International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Big Data using Hadoop

Dinesh D. Jagtap[1]
CSE Department,
Everest Educational Society's Group of Institutions
Aurangabad,Maharashtra, India
dinesh.jagtapd@gmail.com

Prof.B.K.Patil[2]
CSE Department,
Everest Educational Society's Group of Institutions,
Aurangabad, Maharashtra,India.
cseroyal7@gmail.com

Abstract—*:* "Big Data" is data that becomes large enough that it cannot be processed using conventional methods.
The term Big Data concerns with the huge volume, complex and rapidly growing data sets with multiple, independent sources .Due to fast development of networking ,data storage and data collection capacity the concept of big data is now rapidly expanding in all science and engineering domains including biological, physical and biomedical sciences. Social networking sites, mobile phones ,banking and stock exchange sectors, sensors and science contribute to production of peta bytes of data daily. That's why Big Data analysis now drives almost every aspect like mobile services, retail, financial services, manufacturing and life sciences. We all have heard a lot about "big data," but "big" is actually a red herring. Telecommunications companies, Oil companies, and other data-relevant industries have had vast datasets for a long time. And as storage capacity continues to enlarge, today's "big" is certainly tomorrow's "medium" and "small." In next week. The best meaningful definition of "big data" is when the size of the data itself becomes part of the problem.

Keywords— Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations

I. INTRODUCTION

In the last few years, there has been tremendous increased in the amount of data that's available. Whether we're talking about tweet streams, web server logs, records of online transactions, government data, or some other source data. The problem is not only finding data, it's figuring out what to do with the available data. And it's not just companies using their own data, or the data contributed by users of that company. Data mining allows users to examine the data from many dissimilar magnitudes or angles, sort it, and summarize the associations identified. Strictly, data mining is the process of finding correlations or patterns among dozens of fields in big relational databases. Another fundamental characteristics of the Big Data is large volume of data is represented by heterogeneous and diverse dimensionalities. This is because of different information collector prefers their own schemata for recording the data and also the nature of application also results in diverse representation of data. When the size of data increases obviously the complexity and relationships underneath the data.Hadoop is an open source software project that enables processing of large data sets distributed across the clusters of product servers. We're discussing data problems that are ranging from gigabytes to petabytes size of the data. At particular point, conventional techniques for working with data run out of the stream.

Information platforms are somewhat like as traditional data warehouses, but different. They describe rich APIs, and are designed for exploring and considering the data rather than for traditional analysis and reporting. They allow all data formats, with the most messy, and their schemas grow.

Figure. 1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

II.REQUIREMENT

Most of the organizations that have built data platforms have establish it necessary to go further than the relational database model. Conventional relational database systems stop being valuable at this balance. Managing shading and replication across a mass of database servers is difficult and slow. The need to define a schema in advance conflict with reality of numerous, formless data sources, in which you may not know what's important until after you've analyzed the data. Relational databases are premeditated for uniformity, to support complex transactions that can easily be rolled back if any one of a composite set of operations fails.

To store vast datasets efficiently, we've seen a new type of databases appear. These are normally called NoSQL databases or Non-Relational databases, while neither term is very practical. Many of these databases are the logical offspring of Google's Big Table and Amazon's Dynamo, and are intended to be distributed across many nodes, to provide "ultimate uniformity" but not absolute consistency, and to have very flexible schema. Whereas there are two dozen or so products available (about all of them open source), a few leaders have recognized themselves:

III.OBJECTIVES

Data is only useful if you can do something with it, and massive datasets introduces computational issues. Google popularized the MapReduce approach, which is mainly a divide-and-conquer policy for distributing an tremendously large problem across an very large computing cluster. In the "map" stage, a programming task is divided into a number of equal subtasks, which are after that distributed across many processors; the halfway results are then combined by a single shrink task. In perception, MapReduce seems like an clear solution to Google's major trouble, creating large searches. It's so easy to allocate a search among number of processors, and after that merge the results into a single set of answers. What's less understandable is that MapReduce has proven that to be broadly valid to many large data troubles, ranging from searching to machine learning. Architecturally, the cause you're able to deal with lots of data is because Hadoop spreads it out. And the reason you're able to ask complicated computational question is only because you've got all of these processors, working in parallel, harness mutually.

IV.THEME

Using data effectively requires something different from traditional statistics, where actuaries in business suits perform arcane but fairly well-defined kinds of analysis. What differentiates data science from statistics is that data science is a holistic approach. We're increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form,making it tell its story, and presenting that story to others.To meet the challenge of processing such large data sets, Google created Map-Reduce. Google's work and Yahoo's creation of the Hadoop MapReduce implementation has spawned an ecosystem of big data processing tools.

A. Literature Survey

Data is everywhere: your administration, your web server, your business partners and even your body, we are finding that almost everything can be instrumented. At O'Reilly, we normally merge publishing industry data from Nielsen BookScan with our own sales

data, openly available Amazon data, and even job data to see what's happening in the publishing industry. Sites like Infochimps and Factual gives access to numerous large datasets, including weather data, MySpace activity.

Storage Map Reduce Big data" is data that becomes huge enough that it cannot be processed using straight methods. Social networks, mobile phones, Banking sector and government agencies contribute to peta bytes of data created daily.

[25] To face the number of challenge of processing such kind of huge data sets, Google invented Map Reduce. Google's work and Yahoo's creation of the Hadoop MapReduce implementation has spawned an environment of big data processing tools.

[26] As MapReduce has grown-up in reputation, a stack for big data systems has invented, comprising layers of Storage, MapReduce and Query (SMAQ).

[27] SMAQ systems are normally open source, distributed, and run on commodity hardware.

Figure.2. SMAQ systems

[28] Created at Google in response to the difficulty of creating web search indexes, the MapReduce framework is the thrust behind most of today's big data processing.

[29] The key improvement of MapReduce is the capability to take a query over a data set, divide it, and run it in parallel over many nodes.

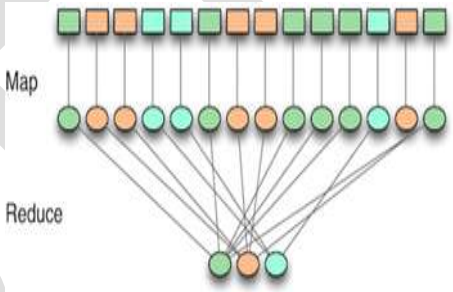

Figure. 3. Map Reduce Technique

[30] Loading the data—This operation is more properly called Extract, Transform, Load (ETL) in data warehousing language. Data should be extracted from its source, prepared to make it ready for further processing.

[31] MapReduce—This segment will retrieve data from storage, process it, and transfer its results to the storage.

[32] **Extracting the result**—Once processing is completed, for the result to be useful to humans, it must be retrieved from the storage and presented.

[33] Many SMAQ systems have characteristics designed to solve the operation of each of these stages.

[34] Storage-MapReduce requires storage from which to retrive data and in which to store the obtained results of the computation. The data predicted by MapReduce is not the relational data as generally used by conventional database system. Instead, data is consumed in chunks, which are then divided among nodes and fed to the map phase as key value pairs. This data does not need a schema, and may be formless.

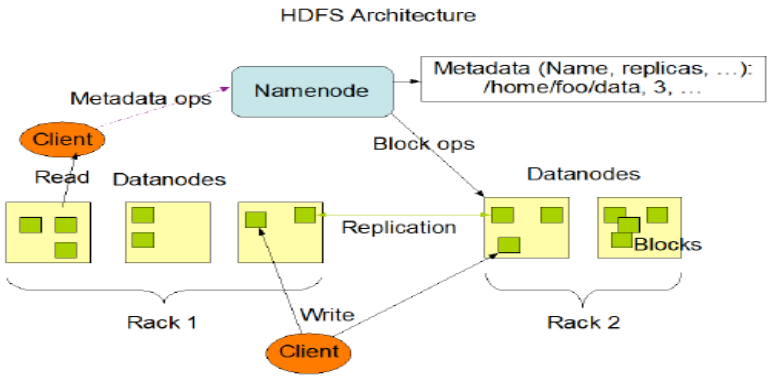

Figure.4 HDFS Architecture

[35] **Hadoop** is leading open source map reduce implementation created by yahoo emerged in 2006 creator is Doug cutting. To communicate between node in 2nd generation uses replication factor Hadoop and HDFS utilize a master- slave architecture. HDFS is written in Java, with an HDFS cluster consisting of a primary name node a master server that manages the file system namespace and also regulates right of entry to data by clients. An elective secondary Name Node for fail over purposes also may be configured. Consecutively. HDFS has many goals. Here are some of the most prominent:

- Fault tolerance is easy to find by detecting faults and applying rapid and automatic recovery.
- Data access via MapReduce streaming.
- Processing logic is close to the data instead of the data close to the processing Logic.

V.BIG DATA CHARACTERISTICS: HACE THEOREM

HACE Theorem. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.These above mentioned characteristics make it an extreme challenge for taking out meaningful facts from the Big Data. In a naı¨ve sense, we can imagine that there are number of blind men are trying to size up a giant elephant (see Fig. 1), which will be the Big Data in this context. The main aim of each blind man is to construct a picture of the elephant on the basis of the part of information collects during the process. Because each person's view is restricted to his local region, so the blind men will each conclude alone that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make this scenario more complex let us imagine that the elephant is growing rapidly and its pose changes constantly, and) each blind man may have his own information sources that tell him about biased knowledge about the elephant (e.g., so one blind man can share his feeling about the present pose of elephant with another blind man, where the knowledge which is shared is inherently biased. Describing the Big Data in this scenario is corresponding to aggregating heterogeneous information from number of sources (blind men) to help draw a best possible picture of the elephant in a real-time position.[3]

5.1 Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the enormous volume of data is represented by heterogeneous and varied dimensionalities. Because different information collectors prefer their own schemata or protocols for recording of the data, different applications and their nature also results in miscellaneous data representations. The simple example is, each human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on. For X-ray and CT scan examination of each patient, images or videos are provides visual information used for doctors to carry detailed examination that's why images and videos are useful entity. Under such situation, the heterogeneous features refer to the representations for the same individuals in different types, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations or health practitioners can have their own kind of schemata to represent each individual, if we want to enable aggregation of data by combining data from all sources then the data heterogeneity and diverse dimensionality are become major challenges.[3]

5.2 Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentra-lized controls aare a main quality of Big Data applications. Being autonomous there is no any centralized control on it so, each data source can produce and gather information without involving and relying on any centralized control. This is same as World Wide Web (WWW) setting where each web server is independent and provides a certain amount of information without necessarily depanding on other servers.The enormous volumes of the data can also

make an application more vulnerable to malfunctions, if the whole system has to be depended only on single centralized control unit.Todays well known social sites such as Google, Facebook, and Walmart, has set of large number of server farms which are situated all over the world to ensure nonstop services and quick responses for local markets. More particularly, the local government regulations also impact on the wholesale management process and it result in reorganized data representations and data warehouses for local markets.[3]

5.3 Complex and Evolving Relationships

While the amount of the Big Data increases, so the complexity and the relationships under the data. In the early stage of data centralized information systems, the main aim to finding best feature values to represent each observation. This is same as using a number of data fields, such as gender,age, income, education background, to describe each individual. This type of representation of sample-feature inherently treats each individual as an independent entity without considering their societal relations, which is one of the most important factors of the human society. In real world our friend circles may be formed based on the frequent hobbies or people are connected by biological dealings. Such social connections also are very popular in cyberworlds. For example, major social networking applications, such as Facebook or Twitter, are mostly characterized by social functions such as friend-connections and followers (in Twitter). In the sample-feature representation, individuals persons are regarded alike if they are sharing similar feature values, whereas in the sample-feature-relationship representation, two persons can be linked together even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to sequential, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the complex data relationships along with evolving changes into consideration, to find out valuable patterns from Big Data collections. [3]

VI.CONCLUSION

Real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task.Term Big Data is accurately concerns about volume of dat, our HACE theorem suggests the key characteristics of the Big Data that are. First is, Large with heterogeneous and diverse data sources,Second is autonomous data sources with distributed and decentralized control, and third one is complex and evolving in data and knowledge associations. Such kind of mutual characteristics suggest that Big Data need a “big mind” to combine data for maximum values. To describe the concept of Big Data, we have review several challenges at the data, model, and system levels. To support Big Data mining, high-performance powerful computing platforms are required, which impose regular designs to unleash the full power of the Big Data. At the data level the autonomous information sources and the range of the data collection environments, often gives result in data with complex situation, such as missing values. In certain situations, solitude concerns, noise, and errors can be introduced into the data, to create tainted data copies. Developing of a secure and sound information sharing protocol is a major challenge. At the model level, the key challenge is to create universal models by combining nearby discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between scattered sites, and combine decisions from several sources to gain a best model out of the Big Data. At the system level, the important challenge is that a Big Data mining framework needs to consider complex relations between models, samples and data sources, along with their evolving changes with time and other likely factors. A system needs to be watchfully designed so that formless data can be linked through their compound associations to form helpful patterns, and the growth of data volumes and item relationships should help form rightful patterns to estimate the tendency and view. We stare Big Data as an capable style and the necessity for Big Data mining is arising in all science and engineering fields. With the use of Big Data technologies we will with any luck be able to give the best part of applicable and most precise social sensing feedback to improved realize our society at realtime. We can additionally stimulate the association of the public audiences in the data building loop for community and economical events. The period of Big Data has arrived.

REFERENCES:

[1] Apache HBase http://hbase.apache.org
[2] Apache Accumulo http://accumulo.apache.org
[3] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, IEEE“Data mining with Big Data” Transaction on knowledge and data Engineering vol. 26 No.1 January 2014.[4] J. Kepner and S. Ahalt, “MatlabMPI,” Journal of Parallel and Distributed Computing, vol. 64, issue 8, August, 2004.
[5]B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A.D. Joseph, R.Katz, S. Shenker and I. Stoica, "Mesos: A Platform for Fine-Grained
[6] N. Bliss, R. Bond, H. Kim, A. Reuther, and J. Kepner, “Interactive grid computing at Lincoln Laboratory,” Lincoln Laboratory Journal, vol. 16,no. 1, 2006.

[7] J. Kepner et al., "Dynamic distributed dimensional data model (D4M) database and computation system," 37th IEEE International1989.
[8] A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44, 2009.
[9] A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44,2009.
[10] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[11] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[12] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012