OPEN ACCESS

Research Article

Big data investment and knowledge integration in academic libraries

Saher Manaseer[1,*] Afnan R. Alawneh[2], Dua Asoudi[2]

[1] Department of Computer Science, The University of Jordan, Amman, Jordan.

[2] Department of Business Administration, The University of Jordan, Amman, Jordan

[*] Email: saher@ju.edu.jo

Abstract

Recently, big data investment has become important for organizations, especially with the fast growth of data following the huge expansion in the usage of social media applications, and websites. Many organizations depend on extracting and reaching the needed reports and statistics. As the investments on big data and its storage have become major challenges for organizations, many technologies and methods have been developed to tackle those challenges.

One of such technologies is Hadoop, a framework that is used to divide big data into packages and distribute those packages through nodes to be processed, consuming less cost than the traditional storage and processing methods. Moreover, it allows an organization to store all its data, unlike the traditional methods of selecting and discarding some data.

In this study, researchers investigate the investment of Hadoop in managing the academic libraries' big data, focusing on three Vs (velocity, variety, and volume).

The studied population is academic libraries in Jordanian universities, and the results show how Hadoop framework using the map/reduce technique can be used to manage the big data of such libraries.

Keywords:

Big data, knowledge integration, Hadoop, libraries, data management

مجلة دراسات المعلومات والتكنولوجيا
جمعية المكتبات المتخصصة
فرع الخليج العربي
JIST - SLA - AGC

1. INTRODUCTION

In the field of data storage, a new paradigm regarding innovation, huge development, and growth has recently emerged. According to McKinsey Global Institute report in 2011, big data has evolved and is an essential part of all life aspects.[1, 2]

Due to the enhancement and improvement in the ICT sectors and in big data applications, huge volumes of big data have emerged.[2] The traditional data mining techniques are no longer adequate to process such huge data size and extract the needed information.[2, 3]

Big data is mainly discussed in terms of three Vs (volume, variety, and velocity) and the information extraction and analysis techniques.[3, 4] Cloud computing represents an option with less cost that may help in efficient analysis of big data.

Many software and solutions have been developed to manage and analyze big data. This study

http://doi.org/10.5339/jist.2019.3

© 2019 The Author(s), licensee HBKU Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution license CC BY 4.0, which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

كيوساينس
QSCIENCE
دار جامعة حمد بن خليفة للنشر
HAMAD BIN KHALIFA UNIVERSITY PRESS

Cite this article as: Manaseer S, Alwaneh A, Asoudi D. Big data investment and knowledge integration In academic libraries, Journal of Information Studies and Technology 2019:1,3. https://doi.org/10.5339/jist.2019.3

investigates the investment of Hadoop in managing academic libraries' big data, while focusing on four Vs (velocity, variety, veracity and volume).

2. RELATED WORK (LITERATURE REVIEW)

The rapid growth of online data such as those of WWW, the Internet, social media, scientific and engineering applications, graph processing, machine learning, and behavioral simulations[5] has increased the need of huge storage tools and huge data processing algorithms or operations.

Consequently, big data has been defined as a large-scale data with size in orders of terabyte, petabyte, exabyte, zettabyte, or even yottabyte. The following four "Vs" are used to describe big data: volume: the scale of the data; velocity: analysis of streaming data; veracity: uncertainty of data; and finally variety: the different forms of data. Furthermore, new challenges have emerged related to the relational databases for processing big data. In the relational database tables, relationships need to be built, which consumes huge time and efforts.[2, 3, 6]

Big data control and analysis have been considered as significant challenges in the data management field. Organizations need to efficiently analyze data and bring out available opportunities from this data in order to gain a sustained competitive advantage. Data mining is the process of searching through a large amount of data. The data mining techniques should be sufficient to extract the valuable information from the large volumes of data.[2, 6]

Challenges in applications for managing big data are mainly on how to extract the exact needed data from huge volumes.[7] One of the main challenges involved with big data is the management of such huge and tremendous size and amount of collected information in a way that guarantees efficient use of software and hardware (minimal requirements needed) while optimizing process expenses. In addition, one of the main targets of big data analysis is to produce manageable, easily accessible, and secured dataset.[8, 9]

In such giant volumes of data, errors, noises, incomplete data may occur or be collected. To ensure quality information is extracted from this big data, a cleaning process should be conducted, and the reliable parts should be verified. The complexity of big data is represented by its velocity, volume, and variety. Managing this triangle has been considered the main challenge in big data issues.[10]

Hadoop is a software that was developed to solve the problems of the traditional big data processing methods, especially the low performance and high complexity. Hadoop's significant advantage is its rapid capacity to process large amounts of data, which is done through parallel clusters. Another important advantage is that Hadoop performs the processing and computations where the data are stored in contrast to the traditional methods that consume more space, as the files are copied and executed in different spaces; thus, Hadoop reduces communications load in the network.[11,12]

Another property that makes Hadoop a powerful big data analysis tool is that it has two main subcomponents: Hadoop distributed file system (HDFS) and MapReduce framework. Moreover, Hadoop enables users to add modules according to their requirements.[7, 13]

2.1. Hadoop distributed file system

The HDFS is a storage system that helps to save storage costs and increase the reliability of the storage capabilities. The ability to hold tremendous volumes of data is the main significant advantage of this system. Another main advantage of this system is the portability of heterogeneous equipment and requirements of hardware and software. Moreover, it helps to reduce the congestion and load of the network, thus enhancing the overall system performance. This system is developed by having two kinds of nodes (master and slave[14]).

The master node is called the NameNode and is responsible for managing the operations of the system files and for saving users' data in blocks, where each block's size is equal to 64 megabytes, and all the blocks are replicate three times in different DataNodes to preserve the data in case one node fails. The Hadoop platform splits each DataNode into racks. The DataNode periodically sends a report about the data and also sends a heartbeat to the NameNode. The DataNodes are heterogeneous, which means each DataNode has different characteristics from the others in various aspects, such as memory size, disk space, and processor capability. Moreover, each DataNode can be located in a different place.

The slave DataNode is responsible for coordinating data storage for each compute nodes. It controls the DataNodes since all the information about each DataNode is saved to the NameNode. If any user wants to retrieve some information, the NameNode retrieves the data. There is another type of NameNode known as the secondary NameNode. It is a copy of the NameNode and is automatically used if the NameNode fails.[15, 16]

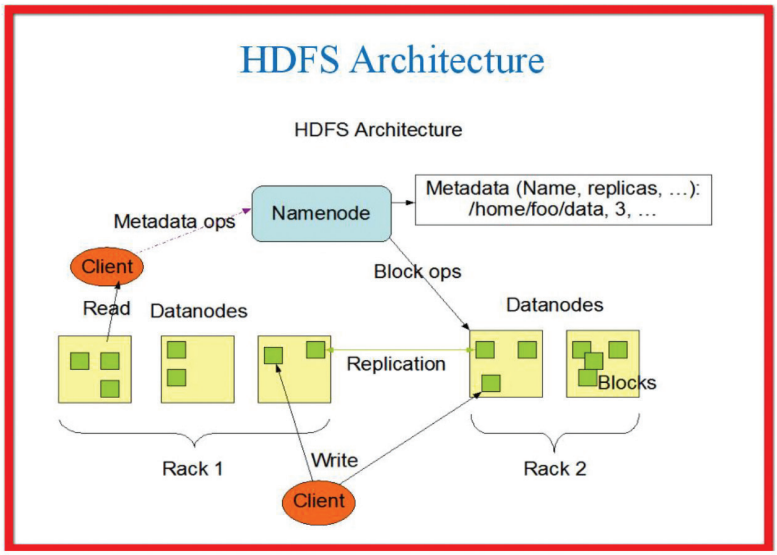

Figure 1. HDFS Architecture.

2.2. The MapReduce framework

According to Mall,[16] the MapReduce framework is considered as one of the first developed models and tools for big data management. What makes this model significant and popular is that it has an efficient and cost-minimizing mechanism and simplifies the processing of massive and huge data volumes. It also supports parallel processing through two main computations functions: Map function and Reduce function. MapReduce is an algorithm that is used for analytical purposes. Generally, the MapReduce algorithm consists of two phases: the map phase, also known as mapper phase, and the reduce phase or reducer phase. The main purpose of the mapper is to gather similar data as key/value pairs. However, to save time, many mappers work in parallel. The reducer phase is started after shuffling and sorting the data. From a particular perspective, sequentially processing the data takes a long time.[17, 18, 19]

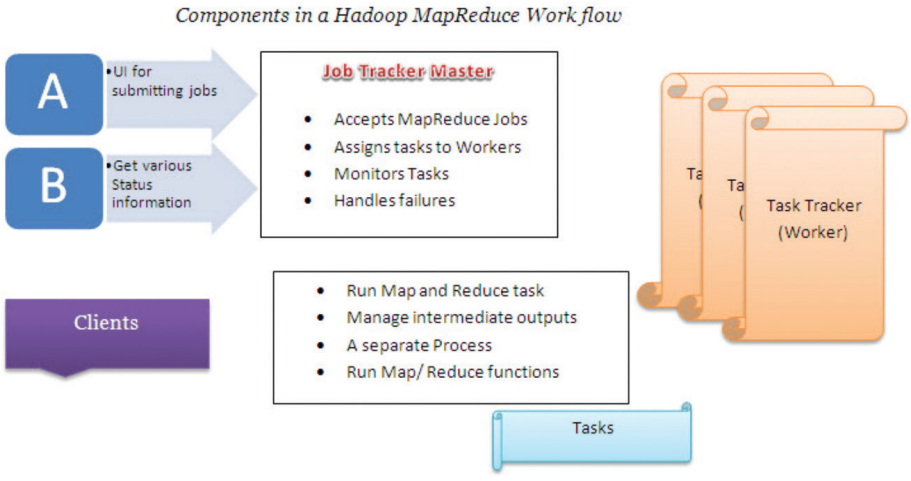

Figure 2. Components of Hadoop MapReduce Algorithm.

2.3. Importance of analyzing big data

Analyzing and managing big data has gained importance and recently attracted the focus of researchers. Some benefits gained by the analysis and efficient management of big data are as follows:[20]

- It helps in decision making.
- It attracts Investors.
- It brings satisfaction.
- It gives organizations a competitive advantage.
- It makes organizations abreast with the ongoing news all over the world.
- It broadens organizations' perspectives toward any issue by analyzing the data of social media.

2.4. Hadoop in libraries

The **applicability** of big data in libraries, especially academic ones, mainly lies in the following concepts:[21]

Resources

The resources of libraries vary; they include not only academic resources but also financial resources, budget, staff details, social media pages of the library, and any other info about the library. Moreover, the academic content and the catalogue of a library are considered resources for data and form a big data store, which needs to be properly managed and analyzed to help users or decision makers reach the needed information.

Gaining a competitive advantage through benchmarking

An academic library, in its strategy, should assign importance to the process of benchmarking and compare the results with those of other libraries, especially in regard to the e-content that a library offers, in order to remain in the field of competition.

React and predict

The data collected and analyzed should give librarians a clear insight about the needs of the libraries' patrons; moreover, it can help in better strategic planning for the next level.

Service enrichment

The data may provide insights on the problems that patrons face, their comments, and needs, which can help solve many key issues and satisfy the patrons, thereby enhancing the services provided by the library to its patrons.

3. FINANCIAL BENEFITS OF USING HADOOP

Hadoop uses a distributed file system, and the processors are distributed on the same space that contains the data; moreover, Hadoop is also capable of parallel processing, which makes its data processors faster than those of the traditional methods.[22]

In terms of financial benefits, it has been estimated that operating Hadoop allows organizations to store their data using 300 TB for over three years with costs of around 1.05 million dollars.[23] At the same time, Oracle has sold a database with over 168 TB in returns of 2.33 million dollars, not including the operating expenses.[24]

This reveals that Hadoop saves costs and storage, as well as provides faster data processing. It also allows using many models for storing and analyzing data.

4. METHODOLOGY AND TOOLS

MapReduce works by splitting the data file into smaller files and then processing in parallel[25]. Each map processes a small amount of data. However, the output data calls intermediate records. These records are key/value pairs, whereby the key is the store name and the values are the total sales of each store name. After this, a phase called shuffle and sort will start. Shuffle involves moving the data from all the mappers into the reducer, and sort involves putting all the keys together. Finally, at the same time, each reducer deals with one key and all its values. The reducer processes those values; for example, the reducer may sum all the total sales of one store. Hadoop uses two controllers: job-tracker and task-tracker. The job-tracker locator is inside the NameNode. It generates

the map tasks, splits the task, and finds the suitable task-tracker. The task-tracker location should be close to the data location.[26, 27]

The Map Code

```
import sys
import string
import hashlib

while True:
  line = sys.stdin.readline()
  if not line:
    break

  line = string.strip(line, "\n ")
  sale, cost = string.split(line, "\t")
  print "\t".join([sale, cost])
```

The Reduce Code

```
import sys
import string
import hashlib


        salestotal=0
        oldkey=None

        for line in sys.stdin:
                Data=line.strip().split("\t")
                if len(Data)!= 2:
                        continue
                thiskey, thissale = Data

                if oldkey and oldkey!= thiskey:
                        print "{0}\t{1}".format (thiskey,salestotal)
                        salestotal=0

                oldkey = thiskey
                salestotal += float (thissale)

        if oldkey!= None:
                print "{0}\t{1}".format (oldkey,salestotal )
```

5. DISCUSSION

The code is explained in detail:

Example: Word-Count[28]

Counts occurrences of each word across different files

Two input files:

- File1: “hello world hello moon”
- File2: “goodbye world goodnight moon”

Three operations will happen:

Mapping, combination, and reduction

The results will be as follows:

Mapping

First map:	Second map:
‹ hello, 1 ›	‹ goodbye, 1 ›
‹ world, 1 ›	‹ world, 1 ›
‹ hello, 1 ›	‹ goodnight, 1 ›
‹ moon, 1 ›	‹ moon, 1 ›

Combination

First map:	Second map:
‹ moon, 1 ›	‹ goodbye, 1 ›
‹ world, 1 ›	‹ world, 1 ›
‹ hello, 2 ›	‹ goodnight, 1 ›
	‹ moon, 1 ›

Reduction

‹ goodbye, 1 ›

‹ goodnight, 1 ›

‹ moon, 2 ›

‹ world, 2 ›

‹ hello, 2 ›

6. CONCLUSIONS AND RECOMMENDATIONS

This paper aims at increasing the awareness of how big data management has become a serious issue and how it can affect the development and enhancement of an academic library’s services and processes.

Moreover, the paper shows that big data management is a crucial part in decision making and strategic planning of a library.

The study investigates the use of Hadoop as an analytical tool to effectively store, analyze, and manage big data resources, with advantages of reduced cost, relatively low storage space requirement, and parallel processing, a characteristic that speeds up data analysis.

Big data management is important for a library to gain a competitive advantage. It is important to analyze the data and extract the right information, as this leads to making better decisions, enhancing the level of services, and adopting the latest quality standards. Moreover, analyzing the big data in a correct efficient manner can enable libraries remain abreast with the latest trends and solve problems that may face their users and patrons.

REFERENCES

1. Jach, T., Magiera, E., & Froelich,W. (2015). Application of HADOOP to Store and Process Big Data Gathered from an Urban Water Distribution System. Procedia Engineering, 119, 1375-1380.
2. Cheng, Y., Chen, K., Sun, H., Zhang, Y., & Tao, F. (2017). Data and knowledge mining with Big Data towards smart production. Journal of Industrial Information Integration, 9, pp.1–3. DOI:http://dx.doi: 10.1016/j.jii.2017.08.001.
3. Samiya, K., Xiufeng, L., Shakil, K., & Alam, M. (2017). A survey on scholarly Data: from Big Data perspective. Information Processing & Management, 53(4), pp.923–944. DOI: https://doi.org/10.1016/j.ipm.2017.03.006.
4. Chen, C. P., & Zhang, C. (2014). Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data. Information Sciences, 275, 314–347. DOI: http://dx.doi.org/10.1016/j.ins.2014.01.015.
5. Sledgianowski, D., Gomaa, M., & Tan, C. (2017). Toward integration of Big Data, technology and information systems competencies into the accounting curriculum. Journal of Accounting Education, 38, pp.81–93.
6. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge discovery and data mining: towards a unifying framework, In KDD Proceedings (Vol. 96, pp. 82-88).
7. Oussous, A., Benjelloun, F., Lahcen, A., & Belfkih, S. (2017). Big data technologies: A Survey. Journal of King Saud University-Computer and Information Sciences, 30(4), pp.431–448. DOI:https://doi.org/10.1016/j.jksuci.2017.06.001.
8. Chen, M., Mao, S., & Liu, Y. (2014a). Big Data: a survey. Mobile Networks and Applications, 19(2), pp.171–209.
9. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in Big Data analytics. Journal of Big Data, 2(1).p.1.
10. Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., Shiraz, M., & Gani, A. (2014). Big Data: survey, technologies, opportunities, and challenges. The Scientific World Journal. vol. 2014, Article ID 712826, 18 pages, 2014. DOI: https://doi.org/10.1155/2014/712826.
11. Usha, D., & Aps, A. J. (2014). A survey of Big Data processing in perspective of Hadoop and MapReduce. International Journal of Current Engineering and Technology, 4(2), pp. 602–606.
12. Maheswari, N., & Sivagami, M. (2016). Large-scale data analytics tools: apache hive, pig, and hbase. In Data Science and Big Data Computing (pp. 191-220). Springer, Cham.
13. Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., & Chen, D. (2013). G-Hadoop: MapReduce across distributed data centers for data-intensive computing. Future Generation Computer Systems, 29(3), pp.739–750.
14. Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. (2016). The anatomy of Big Data computing. Software: Practice and Experience, 46(1), pp.79–105.
15. White, T. (2012). Hadoop: the definitive guide. San Francisco, CA: O'Reilly Media Inc.
16. Mall, N. N., & Rana, S. (2016). Overview of big data and Hadoop. Imperial Journal of Interdisciplinary Research, 2(5).
17. Ghoting, A., Krishnamurthy, R., Pednault, E., Reinwald, B., Sindhwani, V., Tatikonda, S., ... Vaithyanathan, S. (2011). SystemML: Declarative machine learning on MapReduce. In 2011 IEEE 27th International Conference on Data Engineering (pp. 231-242). IEEE.
18. Wang, G., Salles, M. V., Sowell, B., Wang, X., Cao, T., Demers, A.,... White, W. (2010). Behavioral simulations in mapreduce. Proceedings of the VLDB Endowment, 3(1-2), 952–963.
19. Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2014). A scalable two-phase top-down specialization approach for Data anonymization using mapreduce on cloud. IEEE Transactions on Parallel and Distributed Systems, 25(2), 363–373.
20. Big-Data for development facts and figures. (2014). Retrieved October 5, 2017 from http://www.scidev.net/mena/Data/feature/Big-Data-for-development-facts-and-figures-AR.html.
21. Constance, M. (2017). Libraries and the Big Data revolution [pdf]. Retrieved from https://www.oclc.org/content/dam/oclc/events/2017/EMEARC2017/EMEARC-2017-Plenary-Session-2-Libraries-and-the-Big-Data-Revolution-Constance-Malpas.pdf/
22. Hadoop importance in handling big data. (2016). Retrieved on October 9, 2017, from http://blog.BigDataweek.com/2016/08/01/hadoop-important-handling-Big-Data/
23. Compression tames Big Data on Hadoop. (2013). Retrieved October 9, 2017, from https://www.slideshare.net/rainstor/big-dataanalyticsonhadoopinfographic.
24. What makes Hadoop special? (2013). Retrieved October 9, 2017, from https://hyperstage.net/2013/08/what-makes-hadoop-special/.

25. Doulkeridis, C., & Nørvåg, K. (2014). A survey of large-scale analytical query processing in MapReduce. The VLDB Journal, 23(3), 355–380.
26. Eldawy, A., & Mokbel, M. F. (2013). A demonstration of spatial hadoop: An efficient mapreduce framework for spatial data. Proceedings of the VLDB Endowment, 6(12), 1230–1233.
27. Li, F., Ooi, B. C., Özsu, M. T., & Wu, S. (2014). Distributed data management using MapReduce. ACM Computing Surveys (CSUR), 46(3), 31.
28. Chavan, V., & Phursule, R. N. (2014). Survey paper on big data. International Journal of computer Science and Information Technologies, 5(6), 7932–7939.