IJCTA, 9(23), 2016, pp. 57-62

Towards Addressing the Challenges of Data Intensive Computing in Big Data Analytics

Niteesha Sharma* and B.Namratha**

Abstract : As the speed of data progress exceeds in this new era, excessive data is making huge troubles to human beings. However, there is much of the potentially used values hidden within the massive quantity of data. Tremendous data has drawn significant concentration from determination makers in potential sciences, coverage and selection makers in governments and firms. A tremendous quantity of fields and sectors, opening from economic and trade firms to public administration, from national security to scientific researches in tons of areas, involve with huge data problems. On one hand, tremendous data is extremely useful to give productivity in firms and evolutionary breakthroughs in scientific disciplines, which provide us quite a few possibilities to make fine progresses in tons of fields. Big Data has modified the way in which we adopt in doing corporations, managements and researches. Data-intensive science mainly in data-intensive computing is coming into the arena that aims to provide the tools that we have to manage with the massive knowledge problems. Information-intensive science is rising because the fourth scientific paradigm in phrases of the prior three, specifically empirical science, theoretical science and computational science. In this paper we aimed to illustrate a view about the tremendous information in relation with the context of Data Intensive Computing. Optimizing data entry is a method to make stronger the performance of data-intensive computing; these techniques comprise Data redundancy, migration and distribution of data, and access parallelism.

Keywords : Big Data, Data intensive computing, Analytics

1. INTRODUCTION

Big Data is a collection of massive data sets that cannot be processed using usual computing approaches. Big Data is not merely a data instead it has become an entire subject, which includes quite a lot of tools, strategies and frameworks.

Big Data is probably becoming the most emerging technology in real world now a day. The actual challenge the big organizations are facing is to get maximum out of the data already on hand and predict what type of information to accumulate in future. The key discussion is –"How you can take the prevailing data and make it meaningful that it provides us correct insight prior to the past data in many of the executive meetings in corporations". With the explosion of the information the challenge has long lasted to the next level and now Big Data is been the reality in many organizations.

The corporations have grown with the data related to them which has grown exponentially and today there are plenty of complexities to their data. Many of the tremendous businesses have data in multiple functions and in different formats. The data can also be spread out a lot that it is hard to categorize with a single algorithm or good judgment. Big companies are indeed facing challenges to hold the entire data on a platform which give them a single consistent view of their data. Big Data world is facing this particular challenge where in data coming from different sources and deriving the valuable expertise out of its revolution is making greater sense.

* Department of Information Technology Anurag Group of Institutions, Hyderabad, Telangana, India nitishasharma@cvsr.ac.in

** Department of Information Technology *Anurag Group of Institutions, Hyderabad, Telangana, India* namrathait@cvsr.ac.in

Big Data has got enormous attention from the understanding system self-discipline over the last few years with a few latest commentaries, editorials and certain issue introductions on the various forms of data as listed below:

- Structured data: Relational data.
- Semi Structured data: XML data.
- Unstructured data: PDF, Text, Word, Media Logs.(80% of Data is unstructured)

Defining Big Data : The 5V's of Big Data are **V**ariety, **V**elocity and **V**olume, Value and Veracity.

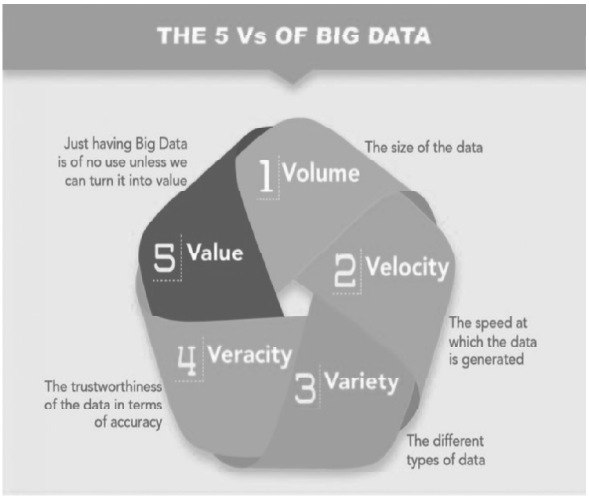

Fig. 1.

"Variety makes different type of data. Different Variety of data include the text, audio, video, log files, sensor data etc.

Volume represents the size of the data in terabytes and petabytes.

Velocity defines the speed at which data is streamed.

Value refers to change the data into value.

Veracity refers to trustworthiness of data in terms of accuracy". (August 2014 Nada Elgendly and Ahmed Elragal)

The kind of datasets considered in big data : Various examples involves social media community which analyzes their individuals' data to be taught more about the data and fasten them along with the advertising as well as content to their interests or search engines like Google which relates questions and answers to present better solutions to users queries.

There are two biggest sources of the data available in huge portions they are:

1. Transactional Data.
2. Sensor Data.

Transactional data includes from stock market prices to bank information to persons merchants purchase histories.

Sensor Knowledge is the statistical measure of data coming from the wired or the wireless sensors. This sensor data can be a measurement taken from the robots to the data on a mobile cellphone network, to instant electrical usage in houses and firms.

Big data analytics : Big Data analytics is where evolved analytic techniques operate on enormous data units. Therefore, big data analytics is really about two things-big data and analytics plus how both have teamed up to create one of the most profound tendencies in business intelligence (BI) today. To solve the big data analytics we need to distribute the data across multiple nodes stored physically in some location. This can be done using Distributed File System which is used to access the data of all the multiple nodes logically into a single file system. The most effective and established tool for big data analytics is Apache Hadoop.

Objective : Big data is becoming an increasingly important asset for decision makers. Large volumes of highly detailed data from various sources such as scanners, mobile phones, loyalty cards, the web, and social media platforms provide the opportunity to deliver significant benefits to organizations. This is possible only by optimizing the data with the help of Data Intensive Computing. And to do this the evaluation of, Data preparation *i.e.*, merging, replication, distribution, Data visualization etc are carried out.

2. RESEARCH METHODOLOGY

Data Intensive Computing deals with huge voluminous data and to analyze it various computational methods and their architectures are generated in many application domains. Since the data is becoming big and growing exponentially the factors being affected with it are scalability, reliability, high availability, elasticity and lower cost. Therefore there is a need of Data intensive computing which has become one of the important research fields in the information technology era.

Managing and processing exponentially increasing data volumes often are arriving in time-sensitive streams from arrays of sensors and instruments or as the outputs from the simulations; and significantly reducing data analysis cycles so that researchers can make timely decisions. Data intensive computing is mainly concerned with creating scalable solutions for capturing, analyzing, managing and understanding multi-terabyte and petabyte data volumes. Simply capturing the data at the rates it is emitted from a complex simulation, high resolution instrument or high-speed network is alone a challenging problem. To address this, the first stage of processing typically applies techniques to reduce the data in size, or process it so that it can be more efficiently manipulated by downstream analytics [1].

And to address this Data Intensive computing systems are built on the distributed file systems such as Hadoop Distributed File System (HDFS).

Apache Hadoop framework allows the distributed processing of huge data sets across the clusters of different computers making use of a simple programming model. The Digital data is used by Hadoop.

The Core Components of Hadoop are:

1. HDFS-Hadoop Distributed File System (Storage)
2. Map-Reduce

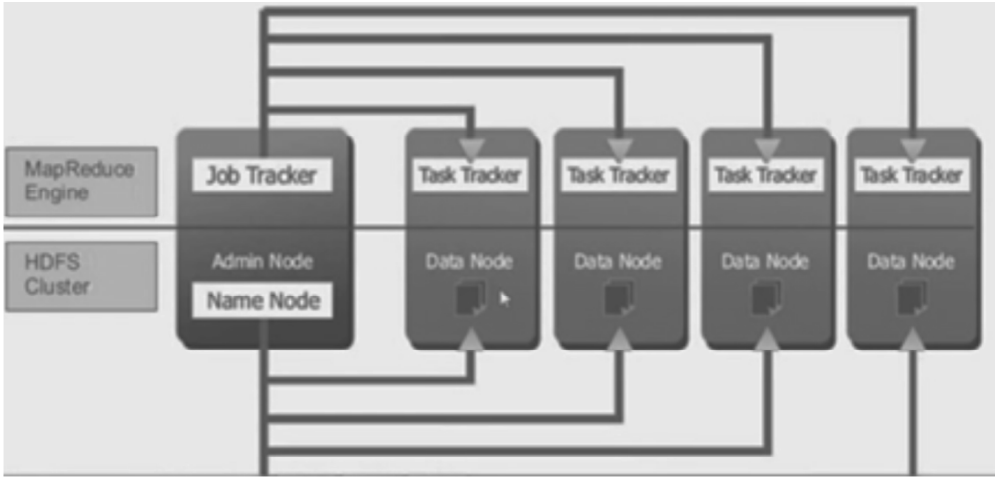

Fig. 2. Master-Slave Type of Configuration.

HDFS Cluster : It is the name given to the master-slave configuration shown in the figure where the large data sets are stored.

MapReduce : Map Reduce engine is the programming model which is used to retrieve and analyze the data.

Master Nodes : Name Nodes.

Slave Nodes : Data Nodes.

HDFS Architecture : Hadoop Distributed File System or HDFS can store massive amounts of data sets as well as informationthat can be scaled up incrementally to survive with the failure of significant parts of the storage infrastructure without losing the original data .The figure below demonstrates HDFS Architecture where hadoop creates a clusters of different machines and it helps to coordinate the work among them. If one of the systems fails to work or coordinate with different systems,hadoop continues to coordinate the clusters without losing the data or shifting the work or interrupting the work to other systems. HDFS manages to store the data on clusters by breaking the incoming data into pieces which are called as blocks and then storing each block of data repeatedly across the pool of servers. As a practice HDFS is able to store three complete sets of copies of each file by copying each and every piece to three individual different servers.

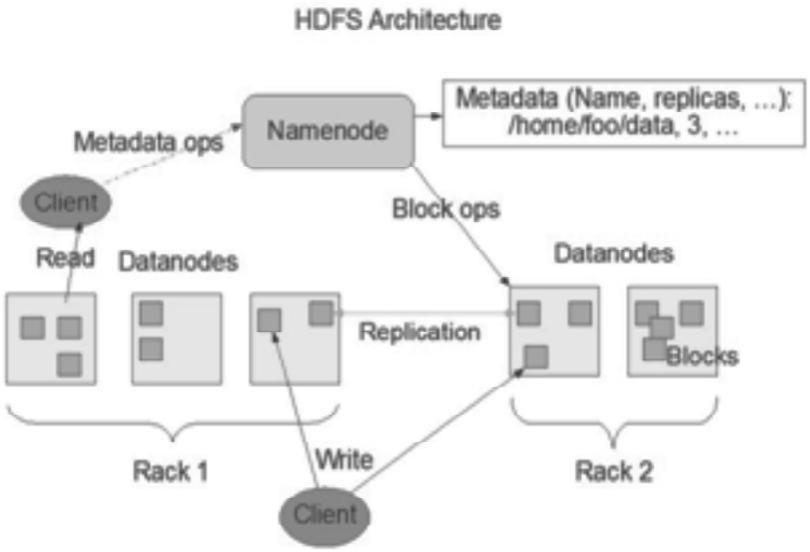

Fig. 3. HDFS Architecture.

Since HDFS is deployed on low cost commodity hardware server failures are common therefore the following are the features using which the file system is designed.

Features of HDFS :

- Highly Fault Tolerant.
- High Throughput.
- Suitable for Distributed Storage and Processing.
- Suitable for applications with large Datasets.
- Access to file system data.
- Provides file accessibility permissions and authentication.

Hadoop **Map Reduce** is a software architecture which process multi-terabyte data sets in parallel on huge clusters of the commodity hardware by considering the functionalities of reliability, Fault-tolerant and throughput.

Map : "Definition-the function takes key/value pairs as input and generates an intermediate set of key/value pairs".

Reduce : "Definition-the function which merges all the intermediate values associated with the same intermediate key".

The job of map reducer is to divide the input data sets into chunks that are independent of nature and are processed using the map tasks completely in a parallel manner. This framework usually sorts the map outputs that are then taken as input to reduce the tasks. All the input as well as the output data are stored in one particular file system using which the framework takes care of scheduling tasks, monitoring them and the failed tasks are re-executed. The architecture consists of one Job Tracker called as master and other Task Tracker called slave for each cluster-node. The master node is one who is responsible for rearranging the jobs' component tasks on the

slaves, and monitors them by re-executing the tasks that are failed. The slaves now executes the given tasks directed by the master Map Reduce framework and the Hadoop Distributed File System runs on the similar set of nodes (as shown in the figure above). Thus the above configuration of master-slave helps to effectively schedule the tasks on nodes which contain data, to result in and with high aggregate value, very high bandwidth value in and across the cluster.

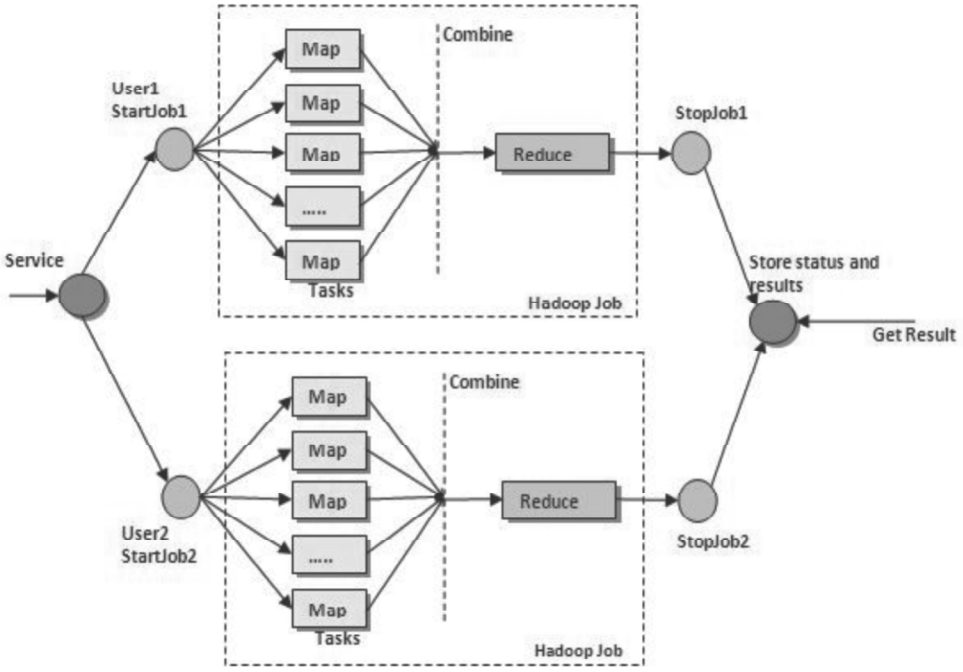

Fig. 4.

3. LITERATURE SURVEY

In the current trend of information society large amount of data is available easily to the decision makers. Big data is a set of datasets which are not only big in nature but also that are high in interactions with the various components of the Big Data like the volume, variety, velocity, value, veracity which makes the datasets difficult to handle using the traditional tools and found techniques. Due the above mentioned issues solutions are needed as such we can extract as well as study the value, need and knowledge of different datasets. Also the decision makers should be able to grasp the recent advancements of rapidly changing data from daily day to day interactions for the entire customer as well as the social media networking interactions. To solve such issues Big Data analytics- Study of advanced analytics techniques, Data Intensive methodologies have to be used.

This section dealt with the description of the contents of Big Data like scope, methodology, examples, advancements, challenges, which are very essential in forming the different datasets for decision makers. The most critical issues of Big Data are privacy, safety and security. From this paper we can make a conclusion that any organization or industry with big data can be benefited by carefully analyzing the problem using Big Data Analytics and solving the purpose. The major challenge is to extract useful information from the collected data [1].

"Big Data Analytics is an analysis of huge amount of data to get the required useful data, extracting from the patterns which are hidden. Big data analytics refers to the Mapreduce Framework which is developed with the aid of the Google. The open source tools used were Apache Hadoop for implementing Map reduce model [2].

"According to 2013, Facebook has 1.11 billion humans active money owed from which 751 million use facebook from a cell. Yet another instance is flicker having function of unlimited snapshot uploads (50MB per picture), limitless video uploads (90 seconds max, 500MB per video), the capacity to exhibit HD Video, unlimited storage, limitless bandwidth. Flickr had a total of 87 million registered participants and more than 3.5 million new snap shots uploaded daily [3].

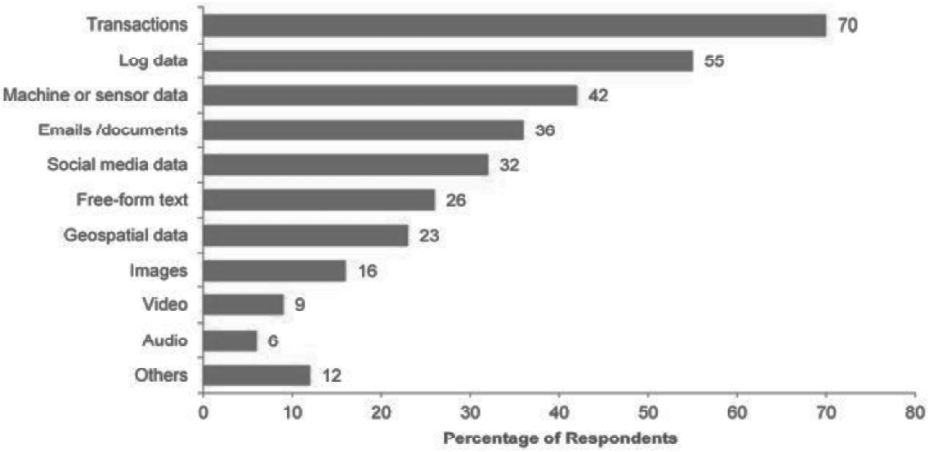

Fig. 5. Big Data Sources.

Reports the practical work on the issues of Big Data. It describe the premiere solutions making use of Hadoop cluster, Hadoop distributed File System (HDFS) for storage and Map scale back programming framework for parallel processing to process tremendous information set of data [4].

Previously the information was once much less and readily treated by means of RDBMS but not too long ago it's tricky to manage gigantic data by way of RDBMS tools, which is preferred as large information. On this they instructed that bigdata differs from other information in 5 dimensions equivalent to volume, speed, type, value and complexity. They illustrated the hadoop structure including identifying various nodes. Hadoop architecture manage enormous knowledge units, scalable algorithm does log management utility of enormous information will also be discovered in monetary, retail industry, health-care, mobility, coverage. The authors also interested by the challenges that have to be faced with the aid of businesses when dealing with giant information: - data privacy, search analysis, and so on [5].

Data Intensive computing is managing, analyzing and understanding data at volumes and rates. In recent twenty years there is an explosive growth of the data in all over the world. The various Data intensive applications which include Internet text data processing, inverse data processing, scientific research data processing and large scale graph computing as well as the different architecture design issues are discussed [6].

4. CONCLUSION

In this research paper we have examined the innovative field of current information trend Big Data where voluminous amount of data are being produced daily and within them there are intrinsic number of hidden patterns which should be extracted and utilized. For this the Big Data Analytics is applied with parallel processing of data called the Data Intensive Computing which manages, analyses and understands the data by optimizing the data access and hence it improves the performance of Data Intensive Computing.

5. REFERENCES

1. N.Suresh Goud, "Data Intensive computing in Clouds".
2. Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review".
3. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012)"Shared disk big data analytics with Apache Hadoop
4. Real Time Literature Review about the Big data.
5. Aditya B. Patel, Manashvi Birla, Ushma Nair (*6-8 Dec. 2012)* "Addressing Big Data Problem Using Hadoop and Map Reduce.
6. S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/ IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).
7. Yanhui Wu, Guoqing Li1, Lizhe Wang, Yan Ma1, Joanna Ko³odziej3 and Samee U. Khan "A Review on Data Inteensive Computing".