BIG DATA: A New Technology

Farah DeebaHasan
Student, M.Tech.(IT)
Amity School of Engineering and Technology
Amity University, Noida, U.P. India

Anshul Kumar Sharma
Student, M.Tech.(IT)
Amity School of Engineering and Technology
Amity University, Noida, U.P. India

Abhilasha Singh
Assistant Professor
Amity School of Engineering and Technology
Amity University, Noida, U.P. India

Abstract - This paper describes the concept of BIG Data. Its definition, when it first time used and came into the IT world. This paper characteristics of big data i.e. volume, velocity, variety have been described. Challenges and business issues have been studied. Since Big data is not a simple concept so its processing is also in many steps so this paper tries to describe these steps in simple language along with its difficulties at each step. At last, the software which are been currently used i.e. Hadoop, MongoDB areexplained in short.

IJOART

Keywords: business analyst, volume, velocity, variety.

I. INTRODUCTION

Big Data – A new word in the information world. The amount of data has exploded in the past few years because of new social behaviour, social media along with spreading software systems. Big Data: From the Business perspective focus on who, where, what, why and when it is required. Big Data: From the Technology Perspective deals with the some basics about Big Data open source technologies like Hadoop.

Since Data is very important in today's fast changing world, business executives need concise and needful information in less time to take important decision.Big Data is not just revolutionizing not just businesses or research, but also education system. Researches show that using information technology can reduce the cost of healthcare while improving its quality [1]. Similarly Big Data help in many areas like urban planning, environmental modelling, energy saving, computer security and many other fields.

Section II of this paper gives the basic definition about Big Data and explained about the characteristics of it. Section III describe that who exactly is in charge of big data earlier it was IT personal but now business analyst are.In Section IVchallenges and real business issues are being explored and presented. Section V describes about the phases of big data analysis. And lastly in section VI introduces the software in brief.

II. BIG DATA CONCEPT

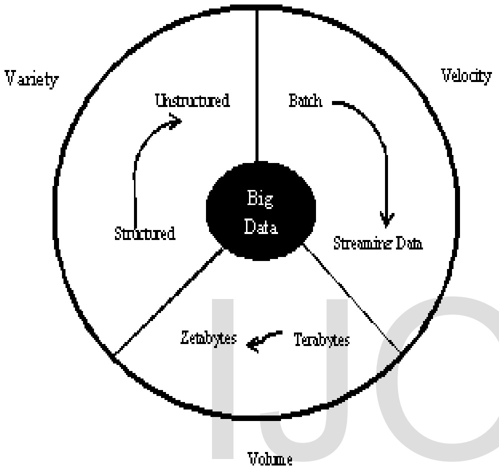
Big Data is a word that covers a wide range from technological aspect to business model. Big Data can be defined as:

"*Big Data "is a term encompassing the use of techniques to capture, process analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies" [2].*

Cost of Big Data lies within economic cost of storing data and processing these datasets. Cost of storage has been reduced along with amplification by cloud business model has lowered the upfront IT investment costs of all businesses. In short, we can say that data is saved in most raw form or in semi structured or unstructured format, so it is difficultto find out the required information from it. Till when the amount is less there is no major issue, but as the amount increases fetching desirable information from it is difficult. At this time Big Data concept is very helpful for business executives. So Big Data is not a new concept, but can be seen as moving target linked to a technology context[2].

Characteristics of Big Data: 3 V's of Big Data.

i) Volume of Data.
ii) Variety of Data.
iii) Velocity of Data.



IJOART

Figure 1 : Characteristics of Big Data

These 3 words together defines what Big Data is all about and how it provide a modern approach to deal with today's overwhelming amount of data and ability to control it along with taking action according to it. Below these words are being defined.

a) **Volume of Data:** The volume of data being generate and being stored is increasing day by day. Some statistics of volume of data being generated are as follows[3]:
   1) 2.7 Zeta bytes of data exist in the digital universe today.
   2) 235 Terabytes of data has been collected by the U.S. library of Congress in April 2011.
   3) Facebook stores, access, and analysis 30+ Petabytes of user generated data.
   4) More than 5 billion people are calling, texting, tweeting on mobile phone worldwide.
   5) Wal-Marthandles more than 1 million customer transactions every hour, which is imported into database estimated to contain more than 2.5 petabytes of data.

b) **Variety of Data:** With the increasing amount of volume new challenges have come into picture. These new challenges are that with advent of e-mail, social media,text data, smart devices, new technologies and so on . The data now is not only in traditional format but also in complex,semistructured or unstructured from all over internet.

   Variety can be better understood by this only that the sheer number of different types of data is very high due to different types of devices and sensors. With respect to Big Data bigger challenge is to join this entire different data set into one unified analytic.

c) **Velocity of Data:** It is an important about Big Data. It defines not only how fast data is arriving and storing but also how quickly data is being changing day by day. Therefore it is more suitable to say data to be in motion. Some statics about data velocity are as follows[3]:
   1) According to estimates, the volume of business data worldwide, across all companies, doubles every 1-2 year.
   2) YouTube users upload 48 hours of new video every minute of the day.
   3) Brands and organizations on Facebook receives 34,722+ likes every minutes of the day.
   4) Every person in the world having more than 215m high- resolution MRI scans a day.

   As these statics shows the velocity with which data is being generated and stored so if it is not taken into account then analytic results may be invalid. Especially in term of stock market or in telecom companies where data records only remains relevant for a day.

III. WHO'S INCHARGE OF BIG DATA?

Big Data despite industry hype,organisations have yet to develop, implement or execute a big data strategy. Leading organisations first implement the new technology while others wait for the success or failure of the initiatives. The survey found that 12% of organisations are currently implementing or executing a big data strategy while 71% of participants have yet to begin planning [5]. Most important reason is that they don't know enough about this new technology and can't understand the benefits of it.

As survey shows that only some organisations are implementing big data strategy so the question "How to make sense of this information?" is a major topic. Along with this a major challenge is who in-charge of data management strategy is? Earlier IT personnel were responsible for it but now as the trend changes new challenges have arisen so now business analyst and business managers are mainly responsible. Studies shows that single department or group are not the primary generator or consumer of data, it covers a wide spectrum. Therefore business analyst and manager who are responsible for the overall business success are in-charge of big data technology implementation , its usage along with extracting information from it.

IJOART

IV. Challenges and real Business issues with big data

Since the advent of the big data technology a lot of researches have been done, blogs, white papers have been written; experts have given their views, ideas in its advancement. Still today also there are lot of challenges in fully utilising this technology. Some of the issues are [3]:

1) 39% of marketers say that their data is collected "too infrequently or not real-time enough".
2) There are too few people with deep analytical skills to fill the demand of Big Data jobs .
3) Poor data can cost business 20%-35% of their operating revenue.

These issues are just a few. With big data a technical challenge is the size of data, along with velocity and variety. Theseword are used to define big data but they are to be taken special care and also data privacy and usability is to be considered. In order to extract information from such huge amount of data many distinct phases have to be followed and each of its phase present a challenge to analysts. For example, when it comes to merging different data format then a major problem is how to bring that entire different format to a same format for processing. Other challenge that come into account is that which data to be considered and which not to be to get better result.

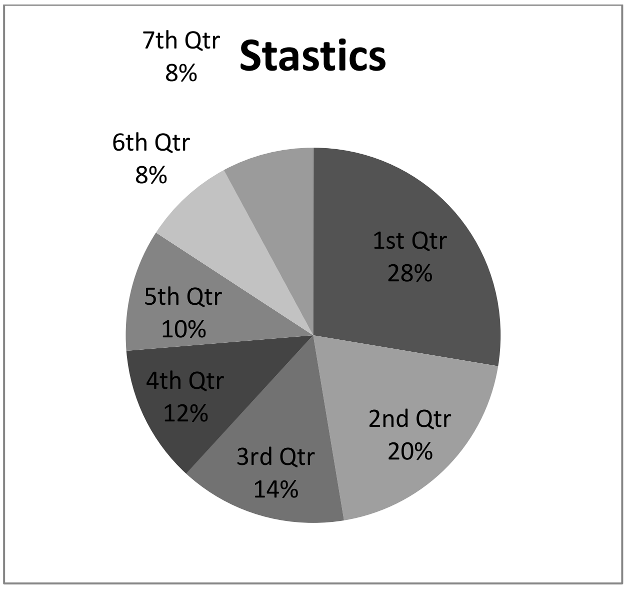


Figure 2 : Reason why organisation are not implementing Big Data for decision making

1stQtr.: Don't know enough about big data
2ndQtr.: Don't understand the benefits
3rdQtr.: No reason
4thQtr.: Lack of business support
5thQtr.: Poor data quality in current systems
6thQtr.: Lack of executive commitment
7ThQtr.: cost/ initial resources

V. Phases in Data Analysis

As till now it is clear that big data is not a simple technology but the integration of different technology in order to get fast reliable and time, cost saving system. So it also involves different phases which can be described as follows:

1) **Data acquisition and storing:** huge amount of data is generated every day; much data is of no use. So here a major issue is to make decision which data is to be kept and which not. One other thing to be considered is to produce metadata of this stored data.

2) **Information extraction and cleaning:** Extracting information from different sources is not an easy task. Data is saved in form of text, images, videos etc. Now in order to extract information all these data has to be change to a same format and unwanted data has to be removed. Therefore it extraction and cleaning is a technical challenge.

3) **Data integration, Aggregation and Representation:** After extraction data bringing it to same format so that it can be processed by the system. Aggregating it, removing all debris and bringing it to a representable format so that any user an understand it and use it to get what information is needed.

4) **Query processing and data Analysis:**Querying and processing method of big data are much different from other statistical analysis of same format. Noisiness, heterogeneity, inter-relation are challenges that are related to big data. Analysis is done by carrying out SQL query
But still problem with current processes that data is exported from database which means performing a non- SQL process and then bring it back to database.

5) **Interpretation: T**he result of analysis has to be interpreted by the decision maker, it involves considering all assumption made and retracing the analysis. In others words, we can say that supplementary information is also provided so as to explain how result was delivered and upon what input to help better understood by the decision makes.

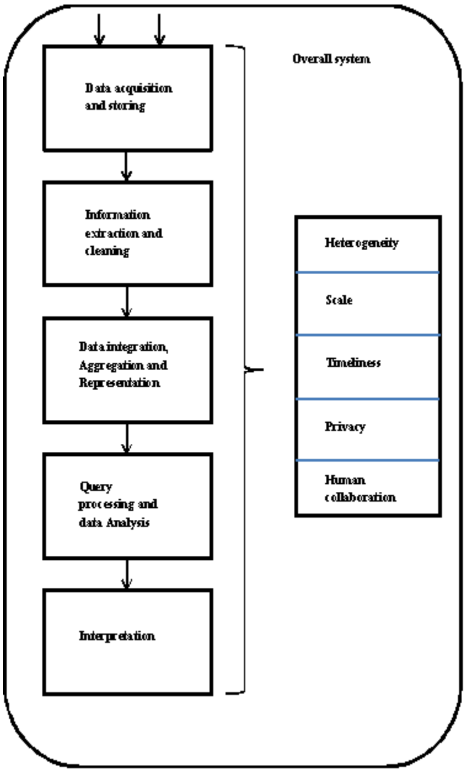

Data acquisition and storing
Information extraction and cleaning
Data integration, Aggregation and Representation
Query processing and data Analysis
Interpretation
Overall system
Heterogeneity
Scale
Timeliness
Privacy
Human collaboration

Figure 3 : overall system of Big Data

IJOART

VI. Big data software

Some software widely used are hadoop and MongoDB.

Hadoop: It is a top-level Apache project in the Apache software Foundation that's written in Java. It was first shipped as part of the IBM InfoSphere BigInsights platform. It is a open source project for massive data analysis. Therefore application written in Hadoop will be running on BigInsights.

Hadoop was inspired by Google File System (GFS) and MapReduce paradigm which use the concept of mapper and reducer for manipulating data stored across a cluster of server.

MongoDB :is a cross-platform document- oriented db system. 10gen first developed it in oct. 2007 as a component of planned platform as a service product. In 2009 open source model was developed.

It is currently adopted as a backend software by a no. of major websites and services including ebay, foursquare, The New York Times. Till today it is the most popular MySQL database system. Why MongoDB is used[7]:

1) It provides a simple way to evolve the db with application.
2) It combines structured and unstructured data in a single data store.
3) Native analytics makes it easy to deliver actionable insights in real-time.
4) Horizontal Scaling allows organization to support massive data volumes and high ingestion rate.

VII. Conclusion

At last, Big data is not any new technology but the combination on chained processing used in context to huge amount of data. Since world is not stable so also the data therefore a technology that can cope up with this changing and fast growing data was needed. Big data is the result of this need only. Hadoop, MongoDB, splunk are being widely adopted by leading organisation and it its hoped that in near future big data will combine with SAP, real time processing and other technology and till then small organisation will also be implementing it .

IJOART

References

[1] White paper “Challenges and Opportunities with Big Data”
[2] Big Data A New World of Opportunities Jun Hou, Lei Xu,”A Testing Tool for Composite Web Services based on data flow”,Sixth web information systems and applications conference,2009
[3] http://wikibon.org/blog/big-data-statistics/Boris Beizer, “software testing techniques”, International Thomas Computer Press,1990
[4] http://www.lavastorm.com/blog/post/taming-data-variety-and-volatility-is-key-for-big-data-analytics/A.Rembiszewski “Data flow coverage of object programs” Msc thesis,Institute of Computer Science,Warsaw university of technology,2009
[5] 2013 big datasurver research brief by SAS
[6] http://en.wikipedia.org/wiki/Big_data
[7] http://www.mongodb.com/use-cases/big-data

Copyright © 2014 SciResPub.