Advances in Aeronautical Science and Engineering
ISSN: 1674-8190

Exploring Big Data Processing: A Comparative Approach with Hadoop

Sebastian Lenz
Research Scholar, Heidelberg University, Baden-Württemberg, Germany

ABSTRACT

Big data is an important concept in the field of information technology where companies and organization take advantage of the data that they have stored to find meaningful patterns and predictions to help them in making informed decisions. Big data analysis involves the use of advanced tools and techniques that are used in the processing the large volumes of data that is produced by the organization (Sammer, 2012, p. 23). The Hadoop framework is an important framework in enabling easy and efficient processing of large data sets thus making it one of the most popular big data analytics frameworks available. In this research, a comparative analysis of big data will be made through the use of Hadoop specifically concentrating on the Hadoop architectures.

***Keywords**: Big Data, Hadoop Framework, Data Mining.*

I. INTRODUCTION

The Hadoop framework is an open source distributed processing framework that is used in the management of storage and data processing for big data applications which run on clustered systems. The core components of Hadoop include the Hadoop distributed file system, MapReduce, YARN, and Hadoop common (Sammer, 2012, p. 29). There are also architectures that use the Hadoop framework which include the Horton works distribution Cloudera distribution, IBM InfoSphere Big Insights Distribution, and Pivotal HD distribution.

All the architectures that are based on Hadoop are designed and edited by different organizations such as yahoo oracle, face book and Google depending on the requirements of the editors of the framework. Other factors considered when evaluating the architectures include the administration console, base and the available editions of the architecture. In this research, we conducted a comparative analysis on Hadoop architectures based on the editors, base, available versions and administration console.

II. LITERATURE REVIEW

There are several major architecture distributions of Hadoop. These distributions use the Hadoop framework in manipulating big data and help in the management of its components. These distributions include HortonWorks, Cloudera, Pivotal HD, and IBM Infosphere Big Insights. Horton Works is a distribution that is formed from the Hadoop framework. Al its components are open source and have been licensed from apache thus facilitating the adoption of Apache's Hadoop framework ("Digging into Hadoop-based Big Data Architectures," 2017, p. 53). It contains elements such as querying, Heart Hadoop, Integration services, planning, coordination, learning, script platform, supervision and management.

IBM InfoSphere Big Insights Distribution initially had two versions when it was being launched. These versions included the basic version and the enterprise edition. The basic version was a free download of Apache Hadoop that is integrated with a web management console. Another quick start was later launched which provides massive data and volume analysis capabilities on the organizations or business platform. This new version combines both the Hadoop's open source solutions with the enterprise functionality thus providing a large-scale analysis that has capabilities of being resilient and fault tolerant.

Pivotal HD Distribution is a distribution of Apache Hadoop that is commercially supported by Apache Hadoop. This distribution incorporates an integration of Greenplum which massive parallel processing database with the Apache Hadoop thus ensuring that it is a cost-effective method and ensuring that it is an open source and flexible big data platform ("Digging into Hadoop-based Big Data Architectures," 2017, p. 55). This distribution is among the most efficient deliveries that uses the Hadoop framework.

Cloudera distribution is a distribution that has largely borrowed and inherited the components of Apache's Hadoop. It also uses the house components for the management of clusters. It also offers a fully open source version of its platform. It also has another business model of selling licenses and also selling training and support.

III. METHOD & MATERIAL

To evaluate the distributions, it is necessary to conduct a comparative study to identify the strengths and weaknesses of the distributions of Hadoop. The main focus is on the Forrester wave which was conducted on the Hadoop distributions. The criteria for comparison mainly focuses on the editor used in the distribution, available edition, basis for publishing, tools used in the management and administration of the distributions, and the components of the solution.

A comparative analysis of the editor of the distribution involves evaluation of the companies that were involved in designing and developing the adobe distributions and whether they commercialized the distribution. The available edition comparative evaluation the different versions of the distribution as presented by the publishers or the editors of the distribution. The base or basis for publishing compares the basis in which the publishers based themselves in the improvement of their solutions with the publishers having to concentrate on the Hadoop distributed file system ("Digging into Hadoop-based Big Data Architectures," 2017, p. 56). Comparative evaluation of the administrative consoles of the distribution with it specifically focused on the tools that are used in the management of Hadoop. Such tools help the distributions with configuration, automation, tracking, reporting, robust troubleshooting, and simple maintenance of the distributions. A comparison of the solution evaluates the elements that make up each of the distribution solutions.

IV. RESULT & DISCUSSION

This section evaluates the results from conducting a comparative analysis of the four main Hadoop distribution architectures. These includes Cloudera distribution, Horton Works Distribution, IBM InfoSphereBigInsights Distribution and pivotal HD Distribution. The comparative analysis is based on the editors of the distributions, available editions. The analysis is explained according to each distribution.

Comparative analysis on the editors

- The editors of Cloudera distribution are Hadoop experts from Oracle, Facebook, Google, and Yahoo. They majorly used the components of the Apache Hadoop which are integrated with house components for cluster management (Kaur, Chauhan, & Mittal, 2018, p. 23).
- The editors of Horton Works Distribution are the member of the Yahoo team that is in charge of Hadoop projects. They made all the components of the distribution to be open source.
- IBM developed IBM InfoSphereBigInsights Distribution in 2011 with two versions which are the basic free version and the enterprise version.
- The Pivotal Software, Inc. write pivotal HD Distribution; this is a software services company which is based in San Francisco with a department for developing big data solutions.

Comparative analysis of available editions

- Cloudera has Cloudera Express and the Cloudera enterprise which has a basic edition, FlexEdition, and Data HubEdition (Sammer, 2012, p. 36).
- Horton Works contains one single version which is the HortonWorks Data platform 2.5 that includes Apache Hadoop and is used for storage, analyzing and processing large volumes of data.
- IBM InfoSphereBigInsights Distribution has three significant versions that it has produced which include Quick Start Edition, Standard Edition, and Enterprise Edition (Vohra, 2016, p. 37).
- Pivotal HD Distribution has a single version which is the Pivotal HD Enterprise which is used in operationalizing and analyzing data produced by an enterprise.

Comparative analysis of the base

All the distributions use YARN. This is a job scheduling and resource management technology that is used in the Hadoop open source distributed processing framework ("Hadoop Introduction," 2016, p. 9).

Comparative analysis of administration console

Cloudera uses the Cloudera manager who is an end to end application manager for the management of CDH clusters. Horton Works uses Ambari in to help with its administration where it uses guesswork out of making operations on Hadoop (Vohra, 2016, p. 34). Pivotal HD Distribution uses a command center in its administration. IBM InfoSphereBigInsights uses a web console that allows commands to be executed to administer the distribution.

Comparative analysis of the components

- Cloudera has components such as Hive, Pig, Zookeeper, HBase, Hue, Sqoop, Avro, Whirr, Flume, Yarn, Mahout, Cloudera, Cloudera Manager, Impala, Oozie, r, Apache Sentry (Kaur, Chauhan, & Mittal, 2018, p. 13).
- Horton Works contains components such as Hive, Pig, HBase, WebHDFS, Hue, Tez, Whirr, Yam, Zookeeper, Mahout, Sqoop, Oozie, Storm, Apache Ganglia, Flume, ApacheFalcon, NFS, Spark, Accumulo, and Knox (Kaur, Chauhan, & Mittal, 2018, p. 17).
- IBM InfoSphereBigInsights contains components such as Yarn, Zookeeper, Spring Oozie, HBase, Spring XD, HDFS, Sqoop, HAWQ, and GemFire XD.
- Pivotal HD Distribution BigSheets, Integrated Installer, Dashboard & Visualization, Text Analytics, MapReduce, Hive, Workflow, Pig, Text processing Engine &Extractor Library, Jaql, R.

From the results, it is clear that all the distributions have been edited by different organizations such as yahoo, IBM, Facebook, and Google. Each of the distributions has already released several versions for usage. Cloudera and HortonWorks distribution have some similar components such as HIVE and PIG with the rest of the distributions each having their own components. Each distribution contains its own administration console. All the distributions use the base YARN.

V. CONCLUSION

Hadoop is an open source distributed processing framework. It contains several distributions such as Cloudera, Horton Works, IBM InfoSphereBigInsights, and Pivotal HD Distribution. These distributions may be compared based on their editors, available edition, the basis for publishing, tools used in the management and administration of the distributions, and the components of the solution. Results show that the distribution contains different comparisons from the basis previously set, but all use YARN as their base.

In the future, it is necessary for more developers to integrate and form new distribution that will ensure better management of big data. Integrating and forming the new distributions that are compatible will ensure that there is flexibility with how big data may be managed (Shrivastava & Deshpande, 2016, p. 57). Hadoop also needs to evaluate the possibility of integrating machine learning components to its framework to help distributions and those managing big data with the framework to get better results. It is also necessary for the framework to continuously update its security features in the future to ensure that the data managed using the framework is safe from unauthorized people("Hadoop Security," 2016, p. 133).

REFERENCES

1. *Digging into Hadoop-based Big Data Architectures. (2017). International Journal of Computer Science Issues, 14(6), 52-59. doi:10.20943/01201706.5259*
2. *Hadoop Introduction. (2016). Professional Hadoop®, 1-14. doi:10.1002/9781119281320.ch1*
3. *Hadoop Security. (2016). Professional Hadoop®, 109-140. oi:10.1002/9781119281320.ch6*
4. *Kaur, R., Chauhan, V., & Mittal, U. (2018). Metamorphosis of data (small to big) and the comparative study of techniques (HADOOP, HIVE and PIG) to handle big data. International Journal of Engineering & Technology, 7(2.27), 1. doi:10.14419/ijet.v7i2.27.11206*
5. *Sammer, E. (2012). Hadoop Operations. Sebastopol, CA: O'Reilly Media.*
6. *Shrivastava, A., & Deshpande, T. (2016). Hadoop Blueprints. Birmingham, England: Packt Publishing.*
7. *Vohra, D. (2016). Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. New York, NY: Apress.*

.