iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70585

A Comparative Study of Big Data Analytics Tools: Hadoop, Spark, TensorFlow, SAS, and Tableau in Modern Data Pipelines

Kiran Vishwakarma[1], Jasleen Kaur[2], Dr. Goldi Soni[3]
[1]*Amity Institute of Information Technology, Amity University Chhattisgarh, Raipur, C.G - 493225, India*
[2]*Amity Institute of Information Technology, Amity University Chhattisgarh, Raipur, C.G - 493225, India*
[3]*Amity School of Engineering & Technology, Amity University Chhattisgarh, Raipur, C.G - 493225, India*

Abstract: With the exponential growth of data in today's digital world, proper analysis has become crucial for extracting meaningful insights that enable predictive capabilities. These capabilities prove invaluable across industrial and business sectors for informed decision-making, risk assessment, and strategic planning. Big Data Analytics serves as the cornerstone of this process, employing specialized tools and techniques to handle massive, complex datasets characterized by volume, velocity, variety, veracity, and value. This paper examines the pivotal role of Big Data Analytics and its transformative impact across various domains. We focus on five powerful analytical tools - Hadoop, Spark, SAS, TensorFlow, and Tableau - each addressing distinct aspects of the data processing pipeline from storage and computation to advanced modeling and visualization. The study provides a comprehensive comparative analysis of these technologies, evaluating their architectural differences, performance efficiency, and real-world applicability in key industries including finance, healthcare, e-commerce, and artificial intelligence.

Keywords: Big Data, Data Analytics, Hadoop, Spark, SAS, TensorFlow, Tableau.

I. INTRODUCTION

Big data encompasses large and complex datasets that traditional analytical methods are unable to process effectively. It is described as an aggregation of enormous datasets, which may include structured, unstructured, or raw data produced in today's digital landscape by governments and businesses. The defining features of big data are often summarized by the 5 Vs: Volume (large scale), Velocity (rapid generation), Variety (various data types), Veracity (issues related to data quality), and Value (the process of deriving significant insights).

According to the definition given by the scientist Gartner “Big Data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision-making, and process automation." Conventional analysis methods are inadequate for managing this data; it requires sophisticated techniques that can identify hidden patterns to enhance predictive accuracy and support real-time decision-making. Big data analytics plays a crucial role in managing and analyzing these datasets, allowing for the extraction of the valuable insights.

This approach allows for the identification of trends, patterns, predict future outcomes, and assess potential risks and correlations in vast amounts of raw data, which helps analysts make decisions based on data-driven insights. It also helps in operational analysis, financial planning, customer analysis and much more. By examining historical data alongside current market trends, organizations can develop strategic plans that highlight both advantages and disadvantages. This factor necessitates the use of specialized tools to effectively store, process, analyse, and visualize extensive datasets. Below are five crucial tools commonly employed in the realm of Big Data analytics:

1) *Hadoop*: Hadoop is an open-source framework that facilitates the distributed storage and processing of large volumes of data across numerous machines. It utilizes the Hadoop Distributed File System (HDFS) for storing substantial datasets and the MapReduce programming model for parallel data processing, which enhances its scalability. Additionally, Hadoop can integrate with various other Big Data tools such as Hive, Pig, and HBase, thereby improving its data management functionalities. It is extensively utilized in sectors such as finance, healthcare, and retail for data storage, ETL (Extract, Transform, Load) operations, and large-scale data analysis. [4]

2) *Apache Spark*: Apache Spark is a robust, open-source data processing engine tailored for both real-time and batch processing. In contrast to Hadoop, Spark employs in-memory computing, which greatly accelerates data processing speeds. It accommodates several programming languages, including Python, Java, Scala, and R, making it a flexible option for data scientists and engineers. Spark integrates with MLlib (machine learning), SQL, GraphX (graph processing), and Spark Streaming, enabling users to conduct complex analytics with efficiency. Its rapid data processing capabilities render it particularly suitable for applications such as fraud detection, recommendation systems, and real-time analytics. [3] [6]
3) *SAS (Statistical Analysis System)*: SAS is a robust statistical software suite that is extensively utilized for advanced analytics, data management, and predictive modelling. It offers a comprehensive array of tools for data mining, machine learning, and analytics driven by artificial intelligence, enabling organizations to derive significant insights from large datasets. SAS is capable of processing both structured and unstructured data, rendering it applicable across various sectors, including finance, healthcare, and marketing. Companies employ SAS for activities such as risk assessment, fraud detection, and business intelligence reporting, facilitating informed decision-making with precision and effectiveness. [3]
4) *TensorFlow*: Tensor Flow is an open-source framework for machine learning and deep learning, created by Google. It is commonly applied in Big Data contexts that require artificial intelligence and deep learning capabilities, including image recognition, natural language processing (NLP), and predictive analytics. TensorFlow is designed for scalability and adaptability, offering support for both CPU and GPU acceleration to achieve high-performance computations. It is integral to numerous AI technologies, such as autonomous vehicles, voice-activated assistants, and automated recommendation systems. Its proficiency in managing extensive unstructured data makes it an essential resource in contemporary data science and artificial intelligence initiatives. [9]
5) *Tableau*: Tableau stands out as a premier data visualization platform that empowers users to examine and showcase insights derived from Big Data via engaging and user-friendly dashboards. Its drag-and-drop functionality facilitates the creation of real-time reports, eliminating the need for advanced technical expertise. Tableau is compatible with a wide range of data sources, such as Hadoop, SQL databases, and cloud storage, which allows for smooth data integration. Organizations leverage Tableau for purposes such as business intelligence, financial reporting, and performance tracking, thereby simplifying data analysis and enhancing accessibility for decision-makers. [8]

The main objective of this study is to enhance understanding of the significance of these tools and their respective areas of application. This understanding will enable them to choose the most appropriate tool to meet their needs. These tools are widely adopted across different industries, companies, and enterprises to analyze large datasets and derive meaningful insights for their growth. the demand of these tools varies based on specific requirements. The following comparative analysis will hen to use each tool effectively.

TABLE I
COMPARISON OF BIG DATA W.R.T IMPORTANCE OF THE TOOLS

Characteristics	Hadoop	Spark	SAS	TensorFlow	Tableau
Developers	Apache Software Foundation	Apache Software Foundation	SAS Institute	Google Brain Team	Tableau Software (Acquired by Salesforce)
Importance	Foundation of distributed storage	Unified analytics engine	Gold-standard regulated analytics	Leading AI/ML framework	Business intelligence democratization
Use Case	Big Data + Batch	Big Data + Speed	Statistical analysis	Deep Learning & AI	Data Visualization & BI Dashboards
Programming Language	Java, Python	Python, Java, R, SQL	SAS language	Python, C++	
Ideal Users	Finance, Log Analysis	E-Commerce, IoT	Banks, pharmaceutical companies	Autonomous Vehicles, NLP	Sales, HR, Supply Chain
Integration	Hive, Pig, HBase	Kafka, Cassandra, TensorFlow	Excel, RDBMS, Hadoop	Keras, PyTorch	SQL, Snowflake, Salesforce

II. 5VS OF BIG DATA

The 5 Vs of Big Data—*Volume*, *Velocity, Variety, Veracity*, and *Value*—are essential attributes that characterize big data. *Volume* pertains to the enormous quantities of data produced from diverse sources, while *Velocity* emphasizes the rapid pace at which this data is generated, gathered, and analyzed. *Variety* refers to the various types of data, including structured, semi-structured, and unstructured formats such as text, images, and videos. *Veracity* concerns the reliability and accuracy of the data, tackling challenges related to inconsistencies and overall data quality. Lastly, *Value* highlights the insights and advantages that organizations can gain from analyzing big data, transforming raw information into practical knowledge. [2]

A. Comparison between tools on the basis of 5vs

1) *Volume*: The term volume in relation to big data tools signifies the capacity of a tool to effectively manage, process, or analyse data. *Hadoop* is tailored for extensive data storage and batch processing, capable of handling petabytes of both structured and unstructured data through HDFS. *Apache Spark*, while also adept at managing large data volumes, stands out for its in-memory processing capabilities, which allow it to outperform Hadoop in iterative tasks involving vast datasets. *SAS* is a robust analytics platform, but it generally focuses on smaller volumes of structured data, primarily within enterprise settings. *TensorFlow*, which is mainly utilized for machine learning and deep learning applications, processes large volumes of data in tensor format, emphasizing model training over raw data management. Finally, *Tableau* is a data visualization tool that is most effective with moderate data volumes and typically depends on backend systems like databases or data warehouses to handle larger datasets. In conclusion, Hadoop and Spark are best suited for high-volume big data, TensorFlow is designed for high-volume model inputs, SAS excels with structured enterprise data, and Tableau functions as a visualization layer above other data system.
2) *Velocity*: The velocity of big data tools pertains to the rate at which data is processed or analysed. *Hadoop* operates on a batch processing model, which results in slower performance for real-time applications, although it excels at managing large volumes of data simultaneously. In contrast, *Spark* boasts high velocity, capable of processing data in real-time or near real-time through in-memory computing, significantly outpacing Hadoop. *SAS* provides reasonable speed for analysing structured data, but it is not optimized for real-time big data streaming [5]. *TensorFlow* achieves high velocity during the training and deployment of machine learning models, particularly when utilizing GPUs; however, its speed is influenced by the size of the model and the hardware used. *Tableau*, on the other hand, does not process data directly but focuses on visual representation; its speed is contingent on the performance of the underlying data source, being quicker when linked to real-time data systems. In summary, Spark and TensorFlow deliver rapid processing capabilities, Hadoop is slower due to its batch processing approach, SAS is effective for conventional analysis, and Tableau's speed is determined by its data source connections.
3) *Variety*: The concept of variety in big data tools refers to the different types and formats of data that these tools are capable of handling, which encompasses structured, semi-structured, and unstructured data. *Hadoop* accommodates a broad range of data types, such as text, images, videos, and logs, making it ideal for various big data applications. *Spark* is also capable of managing all data types efficiently, supporting both structured data (like tables) and unstructured data (such as social media content or sensor information). *SAS* primarily focuses on structured data, including tables and spreadsheets, and has limited capabilities for unstructured data. *TensorFlow* is designed for numerical and unstructured data relevant to AI applications, such as images, audio, and text used in deep learning. *Tableau* excels with structured and semi-structured data sourced from databases or spreadsheets, but it is not equipped to directly process raw unstructured data. In summary, Hadoop and Spark offer the most extensive support for various data types, TensorFlow is tailored for unstructured AI data, SAS is predominantly for structured data, and Tableau is effective for structured visualizations but has constraints with unstructured data.
4) *Veracity*: The veracity of big data tools pertains to their ability to maintain the quality, accuracy, and reliability of data. *Hadoop* is capable of storing and processing extensive, unstructured datasets; however, it lacks integrated features for data quality assessment and depends on external tools or custom programming. *Spark* provides enhanced capabilities for data cleaning and transformation via its APIs, yet, similar to Hadoop, it necessitates manual intervention to verify data accuracy. *SAS* excels in data governance, validation, and quality control, making it a highly reliable choice for both business and regulatory applications. *TensorFlow* is primarily focused on model training, which means that data quality must be assured prior to input, as it does not address veracity directly and is particularly sensitive to low-quality data [9]. *Tableau* aids in data visualization and can identify inconsistencies, but it does not perform data cleaning or correction. In summary, SAS is the top choice for managing reliable, high-quality data, Spark provides flexibility, Hadoop requires additional support, TensorFlow is contingent on the quality of input data, and Tableau merely presents existing data without modification. [7]

5) *Value*: The significance of big data tools lies in the insights, results, and advantages they provide from data analysis. *Hadoop* contributes value by facilitating economical storage and batch processing of extensive datasets, which is beneficial for data archiving and long-term evaluations. *Spark* enhances value with its rapid, real-time analytics capabilities, enabling quicker decision-making and handling of complex data processing tasks. *SAS* offers substantial business value through its sophisticated analytics, statistical modelling, and user-friendly interfaces, making it a favoured choice in finance, healthcare, and government sectors. *TensorFlow* generates value by driving AI and machine learning models, assisting in addressing intricate challenges such as image recognition, natural language processing, and predictive analytics. *Tableau* provides value by transforming data into easily interpretable visualizations, allowing businesses to swiftly identify trends and make well-informed decisions. In conclusion, Hadoop enhances storage and processing capabilities, Spark offers rapid analytics, SAS facilitates expert analysis, TensorFlow enables AI applications, and Tableau renders data visually accessible.

TABLE II
COMPARISON OF BIG DATA W.R.T 5 V's

Characteristics	Hadoop	Spark	SAS	TensorFlow	Tableau
Volume	Petabyte-scale	TB-PB scale	TB-scale	TB-PB	TB limit
Velocity	Batch processing	Real-time	Moderate	Variable	Instant queries
Variety	Structured + Unstructured	All types	Primarily structured	Structured + Unstructured	Structured
Veracity	Requires preprocessing	Needs data cleaning	High-audit trails, governance	Sensitive to noisy data	Visual quality control
Value	Historical trends	Real-time analytics	Regulated insights	AI predictions	Business decisions

III.5Cs OF BIG DATA

The 5 Cs of Big Data in data analytics—*Computation*, *Complexity*, *Compliance*, *Capability*, and *Cost*—underscore essential factors for the effective management and analysis of extensive datasets. *Computation* pertains to the necessary processing power and infrastructure to efficiently manage large volumes of data. *Complexity* refers to the difficulties associated with handling various data types, sources, and structures, which often necessitate sophisticated tools and algorithms. *Compliance* emphasizes the importance of following legal and ethical guidelines concerning data privacy, security, and governance. *Capability* relates to the technical expertise, tools, and technologies required to derive valuable insights from big data. Finally, *Cost* involves the financial resources needed for data storage, processing, and skilled personnel, highlighting the importance of cost-effective strategies for sustainable data analytics. [2]

A. Comparison of tools on the basis of 5cs

1) *Computation*: When evaluating Hadoop, Spark, SAS, TensorFlow, and Tableau in terms of computation, each tool presents distinct advantages tailored to various data processing requirements. *Hadoop* operates on a batch-processing framework utilizing MapReduce, which is effective for handling large datasets across distributed systems, though it may lag in speed for real-time or iterative operations. In contrast, *Apache Spark* enhances performance with its in-memory computing features, making it considerably quicker than Hadoop for iterative algorithms and real-time analytics. *SAS (Statistical Analysis System)* is a robust proprietary analytics platform known for its high-performance computing capabilities, particularly in statistical analysis, but it may not scale as effectively as Spark or Hadoop in distributed settings. *TensorFlow*, mainly designed for deep learning and machine learning applications, excels in executing complex numerical computations and leverages GPUs and TPUs for improved performance. Conversely, *Tableau* serves as a data visualization tool with limited computational power; it depends on external data sources for processing and emphasizes the presentation of insights over intensive data manipulation. Thus, while Spark and TensorFlow provide rapid and sophisticated computation for big data and AI tasks, Tableau is optimized for visualization, with SAS and Hadoop offering a balance of strengths in statistical analysis and distributed batch processing, respectively.

2) *Complexity*: *Hadoop* is often viewed as complex due to its distributed system architecture and reliance on MapReduce programming, necessitating considerable expertise in Java and system setup. *Spark* alleviates some of Hadoop's intricacies by providing APIs in several languages (Scala, Python, Java) and enabling in-memory processing; however, it still requires a strong grasp of distributed computing principles. *SAS*, as a commercial product, is generally less complex for users because of its intuitive interface and built-in functionalities, although its proprietary nature may restrict flexibility and integration options. *TensorFlow* presents significant complexity, especially for novices, as it demands a thorough understanding of machine learning principles, neural networks, and low-level programming, although high-level APIs like Keras can help mitigate this complexity. Conversely, *Tableau* stands out as the simplest option among the five, designed for users with limited technical skills, allowing them to create interactive visualizations using a drag-and-drop interface. In summary, Tableau and SAS are more accessible, while Hadoop, Spark, and TensorFlow require more advanced technical knowledge and present greater challenges in deployment and operation. [1]
3) *Compliance*: Compliance across Hadoop, Spark, SAS, TensorFlow, and Tableau revolves around how effectively each tool supports data governance, enforces security protocols, and aligns with regulatory standards. *Hadoop* establishes a robust compliance framework with features such as HDFS encryption, Kerberos authentication, and audit logging; however, the implementation of these features can be intricate and may necessitate specialized configuration. *Spark*, which is frequently utilized alongside Hadoop, shares similar compliance functionalities, but ensuring security and compliance on a large scale can be difficult without adequate integration and management. *SAS*, recognized as an enterprise-level analytics solution, stands out in compliance, being widely utilized in sectors like healthcare and finance due to its comprehensive built-in controls for data governance, audit trails, and regulatory reporting. *TensorFlow*, mainly a machine learning platform, does not offer inherent compliance features, compelling developers to independently establish security and privacy protocols, which may heighten the risk of non-compliance in sensitive contexts. *Tableau*, as a data visualization tool, aids compliance through functionalities such as user authentication, role-based access, and compatibility with enterprise security frameworks, making it effective for managing sensitive information in a regulated setting. In conclusion, SAS and Tableau provide more comprehensive and ready-to-use compliance support, while Hadoop and Spark necessitate more manual configuration, and TensorFlow significantly depends on the developer's efforts to ensure compliance.
4) *Capability*: *Hadoop* excels in the storage and processing of vast amounts of both structured and unstructured data across distributed systems, making it particularly suitable for large-scale batch processing tasks. *Spark* builds on this foundation by offering in-memory computing and facilitating advanced analytics, including machine learning, graph processing, and real-time streaming, thus providing a more comprehensive and rapid data processing environment. *SAS* is particularly strong in statistical analysis, predictive modelling, and reporting, equipped with powerful built-in algorithms and tools that cater to data scientists and analysts, especially in sectors that demand accuracy and regulatory compliance. *TensorFlow* is distinguished by its ability to create and implement sophisticated machine learning and deep learning models, effectively handling neural networks, natural language processing, and computer vision applications with high efficiency across CPUs, GPUs, and TPUs. In contrast, *Tableau* is not designed for data processing or advanced analytics but excels in data visualization and dashboard development, allowing users to interactively explore and present data insights in an intuitive manner. In conclusion, while TensorFlow is at the forefront of AI capabilities, Spark is recognized for its speed and flexibility, Hadoop for its scalable storage and batch processing, SAS for its statistical rigor, and Tableau for its strengths in visualization and business intelligence.
5) *Cost*: Considering the cost aspects, the five tools—Hadoop, Spark, SAS, TensorFlow, and Tableau—exhibit considerable differences in financial requirements, licensing, and infrastructure demands. Both Hadoop and Spark are open-source platforms, allowing free usage; however, they typically necessitate significant infrastructure investments and skilled personnel for deployment and maintenance, which can elevate operational expenses. In contrast, SAS is a commercial solution with substantial licensing costs, yet it provides comprehensive support, reliability, and ready-to-use analytics capabilities, which may be warranted in enterprise settings. TensorFlow, also open-source and supported by Google, is free and highly scalable, but like Hadoop and Spark, it requires technical expertise and considerable computing resources (particularly GPUs or TPUs) for extensive machine learning applications. Tableau employs a tiered pricing model based on user roles (Creator, Explorer, Viewer), and while it is designed for ease of use, the licensing fees can accumulate significantly for larger organizations. In summary, while open-source tools such as Hadoop, Spark, and TensorFlow are economical in terms of software costs, they may lead to higher indirect expenses due to infrastructure and expertise requirements. Conversely, SAS and Tableau entail direct software costs but provide greater built-in support and user-friendliness.

TABLE III
COMPARISON OF BIG DATA W.R.T 5 C's

Characteristics	Hadoop	Spark	SAS	TensorFlow	Tableau
Computation	Batch processing	Real-time & Batch	Single-server analytics	GPU/TPU-accelerated	Instant visualization
Complexity	High	Moderate	High	Very high	Low
Compliance	Moderate	Moderate	High	Medium	High
Capability	Distributed storage & batch analytics	Unified engine (SQL, ML, Streaming)	Statistical modeling & regulated BI	Neural networks & production AI	Interactive dashboards & business insights
Cost	Open Source	Open Source	Very Expensive	Open Source	Expensive

IV. CONCLUSION

Big Data Analytics plays an important function in today's environment. It is a key aspect of the organization's growth and development in response to developing market trends, as well as maintaining track of the past, present, and future. Handling enormous datasets necessitates the use of appropriate analytical tools for thorough data evaluation, ensuring that the output from those datasets is valuable. The above comparative analysis of several trending and demanding tools will assist users and data analysts in selecting the finest tool that meets their needs. Corporates, industries, and diverse enterprises rely heavily on these technologies to better their projects.

V. FUTURE SCOPE

This documentation will help readers are get to know how big data are essential to analysis for the better output and how it impacts in our life. Comparative research can be broadened to assess the effectiveness of tools in new domains such as real-time analytics, edge computing, and IoT data processing. Furthermore, studies may investigate models for cost efficiency, scalability, security improvements, and the ecological effects of big data processing. As sectors progressively depend on data-driven approaches, continuous development and assessment of big data tools will be essential for sustaining competitive edge and promoting innovation.

REFERENCES

[1] Charles, V., Emrouznejad, A., Gherman, T., & Cochran, J, "Why data analytics is an art", November 2022.

[2] J. Vijayaraj, R. Saravanan, P. Victer Paul, R. Raju, "A Comprehensive Survey on Big Data Analytics Tools" November 2016.

[3] Swetha Chinta, "Integrating Machine Learning Algorithms in Big Data Analytics: A Framework for Enhancing Predictive Insights",10, October-2021.

[4] Anayo Chukwu Ikegwu[1]. Henry Friday Nweke[2] Chioma Virginia Anikwe[1]. Uzoma Rita Alo[1]. Obikwelu Raphael Okonkwo[1,3], "Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions", 12 March 2022.

[5] Sivananda Reddy Julakanti, Naga Satya Kiranmayee Sattiraju, Rajeswari Julakanti, "Implementing Spark Data Frames for Advanced Data Analysis", 25 February 2021.

[6] Sainath Muvva, "Optimizing Spark Data Pipelines: A Comprehensive Study of Techniques for Enhancing Performance and Efficiency in Big Data Processing", 20 December 2023.

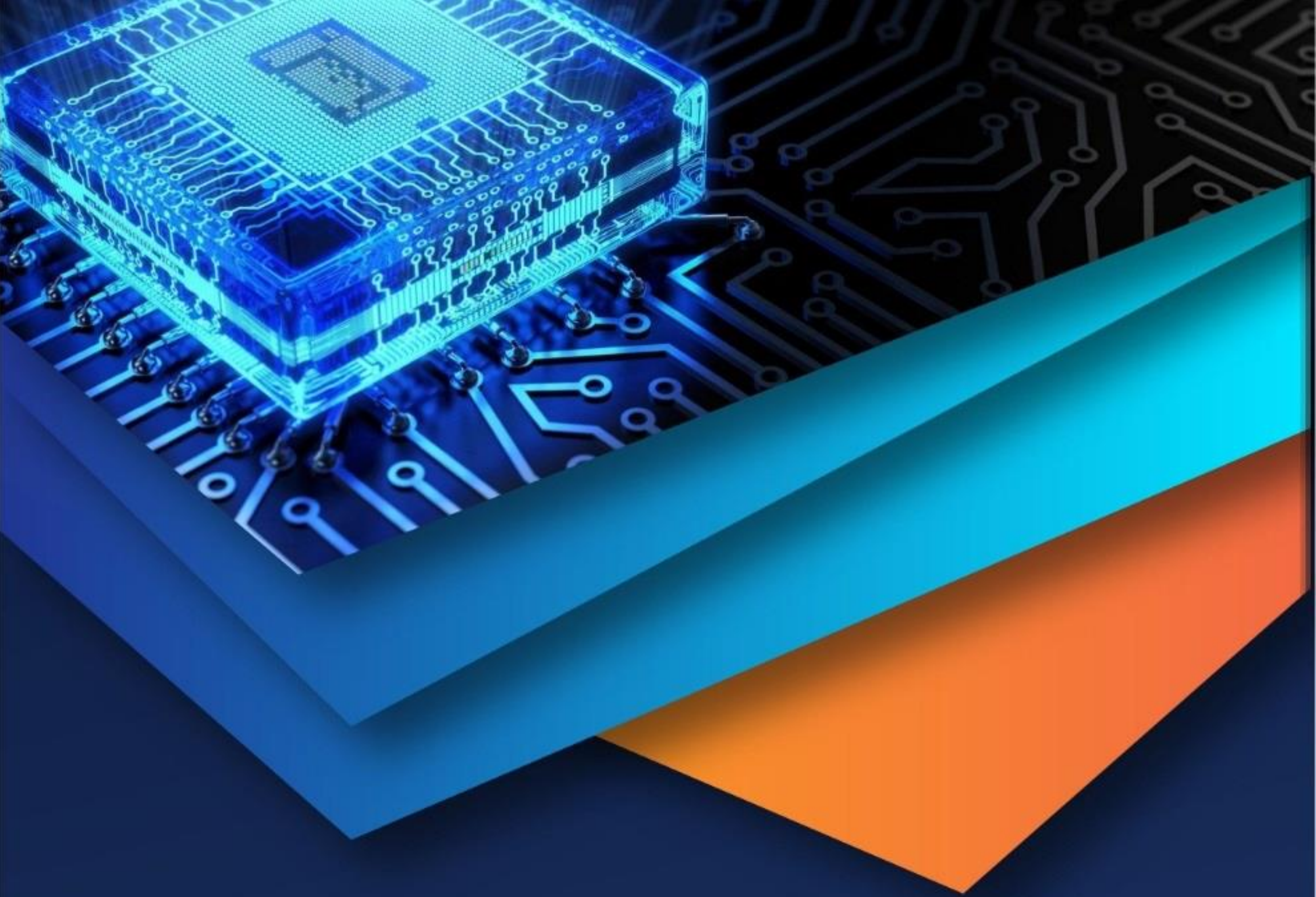
[7] Addepalli Lavanya 1, Sakinam Sindhuja 2, Lokhande Gaurav 3, Waqas Ali 4, "A Comprehensive Review of Data Visualization Tools: Features, Strengths, and Weaknesses", 28 January 2023.

[8] Mrs. Kanchan A. Khedekar, "Data Analytics for Business Using Tableau", 2021.

[9] Peter Goldsborough Fakultät für Informatik, "A Tour of TensorFlow Proseminar Data Mining", October 2016.

[10] Koustubh Sharma[1], Aditya Shetty[2], Arnish Jain[3], Ritesh Kumar Dhanare[4], "A Comparative Analysis on Various Business Intelligence (BI), Data Science and Data Analytics Tools", 27-29 January 2021.

[11] Bharath Muddarla[1] and Vineeth Reddy Vatti[2], "Optimizing Cloud Resources for Machine Learning Applications: A Comparative Study of SQL-Driven and Python-Driven Workflows", 2024.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY