Big Data: Technical Review of Hortonworks, Cloudera and MapR Hadoop Distributions

Muhammad Bukhari Bin Burhanuddin
School of Computer Sciences
Universiti Sains Malaysia
USM, 11800 Georgetown, Penang, Malaysia
P-COM0071/19

Kalai Yarasi Muthu Krishnan
School of Computer Sciences
Universiti Sains Malaysia
USM, 11800 Georgetown, Penang, Malaysia
P-COM0056/20

Dheniesh Thomas
School of Computer Sciences
Universiti Sains Malaysia
USM, 11800 Georgetown, Penang, Malaysia
P-COM0040/20

***Abstract*—Hadoop is an open source distributed system for data storage and parallel computations. Due to the exponential growth of big data, there is an increasing number of people and organizations that are adapting Hadoop. In order to meet the demands of the industries and users to process big data, many distributions emerge offering their own solutions. The aim of this study is to review and compare Hortonworks, Cloudera and MapR Hadoop distributions as they are the market leaders with huge market presence. Each of the distributions is reviewed in terms of their history, features and performance. This literature finds that there is no significant evidence which could completely determine one distribution is more superior than the other in terms of performance. The application of different Hadoop distributions in healthcare, education and transportation domains are also discussed.**

***Index Terms*—hadoop, big data, hortonworks, cloudera, mapr**

I. Introduction

As the world is getting more digitised, there is an exponential growth of data at an accelerated pace that could give beneficial information to interested stakeholders when they are processed correctly with modern computer technologies. The phrase "big data ” was first coined by Roger Magoulas in 2005 referring to a variety of massive datasets that are almost difficult to handle and process using standard data processing tools due to their size and complexities [1]. A perfect example of massive and complex datasets could be seen from various social media platforms such as Facebook and Twitter where thousands of comments, videos and pictures are posted on a daily basis [2].

In order to process all of this unstructured and multidimensional data to make it meaningful, the traditional data management technique would not work. In the interest of addressing this issue, Apache Hadoop which uses simple programming models was developed by Doug Cutting and Mike Cafarella in 2008 [3]. The Hadoop framework allows distributed processing of large datasets across clusters of computers was developed by Apache Software Foundation.

Software enterprises such as Hortonworks, Cloudera, IBM and Microsoft then began offering their own Hadoop solutions to meet the needs of other organizations with their big data requirements. The offering of multiple Hadoop solutions often beg the questions within the spectrum of: what is the difference between them? and which is the best among them? Thus, the purpose of this paper is to review different Hadoop distributions where the focus will be specifically on Hortonworks, Cloudera and MapR as according to a research done by The Forrester Wave in 2014, they are the industry leaders and have the biggest market presence [4].

II. Background Study

This section consists of two subsections where the first one elaborates on the history and evolution for each of the distributions. While the second one gives highlights on the distributions and their components.

A. History and Evolution of the Distributions

1) Hortonworks: Hortonworks was formed by 24 engineers from the original Hadoop team at Yahoo! in January 2011 and it was then successfully acquired by Cloudera in January 2019 for the betterment of open source standards in the Hadoop ecosystem although they remain separate distributions [5]. Hortonworks Data Platform (HDP) is an open-source architecture platform for distributed storage and processing multiple and large volumes of data. Since Hortonworks is a completely open source Apache-licensed platform, their business value comes from providing expert technical user support, training and services catered for their partners [6].

2) Cloudera: Cloudera is a commercial company which was founded by technological experts from Google, Facebook, Yahoo! and Oracle in 2011 [7]. Cloudera has the ability to install and manage Apache Hadoop and related components. It analyses and manipulates the data stored and protects it with its encrypted data storage [7].

3) MapR: MapR was founded in 2009 by former Google members and with an initiative to improve the performance of Apache Hadoop [8]. MapR can store, process and analyse real-time high volume data at high speed. MapR creates its own file system such as MapR File System (MapR-FS), the MapR-DB NoSQL database management system, MapR Streams, and MapR Control System (MCS) user interface [9].

B. Highlights of the Distributions & Their Components

1) Hortonworks: Hortonworks Data Platform (HDP) is an open source framework for distributed data storage that is able to process large and multi-dimensional datasets [10]. HDP is divided into three interrelated layers of main components which are Core Hadoop 2, Essential Hadoop, and Hadoop Support. Core Hadoop 2 is the basic component of Apache Hadoop 2x which consists of 3 sub-components such as Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN) and MapReduce 2 (MR2). Essential Hadoop consists of a set of Apache components which was designed to support Core Hadoop.

Essential Hadoop houses Apache Pig, Apache Hive, Apache HCatalog, WebHCat, Apache HBase, and Apache ZooKeeper. The third component is Hadoop Support which consists of multiple parts that support Hortonwork in monitoring Hadoop's installation connection with larger compute environment [10]. They are Apache Oozie, Apache Sqoop, Apache Flume, Apache Mahout, Apache Knox, Apache Storm, Apache Spark, Apache Phoenix, Apache Tez, Apache Falcon, Apache Ranger, Apache DataFu, and Apache Slider [10].

2) Cloudera: Cloudera provides varieties of products and tools such as Cloudera Distribution Hadoop (CDH). CDH is an Apache-licensed open source platform that supports processing of data on a massive scale. CDH uses the Hadoop framework to work with multiple components such as Hive, Spark, Pig, MapReduce, Impala, Solr, YARN, and Kite. The uniqueness of CDH is that it is the most tested and complete set of Hadoop distribution which makes it one of the most preferable choices for users [11]. CDH is time and effective due to its powerful performance of being able to process critical data efficiently.

Apache Impala is one of the Hadoop components that is offered by Cloudera. Impala provides a familiar SQL interface, real-time and batch oriented queries and also has the ability to process Business Intelligence (BI) style data files [12]. Apache Hive is another integrated part of CDH which provides batch processing for Hadoop and compared to Impala it provides fault-tolerant measure which the former does not due to its use of in-memory based operation Impala does not provide fault-tolerant measure [12].

Cloudera Search is based on Apache Solr architecture that is fully integrated with CDHproviding real-time access to data stored in Hadoop and HBase [13]. Among the main highlights of this component is simplified infrastructure and interaction which gives quicker insights to users without requiring them having advanced programming skills. Cloudera Manager that offered in Cloudera distribution is a centralized application for managing CDH clusters and it provides administrative functions such as deploy, manage, monitor and diagnostic [13]. The administrative functions are beneficial toward user experience and also helpful for users in optimizing cluster performance. Cloudera Navigator which provides data-management and security systems for the CDH framework enables users to explore large amounts of data in Hadoop with encryption keys [13].

3) MapR: MapR Data Platform offers several components and packages such as MapR-XD, MapR Analytics, MapR-DB, MapR-ES and MapR-Edge which help to store, manage, process, apply and perform big data analytics on data in large scale [14]. MapR helps companies explore intelligent ideas with its platform which enables the creation of powerful applications with real time data.

MapR-XD is a distributed data storage combining analytics and operation into one platform and is also often used with Apache Spark to store data at an exabyte scale [15]. MapR Analytics is described as a container management platform for running artificial intelligence/machine learning (AI/ML) and data analytics with open source security projects named Secure Production Identity Framework for Everyone (SPIFFE) and SPIFFE Runtime Environment (SPIRE) [8] [16].

MapR-DB is a NoSQL database which is equivalent to HBase but it is a proprietary product offered by MapR which they claimed to be more efficient, faster, more scalable and can be queried from dbshell for simple interaction [8]. MapR-Edge is a powerful clustering architecture designed by MapR that could run on a small commodity hardware to send data to the cloud at a fast rate as well enable it to capture, process, and analyze IoT data close to the source [17].

III. Comparative Analysis

This section will elaborate on the comparative analysis that was done on the three Hadoop distributions based on two perspectives which are their features and performance.

A. Features Comparison

Features comparison between Hortonworks, Cloudera and MapR distributions is very generic and have been done extensively in multiple research papers such as in [18] and [7]. In addition, the components for each distribution have been elaborated in the previous section. Hence, in this subsection the features comparison is looked at the important features that are deemed significant and it should be noted that a lot of the syntaxes that are used in the table are referred to and adapted from other studies particularly from [18] and [19]. The features comparison is summarized in TABLE I.

One of the interesting points to note from the table is supporting operating systems where only Hortonworks distribution is available for both Windows and Linux which put them at an advantage in expanding their user base compared to Cloudera and MapR that is only available for Linux. Cloudera used to be an open source platform offered with additional licensed proprietary/enterprise features but its business approach has been to provide its distribution completely open source ever since its merger with Hortonworks.

TABLE I: Features comparison between the distributions

	Hortonworks	Cloudera	MapR
File System	HDFS, read-only NFS	HDFS, read-only NFS	HDFS, read/write NFS (POSIX)
Management Tool	Ambari	Cloudera Manager	MapR Control System
Operating System	Windows, Linux	Linux	Linux
GUI Availability	Yes	Yes	Yes
Execution Environment	Local or Cloud	Local or Cloud	Local or Cloud
License	Open source	Open source	Proprietary

TABLE II: Analytical investigation setup

Distributions version	Cloudera CDH 4.3 Hortonworks HDP-1.3 MapR M3 v3.0
Test environment	Virtualized cloud infrastructure Provided by ProfitBricks
Cluster configuration	Four CPU cores for each node RAM: 16 GB Virtualized disk space: 100 GB Custer size: 4 - 16 nodes
Monitoring tool	Ganglia Monitoring System Monitor CPU, disk, RAM, network and Java Virtual Machine (JVM) parameters
Test data size	1.6 TB

B. Performance Comparison

Analytical investigation that had been done by Altoros Systems will be heavily referred to in this subsection for evaluating the performances of the three Hadoop distributions [20]. Evaluating the performance of a Hadoop cluster proved to be a challenging task but the study approaches this problem by setting standardized configuration/test conditions and micro-benchmarks to compare the distributions' performance in processing data (MapReduce job).

Micro-benchmark allows for testing parts of Hadoop infrastructure and with this method, the scalability of a cluster of any size is tested by putting load on its CPU, disk, RAM and network [20]. The micro-benchmarks that were measured are Bayes, Distributed File System I/O (DFSIO), HIVEAGGR, PageRank, Sort, TeraSort and WordCount. The setup for the analytical investigation that was used is tabulated in TABLE II.

For the overall cluster performance, the results show that the movement of Cloudera mirrors Hortonworks and vice versa. In three of the benchmark tests (Bayes, PageRank, WordCount), Cloudera performed better than Hortonworks while Hortonworks performed better than Cloudera for the other four tests (DFSIO, HIVEAGGR, Sort, TeraSort). However, the gap or difference in performance between the two distributions are very little and not significant to conclude that one is more superior than the other as shown in Fig. 1.The y-axis indicates the throughput scalability where the higher the value, the better and x-axis indicates the number of nodes. The throughput is measured in bytes per second.

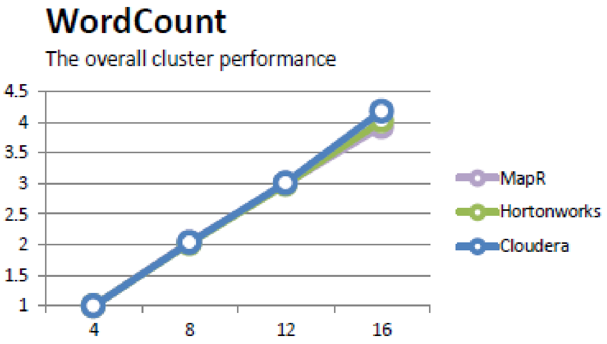

Fig. 1: Overall cluster performance for WordCount benchmark

Meanwhile, the performance of MapR either degrades or plateau in 12-node and 16-node clusters for most of the benchmarking tests. An example of this can be seen from Fig. 2. The research concluded that the choice of Hadoop distributions has no impact that can be considered having a major effect on the overall system throughput and their difference in performance was within the limit of an experimental error [20]. The authors also highlighted that more emphasis should be put in configuring MapReduce task parameters to fully take advantage of CPU, RAM, disk and network utilization. The overall cluster performance for each of the micro-benchmarking tests are compiled in APPENDIX B and could also be referred from Main Ref V.

C. Strength and Weakness

TABLE III describes the strength and weakness between Hortonworks, Cloudera and MapR. The points tabulated in the table are extracted and adapted from multiple researches that had been done by Forrester Research group on big data Hadoop solutions [4] [21] [22].

IV. Discussion

The implementation of Hadoop technology from different vendors is not limited to the technology industry. The following subsections give details on recommendations that

TABLE III: Strength and weakness between the distributions

	Hortonworks	Cloudera	MapR
Strength	Scalable architecture and continually evolving open source ecosystem. Strong innovation strategy through open source community.	Favored among cloud/Saas providers that have deployed Hadoop-based services. Firm cooperation with other technology vendors.	Offered feature-rich proprietary software. Enhanced disaster recovery and high-availability features.
Weakness	Data modeling and data transformation efficiency lags behind other competitors.	Data modeling, data governance and security lags behind other competitors.	Market awareness lags behind other distributions vendors.

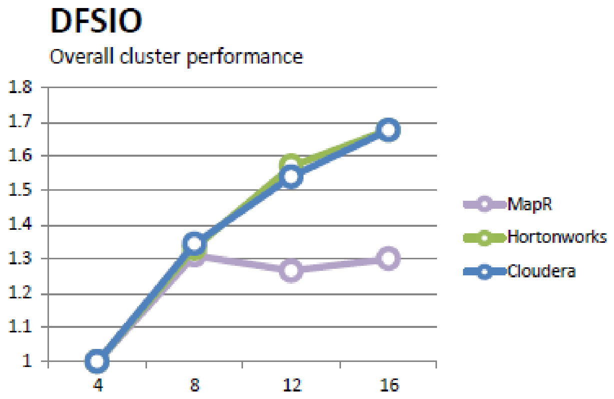

Fig. 2: Overall cluster performance for DFSIO benchmark

could be done from big data analytics and their application in healthcare, education and transportation domains using different Hadoop distributions.

A. Healthcare

Based on research done by Allied Market Research, the utilization of big data analytics in the healthcare market is projected to grow to US$67.82 billion by 2025 and in 2017, it was valued at US$16.87 billion [23]. The usage of Hadoop in the healthcare sector is highly recommended due to its fault-tolerant architecture and high security in preserving as well as protecting patients' medical records.

Hadoop framework also could be utilized by healthcare providers to make informed clinical and business decisions by building a predictive care model from big data analytics, machine learning and artificial intelligence [23]. Such predictive models using big data analytics could be further extended by processing and analyzing massive historical data to avoid preventable diseases and potential outbreak of epidemics.

For example, Cardinal Health deploys HDP that provides next-generation data architecture and rapid data management in order to improve its patient care [24]. Cardinal Health also chooses HDP compared to the rest of Hadoop vendors due to it being an open-source platform where they could leverage existing data assets and spread existing investments into other tools and processes [24].

Cloudera claims that more than 250 healthcare organizations worldwide such as Clearsense employs its end-to-end data management platform to offer better health services, minimize care costs, clinical prediction, precision medicine and agile drug discovery [25]. Among prominent features that were highlighted by Cloudera are hybrid and multi-cloud environments that enable rapid deployment and data architecture that supports Health Insurance Portability and Accountability Act (HIPAA) compliance [25].

The authors in [26] suggest the usage of MapR (MapReduce version 2) to implement machine learning algorithms combined with Apache Spark for fast processing of continuous data streams. The processed data will then give meaningful healthcare data suggestions to the caregivers [26].

B. Education

In the education domain, big data analytics could serve as a catalyst to improve learners' experience. There are many studies such as in [27] and [28] stated that data trials that are left by students when they interact with technologies could be used to examine patterns of students performance over time and identify potential issues related to academic programming.

For example, authors in [29] developed a prototype called HESSEM to assess the United States of America higher education service with a ranking system based on a big data approach. The prototype deployed Hortonworks in its framework in order to map, sort and aggregate data that was feeded from web crawler, jsoup, ETL modules and etc. in real time.

Another example of big data analytics used in education was presented in [30] although no mention of which distribution that was used. The authors developed a recommender system by collecting the students' social networking data that could provide recommendations revolving around their area of interest which benefit their studies [30]. The system was built based on machine learning with Hadoop as its backbone.

C. Transportation

Big data analytics have a big potential in transforming the transportation sector especially in terms of improving traffic management and road infrastructure. With the rapid increases of vehicles on the road year by year, real-time monitoring operations of traffic are now deemed important [31]. The road and traffic department in the City of Dublin for example was able to enhance traffic management leading to 10 - 15% reduction in journey times. This achievement was made possible by analyzing and processing a stream of big data from an array of sources such as bus timetables, closed-circuit television (CCTV) cameras, inductive-loop traffic detectors and many more [31].

Accessing, retrieving and understanding mass spatio-temporal trajectory data is crucial in order to understand traffic congestion. However, such activities are resources intensive and require high computational capabilities as well as efficient data organization mechanisms. As such, the authors in [32] propose ESTRI (Efficent spatio-temporal data retrieval method based on Impala) which is based on the Cloudera Impala query engine to optimize and enhance mass spatio-trajectory data retrieval/sharing. The experiment that was conducted shows that the performance of ESTRI in retrieving massive spatio-temporal trajectory data volumes (50 million, one 100 million, 150 million, and 200 million) is approximately seven times higher than MongoDB [32]. The study concluded with the bus distribution and paths in Taiyuan city being mapped which could help traffic management of public transformation systems in the future [32].

V. Conclusion

The study that we have conducted allow us to define the similar features and specifications of different Hadoop distributions. Thus, the goal of this study is met. Big data technologies are improving and businesses are more data-driven more than ever. Apache Hadoop is a framework provided to users in order to process unstructured datasets and overcoming limited computing capabilities for big data analytics. Results from multiple studies show that there is no certainty to be said that one distribution is more superior than the other as the conditions to measure their performance need to be strictly standardized and each distribution has their own parameters need to be set for the clusters depending on various factors such as dataset size.

In our opinion, the three Hadoop distributions that are described in this literature will continue to be the market leaders that offer more innovative solutions in the future for industries that are moving toward data-driven goals. However, before adapting Hadoop one has to ponder many questions such as whether the provided dataset is suitable and whether they have the expertise to implement the technology. Thus, traditional data management architecture will continue to be relevant.

Acknowledgement

We would like to express our gratitude and appreciation to Dr. Chew for her teaching and guidance. Please refer to APPENDIX A for details regarding group members contribution.

References

[1] M. R. Wigan and R. Clarke, “Big data’s big unintended consequences,” *Computer*, vol. 46, no. 6, pp. 46–53, 2013.

[2] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.

[3] S. Alkatheri, S. A. Abbas, and M. A. Siddiqui, “A comparative study of big data frameworks,” *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 17, no. 1, 2019.

[4] M. Gualtieri, N. Yuhanna, H. Kisker, and D. Murphy, “The forrester wave: Big data hadoop solutions, q1 2014,” 2014.

[5] I. Cloudera, “Hortonworks data platform,” Dec 2020.

[6] I. Analytics, “Hortonworks data platform: An open-architecture platform to manage data in motion and at rest,” tech. rep., IBM Corporation, 2017.

[7] A. Erraissi, A. Belangour, and A. Tragha, “A big data hadoop building blocks comparative study,” *International Journal of Computer Trends and Technology. Accessed June*, vol. 18, 2017.

[8] C. Preimesberger, “Mapr technologies: Product overview and insight,” tech. rep., TechnologyAdvice, 2018.

[9] H. Enterprise, “Mapr 5.2 documentation,” tech. rep., Hewlett Packard Enterprise Development LP, 2020.

[10] I. Hortonworks, *Hortonworks Data Platform: Installing HDP on Windows*. Hortonworks, Inc., Jul 2015.

[11] I. Corporation, “Cloudera data hub with ibm: Enterprise-grade open source for machine learning and analytics,” tech. rep., IBM Corporation, 2019.

[12] X. Liu, N. Iftikhar, and X. Xie, “Survey of real-time processing systems for big data,” in *Proceedings of the 18th International Database Engineering & Applications Symposium*, pp. 356–361, 2014.

[13] I. Cloudera, *Cloudera Introduction*. Cloudera,Inc, 5.9.x ed., Jul 2020.

[14] I. MapR Technologies, “Mapr: What’s included,” Dec 2020.

[15] R. Liu, H. Isah, and F. Zulkernine, “A big data lake for multilevel streaming analytics,” in *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, pp. 1–6, IEEE, 2020.

[16] A. Woodie, “Hpe unveils ‘ezmeral’ platform for next-gen apps,” tech. rep., Datanami, Jun 2020.

[17] B. Lalitha *et al.*, “Recover the missing data in iot by edge analytics,” *i-Manager’s Journal on Software Engineering*, vol. 13, no. 2, p. 25, 2018.

[18] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, “Big data technologies: A survey,” *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[19] A. Erraissi, A. Belangour, and A. Tragha, “Digging into hadoop-based big data architectures,” *International Journal of Computer Science Issues (IJCSI)*, vol. 14, no. 6, pp. 52–59, 2017.

[20] V. Starostenkov and K. Grigorchuk, “Hadoop distributions: Evaluating cloudera, hortonworks, and mapr in micro-benchmakrs and real-world applications,” tech. rep., Altoros Systems, Inc., 2013.

[21] J. G. Kobielus, S. Powers, B. Hopkins, B. Evelson, and S. Coyne, “The forrester wave™: Enterprise hadoop solutions, q1 2012,” tech. rep., Forrester Research, Inc., jun 2012.

[22] N. Yuhanna, G. Leganza, and J. Lee, “The forrester wave™: Big data warehouse, q2 2017,” *Adoption Grows As Enterprises Look To Revive Their EDW Strategy*, p. 17, 2017.

[23] V. Gaul, “Big data analytics in healthcare market,” *Allied Market Research*, Dec 2018.

[24] K. Rose, S. MSL, K. Giannini, and A. Vaverva, “Cardinal health selects hortonworks to power next-generation data platform for healthcare,” May 2013.

[25] S. Strauss, “Healthcare’s big data challenge: How a hybrid data platform can help,” *Cloudera Blog*, May 2020.

[26] M. S. Hossain and G. Muhammad, “Healthcare big data voice pathology assessment framework,” *iEEE Access*, vol. 4, pp. 7806–7815, 2016.

[27] A. G. Picciano, “The evolution of big data and learning analytics in american higher education.,” *Journal of asynchronous learning networks*, vol. 16, no. 3, pp. 9–20, 2012.

[28] B. Daniel, “B ig d ata and analytics in higher education: Opportunities and challenges,” *British journal of educational technology*, vol. 46, no. 5, pp. 904–920, 2015.

[29] R. G. Qiu, Z. Huang, and I. C. Patel, “A big data approach to assessing the us higher education service,” in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–6, IEEE, 2015.

[30] A. Jagtap, B. Bodkhe, B. Gaikwad, and S. Kalyana, “Homogenizing social networking with smart education by means of machine learning and hadoop: A case study,” in *2016 International Conference on Internet of Things and Applications (IOTA)*, pp. 85–90, IEEE, 2016.

[31] J. Fernández-Lozano, M. Martín-Guzmán, J. Martín-Ávila, and A. García-Cerezo, “A wireless sensor network for urban traffic characterization and trend monitoring,” *Sensors*, vol. 15, no. 10, pp. 26143–26169, 2015.

[32] L. Zhou, N. Chen, S. Yuan, and Z. Chen, “An efficient method of sharing mass spatio-temporal trajectory data based on cloudera impala for traffic distribution mapping in an urban city,” *Sensors*, vol. 16, no. 11, p. 1813, 2016.