Big Data Analytics: Extracting Value from Chaos, Challenges and Opportunities

Sandhya Rai[1] and Anil Kumar Pundir[2]

[1]IILM Graduate School of Management, 16 Knowledge Park – I, Greater Noida – 201306, U.P., India, Email: raisandhya@gmail.com
[2]Guru Jambheshwar University of Science and Technology, Hisar - 125001, Haryana, India, Email: anil.pundeer@gmail.com

Abstract

Big data has become the life blood of the organizations. The organizations are able to understand that if they are able to capture all the data that is streams into their businesses, then they can apply insight to get valuable information. The thought of data creating value is not new, business have always wanted to derive insight from data for making real time, fact based decisions. But the speed with which the data is generated and the variety in which it is available is tremendous. The aim of this paper is to understand the concept of big data and challenges and opportunities associated with the same. The paper also discus in details the steps involved in big data analytics and the relevance of each of these stages.

Key Words: *Big data, big data analytics, data preparation, data visualization, data discovery, data scientist, IoT (Internet of Things), cloud, software.*

Introduction:

Big data has become the life blood of the organizations. The organizations are able to understand that if they are able to capture all the data that is streams into their businesses, then they can apply insight to get valuable information. The thought of data creating value is not new, business have always wanted to derive insight from data for making real time, fact based decisions. Big data can be defined as the dynamic, large and disparate volumes of data being created by people, tools and machines. It includes data gathered from day to day operations of the organizations, data from internet enabled devices like tablets, smart phones, information gathered from social media, voice and video recordings, call details, network data, machine data. These data may be available in the numeric, text, graphics, JPG, MP3, MP4 format. It is usually characterized by the four "V's":

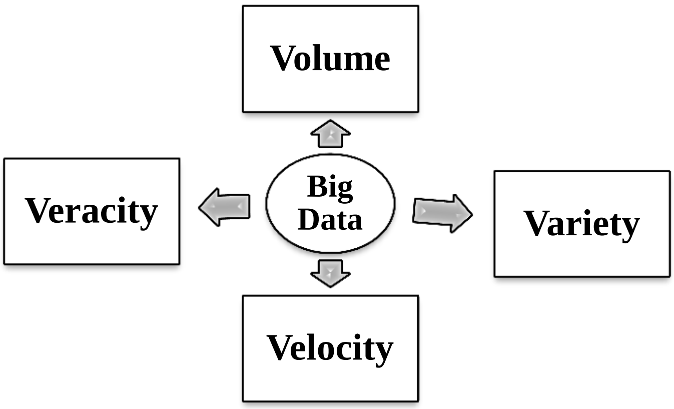


Fig 1: Big Data Characteristics

Volume: The volume of data being stored today is exploding. In the year 2000, 800,000 petabytes of data was stored in the world that is expected to become 163 Zetabytes by 2025 (Forbes APR 13, 2017**).**

Variety: It not only the volume of data that is huge but the data is available in different varieties. With the exploration of sensor and smart devices, and social collaborations, the data is not only available in structured form but in semi-structured and unstructured form also. It comes from different sources and is being created by machines as well as people. It includes data from web pages, web logs, search indexes, social media, e-mail, document and so on.

Velocity: As if the variety and the volume were not sufficient, these data are generated at extremely fast speed and in a continuous manner hence it never stops. The traditional system cannot handle this pace of data and at times this data has a very short shelf life and hence should be analyzed very quickly to get an edge over competitors or for the identification of trends, problems or opportunities.

Veracity: As big data is generated using various sources, one needs to test its trustworthiness. Veracity refers to the trustworthiness of the data. The data that is generated can have lots of noises, abnormalities and biasness that can hamper the quality of the output. Hence quality of the data is another characteristic of big data. The data used should be relevant to the problem under investigation.

Big Data Analytics

The concept of using data to generate value is not a new things, businesses have been using them for years, what's new is, the volume of data accumulating at a fast rate and in various form i.e. Big data. This big data that is available in its natural form is difficult to process because of the above mentioned characteristics and hence needs to be refined further in some structured form. As per the research published by Forbes in Mach 2016 (Gil Press, 2016) the most time consuming task in big data analysis is the cleaning of the data. Data scientist spends almost 60% of their time on cleaning and organizing data whereas almost 20% of the time is spent on collecting them.

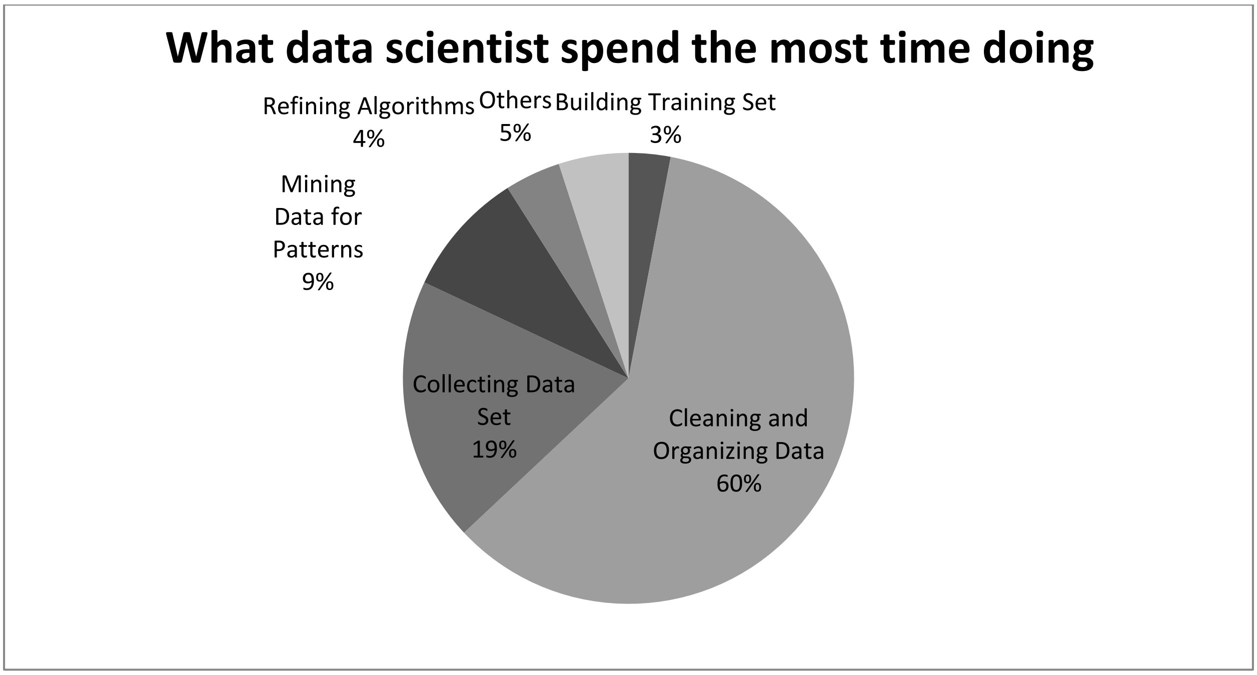


Fig 2: What data Scientist spend the most time doing
(Source: https://blogs-images.forbes.com/gilpress/files/2016/03/Time-1200x511.jpg?width=960)

Hence on an average 80% of the time is spend on preparing and managing the data for analysis and only 20% of the time is utilize on real analysis and interpretation. This makes the big data analysis further more challenging.

It is because of these challenges, the analysis of big data follows a step by step methodology to organize analyze and repurposing the data. The major steps of the big data analysis lifecycle are as follows:

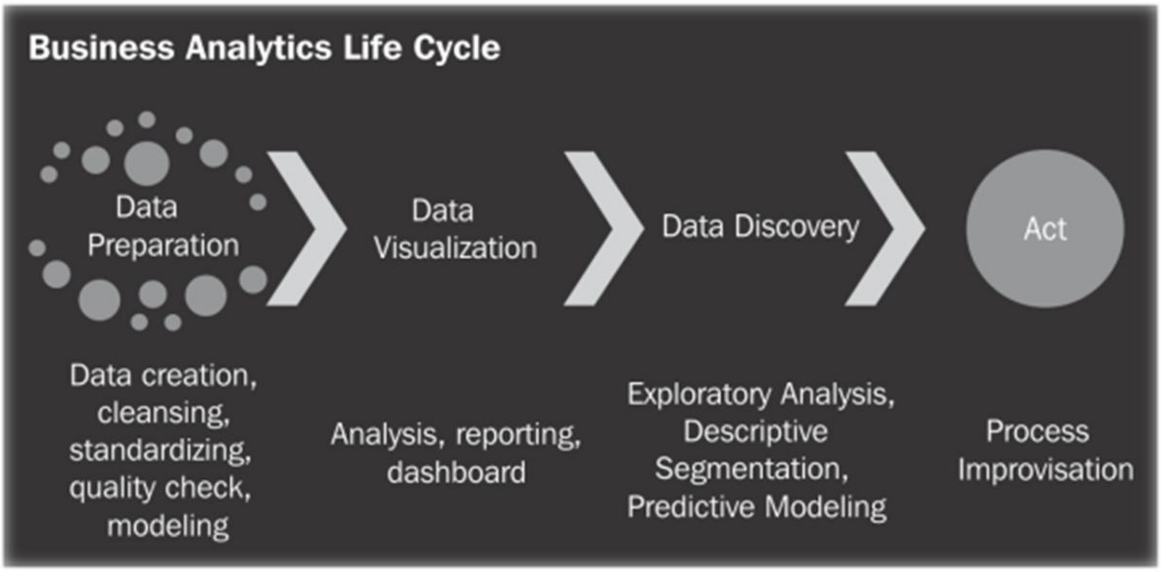

Business Analytics Life Cycle
Data Preparation
Data Visualization
Data Discovery
Act
Data creation, cleansing, standardizing, quality check, modeling
Analysis, reporting, dashboard
Exploratory Analysis, Descriptive Segmentation, Predictive Modeling
Process Improvisation

Fig 3: Business Analytics Life Cycle
(Source: Pentaho for Big Data Analytics, Manoj R Patil, Feris Thia, November 2013, Packt publishing)

Data Preparation: The first stage of the data analytics lifecycle is data preparation. At this stage, data is gathered from various sources and the entire data is then bought to a common platform. Here the data is also checked for the quality and various noises or unwanted or corrupt information are removed and the data is cleaned. The data that is subjected to cleaning may come as a collection of files from various data providers or may require API integration such as with Twitter or Instagram. The data classified as corrupt includes record with missing values or incomplete information or an invalid data type. There is also a possibility that the data that has been discarded in one analysis have value in the other, hence the original data set is always stored as a verbatim copy.

Thus the major purpose of data preparation (sisense.com**,** 2017) includes:

- Management of unstructured or inconsistent data
- Combining data from multiple sources
- Managing unstructured data like image, text, PDF data etc.
- Reporting on data that was entered manually

It is the structure of the filtered data that helps in determining the analytical technique to be used. The analysis can be exploratory or descriptive depending upon the purpose of analysis. It can be descriptive followed by exploratory too.

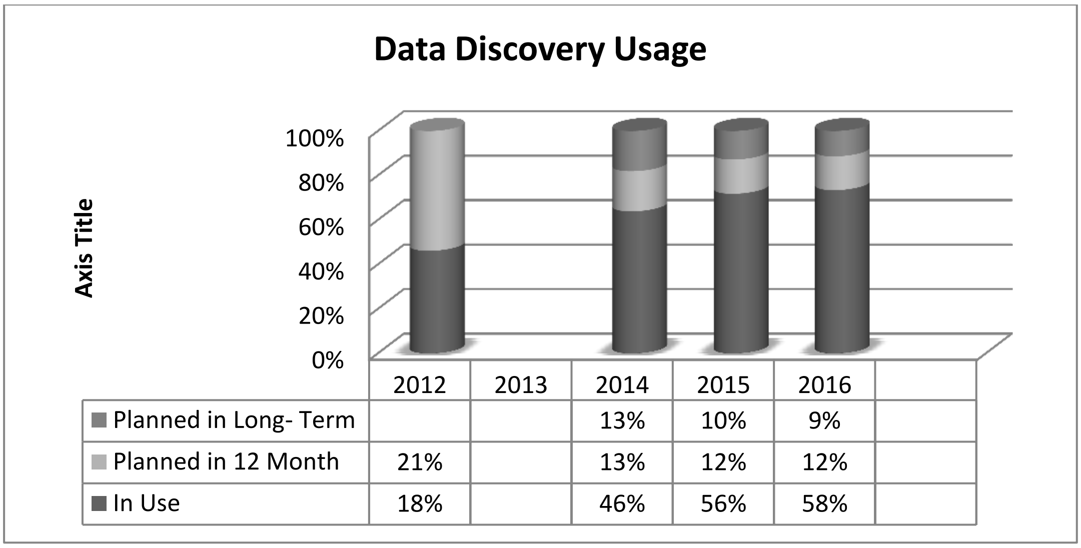
Data Visualization: Once the data is ready for analysis, the next stage is to identify the pattern in the data. The data can be presented in the form of graphical presentation, pictorial representation for the identification of the pattern. The analysis can be further drilled down to charts and graphs and one can also experiment with different scenario by making adjustment in the variables. Data visualization can also help in identifying areas that requires consideration and improvement and thus helps in placing the product or service in a better way and thus predicting their sales. According to Bernard Marr, 2017, seven best data visualization tools are:

- **Tableau**
  With a customer base of more than 57,000 across industries, Tableau is one of the most popular tools for data visualization. Its popularity is attributed to its simplicity of usage, speed of deployment and ability to produce interactive visualization far beyond the one produced by general business intelligence solutions. It is well suited for handling large volume of data set used in machine learning applications, artificial intelligence and other big data applications. Its integration with SAP, Hadoop, My SQL, Amazon AWS and Teradata had made it a very popular tool.
- **Qlikview**
  With a customer base of 40,000 across hundred countries, another very popular data visualization tool is Qlik with their Qlikview. It is giving a tough competition to tableau. Most of its user considered its highly customized setup and wide feature range as a key advantage. It not only offers data visualization capabilities but also offers analytics, BI and enterprise report capabilities. It is very popular along with Qliksense, a data exploration and discovery by the same company.
- **FusionCharts**
  Fusion Chart is a JavaScript-based charting and visualization package and has the ability to produce 90 different chart types. It is also capable of integrating with different types of platforms and frameworks. It also gives user the flexibility to choose visualization form from a range of “live” example templates; users just have to put their own data sources as needed and can get the required visualization for the same. It is paid software and requires a license for commercial use.
- **Datawrapper**
  Datawrapper is another popular tool of data visualization gaining popularity among the organization that are in frequent need of presenting various types of chart and statistical data. Its interface enables the users to upload csv files and create charts and maps the results into reports.
- **Highcharts**
  This is also paid software and like FusionChart, requires a license for the commercial usage however for the non-commercial and for the personal usage, free version is also available. It provides a fast and flexible solution and be rolled out with little data visualization training to the specialist. .
- **Plotly**
  Another popular tool for data visualization is Plotly. It is available both in commercial and non-commercial package where the non-commercial version is free. It is popular because of its ability to integrate with analytics languages like R, Python and Matlab.
- **Sisense**

Sisense is another popular data visualization tool. It provides drag and drop feature that allows the user to create interactive charts and dashboard. Not only this, it also enables the users to collect data from multiple sources at one repository, which can be later accessed for various analyses.

Data Discovery: Data discovery is the process of detecting patterns and outliers in the data by applying analytics. The analytics used may be confirmatory or exploratory. The analysis wherein the phenomenon being investigated is proposed beforehand is called confirmatory analysis. In such cases a hypothesis is proposed and the analyst tries to disprove it. This way they try to find out the answer to specific question. Where as in case of exploratory data analysis, there are no presumptions; data is analyzed to understand the cause of the phenomenon. This method helps in identifying patterns and anomalies in the data. Depending upon the purpose of the analysis, the kind of analysis is determined. This stage can be as simple as querying a data set to compute an aggregation of comparison or it can be as complex as combining complex statistical tools and data mining to discover patterns and depict relationship between variables (Thomas, et al., 2016).

According to the report published by BI-Survey.com (2017) the number of user, using data discovery have significantly increase over the years. As compared to 18% users suing data discovery in the year 2012, there are 58% user in 2016. Not only this, there are almost 21% new users who are planning to use data discover in future, that will lead to almost 80% users



	2012	2013	2014	2015	2016	
■ Planned in Long- Term			13%	10%	9%	
■ Planned in 12 Month	21%		13%	12%	12%	
■ In Use	18%		46%	56%	58%	

Fig 4: Data Discovery Usage
(Source: https://bi-survey.com/data-discovery)

The major reason for the popularity of data discovery includes:

It is expected to provide high value for innovation and success and hence it has suddenly become popular. Not only have this, the availability of cloud based services made it easier to implement. These services have disrupted the usage of data discovery by providing services having the attributes of:

Ease of Usage: These days there are so many tools that are available capable of performing data discovery without much of coding and statistical knowledge and hence any one can use them with little training.

Agile and flexible: Because of the availability of cloud based services, the companies can opt for these services without much reliance on their IT and at the same time not much technical setup is required.

Ease of data Handling: Now a days a number of technologies are available that helps in handling large volume and

Act: Depending upon the results obtained at data visualization and data discovery stage, the decision for the problem under investigation is taken.

Challenges of Big data

Though big data has become very popular, it has a long way to go. The major challenges faced by big data include:

Data storage

The pace with which new data is added every microsecond is very fast. Data is getting constantly created in various forms and across various platforms, it is very difficult to store this data at one place and hence data storage is becoming a challenge for the companies. According to IDC it is estimated that the amount of data stored in the world's IT system is getting double in every two years and by 2020, this data will be enough to fill a stack of tablet to reach from earth to moon 6.6 times.

Recruiting and Retaining Talent

Data analysis plays a major role in making sense out of the huge pool of data and this activity is carried out by Data Analyst and Data Scientist. The shortages of quality data scientist and analyst have made a demand supply gap in the job market and companies witness frequent movement of these individuals. This has created a challenge for the organizations. Many organizations are spending a lot on training their workforce to perform data scientist task.

Quality of Data

The entire big data analysis is based on the data available; hence it is very important to make available unbiased and accurate data. As discussed earlier almost 80% of the time is spent on making the data suitable for usage, this stage proves to be very expensive for the organizations. As most of these data is unstructured, combining it to a data warehouse create problems like data inconsistency, logic conflicts, duplicity of data and missing data. All these possess a challenge to the quality of data.

Security and privacy of the data

Security is another challenge for the companies as these places can always be attacked by hackers. Also the use of open software and tools for the data analysis make these data more susceptible to theft. Hence there is challenge to keep the data secure.

Various Sources of Data

The data that is being used in the big data analysis is use to come from various sources. It includes external sources like social media, competitor data, data from financial institutions and internal sources like finance, marketing and operations department. It is not only the sources are different, the types of data is also different. The data can be in Excel, csv, text, image, pdf, video or audio form; all these are required to be stored and analyzed, which is a challenge.

Opportunities in Big Data

In-spite of all the challenges discussed above, big data provides a plethora of opportunities for organizations and individuals and these include:

Lower costs: Big data is helping in reducing cost. According to Basel, et al, 2013, big data analytics is helping US health industry save up to $450 billion by identifying and suggesting treatment to the patients based on demographics history, lifestyle choices, symptoms and other patterns.

Innovations and new Product Development: By analyzing the trend and customer preference, companies can innovate and offers new products and services to its customers.

Providing real time information: Companies can use real time data to offer better products and services to its customer. For example based on the place one is traveling, companies can share information related to the popular places to stay, eat and spend time on based on the customer's history. In future with the machine learning becoming faster, IoT (Internet of Things) will provide more reliable information

Lower barrier to entry: With time Big Data analytics will be integrated into day to day life of individuals. Predictability and scalability will grow many folds

Business Proliferation: The scope of usage of big data will not only be limited to customer retention and sales but other areas like product designing and development, product innovation and testing will also be made more faster with the usage of big data.

Conclusion

There is no doubt that the amounts of data gathered by the companies are increasing day by day. As it is said, data is new gold; the companies are required to store these data for the future usage. For this they need to find innovative data storage solution. With the development of cloud based solutions, this problem is taken care to a great extent. Though the challenges like quality of data, security of data, sources and format of data is there but they are required to find out the solutions of these challenges and the opportunities are great. The use of data in the decision making will keep growing and hence companies must focus upon building their dig data infrastructure.

References

Basel Kayyali, David Knott, and Steve Van Kuiken, The big-data revolution in US health care: Accelerating value and innovation. Retrieved from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

Bernard Marr (2017), The 7 Best Data Visualization Tools In 2017. Retrieved from https://www.forbes.com/sites/bernardmarr/2017/07/20/the-7-best-data-visualization-tools-in-2017/#27c2d9176c30

Big Data Challenges and Opportunity. Retrieved from https://www.qubole.com/resources/big-data-challenges/

BI-Survey.com (2017), Data Discovery: A Closer Look at One of 2017's Most Important BI Trends. Retrieved from, https://bi-survey.com/data-discovery

Data Visualization: What it and why it matters. Retrieved from https://www.sas.com/en_us/insights/big-data/data-visualization.html

Data Preparation for Analytics: Use the Right Methods and Tools to Effectively Prepare Your Data for Analysis (2017). Retrieved from, https://www.sisense.com/bi-insights/data-preparation-analytics/

Gil Press (2016, March 23), Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task. Retrieved from https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#471805086f63

IDC (2014), Digital Universe report. Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/index.htm

Rhea Sharma (2017, July 10), Top 5 Challenges in Big Data Analytics. Retrieved from https://upxacademy.com/big-data-analysis-top-5-challenges/

Shweta Iyer (2016, April 29), Big Data Analytics: Challenges and Opportunities. Retrieved from https://www.knowledgehut.com/blog/bigdata-hadoop/big-data-analytics-challenges-and-opportunities

Thomas Erl, Paul Buhler, Wajid Khattak (2016, Feb 08), Big Data Fundamentals: Concepts, Drivers & Techniques, Retrieved from http://www.informit.com/articles/article.aspx?p=2473128&seqNum=11