Big Data Analytic Tools, Applications and its Significance

[1]Khalid Hasan Mahmoud Alsinjlawi

[1]Lecturer
[1]Department Computer Science
[1]Jazan University, Jazan, Kingdom of Saudi Arabia

Abstract: Big data analytic tools are used to extract information from large amounts of data, which can be an overwhelming process for humans. These tools allow you to identify trends and patterns in your data and make predictions about future behaviour based on what you learn from your past actions. Data analytics tools are used to help us understand the meaning of large amounts of data. There are many different types of big data analytic tools, each using different algorithms and techniques to gain insight. The most commonly used big data analytic tools are artificial intelligence (AI) and machine learning (ML). AI is a subset of big data analytics that uses algorithms to learn from past experiences, in order to make decisions in the future. ML is the use of computers to find patterns in datasets by analyzing data and making associations between variables. Big data analytic tools allow companies to collect, process and analyze large amounts of data. The size of the data sets can be in petabytes, exabytes, or even zettabytes. The tool helps companies to make decisions based on the analysis of this data. This paper discusses the different big data analytic tools, their applications, features and availability to help organizations make better decisions.

Index Terms: Big data, analytics tools, artificial intelligence, machine learning, datasets

I. INTRODUCTION

Big data analytic tools are an important part of the emerging field of data science. Big data analytics is the process of exploring and analysing large datasets to uncover hidden insights, trends, and associations. It relies on advanced data mining, machine learning, and statistical techniques that can be applied to a wide range of industry sectors including healthcare, financial services, retailing and more. These tools are used to analyse large amounts of unstructured data [1]. These tools are used in many different fields and industries, including healthcare and government. This tool helps to uncover patterns and trends from large sets of data. The analysis can then be used to develop new methods or make predictions about future events. This has become a key component of many organizations' business intelligence initiatives. They allow for large-scale analysis and processing of massive amounts of data, which is crucial to the development of new insights, models, and algorithms. Big data analytic tools can be used in a wide variety of applications including machine learning, natural language processing, computer vision, and speech recognition. These tools can also be used to aid in research on topics such as human-computer interaction or human-robot interaction. They can also be used to make decisions or recommendations based on the analysis of large amounts of data.

No matter how small or large your data is, there is a tool for you to find patterns and visualize them to make better decisions. Big data analytics tools can serve as an easy-to-use platform that allows almost anyone to query their organization's novel, structured or unstructured datasets and gain insight from those sets with minimal programming expertise. The choice of which tool to use can be overwhelming, but choosing the right one can save time and money while also improving the quality of your analysis. While there are many big data analytic tools available, each one has its own set of features that make it unique. Some tools may offer greater flexibility than others, and some may be more focused on specific types of analysis. However, there are also some major differences between different tools that can affect how you choose which one is right for you.

One important consideration is whether or not the tool supports multiple sources; some only support one source at a time, while others can handle multiple sources at once. Another factor is whether or not the tool allows users to collaborate on their work; this is particularly relevant when working with large datasets because it allows multiple people who have different skillsets (e.g., programmers vs statisticians) to contribute without having to wait for someone else's approval before continuing their work on a project together with colleagues from across departments who might not otherwise interact with each other outside of meetings.

Big data analytic tools are a method for large-scale data processing. The architecture of these tools is based on the concept of a distributed system, where all the tasks are split into several parts that are processed in parallel. The features of this kind of architecture include scalability, fault tolerance and load balancing. The main feature of big data analytic tools is their ability to process massive amounts of data in parallel, which allows them to handle extremely high workloads. They also have built-in fault tolerance and load balancing capabilities, which means they can handle failures gracefully and make sure that all of your tasks get completed efficiently. Big data analytics is a rapidly growing field that uses complex algorithms to process large amounts of data.

As the Internet of Things (IoT) grows and expands, so does the amount of data that needs to be processed by big data analytic tools. Big data analytic tools include the following but are not limited to Hadoop, Spark, Python, R, and Julia [1]. These tools are used for analysing large amounts of data stored in multiple formats and locations. The architecture of these tools includes the ability to distribute workloads across multiple machines using a cluster or grid computing model. This allows them to handle large amounts of data without overwhelming individual machines.

Big data analytic tools can be divided into two types: batch processing and stream processing. Batch processing is used for offline analysis, while stream processing is used for online analysis. While both types of big data analytic tools offer advantages over

traditional analytical methods, stream processing offers some unique advantages over batch processing. Stream processing allows users to interact with their data in real-time, which is not possible with batch processing. It also reduces the need for pre-processing steps or lengthy queries that would otherwise be required by batch-based systems.

The paper is sectioned as follows. It begins with an introduction to big data and then reviews the literature on the various different tools and techniques that can be used to analyse big data, including analytics, artificial intelligence, machine learning, and natural language processing. The paper concludes with an analysis of each tool's limitations and how these limitations might be overcome through further research.

II. BIG DATA ANALYTICS TOOLS

Big data analytical tools are a major breakthrough in the field of data analytics. They are used to solve big data problems by analysing the large amount of data that is generated by organizations. There are many tools available to help you with your big data analytics projects. Some of these tools include:

Hadoop [3] is an open-source framework licensed under the Apache License for processing large amounts of data, which is distributed across multiple machines. It is used for storing data in a distributed manner over several machines so that it can be accessed quickly by any user at any time without much delay or lag in performance. It consists of various modules that can be used independently or together. Apache Hadoop is an open-source software framework for storing and processing large quantities of data on computer clusters built with commodity hardware. Features of Hadoop include distributed processing, scalability, fault tolerance, and inexpensive storage. Its components include the Hadoop Distributed File System (HDFS), YARN, and MapReduce. The source code for Hadoop is available for download from the Apache Software Foundation. Hadoop is the community name given to Apache Hadoop, which is an open-source implementation of the Map-Reduce framework for distributed data processing over clusters of commodity hardware.

MapReduce [4] is a programming model for processing large amounts of data using parallel processing on multiple machines (clusters). MapReduce is an open-source project, which means that its source code is available for free use. Users of the Hadoop platform can freely download the underlying software platform from the Apache Software Foundation website. MapReduce is open source and is available in Apache Hadoop as a subproject.

Tableau [5] is an analytics tool that makes it easy to explore and visualize data in many formats – whether you're dealing with spreadsheets, databases or specialized applications. With powerful features like drill-through, fast filtering and simple data blending, Tableau allows you to answer the questions that matter most. Tableau can be beneficial because it allows you to leverage your existing data, and provides the option of getting new data. The overview page is designed to make some of the other features more user-friendly. It allows you to access your dashboard and data sources, as well as even search through your interactive visualizations. Tableau allows you to create easy-to-digest reports, charts, and graphs that don't require too much technical knowledge or coding experience in order to use the content effectively. The strengths of this tool are its easy installation, user-friendliness, and visual representation. From a weakness standpoint, users have reported difficulty in performing very complex queries that require multiple steps.

Power BI [6] is a cloud-based data analytic tool that provides companies with the capabilities and benefits of an enterprise analytical platform, without requiring large teams of business analysts. Power BI enables users to explore and visualize data from multiple sources on any device without requiring developers or IT resources. Users can work with data from hundreds of SaaS applications, including Azure SQL Database and Relational Database Service (RDS). Users can also create their own datasets using PowerShell scripts or by importing Excel files into the Power BI desktop application. The visualizations are very flexible, as they allow for multiple combinations of data visualization types. This can help you find interesting insights even in the most complex data sets if you know what questions to ask of your data.

Apache Spark [7] is an open-source framework for processing large amounts of data, which is distributed across multiple machines (and even multiple clusters). It can run on top of Hadoop or other frameworks like Storm or Flink to provide a fast execution engine for complex operations like machine learning algorithms. The components of Spark are Spark SQL, Apache Spark core, Spark Streaming, Mlib, Graphx [8]. This tool is used for machine learning and data analysis. It uses an in-memory cluster manager to process large amounts of data quickly. Spark has been used by some large companies such as Yahoo! and IBM.

Apache Hive[9] helps users query large datasets stored on distributed storage systems like HDFS (Hadoop Distributed File System). Hive makes use of SQL queries so that users can easily interact with their data without having any knowledge of how it was stored or indexed (e.g., using a key/value pair).

Splunk [10]is a tool which is used to search, monitor and analyse data. The main features of Splunk consist of indexing, searching, a dashboard and reports.

Flink [11]is a distributed processing framework following a layered architecture. It can resolve accurate and approximate results from bounded and unbounded data streams. Flink programs are converted into data flow graph. Flink is similar to Spark in that it can be used for both batch processing and stream processing. It also allows users to write applications using Java or Scala instead of Python or R which is common with other tools like Spark or Hadoop MapReduce (HMR). Flink has been used by companies such as Uber and eBay.

Tools	Features	Strengths	Limitations	Programming support	Open source or free
Power BI	Cloud-based Drag and drop	Easy to use Faster User friendly Interactive dashboards	Difficult to handle big data Performance issues Limited sharing of	M Language DAX (Data	1 GB data storage in the free version Not open source

		Handle complex data sets	data Complex in nature	Analysis Expressio n)	
Tableau	In memory data Data sources available	User friendly Non-technical users	not supports uncleaned data Lacks data modeling	C C++ Java Python	Free but not open source
Apache Storm	Process big data Reliable paralleliz able	Process big data Fast Reliable Scalable and fault tolerant	Tedious Difficult to scale	Ruby Python	Open source
Apache Hadoop	Distribut ed file system Secure storage	Scalable Large scale processi ng of data sets	No real time processin g File concerns Slow processin g speed	R Python Ruby	Open source
Apache Spark	Supports sophistic ated analysis	Big data workloads Flexible Fast	No real time processing Latency Iterative	Scala Java R Python	Open source
	Batch analysis	processing	processing		
Splunk	Searching Create user interfaces	Interactive charts Scalable Easy to implement	Less reliable Expensiv e Difficult to optimize the search	Python	Not open source
Flink	• High throughput • Parallel computations • Stream and batch processing	Low latency Process in real time Good perform ance level	Fault tolerant Memory managem ent limitation s	Java Scala	Open source
MapReduce	Manage structured and unstructured data	Fast Scalable Parallel processing	No real time processin g Technical expertise	R Python	Open source
Apache Hive	Manage structured data	Scalable Fast Cost efficient	Supports only online analytics No delete operation	Java Ruby Python	Open source

III. APPLICATIONS

The following are some of the applications for big data analytic tools [2]:

Information Retrieval: This involves finding the most relevant information from large databases or collections of documents in order to answer a question or solve a problem. It's often used by companies looking for relevant keywords when creating ads online or by researchers who want to find papers related to their topic area quickly so they can start reading them right away instead of having to search through many different sources individually first before figuring out which ones might be most useful for their current needs).

Predictive Analytics: This involves analysing historical data sets in order to predict future events based on trends within those datasets (i.e., which customers are likely to buy more products over time based on what type of products they've already purchased before). This technique is popularly used by companies like Amazon.

The major strengths mainly include allowing for the exploration of data in a way that is not possible with other tools. It is able to handle large datasets and can be used on diverse sources of data (e.g., unstructured, semi-structured, structured). The key weakness is that these tools are not as easily accessible as some others, which may limit their ability to be utilized by a broad range of users. In addition, they require extensive expertise in statistics and data science.

IV. DISCUSSION

As organizations begin to integrate big data analytics into their operations and business processes, it is important to consider the different tools available for analysis. Each tool has its own features, strengths and weaknesses. However, before we can make an informed decision about which tool is best suited for our needs, we must first understand what these tools are capable of doing and how they are used in an organization. Table 1 shows the different tools in terms of features, strengths, weaknesses, and availability.

Cyber-attacks, data security and privacy are major concerns in big data analytics. Some of the tools still need to improve the processing of data files. Constantly changing data needs systems to handle computing, analyze data and difficulty in determining good data.

V. CONCLUSION

Big data analytic tools are essential for companies that want to extract insights from their vast collections of data. This article aims to provide an overview of big data analytic tools and discuss their research directions. Big data analytic tools can be used to analyse large amounts of data to identify patterns, trends, and insights. This information can then be used to make decisions that affect the organization's operations. Big data analytics tools can be used for a variety of purposes, such as improving business efficiency, detecting and preventing fraud, and improving customer service. They can also be used to improve decision-making and to improve the accuracy of predictions. Big data coupled with artificial intelligence might improve the inconsistency in data or missing data

References

1. "IBM," [Online]. Available: https://www.ibm.com/analytics/big-data-analytics.
2. "Simplilearn," [Online]. Available: https://www.simplilearn.com/what-is-big-data-analytics-article. [Accessed 2022].
3. C. V. a. S. T. S. Rajkumar Buyya, Mastering Cloud Computing, ScienceDirect, 2013.
4. O. I. K. Y. A. &. S. A. Surendran Rajendran, "MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network," *Scientific Reports,* vol. 11, 2021.
5. T. G. H. T. Steven Batt, "Learning Tableau: A data visualization tool," *The Journal of Econoimc Education,* vol. 51, no. 3-4, pp. 317-328, 2020.
6. P. T. a. F. Thabtah, "Data Analytics Tools: A User Perspective," vol. 18, no. 1, 2019.
7. A. Mahmood, "IBM," 2021. [Online]. Available: https://developer.ibm.com/articles/introduction-to-big-data-analysis-withpyspark/.
8. R. D. X. C. P. X. P. &. J. Z. H. Salman Salloum, "Big data analytics on Apache Spark," *International Journal of Data Science and Analytics volume,* no. 1, pp. 145-164, 2016.
9. A. C. ,. A. G. ,. E. K. ,. O. O. ,. V. G. ,. Z. H. ,. S. S. ,. P. J. ,. S. S. ,. D. J. ,. S. B. ,. N. B. Jesús Camacho-Rodríguez, "Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing," in *SIGMOD '19: Proceedings of the 2019 International Conference on Management of Data*, 2019.
10. K. P. B. H. B. B. B. P. Balaji N, "Data Visualization in Splunk and Tableau: A Case Study Demonstration," in *Journal of Physics: Conference Series*, 2020.
11. M. K. a. M. M. Yousaf, "A Comparative Analysis of Big Data Frameworks: An Adoption Perspective," *Applied Sciences,* vol. 11, no. 22, 2021.
12. "Influence of big data in smart tourism," in *Hybrid Computational Intelligence*, 2020.