RESEARCHERID
THOMSON REUTERS
[INIT: January 2017]
ISSN 2348 – 8034
Impact Factor- 4.022
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

BIG DATA: AWARENESS AND OPTIMAL USAGE

Mohamed Afsal[*1] & Avarsha K. S[2]
[*1&2] Acharya Bangalore B-School, Andrahalli main road, Off Magadi road, Bangalore – 560 091

ABSTRACT

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis,capture,data curation,search, store,transfer,visualization and information privacy.It may well be the next big thing in IT world.It has many characteristics such as volume,variability,velocity and complexity.Now a days it is a big problem to maintain historical data and analyses it.Our whole world is full of data in fact world is itself the data,Maintaining data of daily activity such as ATM Transaction,Online Booking,Online shoping, Scientific Researchcenter etc has been drastically increasing .it is the big question in IT just like garbage in silicon city.

The collection of data is in different format like flat file,xml, SQL tables,Ms-access ect and these two ambiguity is the big question in It.These ambiguity can be solved by different ETL tools like PowerCenterInformatica, Oracle Warehouse Builder (OWB),SAP Data Services.,IBM Infosphere Information Server,SAS Data Management,Elixir Repertoire for Data ETL.Data Migrator (IBI).

The use of appropriate Data Warehousing tools can help ensure that the right information gets to the right person via the right channel at the right time.

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.

Keywords- *BigData,ETL tools, Informatica PowerCenter, SAP Data Services, mappings, Challenges, big data analytics*

I. INTRODUCTION

Big Data is a term for data sets that are very large and complex that traditional data processing application softwares are inadequate to deal with them. The term "big data" often refers simply to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB

ISSN 2348 – 8034
Impact Factor- 4.022

.

- 1.23 billion People log onto Facebook daily active users (Facebook DAU) for September 2016, which represents 18% increase year over year.

- As of 2016, Google was handling roughly 2 trillion searches annually, over 100 billion searches per month and an average of 2.3 million searches per second!
- Oracle NoSQL Database has been tested to past the 1M ops/sec mark with 8 shards and proceeded to hit 1.2M ops/sec with 10 shards.

Big data can be described by the characteristics **Volume:**The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.**Velocity:** Big data is often available in real-time. **Variety:** Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion. **Variability:** Inconsistency of the data set can hamper processes to handle and manage it

II. WHY WE NEED BIG DATA?

Creating Smarter, Leaner Organizations:
A well thought out and executed Big Data and analytics strategy helps in making organizations smarter and more efficient. Today, Big Data is being used in many industries from criminal justice to health care to real estate with powerful outcomes. The same common sense approach to Big Data should be employed by organizations who need similar results.

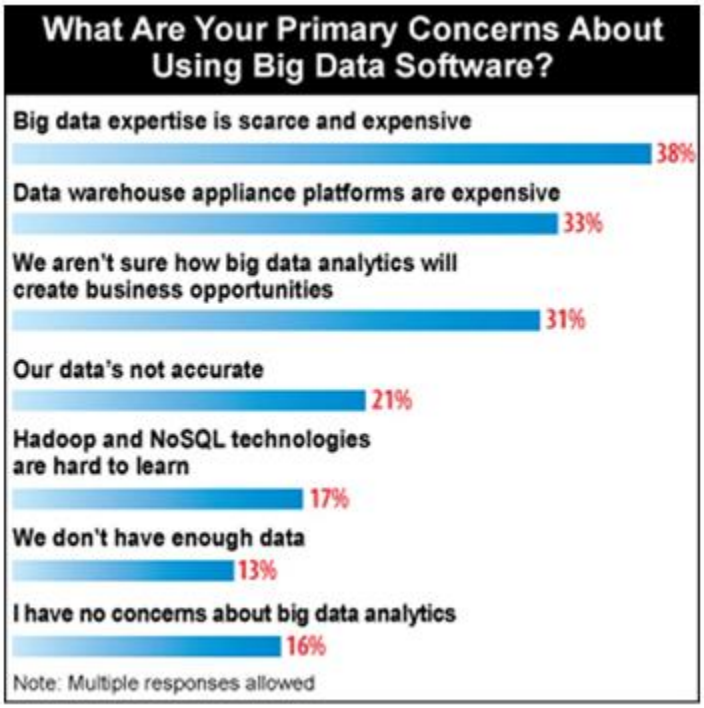
ISSN 2348 – 8034
Impact Factor- 4.022

Challenges

- analysis
- capture
- search,
- sharing
- storage
- transfer
- visualization
- Querying, updating
- and information privacy
- Can be misleading.
- Complex task to gain insight.
- Bespoke solution to query is not possible.

Sampling Big Data

- An important research question that can be asked about big data sets is whether you need to look at the full data to draw certain conclusions about the properties of the data or is a sample good enough.
- The name big data itself contains a term related to size and this is an important characteristic of big data. But sampling (statistics) enables the selection of right data points from within the larger data set to estimate the characteristics of the whole population.



- For example, about 100 million users login to Twitter per day. Is it necessary to look at all of them to determine the topics that are discussed during the day? Is it necessary to look at all the tweets to determine the sentiment on each of the topics? In manufacturing different types of sensory data such as acoustics, vibration, pressure, current, voltage and controller data are available at short time intervals. To predict down-time it may not be necessary to look at all the data but a sample may be sufficient.

ISSN 2348 – 8034
Impact Factor- 4.022

Solution
ETL Tool is the one of best solution now a days to deal with Big data

The list of some ETL tools.

- Oracle Warehouse Builder (OWB)
- SAP Data Services.
- IBM Infosphere Information Server.
- SAS Data Management.
- PowerCenterInformatica.
- Elixir Repertoire for Data ETL.
- Data Migrator (IBI)
- SQL Server Integration Services (SSIS)

About Informetica ETL tool

- **InformaticaPowerCenter** is a tool, supporting all the steps of Extraction, Transformation and Load process. A whole lot of product offerings are orchestrated around PowerCenter's
- Ability to connect to different technologies ranging from mainframe to CRM to Big Data.
- InformaticaPowerCenter is an easy to use tool. It has got a simple visual interface like forms in visual basic. You just need to drag and drop different objects (known as transformations) and design process flow for data extraction transformation and load. These process flow diagrams are known as **mappings**.
- Once a mapping is developed, it can be scheduled to run as and when required. In the background Informatica server takes care of fetching data from source, transforming it, & loading it to the target systems/databases.

About Informetica ETL tool

- PowerCenter can communicate with all major data sources (mainframe, Big Data, RDBMS, Flat Files, XML, SAP, Salesforce & the list goes on)
- Can move/transform data between them. It can move huge volumes of data in a very effective way, many a times better than even bespoke programs written for specific data movement. It can throttle the transactions (do big updates in small chunks to avoid long locking and filling the transactional log).
- It can effectively join data from two distinct data sources (even a xml file can be joined with a relational table). In all, Informaticahas got the ability to effectively integrate heterogeneous data sources & converting raw data into useful information.
  The Main primary concern about using Big data software according to our survey about data expertise, business opportunities, data accuracy, data analysis and technologies have shown as above the bar graph.

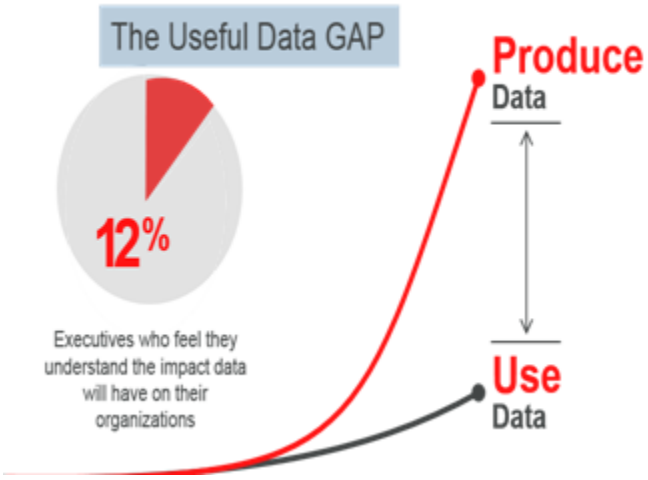
Benefits of ETL for big data analytics

- It is particularly advantageous to use an ETL tool in the following situations:
- When there are many source systems to be integrated
- When source systems are in different formats
- When this process needs to be run repeatedly (e.g. daily, hourly, real time)
- To take advantage of pre-built warehouses/marts.

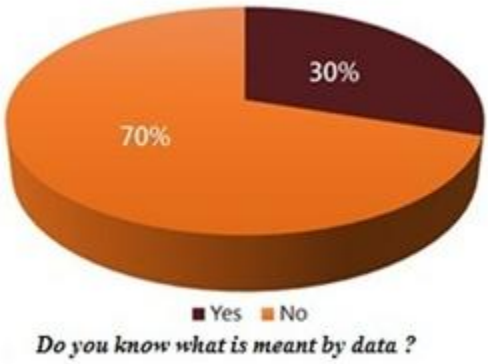
III. ANALYSIS AND REPORT

As per our survey and research, the results were shocking. Everyone using/creating a bunch of data but most of them doesn't know actually what data is. result of our survey shows that only 30% of the students know about data – since they have the topics in their curriculum. Rest of the 70% doesn't know that, but still they are using and producing data every moment of their life.

ISSN 2348 – 8034
Impact Factor- 4.022



Here the importance of awareness matters. The unawareness leads to wastage of resources and inappropriate use of data. Without knowing about



data no one can use it effectively. It is dangerous to waste the high potential resources. If they are aware about it, they can use it wisely and develop their own business as well as our country.

As we mentioned above only technical students who are studying about data are aware of it, the rest of the students who are studying other courses like BBA, B.com etc. are don't have the basic knowledge of it. This is because they are not getting sufficient knowledge and training on the topic.

Most of the universities are not giving any importance for data on their curriculum for the courses apart from technical courses. But the fact is – not only technical peoples are related to data, in this emerging world all are related, used and effected by data. So everyone should know what they are using/producing how they should handle that.

The issue what we noted is the redundancy of data. The root of this problem also same as lack of knowledge. For examples, when social record created, there are two different columns for Age and Date of birth. But in fact these are depended, by knowing the date of birth we can apply business logic to find the age. This saves millions or billions of shells as well as storage. Like these small things can be find solution for the biggest challenge of storage up to an extent.

ISSN 2348 – 8034
Impact Factor- 4.022

IV. CONCLUSION

Big data takes an important role to improve organization, society, company or whole world's growth with the historical data analysis and planning based on Big data. Awareness among the society about the impact, efficient use, maintenance and security factor of big data to overcome the challenges of which we are facing in globally.

Arrangement of the data in a proper manner to avoid the redundancy can fulfil the big data challenges.

Big data era brings new challenges to various aspects of GIS including data collection, management, processing, and visualization.It derives innovative solutions, Accesses vast information via survey and Delivers answer for any query.

REFERENCES

1. *Pete Warden, "BIG DATA GLOSSARY"*
2. *SherifSakr , Mohamed MedhatGaber – "LARGE SCALE AND BIG DATA, PROCESSING AND MANAGEMENT"*
3. *Jure Leskovec, AnandRajaraman, Jeffrey D. Ullman - "Mining of Massive Datasets"*
4. *www.wikipedia.org*