Review Paper on BIG DATA and HADOOP

Anjali Goel[1], Vishwachi[2]
[1,2]Scholar, Department of Information Technology, ABES Institute of Technology, Uttar Pradesh, India
[1]anjaligoel4682@yahoo.in, [2]vishwachi.choudhary@abesit.in

Abstract:
BIG DATA refers to the huge volume of data which is being processed on a daily basis and needs to be stored, distributed and managed. It can be structured, semi-structured or unstructured. Parallelism is used to process this data in an efficient manner. BIG DATA demands a platform and a technique which can analyze this data and extract hidden knowledge from it. HADOOP is an open source software, a core platform to structure BIG DATA. It offers a distributed file system known as HDFS (HADOOP Distributed File System) and Map Reduce structure to deal with large data and provide a high degree of fault tolerance.

1. Introduction

A. Definition:

Huge Data alludes to extensive informational indexes that can be broke down computationally to consider examples and patterns particularly identifying with human conduct and cooperation's. It depicts any voluminous measure of organized, semi organized, and unstructured information created once a day. Informational collections that are so voluminous and complex that conventional information preparing strategies, for example, social databases, are inadequate to manage them. There are five measurements (5 V's) to enormous information known as Volume, Variety, Velocity, Veracity and Value which depict it'sproperties. Figure 1 gives Layered Architecture of BIG DATA System. It can be isolated in three layers, Infrastructure Layer, Computing Layer, and Application Layer start to finish.

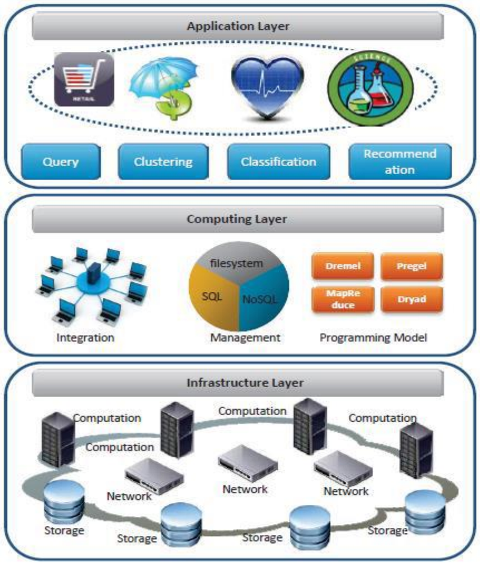


Fig. 1. Application, Computing and Infrastructure Layers

B. 5 V's of BIG DATA

Volume of data: Volume refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, photographs, video, etc. The increasing amount of data has made it difficult to store data in traditional format. Now the technology used, stores data in different locations and combined back by software.

Variety of data: Variety is defined as the different types of data available. Today's data cannot be put into a table, as it's highly unstructured. To be more precise, 80% of the world's data comes in this category, including video sequences, social media updates, photos, etc.

Velocity of data: Velocity refers to the speed at which data is being generated, collected and analyzed. The number of emails, social media messages, photos, etc. are increasingly rapidly on daily basis. Not just analyze, but the speed of transmission must also remain constant to allow for faster access.

Veracity of data: Veracity is a term used to define the quality of data. Storing data in large amounts has no use if the data is not accurate.

Value of data: Value is the worth of the data that is being extracted. Storing huge amounts of data is pointless, unless it can be turned into value. The important part of BIG DATA study is to find out whether it is feasible and economic or not.

C. Problem with BIG DATA

a. Heterogeneity and Incompleteness:

Systems cannot work with data that is complex or unstructured. So, another challenge is to make it aligned and Compatible.

b. Scale:

Managing large and rapidly increasing volumes of data has been challenge since long. In thepast, it was mitigated by processors getting faster, following Moore's law, to provide us with the resources required to work with increasing amount of data. But, there's a change now: data is being accumulated at a faster speed, than what the systems can process.

c. Timeliness:

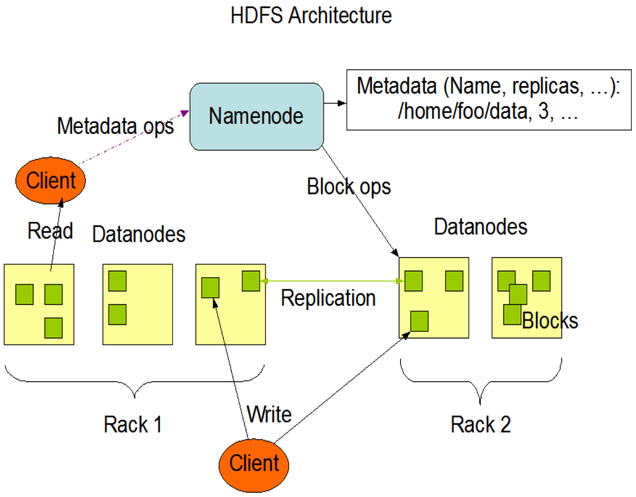
The other thing with size is speed. The bigger amount of dat, the more amount it will take to process. However, it is not just in terms of speed, but also about the rate at which processing of the data takes place in the system.

d. Privacy:
This is another major factor to be concerned about. When it comes to BIG DATA, privacy concerns increase. However, there is a fear in people regarding the inappropriate use of personal data. Managing privacy is a technical and a sociological problem.

e. Human Collaboration:
There are situations where humans are capable of solving an algorithm much faster than system algorithms. Hence, it is important that apart from technical support, humans should also deal with the real time problems to reach an efficient solution which is best for the situation.

2. HADOOP Solution for BIG DATA:

HADOOP is a Programming structure used to help the handling of vast informational collections in a dispersed registering condition. HADOOP was created by Google's MapReduce that is a product structure where an application separates into different parts. The Current Apache HADOOP biological community comprises of the HADOOP Kernel, MapReduce, HDFS and quantities of different segments like Apache, hive and Base of HDFS, HDFS is able to store huge amount ofdata.



Fig, 2. HDFS Architecture

A. HDFS Architecture:
HADOOP includes a fault-tolerant storage called the HADOOP Distributed File System or HDFS which reduces the chances of data failure. HDFS is able to store huge data in a systematic manner so files can be
easily retrieved.

B. MAPREDUCE ARCHITECTURE
The pillar in the HADOOP system is the MapReduce framework. This framework allows the operation to be applied to a huge data set, divide the data, and run it in parallel. From another point of view, this can occur on multiple possibilities. For example, a large dataset can be reduced in a smaller set where processes can be applied. In HADOOP, operations are written as MapReduce work in Java. There are higher level languages like Hive and Pig that make writing these programs easier.

There are two functions in MapReduce as follows:

Map: the function takes key/value pairs as input and generates an intermediate set of key/value pairs.
Reduce: the function which merges all the intermediate values associated with the same intermediate key.

3. Literature Review

S. Vikram Phaneendra & E. Madhusudhan Reddy et.al. [6] stated that in olden days the data was less and easily handled by RDBMS but now it is difficult to handle huge data, which is preferred as “BIG DATA”. They told that BIG DATA differs from other data in in terms of volume, velocity, variety, value and complexity. They illustrated the HADOOP architecture consisting of name node, data node, edge node, HDFS to handle BIG DATA systems. They also focused on the challenges that need to be faced by enterprises when handling BIG DATA: - data privacy, search analysis, etc.

Albert Bifet et.al. [3], [5] Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created every day is “BIG DATA”. The mining tools used for BIG DATA are apache HADOOP, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc.

Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N. et. al [1], [7] analyzed big and said sources like business processes, web servers, transactions, social networking sites, etc. in structured and unstructured form analyzing the loads of data or extracting information is a challenging task. The term “BIG DATA” is used for large data sets. BIG DATA sizes are a constantly moving target from a few terabytes to many peta bytes of data in a single dataset. Difficulties include capture, storage, search, sharing, analytics Typical examples of BIG DATA include web logs, RFID generated data, satellite and geo-spatial data, sensor networks, social data from social networks, Internet text, Internet search indexing, call detail records, atmospheric science, genomics, biogeochemical, biological, and other complex and interdisciplinary scientific project, military.

Kiran kumara Reddi & Dnvsl Indira et. al. Stated that BIG DATA is combination of structured, semi-structured, unstructured homogenous and heterogeneous data. They suggested to use model to handle transfer voluminous data over the network. Under this model, these transfers are relegated to low demand periods where there is ample, idle bandwidth available. The Nice model uses a store –and-forward approach by utilizing staging servers. The model accommodates differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer BIG DATA and to solve issues like security, compression, routing algorithms.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk BIG DATA analytics with Apache HADOOP" BIG DATA analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. BIG DATA analytics refers to the Mapreduce Framework which is developed by the Google. Apache HADOOP is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces. SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

4. Conclusion

We have entered a time of BIG DATA. This paper talks about the concept of BIG DATA with 5 Vs, Volume, Velocity, Variety, Veracity and Volume of BIG DATA. The paper also deals with BIG DATA processing problems. These challenges need to be solved for efficient and fast processing of BIG DATA. The challenges include scale, heterogeneity, lack of error-

Handling, timeliness, privacy, provenance, structure, and visualization. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes HADOOP which is an open source software used for processing of BIG DATA, along with its features HDFS and MAPREDUCE.

REFERENCES

[1] 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).

[2] Umasri M. L, Shyamalagowri. D, Suresh Kumar. S "Mining BIG DATA: Current status and forecast to the future"

[3] Kenn Slagter, Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013

[4] Balaji Palanisamy, Member, IEEE, Aameek Singh, Member, IEEE Ling Liu, Senior Member, IEEE" Cost-effective Resource Provisioning for MapReduce in a Cloud" gartner report 2010, 25

[5] Sameer Agarwal, Barzan MozafariX, Aurojit Panda, Henry Milner, Samuel Madden X, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data"

[6] Albert Bifet "Mining BIG DATA in Real Time" Informatica 37 (2013) 15–20 DEC 2012

[7] Bernice Purcell "The emergence of "BIG DATA" technology and analytics" Journal of Technology Research 2013.