A REVIEW PAPER ON BIG DATA AND USING TECHNIQUES

[1]CHIRAG GOYAL,[2]RAMESH, [3]Er.HARKIRT SINGH BRAR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MIMIT, MALOUT

Abstract:-In this **REVIEW PAPER,** we explained **Big Data and Techniques** with the reference to cloud computing. **Cloud Computing** is based on **Internet** by which shared resources, software and information are present to computer and another device on demands. **Cloud computing** supports technology infrastructure "**in the cloud**" that hides the details from the users. **Cloud Computing** is **win-win** situations for users as well as providers. The main aim of the **Big Data** is to **Data** Expansion Day by day amount of data increase expontially because of today various data production source like smart electronics devices. As per **IDC (International Data Corporation)**.Every Day we create 25 quintillion byte of data so much that 95% of the data is world present has been used in the last year also .This data becomes everywhere sensor used to geather climate information post of the social media site Digital picture and videos, purchases transaction record. Report new data created per each person in word per second by 2020 will be 1.7 Mb The amount to total data in the word by 2020 a each Around 44 Zetta bytes and 175 Zetta Bytes 2025.

Keywords: - Big Data,5 V's, SNA,Data Mining,Hive.

I.INTRODUCTION

BIG DATA

Big Data is also data but with a huge size. Big data refers to the big, diverse set of information that grow at increasing data. It compasses the volume of details, the velocity or fast at which it is create and collectede, and the variety or scopes of the data points being covering.big data is a mostly buzzward used to describe a massive volume of the both structured and unstructured data and semi-structured.which is very large and complex to process using traditional database and software techniques.Data can be generated on web in various forms like text,image or video or social media post. Data has extra-large Volume,comes of Variety of Source,variety of formats and comes at us with a great Velocity normally refer to has big data .[1]

Big data is three types of data these are-

1.Structured Data – Data that has a well defined structure falls under structured data.This data has record divided into rows and columns.As a result,it is easy to read and managing data.This type of data constitutes about 10% of the today's total data and is accessible through database management systems. As technology performance has continued to improve, mostly numericaly, wherethe meaning of each data item is informed.the today's all data and is accessible through database manage system.

Example source of structured (or traditional) data include official registers that are created by private institution to stored data on individual, enterprising and real estates in industries that collect data about the processes. such all kind of database and also deriving value out of it. However, nowdays, we are forese issues when a sized of such data grows to a more extent, typical size are being in the rage of multiple zettabytes.

Example of Structured Data-

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
401	Chirag Goyal	Male	Finance	650000
404	Ramesh Kumar	Male	Admin	650000
405	ShushilSingla	Male	Admin	500000
406	ShubhojitGoyal	Male	Finance	500000
407	PriyaGoyal	Female	Finance	550000

Fig.1 example of structured data

2.Unstructured Data- Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

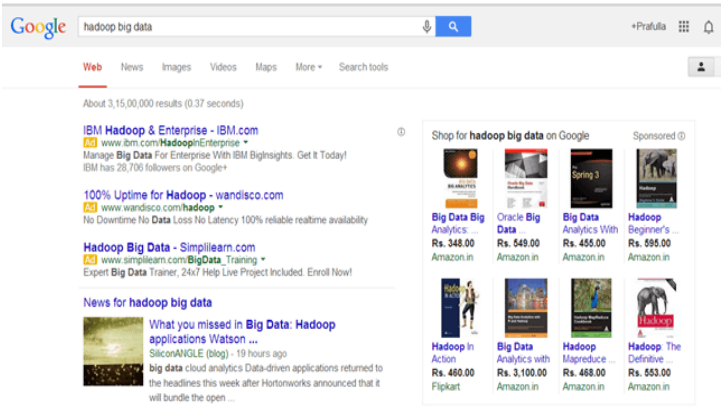

Fig.2 example of unstructured data

3.Semi-structured - data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.Personal data stored in an XML file-

```
<rec><name>Chirag
Goyal</name><sex>Male</sex><age>35</age><
/rec>
<rec><name>Ramesh
ku</name><sex>Female</sex><age>41</age></
rec>
<rec><name>PayalGoyal</name><sex>Male</se
x><age>29</age></rec>
<rec><name>Jerry
Singla</name><sex>Male</sex><age>26</age>
</rec>
<rec><name>KunalGoyal</name><sex>Male</se
x><age>35</age></rec>
```

Fig.3 example of semi-structured data

There are five characterstics for big data. They are Volume,Velocity,Variety,Value and Veracity.

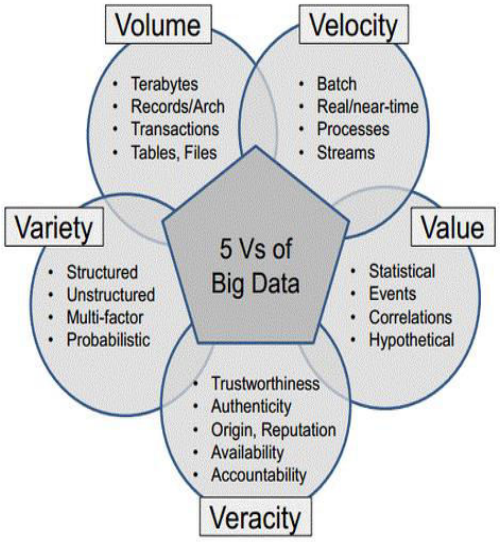

Fig.4 5V's of BIG DATA

5 V's of Big Data

In recent years, Big Data was defined by the "*3Vs*" but now there is "*5Vs*" of Big Data which are also termed as the characteristics of Big Data as follows:

1. Volume:

- The name 'Big Data' itself is similar to a size which is vast.
- Volume is a huge amount of data.
- Nowdays data volume is increasing from gigabytes to petabytes [2].
- To determine the value of data, size of data plays a very cirucial role. If the volume of data is very large then it is actually considered as a 'Big Data'.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the present year we will have almost 40000 ExaBytes of data.

2. Velocity:

- Velocity refering to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like 'velocity'.
- *Example:* There are more than 4.5 billion searches per day are made on Google Browser. Also, FaceBook users are increasing by 22%(Approx.) year by year.

3. Variety:

- Variety is another important characterstics of big data.
- It refers to the types of data. Data may be in different forms such asText,numerical,images,audios,videos,social media data[2].
- On twitter 500million tweets are sent per day and there are 220 million active users on it[3].
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that

Value! is the most important V of all the 5V's.

Technologies for Big Data

Social network analysis:-Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.

Examples	Application
Friendship Networks	College/school students, organizations or web (Facebook, Myspace etc.)
Followers networks	Twitter,Linkedin,Instagram etc.
Preference similarity networks	Pinterest, Twitter etc.
Interaction networks	Phone calls, Messages, Emails, Whatsapp etc.
Spread networks	Epidemics,Information,Rumours etc.
Co-actor networks	IMDB,etc.

TABLE1 Some examples of social network.

Management consultants use this methodology with their business clients and call it Organizational Network Analysis [**ONA**]. ONA allows you to x-ray your organization and reveal the managerial nervous system that connects everything.To understand networks and their participants, we evaluate the location and grouping of actors in the network. These measures give us insight into the various roles and groupings in a network -- who are the connectors, mavens, leaders, bridges, isolates, where are the clusters and who is in them, who is in the core of the network, and who is on the periphery?

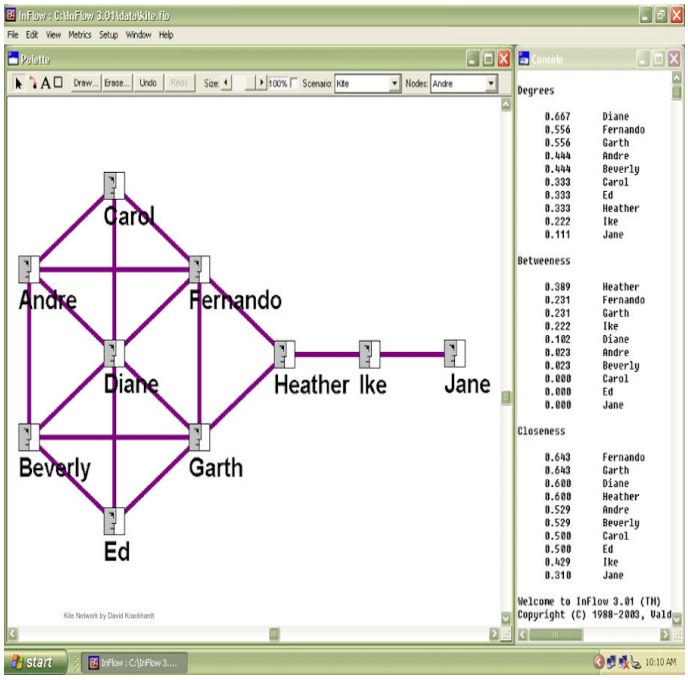

Fig.5 Social network analysis.

We look at a social network -- the "Kite Network" above -- developed by David Krackhardt, a leading researcher in social networks. Two nodes are connected if they regularly talk to each other, or interact in some way. Andre regularly interacts with Carol, but not with Ike. Therefore Andre and Carol are connected, but there is no link drawn between Andre and Ike. This network effectively shows the distinction between the three most popular individual centrality measures: Degree Centrality, Betweenness Centrality, and Closeness Centrality.

CONCLUSION

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to

analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.I actually don't really care about the promise of data unless they can deliver on that promise that comes with the data.

REFERNCE

[1] Harshawardhan S. Bhosale, Prof.Devendra P. Gadekar "A ReviewPaper on Big Data and Hadoop" in International Journal ofScientific and Research Publications, Volume 4, Issue 10, October2014.

[2]SMITHA T, V. Suresh Kumar "Application of Big Data in DataMining" in International Journal of Emerging Technology andAdvanced Engineering Volume 3, Issue 7, July 2013).

[3] IBM Big Data analytics HUB,www.ibmbigdatahub.com/infographic/four-vs-big-data.