IJCSC
0973-7391
Volume 7 • Number 2 March 2016 - Sept 2016 pp. 80-84 available online at www.csjournalss.com

Comparison of MapReduce and Spark Programming Frameworks for Big Data Analytics on HDFS

Jai Prakash Verma[1], Atul Patel[2]
[1]Assistant Professor, CSE Department, Institute of Technology, Nirma University, Ahmedabad, India
[2]Professor & Dean, CMPICA, CHARUSAT University, Changa, Gujarat, India

Abstract: Use of internet and all the types of computer automated systems generates large amount of data in different forms. Due to large volume, different types of varieties, and high velocity of this type of data emerges the Big Data Problem. Spark and MapReduce programming frameworks provide an effective open source solution for managing and analyzing the Big Data. Today researcher are comparing both the frameworks and making many interpretations which also generates many misconceptions about the performances and efficiency. In this paper we are discussing the working model of both the programming frameworks and by experimental analysis, we are also finding that Spark is three to four times faster than MapReduce paradigm on single node implementation of Hadoop Distributed File System.

Keywords: Big Data Analytics, Hadoop, Mahout, MapReduce, Spark Framework.

1. INTRODUCTION

Big Data cannot be a considered as a precise term with a proper definition. Instead, it is a characterization of the infinitely long accumulation of various kinds of data, mostly unstructured. This represents data which is too vast that relational database management systems won't be able to analyses such data, because of the sheer size and unstructured nature. Any business should be capable of making use of data in the right way to create value on the table – value that can create better financial resources for the company or better experiences for the customers. Big Data Analytics provide a more precise solutions to researchers and analysts to use the previously unknown and unusable large amount of data available. Using advanced analytics techniques such as predictive analytics, text analytics, statistics, data mining, machine learning, and natural language processing, businesses can apply on these types of previously unknown and unused large dataset to find more valuable and useful insight which helps enterprises to make right decision as right time.Open source technology like Hadoop/MapReduce and Spark provides an effective solution for Big Data Analytics[10, 11, 12]. In this paper we are discussing the comparison between Spark and MapReduce framework on Hadoop Distributed File System.

In in this paper Section –II covers related work done by different researchers to compare these two programming frameworks. As per environment setup and data size work presented in this paper is completely different from these papers. Section-III discusses the concepts of Big Data and Big Data Analytics. Section-IV covers open source solution Apache Hadoop with MapReduce Framework for storage and retrieval of large scale dataset. Section-V discusses about Mahout, the machine learning algorithm implemented in MapReduce programing concepts witch can run in distributed manner with Hadoop framework. Section-VI and VII discusses about Spark programming paradigm and MLlib library available with Spark for machine learning algorithms. Section –VIII produces the results and discussion generated by experimental study done with different size file with Spark and MapReduce programming framework on HDFS with single node Hadoop implementation.

DOI: 10.090592/IJCSC.2016.113

2. RELATED WORK

Aaron N. Richter, Taghi M. Khoshgoftaar, Sara Landset, and Tawfiq Hasanin [5] proposed acomplete multidimensional examination of different open source devices likes Mahout, MLlib, H2O, and SAMOA for machine learning with huge information. An assessment standard is proposed alongside correlations of the structures talked about these open source technologies.

Satish Gopalani and Rohan Arora [6] gives the analysis between Hadoop Map Reduce and the as of late presented Apache Spark utilizing a standard machine learning calculation for K-Means clustering.

Juwei Shi, YunjieQiu, Umar Farooq Minhas, Limei Jiao, Chen Wang, Berthold Reinwald, and Fatma O¨ zcan [7] assess the major compositional segments in MapReduce and Spark systems including: merging, execution time, and storing, by utilizing an arrangement of critical investigative workloads.

Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter* and Tawfiq Hasanin, [8] gives a rundown of criteria to making determinations of devices for Big Data Analytic alongside an investigation of the focal points and downsides of each.

Jai Prakash Verma, Bankim Patel, and Atul Patel, [9] give execution of information investigation utilizing Hadoop Framework for content dataset.

3. BIG DATA AND BIG DATA ANALYTICS

There are endless supply of both structured and unstructured information lying in the World Wide Web and this information could have answers to a considerable measure of inquiries, which haven't been encircled in the bothers of business pioneers. On the other hand these organizations don't figure out how to utilize Big Data to the best point of their interest, the open source technologies like Hadoop and Spark can be used to finding knowledge for the organizations. Organizations must discover powerful approaches to find knowledge from the hugeamount of information being created each moment by making sense of what is in it and what to do with it. And there have been a few advancements in technologies which will add to the organizations making benefit out of Big Data.

These Big amount of data allude to five dimensions that are volume, velocity, variety, veracity and value. Volume: Enterprises are overwhelmed with regularly developing information of different kinds, effortlessly accumulating Exabyte—even Zettabytes—of data [1,2,3,4]. According to [4] 12 terabytes of Tweets made every dayinto improved product sentiment analysis, Convert 350 billion annual meter readings to better predict power consumption.Velocity:Data is increasing at unparalleled speed. We should think of what will happen in next coming years. It is a challenging task. Variety:As the increasing in technology, the data that comes today is in all type of formats. For example, audio, video, picture, numeric and structured data in traditional database. It is also becoming problem for the organizations nowadays.

Veracity: As per the different variety of data business analyst can trust on the information and knowledge generated. Because some time this types of data are not authentic as well as some time this types of data are sponsored or spam. As per [4] 1 in 3 business leaders don't trust on the information or knowledge generated by computer automated systems for making decisions. Value: As value dimension researchers try to identify importance and usability of this huge amount of Big Data for enhancing business and living standard. Big amount of biometric data also follows these five dimensions, that why we can solve storing, managing, and retrieving issues using Big Data problem.Big data analytics, it is the process in which big data is examined that contains various kinds of data of various data types. The analysts find the hidden pattern in the big data and research the solution.The example of big data analytics done can be through text analysis, data mining etc.The steps involved in Big Data Analysis are: 1) Data Acquisition and Gathering 2) Data Cleaning and Extraction 3) Data Aggregation and Integration 4) Analysis and Modelling 5) Interpretation.

4. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

Apache Hadoop/MapReduce is an open source solution for handling this large amount of data known as Big Data. Hadoop/MapReduce framework processes Big Data in parallel and in a fault tolerant manner. As per MapReduce programming paradigm a map/reduce job spilt the input data into number of memory chunks, and assign to different map tasks for processing completely in parallel manner. The MapReduce framework is responsible for sorting the output of map tasks and fed as input to the reduce tasks [1].

Apache Hadoop project includes following modules:

a) Hadoop Common: The common utility libraries which help to interact with other Hadoop components.
b) Hadoop Distributed File System (HDFS™): HDFS is a distributed file system which helps processes to access application data in high-throughput, in a fault tolerant and parallel manner.
c) Hadoop YARN: This framework is responsible for job scheduling and cluster resource management during a processes execution.
d) Hadoop MapReduce: It is a YARN-based system which responsible for parallel processing of large amount of data sets known as Big Data.

.

5. MAHOUT WITH MAPREDUCE FRAMEWORK

Apache Mahout is a machine learning and data mining framework for classification, clustering, Frequent Itemset and recommendation which can run on Hadoop/MapReduce framework in distributed manner. It is an open source, scalable machine learning libraries which can be used for analyzing big data on a distributed manner. Apache Mahout

Software includes three major features: First: Mahout is a simple and extensible programming environment and framework for building scalable algorithms. Second: Mahout provides a wide variety of premade algorithms comparative to Scala + Apache Spark, H2O, and Apache Flink. Third: Mahout is an establishedset of libraries which implemented classification, clustering, Frequent Itemset and recommendation algorithms that can run on Hadoop MapReduce environment.

6. SPARK FRAMEWORK AND ARCHITECTURE

Apache Spark (figure-1) is an alternative open source solution for Big Data Analytics which can be work in distributed manner in the way of data stoppage as well as analysis. Apache spark introduce RDD (Resilient Distributed Dataset) which provides an application programming interface centered on a data structure. RDD is an immutable fault-tolerant, distributed collection of objects that can be operated on in parallel. In MapReduce cluster computing paradigm each map and reduce phase disk read and write operation were performed which causes extra delay in execution. Spark's RDDs provides the functionality where programmer can write a program in a distributed manner that offers a (deliberately) restricted form of distributed shared memory because RDD helps to persist the data in RAM and help for effectively process these data. It is the technology that efficiently utilize in-memory LRU cache with possible on-disk eviction on memory full condition. And it puts all the dataset data on the local file systems during the “shuffle” process [13,14].

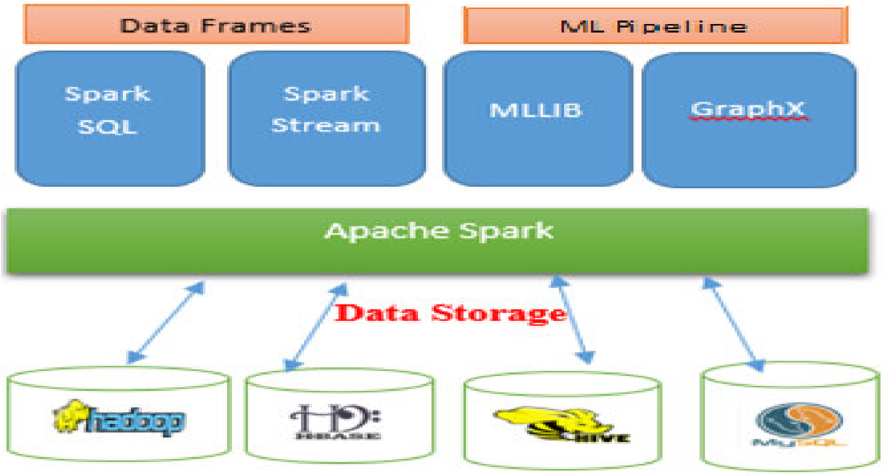

Figure 1:- Spark Architecture

7. MACHINE LEARNING LIBRARIES (MLlib) WITH SPARK

Spark additionally includes libraries for doing things like machine learning, streaming, graph programming and SQL (see the figure -1). This additionally included libraries makes things much simpler for programmer and developer. These libraries are incorporated, so many enhancements in Spark as per time period which gives advantages to the additional packages as well. Most data analyst would somehow or another need to fall back on utilizing heaps of other random tools to complete their work, which makes things complex. Spark’s libraries are intended to all work together, on the same bit of information, which is more coordinated and less demanding to utilize. Spark streaming specifically gives an approach to do ongoing stream handling. The previously stated. Spark will empower programmer to do real-time analysis of everything from data available with an enterprise. The data may be trading data or web clicks, Spark provide an ease and faster open source solution to handle the Big Data problem for an enterprise [13,14].

8. EXPRIMENTAL CASE STUDY

For experimental analysis we are running wordcount program with different sizes files on Spark and MarpReduce Framework. Here we are running both the open source tools on single node Hadoop installation on Ubuntu machine.

A. Dataset Selection: Here we are using a large text file that represents customer review and feedback about the different products. We split the same file in different sizes for performance analysis.

B. Methodology and Algorithm: Here we executed wordcount program written on MapReduce pragramming paradigm and another wordcouknt program written on spark programming paradigm. Both the open source tools were running on single node Hadoop Distributed File System implemented with Ubuntu operating system.

C. Findings:Table - 1 shows the CPU execution time for wordcount program on both the programming frameworks. Figure -1 shows the comparison of spark and mapruduceprogram execution. Here the graph shows that Spark is three to four times faster than mapreduce framework. Figure – 2 shows the graph between CPU execution time for both the tools with different sizes files.

DOI: 10.090592/IJCSC.2016.113

TABLE I. PERFORMANCE EVALUATION FOR WORDCOUNT PROGRAM

Sr. No.	Text File Size	Text File Size (in Bytes)	MapReduce Framework	Spark Framework	%
1	153.3 kB	156979.2	3.15	1.053078	2.991231
2	260.1 kB	266342.4	3.67	1.110998	3.303336
3	358.4 kB	367001.6	3.52	1.132921	3.107013
4	769.3 kB	787763.2	4.5	1.240024	3.628962
5	1.0 MB	1048576	5.06	1.38958	3.641388
6	2.8 MB	2936012.8	7.08	1.770073	3.999835
7	5.1 MB	5347737.6	8.83	2.123277	4.158666
8	10.5 MB	11010048	10.46	3.132935	3.338722
9	15.6 MB	16357785.6	12.5	4.874238	2.564503
10	26.1 MB	27367833.6	16.81	5.329568	3.154102
11	55.1 MB	57776537.6	26.83	9.652699	2.779533
12	100.3 MB	105172173	54.86	20.168503	2.720083
13	255.9 MB	268330598	131.46	33.377112	3.938627
14	329.3 MB	345296077	161.76	49.030878	3.299145
15	1.2 GB	1288490189	610.65	147.019231	4.153538
16	1.6 GB	1717986918	791.09	187.673476	4.215247
17	2.8 GB	3006477107	1395.84	337.269451	4.138649
18	4.7 GB	5046586573	1731.95	550.205804	3.147822
19	7.5 GB	8053063680	3604.45	835.431703	4.314476

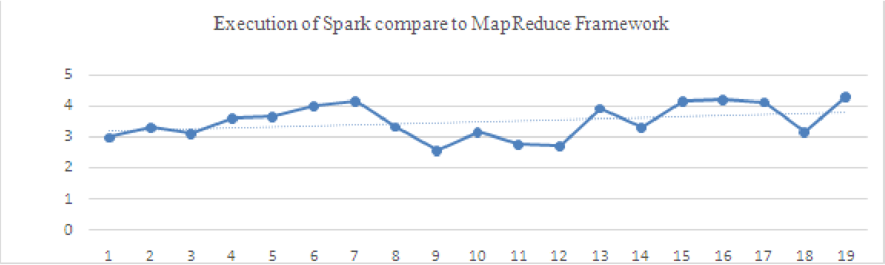

Figure 2:- Comparison of Wordcount program execution between Spark and MapReduce Framework

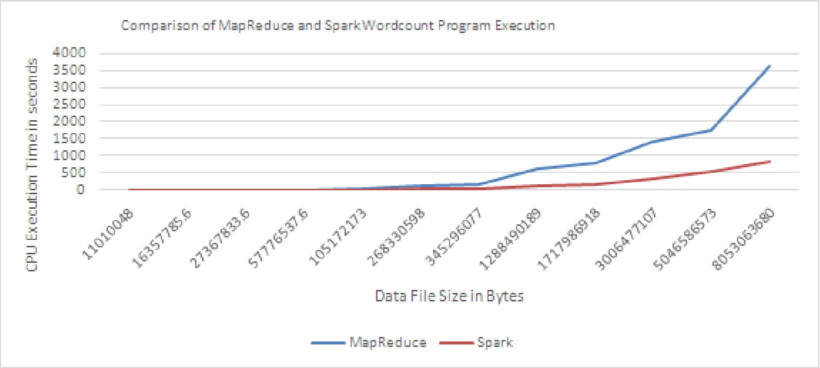

Figure 3: Comparison of MapReduce and Spark Wordcount Program Execution

DOI: 10.090592/IJCSC.2016.113

D. Discussions:

As per architecture and workflow of Spark and MapReduce frameworks, Spark is much faster than MapReduce because it perform IO (Input/Output) operation for read and write data on HDD once at the times of shuffles but MapReduce programming model perform at the time of every map and reduce task. In MapReduce workflow there are series of map and reduce jobs, each of which perform data read and write operation to HDD between iterations. With the supports of DAGs and pipelining Spark performthe complex workflows without intermediate data materialization (unless you need to “shuffle” it) Caching. In general the concept of in memory execution enhance the performance of Spark over Hadoop/MapReduce framework. Here the meaning of in memory execution is caching the read and write operation without interaction with HDD. As per experimental study done with Spark on single node Hadoop implementation in section-8 shows Spark can be executed 3 to 4 times faster than MapReduce programming framework (Figure -2).

The key difference between Spark and MapReduce programming concept is spark’s in memory execution of read and write operation require as the time of shuffling of data. For every MapReduce job data are read from HDFS and write to HDFS in each data shuffle iteration but in spark it performs once.

9. CONCLUSION

Spark and MapReduce framework both are popular distributed computing paradigm for providing an effective solution for handling this large amount of data called Big Data. Today there are many misconception about the comparison of spark and mapreduce framework. Many researchers and bloggers are claiming that Spark is 10 to 100 times faster than mapreduce framework. In this paper we are also comparing both the programming frameworks. We have run wordcount program on both the programming framework with different size of files and found that Spark is 3 to 4 times faster in single node Hadoop implementation.

REFERENCES

[1]. Alex Holms, “Hadoop in Practice”, 2012, Manning Publications co.

[2]. Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman, “Mahout in Action”, 2012, Manning Publications co.

[3]. Pete Warden, “Big Data Glossary a guide to the new generations of data tools”, 2011, O’Reilly.

[4]. “Online Shopping touched new heights in India in 2012”. Hindustan Times. 31 December 2012. Retrieved 31 December 2012.

[5]. Aaron N. Richter, Taghi M. Khoshgoftaar, Sara Landset, Tawfiq Hasanin, “A Multi-Dimensional Comparison of Toolkits for Machine Learning with Big Data”, 2015 IEEE 16th International Conference on Information Reuse and Integration

[6]. Satish Gopalani and Rohan Arora, “ Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means”, International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 1, March 2015

[7]. Juwei Shi, Yunjie Qiu, Umar Farooq Minhas, Limei Jiao, Chen Wang, Berthold Reinwald, and Fatma O¨ zcan, “Clash of the Titans: MapReduce vs. Spark for Large Scale Data Analytics”, Proceedings of the VLDB Endowment, Vol. 8, No. 13 Copyright 2015 VLDB Endowment 2150 8097/15/09

[8]. Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter* and Tawfiq Hasanin, “A survey of open source tools for machine learning with big data in the Hadoop ecosystem”, Landset et al. Journal of Big Data (2015) 2:24 DOI 10.1186/s40537-015-0032-1

[9]. Jai Prakash Verma, Bankim Patel, and Atul Patel, “Big Data Analysis: Recommendation System with Hadoop Framework”, 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 978-1-4799-6023-1/15 © 2015 IEEE

[10]. Anindita A Khade, “Performing Customer Behavior Analysis using Big Data Analytics”, 7th International Conference on Communication, Computing and Virtualization 2016

[11]. Victor Changa and Gary Wills, A model to compare cloud and non-cloud storage of Big Data, Future Generation Computer Systems 57 (2016) 56–76

[12]. Yi-Cheng Huang, Wenwey Hseush, Yu-Chun Lai, Michael Fong, “BigObject Store: In-Place Computing for Interactive Analytics”, 2014 IEEE International Congress on Big Data

[13]. Web Content Available, [https://www.mapr.com/blog/parallel-and-iterative-processing-machine-learning-recommendations-spark?source=Email&campaign=Spark%20Nurture&utm_medium=Email&mkt_tok=eyJpIjoiWWpreVlqTmpZVGcxTWpZeSIsInQiOiJXZW5PalFsb3YwdGZXZnd1VThiZVVVZVQyUU5ZV0REWTdTVGd0OFA5TWYrcW5yMGZUd2dUT0tWbk9lbnViWEFlQjluVHdrNmdubGlLV0V5K2dTNzVJOU1ZYjlKRTVzak11RTE1Wm9CdEV1cz0ifQ%3D%3D&aliId=23187144] as on date - 11-06-2016

[14]. Web Content Available, [https://www.mapr.com/blog/5-minute-guide-understanding-significance-apache-spark], as on date - 11-06-2016