Big Data

Ali Ahmed

Baghdad, Monsur, Iraq

aliahmed9683@yahoo.com

Abstract: The data is growing day by day all over the globe and hence it results into the concept as Big Data . The term Big data is a collection of data sets which is very large in size as well as complex. Generally, size of the data is Petabyte and Exabyte. As the internet is growing, amount of big data continues to grow.

This research paper discusses about the various aspects of big data in terms of volume velocity and variety. This includes the three V's of big data which are velocity, volume and variety. This paper also briefs on phases of big data, challenges and tools used for big dat. Also, this research is intended towards the merits and demerit of Big data. The review of literature by various scholars, researcher, journal are also described in this paper. The use of big data is also forecasted in the present study.

Keywords: Big Data , Velocity, Volume, Variety, Petabytes.

I. Introduction:

Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. The term Big Data is now used almost everywhere in our daily life. The term Big Data came around 2005 which refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.[1]

II. 3 V's of Big Data :

The 3Vs that define Big Data are Velocity and Volume, Variety

- **Volume:** There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.

- **Velocity:** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

- **Variety:** Today, data comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

III. Review of Literature :

A research paper entitled "Challenges and Opportunities with Big Data" by the researchers Agarwal D., et.al. (2012). The paper elaborates the challenges and opportunities of Big Data and also describes the phases in the processing pipeline which contains Data Acquisition and Recording, Information Extraction and Cleaning, Data Integration, Aggregation and Representation, Query Processing, Data Modeling and Analysis, Interpretation. The author also gives the conceptual view of the Big Data Analysis Pipeline and also highlights the challenges include not just the issues of heterogeneity but also issues of scale, lack of structure, error handling, privacy, timeliness, provenance and visualization.

Furthermore, authors suggest that challenges will require transformative solutions and will not be addressed naturally by the next generation of industrial products. [3]

The authors Bhosale H. and Gadekar D. (2014) entitled a paper "A Review Paper on Big Data and Hadoop". The paper elaborated the 3 V's of Big Data Velocity, Volume and Variety. They also describe the problems with Big Data processing like Heterogeneity and Incompleteness, Scale, Timeliness, Privacy and Human Collaboration. Researchers also gives conceptual framework of Layered Architecture of Big Data System and Hadoop Distributed File System(HDFS). [4]

S. Justin Samuel et.al. (2015), Started that Big Data is comprised of large datasets that is hard to process

with convential data processing Systems. In this paper, authors have done an elaborate study on Big Data and its research Challenges. They have highlighted the existing problems and have presented the research opportunities. The authors have briefed about the various research areas in this field through diagram. The paper has highlights the ongoing research in Big data analytics through a structured tabular form.

Researchers also describes conceptual view of Big Data classes and the major research areas like Applied Ontology, Security, Storage and Transport, Accessibility, Inconsistencies and Mobility. [5]

IV. Phases of Big Data

1. **Data Acquisition**: The first phase in Big Data is acquiring the data itself. Because of use of internet, smart phone and social network, and a wide array of sensors the rate of data generation is rising exponentially. Most of this data are not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data can play very important role in making decision when merged with other valuable data.

2. **Data Extraction**: All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user's movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the data was initially generated. For instance, footage of a security camera with the same frames may be discarded however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor.

   Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under common platform to standardize data extraction is a major challenge.

3. **Data Collation**: Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user. Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.

4. **Data Structuring**: Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.

5. **Data Visualization**: Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore "humanizing" the data becomes all the more important.

6. **Data Interpretation**: The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: Retrospective Analysis includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas. Prospective Analysis includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis. Problems accruing from such interpretations pertain to fallacious and misleading trends being predicted. This is particularly dangerous due to an increasing reliance on data for key decisions. For example, if a particular symptom is plotted against the likelihood of being diagnosed with a particular disease, it might lead to

misinformation about the symptom being caused due to the particular disease itself. Insights gained from data interpretation are therefore very important and the primary reason for processing big data as well. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.[4]

V. Challenges of Big Data:

There are numerous challenges facing Big Data:

- The first challenge for organizations is to choose and select the relevant and important data. With such high volumes of data, it becomes important for organizations to able to separate the relevant data.

- The second challenge is that even now, in organizations, many data points are not connected. This problem of connectivity is a severe hurdle. Big Data is all about collection of data from various transaction points. Organizations need to be able to manage data from across its enterprises.

- To leverage Big Data, one has to work across departments such as IT, Engineering and Finance. Thus the ownership and procurement of this data has to be a co-operative endeavor across these departments. This proves to be a significant organizational challenge.

- There is a security angle related to Big Data collection. This is a major obstacle preventing companies from taking full advantage of Big Data Analysis.

- Several issues will have to be addressed to capture the full potential of big data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. Organizations need not only to put the right talent and technology in place but also structure workflows and incentives to optimize the use of big data. Access to data is critical companies will increasingly need to integrate information from multiple data sources, often from third parties, and the incentives have to be in place to enable this.[1]

VI. Pros and Cons of Big Data

The term “Big data” refers not only to large data sets, but also to the frameworks, techniques, and tools used to analyze it. It can be collected through any data-generating process such as social media, public utility infrastructure, and search engines. Big data may be either semi-structured, structured, or unstructured.

Typically big data is analyzed and collected at specific intervals, but real-time big data analytics collect and analyze data constantly.

The purpose of this continuous processing loop is to offer instant insights to users.

- **Pros:**
  1. It allows Businesses to detect errors and fraud quickly, which significantly mitigates against losses.
  2. It provides major advantages from a competitive standpoint. Real-time analysis allows businesses to develop more effective strategies towards competitors in less time
  3. The data collected is valuable and offers businesses a chance to improve profits and customer service.
  4. Proponents of big data point out that healthcare organizations can use electronic medical records and data from wearables to prevent deadly hospital infections.

- **Cons:**
  1. Companies hoping to use big data will need to modify their entire approach as data flowing into the company becomes constant rather than periodic: this mandates major strategic changes for many businesses.
  2. One of the biggest concerns many laypeople and politicians have about real-time analysis of big data is privacy.
  3. Civil liberties advocates have attacked the use of big data from license plate scanners and drones.[6]

VII. Tools of Big Data

1. Pentaho Business Analytics

Pentaho is another software platform that began as a report generating engine. You can hook up Pentaho's tool to many of the most popular NoSQL databases such as MongoDB and Cassandra. Once the databases are connected, you can drag and drop the columns into views and reports as if the information came from SQL databases. Pentaho also provides software for drawing HDFS file data and HBase data from Hadoop clusters. One of the more intriguing tools is the graphical programming interface known as either Kettle or Pentaho Data Integration. It has a bunch of built-in modules that you can drag and drop onto a picture, then connect them. Pentaho has thoroughly integrated Hadoop and the other sources into this, so you can write your code and send it out to execute on the cluster.

2. Karmasphere Studio and Analyst

Many of the big data tools did not begin life as reporting tools. Karmasphere Studio, for instance, is a set of plug-ins

built on top of Eclipse. It's a specialized IDE that makes it easier to create and run Hadoop jobs. I had a rare feeling of joy when I started configuring a Hadoop job with this developer tool. There are a number of stages in the life of a Hadoop job, and Karmasphere's tools walk you through each step, showing the partial results along the way.

Karmasphere also distributes a tool called Karmasphere Analyst, which is designed to simplify the process of plowing through all of the data in a Hadoop cluster. It comes with many useful building blocks for programming a good Hadoop job, like subroutines for uncompressing Zipped log files. Then it strings them together and parameterizes the Hive calls to produce a table of output for perusing.

3. Talend Open Studio

Talend also offers an Eclipse-based IDE for stringing together data processing jobs with Hadoop. Its tools are designed to help with data integration, data quality, and data management, all with subroutines tuned to these jobs.

Talend Studio allows you to build up your jobs by dragging and dropping little icons onto a canvas. If you want to get an RSS feed, Talend's component will fetch the RSS and add proxying if necessary. There are dozens of components for gathering information and dozens more for doing things like a "fuzzy match." Then you can output the results.[2]

4. Skytree Server

Not all of the tools are designed to make it easier to string together code with visual mechanisms. Skytree offers a bundle that performs many of the more sophisticated machine-learning algorithms. All it takes is typing the right command into a command line.

Skytree is more focused on the guts than the shiny GUI. Skytree Server is optimized to run a number of classic machine-learning algorithms on your data using an implementation the company claims can be 10,000 times faster than other packages. It can search through your data looking for clusters of mathematically similar items, then invert this to identify outliers that may be problems, opportunities, or both. The algorithms can be more precise than humans, and they can search through vast quantities of data looking for the entries that are a bit out of the ordinary.

5. Tableau Desktop and Server

Tableau Desktop is a visualization tool that makes it easy to look at your data in new ways, then slice it up and look at it in a different way. You can even mix the data with other data and examine it in yet another light. The tool is optimized to give you all the columns for the data and let you mix them before stuffing it into one of the dozens of graphical templates provided.

VIII. Future of Big Data

Ongoing development of best practices for real-time analysis of big data should continue to be a priority for businesses and government agencies.

Each company will need to carefully assess whether pros of such big data use outweigh the cons for their particular case.

We must support and encourage fundamental research towards address in these technical challenges if we are to achieve the promised benefits of Big Data.

There is immense scope in Big Data and a huge scope for research and Development.

IX. Conclusion:

Big Data is changing the way we perceive our world. The impact big data has created and will continue to create can ripple through all facets of our life. We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, privacy ,security and all stages of the analysis pipeline from data acquisition to result interpretation. But, Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises Big Data is also changing things in the business world.

Companies are using big data analysis to target marketing at very specific demographics. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude.

X. References:

1. http://www.ijirae.com/volumes/Vol2/iss2/03.FBCS10080.pdf

2. http://www.infoworld.com/article/2616959/big-data/7-top-tools-for-taming-big-data.html

3. Challenges and opportunities of big data:http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

4. Bhosale H. and Gadekar D. (2014) “A Review Paper on Big Data and Hadoop”:http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf

5. S. Justin Samuel et.al.2015. “A SURVEY ON BIG DATA AND ITS RESEARCHCHALLENGES”: http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0515_1931.pdf