Big Data: A Trouble or A Real Solution?

Ms. Aakanksha Chopra[1] ,Ms. Suman Madan[2]

[1] **Assistant Professor (IT), Jagan Institute of Management Studies (JIMS),**
Sec-5, Rohini, New Delhi- 110085, Affiliated to GGSIPU, Dwarka Sec-16 New Delhi-110078.

[2] **Assistant Professor (IT), Jagan Institute of Management Studies (JIMS),**
Sec-5, Rohini, New Delhi- 110085, Affiliated to GGSIPU, Dwarka Sec-16 New Delhi-110078.

Abstract

Information today has gone from scarce to superabundant which brings immense new benefits but complementary big headache too. We are surrounded only data. From where this data has flooded we hardly have time to track that. Today every bit of data is important and hence stored. The actual problem is that more the data more accurate analysis and forecast is possible but the dark truth behind it is that the actual data is surrounded by elephantine amount of uncertain data. This paper focuses on the biggest problem called data deluge, need of big data, components of big data, business intelligence versus big data analysis, business concerns in respect to big data, how big data overcomes the problems faced by business, techniques for analyzing big data, types of big data analysis, complex challenges for an organization to shift towards big data.

Keywords- data deluge, big data, big data analytics, components, techniques.

1. Introduction

William Gibson once said that- "The future is here, but it's just not evenly distributed yet." According to *The Economist* [1]- In 2010 the digital universe was 1.2 zettabytes of data, in a decade the Digital Universe grew to 35 zettabytes, and in 2011 it shoot to 300 quadrillion files in which 90% of digital universe data is Unstructured- which is an alarming issue**.** In real world if we need to get the right answer then we must first have the right question to be asked, similarly to get the correct analytics from the data we must first have the real data rather than uncertain data

Alex Szalay, an astrophysicist at Johns Hopkins University, states that "The procreation of data is making them progressively unapproachable." He says that people must be trained not only the scientist or the computer professionals or the industry or government but including all of them who are contributing towards this never ending data creation must be trained in -How to make sense of all these data?

James Cortada of IBM said - "We are at a different period because of so much information." The information available is next to infinite which is getting difficult to manage and tackle. According to *John Easton, IBM Distinguished Engineer- Advanced Analytics Infrastructures* 80% of the data available by 2015 will be uncertain [2]-hence creating problem to handle (See figure 1). *Joe Hellerstein, a computer scientist at the University of California in Berkeley*, calls it "the industrial revolution of data." The effect is being felt everywhere, from business to science, from government to the arts. Scientists and computer engineers have brainstormed a new term for the phenomenon: "BIG DATA". The uncertain data is replicating at a speed which is almost ten times more as compared to the real data.

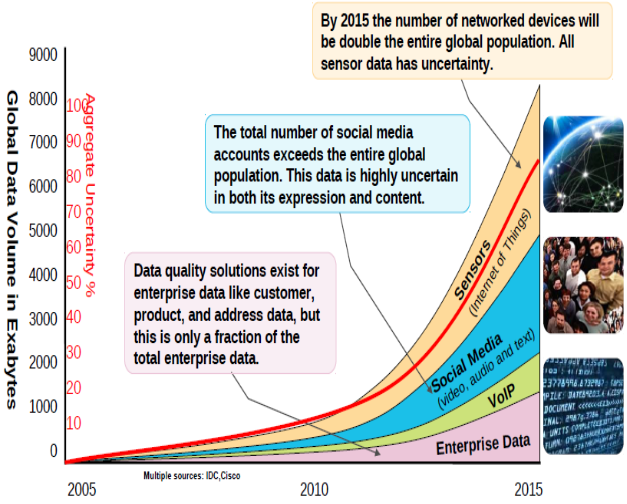

Figure 1: How Data Uncertainty is increasing (Source [2])

The term "BIG DATA" was given by *Roger Magoulas* from O'Reilly media in 2005[3]. He explained that due to its whopping size and complexness, wide range of data sets is almost becoming insoluble to handle and manage through traditional data management tools. Big data can be seen in Banking and business for Inventory Management, Customer Behaviour, Market Behaviour. Big data is also seen in life sciences for analyzing and advance research in Genome sequencing, clinical data and patient data areas.

Big data could also be seen in other areas like Astronomy [1,3] and Oceanography[3].

Taking few examples like Wal-Mart which is a retail giant, handles more than 1 million customers transaction every hour, fattening databases approximately to more than 2.5 petabytes which is equivalent to 167 times the America's Library of Congress [1]. World's most famous social networking website Facebook is a home to 40 billion photos. The Human Genome (which is a complete set of genetic information encoded as DNA sequencing of humans) involves analysing 3 billion base pairs- which took ten years in 2003 when it was first time done, but can now be unbelievably conquered in a one weeks' time span.

Above examples conclude the same phenomenon: that the universe embraces an indescribable mammoth extent of digital information which is clutching gigantic (infinite) more speedily. The amount of digital information accelerates tenfold every five years [1]. According to Moore's law, which IT sectors takes for granted, says that 'the processing power and storage capacity of computer chips double or their prices halves roughly every 18 months.'

Due to hasty outburst and combustion of data & the demand for digital collaboration everywhere, IT people know that the conventional data management techniques are no longer helpful in managing and utilizing their data, so they are moving towards finding advance solutions to secure their data. To have a grip over this exploded data we need to realize new opportunities i.e. we need to explore beyond the conventional sources of data like- Enterprise content, Social Data, Machine Data, Transactional and Application Data [5].

The key for IT with Big Data is to get past the hype & to learn more about the practical benefits, like finding exposed data, flagging malicious activity, finding illegitimate users, detecting frauds & identifying excessive access. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account.

Looking into Figure 2 we can very well imagine how this huge data has originated a big problem of storage for data. Despite the abundance of tools to capture, process and share all this information in recent years it has already exceeded the available storage space. In addition to this security of data and protecting the privacy is becoming callous because the information replicates and is exchanged abruptly across the whole world. Due to this the limitations of existing Data Analytics Architecture is that A slender 10% of the ~ 2 petabytes of data is available for Business Intelligence but the stupefying fact is that 90% of ~2 petabytes of data is archived and hence it is neither opened nor analyzed properly resulting in premature data death[7].

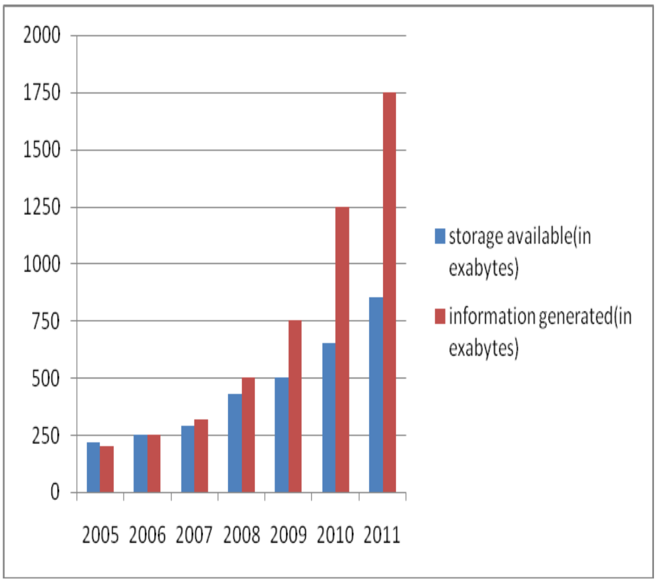

Figure 2: Relation between storage available and Information created

2. The Real Problem: Data Deluge

John Naisbitt quoted that - "We have for the first time an economy based on a key resource [Information] that is not only renewable, but self- generating. Running out of it is not a problem but drowning into it is." Data deluge is very hazardous as there is a lot of risk associated with it; as there is data and only data everywhere. For example there are many cases in which the data goes missing, stolen, or breached resulting in identity theft, frauds, illegitimate access of data.

The most suitable approach to overcome the drawback of data deluge is to understand what big data is actually. *Clive Humb and Dunn Humby says that* "Data is the new oil. Data is just like crude. It's valuable, but if unrefined it cannot really be used." Hence, Big data is not about what we have? It's about what we are going to do with what we actually know. Therefore, firstly, we need to create value from the data and secondly, we need to create pattern from the data to create an insight.

In order to create value from the big data the data should be appropriately utilized by shedding greater transparency in many fields. For this three factors should has to be considered-

2.1 Users Control

It's time to give the users an upper hand on the control and access over the information held about them, also including with whom it is shared with, with various customizations allowed.

2.2 Taking Security Issues seriously

Organizations should start discussing and disclosing their hidden security policies now, in order to control security breaches within the organization.

2.3Yearly examining Security points

Making the security policies to be audited on yearly basis will help organizations to find out the loop holes and rectify them with more security measures for future furthermore it will help the organizations to keep their security measures up to date and help in taking a proactive step for any illegitimate entry.

Consecutively following above three factors the organizations can gain higher market initiative as compared to those not following them. Users will get more control over their data, will be secured and have freedom from complicated regulations that could starve them from innovations if, high level of transparency is provided to them.

3. Why Big Data Should Matter To You?

The actual problem is not acquisition of hefty data but the undeniable need of this data. In simple words we can say large data is directly proportional to higher accurate analyses [10]. Higher accurate analyses may result in more dauntless decision making. And of course a smarter decision making can lead to an improved operational productivity, reduced costs, reduced time [11] and reduced risk. By entangling big data and high-powered analytics it is possible to-

3.1 Extracting the actual reason behind defects, flaws, failures in the real- time to save losses in future.

3.2 Analyze millions of Stock keeping units to determine prices that boost profit and clear all the stock.

3.3 Fully Optimize routes followed by package delivery vehicles while they are on the road.

3.4 Customers who have higher priority (more important) should be swiftly recognized.

3.5 Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.

3.6 Have a system those analysis current and past purchases of various customers to generate attractive vouchers for them.

3.7 Reanalyze all the risk factors or points within minutes.

3.8 Use Clickstream analysis and data mining to unmask fraud nature & behaviour.

4. Understanding Big Data

4.1 Components of Big Data

Big data analytics has the capacity to process any variety, volume and velocity of information and to derive an insight into data [5]. Big data is a way to find insights in evolving types of data and information to find out answers to the questions which where earlier unanswerable or unreachable. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data “BIG DATA” to discover patterns and important information. Big data analytics also help in discovering and extracting data which is helpful in making future business decisions in addition to helping understanding the information within data. Big data analysts basically want the refined data actually useful that comes from analyzing the data. BIG DATA analytics is about joining trusted, internal information with new data types to create value bringing new source of unstructured info to existing core data to create insight. What is this New data we are talking about - It’s the information that is already there but we never used it. Like Email, Blog, Stock Market, Sensors, Mobile Phone GPS etc.

The four V’s of Big Data are-

4.1.1 Volume

Increase in data volume is caused by many factors. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data. Enterprises are flooded with ever-growing data of all types, easily accumulated to terabytes - even petabytes -of information. The handle volume Big Data turns 12 terabytes of data (tweets) created each day into improved product sentiment analysis. Secondly, Converts 350 billion annual meter reading to better predict power consumptions etc.

4.1.2 Velocity

Data is streaming in at unprecedented speed and must be dealt with in a timely manner. Processing should be fast and quick for time- sensitive processes such as catching frauds. 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. Big data must be used to firstly, Scrutinize 5 million trade events created each day to identify potential frauds. Secondly, analyze 500 million daily call detail records in real-time to predict customer churn faster.

4.1.3Variety

Data today 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. Managing, merging and governing different varieties of data is something many organizations still grapple with. Big data must Firstly, monitor 100's of live feeds from surveillance cameras to target points of interest. Secondly, exploit the 80% data growth in images, videos and documents to improve customer satisfaction.

4.1.4Veracity

In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

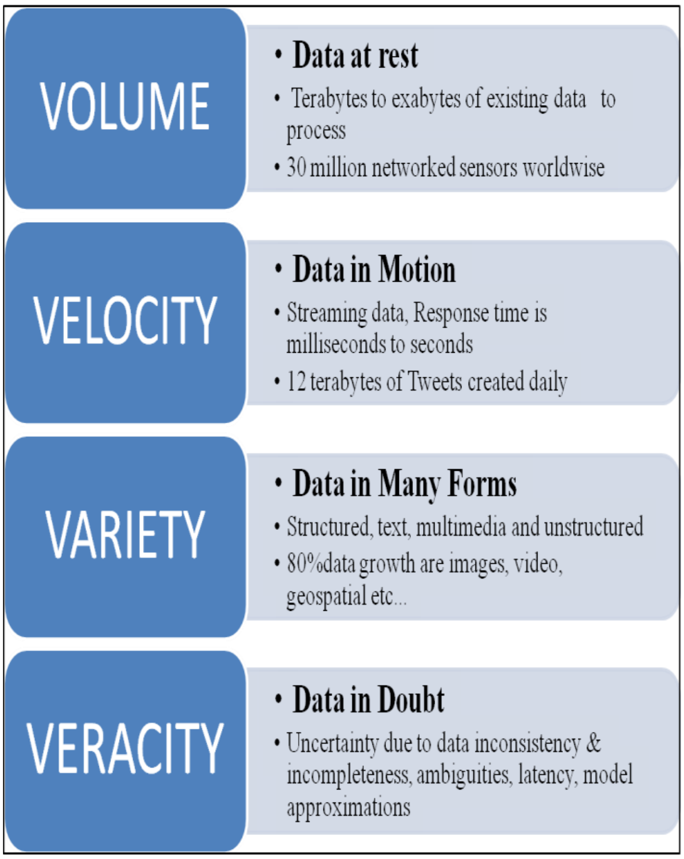
VOLUME
• Data at rest
• Terabytes to exabytes of existing data to process
• 30 million networked sensors worldwise
VELOCITY
• Data in Motion
• Streaming data, Response time is milliseconds to seconds
• 12 terabytes of Tweets created daily
VARIETY
• Data in Many Forms
• Structured, text, multimedia and unstructured
• 80%data growth are images, video, geospatial etc...
VERACITY
• Data in Doubt
• Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, model approximations

Figure 3: Four Dimensions of Big Data

Matt Eastwood, IDC [2] says "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of wide variety of data, by enabling high velocity capture, discovery and/or analysis."

In today's scenario the largest challenge within organizations with respect to volume, variety, velocity and veracity is the need of the hour. Challenge with volume is – since the volume of data is just increasing the difficulty in the roadmap is that what if it becomes faster as compared to what database can handle? With Variety of data available be it Structured, Semi- structured, or unstructured data what is the correlation from multiple sources and/or signals, videos, audios or other non-relational data types? Velocity is data in motion, till now response time of data was from millisecond to seconds, but the bigger challenge is to stream data and reduce the response time under(much lower than) milliseconds. Veracity is ambiguity of data, the incompleteness, and the inconsistency in data. The organizational challenge is to establish and build trusted information so that users can trust the application [5].

SAS, introduced additional dimensions **Complexity** while thinking about big data. **Complexity-** Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

4.2 Misconception and Truth

A common misconception is when customers confuse the term Big Data with having to deal with lot of data. But the Truth is that volume is clearly a part of big data solution but Big Data is more about unlocking the potential of Structured & Unconstructed information, inside & potentially outside of our firewall & doing it in right time.

5. Difference Between Business Intelligence & Big Data

Business Intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purpose. A few years ago such technologies today commonly known as Business Intelligence were available to only to the world's biggest companies. But as the price of computing and storage has fallen and the software systems have got better and cheaper, the technology has moved into the main stream. Companies are collecting more data than ever before. In the past they were kept in different systems that were unable to talk to each other, such as finance, human resources or customer management. Now the systems are being linked, and companies are using data mining techniques to get a complete picture of their operations- “a single version of the truth” as the industry likes to call it. That allows firms to operate more efficiently, pick out trends and improve their forecasting [1].”

There is confusion b/w Business Intelligence & Big Data- data is only truly in relation to its Volume, Velocity, Variety and Veracity. A Big Data solution has to be able to cope with all 4 of these. Traditional Business Intelligence & Data Warehouse Solutions are not engineered for this type of dynamic & Unstructured Data [6].

PARAMETER	BUSINESS INTELLIGENCE	BIG DATA ANALYTICS
VARIETY	Structured	Structured, unstructured and semi-structured but Mostly semi-structured
WORKLOAD	Repetitive	Ad-hoc and Experimental
VOLUMES	Generally GBs to 10s of TBs	10s of TB to 100s of PBs
SOURCES	Operational	External and Operational

Table 1: Big data is different from Business intelligence

6. Business Concerns

Initially, the fears associated with business success were were what happened and why it happened? These doubts were eradicated by applying data mining techniques on the data collected from different reports, dashboards etc. With span of time these tensions shifted to why it is happening? And the technology gradually changes from data mining to real-time data mining technique which is quick and spontaneous. With the growth in technology, the business concerns moved to what is likely to happen in future? Today, data modelling and forecasting are used to determine future possibilities. Contradictory to traditional thinking those organizations evolved from descriptive analytics to predictive to prescriptive analytics, but today, all kinds of analytics needs to be done in right mix to enable the holistic contrivance of insights.

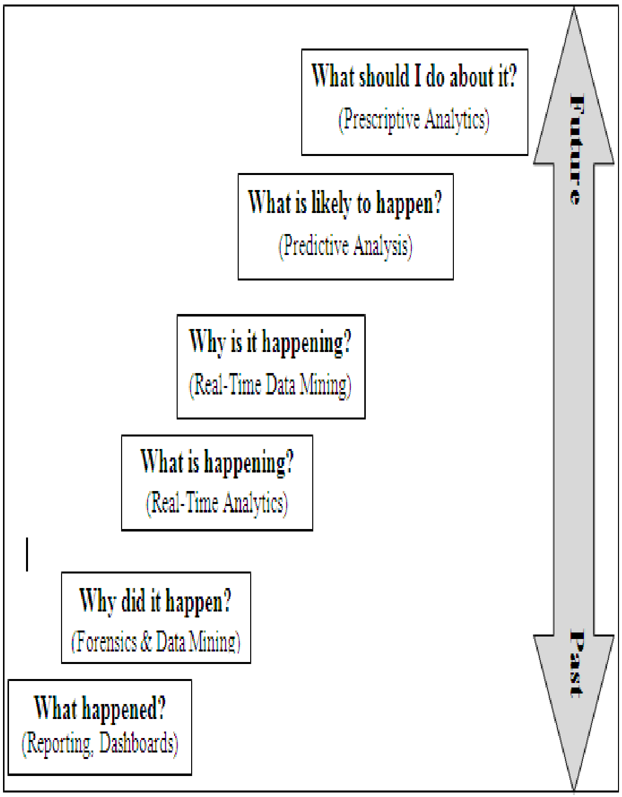

Figure 4: Changing business concerns

7. Advantages Of Big Data To Business

Big Data has gained a top spot on the agenda of business leaders for the real value it has began to create. Today the technologies and skill used to leverage by data for business purposes have reached a tipping point. New types of data, supported by better tools to leverage it enable companies to find financial & competitive benefits from their data. Big data will impact every aspect of the business- it will know everything about the customers, it will run Zero-latency operations, Innovate new products at speed and scale, Instant Awareness of frauds and risk, Exploit instrumental Assets [5]. Figure 5 depicts various financial services, healthcare, Government, Retail, Manufacturing; Web/Social/Mobile sectors are embracing big data.

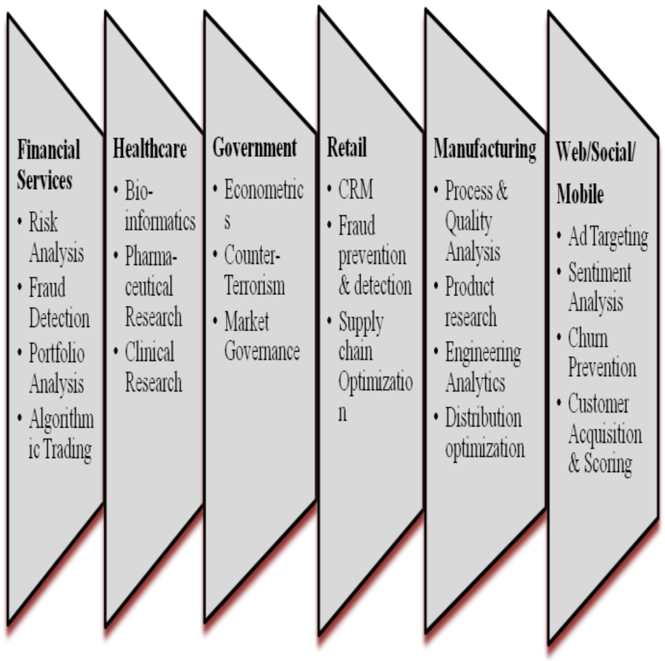

Figure 5: Various opportunities with industries to adapt big data

8.Types of Big Data Analysis

There are many different types of analysis that can be done in order to retrieve information from big data. Each type of analysis will have a different impact or result. Which type of data mining technique you should use really depends on the type of business problem that you are trying to solve. Different analyses will deliver different outcomes and thus provide different insights. One of the common ways to recover valuable insights is via the process of data mining. Data mining is a buzzword that often is used to describe the entire range of big data analytics, including collection, extraction, analysis and statistics. Types of big data analysis are explained as follows-

8.1 Anomaly or Outlier detection

Many businesses require real –time outlier detection i.e. searching for data items in a dataset that do not match a projected pattern or expected behaviour. These outliers also called Anomalies, exceptions, surprises or contaminants and they often provide critical and actionable information. An outlier is an object that deviates significantly from the general average within a dataset or a combination of data. Thus, outliers should be carefully designed to deal with the complexity, speed, volume and variety involved in big data. They need to be accurate and minimize false positives or false negatives due to the cost of analyzing each anomaly. Anomaly detection is used to detect fraud or risks within critical systems, fraudulent actions, flawed procedures or areas and they have all the characteristics to be of interest to an analyst, who can further analyze the anomalies to find out what's really going on.

8.2Association rule learning

Association mining means finding frequent patterns, associations, correlations or causal structures among sets of items or objects in transaction databases, relational databases and other information repositories. Thus, Association rule learning enables the discovery of interesting relations, uncovers hidden patterns between different variables in large databases and the co-occurrences of different variables that appear with the greatest frequencies. Association rule learning is often used in the retail industry when finding patterns in point-of-sales data. Association rule learning is being used to help: place products in better proximity to each other in order to increase sales, extract information about visitors to websites from web server logs, analyze biological data to uncover new relationships, monitor system logs to detect intruders and malicious activity**.**

8.3Clustering analysis

The larger data sets before use needs to be categorized in an effective, fast and automatic manner. One of the most commonly used systems is a series of statistical techniques called Cluster Analysis (CA), which is able to group data sets according to their "similarity". Thus, clustering analysis is the process of identifying data sets that are similar to each other to understand the differences as well as the similarities within the data. Clusters have certain traits in common that can be used to improve targeting algorithms.

8.4Classification analysis

Big data problems are often complex to analyze and solve. The sheer volume, velocity, and variety of the data make it difficult to extract information and business insight. Thus this technique classifies the big data problem according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store i.e.

finding important and relevant information about data, and metadata – data about data. The classification analysis helps identifying to which of a set of categories different types of data belong. Classification analysis is closely linked to cluster analysis as the classification can be used to cluster data. Categorizing big data problems by type like web & social data, machine & human generated data etc., makes it simpler to see the characteristics of each kind of data. These characteristics can help us understand how the data is acquired, how it is processed into the appropriate format, and how frequently new data becomes available. Data from different sources has different characteristics; for example, social media data can have video, images, and unstructured text such as blog posts, coming in continuously.

8.5Regression analysis

At a basic level, regression analysis involves manipulating some independent variable to see how it influences a dependent variable. It describes how the value of a dependent variable changes when the independent variable is varied. Independent variables can be affected by each other but it does not mean that this dependency is both ways as is the case with correlation analysis. A regression analysis can show that one variable is dependent on another but not vice-versa. Regression analysis is used to determine different levels of customer satisfactions and how they affect customer loyalty and how service levels can be affected by for example the weather. It works best with continuous quantitative data like weight, speed or age.

9. Techniques for Analyzing Big Data

Big data analytics is a process of elicitation, organizing and analyzing “BIG DATA” to form patterns and valuable information. Big data analytics also help in discovering and extracting data which is helpful in making future business. Big data analysis is actually making “sense” out of large volumes of varied data that is lacking in unrefined or raw data. Big data analyst further refines the analyzed data. Big data analytics is a juncture between trusted, internal information with new data types to create value; bringing new source of unstructured information to existing core data in order to create insight.

There are several new issues you should consider as you embark on this new type of analysis:

9.1 Discovery

The biggest misconception about the data is what is data all about? In many situations you don’t know what data you have and how variant data sets are related to each other? This must be sorted out by the process of probation, exploration and discovery.

9.2 Iteration

Since the exploration and discovery process in not a onetime process and it keeps on going hence, the actual relationships cannot be well know in advance, uncovering insight is sometimes an iterative process as you keep on finding the answers that you seek. The process of iteration is that sometimes it leads to a goldmine and sometimes it leads to a dead end. We can conclude that experimentation is a part of iteration process only.

9.3 Flexible Capacity

Big data analysis is iterative in nature which keeps on extracting new relationships between the data sets while discovery therefore, you can say that one needs a flexible time, value and resources to utilize and make problems solvable.

9.4 Mining and Predicting

Big data analysis is not black and white. You cannot always predict the relationship between different data elements. As you do data mining, the data to discover patterns and relationships, predictive analytics can yield the insights that you seek.

9.5 Decision Management

If you are considering the transaction volume and velocity while using big data analytics, then you need to consider how to automate and optimize the implementation of the actions taken to make operational decisions.

10. Complex Challenges For Organizations To Shift To Big Data

Apart from confusion around the definition, the complexity of big data is also stopping organisations from already realising its benefits and taking a back foot from adopting the Big data analytics.

10.1Volume

According to a study by IBM by 2015, 80% of data will be uncertain data [2]. Portal's Beeston says: "Since a ponderous quantity of information is unstructured, and an equal proportion of customers have no unstructured analytical capability, the volume concern is a big issue, made worse by the variety of data types that are emerging."

10.2 Storage and Protection

How, where will be this large data stored on server and what measure would be taken to protect this mammoth of data.

10.3 Swiftly changing data

Customers are also horrified at the speed at which information changes. Analyzing vast amount of data, or indeed achieving golden insight, is pointless unless the output is presented at the right time for the user.

10.4 Backup and Restoring

How much backup is to be taken and after how much time, what chunks of data to be restored or whole data is to be restored, how many copies of a data file is to be created.

10.5 Organizing and Cataloguing backup data

Organizing the backup data is another point to be concentrated upon; how will the backup data be catalogued how to create metadata of backup data.

10.6Skills

There is a need of data scientist and analysis specialist managers, also there has to be a flexible approach from totally hiring from outside to in-house skills. *Varonis's Sobers* agrees that customer's biggest worries are skills. Biggest concern is staffing – 'Do I have to hire data scientists and programmers to implement a big data solution?'

10.7 Value

The key is exploration, finding the value and building the skills at an incremental cost.

10.8 Cost

How will be the cost be kept low in keeping in mind that all the crucial data is on time available when they are needed?

There are various risks also involved with big data technology which is further stopping organizations to adapt it. Firstly, since it is a new technology for many organizations, and also since there is lack of skilled people for understanding big data analytics it may introduce new vulnerabilities. Secondly, since the implementation of big data is open source it may lead to an open entry for some unwanted access. Thirdly, there might be dearth of control over user authentication and access to data from multiple locations, as users logging in from multiple locations might become difficult to trace. Fourthly, there is a significant chance for malignant data input and insufficient data validation [9].

11. References

1. https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf
2. http://www.thebigdatainsightgroup.com/site/system/files/private_1
3. http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf
4. Dilpreet Singh and Chandan K. Reddy, "A Survey on Platforms for Big Data Analytics", Journal of Big Data, 1:1, 8, 2014. http://dmkd.cs.wayne.edu/Papers/JBD14.pdf
5. https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Calgary_Keynote_%20David_%20Corrigan%20-%20v1/$file/Calgary_Keynote_%20David_%20Corrigan%20-%20v1.pdf
6. http://www.snia.org/sites/default/education/tutorials/2012/fall/big_data/RobPeglar_Introduction_to_Analytics_Big_Data_Hadoop.pdf
7. http://www.slideshare.net/EdurekaIN/learn-big-data-hadoop
8. http://data-informed.com/manage-big-datas-big-security-challenges/

9. http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view#

10. http://www.sas.com/en_in/insights/big-data/what-is-big-data.html

11. http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf [Filename- Big data- big companies(Paper)] http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

12. Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014,Article ID 712826, 18 pages. http://dx.doi.org/10.1155/2014/712826. http://www.researchgate.net/publication/256082290_Addressing_Big_Data_Issues_in_Scientific_Data_Infrastructure

13. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, Volume 275, Issue null, Pages 314-347 C.1L. Philip Chen, Chun-Yang Zhang, http://www.sciencedirect.com/science/article/pii/S0020025514000346

14. Karthik Kambatla, Giorgos Kollias, Vipin Kumar, Ananth Grama, Trends in big data analytics, Journal of Parallel and Distributed Computing, Volume 74, Issue 7, July 2014, Pages 2561–2573, http://www.sciencedirect.com/science/article/pii/S

0743731514000057