International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-2, July 2019

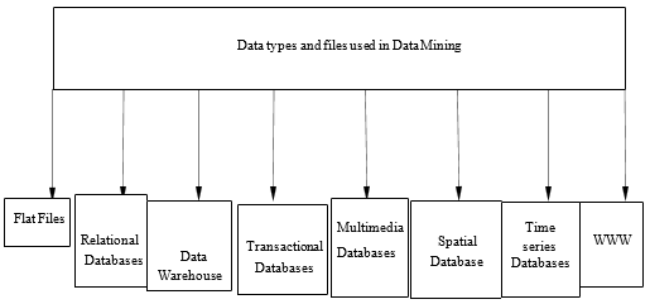
Collaborating Data Mining Modeling with Big Data Analytics for Disaster Prediction

Sumita Mukherjee, Prinima Gupta, Felix Musau

Abstract: ***Data mining is the procedure of bringing out the earlier unfold justifiable, logical, intelligible, functional information from large databases, big data to deliver accurate prediction, decision and implementation systems in engineering, business, research and education world. Data mining will effectively introduce the computing strategies and techniques to retrieve the applicable and convenient information from combined large databases known as big data. This paper signifies and explains Big data, Data mining and the importance and ease of Data mining using big data as a back end for delivering appropriate forecasts, prediction and experimental prospective solution as a front end.***

I. INTRODUCTION

Big data reflects to be a strong alternative when traditional procedure of grasping data with various techniques are unable to handle the data in depth for activating a satisfying solution. Big data caters many big and massive databases together from different sources (business applications, public



web, social media, and sensor data and so on) of any size and requires readily available hardware to store for further solution. The massive scale, the speed of consuming and accessing, and the nature, features of the data that must be dealt with an easy concept and solutions to manipulate the new challenges at each stage of the process. Data Mining plays a very strong role of extracting information from the big data. In other words, we can say that, data mining is mining the required information, reference to deliver an accurate result from data. To successfully access and process the big data like knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, regression etc. techniques of data mining becomes handy and comparative to activate the goals through big data analysis.

Revised Manuscript Received on July 09, 2019.
***** Correspondence Author
Sumita Mukherjee*, Research Scholar, Manav Rachna University, Faridabad (Haryana), India.
Dr. Prinima Gupta, Professor, Manav Rachna University, Faridabad (Haryana), India.
Dr. Felix Musau, Professor, Riara University, Nairobi, Kenya.

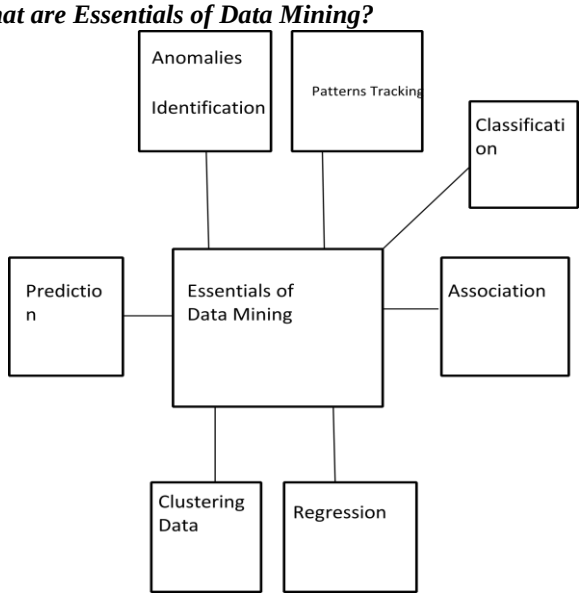
Definition of Data Mining: It is the enactment of natural scanning through massive reservoir of data for finding the common patterns and trends which is not meaningful and accurate with general analysis. It involves enlightened arithmetic algorithms to analyze the data and classify the degree of chances of future occurrences. It is commonly known as Knowledge Discovery in Data (KDD). Following queries can be answered by Data Mining a) Automatic Discovery Patterns b) Prediction of likely outcomes c) Creation Of actionable information d) Focus on large data sets and databases.

Data types and files used in Data Mining:

1. Flat Files: These files deal with a structure having text or binary form. No relationship is observed between tables if the tables are stored in flat files. Data dictionary signifies flat files. Example CSV file. Application: Oracle, SQL etc.
2. Relational Databases: The two aspects of relational databases physical schema deals with table's structure and logical schema deals with table's relationship. Thus, when data are kept in tables which is a representation of rows and columns is known as relational database. Example SQL, Application: Data Mining, ROLAP model, etc. 3. Data Warehouse: When data is composed, integrated from different sources will be utilized for queries and decision-making processes is known as data warehouse. Enterprise, Data mart, Virtual are three categories of Data warehouse. Example Oracle, Amazon Web services etc. Application: Business decision making, Data mining etc.
3. Transactional Databases: Organization, collection of data by time factors, day, date etc. constitute transactional database. If the transaction is not finished, the database has a scope to lay off or reverse its operation. Example: Sale on a credit to customer, Purchase consumable supplies from a supplier etc. Application: Banking, Distributed systems, Object databases, etc.
4. Multimedia Databases: They are capable of handling complex data, complex formats and Object-Oriented Databases, audio, video, images and text media. Example: Stack Exchange, Mongo DB etc. Application: Digital libraries, video-on demand, news-on demand, musical database, etc.
5. Spatial Database: It comprises of different types of coordinates, topologies, lines, polygons and specialized in storing geographical data. Example: Geodatabase of countries, administrative divisions, cities etc. Application: Maps, Global positioning, etc.
6. Time-series Databases: A real time analysis is capable of storing arrays of number sorted by time, date, day month etc. Example: Stock exchange data, user logged activities etc. Application:

eXtremeDB, Graphite, InfluxDB, 8. WWW: World wide web is a storehouse for data collecting from different sources. As the information is growing, changing continuously it is treated as a dynamic process. It can hold various records like audio, video, picture, text etc. linked by Hyper Text Manipulation Language pages and can be processed through the network. Example: www.cs.ucl.ac.uk,www.riarauniversity.ac.keetc.
Application: Online shopping, Job search, Research, studying, etc.

What are Essentials of Data Mining?

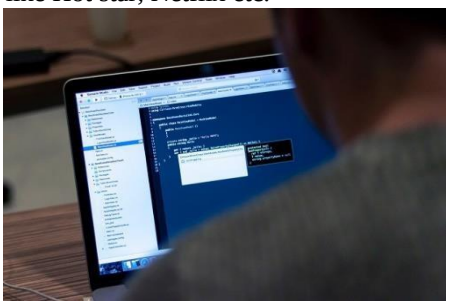


a) Anomalies Identification: No fixed trends are maintained. Business should have a proper understanding of the trends so that it can avail the trend to receive maximum advantage. For example, that a company's target demographic is old men a large the number of young women buyers sharply increases, and then returns to normal.
b) Patterns Tracking: This plays a very vital role for the business to grow. Observing a common pattern is potential and reliable for the growth of the business for example during cricket world cup sport shoes show maximum sales in most of the shops globally.
c) Classification: When we keep the data having same features and characteristics in specific group justifies classification for example aquariums, fishes, fish food, water pumps will comprise one group named “fish items”. Thus, classification builds links among elements in a data sets and also develops a mathematical algorithm for better prediction.
d) Association: Classification and association go hand in hand. Though some data elements show connections but not logically the same data elements can be categorized as one group. For example, buying skirt, oil, mutton, cups have no association and connection between these items.it can be explained as while checking departmental store a designer store is also visited. Thus, Association can be handy for marketing plans and dealing with customers.
e) Prediction: If a company accesses data properly on common patterns and abnormality then the company can make an accurate prediction for future of a significant change and adjustments. For example, in bank, loan approval for the client studies his history of money balance, his credit and debit history, default record etc. must be analyzed. If any abnormal pattern is detected, then it should be reported as a “doubtful client”.
f) Clustering Data: Clustering data is similar to classifying data, so much so that some people might even confuse the two. It's an admittedly and honest mistake, but it's one that must be nipped in the bud no less. Clustering data can be thought of as much less general than classifying data. Classifying data sorts items into specific classes, while clustering data organizes data by similarities. These similarities don't have to be extremely significant. Imagine that you were going shopping for pasta and made your way to the pasta aisle in the store. Rigatoni is clearly not fettuccine, yet you'll find both in the same aisle. Finding both items in the same aisle makes shopping for pasta much simpler than it would be otherwise.
g) Regression: A band of continuous values are used in a specified dataset for analysis and future prediction. Regression plays very vital role for financial forecasting, understanding business, data modelling, forecasting business etc. for example, the value of a property based on its amenities, location, probability of increasing price per square feet, level of commuting factor, nature of neighborhood etc.

II. EXAMPLES OF DATA MINING

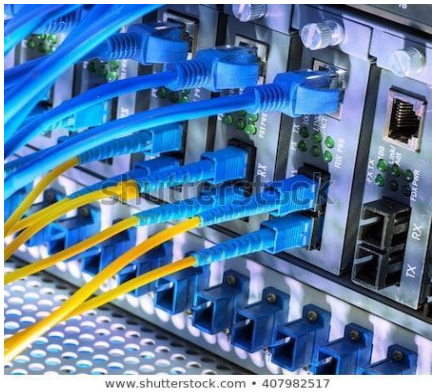
1.Crime Prevention Agencies: These agencies use the mined data studies thoroughly and understands the trend and common pattern and predicts future crime and terrorism activities (where, when, how many are involved, when who is responsible etc.)

2. Artificial Intelligence and Machine Learning:
Most of the e-commerce websites like Amazon, Flip cart etc. studies and analyzes past data and behaviors, trends and also display the products in a specified category for sell and purchase analysis. They use AI to depict the recommendation of their product, various audio video platforms like Hot star, Netflix etc.

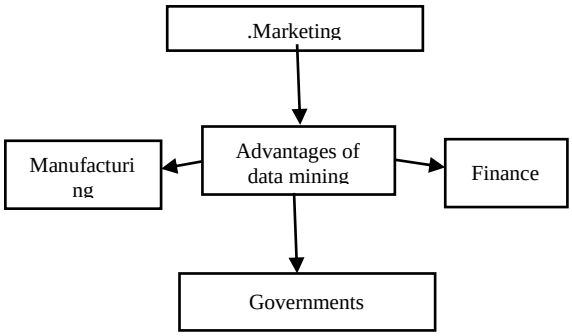


3. Supermarkets and Retail Stores: The Managers and owners can realize the choices and preferences of the customer. Studying the history, performances, movements, actions etc. the shopkeeper can conclude that the female customer
4. is going through menopause. Such is the power of data, patterns, and analysis. In general, these retail stores divide the customers into different categories like strong buyer, medium buyer and no buyer and each buyer will be subject to different style of attracting offers.

Service Providers: Data mining helps service provider to withhold the customers to them and provides all the facilities like personalized attention, discounts, billing information, website visits, easy payment schemes free downloading of apps etc. all possible facilities so that customers stick to the original service provider and don't opt for another vendor. Today, every service provider has terabytes of data on their customers.



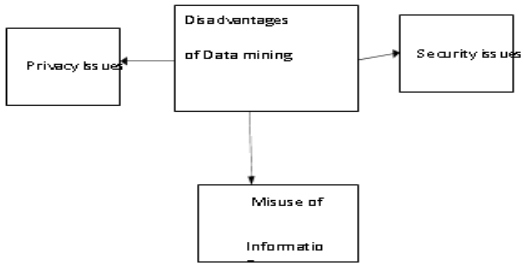
Advantages of Data Mining:



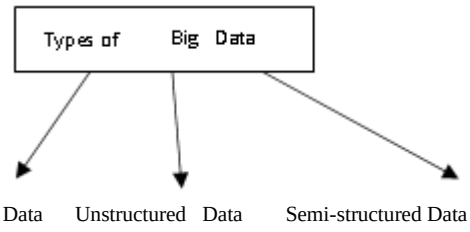
1.Marketing / Retail: Using data mining, historical data develops a model the marketing companies to predict the responses of customer, the effect of marketing campaign, profitability, frequent buying capability of customers, need of discount and what percent discount, or other means to have a foot fall of more customers etc.

2. Finance / Banking: Using data mining, historical data develops a model for catering an appropriate information on debit, credit history. It also helps to determine the dead and effective loans. It is very useful to recognize the fraud customers.
3. Manufacturing: Using data mining the faulty equipment, most suitable equipment, required parameters, manufacturing condition, quality and quantity control etc. can be predicted efficiently and is effective for manufacturing unit.
4. Governments: Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

Disadvantages of Data mining:



1. Privacy Issues: People are not very comfortable of their personal information due to privacy issues, as many times social networks, e-commerce, forums, blogs etc. are hacked and many unethical practices are observed causing troubles. The personal information is gathered during the survey form for the business operation and once this information is available can be leaked and misused.
2. Security issues: Security is a big issue. Every employee's and customer's information like employee number, birthday information, payroll information, mobile number, address etc. are easily available and accessed. It becomes very easy for the hacker to access the personal information and misuse like credit card stolen and used, troubling by making fake calls etc.
3. Misuse of information/inaccurate information: Data mining uses information can be misused unethically, can be exploited by hackers, unethical platform to take an advantage of easy going and simple people. There is not hundred percent accuracy observed through data mining, thus, inaccurate information might cause wrong decision-making and might not have a pleasant outcome.



Defining Big Data:
The big data is an idiom which expresses a voluminous data comprising structured, semi-structured and unstructured data and information.

Big data also narrates the massive data with a difficulty to process with conventional database and software techniques. Though it is very convenient to mine the big data and recommends a big contribution to analytical and prediction methods. Types of Data:

Data Semi-structured Data

Book_ID	Author_Name	Publication	Category	Shelf_ Number
2365	Rajesh Kulkarni	Khanna Publishing House	Finance	D-4 right-red
3398	Pratibha Joshi	Vikas Publishing House	Admin	B_2left_ red
7465	Shushil Roy	White Globe Publishing	Admin	C-5left - green
7500	Shubhojit Das	Publishing India Group	Finance	B-3right-blue
7699	Priya Sane	Pioneer Science publication	Finance	A-2left-red

Structured Data: A specified protocol is defined to store, process and retrieve data. The information is categorized and preserved in a systematic manner and can be easily acquired by simple search engine algorithms. For instance, the book table in a library database structured as the book details, category, rack number, synopsis, publisher etc., will be available in a methodical scheme. Few more examples are data from sensors such as GPSs, RFID tags, medical devices, data from network a web logs, retail and e-commerce data.

Unstructured Data: A defined protocol is not applied to the data structure and it becomes hard and prolonged to refer the data to retrieve a productive and analyzed report. Examples are e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Semi-structured Data: This data is the combination of both structured and unstructured data. Here data will not be placed under a particular database, but the extent of Information retrieved will be able to categorize different attributes within the data. The data represented by semi structured data holds a few parts of data (5 to 10%) so the last data type is strong one: Examples are CSV but XML and JSON documents are semi -structured documents, NoSQL databases are considered as semi structured.

What are the essentials of Big Data?

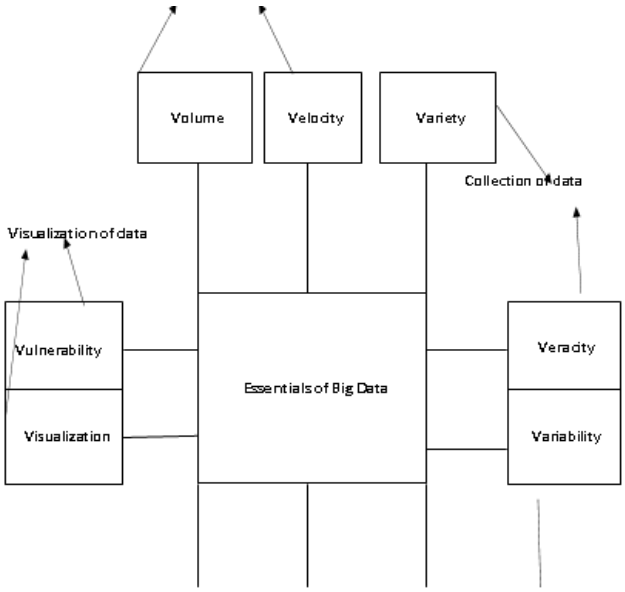
Processing of data

a) Volume: 95 percent of all present data was collected and reserved in the old records. The latest amount of data available is amazing and fascinating. By channelizing the data, the data can be stored in various locations and can be retrieved by an appropriate software. Examples are: incredible amounts of data generated every second and minute from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc.

b) Velocity: It refers to the fast rate at which the huge data are collected, processed, analyzed and refreshed. Around the globe, the number of emails, twitter messages, photos, video clips, etc. are increasing so heavily that data is analyzed during generation and then stored in the database through big data. Examples are Facebook claims 600 terabytes of incoming data per day. Google alone processes on average more than "40,000 search queriesevery second," which roughly translates to more than 3.5 billion searches per day.

c) Variety: It means a variety of data can be used for big data. This refers to absorption of structured data, semi structured and mostly unstructured data. Examples are 80 percent of data accommodate, includes photos, video sequences, audio, images, social media updates, etc. It can also include other text formats like log files, click data, machine and sensor data, etc.

d) Variability: Multiple data types, data sources, data formats are responsible for data variability of Big data. The data collected as big data are conflicting as Big data observes inconsistences among data before it is loaded to databases due to fast speed involvement. Example is Suppose a pastry shop has 5 different types of pastry and tastes same whereas if they taste different is an example of variability. An outlier detection method is applied for a meaningful result.



e) Veracity: It means the accuracy and confidentially of data. It depends directly on the meaningful data source; its contents so that the strong analysis can be carried out based on sources. It is not directly proportional to other characteristics. One good example is the reference of Global Positioning System data. Mostly the GPS will “drift” and manipulate the situation .GPS loses all the signals as they have to travel through bigger structures and buildings .When this situation arises the location has to be changed and moved to another road structure.

f) Validity: Validity refers to the degree of required exactness, fullness, rapport of need that data users or corporates have for accessing big data. It is used to outline the data satisfactorily within user-defined conditions or a user-defined range.

One example is earthquake prediction satellite indicates that there is a probability of tremors in one part of the world. What would be the impact of the tremors on people? The pouring responses on Twitter makes it possible to predict tremors and analyze the impact of a tremors on local populations.

g) Vulnerability: It is the incapacity to withstand a risk or to acknowledge a calamity. To explain this feature, we take an example of those who live in polluted areas are more vulnerable to asthma than those who live in less polluted area. Big Data is stored either in the cloud or in physical storage including various systems. Losing track of data forces to protect the open and shaky data.

h) Volatility: This means the validity and longevity of data which is stored. In a standard data setting, we can retain data for decades because we with time can analyze the required data and formulate rules, with the available data in such a way so that mapping of data is easily established to deliver a required output in our businesses. We can dispose the data once the data validity expires. For example, an online retail company erases the discount scheme offered to customers as after one year the discount offer becomes null and void.

i) Visualization: Another characteristic of big data is how challenging it is to visualize the response time, performance, expandability. It also depicts how data association, data clusters, decision tree, common pattern, regression, network presentation will visualize a better analysis and prediction through big data. For example, if we consider a particular air service company for better service through visual analysis. Six variables are plotted: the number of aircrafts, its flying location on a two-dimensional surface (x and y), fight duration, time of flying, kind of services.

j) Value: The other interesting characteristic is value which is always traced in big data and reflects the importance of data being abstracted. The value enforces better understanding of data user, their requirements, performance. Using optimal process value can bring improvisation on big data. Example is if a retail shop is in a position of analyzing data can save costs and offer better benefits.

How much data does it take to be called Big Data?

Generally, when data, is equal to or greater than 1 Tb known as Big Data. Analysts predict that by 2020, there will be 5,200 Gigabytes of data on every person in the world. Examples of Big Data:

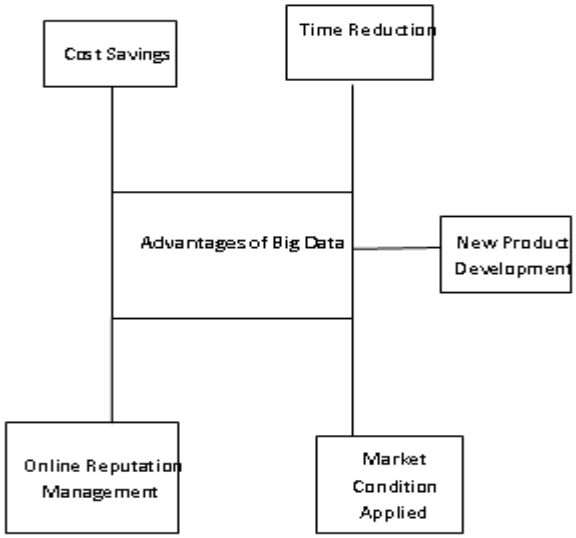
1.The Indian Stock Exchange uses about one gigabyte per day for just one trade transaction. Imagine the number of trades handled per day and the data consumption made by ISE per day.

2.The statistic shows that 500+terabytes of new data every day get pushed into the databases of social media site Facebook. This data mainly consists of photo and video uploads, message exchanges, putting comments, on average, people spend about 50 million tweets per day, etc.

3.A railway enabled train station in India uses 350 MB data per session. Imagine the number of sessions a railway station goes through and consumption of data reaches to few terabytes.

4. A Jet engine can generate 10+terabytes of data in 30 minutes of flight time. With many thousand flights per day, generation of data extends to many Petabytes.

Advantages of Big data

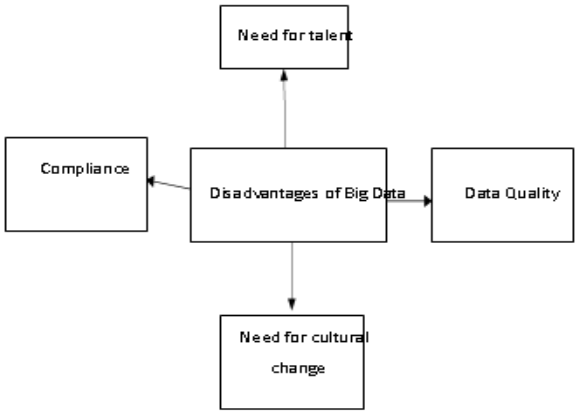


1. Cost Savings: Big Data uses tools for example Hadoop and Cloud-Based Analytics which can render cost savings to business specially when a huge amount of data is stored and accessed. These tools also help in identifying more efficient ways of doing business. It has two components a) Better decision-making b) Reduce costs

2. Time Reduction: Tools like Hadoop with high speed and in-memory analytics identify new sourced data and are utilized for business analysis and induce the fast decision based on the concepts and principles. It also has two components a) Increased Productivity b) Improved Customer Service

3. New Product Development: By understanding the customer's choice, needs and desire of the customer, applying analytics one can design and supply products according to customer's demand. It has two components. a) increased revenue b) increased agility
4. Market Condition Applied: By grasping the big data it is easy to figure out the retail conditions. For example, by understanding and analyzing customers' purchasing pattern, a business house can apprehend the nature of fastmoving products. Competitors can also be challenged. It has two components. a) Faster speed to markets b) Greater Innovation
5. Online Reputation Management: This can be handled largely by the feedback from customers and people surfing the online. Big data tools are very helpful to monitor and improve the online visibility. It also has two components. a) Sentiment Analysis b) Fraud Detection.

Disadvantages of Big Data



1. Need for talent: We need a very trained staff to handle the big data. Data scientists and big data experts are among the most highly coveted —and highly paid — workers in the IT field though the survey shows that there is a lack of a trained big data skilled personals. It thus becomes challenging for the business house to appoint the skilled persons.
2. Data Quality: Data quality plays a big concern in big data. The data analysts and the data scientists must emphasize on the correctness, relevance and accuracy of data so that proper format of analysis is observed. The improper quality is not only harmful but also will bring wrong report. It covers two aspects. a) Cyber security risks b) Costs
3. Need for cultural change: Most of the organizations Many of the organizations like to use big data analytics to demonstrate a data driven culture in the company. Companies want to create a change in the culture so that by using analytics creation of a data-driven culture throughout the organization is observed. It will take time to change the entire culture. It covers two aspects. a) Rapid change b) Hardware needs
4. Compliance: The big data implications depend on or complied with Government regulations the business houses must ensure and implicate the industry standards and government rules and regulation for storing and handling data. Many surveys depict data governance, is the most strategic barrier to working with big data. This covers one important aspect. a) Difficulty integrating legacy systems.

Differences between Datamining and Big data

Heading	Data Mining	Big Data
Focus	Mainly concentrates on huge data on their description.	Mainly concentrates on huge data on their relationship.
View	This depicts the near proximity view of data.	This depicts the larges sizeable view of data.
Data	It expresses the whereabout of the data for consideration. It means what of data.	It expresses the need of data for consideration. It means why of data.
Volume	It refers to small, medium large volume of data for usage.	It refers to only large voluminous data for usage.
Definition	This can be defined as a method for data analysis.	This can be defined as a concept for understanding data analysis.
Data Types	Relational, (Db2 Express, Oracle Database XE) Dimensional (SPSS, SDAR) and Structured data (MATLAB, IBM Watson are considered.	Unstructured, (SPSS, NLP) Semi structured (HTML, SQL/BI) Structured data (are considered.
Analysis	It utilizes statistical calculations for various kind of predictions and detection of features for business development on small degree of data.	It utilizes statistical calculations for various kind of predictions and detection of features for business development on large degree of data.
Output	It reflects on critical decision making. It is used as a handler of	It reflects on indicator panel and predictive measures. It is the asset.
Kind of Information	 Lots of detail Information	 Lots of relationships

III. CONCLUSION

As we foresee that in coming years demand of big data will be enormous and it will be difficult for data scientist to manage in terms of both hardware and software assistance. Not only the data will be larger, the diversification of data will be a big concern and challenges for everyone. The KNN classifiers will not compliant to the applications of big data. VFDT data will have also restrictions to produce quality results. Thus, data mining will be the most suitable for practical and theoretical analysis of big data by its powerful different techniques. Though, big data will throw an insight to discover unusual facts which was never treated so easily earlier. Data mining is going to apply its own techniques and will filter the big data in such a manner that only the productive, useful and applicable data are extracted from the big data to access easily, effectively in an understandable and logical manner. Both will work hand in hand in future to use for forecasting, predicting, foreseeing the consequences in almost all fields like science, research, retail, commerce, hospitals, environment etc. so that a proper preparation are taken care by the stockholders. Numerous technology progress has been made such as the classification method for big data with both categorical and numerical attributes, heterogeneous mixture learning, and the online feature selection (OFS) algorithm, etc.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my guide Dr. Prinima Gupta, and co -guide Prof. Felix Musau for their guidance and support. I will forever remain grateful for the constant support and guidance, discussion by them and helped to solidify my ideas to write this paper. I would like to extend my thanks to my dear colleagues Ms. Carolyne, Maryanne Gichuhi, and Mr. Faithful Wachirafor co-operating and motivating me.

REFERENCES

1. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg ,2011
2. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 ,2011
3. R. Vrbić, “Data mining and cloud computing,” Journal of Information Technology & Applications, Vol. 2, No. 2, pp. 75-87, 2012.
4. W. Fan, A. Bifet, “Mining Big Data: Current Status, and Forecast to the Future,” ACM SIGKDD Explorations, Vol. 14, No. 2, pp. 1-5, December 2012.
5. A. Bifet, “Mining Big Data in Real Time,” Informatica, Vol.37, pp. 15–20, 2013.
6. G. Krempl, I. Zliobaite, D. B. Nski, E. H. Ullermeier, et. al., “Open Challenges for Data Stream Mining Research,” ACM SIGKDD Explorations, Vol. 16, No. 1, pp. 1-10, 2013.
7. B. Thakur, M. Mann, “Data Mining for Big Data: A Review,” International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, No. 5, pp. 469473, 2014.
8. Anand S, Narayana K. Earthquake Reporting System Development by Tweet Analysis, International Journal of Emerging Engineering Research and Technology. 2014 Jul; 2(4):96−106.
9. V. Nekvapil, “Cloud computing in data mining – a survey,” Journal of Systems Integration, No. 1, pp. 12-23, 2015
10. Ikram, A. and U. Qamar, Developing and expert system based on association rules and predicate logic for earthquake prediction, Knowledge-Based Systems, 75, 87-103, 2015 [11] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In:
11. IEEE translations and con- tent mining is permitted for academic research. 2016.
12. Priyanka Jain, NilayKhare,ManasiGyanchandani, Big data privacy: a technological perspective and review DOI: 10.1186/s40537-016-0059-y,2016
13. Abdullahi,Abubakar Imam, ShuibBasri,Rohiza Ahmed New generation databases also called NOSQL(Not only SQL) databases are highly scalable, flexible, and low latent published in Journal of big data,6th december,2018.
14. Phillip M.Lacasse,WilkistarOtieno, Francisco P.MaturanaA hierarchical, fuzzy inferenceapproach to data filtration and feature prioritization in the connected manufacturingenterprise published in Journal of big data 19th November 2018.