Big Data Analytics for Social Network - The Base Study
Miss. Payal Rajkumar Rathi #1, Prof. A S. Bhala #2, Miss. Sarita G. Rathi #3

#1 Student of M.E. CSE Department, Sant Gadge Baba Amravati University, India
#2 Prof. of CSE Department, Sant Gadge Baba Amravati University, India
#3 MBA BE(IT) Sant Gadge Baba Amravati University, Amravati India

Abstract- Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy. Big data is a term that describes the large volume of data – both structured and unstructured. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Unstructured data like email, HTML, social data and images. Big data is the vast sets of information gathered by researchers at companies like Facebook, Google and Microsoft from patterns of cellphone calls, text messages and Internet clicks by millions of users around the world. The challenging part about big data is not the collection, but the management of the data. The analysing of such data is challenging issues. To processing of huge amounts of data, companies will need to incorporate big data into there previous information management system.

Keywords: Big Data Analytic Tools, Data Mining, Hadoop and MapReduce, unstructured data, HBase and Hive tools, User-Friendly tools

Introduction:
Big Data term appeared first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of Big Data and the Next Wave of Infra Stress. It is complicated and large data set to process to the traditional data management tools[1].Data are collected and analyzed to create information suitable for making decisions. Every time new data medium was invented to access data from database. Hence data provide a rich resource for knowledge discovery and decision support. Simply Big Data as the amount of data just beyond technology’s capability to store, manage and process efficiently. Big data help to extract information that best matches with user interests. Big data is a heterogeneous collection of both structured and unstructured data. Structured: Structured Data are numbers and words that can be easily categorized and analyzed.
Unstructured: Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. Big data used data sets to handle the big amount of data from different simulation models, observation sources and different man-based products. There are different definitions of Big Data (Manyika et al 2011, Gartner 2011, 2014, NESSI 2014, Kitchin 2014, IBM 2014). Big data has been used to convey all sorts of concepts like, huge quantities of data, social media analytics, next generation data management capabilities, crawled documents, real-time data, emerging requirements for real time information and much more. Here we can understand that “big data” covers all the technologies and techniques that is too complex to be processed by traditional means in an acceptable way. The complexity comes from the multiple dimensions of data (Ishwarappa 2015, NESSI 2014, Manyika et al 2011 Gartner 2014, IBM2014), and particularly from Volume, Variety, Veracity or Velocity, the so-call four “Vs” of big data. Doug Laney[2] was the first one in talking about 3 V’s in Big Data management: We define Big Data as the amount of data just beyond technology’s capability to store, manage and process efficiently. These imitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and methods) used to analyze it [3]. Consider the Internet data. Google has been indexed one million web pages in 1998, but in 2000 it rapidly reached to 1 billion and in 2008 it reached to 1 trillion. The social networking applications, such as Facebook, Twitter, etc. [5 v’s] . large amount of data are generated within minute. A recent study estimated that every minute, Google receives over 4 million queries, e-mail users send over 200 million messages, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content, and Twitter users generate 277,000 tweets [4] [5].Big data have different sources, like, social networks that are made of social entities. For example, individual social entities can be linked by friendship. Another example, corporations can be linked by financial exchange. In these examples, a social entity is connected to another social entity as his next-of-kin, friend, collaborator, co-author, classmate, co-worker,
team member, and business partner. All entities depend on another entities. Map Reduce is a programming model. Map Reduce an associated implementation for processing and generating large data sets with a parallel. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Nowadays, social networking sites or services such as Face-book, Google+, LinkedIn, and Twitter [6, 7, 8] are commonly in use. There are number of social networking sites, among them Facebook allows users to create a personal profile, add other Facebook users as friends, send them messages, pictures and exchange data. In addition, Facebook users can also join common groups and categorize their friends into different customized lists (e.g. classmates, co-workers).

II. BACKGROUND & RELATED WORK

The data mining is finding the value from volume, data processing and prediction. In any business and technology, the higher level data generated during its day-to-day operation is large in volume, velocity, variety 3 V’s. The definition of high volume, velocity and variety information assets that demand, the new forms of information processing for enhanced insight and decision making. For maintaining reasonable processing times Big Data provides a scalable solution that is terabytes and petabytes are the unknown capacity of data that was measured in megabytes and gigabytes. “Big data allows to organization tools, processes and procedures that is manipulate and manage very large data sets”. The issue solved by the data warehouse are measuring large volumes of data and have ability to manage large Big Data sets. For usefulness of Big Data mining, we would like to mention the work that Global Pulse is doing [9] using Big Data to improve life in developing countries.

Data mining attempts to extract valuable information from database. Data mining itself helps to improve the quality and trustworthiness of the data, open source technology toolkit for analyzing real-time data and sharing and provide intelligent querying functions. Mining requires integrated, cleaned, trustworthy, global network of Pulse Labs and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and Big Data computing environments, to pilot the approach at country level. The industrialized world can access the Big Data mining revolution. For the future queries Big Data will be automatically generated. In some cases result of analysis is required immediately. The privacy of data is the large concern. For example, a user’s location information can be tracked through several stationary connection points (e.g., cell towers). Data Mining developed new technologies that have different parts like capturing devices, sensors, and mobile applications.

III. APPLICATION OF BIG DATA AND DATA MINING

The term ‘Big Data’ appeared for First time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of “Big Data and the NextWave of infraStress” [10]. The large amount of data stored in Big Data which created in day to day. Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization’s storage [11]. Machine-to-machine interactions, call detail records, RFID systems; all these are create their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast growing streams of unstructured and semi structured data from social media. More complex information stored in Unstructured Data. These data cannot be separated into categories easily.

There are huge amount of data, but it is the reality for many organizations. All sectors organizations have at least 100 terabytes of data, and more than a petabytes. Big Data have data management capabilities. There are (3Vs) Big Data vectors for volume, velocity, and variety [13]:

![Image 1: Big Data Organization](image1.png)

![Image 2: Three V's of Big Data Vectors](image2.png)
VOLUME (SCALE): The amount of data in addition to data volume is increasing exponentially. There are 44x increases from 2009 to 2020 from 0.8 zeta byte to 35zb.

VELOCITY (SPEED): Data comes in the form of stream of bit. Speed rate in collecting processing of data. The data is begin generated fast and need to be processed fast for obtaining useful information from it in real time, online data analytics. Eg : e-promotions, healthcare monitoring.

VARIETY (COMPLEXITY): The different data types along with latest insights are found when analysing together, having various formats, types, and structures. Text, numerical, images, audio, video, graph, sequence, time series, social media data, multi-dimensional arrays, etc. and Static data vs. streaming data, a single application can be generating and collecting many types of data, to extract knowledge-all these types of data need to correlate.

IV. MEASURING THE VALUE AND POTENTIAL YIELD OF BIG DATA

The people are capturing and digitizing more information than ever before. Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different Formats, from independent or connected applications. The flow of data flood increases in recent years. According to IDC, the world produced one zeta byte 1,000,000,000,000 gigabytes of data in 2010. this data explosion are over five billion mobile phones, 30 billion pieces of content shared on Facebook per month, 20 billion Internet searches per month, YouTube having over 72 hours of video data uploaded every minute, Twitter users generate 277,000 tweets and millions of networked sensors connected to mobile phones, retail packaging and more.

Apache Hadoop is a open source project that processing of large scale data. The Apache Hadoop platform has proven successful in the past in the storage and processing of large scale unstructured and semi structured data with technologies such as HDFS, MapReduce and HIVE. MapReduce (implemented on Hadoop and Yahoo ) is a framework for parallel distributed processing large volumes of data, developed by Google. To process high volumes of big data, researchers have used a high-level programming model called MapReduce [14] over the past few years. The model uses parallel and distributed computing on large grids of nodes. Big Data application analysis framework include the following: support for multiple data types, handle batch processing and real time data streams, overcome low latency, provide cheap storage, and integrate with conventional deployments.

V. FUTURE WORK OF BIG DATA

In this paper, we proposed Big Data Analytics in social network. There are many future important challenges in Big Data management and analytics, that depends upon the nature of data. There are 84 different big data programs spread across six departments [15]. Private Sector: Wal-Mart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data. In Facebook handles 40 billion photos from its user base. Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide [16].

VI. TOOLS: OPEN SOURCE REVOLUTION

The open source software revolution is an important part of Big data. It is a good tool to promote a company’s image, including its commercial product. No. of companies like Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source project.

Hadoop : Hadoop is an open-source software framework for storing data. It provides large storage for any kind of data. Apache Hadoop is a open source project that allows reliable, faster and distributed processing of large scale data. processing of large scale unstructured data with technologies as below:

HBase , Map Reduce,HDFS and Hive.

HBase: Hadoop is an open source project, which leverages current hardware in a way that minimizes cost while maximizing resource utilization. For the Map Reduce jobs both the input and output serve by HBase in Hadoop.
Map Reduce: Apache Hadoop is one of the popular open-source implementation of Map Reduce framework. An average of one hundred thousand Map Reduce jobs are executed on Google’s clusters every day, parallel and distributed implemetation used for the clusters.

Hive: Hive is a popular data warehousing software that is built over the Hadoop framework. Hadoop provides flexibility i.e. it enables storage and processing of all forms of data (structured and unstructured). Hadoop complete file system stored the analysis of large datasets.

User-friendly tools for Big Data: Hive supports query language SQL that is familar with syntax and structure of SQL, like frameworks to run the queries directly against data stored in Hadoop.

VII. CONCLUSION

We have entered in the world of Big Data. Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year and improving the profitability and success of many enterprises. Technical challenges, business applications, scientific data research, timeliness, visualization, at all stages of the analysis direct from data achievement to result of address in one domain, the main challenges for the future. The challenges of Big Data are limited as compared to potential benefits. It have ability to make connections among the trillions of bytes of data we have access from the data source. Big Data mining will help us to discover knowledge. Big Data is inspirable when it comes to processing of huge amounts of Data in actual.

REFERENCES