ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

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

Listone Kaputula [1], Dani Eliya Banda [2]

P.G Student, Department of Electrical and Electronic Engineering, University of Zambia, Lusaka, Zambia[1]

Lecturer, Department of Electrical and Electronic Engineering, University of Zambia, Lusaka, Zambia[2]

ABSTRACT: The way data is collected and stored has evolved over time due to various developments which has 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 been accommodated in the databases. These new technologies have 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, strong 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.

KEYWORDS: Big Data, Conceptual data modelling, Database, Dataset, NoSQL.

I. INTRODUCTION

In computing, data is defined as known facts that can be recorded and that have implicit meaning [1]. The collection and storage data has over time gone through changes. With more and more developments in database technologies, dataset has also evolved from structured to semi-structured and finally to the unstructured data [2]. By structured data, it means data is well defined, can be relatively be searched, group or sorted in accordance to certain criteria [3]. In semi-structured data, some data are predefined while in unstructured ones, predefined format is not obvious.

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. Thus, Big Data is defined as volume, velocity and variety of data which is beyond today's norm [4]. 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 [5].

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 [6]. 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 [7]. 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**

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

consistency indicates that when nodes are added data is replicated [8]. 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. In this research, the focus was on conceptual data modelling and how it is employed in non-relational databases. Conceptual data model describes the physical and social aspect of the World around us.

II. LITERATURE REVIEW

Dataset is the collection of values either as numbers or strings [9]. These numbers and strings are known as data. 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, organized, structured or presented so as to make them meaningful or useful, they are called information.

It was not common to refer to data as either "small" or "big" before 2008 [10][11] 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. Laney considered dataset to be Big Data when it possesses the three data characteristics commonly known as 3V such as Volume, Velocity and Variety[12]. 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.

Sources of data is categorized into three basic dimensions which includes machine, human interaction and data processing.

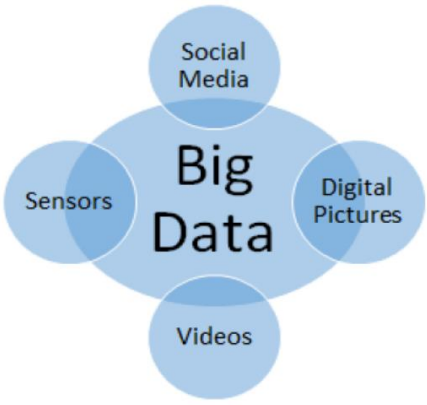

Figure 1: Big Data Sources [13]

Figure 1 shows some sources of Big Data which social media, sensors, videos and digital pictures.

Database is the collection of data with activities of one or more related to an organization. A software designed to manage and utilize the collection of data is called Database Management System (DBMS). To get to where it is today, database has gone through various evolution. The year 1958 is an important year for data. In this year IBM Corporations of United States, based in New York produced a system file that could store and retrieve data [14]. This system was called Format File System (FFS) and was very instrumental to the introduction of automated database management. FFS is referred to be the first generation of database management system.Another IBM engineer named Donald Chamberlin led a team that developed Structured English Query Language (SEQUEL) in 1974 [14]. Three years later, Chamberlin with his colleague Robert Boyce changed SEQUEL to SQL (Structured Query Language). This relational database

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

dominated the field of database management system for period of time. Keith D. Foote said that 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 [15]. This led to the development of a non-relational database called NoSQL. NoSQL means Not Only SQL.

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. Pascal suggested [16] 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.

III. ANALYSIS AND DISCUSSION OF THE 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"[17].

The second paper which was analysed was authored by four authors namely FatmaAbdelhedi, Amal A. Brahim, FatenAtigui 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 group into the same column-family with each row describing as an instance[18].

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[19].

MohanadHalaweh and Ahmed E. Massry proposed a six indicator conceptual model to successfully implement Big Data in an organization. The indicators included Top Management support, Organization Change, Data availability and quality, Infrastructure, Require skills, and Privacy and Security[20].

Noa Roy-Hubara, LiorRokach and BrachaShapira 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 did 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[21].

The last paper which was analyzed 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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

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[22].

IV. METHODOLOGY

The study was carried out both a survey questionnaire 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.

Figure 2 below 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.

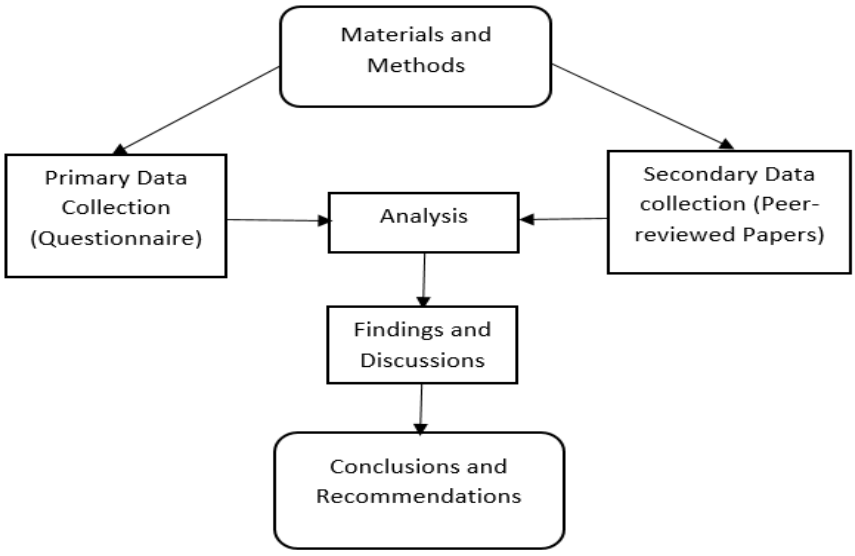

Figure 2 Methodology Approach

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 questionnaires were distributed to forty (40) database users and experts from 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. The researcher applied spreadsheet software known as Microsoft Excel to analyse the statistical values.

To further address the objectives of this study, the researcher surveyed literature between 2010 and mid-2018. The criterial 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 modeling for non-relational databases prior to 2010. Beside the choice of the period of publication, the other two inclusions criterial used were: 1) English published

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

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.

V. RESEARCH FINDINGS

Data Modeller Skills

Table 1 below shows the skills for Data Modellers. Respondents said that all skills in table apart from the knowledge of normalization are important to a Data Modeller.

Table 1: Data Modeller Skills

Modeller 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%**	

Application of Conceptual Data Modelling

As shown in figure 3, organisations use conceptual data modelling to define key issues of problems which needs to be addressed, and closing gaps between a solution model and document requirements. Conceptual data modelling also helps them to address both digital and non-digital concepts.

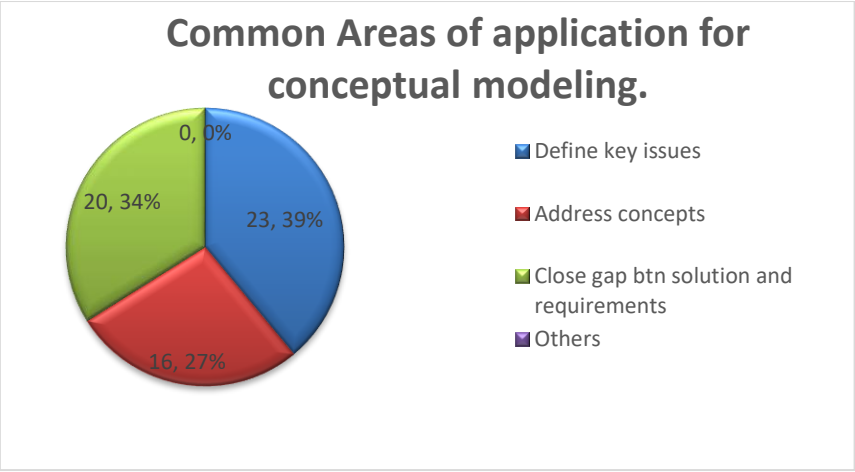

Figure 3. Areas of Application for Conceptual Modeling

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

Database used in local Organisation

The figure below shows the type of databases most organisations use to store data.

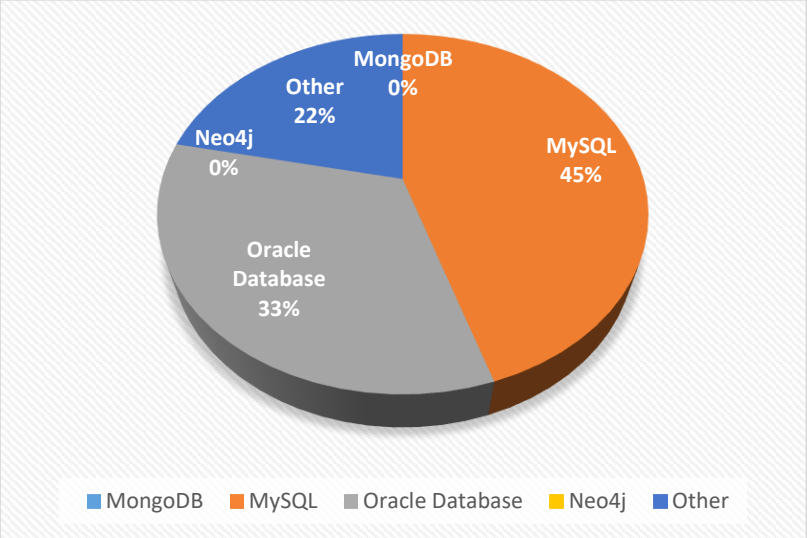

Figure 4. The types of databases in use in local organisations

MySQL is the most popular database in use, which is follow Oracle Database and relational databases. The study, as seen in figure 4, showed that no organisation uses non-relational databases.

Dataset Stored in Organisations

Below is figure 5 showing the types of datasets stored in organisation. The results show that most organisations store structured datasets and a bit of semi-structured.

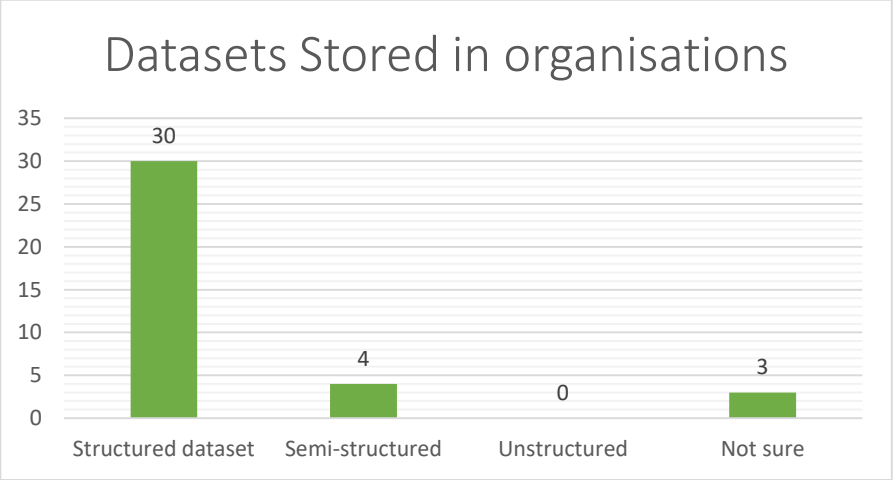

Figure 5. Dataset types stored in organisation's Databases

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

VI. DISCUSSIONS

A. ***Data Modelling Skills***

In the study, it was found that the knowledge of database management system is a key skill for data modelling. Other skills which attracted more than 90% response includes the knowledge of 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 also noted that the skills of knowing when to stop and verbal communication to be essential.

B. ***Area of application for Conceptual Data Modelling***

The survey revealed that organisations use conceptual modelling mostly to define key issues of problems which need to be addressed by the system (39%). Other areas where conceptual data modelling is applied is in addressing digital and non-digital concepts (27%) and closing the gap between a solution model and document requirements (34%). This shows that most organisations are capable of handling Big Data when a right approach for conceptual data modelling for non-relational database is in place. Thus, with an approach for non-relational database systems in place, engaging the business owners for Big Data conceptual data modelling can provide the modellers with a big picture for better communication.

C. ***Types of Database used in Local Organisations***

MySQL database (at 45%) was found to be most popular database in use and this followed by Oracle (33%) and Microsoft SQL server (22%). This confirms to the responses obtained when respondents were asked about the categories of database, where just about a quarter use the licensed while the rest uses either open source database applications or both. It was observed from the survey that neither an individual nor organization are using or have worked on non-relational databases. This entails that the deployment of non-relational databases is still at infancy or it has not commenced.

D. ***Dataset Stored in Local Organisation***

When it comes to storing of dataset, all organisations are fully engaged in the structured ones while only about 7% collect semi-structured data and none for unstructured. The reluctant by our local firms on collecting unstructured data may be attributed to lack of conceptual data model approach for these types of dataset which comes in voluminous, varieties and at high speed.

VII. CONCLUSION

The study revealed that most of organisations in Zambia use relational databases for their data storage; and have limited or no knowledge of NoSQL databases. The insufficient knowledge in non-relational databases may be attributed to lack of skills for creating conceptual data modeling. Further, in the study, it was found that conceptual data modeling is mainly used to define key issues of problems which need to be addressed; and closing the gap between a solution model and document requirements. Conceptual data modeling is also applied in addressing digital and non-digital concepts of an organisation. From the analysis of peer reviewed papers it was found that most their research works involved the use of a certain relational conceptual schema to a particular NoSQL logical model. For example, a graph database schema was proposed by [21] through the transformation of entity-relational diagram (ERD) conceptual schema which involved the process of mapping ERD to a NoSQL graph database.

REFERENCES

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

[1] E. Ramez and N. B. Shamkant, Fundamentals of Databases Systems, 6th ed., Boston: Addison-Wesley, 2011.
[2] J. Hurwitz, A. Nugent, F. Halper and K. Marcia, "Unstructured Data in a Big Data Environment," 2018. [Online]. Available: https://www.dummies.com/programming/big-data/engineering/unstructured-data-in-a-big-data-environment/. [Accessed 17 12 2018].
[3] K. K.V and V. M, "Unstructured Data Analysis-A Survey," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 3, pp. 223-225, 2015.
[4] M. Cameron, "Big Data The definitive guide to the revolution in business analytics," Fujitsu Services Ltd, London, 2012.
[5] D. Edd, Big Data Now, 1st ed., Beijing: O'Reilly Media, 2012.
[6] M. T. Robert, "NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database," in Proceedings of Informing Science & IT Education Conference (InSITE), Denver, 2015.
[7] Z. Y. Abert and S. Sherif, Handbook of Big Data Technologies, 1st ed., Sydney: Springer International Publishing, 2017.
[8] C. D. Ganesh, "Base Analysis of NoSQL database," Elsevier, vol. I, no. 1, pp. 13-21, 2015.
[9] H. Wickham, "Tyde Data," Journal of Statistical Software, vol. 59, no. 10, pp. 1-23, 2014.
[10] R. Kitchin and T. Lauriault, "Small data in the era of big data," GeoJournal, vol. II, no. 3, pp. 463-475, 2015.
[11] R. Kitchin and P. T. Lauriault, "Small data, data infrastructures and big data," The Programmable City, County Kildare, 2015.
[12] T. Erl, W. Khattak and P. Buhler, Big Data Fundamentals- Concepts, Drivers & Techniques, 2nd ed., Boston: Arcitura Education Inc, 2016.
[13] K. A. I. Hammad, M. A. I. Fakharaldien, J. M. Zain and M. A. Majid, "Big Data Analysis and Storage," Orlando, 2015.
[14] C. Olofson, "The Database Revolution," IBM Corporation, New York, 2011.
[15] D. K. Foote, "A brief History of Non-Relational Databases," 2018. [Online]. Available: www.dataversity.net/a-brief-history-ot-non-relational-databases/. [Accessed 2 September 2018].
[16] D. Pascal, "Data Modeling is Dead...Long Live Schema Design!," 2018. [Online]. Available: www.dataversity,net/data-modeling-dead-long-live-schema-design/. [Accessed 1 September 2018].
[17] G. Rossel and A. Manna, "A Modeling methodology for NoSQL Key-Value databases," Database Systems Journal, vol. VIII, no. 2, pp. 12-18, 2017.
[18] F. Abdelhedi, A. B. Amal, F. Atigui and Z. Gilles, "Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems?.," in In Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Paris, 2016.
[19] E. W. David, Stephen and L. W, Big Data—Conceptual Modeling to the Rescue, Utah: Brigham Young University, 2013.
[20] H. Mohanad and E. M. Ahmed, "Conceptual Model for Successful Implementation of Big Data in Organizations," Journal of International Technology and Information Management, vol. XXIV, no. 2, pp. 21-34, 2015.
[21] R.-H. Noa, R. Lior, S. Bracha and P. Shoval, Modeling Graph Database Schema, IEEE Computer Society, 2017.
[22] S. Kwangchul, H. Chulhyun and J. Hoekyung, "NoSQL database design using UML conceptual data model based on Peter Chen," in International Journal of Applied Engineering Research ISSN 0973-4562, 2017.