*Subodh Journal of Recent Trends in Information Technology*
*Volume 10, Issue-02, June 2019* |ISSN No. 0975-9875

# Critical Analysis of Big Data Analytics Tools

**Ms. Suman Choudhary**
*Scholar, Deptt of Computer Science and Application, MA University, Jaipur*

**Dr. Mahaveer Sain**
*Professor, Deptt of Computer Science and Application, MA University, Jaipur*
*mahaveersain@gmail.com*

**ABSTRACT -** Billions of people upload and share information to social media and other platforms every day via mobile phones, laptops and PDAs. This information includes images, locations, goggle map locations, videos, text, and voice messages. These are collections of structured, unstructured, and complex data objects. Traditional data processing techniques are insufficient to process this large, heterogeneous, fast paced data. In recent years, the popularity of electronic commerce and digital marketing has increased, and the business industry is increasingly dependent on online transactions and services. Big data analytics is beneficial for such industries because it helps to extract useful patterns and unknown correlations from potential data sources, consumer preferences, purchase attributes, and many other information has been proven. The purpose of this article is to analysis, review and compare the latest tools for big data analysis.

**KEYWORDS -** *Big Data, Big Data Analytics, Big Data Analytics Tools and Cloud Computing.*

## 1. INTRODUCTION

Big Data Analytics looks at different datasets to help unknown business relationships, hidden patterns, market trends, customer choice, and organizations make business decisions from more data analytics. The process of finding the most useful information. Data can be processed very quickly and efficiently. This includes analysing the data and using the results. This improves efficiency by reducing workloads not possible with traditional business intelligence solutions. With digitalization, data sets are growing rapidly in different ways. If the data or set of information is too large and complex, traditional data processing techniques cannot handle complex data, and data is called big data. Researchers, scientists, business organizations, government agencies, advertising agencies and medical researchers are often more difficult to process decision data. The data available for research should be processed using various data analysis techniques, called large data analysis. These technologies are rapidly changing and benefit from the processing of large amounts of unstructured, structured, or semi-structured data that cannot be processed using traditional database technologies. This article explains the main goals of big data analytics by comparing the various tools available for high data validation.Big data analytics tools are important for businesses and large companies because modern organization tools that use big data tools generate and manage large amounts of data. Great analytical data tools help companies save time and money and gain insight into data-driven decisions [5]. Big Data Analytics is the complete process of collecting, organizing, and analysing large amounts of data (called Big Data) to feed and identify patterns and other useful information needed for business decisions. Large-scale analysis helps organizations better understand the information in the database. People who work with big data analysts often have knowledge of data analytics. The dataanalysis process includes various types of tools such as data analysis, data cleaning, data mining, data visualization, data integration and data storage and management [6].Large data analysis processes require very high performance analysis. Therefore, analysing such large amounts of data requireslarge data analysis processes and application-specific software tools for predictive analysis, data mining, text mining and data prediction and optimization.

## 2. LITERATURE REVIEW

Big data analytics tools are important for enterprises and large enterprises because modern organization tools that use big data tools generate and manage large amounts of data. Big data analytics tools can help companies save time and money and gain insights to make decisions based on data [4].Big data analytics processes require very high performance analysis. Therefore, analysing such large amounts of data requires big data analysis processes and application-specific software tools for predictive analytics, data mining, text mining, forecasting, and data optimization.

The work of [13] **Jung D, Shi. 1. Big Analytical Data: A Literature Review. J Manag Anal** presented an overview of various data analysis tools, technologies and methods by categorizing large literature on analytical data according to their research focus. This paper is different in that it presents a systematic review of the literature, which

focuses on large-scale analyses to "stream" data.**Akter S, Fosso WS, Big data analytics for e-commerce: a systematic review and future research agenda, 2016** [14] authors presented a systematic review of big data analytics in e-commerce. The study explored the characteristics, definitions, business values, types and challenges of big analytical data in the e-commerce landscape.

**Sivarajah U, Kamal MM, Irani Z., Verakkodiya V. Critical analysis of large data challenges and analytical methods. J Bus Res, 2016** also [15] conducted a study that focuses on large-scale data analysis in technology and organizational resource management, specifically focused on reviews that present major data challenges and large data analysis methods. Although they are systematic reviews, the focus is not particularly on large data transmission. Authors [16] **Wienhofen LW, Mathisen BM, Roman D. Empirical Big Data Research, 2015**, presented the status of areas of empirical research and big data application using a systematic mapping method. In the same vein, the **authors Habab RAA, Nassarudin J, Ghani A, Hashim IAT, Ahmed E, Imran M. 2018** [17] also conducted research on large data technologies and machine learning algorithms with a particular focus on anomaly detection. A systematic review of the literature, which aims to determine the scope, application, and challenges of large analytical data in health, was presented by **Host M, Orucevic-Alagic A, 2013** [20].The work of **Sun D, Zhang G, Zheng W, Li K. Key technologies for large data flow for computers, Big Data Algorithms, Analytics and Applications, 2015** [2] presented an overview of four major data streaming tools and technologies. While the study conducted in this paper provided a comprehensive overview of not only the major data streaming tools and technologies, but also the methods and techniques used in analysing large streams of data. In addition, the authors [2] did not provide a clear explanation of the methodological approach for selecting the papers reviewed.

## 1. Methodology of Big Data Analytics [5]

This section describes the various life cycle stages of big data analysis -

**Data identification and collection -** At this stage, various data sources are determined based on the severity of the problem. More data resources mean more opportunities to find hidden correlations and patterns. Tools are needed to capture keywords, data and information from these heterogeneous data sources.

**Data Storage-** Captured structured and unstructured data needs to be stored in a database/data warehouse. A NoSQL database is needed to accommodate big data. Organizations such as Apache, Oracle, and others have developed various frameworks and databases that allow analysis tools to retrieve and process data from these repositories.

**Data Filtering and Noise Cancellation-** This phase is dedicated to removing duplicate, corrupt, empty, and unrelated data objects from the collected information. However, filtered and deleted data may be of some importance in another context or analysis. Therefore, it is recommended to keep a copy of the original data set in a compressed form to save storage space.

**Data Classification and Extraction -** This phase is responsible for extracting inconsistent data and transforming it into a common data format that the underlying analysis tools can use for their purposes. This may also involve extracting related fields or text to reduce the amount of data to submit to the analytics engine.

**Data Cleansing, Validation, and Summarization -** this phase applies business case-based validation rules to confirm the necessity and relevance of extracting data for analysis. Although due to complexity, it may sometimes be difficult to apply validation constraints to the extracted data. Aggregation helps to combine multiple data sets into fewer numbers based on common fields. This simplifies further data processing.

**Data Analysis and Processing -** This phase performs actual data mining and analysis to create unique and hidden patterns to make business decisions. Data analysis techniques may vary from scenario to scenario, ie exploratory, confirmatory, predictive, prescriptive, diagnostic or descriptive.

**Data Visualization** - This phase involves representing the results of the analysis as visual or graphical to make it easier to understand the audience.

## 2. Big Data Analytics Key Technologies[6]

As mentioned earlier, the big data analysis process is not a single activity that involves a large amount of data. Instead, it can be applied to advanced analysis of big data, but in practice, several types of different technologies work together to extract maximum value from information. There are a number of big data analysis tools, and some of the key tools for storing and analysing big data are listed below:

**Apache Hadoop:**
Apache Hadoop, a big data analysis tool, is a Java-based free software framework. Helps you efficiently store large amounts of data in storage locations called clusters. A special feature of this framework is that it runs on the cluster in parallel and can handle large amounts of data on all nodes. Hadoop has a storage system, commonly referred to as Hadoop Distributed File System (HDFS), that helps you split large amounts of data and distribute it across many nodes in the cluster. It also performs a data replication process within the cluster, providing high availability and failover. This improves fault tolerance.

**Features:**
• Improved authentication when using an HTTP proxy server
• Hadoop compatible file system initiative specifications
• Support for POSIX style file system extended attributes
• Provides a robust ecosystem suitable for meeting developer analytical needs
• Brings flexibility to data processing
• Faster data processing possible

**KNIME**

The KNIME analytics platform is one of the leading open solutions for data-driven innovation. This tool can help you discover the potential and concealment of large amounts of data, provide new insights, and predict new futures. KNIME analytics platform tool is a very useful toolkit for data scientists.

**Features**
- Simple ETL operation
- Perfect combination with other technologies and languages.
- Rich set of algorithms.
- Highly available and organized workflow.
- Automate a lot of manual work.
- There are no stability issues.
- Easy to set up.

**OpenRefine:**
OpenRefine was introduced as Google Refine. This tool is one of the effective tools for handling messy and large amount of data, such as data clean-up, conversion of data from one format to another, and performing extension using web services and external data One. Open refine tools make it easy to explore large datasets.

**Features:**
- OpenRefine tool makes it easy to explore largedatasets.
- Can be used to link and extend datasets with various web services.
- Import data in various formats
- Explore data sets in seconds

Apply basic and advanced cell conversion
Can handle cells with multiple values
Create instant links between datasets
- Automatically identify topics using named entityextraction of text fields
- Perform sophisticated data manipulation using sophisticated languages

**Cloudera**
Cloudera is the fastest, simplest and most secure big data platform. This allows anyone to retrieve any data in any environment within a single scalable platform.

**Features:**
- High performance analysis
- Provide multi-cloud provisioning
- Deploy and manage Cloudera Enterprise across AWS, Microsoft Azure, and Google Cloud Platform
- Launch and exit the cluster, pay only when needed
- Data model development and training
- Business intelligence reports, surveys, self-service
- Providing real-time insights for monitoring and detection
- Conduct accurate model scoring and delivery

**RapidMiner:**
RapidMiner tools work using visual programming to manipulate, analyse, and model data. The RapidMiner tool enables all the work of open source platforms (machine learning, data preparation, model deployment, etc.), making the data science team easier and more efficient. With the unification of data science platforms, accelerating the creation of complete analytical workflows in a single environment can significantly improve the efficiency and short-term value of data science projects.

**Features:**
- Allow multiple data management methods
- GUI or batch processing
- Integration with internal database
- Interactive and shareable dashboard
- Big data predictive analysis
- Remote analysis processing
- Data filtering, merging, joining, and aggregation

- Prediction model construction, training and validation
- Save streaming data in many databases
- Reports and triggered notifications

**Table1: Comparative Evaluation of Big Data Analysis Tools**

| Data Analysis Tool | Platform | Verdict | Pricing Model | Customer Types | Deployment |
|---|---|---|---|---|---|
| Apache Hadoop | Cross-Platform | Open-source software for reliable, scalable, distributed computing. | Free | For small, medium and large enterprises. | Cloud Hosted<br>Open API |
| KNIME | Windows, Mac, Linux. | Works with Microsoft Azure and AWS. Easy to learn software. | Quote Based | For small, medium and large enterprises. | On Premise |
| OpenRefine | Microsoft Windows,Linux,MacOS | Open refining tools make it easy to explore large data sets. | Free | For small, medium and large enterprises. Freelancers | On Premise<br>Open API |
| Cloudera | Windows,Mac, Web-based, | Fastest, simplest and most secure big data platform. | Quote-based | For small, medium and large enterprises. | Cloud Hosted |
| Rapid Miner | Cross-platform | System is easy to use. Powerful GUI. Five products to choose from. | Free Annual Subscription Quote-based | For small, medium and large enterprises. | On Premise<br>Open API |

## 3. Some of the Examples/areas using Big Data Analytics Tools:

Big data analytics tools are in great need of business/enterprise. These tools rely on fast, agile decisions to stay competitive, and in most cases, big data analytics tools are critical to business decisions based on their previous business data. Here are some different types of organizations that can use this technology:

**Travel and Hospitality:**
Maintaining customer satisfaction is a very important and critical factor in the travel and hospitality business, but it is difficult to measure customer satisfaction. For example, in resorts and casinos, they will have a short chance to reverse the customer experience. Therefore, big data analytics applications can collect customer data and apply statistical analysis to better understand and improve these businesses.

**Retail:**
Today, customer service has become a huge tree compared to the past few decades, and knowledgeable shoppers are constantly searching and expecting retailers to accurately understand what they want and when they need it. Here, big data analytics technology emerges to help retailers meet customer needs.

**Government:**
Few specific government agencies always face some major challenges, such as how to prepare a budget for the public without any compromise on quality or productivity. As a result, many organizations use big data analytics; this helps them streamline operations while allowing organizations to more accurately understand criminal activity and avoid preparing a viable and good budget.

**Health care:**
Big data analytics can also be used in the healthcare industry. Maintain patient records, their insurance information and health plans, and all other types of information that are difficult to manage. Therefore, the application of big data analytics technology in the healthcare industry is very important.

## 4. Final Words:

In the current situation, as the world's population grows and technology grows, the amount of data is also growing. This is a clear sign/sign of the growing use

and necessity of big data analytics solutions. Big data is not just a technology trend, but a business practice that helps industries/companies maintain a competitive world, enabling their proactive data-driven business decisions to improve sales and marketing team performance and increase revenue.

## 5. Scope for Further Research

The current study includes only five different big data analysis software. However, further research is possible, including the inclusion of more software for comparison purposes. In addition, in addition to predictive analytics, other techniques used in big data analytics (using unsupervised learning tests) can be used to compare software. In addition, different data sets are available for further research.

## REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, "Big data: a survey", Mobile Networks and Applications, vol. 19, No. 2, pp. 171–209, 2014.

[2] Sun D, Zhang G, Zheng W, Li K. Key technologies for big data stream computing. In: Li K, Jiang H, Yang LT, Guzzocrea A, editors. Big data algorithms, analytics and applications. New York: Chapman and Hall/CRC; 2015. p. 193–214. ISBN 978-1-4822-4055-9.

[3] T. Erl, W. Khattak, and P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Prentice Hall, India, pp. 65-88, 2015.

[4] N. Khan et. al, "Big Data: Survey, Technologies, Opportunities, and Challenges", The Scientific World Journal, vol.2014, Issue.4, pp.1-18, 2014.

[5] Online source, [Available] https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/, 2018.

[6] Online source, [Available] https://www.guru99.com/big-data-tools.html, 2018.

[7] Online source, [Available] https://www.octoparse.com/blog/yes-there-is-such-thing-as-a-free-web-scraper/, 2018.

[8]https://data-flair.training/blogs/apache-storm-vs-spark-streaming/

[9] A. Narang, "A review-Cloud and cloud security", International journal of Computer Science and mobile Computing, vol. 6, issue 1,pp. 178-181, 2017.

[10] Oussous A, Benjelloun F, Lachen AA, Belfkih S. Big data technologies: a survey. J King Saud Univ Comput Inform Sci. 2018; 30:431–48.

[11] S Kaushal, J.K. Bajwa, "Analytical Review of User Perceived Testing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, 2012.

[12] S. M. Ali et.al, "Big Data Visualization: Tools and Challenges", 2nd International Conference on Contemporary Computing and Informatics,2016.

[13] Chung D, Shi H. Big data analytics: a literature review. J Manag Anal. 2015;2(3):175–201

[14] Akter S, Fosso WS. Big data analytics in e-commerce: a systematic review and agenda for future research. Electr Markets.2016; 26:173–94.

[15] Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of big data challenges and analytical methods. J Bus Res. 2016; 70:263–86.

[16] Wienhofen LW, Mathisen BM, Roman D. Empirical big data research: a systematic literature mapping. CoRR, abs/1509.03045. 2015.

[17] Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M. Real-time big data processing for anomaly detection: a survey. Int J Inform Manage.2018; 45:289307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006.

[18] Mehta N, Pandit A. Concurrence of big data analytics in healthcare: a systematic review. Int J Med Inform. 2018; 114:57–65.

[19] S. Mujawar, S. Kulkami, "Big Data: Tools and Applications", International Journal of Computer Applications, vol. 115, No. 23, pp. 7-11, 2015.

[20] Host M, Orucevic-Alagic A. A systematic review of research on open source software in commercial software product development. 2013.