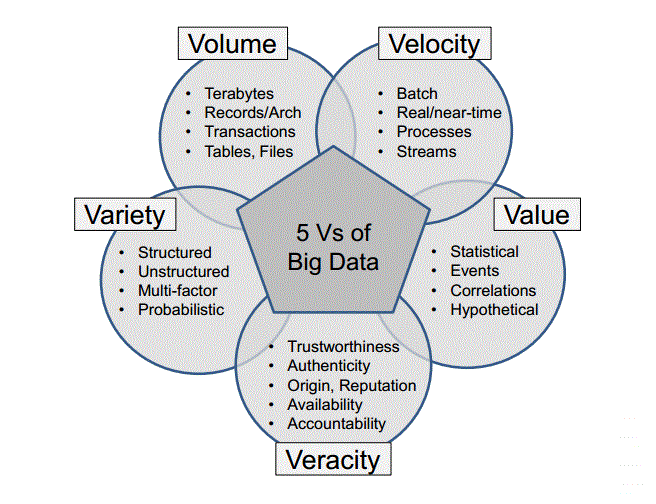
A critical analysis of the data mining problems on big data

# introduction

Big Data is a term that is used for calling the datasets that are so vast in size, which makes it practically impossible for any of the typical database software tools to store, manage or analyze it. They face some unique computational and statistical challenges, which includes scalability and storage issues, noise accumulation, spurious correlation and measurement errors which may cause inaccuracy. These challenges require some different computational and statistical approach. This article talks about the Big data Mining and its issues and challenges pertaining to its distinguished features. It also focuses on some of the methods to deal with big data. The 5 main attributes of big data are ***Volume, Variety,*** ***Velocity***, ***Veracity***, and ***Value***.



# Challenges of big data analysis

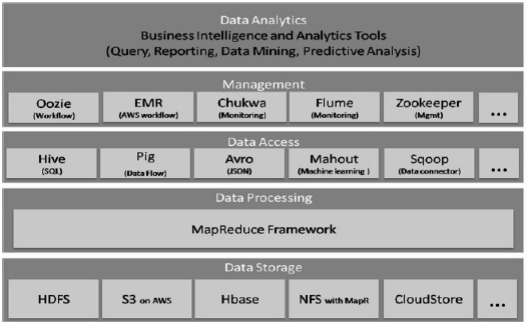
Big data analysis is all about applying some analytics and visualization techniques on these large data sets to analyze the patterns and correlations in the datasets for an effective decision making. It involves multiple stages like acquisition of data and recording, extraction of information and cleaning, integration of data, aggregation of data and its representation, query processing, modeling of data and analysis and interpretation of the data. Each of these stages introduces us to new challenges because of its vast size. Some of the big data mining challenges are

1. ***Heterogeneity*** - The problems of big data mining come not only from its large scale but also from the presence of mixed data which are based on different patterns or rules in the data. These are heterogeneous in nature, which are called as the heterogeneous mixture data. In cases of such complicated heterogeneous data, the data might have numerous patterns and rules, where its properties of the pattern might vary greatly. Another challenge is that the data can be both structured and unstructured. Nearly 80% of the data generated by industries are unstructured. They are very dynamic and may not have a common format. It will be in the forms of images, pdfs, documents, email attachments, medical records, voice-mails, videos, audios, etc. These kinds of data cannot be stored in tabular format as structured data either. Converting this data to a structured format for analysis is a major problem in big data mining.
2. ***Incompleteness*** - A data that is incomplete (missing values) exposes us to uncertainties during data analysis affecting the accuracy of the model. It is caused by several factors such as malfunction of a sensor or some technical error. While most data mining techniques have inbuilt algorithms to handle missing values by removing such fields, it still affects the accuracy as compared to its original data. Hence data imputation is recommended in such cases for better results.
3. ***Scalability & complexity*** - It’s a very daunting task to manage large and rapidly increasing volumes of data. The software tools that are used for managing the traditional datasets are useless in the case of big data as it can’t keep up with the such increase in the volume of data. Data analysis, data retrieval or even data modeling becomes a challenge due to its scalability and complexity.
4. ***Timeliness*** - Larger the dataset, larger the time it is perform an analysis on it. There are cases where the analysis result is required immediately. For example, in the suspicion of a credit card fraud, the transaction should be immediately flagged before the transaction takes place. A thorough analysis of user’s purchase history is not feasible in real-time. We should generate partial results in advance based on a small amount of incremental computation with new data to conclude at a quick determination. Scanning the entire dataset for this is impractical. In such cases, Index structures are developed in advance to find the qualifying elements fastly. The issue with that is index structures are designed to support specific classes of criterion.
5. ***Privacy*** - As big data expands its sources of data, the authenticity of each source of data has to be verified in order to find maliciously inserted data. Information security is a huge problem where large amount of data will be correlated and analyzed for meaningful patterns. Unauthorized information leak, unauthorized modification of data and denial of resources indulge in security violation. Security of big data can be increased by authorization, authentication, encryption and audit trails. Some of the methods for the protection of data are
   1. Use of authentication methods
   2. Use of file encryption
   3. Implementation of access controls
   4. Use of key management
   5. Logging and use of secure communication

# big data mining tools and techniques

Big data can produce important information for organizations to manage their business problems. These datasets are massive and complicated for users to effectively extract information without the help of computational tools. Modern technologies such as the Hadoop framework, Market Basket analysis and MapReduce offer incredible ways to process these data to obtain a meaningful knowledge.

1. ***Hadoop*** - Hadoop is an open source architecture that is scalable with fault tolerant Virtual Grid operating system for the processing and storage of data. It functions on commodity hardware, which uses HDFS (Hadoop Distributed File System) which is a fault-tolerant high bandwidth clustered storage architecture. It also runs MapReduce for distributed data processing and works with both structured and unstructured data. For handling the heterogeneity and velocity of data, it possesses tools like Pig, Hive and Mahout which are a part of Hadoop and HDFS framework. Hadoop and HDFS by Apache is mainly used for managing and storing big data. It consists of distributed file system, analytics platforms and a layer to handle parallel computation, workflow and configuration. It runs across the nodes in a cluster and connects to the file systems on multiple input and output data nodes to make them into a big file system. The Hadoop ecosystem consists of the Hadoop kernel, HDFS MapReduce and components like HBase, Apache Hive, Pig, Oozie and Zookeeper. Some of the big data mining tools can be seen in image below.



1. ***MapReduce*** – It is a programming model which is used for processing the large data sets with a parallel, distributed algorithm on a cluster. Hadoop MapReduce is a software framework and programming model for writing applications to process rapidly large amounts of data parallelly on large clusters of compute nodes. The MapReduce comprises of two main functions, map() and reduce(). Mapper function does the task of sorting and filtering, and reducer function does the task of summarization of the result. There can be many reducers to perform aggregations parallelly. It also enables the users to implement their own logic for process by mentioning a customized map and reduce function. The map function takes an input value pair and generates a list of intermediate value pairs. The runtime of MapReduce system clubs all intermediate pairs together based on the intermediate keys and allows them to reduce function for generating the final results. Map Reduce is used widely for the big data analysis. Controlling thousands of processors and managing parallelization in distributed environments makes it a more tedious process. Map Reduce serves as the key for the mentioned issues as it supports distributed and parallel I/O scheduling. It is also fault tolerant and supports scalability with inbuilt processes for status and monitoring of heterogeneous and huge datasets as in Big Data
2. ***Market Basket Analysis*** - Market Basket Analysis is a business intelligence technique used by for predicting the future purchase decisions of customers. It involves in the process of looking for combination of products that are frequently bought together. This technique functions by studying the customers’ buying patterns/preferences and is based on the idea that people who purchase a specific group of products are either likely or unlikely to buy other groups of products. When it is observed from analysis that customers who purchase one product are likely to buy another, the retailer can cross-sell the products together in order to drive sales performance. The technique used here is called Association Rules, which is to identify relationship between items that are being sold together. Apriori and FP growth are the algorithms used for this association rule.

Some of the common data mining tools that are used for the traditional datasets are ***RapidMiner, Python, R, Weka, SAS Enterprise Miner, Neo4j,*** etc.

# Conclusion

The data is growing exponentially worldwide day-by-day due to several factors. Big Data is becoming a new era for data research. Its analysis helps organizations to make better decisions, to predict and implement changes and opportunities. In this article we talked about the challenges of big data mining and its techniques which enables organizations to better understand their marketplace to take better decisions.

# references

1. Jaseena, Julie, David, 2014. ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING, Vol. 1, 10.
2. Albert Bifet, (2013), “Mining Big data in Real time”, Informatica 37, 15-20
3. Richa Gupta, (2014), “Journey from data mining to Web Mining to Big Data”, IJCTT, 10(1), 18-20
4. Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), “Big Data: Overview”, IJCTT, 9 (5)
5. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”