Big Data- A Review

Roopashree S
Assistant Professor
Department of Computer Science and Engineering
T. John Institute of Technology
Bengaluru, India

Kashif Akhtar
Department of Computer Science and Engineering
T. John Institute of Technology
Bengaluru, India

Akhil Kumar
Department of Computer Science and Engineering
T. John Institute of Technology
Bengaluru, India

***Abstract*— It is now believed that we are leaving a data trace with every phone call made, websites browsed, cards swiped and security cameras passed. No matter where we are in the world and how it's being used, one thing is clear- BIG DATA is EVERYWHERE. Big data analysis is on its way to become a crucial issue with increased adoption among every industry and with more people keen to access even bigger data. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools.Useful data can be extracted from BIG DATA with the help of DATA MINING. Data Mining is a technique for discovering interesting patterns as well as descriptive understandable models from large scale data. Often the requirements for big data analysis are not well understood by the developers and business owners, thus creating an undesirable product. So this paper would primarily deal with the BIG issue of BIG data and provide an insight for organizations to develop expertise and a process of creating small scale prototypes quickly and test them to demonstrate its correctness, matching with business goals. The transition for the data management between the traditional relational data base system and BIG DATA is also focused.**

Keywords* — *Cloud Computing, Hadoop, MapReduce, DataMining, 3V's, Relational Database

I. INTRODUCTION

"Big data" is a big buzz phrase [1] in the IT and business world right now – and there are a dizzying array of opinions on just what these two simple words really mean.

But big data has changed dramatically. The evolution of the Web has redefined:

- The speed at which information flows into these primary online systems.
- The number of customers a company must deal with.
- The acceptable interval between the time that data first enters a system, and its transformation into information that can be analysed to make key business decisions.
- The kind of data that needs to be handled and tracked.

Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.

An IT professional know how difficult it can be to find a solution capable of handling a task like big data management that combines the following benefits:-

- Scalability
- Performance
- Ease of use
- Low total cost of ownership (TCO)

Fig.1. Big Data

II. CLOUD COMPUTING

Cloud Computing [2] is a technology which depends on sharing of computing resources than having local servers or personal devices to handle the applications.

The only thing that must be done at the user's end is to run the cloud interface software to connect to the cloud. Cloud Computing consists of a front end and back end. The front end includes the user's computer and software required to access the cloud network. Back end consists of various computers, servers and database systems that create the cloud.

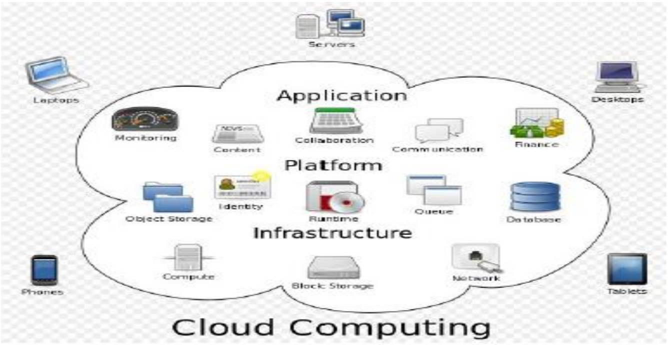

Fig.2. Cloud Computing

III. HADOOP

Hadoop [3] is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, it accomplishes two tasks: massive data storage and faster processing as in Fig.(3). Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant.

Hadoop leverages a cluster of nodes to run MapReduce [6] programs massively in parallel. A MapReduce program consists of two steps: the Map step processes input data and the reduce step assembles intermediate results into a final result.

Both Input and Output are stored in Apache *Hadoop Distributed File System (HDFS*).

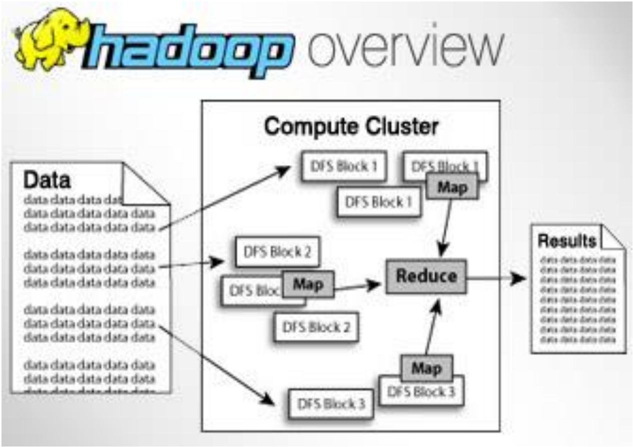

Fig.3. Hadoop Functioning

A. MapReduce

The input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

MapReduce is a framework pioneered by Google for processing large amounts of data in a distributed environment.

Hadoop is the open source implementation of the MapReduce substructure. Due to the simplicity of its programming model and the run-time tolerance for node failures, MapReduce is widely used by companies such as Facebook, the New York Times, etc.

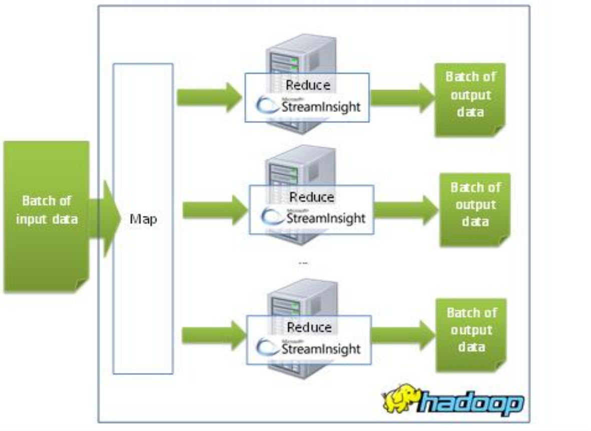

Fig.4. Map Reduce

B. Hadoop Distributed File System

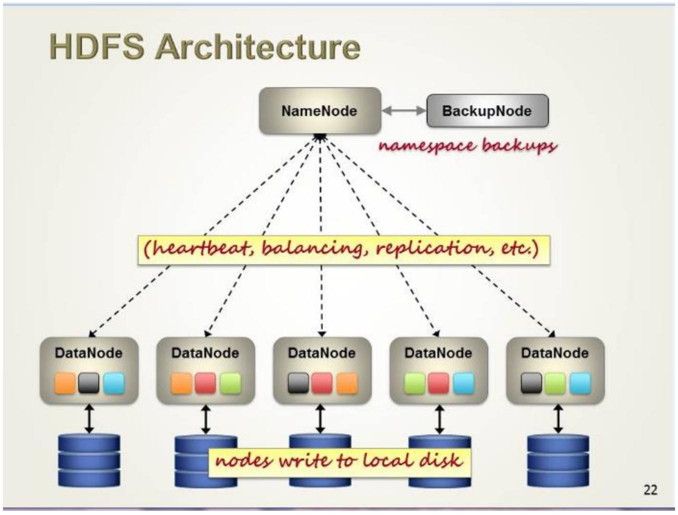

Fig.5. Architecture of HDFS

HDFS [7] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

IV. DATA MINING

Data Mining [4][8] is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) Preprocessing and Analysis, (2) Model building or pattern identification with validation/verification, and (3) Deployment (i.e., the application of the model to new data in order to generate predictions).

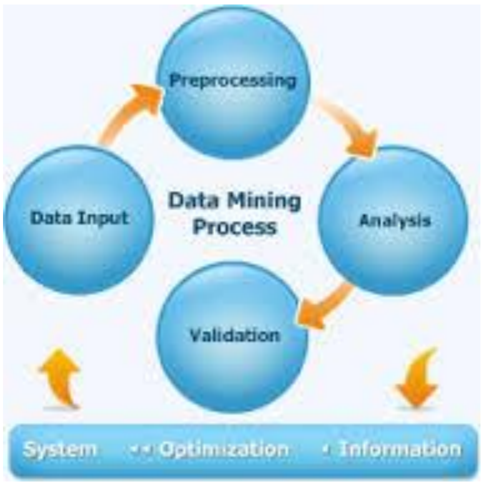

Fig.6. Data Mining

1. *Preprocessing and Analysis:*

This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range .Then, depending on the nature of the analytic problem, the first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

2. *Model building and validation*

This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

3. *Deployment*

That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

TABLE I
DIFFERENCE BETWEEN BIG DATA AND DATA MINING

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

V.3V’s

Big data can be categorized into 3 genres namely volume, variety and velocity, abbreviated as 3V’s [9].

A. *Volume*

Big data implies enormous volumes of data. The volume factor deals with the enterprises capacity to manage the ever-growing data size from some terabytes to petabytes of information. It used to be employees created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is large.

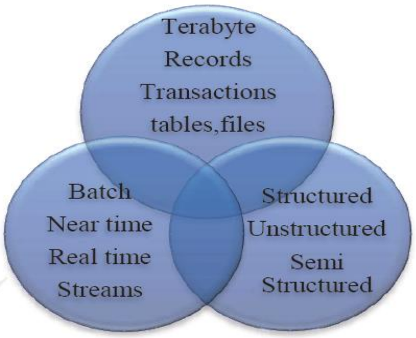

Fig.7.Big Data as 3’Vs

B. *Variety*

It refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

C. *Velocity*

Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and return on investment if you are able to handle the velocity.

V. RELATIONAL DATA BASE

The relational database [5] model is based upon tables or relations. In relational database model the physical implementation of the database is abstracted away from the user. Users query the database using a high-level query language, such as SQL. The relations are made up of columns, which have headings indicating the attribute represented by that column. Tables have key fields, which can be used to identify unique records. Keys relate tables to each other. The rows of the relation are also called tuples, and there is one tuple component for each attribute – or column – in that relation. A relation or table name, along with those relation's attributes, make up the relational schema. Relational Database models are server-centric.

Relational Database [10] is designed by creating a table for each entity type, choosing or inventing a primary key for each table.

Adding foreign keys to represent one-to-many relationships; creating new tables to represent many-to-many relationships; defining referential integrity constraints; evaluating schema quality and by making necessary improvements.

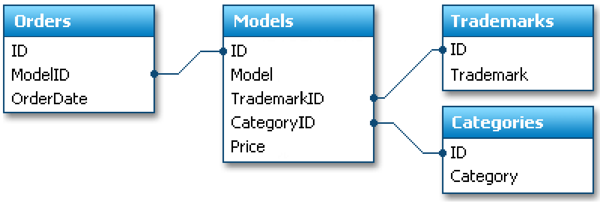

Fig.8.Performing Complex Queries

The real power of relational systems lies in the ability to perform complex queries over the data. Relational systems are well understood, and can be highly optimized in terms of queries, scalability, and storage.

Most of the programming within a RDBMS is accomplished using stored procedures. Often procedures can be used to greatly reduce the amount of information transferred within and outside of a system. For increased security, the system design may also grant access to only the stored procedures and not directly to the tables. Fundamental stored procedures contain the logic needed to insert new data and update existing data. More complex procedures may be written to implement additional rules and logic related to processing or selecting the data.

VI. WHY TRANSITION FROM RELATIONAL DATABASES TO BIG DATA

The following table enunciates the difference between the traditional relational databases and Big Data database systems (Hadoop). Owing to the enormous amounts of data being generated and analyzed real time to provide intelligence to the decision support systems, there is a clear need of the time to transition to Big Data.

TABLE II

DIFFERENCE BETWEEN RDBMS AND HADOOP

	RDBMS .	**Hadoop**
Description	Traditional row-column databases used for transactional systems, reporting, and archiving.	Distributed file system that stores large amount of file data on a cloud of machines, handles data redundancy etc. On top of that distributed file system, Hadoop provides an API for processing all that stored data - Map-Reduce. On top of this basic schema a Column Database, like hBase can be built.
Type of data supported	Works with structured data only	Works with structured, semi-structured, and unstructured data
Max data size	Terabytes	Hundreds of Pitabytes
Limitations	Databases must slowly import data into a native representation before they can be queried, limiting their ability to handle streaming data.	Works well with streaming data

VII. CONCLUSION

Big data isn't just hype – and it's much more than a buzz phrase. Today, companies across industries are finding they not only need to manage increasingly large data volumes in their real-time systems, but also analyze that information so they can make the right decisions – fast – to compete effectively in the market. Security in cloud environments is an important aspect for organizations. Using proposed approaches, cloud environments can be secured for complex business operations.

To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. We regard Big data as an emerging trend and the need for Big data mining is rising in all science and engineering domains. With Big data technologies, MapReduce, HDFS we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCE

[1] White Paper- BY DATASTAX CORPORATION October 2013.

[2] International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[3] CS561-Spring 2012 WPI, Mohamed Y. Eltabakh

[4] Leveraging Massively Parallel Processing in an Oracle Environment for Big Data Analytics

[5] DataStax Enterprise Reference Architecture

[6] ie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[7] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

[8] Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.

[9] Thakur et al., International Journal of Advanced Research in Computer Science and Software Engineering 4(5), May - 2014, pp. 469-473

[10] From Databases to Big Data by Sam Madden – Article published in IEEE Internet Computing magazine.