ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

Future of Big Data Application & Apache Spark vs. Map Reduce

Keerti[1], Kulvinder Singh[2], Sanjeev Dhawan[3]

Department of Computer Science &Engineering, Kurukshetra University, Haryana, India

Abstract-Now a days, Apache project is working in a new system for social networks and healthcare system that is Apache Spark. It is a fast & expressive cluster computing engine system. It is very compatible with Apache Hadoop. Its version 1.2 just get released in December 2014, it requires very less code work than Hadoop and Map reduce. It works in memory system; it rises to become the most active open source project in Big Data. Apache spark can work on multiple platforms. API supports multiple languages like Java, Python, Scala. It is great for small to medium gigs of data. Spark can be only stack you need i.e. No need to run multiple cluster (Hadoop Cluster, Strom Cluster).

Keywords**:** Big Data, Hadoop, ETL, Apache Spark, Map Reduce**.**

1. INTRODUCTION

Eric 14 invented a case study of Apache Spark. Apache Spark is the most widest & important thing happening in Big Data now a days. Spark and Hadoop are great together for the data science. As, everybody aware about the study of Apache hadoop & Big Data. Over from past few years, it is widely used in Data Science. In this section researchers will discuss the concept that how you can use the big data applications in real time system, researchers can use it in social network areas and in health care system. Main frame data become more familiar with Apache Spark. IBM recently working with Apache Spark.

IBM+SPARK = BLAZING FAST ANALYTICS

When Apache Hadoop had first created it was two important innovations, one was scale and storage system the google file system, where reseachers can store the any kind of data that is expensive and very reliable. L. Huang *et al* [13] has applied some computation to balance production as well as performance.

In this paper, let's evaluate the Future of Big Data Application & Apache Spark Vs. Map Reduce. The main objective of this paper is to make comparison between Apache Spark & Map Reduce. This paper is organized as follows: First, let's introduce the map reduce and hadoop and spark. Second section discussed about literature reviews of map reduce and researchers will make its comparison between map reduce and spark. Third section will cover some proposed work with that how big data will work for so many applications of Spark.

2. LITERATURE REVIEW AND COMPARISON TABLE

Apache spark is widely used in organizations to process large datasets. So many large-scale production companies like IBM is working on Spark. It runs everywhere like on Mesos, standalone, on loud and hadoop. It is very easy to use; authors can write its applications easily in Python, Java, Scala.

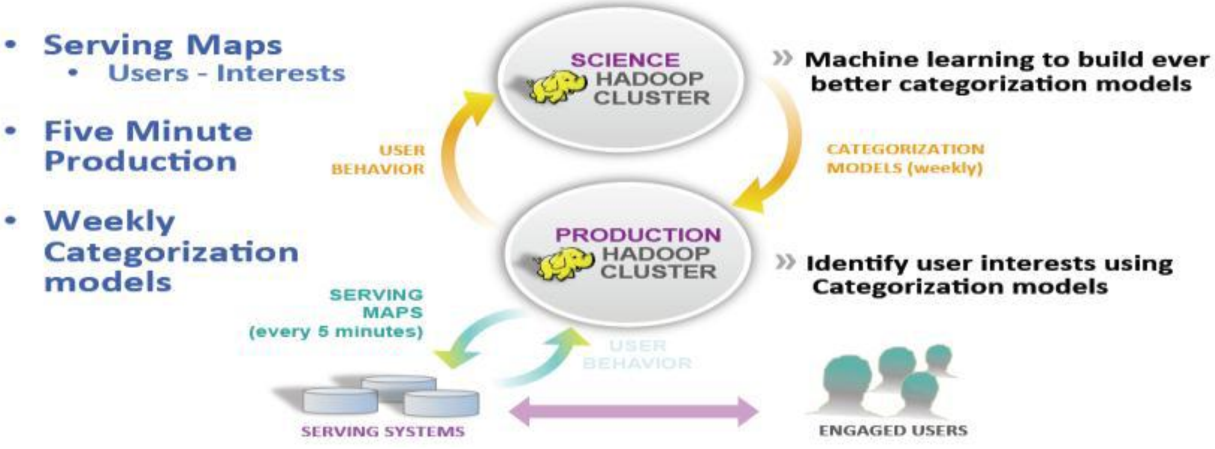
• Serving Maps
• Users - Interests
• Five Minute Production
• Weekly Categorization models
SCIENCE HADOOP CLUSTER
» Machine learning to build ever better categorization models
USER BEHAVIOR
CATEGORIZATION MODELS (weekly)
PRODUCTION HADOOP CLUSTER
» Identify user interests using Categorization models
SERVING MAPS (every 5 minutes)
USER BEHAVIOR
SERVING SYSTEMS
ENGAGED USERS

Fig. 2.1 Build Customized Home Pages With Latest Data (thousands/second)

It runs programs 100 times faster Map reduce in memory. It is basically used in data sciences. It increases its focus in ETL, because science needs data in the right format and place for better job scheduling. Spark has a good speed as it allows compelling interactivity. It is an open source code so it can run well in many languages platforms like cloud and Hadoop.

Spark is used in applications now a day like in 3rd party applications. It has few classes of applications and these are discussed below:

- Custom Solutions: Basically used in internal applications.
- Data Science Tooling: Its collaboration and Reporting. In it research will collaborate all
- The data in fastest rate and report it in output layer.
- Vertical specific Applications: In marketing, retailing, financial, healthcare and gaming. Here authors are discussing about the case study of Yahoo! Homepage.

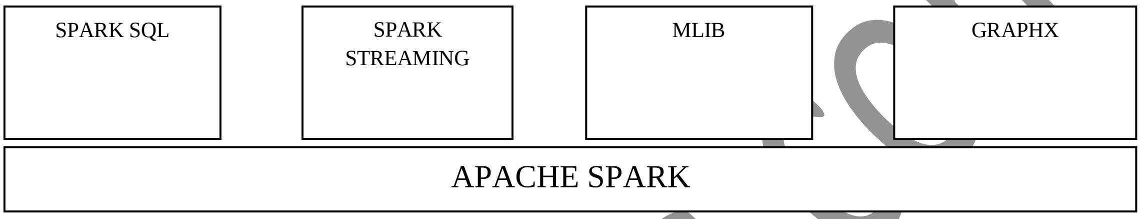

Fig. 2.2 Spark Structure

Table-2.1 Comparison Between Mapreduce And Spark

S. No.	Difference	Map Reduce	Apache Spark
1	Storage	Distributed Storage + Distributed Compute	Distributed compute only
2	Framework	Map Reduce Framework	Generalized Computation
3	Data	Data on disk	Data in Memory
4	Platform	Not ideal for iterative work	Great at iterative work
5	Speed Req.	Batch process	100 times faster for data in memory
6	Coding	Ubantu, Mat lab	Java, Python, Scala

3. ACTIVITY DIAGRAM

The main advantage of using spark is that it supports broad open data source community. Which develops interactive API eases development. Spark has a good speed and runs well at many environments like cloud, hadoop. It needs a very less code set about 7X times less code or you can say 350K code lines comparing to 55K. It is shown below by some activity diagram:

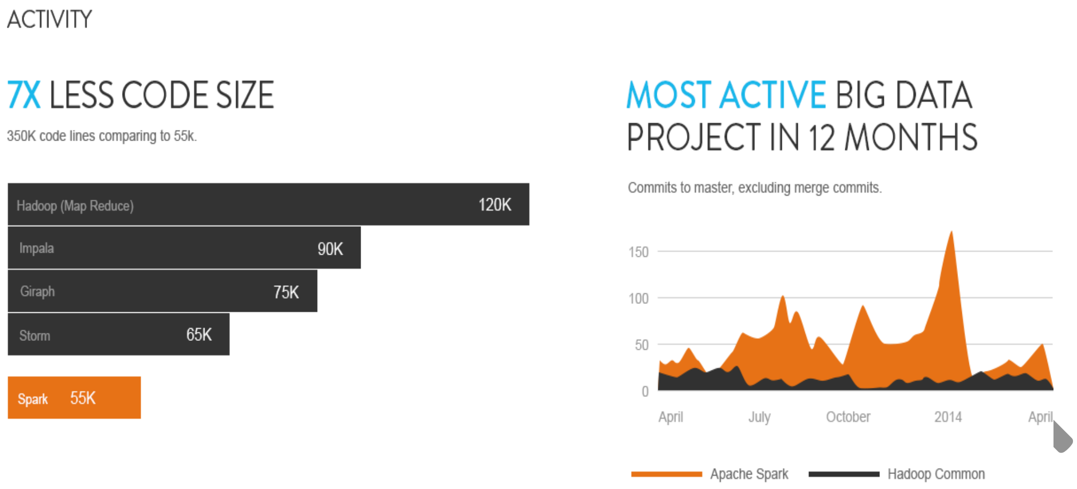

Fig. 3.1 Spark Code / Activity

4. BIG DATA APPLICATION WITH SPARK

Big data application model is explained below:
IMO @Apache Spark is the most exciting thing occurring in big data today. Better uses of tired storage RAM, SSD and Disk. In fig.4 we have explained Big Data Application Model which has three layer structures that is Interactive, streaming, batch layer. The process and layer working is shown in the figure.

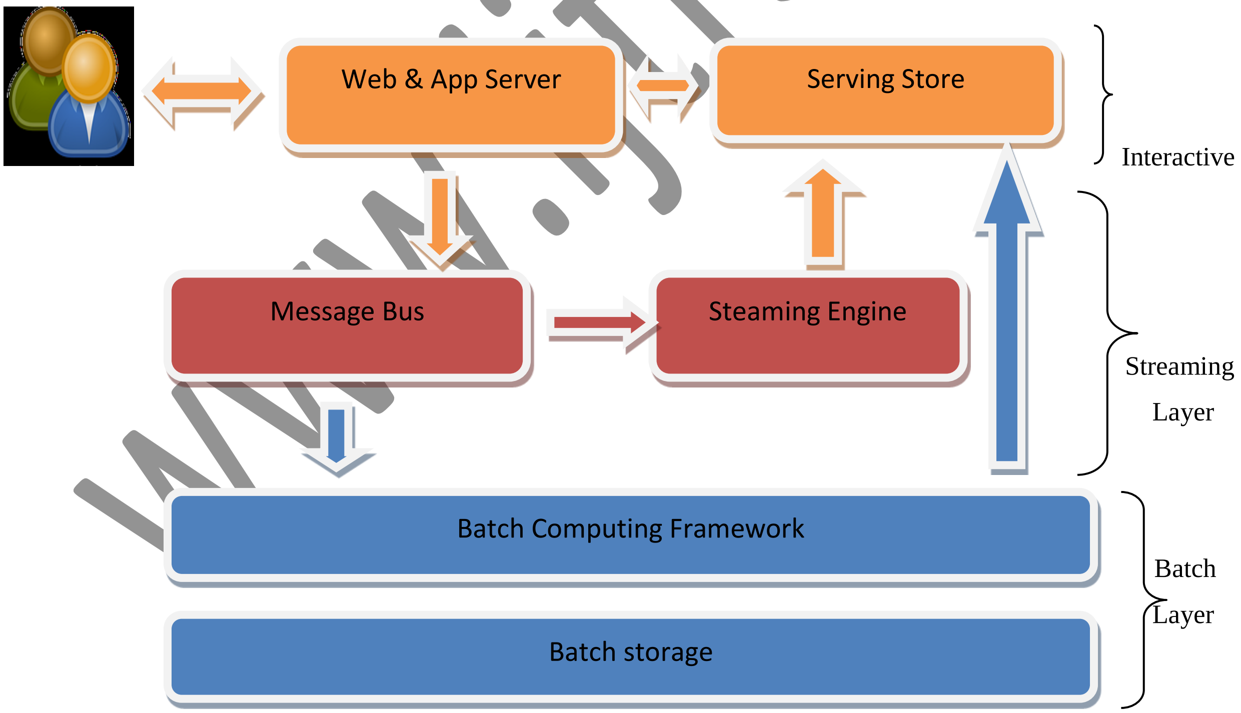

Fig. 4.1 Big Data Application Model

CONCLUSION AND FUTUREWORK

The application performs the operation on big data with spark like counting average speed, and code requirement etc. in most constructive time and producing an output with minimum consumption of resources. The data investigation and handing out is used in a social networking application. Thus providing the mandatory in sequence to the application users with slightest effort. Authors can perform this application by using data processing algorithm and can also implement this for healthcare system

ACKNOWLEDGEMENT

First and foremost I would like to give my particular thanks to Dr. Kulvinder Singh for his guidance and monitoring through the work process. Additionally would like to thanks Assistant Professor Dr. Sanjeev Dhawan for inspiration and motivation that triggered me to this work.

REFERENCES

[1] Dean, J., Ghemawat, S. “MapReduce: Simplified Data Processing on Large Clusters,” Mag. Commun. ACM 50th anniversary, vol. 51, issue 1, 2008, pp.107-113

[2] Ibrahim S., Hai Jin, Lu Lu, Bingsheng He and Song Wu (2011), “Adaptive Disk I/O Scheduling for Map Reduce in Virtualized Environment”, Proc. Fourth IEEE Int’l Conf. on Parallel Processing (ICPP’ 11), pp. 335-344.

[3] ElniketyEslam, Elsayed Tamer and Ramadan E. Hany (2011), “iHadoop: Asynchronous Iterations for MapReduce”, Proc. Third IEEE Int’l Conf. Cloud Computing Technology and Science (CloudCom’11), pp. 81-90.

[4] Yanxin Zhang, Bin Gong, Ying Peng and Hui Liu (2011), “ParallelOption Pricing with BSDEs method on MapReduce”, Proc. Third IEEE Int’l Conf. Computer Research and Development (ICCRD’11), pp. 289-293.

[5] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary R.Bradski, Andrew Y. Ng, and KunleOlukotun (2007), "Map-Reduce for Machine Learning on Multicore", Proc. Int’l Conf. NIPS, pp. 281-288.

[6] J. Ekanayake, S. Pallickara and G. Fox (2008), "Map-Reduce for Machine Learning on Multicore", Proc. IEEE Int’l Conf. on eScience.

[7] O’Reilly Media (2011), “Big Data Now”, pp. 17-75, O’Reilly Media,Inc.

[8] Tom White (2009), “Hadoop: The Definitive Guide”, pp. 41-450,O’Reilly Media, Inc.

[9] J. Talbot, R.M. Yoo, C. Kozyrakis, and Phoenix++: modular MapReduce for shared-memory systems, in: Proc. of the Second International Workshop on MapReduce and its Applications, MapReduce 2011, pp. 9–16.

[10] J. Baker et al., ``Megastore: Providing scalable, highly available storage for interactive services," in Proc. Conf. Innov. Database Res. (CIDR), 2011, pp. 223_234.

[11] P. Bhatotia, A. Wieder, and R. Rodrigues, et al. Incoop: MapReduce for incremental computations, in: Proc. of the 2nd ACM Symposium on Cloud Computing, SoCC 2011.

[12] A. Labrinidis and H. V. Jagadish, ``Challenges and opportunities with big data," Proc. VLDB Endowment, vol. 5, no. 12, pp. 2032_2033, Aug. 2012.

[13] S. Chaudhuri, U. Dayal, and V. Narasayya, ``An overview of business intelligence technology," Communication ACM, vol. 54, no. 8, pp. 88-98, 2011.