 ISBN:978-93-5254-640-4

Big Data and its Technologies: Overview

Mrs. Dipali Latesh Mahajan[#] Mrs. Suvarna Amit Gogate[*]
Department of Computer Scinece[#], Department of Computer Application*
Pratibha College of Commerce & Computer Studies, Chinchwad, Pune, India
dipalimahajan.pune@gmail.com, suvarna.gogs@gmail.com

Abstract:-
In this era of information age, there is vast amount of data which is flowing among internet, systems, telephones etc. Data is coming from everywhere, from social media sites, digital pictures, videos etc. This data is being collected and stored at unprecedented rates. This data is called "Big data", which is large and complex. It contains structured and unstructured data. To store and manage the large volume of data, analyse it and extract meaningful information from it is a big challenge. Several approaches are available for collecting, storing, processing, analysing big data. Data mining is a technique by which useful data can be extracted from big data. Data mining is useful for discovering interesting patterns as well as descriptive, understandable models from large scale data. This paper overviewed types of big data and challenges in big data for future. This paper also gives quick overview of emerging technologies such as the Apache Hadoop framework and Apache Map Reduce.

Keywords*:* - Big Data, Business intelligent, Data mining, 3V's, Apache Hadoop, Apache MapReduce

.

Introduction:

Data has become an essential part of every economy, industry, organization, business function and individual. Each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices and big companies as Google, Apple, Facebook, Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more.
Now a day, the organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. Due to this, a small, but growing group of pioneers is achieving breakthrough business outcomes.

I. BIG DATA

Big data is a buzzword used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. 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 be easily 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.

Three V's in Big Data

Doug Laney was the first one talking about 3V's in Big Data Management. These 3V`s are illustrated in fig 1.

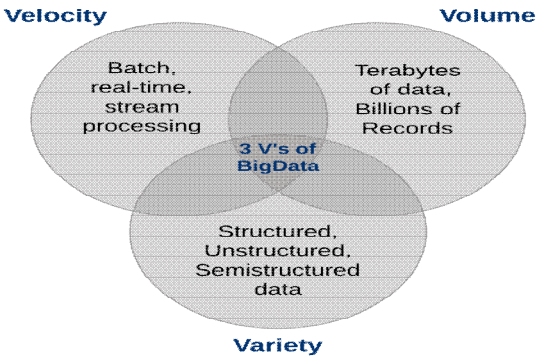

Fig1: Model Expressing Big Data in 3 dimensions

Volume: The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate. Think petabytes instead of terabytes.

Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types,

including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files, shadow data such as access journals and Web search histories and more.

Velocity: Data is generated as a constant stream with real-time queries for meaningful information to be served up on demand rather than batched. Data is always in motion. The speed at which data is created, processed and analyzed continues to accelerate.

Nowadays there are two more V's

Variability: There are changes in the structure of the data and how users want to interpret that data.

Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach. Meaningful insights that deliver predictive analytics for future trends and patterns from deep, complex analysis based on machine learning, statistical modeling, and graph algorithms. These analytics go beyond the results of traditional business intelligence querying and reporting.

II. DATA MINING FOR BIG DATA

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. 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:

1. Classification
2. Estimation
3. Prediction
4. Association rules
5. Clustering
6. Description

1. 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 neighbor classifier, Naive Bayes, 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.

2. 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.

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

4. 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.

5. 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.

6. Description: Description provides a concise and succinct summarization of a collection of data and distinguishes it from others.

Difference between Big data and Data mining

Big data:

1. Big data is a term for large data set
2. Big data is the asset.
3. 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:

1. Data mining refers to the activity of going through big data set to look for relevant information.
2. Data mining is the handler which provides beneficial result.
3. Data mining refers to the operation that involves relatively sophisticated search operation.

III.TECHNOLOGIES BEHIND BIG DATA ANALYTICS

The Apache Hadoop Framework and MapReduce are new technologies are emerging to make big data analytics possible and cost-effective.

The Apache Hadoop framework is evolving as the best new approach. The Hadoop framework redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources.

The Hadoop open-source framework uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In addition to offering high availability, the Hadoop framework is more cost- effective for handling large, complex, or unstructured data sets than conventional approaches, and it offers massive scalability and speed. The fig 2. gives Apache Hadoop Framework :

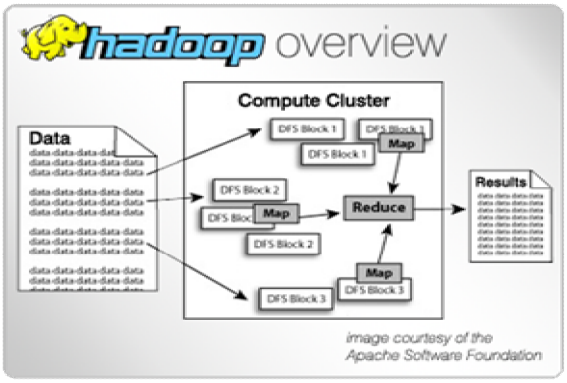

Fig 2: The Apache Hadoop Framework

Hadoop MapReduce (**Hadoop Map/Reduce**) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache **Hadoop** project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.

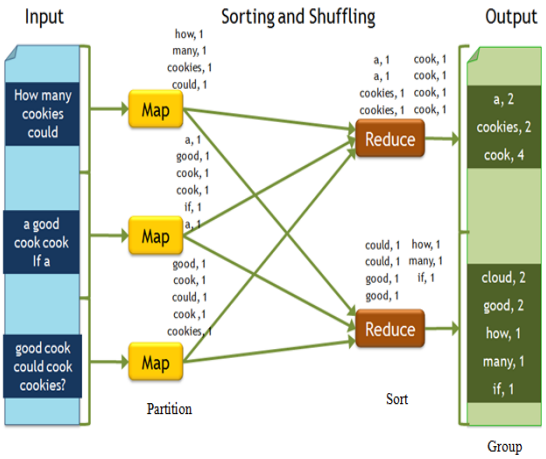

Fig. 3: MapReduce Process

Fig.3 elaborates the process of MapReduce. Here, Map and Reduce are the user programming code.

IV. BIG Data at the Edge

Today big data analytics focuses on managing and analyzing unstructured data from business and social sources such as e-mail, videos, tweets, Face book posts, reviews, and Web behavior. This type of big data analytics promises to provide significant value to organizations. Data from intelligent systems and sensors is some of the largest volume, fastest streaming, and/or most complex big data. The data sources are distributed across the network and data is collected by an enormous variety of equipment, such as utility meters, traffic and security cameras, fitness machines, and medical devices. Everywhere connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Edge data can provide significant value to both the private and public sector as a source of enormous potential for gaining deeper, richer insight faster and more cost-effectively than in the past. In many cases, analysis of edge data can help organizations respond to events and solve problems that were previously out of reach.

V. CONCLUSION

Big data is a collection of complex data sets while Data mining is an analytical process designed to explore data(usually large amount of data-typically business or market related-also known as “big data”) for consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. Big data technologies, like Hadoop will provide most relevant solution in upcoming years merely for all business sectors.

References:

[1] Bharti Thakur, Manish Mann, “ Data Mining for Big Data: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014, ISSN: 2277 128X

[2] Prashant Kumar, Khushboo Pandey, “Big Data and Distributed Data Mining: An Example of Future Networks”, International Journal of Advance Research and Innovation, Volume 1, Issue 2 (2013) 36-39, ISSN 2347 - 3258

[3] Kusnetzky, Dan. "What is "Big Data?"". DNet Vance, Ashley (2010). "Start-Up Goes After Big Data With Hadoop Helper". New York Times Blog

[4] a b c d e f "Data, data everywhere". The Economist (25) 2010 Retrieved (9) 2012

[5] "E-Discovery Special Report: The Rising Tide of Nonlinear Review". Hudson Global Retrieved 2012.by Cat Casey and Alejandra Perez

[6] "What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review" Forbes. Retrieved 2012

[7] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner Retrieved 2001

[8] Richard Waters (2013). "Google search proves to be new word in stock market prediction". Financial Times Retrieved 2013

[9] Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center" National Security Agency Central Security Service. Retrieved 2013

[10] Hellerstein, Joe (2008). "Parallel Programming in the Age of Big Data" Gigaom Blog