The Analytics of Big Data and Hadoop Architecture: A Literature Review

Anchal Pokharana, Krishan Kant Sharma, Samiksha Sharma

Chandigarh University, India

Corresponding author: Anchal Pokharana, Email: Anchalpokharana258@gmail.com

We are living in an immensely created, specialized time, where web is getting to be major need of all people. Today, our social, individual and in addition proficient life is rotating around internet. Along these lines, bringing forth Big Data at a staggering force. Conventional administration apparatuses and systems are demonstrated, un-reasonable while managing Big Data. Big Data has actually been coined for and represents those useful informational attributes of data whose volume, speed and distinctiveness make them complex to be caught, worked with or processed upon. So in order to manage this large data, Hadoop platform comes into picture. Hadoop is an open-source platform that ensures massive data management, data regulation and also takes care of heavily enlarged data applications running in assembled systems. It is at the point of convergence of creating organic network of colossal data which is on a very basic level and is used for advanced assessment exercises, including data mining and AI applications.

Keywords: Bigdata, Hadoop Framework, HDFS, MapReduce, Hadoop Component.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 195–201. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-20

1. Introduction

BigData is an emerging term that depicts volumes of data in big measures which is of organized, semi-organized and completely scattered format. The 3 famous V's because of which Big Data gets a complete description are Volume (that corresponds to the length, breadth & height of data), Velocity (which caters to the rate of change in speed at which data gets converted into useful information) and Variety.

a) Variety

Data and information come in varied formats like audio, video ranging in different extensions like .mp3, mpeg, mp4 which constitutes the variety feature.

b) Velocity

As already discussed that velocity is measured by the rate of change of speed by which data gets transformed into useful facts i.e. information. Such information streams are managed with the help of RFID labels, sensors and other technologically advanced methods.

c) Volume

Volume defines how much data weighs. For example, the data on a particular hard drive occupies 100gb of data. The size of the information and data falls under the Volume category.

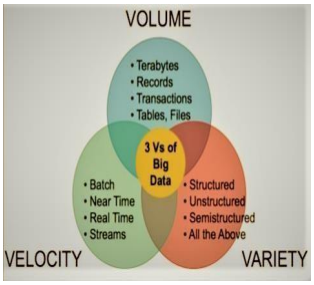

Fig.1: 3 V's of Big Data

2. Management Issues with Bigdata Using Traditional Approach

Big Data is about larger volumes, fast and organized dataset. Relational data base management system do not gel anymore with the current trending technologies due to its inefficient methods in dealing and managing a huge amount of data. RDBMS works better in the case of homogeneous datasets but when comes to the variety and distinctiveness of data, RDBMS is not a wise optionas it becomes quite tedious and complex for the former to manage voluminous data. Be that as it may, Big Data can't be bargain by such an inflexible administration frame work having bunches of if's and buts'. Another disadvantage of RDMBS is that information is investigated in light of connections. In Big Data keeping up connection between unstructured information (pictures, recordings, Mobile created data, RFID and soon) is by outlandish. Aside from above, Big Data Analytics ought to have quick preparing pace like continuous or close to ongoing, which RDBMS doesn't ensure. Along these lines, NoSql with conveyed document frameworkcanbeasuperiormethodologyfordissectingBigData. Last yet not the slightest adjusting,

keeping up and incorporating Big Data is expensive arrangement when we bargain Big Data with conventional methodologies.

3. HADOOP: A necessity to handle Big Data

Hadoop has good scalability over a plenty of hardware centre points and bulks of important information as it continues to execute on group of item servers. A dispersed (forwarding) archive structure is planned to give brisk data access over the center points in a group, notwithstanding accuse lenient capacitiessothatapplicationscancontinuetorunifindependent centers crash and burn. Hence Hadoop is quite popularized in datamanagementandoperatingforoveralongerperiodoftime say in the mid of2000's.

HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) is the most necessary information stockpiling/storage framework utilized by Hadoop applications. It utilizes a NameNode and DataNode engineering to execute a disseminated document framework that gives elite access to information across profoundly adaptable Hadoop groups.

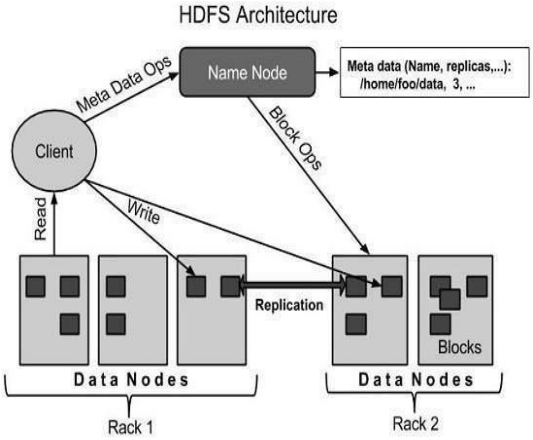

Fig.2: HADOOP Architecture

Being an eminent piece of Hadoop ecosystem system advancements, Hadoop provides a dependable way to oversee pools of enormous information and supporting related big data analytics. Good throughput is achieved by parallel provision of information access. HDFS engineering is comprehensively isolated into following three hubs which are Name Node, Data Node, HDFS Clients/Edge Node.

1. What is NameNode?

A midway set hub, which contains data related to Hadoop record framework. NameNode in HDFS Architecture is otherwise called Master hub. HDFS Namenode stores meta-information i.e., number of information squares, reproductions and different subtle elements. This meta-information is accessible in memory in the ace for quicker recovery of information. Name Node keeps up and deals with the slave hubs, and allocates

undertakings to them.

1.1 *Tasks of NameNode*

- First task is to process the regulation of customer entry into thedata.
- Next task is to perform operations over the entered data like shutting, opening, writing etc.
- It guarantees that the DataNodes are alive. A square report contains a rundown of all squares on adatanode.
- Name Node ensures that no replicated squares and data should be there. That is information must beunique.

2. What is DataNode?

DataNode in HDFS Architecture is otherwise called Slave. In Hadoop HDFS Architecture, Data Nodes to resgenuine information in HDFS. It performs read and compose task according to the demand of the customer. DataNodes can convey on ware equipment.

2.1 *Tasks of DataNode*

- As guided by the name node, datanode looks after the actual execution and operations over the data for example designing, alteration and cancellation of blocks.
  - DataNode oversees information stockpiling of the framework.
  - DataNodes send heartbeat to the Name Node to report the well being of HDFS. As a matter of course, this recurrence is set to 3seconds.

4. MapReduce

MapReduce is a Java based programming model which aims at management and access of large volumes of information in distributed computing environment. MapReduce is the core centric part of Hadoop in data handling. As the name suggests, MapReduce algorithm has two major phases which are named as Map and Reduce. In the map phase, similar data is mapped which is achieved by further reduction of the individual elements into tuples or key pairs. The data which is homogenous gets mapped. In the Reduce phase, it follows a pipeline process where the confined output from the Mapper phase becomes the input of Reduce phase and it combines the mapped tuples or mapped key pairs thereby reducing the size of the data into smallersets.

MapReduce is advantageous as it's scalability is considered relatively very high in data processing. Mappers and Reducers are the names of the processing primitive features of Map Reduce Algorithm.

5. MapReduce Architecture

- Job Configuration sets various parameters like input format, input locations, map & reduce functions etc for the Hadoop framework. This is provided by the user and it basically configures thetasks.
- When the data gets transferred from the mapper to reducer hubs, the input format, output and locations, map and reduce aim at data controlling at both thesides.

InputFormatandInputlocationscomeintopicturewhendata is brought into the system. Input format frames the data arrangement for Map reduce and Input locations provide the address of the particular datafile.

- Map Function maps the individual key pairs and it marks the mapping of tasks.
- Reduce Function combines all the single tuple and key pairs to give the data a confined reduced format.
- Output Format houses the resultant tuples and key pairs as an output and output location provides the area and address of the final output files.

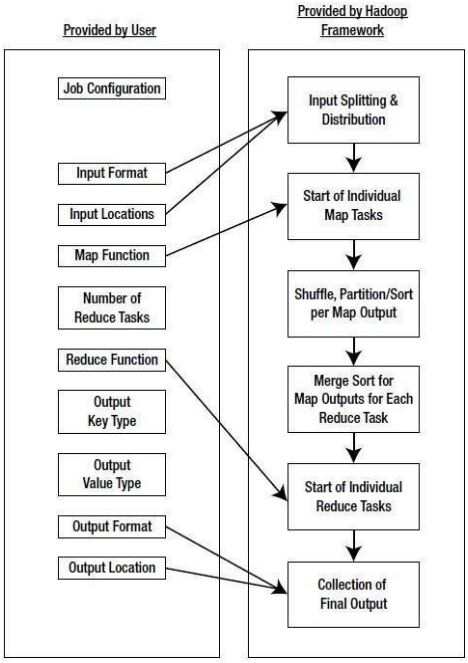

Fig.3: MapReduce Architecture

6. Example Explaining MapReduce Methodology

Referring to figure 4, there are a couple of words and expressions. The motive is to locate and map the quantity of each word event to the other. Let's consider the input words **Cat, Rat and Sea** which are passed into the Hadoop framework. These three words are three separate keys (tuples) and they are called a sinput formats. Their input locations or address will also be passed along with them. We have to locate the quantity of every word event. There are two hubs, mapper and reducer. When **Cat, Rat and Sea** are passed into Map phase, then first of all they will split into individual tuples/keys. Values will continue to split to such an extent till then they get converted into a single value /word. Now every atomic word will be mapped to its corresponding similar value using Merge sorts or any other sort algorithm. Once the values are correctly and completely mapped, they will be sent to the reducer hub. Now, at the reducer hub tuples with the similar key values will be sent to similar cluster. At the Reduce phase, it can be observed that we have two tuples of "**Rat"** and it has been combined into a single cluster. Similarly, we have three tuples of "**Bike**" which are then later combined to similar single hub.

Resultantly we receive 4 output files and then they are finally aggregated as output **"Rat, Bike, Cat, Sea".**

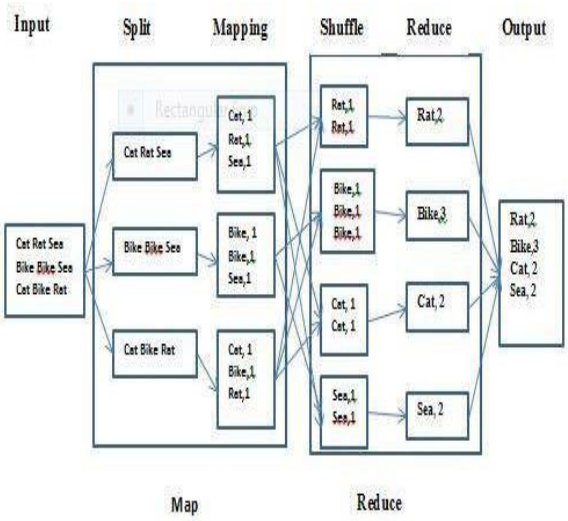

Fig.4: MapReduce Example

7. Conclusion

This paper talks about the need of development of approaches to deal with tremendous measure of information which is flooding at a disturbing rate. To oversee and control information, Big Data Hadoop and MapReduce philosophies have been talked about quickly. A model has additionally been represented that shows how MapReduce capacities. Hadoop engineering and MapReduce Architecture have additionally been examined.

As Big Data Analysis is still in its outset organize, we are certain that this paper causes the specialists to more readily comprehend the ideas of Big Data and its preparing. Enormous Data will bring a noteworthy social change. In spite of the fact that programming dialects like R, SPSS are advancing for Big Data investigation additionally examine is as yet required to guarantee uprightness, security for the substantial informational collections being prepared. Huge Data Analytics ought to be misused for maintainable and unprejudiced society

References

[1] Varsha B.Bobade, " Survey Paper on Big Data and Hadoop**",** IRJET,2016
[2] Hadoop Distributed File System, http://hadoop.apache.org/hdfs
[3] HadoopTutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html
[4] J. Dean and S. Ghemawat, "Data Processing on Large Cluster", OSDI '04, pages 137–150, 2004
[5] HadoopMapreduceTutorialhttps://www.dezyre.com/hadoop- tutorial/hadoop-mapreduce-tutorial-
[6] TheHadoop"https://searchdatamanagement.techtarget.com/definition/Hadoop

[7] V. S. Patil and P. D. Soni,"HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS ",International Journal of Application or Innovation in Engineering & Management (IJAIEM) Vol . 2(2), Feb 2013,pp. 247-250

[8] Hadoop Wiki,"Apache Hadoop", Accessed :http://wiki.apache.org/hadoop

[9] Juniper Networks,(2012),"Introduction to Big Data:Infrastructure and Networking Considerations", JuniperNetworks.Accessed:http://www.juniper.net/us/en/local/ pdf/whitepapers/2000488-en.pdf

[10] D. Harris,(2013),"The history of Hadoop:From 4 nodes to the future of data".Accessed: https:/gigaom.com/2013/03/04/the-history-of-hadoop-from-4- nodes-to-the-future-of-data/

[11]Vance, Ashlee (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times. Archived from the original on 11 February 2010. Retrieved 2010-01-20.

[12] Michael J. Cafarella, Web.eecs.umich.edu. Retrieved 2013-04-05.

[13] Intellipaat. "Hadoop Creator goes to Cloudera". Intellipaat Blog. Retrieved 2 February2016.

[14] Wibibonblog,"A Comprehensive List of Big Data Statistics",Acessed : http://wikibon.org/blog/big-data-statistics/

[15] Vidyasagar S. D," A study of Hadoop in information Technology Era", S.D,Global Research- Analysis,

[16] Prity Vijay, Bright Keshwani, "Emergence of Big Data with Hadoop : A Review", IOSR Journal of Engineering (Vol. 06, Issue 03 (March. 2016)