International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT239065

Big Data Mining : Tools, Technique, Application

Neelakshi Singh[1], Kunal Rajput[2], Rashmi Pandey[3]

[1,2]Computer Science, Institute of Technology & Management, Gwalior, Madhya Pradesh, India

[3]Assistant Professor, Computer Science, Institute of Technology & Management, Gwalior, Madhya Pradesh, India

ARTICLEINFO

Article History:

Accepted: 02 Nov 2023

Published: 20 Nov 2023

Publication Issue

Volume 9, Issue 6

November-December-2023

Page Number

77-81

ABSTRACT

Beyond the capabilities of traditional applications, big data is a vast dataset of incredible complexity. In the modern environment, it includes enormous, intricate, and voluminous structured, semi-structured, and unstructured data, together with hidden data supplied from various domains and origins.

The duties of data extraction, analysis, visualization, sharing, storage, transmission, and retrieval are all included in the issues provided by the management of big data.

As a result, it becomes urgently necessary to develop effective and efficient methods for mining big data.

Keyword: Bigdata, Traditional Data, Analysis, Tools, Business Intelligence

I. INTRODUCTION

Datasets with a defined size, a distinct and well-defined structure, and compatible with relational databases are referred to as small data. The relational database model, which stresses clearly defined relationships between various subjects or entities within the data, is the foundation for traditional analysis approaches for little data. These techniques are generally simple and rely on this model. Small data is a term used in the data management industry to describe well- structured datasets that have historically been kept in data warehouses.

These datasets are small and simple to comprehend. But with the start of the digital age came the era of "big data," which is distinguished by its huge volume, non-relational nature, and complexity. This information flood comes from a variety of sources, including social media, the medical field, business, industry, and academic study.

Big data analytics' main goal is to unlock the potential of this data, not only to store it. To process, analyze, and extract important insights from this enormous, unstructured, and heterogeneous data landscape, big data technologies and analytics methodologies are designed to break it down into "small data" that can be readily noticed and used for strategic decision-making. In addition to standard data tools, big data has plenty of chances for progress across numerous industries. It continues to be a major and relatively new topic of discussion in modern data talks since it offers a revolutionary method for managing huge, complicated

information and encourages innovation throughout the data industry.

Big Data - Definition:

Big data is a term used to describe high- volume, high-velocity, and high-variety information assets that require fresh, cost-effective methods of information processing for improved insight and decision making.

Big data is well known for improving company data by storing, processing, and analyzing previously ignored data due to limits in traditional data management methods. Extraction of value and meaning from this huge amount of information represents the main challenge in dealing with big data, rather than merely storing massive amounts of data.

Big Data Tools:

Structured, semi-structured, and unstructured data types are all included in big data, which frequently has storage capabilities measured in petabytes, exabytes, or zettabytes.

Specialized techniques and technologies capacities.

For the processing and management of huge data streams, major corporations like Google are increasingly turning to specialist big data tools rather than conventional methods. These tools are built to effectively handle a wide variety of complicated data formats. The R programming language, NoSQL databases, MapReduce, and

Hadoop are the four most widely used open-source big data solutions.

1. NoSQL -

NoSQL is free and open-source database software that is useful for managing large amounts of data. It is frequently used with other technologies such as massively parallel computing, columnar databases, and database-as-a- service (DaaS). Notably, Apache Cassandra, a NoSQL database engine, is used by well-known social networks including Facebook, LinkedIn, and Twitter.

2. MapReduce-

Using parallel and distributed algorithms on a cluster, programmers can process and analyze massive datasets using the open-source data mining approach known as MapReduce. Through several MapReduce libraries, this method supports a variety of programming languages, including C, C++, Java, Perl, and others. It makes it easier to create apps that can effectively handle heavy data loads.

The distributed computing framework MapReduce was developed with Google's help. It consists of the two crucial parts Map and Reduce. Within a distributed cluster, data is dissected, filtered, and sorted during the "Map" step. The "Reduce" stage, on the other hand, is a separate operation that combines the interim outcomes and processes them one at a time to ultimately produce a single consolidated result.

3. Hadoop

The open-source Hadoop framework was primarily created in Java. It provides a full range of tools and frameworks made for handling, creating, and running big distributed datasets

and applications.

Hadoop can handle vast, complex, and non-relational information that frequently total thousands of terabytes. It works with several OS, including Windows, Linux, BSD (UNIX), and OS X for Mac.

4 R

Statistical data processing and graphics are the main uses of the programming language R. It is a free software computing environment that Bell Labs' GNU has been working on as an open- source project. R is a computer language that implements S and is useful for handling massive volumes of data.

Data Mining Concept and Types:

Data mining uses both big data and artificial intelligence (AI) simultaneously to enhance the processing of massive and complex information. Making computers mimic human thought and behaviour is the goal to significantly improve the effectiveness and efficiency of data analysis.

To aid in problem-solving and decision- making, data mining entails the process of discovering, analyzing, and extracting useful data from data warehouses.

Knowledge Discovery in Databases (KDD) is the name given to the entire process.

The two main categories of data mining models are descriptive and inferential:

1. Descriptive data mining is frequently used in marketing for tasks like summarization, grouping, and interconnection data mining.
2. Predictive data mining is the second type, and it entails creating models based on current data to analyze and extract more precise classifications.

Using methods like classification, specification, and prediction, this model is frequently used in marketing to make forecasts, such as predicting the popularity of new items.

Data Mining Techniques:

Parallel data mining technologies have been developed to be used for solving numerous issues utilizing a variety of methodologies, including artificial neural networks, decision trees, rule induction, evolutionary algorithms, closest neighbor, and many more. In this section, several of these tools are briefly described.

Artificial neural networks (ANNs), a new advancement in computing, are modelled after how the human brain functions. They are made up of interconnected processing units that imitate neurons, which enables them to tackle problem types by replicating the neurological network of the brain.

The ability to extract and predict data from challenging or ambiguous inputs is a strength of neural networks. They can recognize complex trends and patterns that may evade human observation or other traditional computer technologies.

Decision trees have a tree-like form and are frequently utilized in operations research and decision analysis. They come in three primary node types: choice nodes, chance nodes, and end nodes. choice nodes are depicted as squares, chance nodes, as circles, and end nodes, as triangles.

An efficient way to improve the performance of information retrieval systems is to use genetic algorithms, a type of artificial intelligence, which use natural evolution strategies for optimization and search issues. They have numerous uses, including in the fields of software engineering, bioinformatics, artificial creativity, managing airline revenue, clustering, biology, chemistry, and electrical circuit design.

Big Data Applications :

In the contemporary information age, a variety of big data apps with remarkable uses are accessible to suit our needs for obtaining and analyzing data from various data sources. Most of each application and its most prominent usage are briefly introduced in this section.

Companies:

Major corporations have embraced big data mining, a potent technique that enables them to filter through enormous amounts of data to uncover insightful information. Companies can learn vital details about the behavior of current and potential customers by examining this data. Big data mining tools give businesses the ability to improve their pricing and marketing tactics, enabling them to better engage with their current clientele and draw in new ones. Big data mining ultimately acts as a catalyst for well-informed choices and focused activities that promote customer happiness and business progress.

Heath Care and Medicine Company- Big Data has significantly impacted the relationship between life sciences and the medical industry, enabling predictive analysis and speeding up communication between patients and physicians. It assists in predicting diagnoses, optimizing treatment modalities, predicting surgical results, and streamlining pharmaceutical breakthroughs through substantial data analysis. This synergy improves decision-making through better information, which ultimately benefits patients, the

effectiveness of treatments, and stakeholder cooperation in the medical industry.

Financial Banking-

Financial banks can build a variety of analytical models with the use of big data approaches, with a special emphasis on forecasting consumer behavior and customizing services for each customer group. Numerous potent big data tools are available for commercial applications in the financial banking sector. With the use of these tools, huge amounts of data can be processed and analyzed quickly, allowing for the discovery of patterns, trends, and insights that improve customer service, risk assessment, fraud detection, and tailored product offerings. In the financial banking industry, most these big data solutions mainly rely on categorization and prediction methods. The construction of an intelligent system known as a Business Intelligence (BI) system is supported in large part by these techniques. To evaluate and categorize data, the BI system makes use of classification and prediction algorithms.

Telecommunications-

Big data mining techniques are widely used by telecommunications companies to analyses various elements of their operations. Insights into user behavior, pricing strategies, failure analysis, customer recruitment techniques, forecasting finance requirements, and identifying client loyalty patterns can all be gained by examining records and calls. Mobile user data mining, which has developed into a crucial communication channel in both the work and personal spheres, is a cutting- edge technology in this area. Mobile user data mining is specifically made to examine and forecast the actions of mobile users based on actual user data, providing insightful information that can be used to enhance services and modify products in accordance with customer needs.

Industrial-

Tools for data mining are essential for automating different parts of managing industrial operations. They include programs targeted at crucial areas including process optimization, logistics, and quality control. Industries can improve quality control procedures, optimize production procedures, and streamline logistical operations by using these tools to evaluate enormous volumes of data. This automation increases productivity overall in the industrial sector by reducing errors, increasing efficiency, and reducing costs. Big data has a huge impact on many more fields in addition to the ones mentioned above. It has a revolutionary impact on science and engineering, transforming how data is examined and used to encourage progress. These are only a few examples of the diverse range of domains that benefit from the far-reaching implications of big data, underscoring its broad influence across various sectors. Furthermore, big data influences text and web data mining, enabling a deeper understanding of textual and online content to derive valuable insights and support informed decision-making.

II. Conclusion

We give a succinct explanation of big data that covers its historical context, terminology, defining traits, and important tools. We place emphasis on the shift away from conventional small data and toward the requirement for big data in scientific research. We also present the types of data mining and the methods for searching, extracting, and analyzing various data types. Big data was first introduced as a term in the early 2000s and gained popularity as digital data exploded because of the internet, social media, and developments in data processing and storage technology.

Possibilities for Big Data in Different Sectors:

Financial Sector: Accurate risk assessment, fraud detection, individualized client service, and real- time trade analytics are made possible by big data.

Big data helps with consumer behavior analysis, network optimization, quality of service improvements, and targeted marketing techniques in the telecommunications industry.

Industry: Predictive maintenance, supply chain management, and quality control are all improved by big data applications.

In conclusion, big data has become a disruptive force that presents enormous opportunity across numerous industries. With Data Fusion and Data Binding leading the way, it is poised to transform how we use and integrate data, enabling better analysis, more informed choice-making, and eventually, a more linked and effective digital world.

III. REFERENCES

[1]. Kudyba, S. (2014). Big Data, Mining, and Analytics: Components of Strategic Decision Making. CRC Press.

[2]. Elorie Knilans, “The 5 V’s of Big Data”, Avnet Advantage: The Blog, Soulution- Focused Insight for Growth-Minded VARs. http://blogging.avnet.com/ts/advantag e/2014/07/the-5- vs-of-big-data/#comment-474 (Last seen 05–April–2015)

[3]. The Four V’s of Big Data – IBM http://www.ibmbigdatahub.com/infogr aphic/four-vs-big- data (last seen 05–April–2015). 4.Feinleib, D. (2014). Doing a Big Data Project. In Big Data Bootcamp (pp. 103-123).Apress.

Cite this article as :

Neelakshi Singh, Kunal Singh Rajput, Rashmi Pandey, "Big Data Mining : Tools, Technique, Application", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 6, pp.77-81, November-December-2023. Available at doi : https://doi.org/10.32628/CSEIT239065
Journal URL : https://ijsrcseit.com/CSEIT239065