Review Analysis on Handling Big Data UsingHadoop

NidhikaChauhan[i]DiwakerBadoni[ii]

Assistant Professor,Business Manager,

University Institute of Computing Aaxiom Technology,

Chandigarh University, Gharuan.ChukkuwalaDehradu,

Nidhi29.chauhan@gmail.comDiwaker.badoni @aaxiomtechnology.com

Abstract

Objective: The paper focuses on reviewing concepts of Hadoop framework.

Method and statistical analysis: Secondary data is reviewed to find the meaningful insight related to concept.

Findings: Study examined the issues related to data management. The traditional methods and processing application are not efficient in handling such data. These vast arrangement of information are utilized to get extra data from investigation of a solitary substantial arrangement of related information, when contrasted with independent littler sets with a similar measure of information, demonstrating the connections to be found to centre around business patterns, avert maladies, battle wrongdoing et cetera.

Application/Improvements: Every day, a very huge amount of data is created and the problems it faces where this bulky data should be stored and how to analyse it. The simple way to understand the concept of big data is through the small example of weather forecast what do you think how they predict the upcoming climate change it is done through the collection, processing, and analysis of data. This is Big Data which help to draw the conclusion and make these predictions.Hadoop is one way to manage and manipulate this data

Keyword: Big Data, Hadoop, Name Node, Data Node, MapReduce.

1. Introduction

Hadoop is the infrastructure that helps us to manage this big data. It's like multiple machines act as mainframe computer which were basically used to handle the big data mainly used in larger organisations. But these systems were expensive and it was difficult for everyone to afford them. So there had to be some alternative Hadoop is one of the alternatives. The functioning of hadoop system is similar to parallel processing of computer where multiple devices connect together to work on volumes data. Role of hadoop is to bring together machines into a computer cluster.

When we talk about large data we mean bulky data set. These data sets contain data which cannot be managed, analysed and processed using traditional mediums. So the technology has worked upon new aspects to provide the solution for the mentioned problems. They came up with the concept of Hadoop. Now just imagine your facebook account has so many thinks like audio, video, comments, reviews, images, likes, dislikes etc. in order to mention this data we need to have a system but as the data varies and has volume the process is complex for traditional methods. Here is when Hadoop comes into play.(Elgendy, 2014)

Literature Review

Ha Lee.et.al(2011) emphasises on data processing tools. The key focus is Map Reduce technique and author do conduct various surveys. He introduces different views of Map Reduce technique. (D.Rajasekar, 2015)

B. Patel.et.al(2012) has performed various experiments on Big data problems. Finally they reached the conclusion that Hadoop system would be the most appropriate solution for all these problems. (Patel, 2012)

Shilpa.et.al(2013) discusses about BigData, components of Hadoop and the algorithms used for MapReduce. (Shilpa, 2013)

Sethy.et.al(2015) introduces the concept of big data and then it concentrate of Hadoop system and map reduce technique. The main idea is to implement Hadoop system and handle node failure. (Sethy, July 2015)

Verma.et.al(2015) discusses about techniques to handle large amount of data. Some open source software's are also discussed. (Hooda, 2015)

Rajasekar.et.al(2015) highlights some concepts of Big data, the essential tools and the techniques used for processing, analysing and managing the bulky data. The author has strongly focused on HDFS and MapReduce methods. (D.Rajasekar, 2015)

Jach.et.al(2015) explains the big data needs to be properly stored and managed. A comparative experiment was performed on both MySQl and Hadoop and according to the results Hadoop system is most preferred. (Lee, 2011)

Jha.et.al(2016) the findings gives an overview of important information to the researchers about the trends of big data and its domain. (Jha, 2016)

Ahmed G.et.al(2017) Hadoop system works towards lessning and processing of massive data. Performance of Hadoop is impressive and it is changing modern world. (G, 2017)

Architecture

Data Distribution

In Hadoop we store the data in clusters. A cluster is designed for storing and analysing large amount unstructured data and provide high-throughput access. The data is distributed to the clusters once it gets loaded. The Hadoop Distributed File System (HDFS) divides the large data files into chunks. Once the chunks are created they are managed by various notes of cluster. The chunks are replicated in nodes so as to prevent data lose. The main purpose of Hadoop Distributed File System was to withstand fault-tolerance, provide scalability and

distributed system for storage. Each record is separated into line or into the configurations particular to the application rationale. Each procedure running on a hub at that point forms a subset of these records. The Hadoop system at that point plans these procedures.(Jha, 2016)

HDFS

A **Data Nodes** is a machine in the cluster. A file can be made of several segments/blocks, and it's not necessarily that these blocks are stored on the same machine; the machines which stores each block are selected randomly on a block-by-block basis. In order to access any file on the node we need to connect to or may need to request the machine.(Sethy, July 2015)

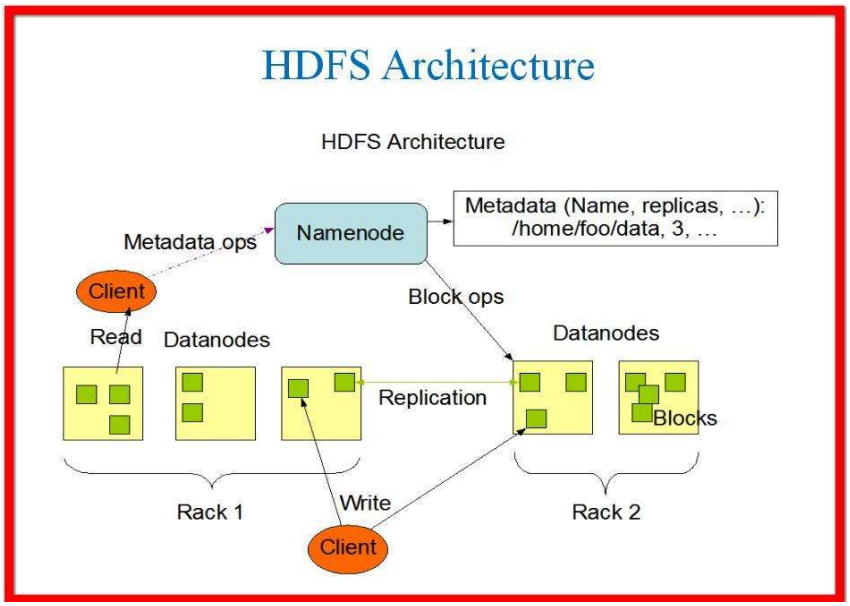

Figure 1 Data Node(Menon, 2016)

NAME NODE

The Name Node is the central part of an HDFS file system. It is also said to be the master node who directs the slave node data node to perform low level tasks. It keeps the all records how the files are broken down into blocks. This node does not store the data or perform any computation of these files.

DATA NODE:

Data Node as the name clearly indicates that this node contain the data it means data is stored in these data nodes. There can be more than one Data Node, with contain the replicated in them. They are basically part of slave machine and take the orders from name node. Data node can read and writes the HDFS file.It communicates with other data nodes in order to replicate data. All the backups are maintained here.

MapReduce:

It refers to two tasks performed by Hadoop. First is outline in which an arrangement of information is taken and after that is changed over to another arrangement of information where each component is separated into key/esteem match. Second is diminish work it accepts the yield from delineate information and join the information key/esteem match into a littler arrangement of keys.Reduce job is always done after the map.(G, 2017)

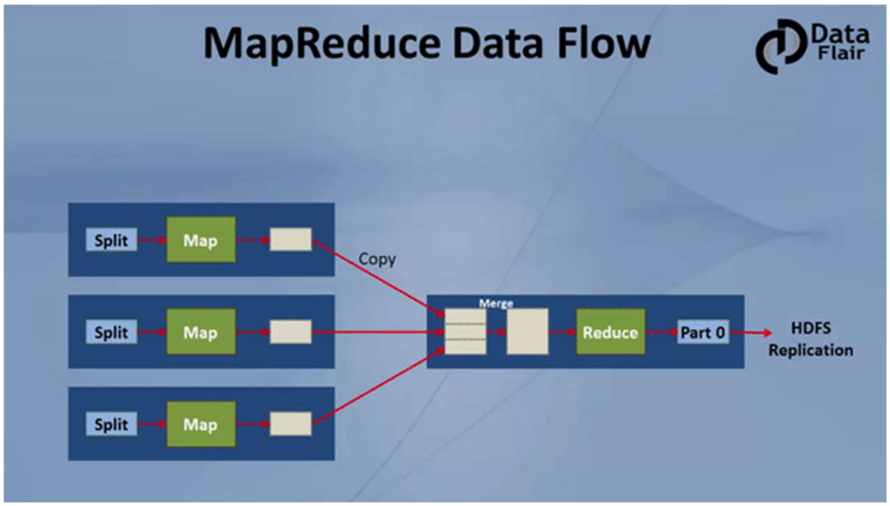

Figure 2MapReduce and HDFS(Team, 2016)

MAPPING LISTS

Mapping is the initial step of a MapReduce sequencer. Mapper transforms each data component independently to a yield information component.

REDUCING LISTS

Reducing is done to total qualities together. A reducer work gets an arrangement of info admirations from an information list. At that point these qualities are joins together and they restore a solitary yield high regard.(Patel, 2012)

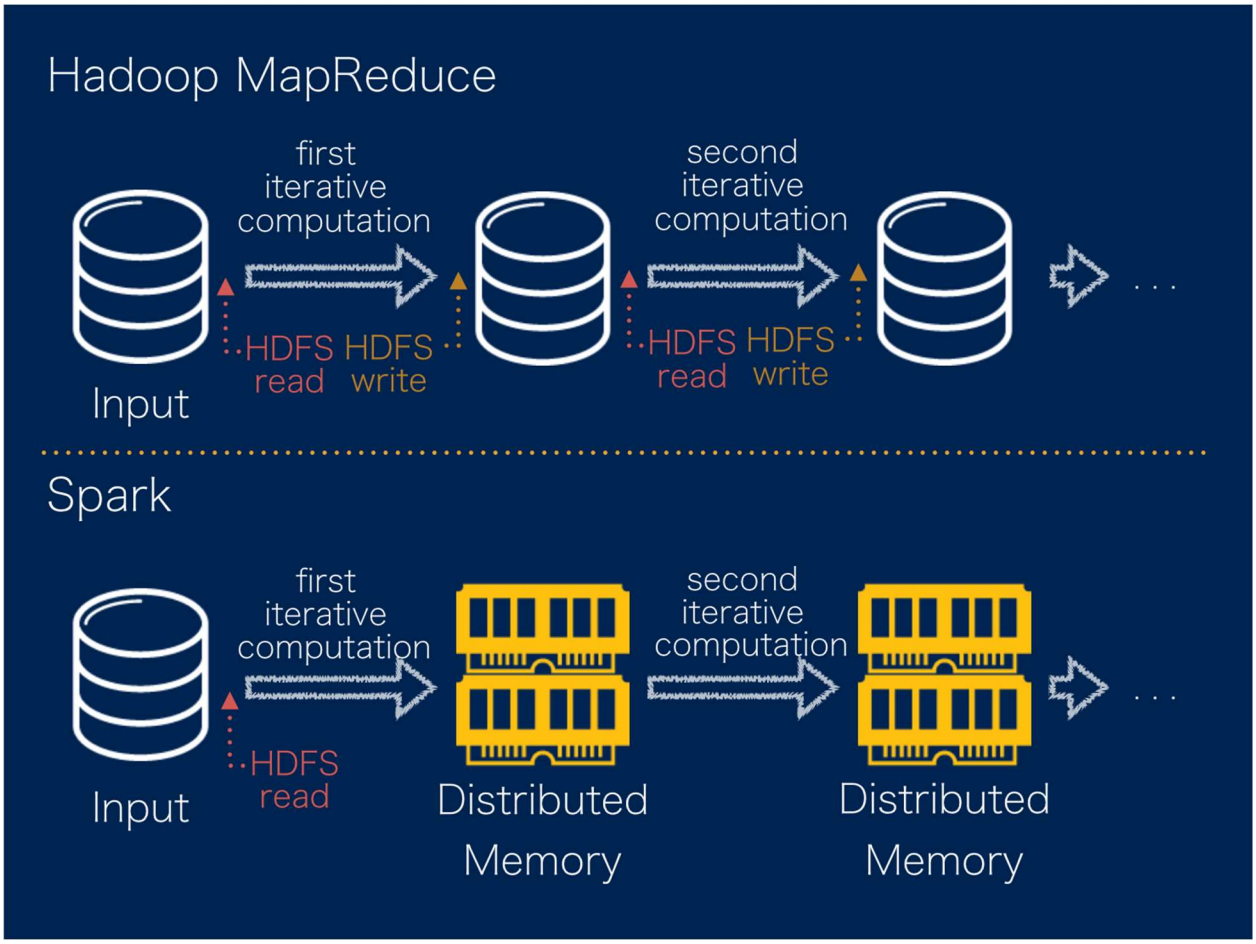

Figure 3MapReduce (J, 2016)

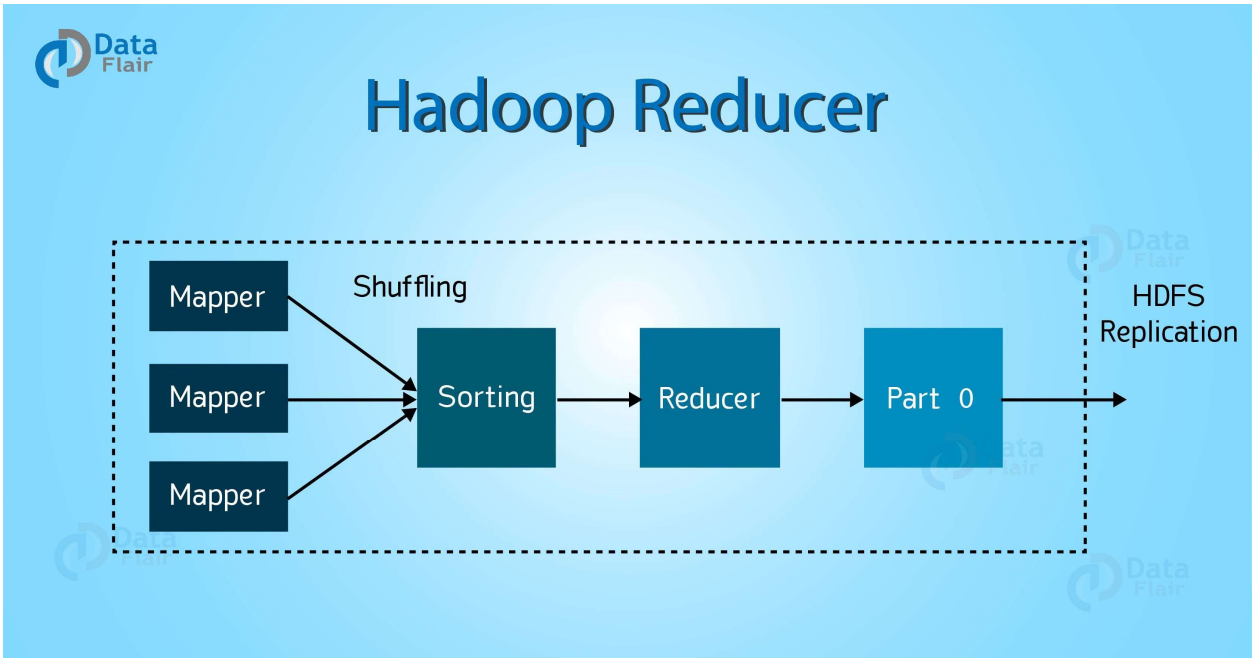

Figure 4 Reducing Function (Team, 2016)

Applications

Companies invest large amount of time and money to store data. The stored data could be used to handle large problems. The data analysis is can be a beneficial part of the business. Some of the examples are listed below.

1. To analyse a risk
   It could be use to analyse any kind of life threatening risk. Be it a natural disaster, a medical emergency etc.
2. Prevention and identification of security breach
3. Prevention of hardware failure.
4. Analysing consumer feedback. (Sethy, July 2015)

Conclusion

Hadoop give the facility to manage the huge data. Data which has to be analysed, studied carefully and then used to make the predictions and computation. The basically it is the concept

of breaking down huge data into smaller sets and later managed by scheduling. This framework is considered to be fault tolerant, reliable and manages thousands of nodes. The problems are broken into pieces and later solved parallel y through HadoopMapReduce approach. This system is the revolution for handling the larger data.

References

D.Rajasekar, C. S. (2015). A Survey on Big Data Concepts and Tools. *International Journal of Emerging Technology and Advanced Engineering, 5*(2).

Elgendy, N. (2014). Big Data Analytics. *researchgate* .

G, S. A. (2017). Mining on Big Data Using Hadoop MapReduce Model . *ICSET*.

Hooda, Y. V. (2015). A Review Paper on Big Data and Hadoop . *International Journal for Scientific Research & Development| , 3*(02).

J, T. (2016, March). .*lab41*. Retrieved from https://www.lab41.org/transformers-rdd-in-disguise/: https://www.lab41.org/transformers-rdd-in-disguise/

Jha, A. (2016). A Review on the Study and Analysis of Big Data using Data Mining Techniques. *International Journal of Latest Trends In Enginaring and Technology* .

Lee, K. H. (2011). Parallel data processing with MapReduce: A survey. *SIGMOD Record, 40*(4), 11-20.

Menon, R. (2016, July 24). *technospirituality*. Retrieved from http://technospirituality.com/2016/07/hadoop-in-5-minutes/: http://technospirituality.com/2016/07/hadoop-in-5-minutes/

Patel, A. B. (2012). Addressing big data problem using Hadoop and Map Reduce. *IEEE*.

Sethy, R. (July 2015). Big Data Analysis using Hadoop: A Survey . *International Journal of Advanced Research in Computer Science and Software Engineering , Volume 5,* (Issue 7,).

Shilpa. (2013). Big Data and Methodology . *International Journal of Advanced Research in Computer Science and Software Engineering , 3*.

Team, D. (2016, nov 23). *Hadoop MapReduce Tutorial – A Complete Guide to Mapreduce*. Retrieved from https://data-flair.training/blogs/hadoop-mapreduce-tutorial/: https://data-flair.training/blogs/hadoop-mapreduce-tutorial/