BIG DATA ANALYTICS

BIG DATA MODELS AND ALGORITHMS

Tripti Tripathi
Department of Computer Science
SRMSWCET
BAREILLY, INDIA
triptitripathi79@gmail.com

Mr. Jitendra Singh
Faculty in Department of Computer Science
SRMSWCET
BAREILLY, INDIA
jitendra.singh@srmswcet.ac.in

***Abstract* —** **In this era of information, huge amounts of data have become available before decision makers. Big Data cannot be defined in TB, PB, and EB. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Additionally, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the relationship between big data and internet of things and big data analytics and metrics.**

Keywords- big data, alorithms, relationship between big data and internet of things, big data metrics.

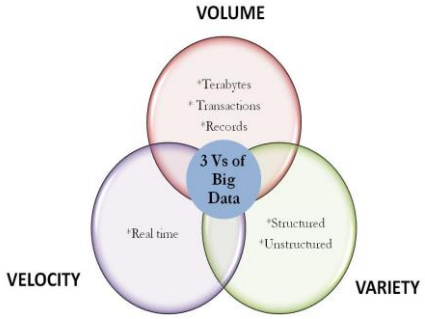
1. Introduction-:

A. Big Data: Definition

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10^12 or 1000 gigabytes per terabyte) to multiple petabytes (10^15 or 1000 terabytes per petabyte) as big data. Figure No. 1.1 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom.

B. 3 Vs of Big Data

Generally, big data is defined as 3Vs, i.e., big data is associated to the volume, variety, and velocity at which the large amounts of data are generated, stored, processed, and analyzed, as shown in the Figure . Dealing with data at this scale is a new frontier .



Volume of data: Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.
Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.
Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Figure 1: Layered Architecture of Big Data System

There are four steps to big data processing: (1) acquisition, which encompasses data captured and acquired from many different sources; (2) access, which includes data indexing, storage, sharing, and archiving, usually based on specific

software framework for integration and organization; (3) analytics, which is related to data analysis and manipulation; and (4) application, which means make decisions and take action[3].

C. Types of Big Data

There are two types of big data: structured and unstructured.

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically.

"Unstructured big data is the things that humans are saying," says big data consulting firm vice president Tony Jewitt of Plano, Texas. "It uses natural language."Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

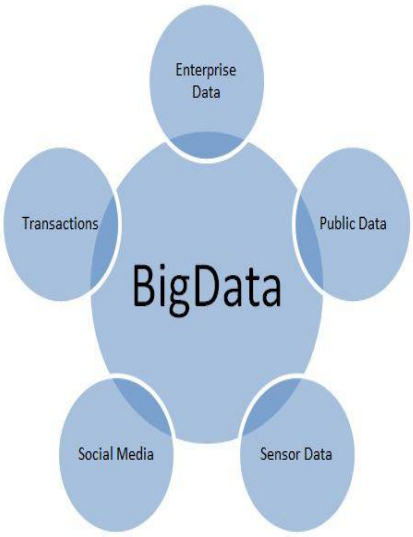


Figure 2: Sources of big data

E. Problem with Big Data Processing

i. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

ii. Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now; data volume is scaling faster than compute resources, and CPU speeds are static.

iii. Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge.

iv. Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

v. Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration[5].

2. Algorithm

Many algorithms were defined earlier in the analysis of large data set. With the growing knowledge in the field of big data, the various techniques for data analysis- structural coding, frequencies, co-occurrence and graph theory, data reduction techniques, hierarchal clustering techniques, multidimensional scaling were defined in Data Reduction Techniques for Large Qualitative Data Sets. It described that the need for the particular approach arise with the type of dataset and the way the pattern are to be analyzed. Nowadays the two main techniques that are in use to analyze big data are:

A. Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points.

i. HDFS Architecture

Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates *clusters* of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster[2].

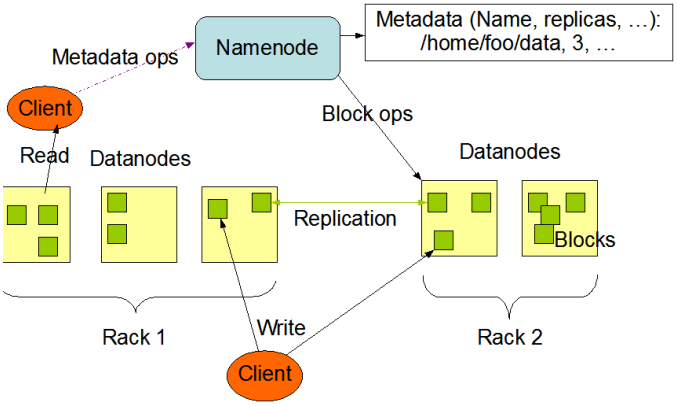


Fig.2 HDFS Architecture

ii. MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. 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.

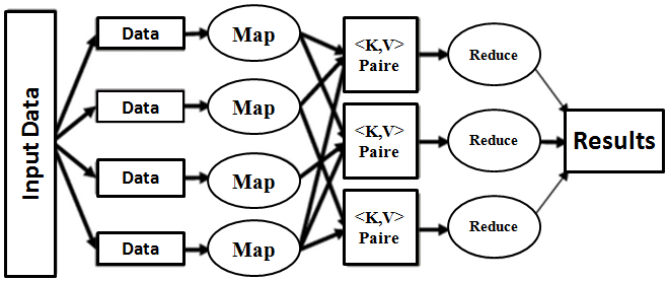


Fig.3 MapReduce Architecture

B. Data Mining

Data Mining is the technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data.

Data Mining has also been termed as data dredging, data archaeology, information discovery or information harvesting depending upon the area where it is being used. The data Mining environment produces a large volume of the data. The information retrieved in the data Mining step is transformed into the structure that is easily understood by its user. Data Mining involves various methods such as genetic algorithm, support vector machines, decision tree, neural network and cluster analysis, to disclose the hidden patterns inside the large data set.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both[1]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Data mining as a term used for the specific classes of six activities or tasks as follows:

i) Classification

ii) Estimation

iii) Prediction

iv)Association rules

v) Clustering

vi)Description

i) Classification -Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbour classifier, Naive Byes, Apriori and

AdaBoost. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples.[4].

ii) Estimation- Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance.

iii) Prediction- It's a statement about the way things will happen in the future, often but not always based on experience or knowledge. P*rediction* may be a statement in which some outcome is expected.

iv) Association- *Rules* An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database.

v) Clustering- Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

Big data	**Data mining**
Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information
Big data is the asset	Data mining is the handler which provides beneficial result**.**
Big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.	Data mining refers to the operation that involve relatively sophisticated search operation

TABLE 1

Difference between Big data and Data mining

3. Internet Of Things

There are volumes of data that will not fit into a standard database and are almost impossible to deal with using current tools. The kind of smart products associated with the internet of things call for complicated logic in relation to security and data privacy as well as interaction and flexibility. This article describes the nature of these new challenges.

The Internet of Things (IoT) is on its way to becoming the next technological revolution. According to Gartner, revenue generated from IoT products and services will exceed $300 billion in 2020, and that probably is just the tip of the iceberg. Given the massive amount of revenue and data that the IoT will generate, its impact will be felt across the entire big data universe, forcing companies to upgrade current tools and processes, and technology to evolve to accommodate this additional data volume and take advantage of the insights all this new data undoubtedly will deliver.

IoT and big data basically are two sides of the same coin. Managing and extracting value from IoT data is the biggest challenge that companies face. Organizations should set up a proper analytics platform/infrastructure to analyze the IoT data. And they should remember that not all IoT data is important.

An IoT device generates continuous streams of data in a scalable way, and companies must handle the high volume of stream data and perform actions on that data. The actions can be event correlation, metric calculation, statistics preparation, and analytics. In a normal big data scenario, the data is not always stream data, and the actions are different. Building an analytics solution to manage the scale of IoT data should be done with these differences in mind.

The Internet of Things consists of three main components:

1. The things (or assets) themselves.

2. The communication networks connecting them.

3. The computing systems that make use of the data flowing to and from our things.

4. Big Data Metrics

Big data is growing rapidly, and it cannot be dismissed as hype; data and analytics are nowadays one of the main topics discussed in organizations. Dealing with data at this scale is a new frontier. The major challenges of big data are to process, aggregate, filter, and organize large amounts of data to transform them into useful information for the companies to get value from data. Value creation can allow companies to gain competitive advantage in a highly competitive world.

Big data potential to create competitive advantage has been influencing how the companies manage their business. In the United States, some authors foresee that big data will rapidly become a key determinant of competition in several economic sectors. Such increasing interest in big data implies that 1.5 million more data-savvy managers will be necessary to analyze big data and make decisions.

Big data can generate significant financial value across sectors. In the United States, for instance, in the health care

sector, it will be generated $300 billion in value per year. In the retail industry, big data has resulted in 60% increase in net margin possible, and in the manufacturing sector, big data has led to up to 50% decrease in product development and assembly costs. Therefore, big data analytics have become more and more important in both the academic and companies of different economic sectors over the past two decades.

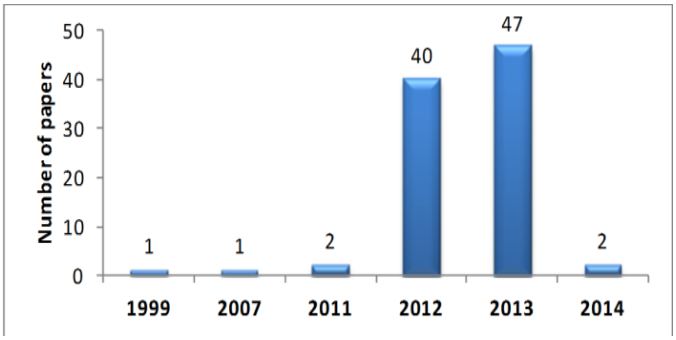


Figure 4: Total papers published per year

A. Main Findings

The 15 articles analyzed were published in 11 different journals. Three articles were published in Harvard Business Review (HBR) magazine. Thomas H. Davenport and James M. Tien were authors of two of three articles published in HBR. The other authors have published only one article on big data.

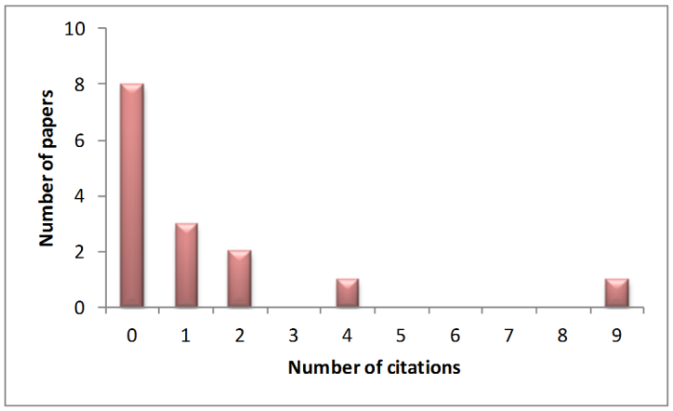


Figure 5: Number of citations of selected articles

Big data is very wide in scope, the magnitude of its potential is huge, and the value it can create varies across sectors. The usage of big data can support the decision-making in a broad range of areas including business, science, engineering, defence, education, healthcare, society, e-commerce, market intelligence, e-government, security, supply chain management, and innovation for long-term sustainability. *In the U. S., major research programs are being funded to* sponsor *deal with big data in all five sectors of the economy: services, manufacturing, construction, agriculture and mining*. Thus, many applications cited in the selected articles pointed out to the influence of big data on PMSs because the two issues support the decision-making.

The main barriers to implement big data are: the appropriate means to collect and store data correctly; to apply the right analytic tools and methods because the data is too voluminous or too unstructured to be managed and analyzed through traditional means; and to make the decision in benefit of organization from the analysis because sometimes managers do not know how the information can be used for making key decisions.

5. Conclusion

We have entered an era of Big Data. The paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. 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 different methodologies associated with different algorithms used to handle such large data sets. And it gives an overview of architecture and algorithms used in large data sets. It also describes about the various security issues, application and trends followed by a large data set. The paper describes Hadoop which is an open source software used for processing of Big Data. This also briefs about the relationship between internet of things and big data.

References

[1] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "On big data and hadoop,"International Journal of Scientific and Research Publications,Volume 4, Issue 10, October 2014 1 ISSN 2250-3153.

[2] Chanchal Yadav, Shuliang Wang, Manoj Kumar,"approaches to handle large data" IJCSN International Journal of Computer Science and Network, Volume 2, Issue 3, 2013 ISSN (Online): 2277-5420.

[3] Raquel Mello, Luciana Rosa Leite Roberto Antonio Martin,"Big data and its performance measurement". Proceedings of the 2014 Industrial and Systems Engineering Research Conference Y. Guan and H. Liao.

[4] Btissam Zerhari, Ayoub Ait Lahcen, Salma, "Algorithms and Challenges".

[5] Bharti Thakur, Manish Mann."Data Mining For Big Data". International Journal of Advanced Research in Computer Science and Software Engineering. Volume 11, Issue 6, September 2014 .