iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53894

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Survey Paper on Big Data Technologies

Priti Joshi[1], Manisha Mathur[2], Priyanka Shah[3]
Computer Engineering Department, Vivekanand Education Society's Institute of Technology,Chembur,Mumbai

Abstract: In the 21st century with the tremendous increase in the usage of internet and extreme advancement in Technology huge amount of Data is generated every second.
This Data is combination of structured semistructured and unstructured data which is called as Big data. Being able to use this data provides huge opportunities and to turn these opportunities into reality, people need to use data to solve problems.
Handling Big data has various challenges like Storage, Search, analysis, sharing, visualization, transfer and privacy violations.
Big data analytics can help better and faster decision making, modeling and predictive analysis for enhanced business intelligence.
In this paper the authors have done the survey on different types of Big data technologies.
Keywords: Big Data, Hadoop , Map reduce, Pig, Hive, Tableau

I. INTRODUCTION

In Today's World so much Data is generated every second due to the surge in internet usage and advancement in technologies. Examples of massive sources of data generation social media is the major contributor other then that medical devices such as MRI and other scan machines generate huge data, stock markets, banks, telecom service providers are the other major sources, Boeing jet generates 1tb of data per flight,.it's a data driven world according to author [3] at twitter, they process approximately 400 billion events in real time and generate pet byte (pb) scale data every day.

Facebook generates 4 petabytes of data per day according to "www.statista.com" the facebook reported over 2.9 billion monthly active users in the first quarter of 2022 .

Social media has contributed a lot towards Big Data …Instagram has 1.628 billion users around the world in April 2023(www.datareportal.com).Over 95 million photos and videos are shared on instagram every day. It is very difficult to store and manage such a huge amount of data using traditional database management systems. Big data size is increasing consistently ranging from terabytes to zettabytes.

The term big data refers to any collection of data which is so large and complex that it becomes difficult to handle using traditional database management systems and data processing tools. Big data is data that can not be processed by currently used traditional databases and software infrastructure.

It is a field dedicated to the analysis, processing and storage of large collections of data which cannot be handled by current technology infrastructure.

Big Data is characterized by 3 v's [6]:Volume, Variety and Velocity. There is no fixed size to classify a dataset as big data or not ,instead 'Volume' dimension refers to a data set which is large enough to be beyond the processing capabilities of traditional data processing tools ,and the dataset can grow up to any size which can be in peta- bytes(1024 terabytes), exa-bytes (1024 peta bytes) or even more.

The 'Variety' dimension is included because the traditional data warehouses store only structured data, but data generated on Twitter etc is highly unstructured, so such data is also beyond the processing capabilities of traditional data processing tools. While traditional data analytics is based on periodic analysis of data, big data is processed and analysed in real-time or near real time. Therefore, the third dimension of 'Velocity' has also been included.

This paper is divided into following sections: section 2 describes the different technologies for big data , section 3 concludes this paper and then referenced materials are listed.

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

II. DIFFERENT TECHNOLOGIES

Big Data Technologies are mainly classified in 4 types

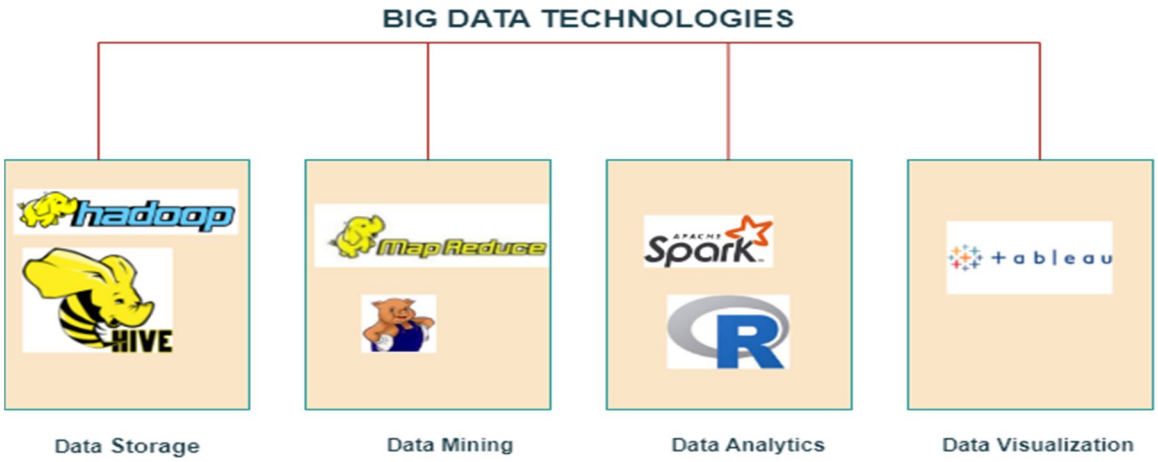


Fig1.Big Data Technologies

A. Data Storage

Big data storage is a technique that makes use of efficient infrastructure which helps in storing , managing and retrieving large amounts of data. Big data is concerned with storing and managing data in and efficient way and fulfilling the requirement of applications which require huge data.

1) Hadoop

Hadoop is a beneficial technology for data analysts. There are many important features of Hadoop which makes it important and user-friendly. The system is able to store and process enormous amounts of data at an extremely fast rate. A semi-structured, structured and unstructured data set can differ depending on how the data is structured. Hadoop Enhances operational decision-making and batch workloads for historical analysis by supporting real-time analytics. Organizations have the facilities to store the raw data and processors filter it for specific analytic uses when required. Hadoop is scalable and therefore organizations can handle more data by adding more nodes into the systems. The components of Hadoop are Hadoop HDFS - In Hadoop Distributed File System data storage and processing is done in distributed parallel processing manner across the clusters. Each data is duplicated many times to make sure data availability. It has two daemons. One for master node— NameNode and for slave nodes —DataNode..Hadoop MapReduce - Hadoop MapReduce executes tasks in a parallel fashion by distributing the data into small blocks..Hadoop YARN - Hadoop YARN is a resource management unit of Hadoop.[4]

2) Hive

Hive is a data storage system used to analysis structured data. Important functionalities for which hive is deployed are data summarization,data analysis and data query. The query language supported by hive is hiveQL. HiveQL translates SQL-like queries into MapReduce jobs for deploying them on Hadoop..Using Hive, we can also avoid the traditional approach of writing complex MapReduce programs. Hive also supports Data Definition Language (DDL), Data manipulation Language (DML), and User Defined Functions (UDF).Features of Hive includes fast and scalable. It is also capable of analyzing large datasets stored in HDFS.Different file formats such as plain text, RCFile, and HBase can be used using hive Due to which it can operate on compressed data that is stored in the Hadoop ecosystem.[13]

3) Mongodb

Mongodb is a data storage technique which is an open-source document-oriented database and is used to store a large amount of data. It is categorized under the NoSQL (Not only SQL) database as the storage and retrieval of data are not in the form of tables. MongoDB itself is a database server and the data is stored in these databases.In other words, the MongoDB environment gives a server that one can start and then create multiple databases on it using MongoDB. As it is a NoSQL database, the data is stored in the collections and documents.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

In the MongoDB server, one is allowed to run multiple databases. Features of MongoDB are no design of schema is required as it uses NoSQL database. It provides more flexibility to the fields in the documents. It provides high performance, availability, scalability. It allows storing nested data. This nesting of data allows creating complex relations between data, storing the nested data in the same document which makes the working and fetching of data extremely efficient as compared to SQL.[15]

B. Data Mining

Huge data is generated every second due to the world wide web .The number of users of the internet is growing exponentially day by day. Massive volumes of data in terms of audio, video and text is being generated and accessed by millions of users throughout the world.If that data is not analysed and used properly it's a wastage in terms of memory storage .This data can prove to be very useful and valuable for different enterprises and businesses. Data can be sorted ,analysed and patterns and relationships can be identified that can prove to be beneficial for solving business problems. Credit Company, Health Care and Medicine Company, Financial Banking, Telecommunications, Marketing and Retail Big data, Industrial, airline, and insurance company can make use of Big Data Mining techniques and tools to get valuable business insights. Data mining techniques and tools enable organisations to foretell future trends and can be useful for making informed and productive decisions. Data mining is also known as Knowledge Discovery in Database (KDD)

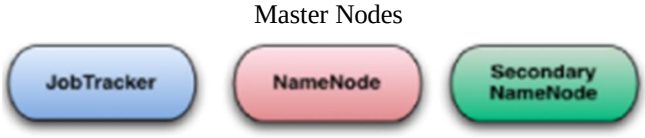
With the analysis of voluminous amounts of data analysts, researchers, and business users can make informed, productive and faster decisions. However, the dramatic increase of data amounts have made the well-known data mining algorithms unsuitable for such huge data sizes .MapReduce and Pig are two of the latest techniques used for processing big data.

1) MapReduce

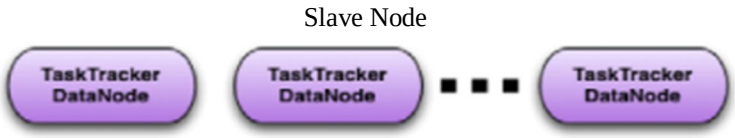
Map-Reduce was introduced by Google in order to process and store large amounts of data on commodity machines. MapReduce provides a programming model for processing large data sets using a distributed ,parallel algorithm on a cluster. MapReduce distributes the task across multiple nodes Many small machines can be used to process jobs that normally could not be processed by a large machine. The Map Reduce programming model works in two phases which are Map and Reduce. The job of the Map function is to perform the task as the master node after taking the input,Map Function divides it into smaller sub modules and distributes sub modules into slave nodes.Due to this a hierarchical tree like structure is formed as slave mode further divides the model into sub modes. The slave node processes the base problem and passes the result back to the master Node[10]. The Map Reduce system arranges together all intermediate pairs based on the intermediate keys and refer them to Reduce nodes for producing the final output. Reduce works as the master node collects the results from all the slave nodes and after combining nes them together forms the output. MapReduce is highly scalable and can be used across many computers.MapReduce provides a clean abstraction for programmers .MapReduce programs are usually written in Java.MapReduce jobs are generally controlled by a software daemon known as JobTracker. The JobTracker resides in master node. Clients submit MapReduce jobs to the Job tracker. The JobTracker assigns Map and Reduce jobs to other nodes on the cluster ,these nodes each run a software daemon called TaskTracker. The TaskTracker is actually responsible for instantiating the Map or Reduce task and reporting progress back to the job tracker.

Map Reduce Components

a) *Name Node:* NameNode is considered as the master node in Hadoop Distributed File System (HDFS) that manages the file system metadata doesn't deal with files directly.
b) *Data Node:* DataNode is a slave node in HDFS that stores the actual data as instructed by the NameNode.
c) *Job Tracker:* Job tracker's job is to manage resources, track the availability of resources and track the progress of fault tolerance. Job tracker with the help of Namenode determine the location of data. Finds the task tracker nodes to execute the task on given nodes so in short it can be said that job tracker schedules, allocates and monitors job execution on slaves
d) Task Tracker: runs Map Reduce operations



©IJRASET: All Rights are Reserved



In small clusters JobTracker, Name Node and Secondary NameNode can all be put on a single machine.
The Non programmers found it difficult to work on MapReduce as every one is not comfortable in writing lengthy Java codes. To understand and implement the concepts of Map,Sort and Reduce concept of MapReduce is not an easy job.It became necessary to think of easier ways to process large data in a less complex and time consuming Java codes.So Apache Pig was developed by Yahoo which uses simple steps to process datasets.

2) Pig

Pig is a scripting platform that runs on Hadoop clusters.It is an open source high-level data flow system, is a layer of abstraction built on the top of Map Reduce.Pig provides high level tool or a platform for analyzing large data sets using a high-level programming language called Pig Latin for writing code for the purpose of data analysis ,Pig Engine which is part of Pig infrastructure will convert these programs into Map and Reduce tasks.Pig is extensively used and accepted by Yahoo!Twitter, Netflix etc.Pig requires its own scripting language named 'Pig Latin'.It uses a query approach that resembles SQL queries which results in reducing the length of the code[7].

a) Pig Latin is SQL like language.
b) It provides many builtIn operators.
c) Pig works on structured,semi-structured and structured data
d) It makes use of the following data types: Tuple, Bag, Map and Atom
e) With Pig a higher level of data abstraction can be achieved.

Pig Component

- Pig Latin: It is a Command based language. Designed specifically for data analysis in Hadoop..
- Execution Environment :The environment in which Pig Latin commands are executed .Currently it supports local and Hadoop modes. In the local mode, the Pig engine takes input from the Linux file system and the output is stored in the same file system. In the Hadoop mode, the Pig engine loads and perform processing operations on the data stored on the HDFS.
- Pig compiler converts Pig Latin into Map Reduce. Compiler strives to optimize execution. Apache Pig has been proven to be the most efficient tool for analyzing structured, semistructured and unstructured data.

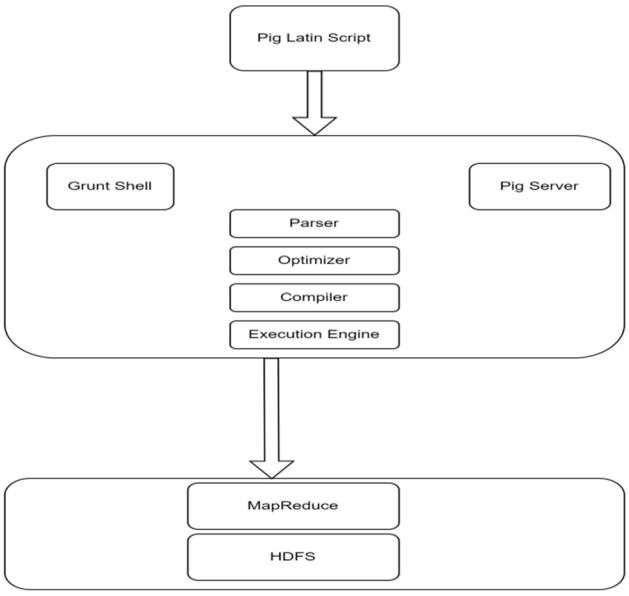


Fig2. Pig Architecture[9]

Pig Architecture: It has following components

- ✓ Pig Latin Script: Programmers write scripts using Pig Latin Script
- ✓ Grunt Shell: Grunt is a Pig's interactive shell which is used to execute Pig scripts which goes to Pig Server
- ✓ Parser: Checks the syntax of Pig Script ,after checking the output will be a DAG(Directed Acyclic Graph)
- ✓ Optimizer: DAG is passed through the optimizer where optimization takes place.
- ✓ Compiler: Converts the DAG into MapReduce jobs
- ✓ Execution Engine: The MapReduce jobs are executed here.
- ✓ MapReduce jobs are processed at HDFS

C. Data Analytics

The process of Examining, analyzing and collecting huge amounts of data to find out market trends, insights, and patterns that can help companies make better business decisions is known as Big data analytics. Data analytics technologies and techniques are widely used in industries to empower organizations to make more-informed business decisions. Various approaches to data analytics include looking at what happened (descriptive analytics), what is exactly going to happen (predictive analytics),why something happened (diagnostic analytics), or what should be solution next (prescriptive analytics).

1) Apache Spark

Apache Spark[5] is a widely used tool for Big data analytics as it gives fast and efficient results while running applications. It is a powerful big data processing platform which adapts the hybrid framework.

A hybrid framework provides support for both batch processing and stream processing capabilities. It has many similar principles with Hadoop's Map Reduce engine.

Apache Spark has many distinguishable features such as Speed, Usability,advanced analytics, flexibility,real time stream processing .

The Apache Spark ecosystem consists of the following main components:

a) *Spark SQL:* Formerly known as Shark. Spark SQL is a distributed framework that works with structured and semi-structured data.

b) *Spark Streaming:* It enables users to process streaming of data in real time.To perform streaming analysis, It enhances the quick scheduling capability of Apache Spark by keeping data into mini batches.

c) *MLlib:* It delivers high-quality algorithms with high speed and makes machine learning easy to use and scale. Several machine learning algorithms such as regression, classification, clustering, and linear algebra are present.

d) *GraphX:* It is a graph computation engine that enables the building, manipulation, transformation and execution of graph-structured data at a large scale.

e) *Spark Core:* It provides a vast range of APIs as well as applications for programming languages such as Scala, Java, and Python APIs to facilitate the ease of development.

f) *SparkR:* It is a package for R which enables data scientists to support the power of Spark from R shell.

2) R Language

R[8] is free open source language which uses integrated development environment (IDE) as R Studio .It is also a software environment which is used to examine statistical information, reporting, graphical representation and data modeling .It is the enactment of the S programming language, which is combined with lexical scoping semantics. It is easy to learn and most effective data analytics programming language R is one of the most important tool which is used by data analyst, researchers, statisticians and marketers for retrieving, cleaning, analyzing, visualizing, and presenting data. It compiles the code and runs on a different UNIX platforms and systems like MacOS and Windows..

R tool is available for anyone to use as it is a free software. The R tool is best suited for people with data oriented problems and not for programmers. The Main task of data science is the way we are dealing with the data: cleaning, feature selection, feature engineering and import. The job of Data scientist's to accept the data, use it, and expose the best approach.. The perfect algorithms can be implemented with R programming for machine learning. TensorFlow and Keras authorize us to develop high-end machine learning techniques .R package has to perform Xgboost. For Kaggle competition Xgboost is one of the finest algorithms

D. Data Visualization

A visual representation of information and data is known as Data Visualization.Visual elements are like graphs,charts and maps. Data visualization tools prepare a way to identify patterns, trends and outliers in data..Our eyes are drawn to colours and patterns our eyes can identify the message easily. When we see a chart, we quickly see trends and outliers.According to the World Economic Forum, the world produces 2.6 quintillion bytes of data every day., and 90% of data has been generated in the past two years.With lots of data, it's become difficult to manage and make sense of it all. It would be quite difficult for any single person to traverse through data line-by-line and recognize patterns and make observations. Data proliferation is a unit of the data science process, which has features of data visualization. Instead of reading thousands of rows on a spreadsheet we use data visualization because a visual summary of information is easy to identify patterns and trends. In this way the human brain works.

The uses of Data Visualization as follows.

- ✓ Explore data with presentable results in a powerful way
- ✓ Pre-processing portion of the data mining process.is the primary way.
- ✓ It provides the data cleaning process by searching incorrect and missing values.
- ✓ It Plays an important role in gathering categories as a part of the data reduction process.

1) Tableau

Tableau[11] is a very advantageous tool. It is a data analysis and visualization tool which is used to connect with various data sources adequately. The big advantage of Tableau is it can create interactive dashboards.Anyone can create dashboards without much coding knowledge with the simple drag and drop interface.Due to its ability to translate data into insightful visual dashboards Tableau has been a very popular tool . It uses application integration like single sign-on applications and JavaScript APIs so that Tableau analytics can be included into basic business applications uniformly. It can create a range of visualizations to present the data and showcase insights. It consists of tools that help it to work upon the data and see the output in a visual format which is very useful and easy to understand. Tableau also comes with real-time data analytics capacity and cloud support.It is the most safe,strong and flexible end-to-end analytical platform for the information from connection via cooperation. It also enhances data power for people. It is the only business intelligence platform which can be scaled for business that turns data into valuable information.The advantage of Tableau is that it can manage millions of rows of data easily. It can create different types of visualization that can be developed with the large amount of data without having another performance of the dashboards.There is a feature of Tableau where users can make two connections to different data sources like SQL . There are various types of visualization modes available in tableau which are atrocious the user experience.

III.CONCLUSION

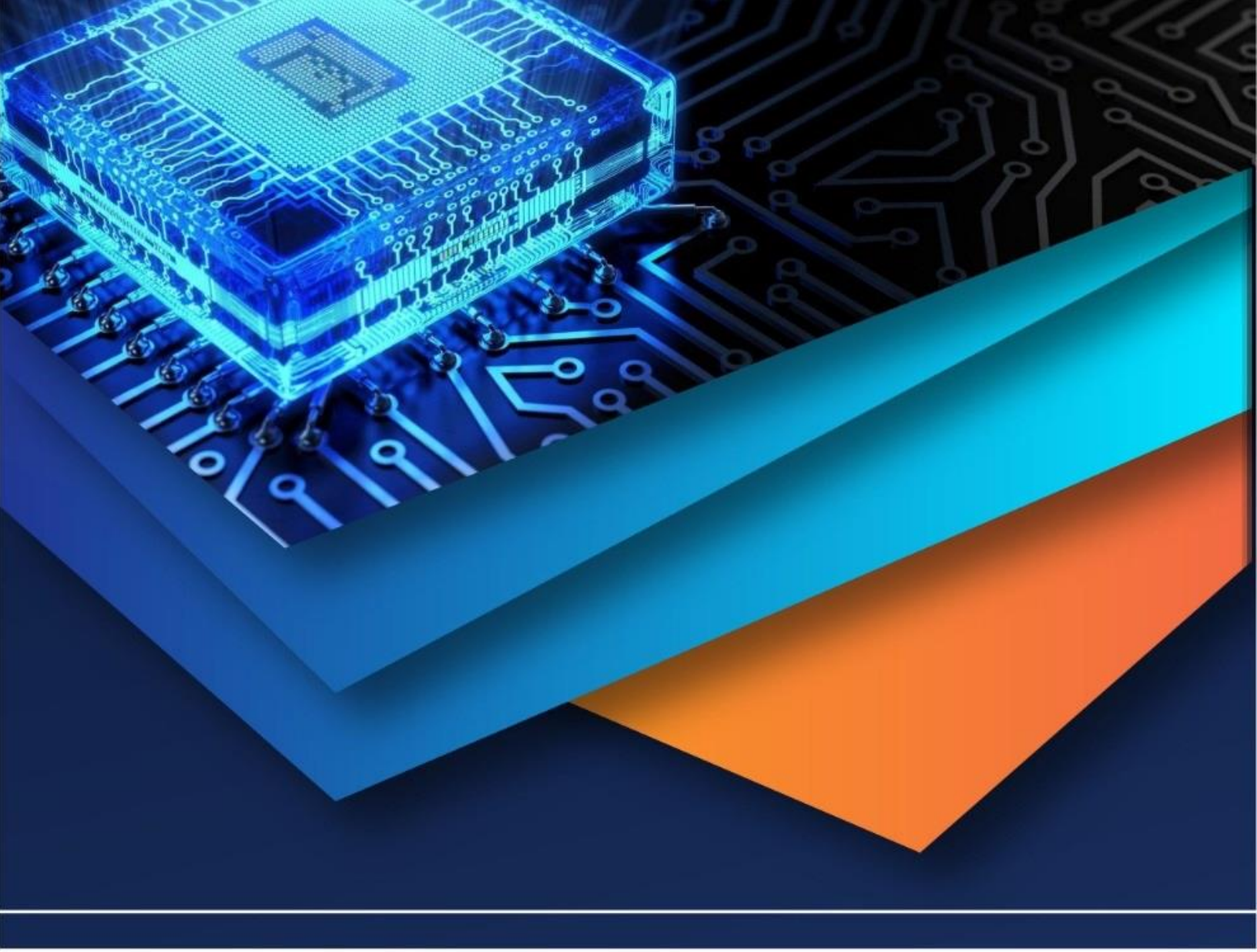
We are at the start of a new era where Big Data will help us to gather knowledge that no one has discovered before. The availability of Big Data, low cost commodity hardware, and new information management and analytic software have produced a unique movement in the history of data analytics. The main aim of this paper is to make a survey of various big data handling techniques like Hadoop, Hive, MapReduce, Pig, Tableau, Spark, MongoDb.

REFERENCES

[1] Y. Demchenko, C. de Laat and P. Membrey, "Defining architecture components of the Big Data Ecosystem," 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, USA, 2014, pp. 104-112, doi: 10.1109/CTS.2014.6867550.
[2] Elgendy, Nada & Elragal, Ahmed. (2014). Big Data Analytics: A Literature Review Paper. Lecture Notes in Computer Science. 8557. 214-227. 10.1007/978-3-319-08976-8_16.
[3] https://blog.twitter.com/engineering/en_us/topics/infrastructure/2021/processing-billions-of-events-in-real-time-at-twitter-#:~:text=At%20Twitter%2C%20we%20process%20 approximately,PB)%20scale%20data%20every%20day.
[4] Abdalla, H.B. A brief survey on big data: technologies, terminologies and data-intensive applications. J Big Data 9, 107 (2022). https://doi.org/10.1186/s40537-022-00659-3
[5] Shaikh, Eman & Mohiuddin, Iman & Alufaisan, Yasmeen & Nahvi, Irum. (2019). Apache Spark: A Big Data Processing Engine. 1-6. 10.1109/MENACOMM46666.2019.8988541.
[6] Laney, "3-D Data Management: Controlling Data Volume,Velocity and Variety," META Group Research Note, February 6,2001.
[7] A. C. Priya Ranjani & Dr. M. Sridhare ANALYSIS OF WEB LOG DATA USING APACHE PIG IN HADOOP ISSN 2348 –1269, Print ISSN 2349-513
[8] Tejas Rajeshirke, Ceena Joseph Thundiyal, Nishi Tiku "STUDY OF R PROGRAMMING "(2017)Volume: 04 Issue: 06 | June -2017
[9] https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop

[10] Manju Lakshmi1 Smita C Thomas2 "Survey on Big data"IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 09, 2017 | ISSN (online): 2321-0613
[11] Nikhat Akhtar , Nazia Tabassum 2, Asif Perwej 3 and Yusuf Perwej 4 Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus): https://doi.org/10.30574/gjeta.2020.3.2.0029
[12] Hive – A Petabyte Scale Data Warehouse Using Hadoop Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy
[13] Evaluating the Performance of Apache Hive and Apache Pig using Hadoop environment Namrata Patel National College of Ireland.
[14] A Review on Various Aspects of MongoDb Databases ,Anjali Chauhan, M.tech Scholar, CSE Department,Rawal Institute of Engineering and Technology, Faridabad, Haryana, India
[15] https://www.geeksforgeeks.org/what-is-mongodb-working-and-features

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)