International Journal of Early Childhood Special Education (INT-JECSE)
DOI: 10.9756/INT-JECSE/V14I2.287 ISSN:1308-5581 Vol 14, Issue 02, 2022

A Systematic Literature Review of Big Data and the Hadoop frameworks

Devishree Naidu[1]*, Adi Thakur[2]
[1]Computer Science and Engineering Department,
Shri Ramdeobaba College of Engineering and Management,India [2]Computer Science and Engineering Department,
Shri Ramdeobaba College of Engineering and Management, India

Author for Correspondence: Devishree Naidunaidud@rknec.edu

Abstract
Big data is a term to define the huge amount of data gathered mostly through new data sources like Twitter, Instagram, Facebook etc. This data is important as its analysis is changing how major businesses work and has the ability to provide the knowledge required to cut back business costs. Most firms are currently using this technology to accurately find trends and predict future events in their various industries. The challenge lies in finding the best way to process, analyze and draw useful insights from this data. This data cannot be handled efficiently by the traditional data management tools and hence required some other advanced data technologies. This is mainly because of its unstructured nature and the five V's – Volume, Variety, Velocity, Value, and Veracity which we mostly use to define big data are the main reason why its handling is a major challenge. Since this data is growing at an exponential rate, it was a necessity a develop technologies to address it. Hadoop, Map Reduce, and No SQL are the major three technologies that were developed to handle the complexities of big data and manage it reliably. This paper discusses the several technologies based on Hadoop which is altogether called the Hadoop ecosystem and their uses in analyzing big data.

Keywords:*Big data, Flume, Map Reduce, Hadoop Ecosystem, Hadoop frameworks*

I. INTRODUCTION

Imagine a world that's is running without data storage. This concept states that every detail of an organization or a couple of individuals is lost after they use it. This loss would affect their ability to perform detailed analysis and gain valuable insight. Data is what represents every minute detail in today's world be it the details of humans, businesses, payments etc. We can't even imagine if this data is lost as, it is created what a big loss it will bring to the major companies.

Living in the 21st century means faster work, precise analysis on the highly processed data. Big data is the new generation of data that demands faster processing and better analysis. Its characteristics such as high velocity, amount of data, and complexity make it difficult to be processed by traditional systems. Volume, Variety, Velocity, Veracity, and Value make up the current 5 Vs of big data. Each standing for its meaning in it. Volume for the huge amount of data, Variety for the variable data that are unstructured, Velocity for the speed of generation of data, Veracity for the variations in the type of data that is being generated on a daily basis.

Due to the rise of the internet and the increasing number of people with knowledge and expertise, the demand for knowledge has become more prevalent. This is why huge amounts of data have become available. To extract value from this data, proper analysis and storage techniques are necessary. Data is the base on which millions of organizations have their foundations on and hence any surge in the amount of data requires an equal surge in the knowledge of data generated.

Examples of big data:

• New York Stock Exchange generates massive amount of data which can go upto about one terabyte of data daily.

• According to the reports, every day more than millions of newly generated data get absorbed into the databases of the sites like Facebook. This massive data is a result conditions of uploading hundreds of images and videos, exchange of messages, commenting etc.

• The black box of a flight, can create 10+terabytes of data in 30 minutes. Thus, creation of data exceeds up too many numbers of Petabytes.

Despite the potential advantages of parallel and distributed computing for Big Data, its characteristics may prevent most common users from using it. Some of these include: Its dependence on data integrity and availability, its task scheduling and cluster load balancing, and its lack of portability.

Originally, the three dimensions were divided into three, which were referred to as the three Vs. Today, they are often divided into four or five Vs. Variety, velocity, and value are the main factors that determine the size of a big data repository.

The concept of MapReduce is a programming paradigm that combines the functions Map and Reduce to generatemaps and answer questions. It is divided into two parts which are the heart of Hadoop – Map and reduce. Both works parallelly to give faster results. Map is used to split the data and create a map of individual words. Reduce is used to reduce the amount generated by map and present a concise output. It works seamlessly across distributed and parallel computing. Google's MapReduce was first introduced in the context of Big Data.

Doug Cutting and Mike Cafarella, who worked for Yahoo!, came up with the concept of Apache Hadoop as a way to hide the complexity of Big Data from users. It is a distributed framework that enables developers to easily create interactive and data-oriented apps. The idea is derived from google that is to hide complexity from the end users. Its is an efficient way to provide answers to store, manipulate and extract information of Big Data. Over the years, Hadoop has developed widely and have attracted the attention of many researchers', hence many papers are being published to extract full use of this framework.

This paper presents a literature review of the Hadoop framework and its related issues.

The main objective of this paper is to review the different papers and literature available on the internet to see what are the different frameworks that are being discussed highly and what are the topics that are still being researched. Thus, we can see what are the areas that might need an improvement and different solutions that these

The remaining of the work in this paper is as follows. First, we discuss the research method that was used, the sources and the criteria. Then we move on to discuss how the different paper were sorted. The main part of the paper is to see the contribution of each paper on our literature and their Hadoop frameworks discussed.

II. Research Method

Systematic review helped to perform a thorough balanced literature review, summing up technical value to its results. The main purpose of a systematic literature review is to present a concise valuation regarding the research topic we choose through the presentation of a dependable, accurate, and reviewed approach. The review starts with the definition of the steps that will lead the development of the research, including research questions and methods. According to the custom, the mandatory steps are as follows

The research questions that the study aims to respond.

1. Approach that was used to search for the other authors papers like search keywords and source to find the papers, journals, and conferences.
2. Inclusion and exclusion criteria for the searched paper got over the internet.
3. Strategy for data withdrawal and following procedure of extracted data.

A. Objective

Parallel and disbursed computing presently has a essential function in statistics processing and records extraction of huge datasets. Over the previous few years, commodity hardware have become a part of clusters, for the reason that x86 platform address the want of getting an common higher cost/overall performance ratio, whilst lowering renovation cost. Apache Hadoop is a framework evolved to take benefit of this technique, the usage of such commodity clusters for garage, processing and manipulation of huge quantities of statistics. The framework turned into designed over the MapReduce paradigm and makes use of the HDFS as a garage report system. Hadoop affords key traits while appearing parallel and disbursed computing, along with statistics integrity, availability, scalability, exception handling, and failure recovery. Even more, those functions are offered in a obvious way to the user, which represents a unique technique to newcomers.

In this context, the goal of this research is to address the question: What are the principle subjects of research associated with Apache Hadoop and its programming paradigm, MapReduce? Although Hadoop is a reasonably followed platform for distributed processing of massive datasets, it nevertheless has room for enhancements and we need to find out what are those areas.

B. Research Questions

1. What are the tools and technology which are used for large stream data flow analysis?
2. What are the prevailing software frameworks structure methods for large data structures in Hadoop?

C. Search strategies

After the research questions had been determined, the search strategies and the search string were defined. The scope of the research and the digital libraries consulted were also defined. Research Keywords: We identify keywords based on research questions. Synonyms for the keywords did not have to be taken into account, as the names were predefined on the basis of the names defined by the project developers. "Hadoop" and "Big Data" were identified as search terms.

D. Sources

The criteria for selecting the sources of the studies were search on the web mechanism; Search mechanisms should enable personalized searches by title and abstract; the complete articles must be available for download via the the digital library;

Importance and relevance of the sources. With the search string defined, we have selected the following digital libraries as sources:

ACM Digital Library. IEEEXplore ScienceDirect, Google Scholar.

E. Study selection criteria

In order to omit irrelevant studies, research work outside the scope, the following exclusion criteria were applied:
EC 1: Articles that do not contain a description of the big data framework or a big data application that uses Hadoop.
EC 2: Article without reference to a field of computer science.
EC 3: Articles are written in a language other than English
EC 4: Workshop article
EC5: Articles that do not (or partially discuss) Big Data architecture or Hadoop

III. Results of the Systematic Literature Review

In this section, the result of our literature review is formulated. The results focus on pre-defined responses to questions considered to be the end result of our systematic review of the literature. Each subsection provides information to answer these questions regarding the purpose of the studies. Some tables are presented to represent the results of this review.

RQ 1. What are the tools and technology which are used for large stream data flow analysis?

Hadoop platforms offer functions and features that enable big data applications to develop, operate, deploy and manage big data streams. Such platforms must be able to extract data streams, process the data and transmit it as a single stream. Various tools and technologies have been used to analyse big data. In response to the growing demand for big data streaming analytics, a variety of alternative big datastreaming solutions have been developed by both the open-source community and enterprise technology.

For instance, (Millman 2014) have discussed some factors to consider when selecting these tools for better data related judgements. They are:

- Shape of data
- Data access
- Availability and consistency
- Workload profile
- Latency check

TABLE 1: PAPER SELECTION STEPS

STEP 1	Apply search query to all sources, collecting results
STEP 2	Exclude invalid and duplicate articles
STEP 3	Apply inclusion / exclusion criteria to article titles, keywords and abstracts
STEP 4	Apply inclusion / exclusion criteria to introductions and conclusions
STEP 5	Examine the selected studies by applying inclusion / exclusion criteria to the text, if applicable

RQ 2. What are the prevailing software framework's structure methods for large data structures in Hadoop?

Apache Hadoop is best known for being the open-source implementation of Google's MapReduce computational model. Being an open-source solution, Hadoop has captured the attention of academic and industry communities. However, the main reasons for its wide adoption are related to its inherent characteristics, namely scalability, fault tolerance, and the ability to perform parallel and distributed computations across low-end compute clusters.

Figure 1. (White 2015) gives a layout of the spread of the Hadoop framework.

The framework includes the following tools, most of them which is deployed by Apache:

HDFS (which is a storage)

So HDFS is high bandwidth clustered storage and it is very useful, what happens here is that we put a suppressed file in our Hadoop cluster, HDFS is broken into blocks, and then spread across the nodes of our cluster and is very useful furthermore we have a fault tolerant concept what is being done here is that HDFS configures the replication factor (which is configured to 3).it sure has 3 copies of each block that makes up this file, which is spread across the entire node in our cluster. It's very useful and important because if we lose a node, it has a sense of what data was in the node. and I will replicate the blocks that were on that node.

Map Reduce

Well, as the name suggests, how we get data through Map Reduce is a two-step process. There's a mapper and the reducer programmers write the mapper function that comes out and tells the cluster which data point we're using. The reducer takes over all data and the added. Hadoop acts as a batch processing tool. Here we are working with all

of the data in the cluster so we can say that Map Reduce is working on all of the data in our clusters. There is a myth that you have to understand Java to get out of clusters completely. In fact, Facebook's engineers created a sub-project called HIVE, which is a SQL interpreter. Facebook wants a lot of people to write ad hoc jobs for their cluster, and they don't force people to learn Java, so the Facebook team-built HIVE, now anyone familiar with SQL can data extract from the cluster.

Hive:

It is the other Data access project extraordinarily famous like pig. Hive is a manner to assignment systems directly to facts internal a cluster so it's miles virtually a database. Data-warehouse constructed on top of Hadoop and it includes a question language Hive QL like SQL to hive question language and it's miles extraordinarily just like sq. Hive is comparable aspect likepig. It converts those queries into map lessen jobs that receives submitted to cluster.

Mongodb is extremely popular, especially with programmers, as it is really very easy to use. It is a document storage model, which means that programmers can take data models and clone them. We call objects in these applications and serialize them directly in Mongodb and with the same ease you can put them back into practice with.**HBase** was based on the google big table which we can use to create a table that has millions of rows in it and we can put indexes on it and we can do some serious data analysis and HBase is a data analysis indexing on them and ready for a high performance of trying to find the data we're looking for and the good thing about HBase is Pig and Hive native to with HBase.

Table 2 provides the tools for large big data stream analysis.

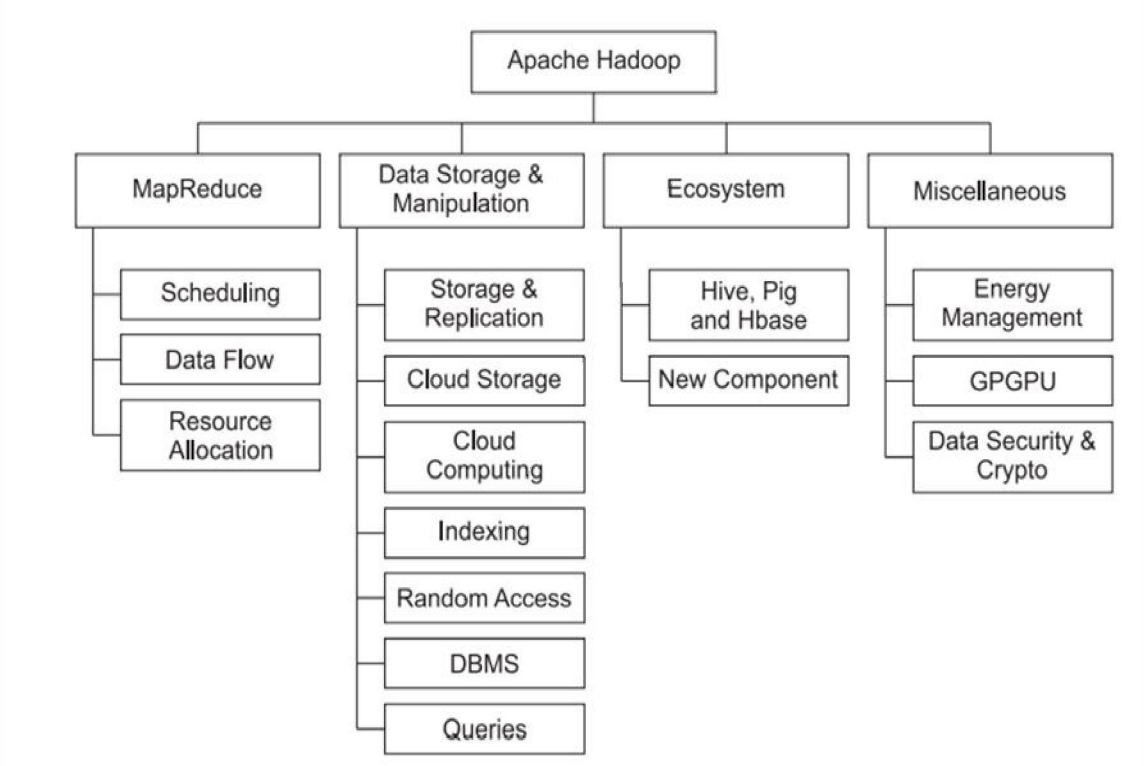

Figure 1. Entire Hadoop Framework Layout (source: (White 2015))

TABLE 2.TOOLS AND TECHNOLOGIES FOR BIG DATA ANALYSIS

TECHNOLOGY	Article
APACHE SAMZA	(Liao 2016)
APACHE STORM	(Liao 2016)
PHOTON	(Liao 2016)
APACHE AURORA	(Avita 2013)
YAHOO! S4	(White 2015)

Discussion

We can now discuss the results of our search and classification of the selected studies. The main point is overview of Hadoop's evolution and introduce its key features, which have changed over time. discuss an overview of the selected studies.Since its first release, Hadoop has changed steadily and significantly in 59 releases - the last was version 2.2.0 over six years of development. In addition to bug fixes, which are contained in each of these releases, new functionalities and modules were developed and integrated, whereby the framework was consolidated as the main solution for processing large amounts of data with the MapReduce paradigm. Hadoop 2.0.x was first released in May 2012. In theory, although still supported in versions 2.x, the gap between version 1x tends to widen and in the future backward compatibility in version 2x will no longer be supported. infrastructure in terms of use and flexibility.

Conclusion

The Apache Hadoop framework has been widely adopted by both industrial and research communities. MapReduce has established itself as the paradigm of distributed computing. The early adoption of Hadoop by several large companies (Yahoo!, Facebook, eBay, Amazon, Adobe, among many others) caught the attention of many IT companies processing data using the MapReduce paradigm due to its key concept of parallel and distributed abstraction.

REFERENCES

1. Avita Katal, Mohammad Wazid, R H Goudar (2013) “Big Data: Issues, Challenges, Tools and GoodPractices”. In IEEE, Contemporary Computing (IC3), Sixth International Conference, pages 404-409, Noida.
2. Davenport, T.H., Barth, P. and Bean, R. (2012) How'bigdata'is different.
3. Liao X, Gao Z, Ji W, Wang Y. (2016) An enforcement of real-time scheduling in Spark Streaming. 6thinternational green and sustainable computing conference, IEEE.
4. M. A. Beyer and D. Laney (2012) The importance of ‟big data‟: A definition, Gartner, Tech. Rep.
5. Millman N. (2014) Analytics for business. Computerworld. 2014. https://www.computerworld.com/article/2475840/bigdata/8-considerations-when-selecting-big-data-technology.html.
6. Russom, P. (2011) Big data analytics. TDWI best practices report, fourth quarter, 19(4), pp.1-34.
7. S. Sagiroglu and D. Sinanc (2013) "Big data: A review," International Conference on Collaboration Technologies and Systems (CTS), pp. 42-47, doi: 10.1109/CTS.2013.6567202.
8. Sadiku, M.N., Olaleye, O.D., Ajayi-Majebi, A. and Musa S.M. (2021), Future of Big Data. IJTRD volume 8(5).
9. Tsai, C.W., Lai, C.F., Chao, H.C. and Vasilakos, A.V. (2015). Big data analytics: a survey. Journal of Big data, 2(1), pp.1-32.
10. White T.(2015*) Hadoop-The Definitive Guide*, 4th Edition.