IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIG DATA ANALYTICS TOOLS AND TECHNOLOGIES USED IN CLOUD ENVIRONMENT

Renu Yadav (*Research Scholar*), School of Computer Application & Technology, Career Point University, Kota, Rajasthan

ABSTRACT

The term "big data" has recently become widespread because many products produce a huge quantity of data at an increasing speed. Big Data has gained prominence and is becoming the choice for new research. Although, big data huge volumes are difficult to store, manage, process, analyze, visualize and extract useful information from these datasets using conventional database approaches. Simultaneously, the cloud computing environment supports big data by providing computational, networking, and storage capacity, as well as functionalities and features that allow applications to develop, operate, deploy, and manage big data. Big Data Analytics is used to gain useful insights hidden in the big data for an enterprise's decision-making process. Proficiency in Big data analysis is necessary to gain knowledge from unstructured data from the web in the form of text, images, videos, or social media posts. In response to the growing demand for big data analytics, a large number of big data analytics tools and technologies are aided to gain insights from the enterprise. So, in this paper, we address the most relevant information regarding big data analytics tools, techniques, and their aiding technologies. We also emphasize the functions of several tools and techniques possibly used to describe from the perspective of big data in a cloud computing environment with recent developments that provide a better understanding to solve real-life challenges.

Keywords: Big data analytics, Machine Learning, Data Centre, visualization, storage, and processing tools.

INTRODUCTION

Big data analytics is essential to proactively process and analyze data and a user can evaluate and derive insights that are helpful for new findings and enable the selection process in an organization. Because of the cloud's elastic resource allocation and high computational capacity, and pay-as-you-go option for managing large data storing and conducting data applications. Cloud is useful for big data analytics, allowing for faster analysis, more timely results, and greater data value. To address the challenges of the cloud and gain valuable information and knowledge, high-performance and scalable computing systems are used under data and knowledge discovery techniques. Big data analytics empowers businesses to easily analyze their data in context and provide real-time analysis through high-performance data mining, predictive analytics, text mining, forecasting, and optimization. Big data analytics could improve business marketing and advertising outcomes, discover hidden economic opportunities, increase customer satisfaction, enhance efficiency, minimize risks, and accomplish other objectives. Although a higher volume of data produces an improved outcome, working with it might be challenging. This paper presents the description of big data and the characteristics of big data, big data analytics, and its types, life-cycle, and processing with aiding tools and technologies in the cloud environment.

Big data analytics (BDA): Big data analytics refers to the critical process of examining large and diverse data sets, or big data, for information such as hidden models, unknown correlations, market trends, and customer preferences that can help organizations (BDA). This analytics process requires the deployment of tools and technology to improve operational efficiency, and strategic potential drives new revenue streams and gains competitive advantages. BDA not only helps us to understand the information contained in the data but also identifies the information that is most important to the organization and future decisions. Organizations today store their data in multi-cloud environments, on-premise, and in data storage collections that would not completely comply with various security requirements. Cloud computing enables the collection of data from remote sources, breaking down legacy storage facilities and feasibly gearing up data for analysis. With high-performance data extraction, analytics, text classification, predictions, and enhancement, Big Data Analytics allows businesses to efficiently assess their data or information as well as provide real-time analysis. There is no single technology that encompasses big data analytics but many tools and techniques that can help to save time, and money, and provide valuable business insights [1].

Types of Big Data Analytics: It could be categorized into four subcategories such as descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics: This technique is the most time-consuming and mostly produces the least value; but even so, it is useful for discovering trends within a specific segment of consumers. The descriptive analysis focuses on providing knowledge about what has happened in the past and we will provide trends to further assess. This articulates past data in an easy-to-read format. This allows the formation of reports such as income statements, profit, sales, etc. This also enables the gathering of data sets.

Diagnostic Analytics: Diagnostic analytics is a technique for determining why something occurred. This is required to determine the cause of the issue. Diagnostic analytics are used by enterprises since they provide a thorough knowledge of a specific problem. **Predictive Analytics:** Most broadly used technique, predictive analytics employs models to predict what may occur in particular situations. This type of analysis examines earlier and present data for forecasting the future. Predictive analysis analyzes current data and makes predictions using mining, AI, and automatic learning. Non-discrete predictions of future states, relationships, and patterns are the focus of this research. Forecasting that isn't discrete. **Prescriptive Analytics:** This technique provides a laser-like emphasis on responding to a specific question. It assists in identifying the appropriate solution among a variety of options, based on known specifications, and indicates alternatives for how to put emphasis on a better future or mitigate uncertainties. It can also be used to highlight the repercussions of each decision to improve decision-making. Perspective analytics can be used with both descriptive and predictive analytics[2].

Processing of Big Data: Data process could also be defined as data-at-rest,data-in-transit, and data-in-use. Data-at-rest is also known as batch processing and it is responsible for the protection of data on storage devices. It is hard for a user to achieve caused of limited physical control over the data. The security of transmission of data can be referred to as data-in-transit. It ensures that all data is not compromised, changed, or removed. Data-in-transit will be more vulnerable than data-at-rest because it moves from one location to another. Data-in-use is also known as stream processing. It refers to a variety of data processing (creation, transformation, or deletion). We needed to use a dedicated ETL tool (Extract, Transform, and Load) to transform the contents of one database into a script, converting it to a different tabular form if needed, and then making a few basic decisions regarding dirty data.

Tools and Techniques used in Big Data Analytics: Numerous big data analytics techniques can assist and reduce time and money while also providing useful business insights. Of course, advanced technology that works together to enhance the value of data is provided.

1. Storage and Processing tools and techniques

HADOOP: This is an approach to examining big quantities of information throughout many node clusters. It's a Programming language open-source program that incorporates Hadoop Distributed File System. HDFS, Map Reduce, Hadoop Common, and YARN are the four major Hadoop components. *Hadoop distributed file system* maintains an enormous amount of information and facilitates appropriate access through storing redundant information throughout different fault-tolerant devices. To interact with HDFS, provided the essential command-line interface. This is designed with limited infrastructure as well as being appropriate for handling big amounts of different datasets. The Map Reduce component is termed for the two primary actions it performs such as retrieving information from a database, transforming it to a suitable format, and executing complex computations. Map Reduce takes a set of data and converts it into tuples, and the reduce job requires the map's result as an input and generates those data tuples into something like a tiny amount of tuples again until the desired outcome is accomplished. *YARN* is a cluster management system it's a significant part of Hadoop's second generation. The basic concept behind the thread is to divide resource management and monitoring work into independent modules. The goal is to have a global resource manager (RM) and a master of application for each application (AM). The Resource Manager and the Node Manager are combined in the YARN data computation architecture. The Resource Manager coordinates resources across all of the system's applications. A scheduler and application manager are included in the Resource Manager. The Scheduler allows resources to the many applications that are currently running. The Application Manager is in charge of all of the nodes' applications. Application Master and container are both parts of the Node Manager. *Apache Spark* process data unlike Hadoop, which stores intermediate data on distributed file systems, it saves data in RAM and queries it repeatedly to improve performance for certain types of applications. Spark SQL is used to work with SQL and DataFrames, MLlib is used to work with machine learning, and GraphX is used to work with graphs and parallel processing. Nowadays, a large number of major providers, such as Microsoft Azure or Cloudera, also provide Spark and Hadoop, allowing developers to choose the best structure for their data analytic applications. *NoSQL* is a database format that contains much more information that's also just structured into tables, rows, and columns, as in a relational database system. Because it is unstructured, unorganized, and does not easily accommodate conventional database frameworks, and NoSql database has now become hugely common in Big Data. Another development carried on by cloud technology in NoSQL databases for data storage and retrieval. *Apache Kafka* is a framework for building a stream-processing-based software bus. It was developed by the Apache Software Foundation as an open open-source platform in Scala and Java. The program's goal is to provide a uniform, high-flow-rate, low-latency platform for processing real-time data flows. Kafka uses a low-latency binary TCP-based protocol, as well as a message set abstraction that groups messages together organically to decrease network costs [3][4].

Cloud Computing: Cloud computing provides organizations to have more value out of their data by providing rapid analytics for a fraction of the previous cost. In essence, it enables enterprises to acquire and store even more data, driving up demand for processing power and creating a circular economy. Aside from its adaptability, Cloud Computing addresses one of the challenges with transmitting data and exchange because data is shared with others. Big Data concerns will support the advancement of cloud computing.

Machine Learning: Machine learning is a part of AI that instructs a machine how to know, enabling it to create models rapidly and automatically that can analyze larger, more complicated data and offer faster, more exact responses on a broad scale. An organization's prospects of discovering profitable possibilities or avoiding uncertainties are also improved by developing precise models. Machine Learning is a technique for creating algorithms that allow systems to evolve new behaviors and organizations to gain valuable insights.

Text Mining: Text mining is the extraction of mining text data from the internet, comment fields, books, and other text-based sources to discover previously unknown insights. Text mining classifies documents, emails, blogs, Twitter and LinkedIn posts, surveys, market intelligence, and other data sources using machine learning or natural language processing technology to store large amounts of data and identify new ideas and patterns.

Data Mining: Data mining explores enormous volumes of data patterns in the data, which may then be used for any further analysis to effectively resolve challenging business issues. It is a method for collecting useful information from data that encompasses clustering, classification, and analysis. It also looks for patterns in huge amounts of data by exploring and analyzing them.

In-memory Analytics: This technology can eliminate analytical processing delays while testing new circumstances and building models; it is not just a simple approach for businesses to stay adaptable and make smarter decisions, but it also allows them to use descriptive and predictive analytics. We can obtain immediate insights from data and respond by examining data from system memory rather than hard disc storage.

Optimization Methods: Optimization technique, as well known as optimization algorithms, is a set of mathematical processes and methodologies that are used to overcome quantitative problems in fields as diverse as physics, biology, economics, and engineering. In various areas of research, optimization methods are used to identify solutions to enhance or eliminate particular research parameters, such as minimizing expenses in the manufacture of a thing or service, maximizing earnings, minimizing raw resources in the development of a better, or maximizing productivity.

Artificial Neural Network (ANN): It is a complex data analysis and optimization technique used in pattern recognition, adaptive control, image analysis, and other applications. It is built around a well-defined architecture of interconnected artificial neurons. However, it also makes use of a distinct learning algorithm that efficiently learns from data using this human-inspired architecture. It is named an artificial neuron because it is a simplified electronic version of a real human biological nerve cell called a neuron.

Social Network Analysis (SNA): The technique of analyzing social structures using networks and graph theory is known as social network analysis. Endpoints and strong linkages are included in this key technique utilized in modern sociology for viewing social relationships. Nodes such as individual actors, individuals, or articles in the network are expressed in terms of edges and relationships to connect them.

Quantum Computing: Quantum computing is the application of powerful quantum physics laws to the processing of data. The 0 and 1 are encrypted into two independent quantum states by a quantum system. Greater quantum computers can solve particular challenges more efficiently and fast than conventional computers. Quantum computing is a kind of computer that uses the ensemble features of quantum systems to accomplish calculations, such as superposition, interference, and entanglement.

Blockchain: It is the actual database technology that supports Bitcoin digital money, and it has the unique ability to ensure that the data is never removed or updated after it has been written. It is a strong security platform that is an excellent match for various big data applications in financial services, accounting, health coverage, medical services, marketing, and other sectors of the economy. Because even though blockchain is in its early stages of development, countless enterprises, including AWS, IBM, and Microsoft, as well as entrepreneurs, have tried all kinds of experimentations to start exploring possibilities for generating blockchain[5].

Xplenty: It is a big data analytics technique that uses the cloud to integrate, analyze, and prepare data from various sources. Its straightforward graphical user interface facilitates ETL, ELT, and replication. Xplenty's toolkit allows for to creation of reduced and also no data pipelines. It would be flexible and scalable clouds with easy access to a variety of data stores and data transformation tools. Using Xplenty's significant expression language, a user can add extensive data preparation procedures.

Qubole: Qubole is a multi-cloud data lake platform that is open, simple, and secure for machine learning, streaming analytics, data exploration, and ad-hoc analytics. It is a cloud-native big data platform, which is useful at any scale because it operates on the concept of separating storage and computation, as well as supporting Apache Hadoop, Apache Spark, Presto, and TensorFlow.

SAMOA: It stands for scalable advanced massive online analysis, and it is a free and open-source analytics technique for streaming and mining that enables the creation of machine learning methods and their execution on a large number of streaming learning devices and stream processing engines. It is a user-friendly, highly scalable, quick, and free big data analytics application. It is designed around the write once, run anywhere architecture.

HPCC: It is an open-source analytical tool that provides a full huge data remedy on a very highly scalable supercomputer. It's written in C++ and Enterprise Control Language, which offers data, pipeline, and system parallelism. It supports parallel data processing and is flexible, reliable, and scalable. It's both affordable and effective.

KNIME: It is Konstanz Information Miner, which is a free and open-source analytical tool. The KNIME Data Analytics Platform is a data-driven platform that aids in the discovery of potentially hidden data, the mining of new insights, and the forecasting of the future. With thousands of functionalities, hundreds of fully prepared instances, a diverse set of interactive tools, and the most powerful set of efficient algorithms in the industry.

Weka: Weka is a collection of machine learning techniques related to data mining. The techniques can be implemented immediately to a set of data or named from Java code. Weka form an effective pre-processing, identification, regression, cluster analysis, classification techniques, as well as visualization. Along with its graphical interface, it is the easiest and quickest way for someone who hasn't coded in a long time to get started in data science.

Pentaho: Pentaho removes the obstructions that prevent organizations from optimizing the value of their data. The platform simplifies the preparation and combining of every data, and it contains a toolkit for analysis, visualization, exploration, reporting, and forecasting. Pentaho is intended to be open, embeddable, and extensible, allowing anyone on a team, from developers to enterprise clients, to adaptive capacities data into the valuation.

Talend: It is a data integration tool for extracting, transferring, and loading data. This also facilitates data planning, quality of data, integration, interoperability, data processing, and big data software tools. It gives a unique product for each of these possibilities, as well as an embedded store for collecting and reusing information. It is an expert in big data integration and offers features like cloud, big data, corporate application interaction, data quality, and data management[6].

2. Visualization and management tools

Open Refine: It is an open-source application that helps to manage and visualize unstructured data, as well as convert, expand, and improve it. It works with operating systems such as Windows, Linux, and Mac OS. Web services as well as other data can be integrated into the data using dataset linking and extension tools.

Elastic Search: It is also an open-source enterprise search engine written in Java and provided under the Apache license. It is a free analytics tool, and one of its best features is that it supports data discovery applications with extremely fast search functionality.

Tableau: It is an analytical tool that provides many integrated services to enable the biggest enterprises in exploring, visualizing, and evaluating their data. It can handle any data size and creates custom visualizations in real time, making it accessible to both technical and non-technical users. It offers excellent and quick support for connecting to a large number of databases.

Rapid Miner: Rapid Miner is a compatibility tool that integrates advanced analytics, pattern recognition, and data modeling into a unified model. It is visual programming software capable of manipulating, analyzing, and modeling data.

Lumify: It is an open-source big data analytics implementation for analyzing and visualizing massive amounts of data. This tool offers comprehensive search, two-dimensional and three-dimensional graphical viewings, automatic templates, multimedia analysis, and real-time project or workplace collaboration.

Datawrapper: It is an online tool for creating interactive data visualization charts. Many journalists and news organizations use Datawrapper to include real-time charts in their stories. It's easy to use and produces greater graphics. This can be used on any device and performs magnificently with smartphones, laptop computers, and tablets are all reasonable alternatives.

Orange: Orange is a tool for data visualization and analytical application that offers interactive workflows and a broad toolset for data analysis and visualization. It includes scattering plots, bar graphs, and trees, as well as dendrograms, networks, and geospatial data, along with other visuals.

NodeXL: It is software for visualizing and analyzing networks and relationships. It performs precise computations. It's a free and open-source network analysis and visualization program. It is one of the best statistical tools for data analysis because of its network monitoring metrics, connectivity to social networking site data distributors, and intelligent systems.

Gephi: Gephi is also an open-source network for analysis and visualization software applications based on the NetBeans platform which is written in Java. Gephi took it a step further by offering accurate calculations.

Google Fusion: This tool is used for data analysis, visualization, and mapping of huge amounts of data. Google's outstanding mapping engine, unsurprisingly, plays a large role in propelling this product to the top of the list.

Infogram: Infogram has over thirty-five interactive charts and five hundred maps to help you visually represent user data. Create a column chart, a bar chart, a pie chart, or a word cloud chart. To fully amaze our audience, we can even include a map in our infographic or report[7].

3. Sentiment Analytics Tools

Open Text: Open Text sentiment classification is a specialized classified engine that detects and evaluates subjective sentiment trends and phrases in textual data. It is done at the relevant, phrase, and documentary on the subject levels. The purpose of the test is to see if text chunks are factual or subjective, and if subjective, if the viewpoint extracted in each of those pieces of information is favorable, negligible, ambiguous, or impartial.

Semantria: It provides a unique service by collecting users' texts, Twitter messages, and other comments and methodically analyzing them to generate practical and extremely knowledgeable information. It is different from lex analytics in that it is accessible through an application programming interface and excel plugins, as well as it is having great knowledge and using deep learning.

Tracker: It is a sentiment analytics tool that examines the key phrases that are being tracked and determine whether the sentiment of the content is favorable, unfavorable, or ambiguous. Because of its rare combination of statistical approaches and rule-based natural language processing methodologies, it could be used to track all mainstream news media and social networking sites to gain insight through trends, hashtag discoveries, automatically generated sentiment classification, and impact rating.

SAS sentiment analysis: It includes produced reports, show trends, and detailed responses, as well as an innovative blend of research findings and regulation speech recognition algorithms for extracting opinions. So that we may concentrate on the emotions that are disclosed as a result of ongoing assessments. We can also update our methods and classifiers to reflect new terms and topics that are beneficial to consumers, enterprises, and industries.

Opinion Crawl: It enables users to analyze web sentiment on a specific topic, individual, incident, enterprise, or product. We can give it a topic and get sentiment analysis in real time. Each topic comprises a graph indicating contemporary attitude, a selection of relevant media articles, several sample display pictures, and a trademark network of core conceptual relevant keywords that the community correlates with the matter. Web crawlers could find the most previous study material on a wide variety of exciting subject areas and actual public concerns, as well as measure sentiment for them regularly for more in-depth analysis.

4. Data Extraction Tools

Octoparse: It is a free and detailed site tracker tool for extracting almost any type of data from a webpage. The point-and-click interface of this tool makes it simple to use for non-programmers. It allows users to use AJAX and Javascript to take all of the text from a page, enabling users to extract practically all of the contents and save them in an organized format. Scheduled Cloud Extraction is a feature that allows you to refresh the website and acquire the most up-to-date information.

Scraper: Scraper is a browser extension with limited data extraction capabilities, but it's essential to conduct web research and export data to a spreadsheet. Scraper is a browser extension with limited data extraction capabilities, but it's essential to conduct web research and export data to a spreadsheet. This tool is intended for both beginners and experts who can use OAuth to transfer files to the template or store them in spreadsheets.

Content Grabber: It is web crawling software that can collect information from practically any webpage and store it in any format we would like, including spreadsheets, reports, online services, XML files, and database systems. Because it provides many complex scripting, editing, and debugging interfaces, it is best suited to people with substantial programming knowledge[8].

DISCUSSIONS AND FUTURE SCOPE

As big data expands at an exponential rate, it places increasing pressure on organizations' storage infrastructure. Cloud computing provides storage for storing, processing, and analyzing large amounts of data. But apart from data backup, cloud backup is important because it provides hyper-scalability, cost efficiency, remote data access, and high availability. The cloud and big data can only achieve the right outcomes if they are accompanied by the appropriate tools and technology. As-a-service offerings are generally the key to ensuring that users can correctly process data in the cloud. Users may also be assured that they have all they need to keep their data strategy secure and compliant by using an as-a-service provider. Data is being generated at a rapid rate, and the diversity and quality of this data need to be analyzed. Batch processing is used for some data, while real-time analytics is used for others. Organizations can obtain better analysis from a large amount of heterogeneous data by using big data analytics in cloud technology. Kubernetes is a vendor-agnostic open-source Big Data platform for clusters and container management from Google. It's a platform for container system automation, deployment, escalation, and execution through host clusters. We can achieve anything with big data if we use the correct analytics, tools, and techniques on the cloud. Researchers can use various techniques such as data mining, machine learning, cloud computing, data stream processing, and quantum computing to efficiently handle big data analytics. The appropriate big data analytics tools and technology will help organizations interact with users on a deeper level and increase the efficiency of their teams. Since the amount of data generated will increase in the future, experts will be required to handle such complicated data, resulting in more job opportunities.

CONCLUSION

The goal of big data analytics, as mentioned in this paper, is to extract meaningful information from a vast collection of heterogeneous data. Having access to large-scale, distributed datasets, on the other hand, expands a variety of techniques for tackling the problems stated. We also discuss the various criteria for big data analytic processes in aspects of data gathering, storing, processing, and distribution. The provision of adequate systems, techniques, and strategies for application monitoring related to big data analytics can help to improve the usefulness of knowledge, but it must be done in a continuous procedure. The era of big data has brought in a significant demand for enhanced data collecting, administration, and analysis systems. Evaluating big data infrastructure tools in the context of current advances gives a good knowledge of how different tools and techniques are utilized to handle actual issues.

REFERENCES

[1]Loris Belcastro, Fabrizio Marozzo, Domenico Talia, Paolo Trunfio,” Big Data Analysis on Clouds”, University of Calabria, Rende, Italy.

[2]Albert Y., Sherif Zomaya, Sakr Sartaj, Sahni, ”Handbook of big data Technologies”,

[3]G. Bharadwaja Kumar, ”An Encyclopedic Overview of ‘Big Data Analytics”, School of Computing Science & Engineering, VIT University, Chennai - 600127, India.

[4]Alkhatib, Ahed J, Shadi Mohammad Alkhatib, and Hani Bani Salameh. ”Prediction of Big Data Analytics (BDA) on Social Media: Empirical Study.” ISSN: 2393-1744, vol. 7, issue 1 (November 2020).

[5]Lynda Kacha, Abdelhafid Zitouni, ”An Overview on Data Security in Cloud Computing”, Lire Labs, Ali Mendjli, 25000 Constantine, Algeria.

[6]Arista, A. Arun Gnanaraj and J. Gnana Jayanthi,” A Comprehensive Study on BigData Analytics-Tools, Techniques, Technologies and Applications”, International Science Press.

[7]D. P. Acharya, S. Dehuri, and S. Sanyal, “Computational Intelligence for Big Data Analysis”, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.

[8]Amira S. Ashour, Chintan Bhatt, Nilanjan Dey, Simon James Fong,” Healthcare Data Analytics and Management”, ISBN: 9780128156360, 0128156368,15 November 2018:Elsevier Science.