Analytical Study of Data Analytics and its Challenges

Neeta Yadav
PhD Scholar

Neelendra Badal
Director, REC Bijnor

ABSTRACT
In the world of information and digitalization data are seen everywhere. Everyday there is a new addition of data in the database. Emerging technologies are directly and indirectly responsible for adding momentum to the continuously multiplying of data in the world. It is also of different variety, volume, and velocity so it creates lots of problems in storage, accessing, preprocessing as well as in extraction procedures.

A gigantic quantity of data has become available on hand to decision-makers. So, it's also creating ambiguity in handling the large scale of data which is different in volume, variety, size, etc.

Due to its gigantic growth, solutions should be studied for analysis, storage, and management. Either the researcher will have to develop a model or algorithm or follow some hybrid methodology for its efficient handling.

Data scientists face many problems when dealing with big data or large-scale data. None of the tools are very efficient, and one of the biggest problems is storage.

The main focus of this paper is to discuss the challenges of data analytics from different perspectives and analyze different methods and tools for handling large volumes of data effectively.

Keywords
Data Analytics, Big Data, Structured Data, Unstructured data, and data mining.

1. INTRODUCTION
Data is the backbone of information technology as well as digitalization. Rising of this technology leads most the data born digitally and exchanged with the help of internet. In tech

driven world data is generated in every second from various resources that creates gigantic picture of data in the repository.

Around 2025 about 1 trillion of devices going to operate with the help of internet (IoT), humans will be moving towards Internet of Everything. It is also responsible for the immense growth of data.

Basically, data which are created from various sources are of different size, type and also grow with different rate. Handling of large and complex data is not easy with traditional database management tools or data processing application.

These are of different kinds like structured, semi-structured and unstructured.
The prime objective of data analytics is to process data of high volume, velocity, variety and veracity using different techniques of data analysis [1].

Mostly, data warehouse has been used to manage large dataset. But extraction and storage of data from gigantic data becomes big issue. Fig No. 1 shows the growth of data from 2010 to 2025.

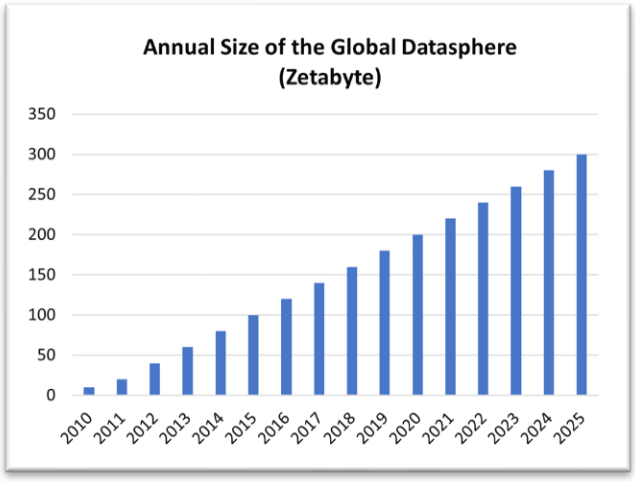

Fig No. 1

This paper discusses the challenges in analyzing large scale of data or big data, tools and techniques available for handling data. In this paper comparative studies of tools are presented, results and analysis also discussed and in last conclusions made.

1.1 Steps of data analysis platforms
There are many core features of a data management platform. These are shown in fig no.2 as follows:

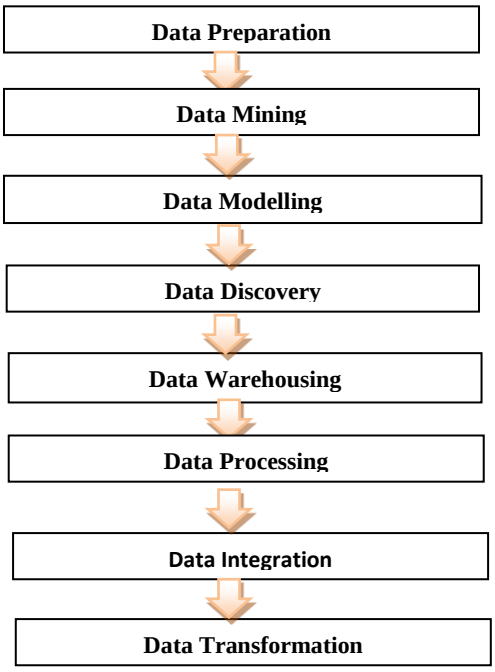

Fig. No.2

2. CHALLENGES IN DATA ANALYTICS

Data analysis is the process of detecting, scrubbing or cleansing, transforming, and modelling data with aim of discovering meaningful information, conclusions, and insight.

It is just the science of analyzing raw data to derive conclusion from the information. The process of data analytics shown in fig no.3.

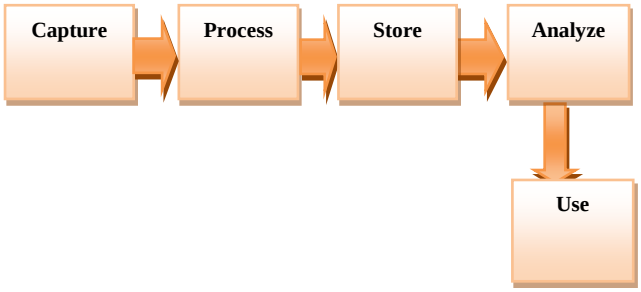

Fig. No. 3

It is a very important task of risk management. It helps a lot for decision making, increase accountability. Some of the major challenges of data analytics which are listed below:

a) Gigantic growth of data
b) Meaningful and real-time data collection
c) Graphical representation of data
d) Analysing data from multiple resource
e) Inaccessible data
f) Noisy data
g) Shortage of skills

3. TOOLS OF DATA ANALYTICS

Data analysis tools help you collect large data sets from various sources and combine them into databases. Data analytics tools can be a speciality software solution meant for data scientists. But many data platforms are easy enough for anyone to use.

Data platforms analyze data to tell us about the things, about your business process. The results from data analysis help us to shape future business decisions. Some of the famous tools of data analytics listed below:

a) Rapid Miner

It is a best data analytics tool. Its main aim to provide in depth business analytics. It is free as well as professional version costs $7500 per user per year.

b) Microsoft Excel

It is a spreadsheet software, mostly used for data collection and wrangling as well as reporting. It is widely used software; its availability is commercial. But it is poor in handling big data or large scale of data.

c) Python

It is an open-source programming language, mostly used with DL/ML for analysis as well as visualization purposes.

d) Jupiter Notebook

It is an interactive authoring software. It is an open-source software. Mostly used for code writing, sharing and presenting work.

e) Apache Spark

It is a free open-source data analytics tool. It was coming into existence in 2009. This software mostly used by data scientist and coders. It is a great processing and streaming tool in real time.
It is best suited with python, Hadoop, free users etc.

f) SAS

It stands for Statistical analysis System. It is developed around 1960s and under constant development. This software used for business intelligence, multivariate, and predictive analysis. But its cost is expensive and graphical representation is poor.

g) KNIME

It is an open source and cloud-based platform. It is a data integration platform. It is mostly used for data mining and machine learning.

h) Sisense

It is data analytics software, developed in year 2004. It handles unstructured data. It is also good in data visualization. It uses machine learning algorithm to compare data sets and discover anomaly and deviations. Some of the other famous data analytics tools are Looker, Qlik, Sisense, Tableau, ThoughtSpot, and compared in table no.1below.

3.1 Comparison of various famous tools of Data Analytics

Table No.1

Tools Name	Year	Invented By	Paid/Unpaid	Best Use For	Pros	Cons
Rapid Miner	2001	Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer	Free/Commercial/ Closed Source	• Predictive Analysis • Large Business • Machine Learning • Data Science	• Flexible • Full Automation • Specialised for business solutions	Limited partioning activities for dataset to training and testing sets

Tableau	2003	Pat Hanrahan, Christian Chabot, and Chris Stolte	Open Source	• Data Visualization • Free Users • Startups • BI	• Great Visualization • Speed • Interactivity • Mobile Support	Poor version control No data preprocessing
KNIME	2004	Michael Berthold	Open Source	• Free Users • Startups • Business Analytics	• Easy to use • Open source	Lacks Scalability
Looker	2012	Lloyd Tabb	Closed Source	• Large Business • Enterprises • BI • Automated Workflows	• Fast • Easy to use • Good Visualization	Does not connect to excel
Qlik	1993	Bjorn Berg, Staffan Gestrelius	Closed Source	• Data Integration • Augmented Intelligence • IoT Data Management	• Cheap • Simple to use	Not suitable for large data set
Talend	2005	Bertrand Diard, Fabrice Bonan	Open Source	• Large Business • Enterprise • Big Data Analytics • Data Pipeline	• Fast • Efficient • Reliable	Scheduling options are limited. ML is not easy to incorporate
ThoughtSpot	2012	Ajeet Singh	Closed Source	• Large Businesses • Ease of Use • Data Diagnostics • Real Time Insight	• Uses advanced • AI and ML	Costly Less market share Not support big data
Microsoft Excel	1985	Charles Simonyi	Closed Source	• Data Wrangling and Reporting	• Widely used • Plugg-ins	Cost Calculations error Poor at handling Big data

4. SOME PROBLEMS SOLVED BY DATA ANALYSIS

- It helps in making use of unused business data
- Minimise misleading revenue models and forecast
- Highlights the minute mistake
- Lowers the challenges in customer service
- Make companies proactive instead of reactive
- Severity of urgent
- Removing the unknown data

Handling traditional data (organizational, academic data) is very fruitful with the available tool, but big data is not handled easily because its size is so large. Some of the differences listed in table no.2 below:

Parameter	**Traditional Data**	**Big Data**
Structure of Data	Structured Data	Structured, Semi-Structured, and Unstructured
Data Volume	Based on business volume, digitalization	Very high in petabytes or even more.
Velocity	Low to Moderate	High Velocity

Flow	Fixed	Continuous (not fixed)
Sources of Data	Organizational Data, Academics etc.	Organizational Data, RFID, social media, Google searches, Facebook, Twitter, WhatsApp etc.
Analytics	Historic view, status report	Real-time, feedback, sentiment analysis
Computing	Centralized	Distributed

5. RESULT AND ANALYSIS

To handle large scale of data is very complex, available tools are not very efficient. For handling data should free from impurities (pre-processed), after that apply hybrid model (combination of tool and best ML/DL algorithm) to handle structured, semi-structured and unstructured data. To reach maximum efficiency.

6. CONCLUSION

This paper presents the analytical study of various tools of data analytics, challenges etc.

It is found that rapid miner is overall good platform in data analytics, Tableau is best visualization tool, If the size of data set is not so large Microsoft excel is also a good option. The Qlik is cheap alternative of rapid miner. Market share ThoughtSpot is nearly 0.01%. The researcher and Data scientist have to choose tool according to their requirement and type of organization or business.

Handling traditional data is quite efficient and successful, but handling big data is quite challenging because it is of different velocity, variety, and velocity. The flow data is continuous and it is in distributed format.

7. REFERENCES

[1] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)

[2] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM InternationalConference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)

[3] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[4] Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)

[5] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analy- sis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)

[6] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011).

[7] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)

[8] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[9] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space- efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)

[10] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innova- tive Data Systems Research, pp. 261–272 (2011).

[11] Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)