IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 6, Ser. II (Nov. – Dec. 2020), PP 31-35
www.iosrjournals.org

A Comparative Outline of Tools Used In Big Data, Data Mining and Data Analytics

Dr. Aruna J. Chamatkar[1], Prof. Sachin Y. Zade[2]& Dr. Pradeep K .Butey[3]
[1]*(Associate Prof., MCA Department, Kamla Nehru Mahavidyalaya, Nagpur)*
aruna.ayush1007@gmail.com
[2]*(Assistant Prof., MCA Department, Kamla Nehru Mahavidyalaya, Nagpur)*
[3]*(HOD, Department of Computer Science, Kamla Nehru Mahavidyalaya, Nagpur)*

Abstract:
We are living in a digital world now. The emergence of web and social networks has led to massive amounts of data being generated every single second, which presents both opportunities and challenges for the data field. The sheer volume of data calls for a change in our understanding of the data and how to extract usable information from this data. While traditional areas of computer science remain important, crunching through the massive volumes of data require new age tools and technologies such as Data Science, Data Mining and Data Analytics. Big data is undeniably a big deal, but it needs to be put in context. Data alone has no value, but the hidden patterns and insights in the data sets are an extremely valuable asset. Data Science, data mining and data analytics are often regarded as a subset of Business Intelligence.

In this paper we discuss the overall comparatives of techniques used in the field of data science, data mining and data analytics.

Key Word: *Big Data, Data Science, Data Analytics, Data Mining*

I. INTRODUCTION

The term Big Data is that describes large volumes of high velocity, complex and changeable data that require advanced techniques and technologies to enable, storage, sharing, management and analysis of the information. The amount of business data that is generated has risen gradually every year and more and more type of information are being stored in digital formats. Data Science a combination of mathematics, statistics, programming, the context of the problem being solved, ingenious ways of capturing data that may not be being captured right now plus the ability to look at things and of course the significant and necessary activity of cleansing, preparing and aligning the data. Data mining is a process used by companies to turn raw data into useful information. Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new view in a shorter time. Data challenges are the challenges related to the characteristics of the data itself. Data analytics can be referred to as the necessary level of data science.Both statistics and machine learning techniques are used to analyze data. Big data is used to create statistical models that reveal trends in data. These models can then be applied to new data to make predictions and inform decision making. Here in this paper we discuss the comparative of Data Science, Big data and Data mining on the basis of their some characteristics.

II. WHAT IS DATA SCIENCE

Data Science deals with both structured and unstructured data. It is a field that includes everything that is associated with the cleansing, preparation and final analysis of data. Data science combines the programming, logical reasoning, mathematics and statistics. It captures data in the most ingenious ways and encourages the ability of looking at things with a different perspective. Likewise, it also cleanses, prepares and aligns the data. simply, we can consider data science is an umbrella of several techniques that are used for extracting the information and the insights of data. Data scientists are responsible for creating the data products and several other data based applications that deal with data in such a way that conventional systems are unable to do.

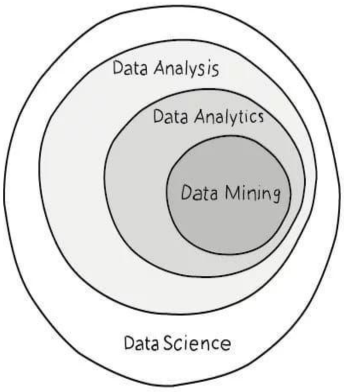

Figure:1Fields of Data science

III. WHAT IS BIG DATA

Big Data mining refers to the activity of going through big data sets to look for relevant information. Big data sets samples are available in different fields like astronomy, atmospheric science, socialnetworking sites, life sciences, medical science, government data, natural disaster and resource management, web logs, mobile phones, sensor networks, scientific research ,telecommunications.

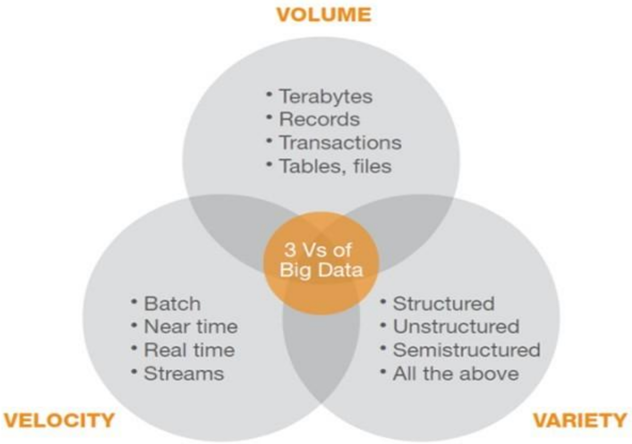

Figure: 2 Characteristics of Big data

Big data are characterized by Volume, Velocity, and Variety i.e. the combination of basic three V’s.

i) **Volume** - The size of data now is larger than terabytes and petabytes. The large scale and rise of size makes it difficult to store and analyze using traditional tools.
ii) **Velocity** – Big data should be used to mine large amount of data within a predefined period of time. The traditional methods of mining may take huge time to mine such a volume of data.
iii) **Variety** – Big data comes from a variety of sources which includes both structured and unstructured data. This heterogeneity of unstructured data creates problems for storage, mining and analyzing the data.

A Big data processing is the method of logical big data to uncover hidden patterns, correlations and other useful Data With big data analytics, data scientists and others can analyze massive volumes of data that usual analytics and business intelligence solutions can't tap. High-performance analytics is essential to process that much data in order to outline out what's significant and what isn't. For most organizations, big data analysis is defied..

IV. WHAT IS DATA MINING

Data mining is the technique or process of extraction of useful knowledge from the large volume of business data or data sets, is a great innovative technology, it is useful for data analysis and decision making. Knowledge is discovered from the dataset and presenting in the suitable form that is easily understood by humans. Data mining using various type of technique to analyze the very large dataset.Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases technology, informational science, visualization and other discipline to address the issue of information extraction from large database.

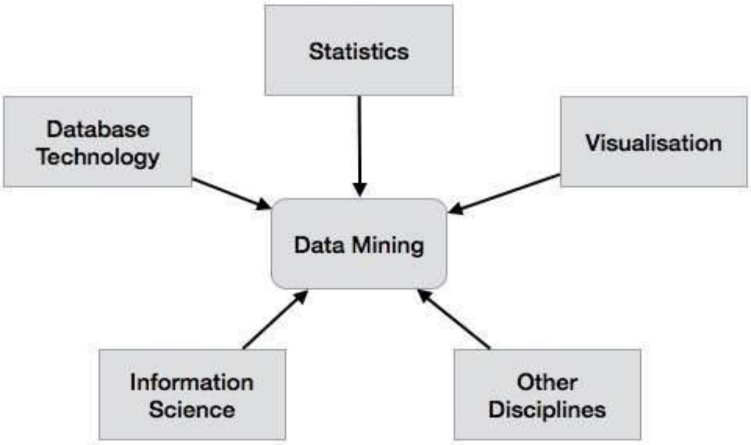

Figure:3 Interdisciplinary field of Data Mining

The growth in the field of data mining and knowledge discovery has been fastened by a variety of factors:

- The growth in data collection, as exemplified by the supermarket.
- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database.
- The availability of increased access to data from Web navigation and intranets.
- The competitive pressure to increase market share in a globalized economy.
- The tremendous growth in computing power and storage capacity.

There are many issues and challenges in data mining here we are considering some challenges that can be classified into the followings :

i) **Distributed data:** The data to be mined is stored in distributed computing environments on heterogeneous format. Both for technical and for organizational reasons it is impossible to bring all the data to a centralized place. Consequently, development of algorithms, tools, and services is required that facilitate the mining of distributed data.
ii) **Distributed operations:** In future more and more data mining operations and algorithms will be available on the grid. To facilitate seamless integration of these resources into distributed data mining systems for complex problem solving
iii) **Massive data:** Development of algorithms for mining large, massive and high-dimensional data sets is needed. Complex data types: Increasingly complex data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are emerging.
iv) **Data privacy, security, and governance**: Automated data mining in distributed environments raises serious issues in terms of data privacy, security, and governance.
v) **User-friendliness:** Ultimately a system must hide technological complexity from the user. To facilitate this, new software, tools, and infrastructure development is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

V. COMPARATIVE OUTLINE OF BIG DATA, DATA MINING AND DATA ANALYTICS

On the basis of some characteristics like data types, analysis style, data volume, expected results, focus ,data structure, output and tools and techniques, distinguishing the big data, data mining and data analytics. Data analytics is the science of analyzing raw data in order to draw conclusions about the information they contains.

Data mining can involve the use of different kinds of software packages such as analytics tools. The process of data science is much more focused on the technical abilities of handling any type of data. Unlike data mining and data machine learning it is responsible for assessing the impact of data in a specific product or organization. While data science focuses on the science of data, data mining is concerned with the process. It deals with the process of discovering newer patterns in big data sets.

It might be apparently similar to machine learning, because it categorizes algorithms. However, unlike machine learning, algorithms are only a part of data mining. However, in data mining algorithms are only combined that tools as the part of a process.

Table No. 1 : Comparison of Big data , Data mining and Data Analytics on the basis of some characteristics

Characteristics	Big Data	Data mining	Data Analytics
Data types	Structured, Semi-structured and unstructured data in NoSql and triple stores.	Structured data in spreadsheet, relational and dimensional database etc.	data analysis can be performed on structured, unstructured, or semi-structured data.
Analysis style	Focus on prediction and discovery of relevant business factors on a large scale using computational intelligence.	Focus on prediction and discovery of relevant business factors on a small sale using computational intelligence.	● Focus on descriptive Analysis, diagnostic analysis, predictiveanalysis. ● Prescriptive analysis.
Data Volume	Large datasets based on distributed and highly scalable processing structure.	Small datasets worked on data sample (small portions) with high data processing cost. on distributed and	Larger volume of the data. Data mining is a step in the process of data analytics.
Expected Result	Dashed board with predictive indicator and strategic recommendation.	Reports with recommendation for strategic decision making.	reports based on analysis and presenting important decisions by identifying various facts and trends
Focus	Knowledge extraction from large data sets with from various sources or kinds of files	Identify data patterns creating new analysis indicators for business intelligence.	focus of data analytics lies in inference, which is the process of deriving conclusions
Data Structure	Structured and unstructured	Very structured	Less Structured
Output	analyze large-scale data through the batch processing technique.	Data Pattern	Develop Models and reslts
Tools and Techniques	● Tableau. ● Apache Hadoop. ● Apache Spark. ● Zoho Analytics. ● MongoDB. ● Xplenty.	● WEKA. ● SAS. ● KNIME. ● Orange. ● IBM SPSS Modeler. ● Apache Spark. ● Rattle.	● R and Python. ● Microsoft Excel. ● Tableau. ● RapidMiner. ● KNIME. ● Power BI. ● Apache Spark. ● QlikView.

VI. CONCLUSION

In this paper we discuss the different Data management technologies, all interrelated through data. Data may be anything which is basic entities in organization. Data has attributes and featured .These data management technologies are interdisciplinary fieldthat includes everything that is associated with the cleansing, preparation and final analysis of data. We studies comparative outline over Data Science vs. Big Data vs. Data Analytics.Discussed minor and major differences between Data Science vs. Big Data vs. Data Analytics such as definition, Data types, Data Structure, focus, Outcomes, tools and techniques used in this fields.

This studies concludes the challenges and comparative of these Big data, Data Science and Data Mining related to these data management technologies. The major challenge in the case of big data to pay more attention and designing system and to elevate well-organized data analysis tools that provide guarantees on the output when data comes from different sources. Data Science provides the best solutions that help to fulfill the challenges of the ever-increasing demand and maintainable future.

ACKNOWLEDGMENT

I am grateful to my Respected guides Dr. PradeepK..Butey, for giving proper guideline and useful suggestion. Ialso happy to acknowledge Dr. Dilip S. Badwaik, Principal of Kamla Nehru Mahavidyalya, Nagpur to provide necessary help to me. Finally last but not the least I am very thankful to my family and my friends for their kind support.

References

[1]. S. Agarwal, Divya and G. N. Pandey, ―SVM based context awareness using body area sensor network for pervasive healthcare monitoring‖, IITM, ACM, New York, (2010), pp. 271-278.

[2]. M. Kendrick, ―Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey‖, http://www.ioug.org/p/cm/ld/fid=91, (Retrieved on September 2, 2015), (2012).

[3]. Ms. Aruna J. Chamatkar,"Data mining classification methods and different Techniques", International Journal of Computer Application ,Issue 4,volume 4(July-August 2014).

[4]. D. Tomar, and S. Agarwal, ―A survey on Data Mining approaches for Healthcare‖, International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, **(2013)**, pp. 241-266.

[5]. RaghavToshniwal ,kanishkaGhoshDastidar Ashok Nath (2015), - Data Security issue and challenge, International Journal of Innovative in Advanced Engineering (IJIRAE) ISSN:2349-2163 ISSUE 2, Volume 2 ,February.

[6]. Khan, N., Yaqoob, I., Hashem, I.A.T. et al., 2014. Big Data: Survey, Technologies, Opportunities, and Challenges. The Scientific World Journal, vol. 2014, Article ID 712826, 18 pages

[7]. Albert Bifet, (2013), "Mining Big data in Real time", Informatica 37, pp15-20.

[8]. Ren (2014)-Information Security in Big Data:Privacy and Data Mining- in IEEE Access, vol. 2, no. ,pp. 1149-1176.

[9]. Ms. Aruna J. Chamatkar, Dr. P.K. Butey" Importance of Data Mining with Different Types of Data Applications and Challenging Areas", May 2014

[10]. NirmalKaur, GurpinderSingh:A Review Paper On Data Mining And Big Data,Volume 8, No. 4, May 2017 (Special Issue).