Understanding tools and applications of Data Analytics

Ravindra Bachate[1] , Supriya Dumbare[2], Gayatri Ladhe[3], Snehal Binawade[4], Srushti Gujar[5]

[1] *Assistant Professor, Department of Computer Engineering, JSCOE, Maharashtra, India*
[2] *UG Student, Department of Computer Engineering, JSCOE, Maharashtra, India*
[3] *UG Student, Department of Computer Engineering, JSCOE, Maharashtra, India*
[4] *UG Student, Department of Computer Engineering, JSCOE, Maharashtra, India*
[5] *UG Student, Department of Computer Engineering, JSCOE, Maharashtra, India*

ABSTRACT

In the current age innumerable amount of data is generated every day. The amount of data created and stored on a global scale is almost unimaginable, and it still grows. It is necessary to analyze such a huge amount of data, so that some useful information can be extracted from it. The information extracted from this 'Big Data' can be used in many fields which includes banking, education, healthcare, manufacturing, business intelligence and robotics. The motive of this paper is to understand that how big data is analyze. This paper also gives a short glance of data analysis implementation in the real world in a conjunction with challenges and advantages. This paper also examine various technologies and algorithms used in data analytics.

Keyword: deriving insights, data analytics, data mining

1. Introduction:

Every day around 2.5 Quintillion bytes of data is generated. This data comes from many sources such as posts to social media sites, purchase transaction, sensors used to gather shopper information, sensors used to gather shopper information, etc. This quite impossible to study this much volume of data. It is necessary to analyze this data so that some useful information can be extracted from it. Data analysis is a use for inspecting, cleaning, transforming and modeling the data to identify useful information and supporting decision-making.

1.1 What is data analysis

"Data analytics refers to qualitative and quantitative techniques and processes that are used to increase productivity and business profit." Data is removed, accepted, and split to identify and analyze behavioral data, techniques and patterns, can be dynamic according to the requirement or requirement of a particular business.

Data Analysis (DA) is a process to check the data set so that they can find out specifically about the information related to the support of specialized systems and software. Data analytics technologies and techniques are widely used in business enterprises to enable organizations to make more informed business decisions and enable scientists and researchers to verify or resist scientific models, theories and concepts.

Data analytics include data mining, which involves sorting through large data sets to identify trends, patterns and relationships; Predicted analytics, which wants to predict customer behavior, device failures and other future events; And learning the machine, an artificial intelligence technique that uses data algorithms, sets data faster than the scientists, through conventional analytical modeling.

1.2 Data analysis in healthcare

Data has always been the basis of a scientific approach to health care: clinical physical support Measurement, laboratory data, and clinical imaging; Analysis of the treatment and the effects of the potential illness causes are based on clinical and epidemiological studies [1]. Study design and data acquisition were the main challenges, while data volumes and data management were not. We expect that this will change rapidly as new sources of health care become more relevant than sources. The generated data sets are high dimensional and abundant; Pure amount is only explosion.

The quantity of data (i.e. volume of data) will be collected and many improvements will be made based on the analysis of these data in health care, as a key goal, with better results at manageable cost in the form of pre-condition to fulfill the full potential, the fundamental changes in health care system may be required and need of data privacy, data ownership, and data security issues must be resolved valid insights and actionable volumes are important solutions with health related data if data on many individuals collected, a statistical analysis, data mining, and train machine learning algorithms.

2. Challenges involved in data analytics

- **Knowledge Discovery and computational complexities**: Knowledge discovery and representation is a major challenge in data analysis. It includes number of sub fields like authentication, archiving, management, prevention, and representation. Since the volume of data keeps increasing exponentially, the available tools may not be able to process the data for getting useful information. It is a great challenge to develop a technology that can deal with computational complexity, Inconsistency and uncertainty in an efficient way [2].
- **Heterogeneity**: it means multiplicity. When human deals with information, there are more chances of heterogeneity which can be handled easily. But in case of computers, they work efficiently only if they are able to store multiple items that are identical in size and structure [3].
- People who understand data analysis: Data Analysis is very important to convert the huge amount of data into useful data. Therefore, there is a great need for Data analysts and Data Scientists. It is important for a data scientist to have skills that are varied as the job is multidisciplinary. This is another challenge faced by companies. The number of data scientists available is very less in comparison to the amount of data being produced.
- **Good quality analysis:** The companies and organizations use big data to make the best decisions. So, the data they are using must be accurate. If the data used to make decisions is not accurate it will result in false decisions that would ultimately be detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue. This requires a lot of resources to ensure the accuracy of the information provided. The process of creating accurate data is very time consuming and requires the use of tools that can be expensive

3. Tools for data analysis

- **KNIME**: KNIME allows you to manage, analyze, and modeling data in an especially intuitive way through visual programming. Essentially, rather than writing blocks of code, you drop nodes onto a canvas and drag connection points between activities. More particularly, KNIME can be continued to run R, python, text analysis.

- **Qlikview:** Qlik with their Qlikview tool is the other main important player in this space and Tableau's biggest competitor. This however can mean that it takes more time to get to grips with and use it to its full

potential. At its root, it is a data visualization tool just like to Tableau with a focus around analytics and statistics that connect to databases to provide a more holistic view for analysts and business managers to inform data supported decisions .The vendor has more than 40,000 customer accounts across over 100 countries, and those that use it repeatedly cite its highly personalized setup and wide feature range as a key advantage.

- **Highcharts:** Highcharts is another big player in the charting space. Like FusionCharts, it also offers a varied range of charts and maps right out of the box. Other than normal charts, it also offers a different package for stock charts called Highstock which is also very useful .It allows ship charts in PNG, JPG, SVG and PDF. Highcharts is free for non-commercial and personal use, but it will need to have to buy a license for deploying it in commercial applications.

- **Plotly**: Plotly enables more complex and sophisticated visualizations, it is integrated with analytics-oriented programming languages such as Python, R and Matlab. It is built on top of the open source d3.js visualization libraries for JavaScript, but this commercial package (with a free non-commercial licence available) adds layers of user-friendliness and support as well as inbuilt support for APIs such as Salesforce.

4. Algorithms used in data analysis

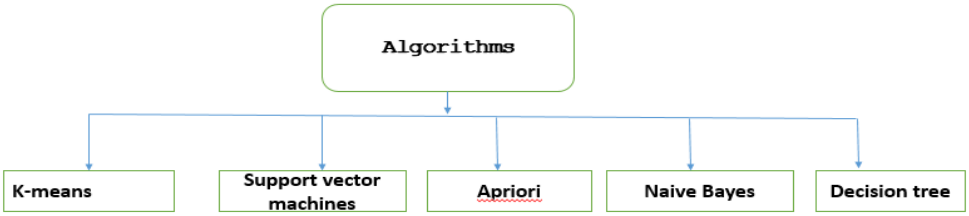

Fig -1: Algorithms used in data analysis

- **K-means:**
  K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way of classifying a given data set through a certain number of clusters (assume k clusters) .In this we define k centers, one for each cluster. These centers should be placed in a smart way because of different location causes different result. So, it is a better choice to place them as much as possible far away from each other. The next step is to take each point having relationship to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as 0barycenter of the clusters resulting from the previous step. After we have these k new centroids, a loop has been generated. As a result of this centers do not move any more.

- **Support vector machines:**
  Support Vector Machine is a discriminative classifier, it uses supervised machine learning algorithm which can be used in classification and regression challenges. But mostly, it is used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space and classify them using hyper-plane

- **Apriori:**
  Apriori algorithm is used for frequent data set mining and association rule learning over database. It identify the frequent items used in data set and extend them to larger item set as long as those items appear sufficiently often in the data set. This frequent item set is used to determine association rules which gives the common trends in the dataset.it is used in market basket analysis.

- **Naive Bayes:**
  Naive Bayes is a powerfull and very useful classifier use to constructs models that assign class labels to problem instances which are presented as vectors of feature values, where the class labels are drawn from some finite set. Naive Bayes classifier gives great results when we use it for textual data analysis. These classifier work on the base of Bayesian theorem

- **Decision tree:**

  A decision tree is a graph that uses a branching method to reveal every possible outcome of a decision. Decision trees are commonly used in operation research and operation management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. It is also use of decision trees is as a descriptive means for calculating conditional probabilities.

5. Applications of data analysis

- Social media: The amount of information available on social media sites is very large and it grow every second. The data from these sites is used by different organizations. Organizations in simultaneously analyze customer data along with behavioral data to create detailed customer profiles.it helps is creating content for different target audiences. Recommending content on demand and measure content performance. Increased use of social media sites like media like twitter, Facebook, weibo is the main cause for generation of big data. Big data analytics in social media era is a big challenge as all of the data generated by social sites is in large amount as well as in unstructured form [4].
- Healthcare: data generated by healthcare organization is very large and it is not in a structured form. Some hospitals are using data collected from a cell phone app, of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. A battery of tests can be efficient but they can also be expensive and usually ineffective. Data analysis helps in uncovering decision making by identifying data patterns and relationships between these patterns using machine learning techniques. Based on these data an improved healthcare is provided for patients. For security of data various encryption algorithms can be sued, so that malicious user cannot access the data. In such cases, map-reduce job helps in identifying that which user is responsible for the leakage of sensitive data [5].

- Business analytics: customer behavior analysis is an upcoming and new field which has a great potential for more betterment. It is important to know which customer is more beneficial buyers because it will help in future business. For data analysis different algorithms like map-reduce, decision tree etc. are used [6].

- Transportation: in recent times, huge amounts of data from location-based social networks and high speed data from telecoms have affected travel behavior. Regrettably, research to understand travel behavior has not progressed as quickly. Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management .Private sector use of big data in transport: revenue management, technological enhancements, logistics and for competitive advantage

6. CONCLUSIONS

In the current age innumerable amount of data is generated everyday. This paper gives a short glance of data analytics, its various challenges and issues and tools to analyze such large volume of data. Through better analysis of the large volumes of data that are becoming available, there is potential for making faster advances in many scientific disciplines and improving the profitability and success. By effectively applying different data analytics algorithms, nearly for every department involving sales and marketing, customer support, business intelligence, operation and maintenance, network construction, etc. can achieve significant benefits. We hope the content discussed in this paper, can be helpful for future analytics.

7. ACKNOWLEDGEMENT

We express true sense of gratitude towards our project guide Prof. R. P. Bachate, Assistant professor Computer Department for his invaluable co-operation and guidance that he gave us throughout our Project. We specially thank our project coordinator Prof. S. H. Patil for inspiring us and for providing us all the lab facilities. We would also like to express our appreciation and thanks to HOD Prof. H. A. Hingoliwala and Principal Dr. M. G. Jadhav and all our friends who have assisted us throughout our hard work.

8. REFERENCES

[1]. Volker Tresp,J.Marc Overhage, Markus Bundschus, Shahrooz Rabizadeh,Peter A. Fasching, and Yu Shipeng "Going Digital: A Survey on Digitalization and Large-Scale Data Analytics in Healthcare"

[2]. By Ravindra Bachate, Sayali Gaikwad, Pranali Nale "Survey on Big Data Analytics for Digital World." 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology

[3]. Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States

[4]. Rekha J.H, Parvathi R," Survey on Software Project Risks and Big Data Analytics", Procedia Computer Science 50 (2015) 295 – 300.

[5]. *J.Archenaa and E.A.Mary Anita,"* A Survey Of Big Data Analytics in Healthcare and Government**",** Procedia Computer Science 50 (2015) 408 – 413.