking | ☐ |
| (e) Energy/Utility | ☐ | (f) Transport | ☐ |
| (g) Academia | ☐ | (h) Others | ☐ |

If others, indicate.....

4) **What is your primary role in the organisation? (Tick one)**

- | | |
|---------------------------------|--------------------------|
| a) Business Analyst | ☐ |
| b) Business User | ☐ |
| c) Database Administrator | ☐ |
| d) Data Modeler | ☐ |
| e) Technical Analyst | ☐ |
| f) Trainer | ☐ |
| g) Help desk agent / technician | ☐ |
| h) Application Developer | ☐ |
| i) others | ☐ |

If others, please specify.....

PART B: TECHNICAL (Tick the correct answer)

1. Conceptual data model helps to describe entity, entity types and entity relationships. Does your organisation apply conceptual modelling when describing its business objects or events?

a) Yes ☐ c) Not sure ☐
b) No ☐

2. Collection of data is called dataset. Do you collect data in your organisation for future use?

a) Yes ☐ c) Not Sure ☐
b) No ☐

3. Big Data is characterized by volume, velocity, variety, veracity and value. Which one(s) among the 5V of big data is considered to be an important aspect of Big Data in your organisation? (tick all that applies)

a) Volume ☐ d) Veracity ☐
b) Velocity ☐ e) Value ☐
c) Variety ☐ f) Others ☐

If others, indicate

4. Data Modelling is the first step in database design. Do you think, data modelling, is always necessary for database design?

a) Yes ☐ c) Sometimes ☐
b) No ☐

5. Which ones among the following are a type of databases you have worked with? (Tick all that applies)

a) MongoDB ☐ d) Neo4j ☐
b) MySQL ☐ e) Others ☐
c) Oracle Database ☐

If there are others, indicate:

6. Do the steps of data modelling (conceptual, logical and physical) seems to be slowing down or hinder the process of application development in in this digital transformation?

a) Yes ☐ c) Sometimes ☐
b) No ☐

7. NoSQL stands for No Only SQL. It is non-relational types databases developed to handle unstructured data. Have you or your organisation started using NoSQL?

a) Yes ☐ c) Plans underway ☐
b) No ☐

8. A conceptual data model helps in identifying high-level key business and system entities and establishing the relationships existing between them. In which ways does conceptual data models help? (Tick all that applies)

- a) Defining the key issues of problems which need to be addressed by the system
- b) To address both digital and non-digital concepts
- c) In closing the gaps between a solution model and requirements document
- d) Others

☐
☐
☐
☐

If there are other, kindly indicate:

.....

9. Facts that represent an object or event and can be recorded or stored in some way are called Data. What database(s) types are used to store data in your organisation? (Tick all that applies)

- a) Microsoft Access
- b) Oracle
- c) MySQL
- d) SQL Server
- e) MongoDB
- f) ArangoDB
- g) Other

☐
☐
☐
☐
☐
☐
☐

If others indicate:

10. what are some of the sources of Big Data in your working environment? (Tick all that applies)

- a) Sensor data e.g. biometric identity
- b) Social media
- c) Public data
- d) Transaction data
- e) Others

☐
☐
☐
☐
☐

If others, indicate:

11. Database is the collection of related data of one or more activities in an organisation. What type(s) of databases are used in your organisation? (Tick all that applies)

- a) Hierarchical database e.g. IMB IMS
- b) Network database e.g. Integrated Data Store (IDS)
- c) Object-Oriented Database e.g. ConceptBase
- d) Relational Database e.g. Oracle
- e) Non-Relational Database e.g. MongoDB

☐
☐
☐
☐
☐

12. Which of the following skills that a data modeler should possess? (Tick all that apply)

- a) Knowledge of the business

☐

- b) Knowledge of data modelling pattern ☐
- c) Knowing the line between ideal and practical ☐
- d) Knowing when to stop ☐
- e) Knowledge of DBMs ☐
- f) Documentation skills ☐
- g) Verbal communication skills ☐
- h) Knowledge of normalization ☐

13. Enterprise Data Modelling is the practice of creating a graphical model of the data used by an enterprise or company. Does your organisation have an Enterprise Data Model? ☐

- a) Not started ☐
- b) Yes, Completed ☐
- c) Partially, Completed ☐

14. If your organisation have Enterprise Data Model, what is the purpose of that model? (Tick all that applies)

- a) Supporting applications integration ☐
- b) Providing starting point for application modelling ☐
- c) Information system planning ☐
- d) Business intelligence ☐
- e) Others ☐

If others indicate:

15. How well is this goal achieved using the Enterprise Data Model?

For each row, please tick only one box. To suggest other big data impacts, please use the space provided.

	Very Well	Somewhat	Not at all	Do not know
Supporting applications integration	☐	☐	☐	☐
Providing starting point for application modelling	☐	☐	☐	☐
Information system planning	☐	☐	☐	☐
Business intelligence	☐	☐	☐	☐
Others	☐	☐	☐	☐

If others indicate:

16. The following are some of the data modelling tools in use, which one(s) does your organisation uses (Tick all that apply)

- | | |
|--------------------------------------|--------------------------|
| a) ER/Studio | ☐ |
| b) Oracle SQL Developer Data Modeler | ☐ |
| c) CA Erwin | ☐ |
| d) InfoSphere Data Architect | ☐ |
| e) MySQL Workbench | ☐ |
| f) Others | ☐ |

If others indicate:

17. Is your organisation using non-relational database technologies such as NoSQL? Which one is it

- | | | | |
|-----------------------|--------------------------|---------------------|--------------------------|
| a) No, not started | ☐ | d) Yes, it is Neo4j | ☐ |
| b) Yes, it is MongoDB | ☐ | e) Yes, others | ☐ |
| c) Yes, it is Redis | ☐ | | |

If others indicate here

18. In your organisation at what level of data modelling does the modeler understand and capture business knowledge from data perspectives?

- | | | | |
|------------------------|--------------------------|--------------------------|--------------------------|
| a) Physical data model | ☐ | c) Conceptual data model | ☐ |
| b) Logical data model | ☐ | | |

19. In your organisation, data about employees or clients whose volume and format is accessible, informative and actionable is considered to be

- | | | | |
|---------------|--------------------------|---------|--------------------------|
| a) Small data | ☐ | c) Both | ☐ |
| b) Big data | ☐ | | |

20. Data captured at toll gates and speed camera contributes to Big Data in Zambia.

- | | | | |
|----------|--------------------------|-------------|--------------------------|
| a) True | ☐ | c) Not sure | ☐ |
| b) False | ☐ | | |

21. The four (4) basic types of NoSQL database includes Key-value, column stores, graph store and document-based stores. If you answered "YES" to Q17, which one(s) is commonly used in your organisation (Tick all the applies)

☐

- | | | | |
|--------------------|--------------------------|--------------------------|--------------------------|
| a) Key-value store | ☐ | c) Graph stores | |
| b) Column stores | ☐ | d) Document-based stores | ☐ |

22. Conceptual modelling displays data independent of the database technology

- | | | | |
|---------|--------------------------|----------|--------------------------|
| a) True | ☐ | b) False | ☐ |
|---------|--------------------------|----------|--------------------------|

23. Big Data is data that exceeds processing capability of relational databases. In your organisation, what size of data is considered to be Big Data:

- | | |
|---|--------------------------|
| a) data bigger than 1TB but less than 1PB | ☐ |
| b) Data whose volume exceeds 1PB | ☐ |
| c) Any data stored in database | ☐ |

Note that 1Gigabytes(GB)= 1×10^9 Bytes; 1Terabytes(TB) = 1×10^{12} Bytes;
1Petabytes(PB)= 1×10^{15} Bytes, etc.

24. Conceptual data models provide big picture for better communication between the modeler and the business stakeholder

- | | | | |
|----------|--------------------------|-------------|--------------------------|
| a) True | ☐ | c) Not sure | ☐ |
| b) False | ☐ | | |

25. There are three types of data models (conceptual, logical and physical) used in relational database design, which one(s) is/are used in your organisation for non-relational (NoSQL) database technologies?

- | | | | |
|---------------|--------------------------|----------------------|--------------------------|
| a) Conceptual | ☐ | c) Physical | ☐ |
| b) Logical | ☐ | d) None of the above | ☐ |

26. During the stage of conceptual modelling, it is important to involve the business owner.

- | | | | |
|-------------------------|--------------------------|---------------|--------------------------|
| a) Yes, it is important | ☐ | c) Not at all | ☐ |
| b) Sometimes | ☐ | | |

27. Datasets are in different forms which includes structured, semi-structured and unstructured. What type(s) or format(s) of dataset do you handle in your organisation?

- | | | | |
|-----------------------------|--------------------------|--------------------------|--------------------------|
| a) structured datasets | ☐ | c) Unstructured datasets | ☐ |
| b) Semi-structured datasets | ☐ | d) Not sure | ☐ |

28. Does your organisation work on Big Data?

a) Yes

☐

c) Not sure

☐

b) No

☐

29. If your answer to question 28 was “Yes”, which areas is Big Data used in your organisation? (Tick all that apply)

a) Marketing

☐

b) Business intelligent

☐

c) Analysis on customer feedback

☐

d) Others

☐

If other indicate:

30. There are generally two sources of data which can be used in data modelling.

These includes primary (data collected by researcher direct from respondents) and secondary source (data from books or journals). Are two the sources of your organisation work on?

a) Yes

☐

c) Not sure

☐

b) No

☐

31. Big Data is data that exceeds processing capability of relational databases in your understanding, what size of data can be considered as a Big Data:

a) Data bigger than 1TB but less than 1PB

☐

b) Data whose volume exceeds 1PB

☐

c) Any data stored in database

☐

Note that 1Gigabytes(GB)= 1×10^9 Bytes; 1Terabytes(TB) = 1×10^{12} Bytes;

1Petabytes(PB)= 1×10^{15} Bytes, etc.

32. Both relational and non-relational databases are available as an open source or licensed software. Which one does your organisation use?

a) Open source

☐

c) Both

☐

b) licensed

☐

33. Which one(s) of the following is/are a type of databases are used in your organisation?(Tick all that applies)

a) MongoDB

☐

d) Neo4j

☐

b) MySQL

e) Others

☐

c) Oracle Database

☐

If there are others, indicate:

34. Is there data that is important to your organisation but it is not in a database because they can't fit in the exist database?

a) Yes

☐

b) No

☐

**Thank you very much for contributing to my Research
YaHUaH barak you!**

Appendix 2: Bio Data Questions coding table

Qsn	Bio Data Variables	Value	code
1	Respondent's Line of Specialty	Computer Science/Eng.	1
		IT	2
		Telecommunications	3
		Others	4
2	Respondent's Professional Certifications	Oracle Certified	1
		MS Certified	2
		IBM Database Certified	3
		Others	4
3	Organisation's Primary Industry	Government Agency/Authority	1
		Health Care	2
		Telecoms/ICT/ISP	3
		Banking	4
		Energy/Utility	5
		Transport	6
		Academia	7
		Other	8
4	Respondent's Primary Role in the Organisation	Business Analyst	1
		Business User	2
		Database Administrator	3
		Data Modeler	4
		Technical Analyst	5
		Trainer/Lecturer	6
		Help Desk Agent/Technician	7
		Application Developer	8
		Other	9

Appendix 3: Technical Data Questions coding table

Qsn	Technical Data Variables	Value	Code
1	Organisation's use of conceptual modelling to describe business objects or events	Yes	1
		No	2
		Not Sure	3
2	Collection of data in organisation for future use	Yes	1
		No	2
		Not Sure	3
3	Important aspect of Big Data (from the five characteristics of Big Data) in respondent's organisations	Volume	1
		Velocity	2
		Variety	3
		Veracity	4
		Value	5
4	Necessity of data modelling in database design	Yes	1
		No	2
		Sometimes	3
5	Database the respondent is familiar with or work on before	MongoDB	1
		MySQL	2
		Oracle Database	3
		Neo4j	4
		Others	5
6	Traditional data modelling steps (conceptual, logical and physical) slowdown or hinder process of application development	Yes	1
		No	2
		Sometimes	3
7	Respondent's organisation started using NoSQL	Yes	1
		No	2
		Plans underway	3
8	Conceptual Models helps in identifying high-level key business and system entities by	Defines the key issues to be addressed	1
		Address digital concepts	2

		Close gap btn solution model and requirements	3
		Other	4
9	Databases used by respondent's organisation to store data	Microsoft Access	
		Oracle	1
		MySQL	2
		SQL Server	3
		MongoDB	4
		ArangoDB	5
		Other	6
10	Sources of Big Data in respondent's working environment	Sensor data e.g. biometry identity	1
		Social media	2
		Public data	3
		Transaction data	4
		other	5
11	Types of databases (in terms of evolution) used in respondent's organisation	Hierarchical Database	1
		Network Database	2
		Object-Oriented Database	3
		Relational Database	4
		Non-Relational Database	5
12	Skills for a data modeler	Knowledge of business	1
		Knowledge of data modelling	2
		Knowing btn ideal and practical	3
		Knowing when to stop	4
		Knowledge of DBMS	5
		Documentation skills	6
		Verbal communication skills	7
		Knowledge of normalization	8
13	Respondent's organisations with Enterprise Data Model	Not started	1
		Yes and completed	2

		Partially completed	3
14	The purpose of Enterprise Data Model in organisations (for those which have EDM)	Supporting application integration	1
		Starting point for application modelling	2
		Information system planning	3
		Business intelligence	4
		others	5
15	Achievement of EDM in each of Q14 purpose	Very well	1
		Somewhat	2
		Not at all	3
		Do not know	4
16	Data modelling tools in use in respondent's organisations	ER/Studio	1
		Oracle SQL Developer	2
		CA Erwin	3
		Infosphere Data Architect	4
		MySQL Workbench	5
		Other	6
17	Organisations using non-relational databases with an example	No, not started	1
		Yes, it is MongoDB	2
		Yes, it is Redis	3
		Yes, it is Neo4j	4
		Yes, other	5
18	Level in data modelling where modeler understand and capture business knowledge from data perspectives	Physical data model	1
		Logical data model	2
		Conceptual data model	3
19	Data about employees or clients whose volume and format is accessible, informative and actionable is considered to be...	Small data	1
		Big Data	2
		Both	3
20		True	1

	Data captured at toll gates and by speed camera contribute to Big Data in Zambia	False	2
		Not sure	3
21	Types of NoSQL used in respondent's organisation (to those who have started)	Key-value stores	1
		Column stores	2
		Graph stores	3
		Document-based stores	4
22	Conceptual modelling displays data independent of the database technology	True	1
		False	2
23	Size of datasets (in bytes) organisations consider to be Big Data	Dataset btm 1TB and 1PB	1
		Dataset bigger than 1PB	2
		Any data stored in database	3
24	Conceptual data models enhance communication between modeler and business stakeholder	True	1
		False	2
		Nor sure	3
25	Data modelling level(s) used by respondent's organisation (those that have started) for non-relational (NoSQL) database technology	Conceptual	1
		Logical	2
		Physical	3
		None of the above	4
26	Importance of engaging/involving the business owner during conceptual modelling	Yes, it is important	1
		Sometimes	2
		Not at all	3
27	Dataset format handled in respondent's organisations	Structured datasets	1
		Semi-structured dataset	2
		Unstructured dataset	3
		Not sure	4
28	Respondent's organisation working on Big Data	Yes	1
		No	2
		Not sure	3

29	Those working on Big Data, what area(s) is it	Marketing	1
		Business intelligent	2
		Analysis of customer feedback	3
		other	4
30	Sources of data respondent's works on in their organisation are primary and secondary sources	Yes	1
		No	2
		Not sure	3
31	Respondent's definition of Big Data in terms of size in bytes	Dataset btn 1TB and 1PB	1
		Dataset bigger than 1PB	2
		Any data stored in database	3
32	Respondent's organisations use	Open source databases	1
		Licensed databases	2
		Both of the above	3
33	Databases used in the respondent's organisations	MongoDB	1
		MySQL	2
		Oracle	3
		Neo4j	4
		Other	5
34	Availability of important data in the organisation stored because it cannot fit in the existing database	Yes	1
		No	2
		Not sure	3

Appendix 4: Journal Published

Listone Kaputula and Dani Eliya Banda, "A study of Conceptual Data Modelling in the Era of Big Data: A Case of Zambia", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 8, Issue 2, February 2019.

[Online]. Available:

http://www.ijirset.com/upload/2019/february/9_A%20Study.pdf