Analyzing the Concept of Big Data using Hadoop's MapReduce with HDFS

Gurpinder Singh
Assistant Professor, Panjab University SSG Regional Centre, Hoshiarpur, India.

Amandeep Kaur
Assistant Professor, National Institute of Technical Teachers Training & Research, Chandigarh, India.

Tanvi Sharma
Assistant Professor, Panjab University SSG Regional Centre, Hoshiarpur, India.

Abstract –Today's world is digital world. Data in this digital world is being generated from various resources e.g. retailer websites, social networks, transactions, digital data (images, videos, audios). The data we usually obtain is unstructured and semi structured. Size of data is too large e.g. 30 billion contents are being shared on Facebook every month. This plethora amount of Structured, Semi-structured, unstructured and hybrid data is known as Big data. Today's data is diverse and large in amount that comes and stored at high speed. It can be gold mine but only if we could process that Data mine. A study done by IBM found that over half of business leaders today realize that they don't have access to insights; they need to do their jobs. To resolve this situation, processing of Big data is only option and the way to do so is Map Reduce (programming paradigm). This Map Reduce we use with File system name HDFS (Hadoop Distributed File System) to store and process Big data. Main focus of this paper is to understand Hadoop Project for data processing through MapReduce framework over HDFS.

Index Terms – Big Data, Hadoop, MapReduce, HDFS.

1. INTRODUCTION

The term Big Data means plethora amount of data which is combination of structured, unstructured and semi-structured data which has been growing exponentially. So basically, Big data is combination of three terms: Volume, Variety and Velocity. Volume means data we are producing nowadays. According to IBM, every day we create 2.5 quintillion bytes of data and 90% we have been created in last two years. Variety means different kind of data we are generating (structured, unstructured and semi-structured and hybrid data) that comes from comment on social networking sites and websites, emails, monitoring data from sensors, documents etc. . Velocity means rate at which we are producing, storing and processing data. So for all this, Analyzing and processing the Big data is difficult for today's structured Databases and software , but if we could process that data then it will be beneficial for business community and society for better understanding the environment to make better and accurate decision for growth.[1]

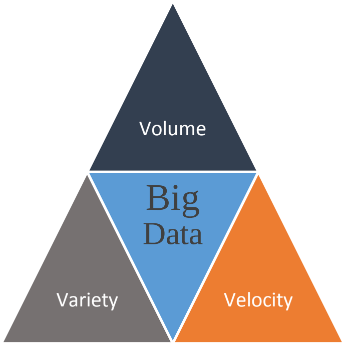

Figure 1 Big Data

2. APPLICATION

Manufacturing and Exploration industry:

Most of companies using sensor network to collect data of activities in industry, to exploring natural resources using sensors. Example of this is Petrochemical industries that are using Big data collected by sensor network because companies can lose millions of dollars by exploring on wrong places. Data they collect are low frequency waves a seismic wave that moves through earth crust due to tectonic plate movement. We use sensors and place these on specific spots in industry, on earth etc. to collect data. SHELL with Hewlett-Packard is using sensor network made up sensors with fiber optics to collect data and send to servers maintained by Amazon and compare it with existing oil field data so that geologist could find accurate site to explore.[2]

E-commerce:

In today's world, everything is online. Like if you want to purchase anything, you don't have to go o the market. Everything is one click away. Most of retailing business is online today. Plethora amount of information is produced every day on social and retailer websites in the form of comments, like about products etc.. We just have to utilize that information to know the trend in market and psychology of customers, so that we can add new product and improve our existing according to customer's interest and region where we are doing business.

Medical field:

At present we are working on molecular level disease finding technique because it is root of our body functioning. Large amount of information is produced by this kind of research. Then to understand the human behavior (psychology) in particular region,we can use data on social websites. This large amount of data we can analyze with only Big Data concept because data ,we are producing is mostly unstructured and semi-structured and hybrid.

Banking System:

Online transaction, internet banking, pay by credit or debit card all these activates produce large amount of data regarding particular region's customer. Bank can use this data to analyze and made their strategy based on this. Interesting thing is that information produced by this is structured and semi- structured.

3. Issues of Big Data

Storage:

As we know, Big data means plethora amount of structured, unstructured and semi- structured and hybrid information, that is produced from various resources. If we want to use this data, then definitely we need storage space to handle this. That's why we need to pay for resources to store and manage Big Data.

Performance:

Most of data come at very high speed and delay of just nanosecond makes a difference e.g. retail sector where we have to compete with other retailer's offer to maintain our market's share. That's why we need fast processing Structure for Big data to make it accessible in less time.[3]

Flexibility:

Our platform for Big data should be flexible to handle structured, unstructured, semi-structured and hybrid information from different sources and then to analyze and manage that information.[3]

Privacy:

The privacy of data is one of the hot topics of Big Data that should be considered. Big data is concept of processing large amount of ever changing data to get analyzed output from different type of input resources. So providing privacy to that kind of data using current technique is quite difficult. Today, a lot of online sites need from us to share private information (e.g. social networking, Online shopping sites etc.), but in case of Big Data ,we can't understand what and how they are going to use our shared data.[4]

Human Source Data:

There are so many websites where we add up information, write down review for special purpose e.g. Wikipedia, review about product and places etc. So to detect error and make summary from that data which we are receiving from people is the responsibility of Big Data.[4]

4. Basic File System of Big Data

File system is the architecture to hold data files and provide access to those file with high throughput rate. HDFS (Hadoop Distributed File System) is the distributed file system which is used to hold large amount of files on different systems to counter failure and to get high throughput for Big Data. It works on concept of master and slave architecture. [5]

HDFS's idea is based on to minimize the seek time, for this we have to minimize the numbers of block of particular data file. But this will make transfer time longer (because large bloc size) than seek time. By default HDFS uses 64MB of block size but some HDFS setups use 128MB block size as according to their requirements. [6]

HDFS works using 2 types of nodes: Name Node and Data Node, in Master Slave network. Name nodes always act as Master Node and Data Nodes as Slave Nodes. Name Node is used to handle File Namespace, trees, directories and their metadata operations like open, close and change the name of files and directories. Data Nodes hold the actual blocks of data and handle operation of data accessing like read and write. Data nodes are also responsible for creation, deletion and replication to maintain availability of data blocks by getting instruction from Name Node. [6][7]

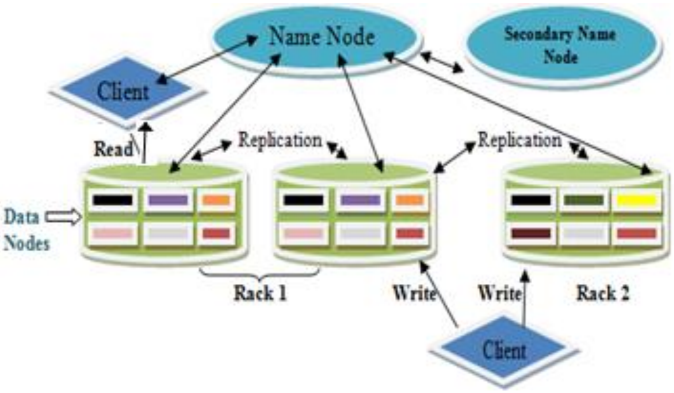

Figure 2 HDFS working

Name Node keeps the records of Data Nodes and blocks of data stored on those for mapping in cluster. Any user can access data Block by communicating to Name Nodes. Name

Node provides the information to user about data blocks in Data Nodes. By this Data Node send those blocks to user directly. Everything on Data Nodes are dynamic means change with time. So we have to update Name Nodes periodically to avoid inconsistencies. [5][7][8]

To tackle with the problem of fault or failure, HDFS setup uses Secondary Name Node that not acts as Primary Name Node But save log periodically from Primary Name Nodes. We always setup Secondary Name Node on different location rather than local. Whenever we lost Primary Name Node, Secondary Name Node act as Primary Name Node after getting metadata from Multipoint on different File Systems because only those points have current metadata. Next is replication of Data Blocks and store in different racks so that if we lost one or more blocks of data, we could get replica from somewhere another rack. One policy is to make three replicas of Data block i.e. replication factor 3; store one replica on local node in local rack, second on different node in remote rack and third one on different node in same remote rack. Replication factor decides the number of copies of each block. [7][8]

Name Node always gets heartbeats periodically (after every 3 seconds) to know whether Data Nodes and secondary nodes are active or not. If Name Node doesn't get heartbeats from any Data Node after 10 Minutes, then it declares that one dead, block communication and start search for replicas. Name Node gets Data Blocks report by every 10th heartbeat. Data Block report contains the information about count of replicas are less or greater than replication factor .Name Node starts creating or deleting replica based on Block report to balance the count of replicas. To decrease the Read latency and bandwidth use, HDFS always tries to select nearest replica to satisfy user's request by Fair Scheduling method which acts as a pool to provides on average fair amount of resources to each job. [6][7][8][9]

5. MapReduce

MapReduce is a processing technique uses programming model for Big Data. Apache Hadoop's MapReduce and HDFS were based on Google work of MapReduce and Google File System; this makes the task of big data analysis easy. [10]

It processes data parallely in distributed environment. MapReduce process is divided into three steps:

Map:

Input Data is transformed in the pair form of <key, value> after getting line by line by splitting.

Shuffle/Sort:

Pairs of same keys are tried to group together by shuffling and sorting by <key>.

Reduce:

At the end, sum up the values of same keys to get final result of input data.

Reduce process can be started before completing the Map process. Below is the figure to explain the working of MapReduce with example to count Alphabets:

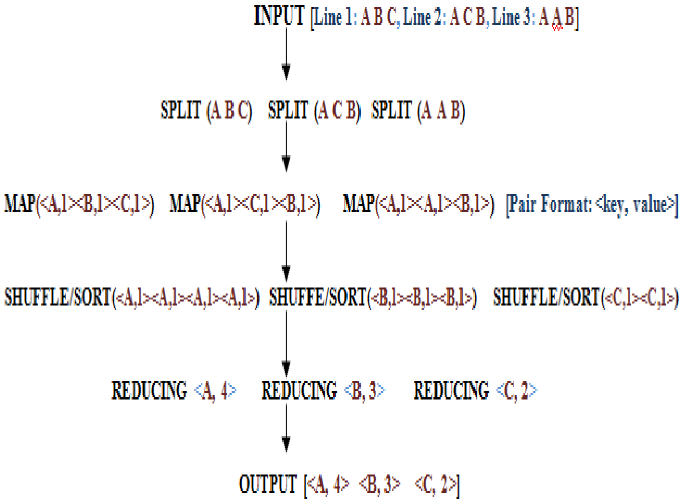
INPUT [Line 1: A B C, Line 2: A C B, Line 3: A A B]
SPLIT (A B C) SPLIT (A C B) SPLIT (A A B)
MAP(<A,1><B,1><C,1>) MAP(<A,1><C,1><B,1>) MAP(<A,1><A,1><B,1>) [Pair Format: <key, value>]
SHUFFLE/SORT(<A,1><A,1><A,1><A,1>) SHUFFE/SORT(<B,1><B,1><B,1>) SHUFFLE/SORT(<C,1><C,1>)
REDUCING <A, 4> REDUCING <B, 3> REDUCING <C, 2>
OUTPUT [<A, 4> <B, 3> <C, 2>]

Figure 3 Process of Word Count by MapReduce

6. MapReduce Architecture

MapReduce architecture is also based on Mater-Slave mode like HDFS. There are four components in this architecture: User, Job Tracker, Name Node, and Task Tracker. The process starts when User send job request to the Job tracker. Job Tracker runs on Master Node with Name Node in MapReduce. It is responsible for monitoring the MapReduce task executed on slave nodes by Task Tracker. Job Tracker communicates with Name Node about location of Data Node where data resides. Job Tracker then communicates with Task Tracker. Task Tracker has actually fixed number of slots for tasks. So it is the responsibility of Job Tracker to find out Task Tracker with free slots for running task. It is always beneficial to find Task Tracker in same rack where data is, to save bandwidth. This procedure is called Rack awareness. After this, Job Tracker sends the splited task to Task Tracker which executes MapReduce process on that. Task Tracker sends status report by heartbeats to Job Tracker after specific period of time. Information in the status report is about phase of task running on slave nodes by Task Tracker, free slots on Task Tracker after finishing task. If Job Tracker doesn't gets heartbeats after specified time, then it declares that Task Tracker dead and find out another Task Tracker to complete the task. at the completion of task, Job Tracker gets the status of success and notify user. After that user can get information.[11][12][13][14]

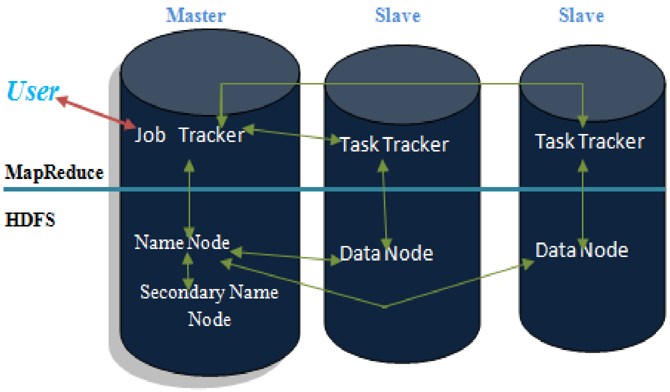

Figure 4 MapReduce and HDFS Working

7. CONCLUSION

Apache's Hadoop MapReduce and HDFS is open source platform used for processing unstructured, hybrid data that is available in abundance. Architectures of MapReduce and HDFS with the ability of fault tolerance/failure are explained in this paper. MapReduce (programming model) with HDFS (Storage unit) is the way to handle large amount of unstructured data by dividing it into independent chunks and process in parallelism to provide desirable result. This paper also explained issues, being faced during implementation of MapReduce with HDFS.

REFERENCES

[1] www.datastax.com/big-data-challenges

[2] www.forbes.com/sites/bernardmarr/2015/05/26/big-data-in-big-oil-how-shell-uses-analytics-to-drive-business-success

[3] Paul C.Zikopoulos, Chris Eaton, irk Droos,Thomas Deutsch,George Lapis ,"Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", IBM, 2011.

[4] http://cra.org/ccc/wp-content/uploads/sites/2/2015/05 /bigdatawhitepaper.pdf

[5] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M, "Efficient Analysis of Big Data Using Map Reduce Framework ",International Journal of Recent Development in Engineering and Technology (ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014.

[6] http://www.datanubes.com/mediac/HadoopArchPerfDHT.pdf

[7] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[8] V. Sajwan, V. Yadav, Dr .M. Haider,"The Hadoop Distributed File System: Architecture and Internals ", International Journal of Combined Research & Development (IJCRD), eISSN: 2321-225X, pISSN: 2321-2241, Volume: 4, Issue 3, April 2015.

[9] T. Cowsalya and S.R. Mugunthan ,"HADOOP ARCHITECTURE AND FAULT TOLERANCE BASED HADOOP CLUSTERS IN GEOGRAPHICALLY DISTRIBUTED DATA CENTER", ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Volume 10, No. 7, APRIL 2015.

[10] https://en.wikipedia.org/wiki/Apache_Hadoop#Fair_scheduler

[11] V. Sajwan1, V. Yadav," MapReduce: Architecture and Internals", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 4, Issue 5, May 2015.

[12] Madhavi Vaidya ," Parallel Processing of cluster by Map Reduce", International Journal of Distributed and Parallel Systems (IJDPS) ,Volume 3, No. 1, January 2012.

[13] K. Srikanth, P. Venkateswarlu, Ashok Suragala ," A FUNDAMENTAL CONCEPT OF MAPREDUCE WITH MASSIVE FILES DATASET IN BIG DATA USING HADOOP PSEUDO-DISTRIBUTION MODE" ,Global Journal of Engineering Science and Research Management, 4(5), May, 2017.

[14] Mohd Rehan Ghazia, Durgaprasad Gangodkara," Hadoop, MapReduce and HDFS: A Developers Perspective", International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014), Interscience Institute of Management and Technology, Bhubaneswar, Odisha, India (Procedia Computer Science 48 (2015) 45 – 50).