DOIs:10.2015/IJIRMF/202205007 --:-- Review Article

Big Data versus Data Mining – A Review

[1]Hetal Chokshi, [2]Himani Bhatt
[1]Assistant Professor, Computer Engineering Department, Dr. Jivraj Mehta Institute of Technology, Anand, Gujarat, India
[2] Lecturer, Computer Department, Parul Institute of Engineering & Technology - Diploma Studies, Parul University, Vadodara, Gujarat, India
Email – [1]hetalshah108@gmail.com, [2]himanibhatt297@gmail.com

Abstract: *The term 'Big Data' might sound scary but it is actually a reality of today's data-filled and data fuelled world. Big data refers to the sets of data that are vast and also highly varied and versatile, which makes them difficult to handle using traditional tools and techniques. As this kind of data is increasing in quantity and number, we have to find new possibilities and solutions to further work upon that data. We need modern tools to process such a kind of data. Also, important functionality derived from such big data is the feature of decision making and value extraction. So it becomes mandatory that such data be evaluated and extracted carefully. We can do so by using Big Data Analytics which serves as a tool to ease our task of filtering and extracting important data. This paper helps reader understand the concepts of Big Data, Data Mining and the potential difference between the two topics which seem the same but are vastly different.*

Keywords: *Big Data, Data Mining, Analytics, Data Cleaning, Algorithm, Orange Data Mining.*

1. INTRODUCTION:

A broad range of techniques can be used to mine data for anomalies, patterns and correlations in order to predict outcomes. This information can be used to increase revenue, cut costs, improve customer relationships, reduce risks, and more. Data mining is the process of finding useful information from given data Using machine learning, statistics, and database systems, data mining is a system of recovering and discovering patterns from large data sets. Mining large data sets involves combining machine learning, statistics, and database systems to prize and identify patterns. But it’s not just the type or quantum of data that’s important; it’s what associations do with the data that matters. Big data can be evaluated for points that can be helpful for improving decisions and using it in business tools. We are going to see the difference between the two most commonly and interchangeably used words: Big Data and Data Mining which are not the same. It is generally considered big data when it is more than 1 Tb in size. Analysts predict that by 2020, there will be 5,200 GB of data per person in the world. The significance of Big Data doesn't mean how important data we've but what would you get out of that data. We can breakdown data to reduce cost and time. Data Mining also known as Knowledge Discovery of Data refers to rooting knowledge from a large quantum of data i.e. Big Data. It's substantially used in statistics, machine learning and artificial intelligence. It is the step of the “Knowledge discovery in databases”. This paper provides an in-depth review of the two techniques and briefs about the usage of both in modern science.

2. DEFINITION AND USAGE OF BIG DATA :

Big data refers to huge amount of data which is not easy to handle with traditional ways. It might be structured, semi- structured or unstructured.
It comprises of 5 Vs-

1. Volume: Refers to amount of data. (can be in quintillions)
2. Variety: Refers to type of data we can use. (structured, unstructured or semi-structured)
3. Value: Refers to the worth of data being extracted.
4. Veracity: It refers to the quality of the data we have.
5. Velocity: Refers to the speed at which our data is growing

Data exists everywhere. There is abundance of data everywhere around us. Starting from a shopping mall to a business to education, data persists everywhere. Businesses are driven by profitability they give in terms of monetary benefits, these tools help in providing meaningful information for making better business decisions and can also be used to study various other things which could benefit humanity in general. The main conception in Big Data on the other hand is haste, source, security of the huge quantum of data at our disposal. Big Data has wide range of applications ranging from Business, Healthcare, Government, Insurance, Information technological, media, etc. Big Data is the collection of data that is huge in volume, growing rapidly with time. It is a data with so large size and complexity that none of traditional data processing or management tools can store it or process it efficiently. Big data is data but with a very large size.

3. DEFINITION AND USAGE OF DATA MINING

Data Mining is a fashion to prize important and vital information and knowledge from a huge set/ libraries of data. It derives correctness by rooting, reviewing, and recycling the huge data to find out pattern and co-relations which can be important for the business. It is similar to the gold mining where gold is extracted from rocks. Data mining involves six common classes of tasks:[5]

- Anomaly discovery (outlier/ change/discovery) – The identification of faulty data records, that might be intriguing.
- Association rule literacy (reliance modeling) – Searches for connections between variables. For illustration, a supermarket might gather data on client purchasing habits. Using association rule literacy, the supermarket can determine which products are constantly bought together and use this information for marketing purposes. This is sometimes referred to as request handbasket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another" analogous", without using given structures in the data.
- Bracket – is the task of generalizing given structure to apply to new data. For illustration, ane-mail program might essay to classify ane-mail as" licit"or as"spam".
- Retrogression – attempts to find a function that models the data with the least error that is, for estimating the connections among data or datasets.
- Summarization – creating a more compact representation of the data set, including visualization and report generation.

Data mining is a pivotal element of successful analytics enterprise in associations. The information it generates can be used in business intelligence (BI) and advanced analytics operations that involve analysis of nonfictional data, as well as real- time analytics operations that examine streaming data as it's created or collected.Effective data mining aids in various aspects of planning business strategies and managing operations. That includes customer- facing functions analogous as marketing, advertising, deals and customer support, plus manufacturing, force chain operation, finance and HR. Data mining supports fraud discovery, threat operation, cyber-security planning and numerous other critical business use cases. It also plays an important part in healthcare, government, scientific exploration, mathematics, sports and further. [14]

4. DATA MINING: THE PROCESS:

Data mining process Data mining is generally done by data scientists and other professed BI and analytics professionals. But it can also be performed by data-smart business judges, directors and workers who serve as citizen data scientists in an association. Its core rudiments include machine literacy and statistical analysis, along with data operation tasks done to prepare data for analysis. The use of machine literacy algorithms and artificial intelligence (AI) tools has automated further of the process and made it easier to mine massive data sets, similar as client databases, sale records and log lines from web waiters, mobile apps and detectors.

The data mining process can be broken down into these four primary stages:
Data gathering. Relevant data for an analytics application is identified and assembled. The data may be located in different source systems, a data warehouse or a data lake, an increasingly common repository in big data environments that contain a mix of structured and unstructured data. External data sources may also be used wherever the data comes from, a data scientist frequently moves it to a data lake for the remaining way in the process.
Data medication. This stage includes a set of way to get the data ready to be trapped. It starts with data disquisition, profiling and pre-processing, followed by data sanctification work to fix crimes and other data quality issues. Data metamorphosis is also done to make data sets harmonious, unless a data scientist is looking to dissect undressed raw data for a particular operation.
Mining the data. Once the data is set, a data scientist chooses the applicable data mining fashion and also implements one or further algorithms to do the mining. In machine literacy operations, the algorithms generally must be trained on

sample data sets to look for the information being sought before they are run against the full set of data. Data analysis and interpretation. The data mining results are used to create analytical models that can help in the decision-making and other business decisions used for various purposes. The data scientist or another member of a data science team also must communicate the findings to business executives and users, often through data visualization and the use of data storytelling techniques.

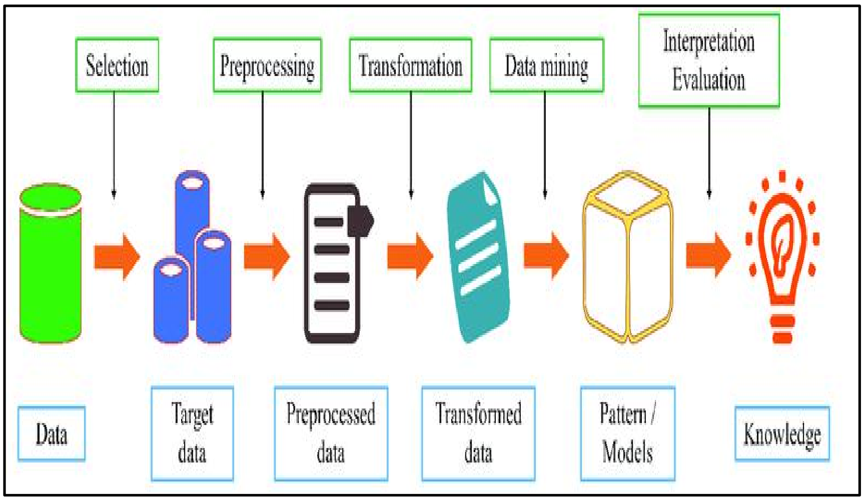

Figure 1: Data Mining Process

Types of data mining techniques:

Various techniques can be used to mine data for different data science applications. Pattern recognition is a common data mining use case that's enabled by multiple techniques, as is anomaly detection, which aims to identify outlier values in data sets. Popular data mining techniques include the following types:

Association rule mining. In data mining, association rules are if-then statements that identify relationships between data elements. Support and confidence criteria are used to assess the relationships -- support measures how frequently the related elements appear in a data set, while confidence reflects the number of times an if-then statement is accurate.

Classification. This approach assigns the elements in data sets to different categories defined as part of the data mining process. Decision trees, Naive Bayes classifiers, k-nearest neighbor and logistic regression are some examples of classification methods.

Clustering. In this case, data elements that share particular characteristics are grouped together into clusters as part of data mining applications. Examples include k-means clustering, hierarchical clustering and Gaussian mixture models.

Regression. This is another way to find relationships in data sets, by calculating predicted data values based on a set of variables. Linear regression and multivariate regression are examples. Decision trees and some other classification methods can be used to do regressions, too.

Sequence and path analysis. Data can also be mined to look for patterns in which a particular set of events or values leads to later ones.

Neural networks. A neural network is a set of algorithms that simulates the activity of the human brain. Neural networks are particularly useful in complex pattern recognition applications involving deep learning, a more advanced offshoot of machine learning.

5. BIG DATA ANALYTICS: THE PROCESS :

5.1. Collect Data

Data collection looks different for every organization. With today’s technology, organizations can gather both structured and unstructured data from a variety of sources — from cloud storage to mobile applications to in-store IoT sensors and beyond. Some data will be stored in data warehouses where business intelligence tools and solutions can access it easily. Raw or unstructured data that is too diverse or complex for a warehouse may be assigned metadata and stored in a data lake.

5.2. Process Data

Once data is collected and stored, it must be organized properly to get accurate results on analytical queries, especially when it's large and unstructured. Available data is growing exponentially, making data processing a challenge for organizations. One processing option is batch processing, which looks at large data blocks over time. Batch processing is useful when there is a longer turnaround time between collecting and analyzing data. Stream processing looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making. Stream processing is more complex and often more expensive.

5.3. Clean Data

Data big or small requires scrubbing to improve data quality and get stronger results; all data must be formatted correctly, and any duplicative or irrelevant data must be eliminated or accounted for. Dirty data can obscure and mislead, creating flawed insights.

5.4. Analyze Data

Getting big data into a usable state takes time. Once it's ready, advanced analytics processes can turn big data into big insights. Some of these big data analysis methods include:

Data mining sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.

Predictive analytics uses an organization's historical data to make predictions about the future, identifying upcoming risks and opportunities.

Deep learning imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

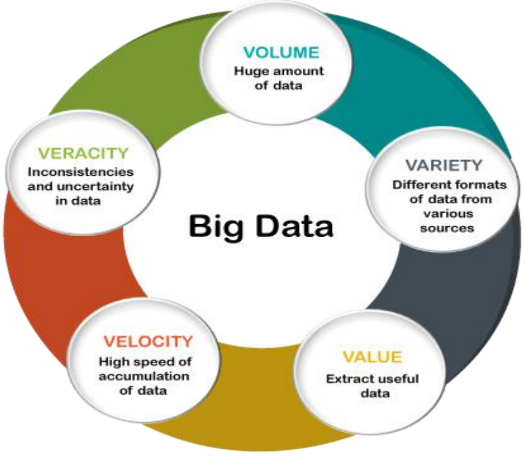
VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

Figure 2: Big data Characteristics

6. TOOLS AND TECHNOLOGICAL COMPARISON:

DATA MINING:

Data mining tools are available from a large number of merchandisers, generally as part of software platforms that also include other types of data wisdom and advanced analytics tools. Listed below are some of the popular tools available in the request.

Rapid Miner

Oracle Data Mining

IBM SPSS Modeler

Knime

Python

Orange

Kaggle

Rattle

Weka

Teradata

H2O

Apache Spark

Sisense

Xplenty

Crucial features handed by data mining software include data medication capabilities, erected-in algorithms, prophetic modeling support, a GUI- grounded development terrain, and tools for planting models and scoring how they perform.

Merchandisers that offer tools for data mining include Alteryx, AWS, Databricks, Dataiku, DataRobot, Google,H2O.ai, IBM, Knime, Microsoft, Oracle, RapidMiner, SAP, SAS Institute and Tibco Software, among others. A variety of free open source technologies can also be used to mine data, including DataMelt, Elki, Orange, Rattle, scikit- learn and Weka. Some software merchandisers give open source options, too.

For illustration, Knime combines an open source analytics platform with marketable software for managing data wisdom operations, while companies similar as Dataiku andH2O.ai offer free performances of their tools.

Orange is a machine literacy and data wisdom suite, using python scripting and visual programming featuring interactive data analysis and element- grounded assembly of data mining systems. Orange offers a broader range of features than utmost other Python- grounded data mining and machine literacy tools.

The rattle is an R language grounded GUI tool for data mining conditions. The tool is free and open- source and can be used to get statistical and visual summaries of data, the metamorphosis of data for data models, make supervised and unsupervised machine literacy models and compare model performance graphically.

7. BIG DATA ANALYTICS:

Big data analytics cannot be narrowed down to a single tool or technology.

Rather, several types of tools work together to help you collect, process, cleanse, and dissect big data. Some of the major players in big data ecosystems are listed below.

Hadoop is an open- source frame that efficiently stores and processes big datasets on clusters of commodity tackle. This frame is free and can handle large quantities of structured and unshaped data, making it a precious dependence for any big data operation.

NoSQL databases arenon-relational data operation systems that don't bear a fixed scheme, making them a great option for big, raw, unshaped data.

NoSQL stands for “ not only SQL,” and these databases can handle a variety of data models.
Map Reduce is an essential element to the Hadoop frame serving two functions.
The first is mapping, which filters data to colorful bumps within the cluster. The alternate is reducing, which organizes and reduces the results from each knot to answer a query.
YARN stands for “ Yet Another Resource Negotiator.”
It's another element of alternate- generation Hadoop. The cluster operation technology helps with job scheduling and resource operation in the cluster.

Spark is an open source cluster calculating frame that uses implicit data community and fault forbearance to give an interface for programming entire clusters. Spark can handle both batch and sluice processing for fast calculation.

PARAMETER	DATA MINING	BIG DATA
Definition	One of the methods in Big Data	Technique to collect, process and maintain large amounts of data
What is it?	Data mining is a part of Knowledge Discovery of the Data.	Extracting the vital information from huge amount of the data. It is a technique of tracking and discovering of trends of complex data sets.
View	It is closer view of the data.	It is a large or overall view of the data.
Goal	The goal is same as Big Data as it is one of the tools of Big Data.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
Type	It is manual as well as automated in nature.	Only Automated.

Focuses On	It only focuses on only one form of data. i.e. structured.	It focuses and works with all form of data i.e. structured, unstructured or semi-structured.
Set of	It is a sub set of Big Data. i.e. one of the tools.	It is a super set of Data Mining.
Describes	It describes "what" about the data	It describes "why" about the data

Table 1: Comparison of Big Data and Data Mining

8. CONCLUSION:

We can conclude that the process of data mining is limited to one form of data which is used to extract and mine important information from a set of big data or library of data. It excavates only the important part from large amounts of data. The result of data mining can be used in knowledge discovery or decision making.

It facilitates the analyst in getting important information regarding some event or business. Thus we can say that the term data mining is related with extraction of information from small amounts of data. Big data in contrast, deals with extremely huge amounts of data comparatively. Nowadays, most of all the big companies and organizations have very large amounts of data where big data comes into picture. It can be in any form, structured, unstructured or semi-structured. Mostly, Hadoop is used for big data processing.

REFERENCES:

Journal Papers:

1. Nada Elgendy and Ahmed Elragal: Big Data Analytics: A Literature Review Paper. In Researchgate.net P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014.
2. D. P. Acharjya, Kauser Ahmed P: A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016
3. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on machine learning; 2004. pp. 1–9.
4. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15 (5):1170–87.
5. Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)
6. Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)
7. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)
8. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)
9. Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)
10. Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)

Web References:

- https://www.javatpoint.com/big-data-characteristics
- https://www.researchgate.net/figure/The-steps-for-data-mining-process_fig3_344166043
- https://www.tableau.com/learn/articles/big-data-analytics
- https://searchbusinessanalytics.techtarget.com/ definition/data-mining