BIG DATA OPEN SOURCE TOOL: HADOOP

Tanvi Panjari
Masters *Of Science in Computer Science Student*
Chikitsak Samuha's Sir Sitaram & Lady Shantabai Patkar College of Arts & Science and V.P. Varde College of Commerce & Economics,Mumbai
Email:tanvipanjari06@gmail.com

Abstract— Hadoop is an open-source software framework for storing and processing big data. It was created by Apache Software Foundation in 2006, based on a white paper written by Google in 2003 that described the Google File System (GFS) and the MapReduce programming model. The Hadoop framework allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It is used by many organizations, including Yahoo, Facebook, and IBM, for a variety of purposes such as data warehousing, log processing, and research. Hadoop has been widely adopted in the industry and has become a key technology for big data processing

Key Words: Big Data; Hadoop; HDFS; MapReduce; YARN

I. INTRODUCTION

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.[2]

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.
While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data. And graph databases are becoming increasingly important as well, with their ability to display massive amounts of data in a way that makes analytics fast and comprehensive.

1. What is Big Data?

Big data consists of structured, unstructured, and semi-structured data. It is formally characterized by its five Vs: volume, velocity, variety, veracity, and value.

Volume describes the massive scale and size of data sets that contain terabytes, petabytes, or exabytes of data.

Velocity describes the high speed at which massive amounts of new data are being generated.

Variety describes the broad assortment of data types and formats that are being generated.

Veracity describes the quality and integrity of the data in an extremely large data set.

Value describes the data's ability to be turned into actionable insights.[3]

IMPORTANCE OF BIG DATA

Big data is important because of its potential to reveal patterns, trends, and other insights that can be used to make data-driven decisions.

From a business perspective, bighelps organizations improve operational efficiency and optimize resources. For example, by aggregating large data sets and using them to analyze customer behavior and market trends, an e-commerce business can make decisions that will lead to increased customer satisfaction, loyalty – and, ultimately, revenue.

Advancements in open-source tools that can store and process large data sets have significantly improved big data analytics. Apache's active communities, for instance, have often been credited with making it easier for newcomers to use big data to solve real-world problems.[3]

Types of Big Data

Big data can be categorized into three main types: structured, unstructured, and semi-structured data.

Structured big data: It is highly organized and follows a pre-defined schema or format. It is typically stored in spreadsheets or relational databases. Each data element has a specific data type and is associated with predefined fields and tables. Structured data is characterized by its consistency and uniformity, which makes it easier to query, analyze and process using traditional database management systems.

Unstructured big data: It does not have a predefined structure and may or may not establish clear relationships between different data entities. Identifying patterns, sentiments, relationships, and relevant information within unstructured data typically requires advanced AI tools such as natural language processing (NLP), natural language understanding (NLU), and computer vision.

Semi-structured big data: contains elements of both structured and unstructured data. It possesses a partial organizational structure, such as XML or JSON files, and may include log files, sensor data with timestamps, and metadata.[3]

The cryptographic consent mechanism associated with the verified identity of network participants can lead to the trusted verification of transactions.

II. BIG DATA TOOLS

Dealing with large data sets that contain a mixture of data types requires specialized tools and techniques tailored for handling and processing diverse data formats and distributed data structures. Popular tools include:

Azure Data Lake: A Microsoft cloud service known for simplifying the complexities of ingesting and storing massive amounts of data.

Beam: An open-source unified programming model and set of APIs for batch and stream processing across different big data frameworks.

Cassandra: An open-source, highly scalable, distributed NoSQL database designed for handling massive amounts of data across multiple commodity servers.

Databricks: A unified analytics platform that combines data engineering and data science capabilities for processing and analyzing massive data sets.

Elasticsearch: A search and analytics engine that enables fast and scalable searching, indexing, and analysis for extremely large data sets.

Google Cloud: A collection of big data tools and services offered by Google Cloud, such as Google BigQuery and Google Cloud Dataflow.

Hadoop: A widely used open-source framework for processing and storing extremely large datasets in a distributed environment.

Hive: An open-source data warehousing and SQL-like querying tool that runs on top of Hadoop to facilitate querying and analyzing large data sets.

Kafka: An open-source distributed streaming platform that allows for real-time data processing and messaging.

KNIME Big Data Extensions: Integrates the power of Apache Hadoop and Apache Spark with KNIME Analytics Platform and KNIME Server.[2]

MongoDB: A document-oriented NoSQL database that provides high performance and scalability for big data applications.

Pig: An open-source high-level data flow scripting language and execution framework for processing and analyzing large datasets.

Redshift: Amazon's fully-managed, petabyte-scale data warehouse service.

Spark: An open-source data processing engine that provides fast and flexible analytics and data processing capabilities for extremely large data sets.

Splunk: A platform for searching, analyzing, and visualizing machine-generated data, such as logs and events.

Tableau: A powerful data visualization tool that helps users explore and present insights from large data sets.

Talend: An open-source data integration and ETL (Extract, Transform, Load) tool that facilitates the integration and processing of extremely large data sets**.[**4]

III. WHAT IS HADOOP?

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Hadoop consists of four main modules:

- Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.

- Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.

- MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.

- Hadoop Common – Provides common Java libraries that can be used across all modules.

How Hadoop Works?

Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed processes against huge amounts of data. Hadoop provides the building blocks on which other services and applications can be built.

Applications that collect data in various formats can place data into the Hadoop cluster by using an API operation to connect to the NameNode. The NameNode tracks the file directory structure and placement of "chunks" for each file, replicated across DataNodes. To run a job to query the data, provide a MapReduce job made up of many map and reduce tasks that run against the data in HDFS spread across the DataNodes. Map tasks run on each node against the input files supplied, and reducers run to aggregate and organize the final output.[2]

Benefits of using Hadoop for big data analytics

Hadoop has advantages from a business perspective, including cost savings and reduced tech failure, the upshots of not relying on in-house hardware. And for many years, there was a robust community of trouble shooters and problem-solvers consistently improving the open-source framework. Hadoop allowed business stakeholders to outsource significant aspects of big data analytics without acquiring separate servers and data scientists to manage the framework.

However, Hadoop has its limitations—otherwise there wouldn't be so many add-ons within the Hadoop ecosystem, like Hive, HBase, and Spark. In fact, as it's aged, the open-source community has dwindled. To still do truly exceptional things with Hadoop, you need expert-level Java skills. And while Hadoop remains a foundational tool of big data analytics, it lacks the versatility to perform more intricate tasks on smaller datasets.[4]

Challenges of using Hadoop

- Complexity: Hadoop can be complex to set up and maintain, especially for organizations without a dedicated team of experts.
- Latency: Hadoop is not well-suited for low-latency workloads and may not be the best choice for real-time data processing.
- Limited Support for Real-time Processing: Hadoop's batch-oriented nature makes it less suited for real-time streaming or interactive data processing use cases.
- Limited Support for Structured Data: Hadoop is designed to work with unstructured and semi-structured data, it is not well-suited for structured data processing
- Data Security: Hadoop does not provide built-in security features such as data encryption or user authentication, which can make it difficult to secure sensitive data.[6]

IV. TOOLS AND APPLICATION OF HADOOP

The Hadoop ecosystem has grown significantly over the years due to its extensibility. Today, the Hadoop ecosystem includes many tools and applications to help collect, store, process, analyze, and manage big data. Some of the most popular applications are:

Spark – An open source, distributed processing system commonly used for big data workloads. Apache Spark uses in-memory caching and optimized execution for fast performance, and it supports general batch processing, streaming analytics, machine learning, graph databases, and ad hoc queries.

Presto – An open source, distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3.

Hive – Allows users to leverage Hadoop MapReduce using a SQL interface, enabling analytics at a massive scale, in addition to distributed and fault-tolerant data warehousing.

HBase – An open source, non-relational, versioned database that runs on top of Amazon S3 (using EMRFS) or the Hadoop Distributed File System (HDFS). HBase is a massively scalable, distributed big data store built for random, strictly consistent, real-time access for tables with billions of rows and millions of columns.

Zeppelin – An interactive notebook that enables interactive data exploration[1]

V. CONCLUSION

In this paper, Hadoop was a major development in the big data space. In fact, it's credited with being the foundation for the modern cloud data lake. Hadoop democratized computing power and made it possible for companies to analyze and query big data sets in a scalable manner using free, open source software and inexpensive, off-the-shelf hardware.

With the introduction of Hadoop, organizations quickly had access to the ability to store and process huge amounts of data, increased computing power, fault tolerance, flexibility in data management, lower costs compared to DWs, and greater scalability. Ultimately, Hadoop paved the way for future developments in big data analytics, like the introduction of Apache Spark.

VI. References

[1] https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/

[2] https://www.javatpoint.com/what-is-hadoop

[3] https://sisudata.com/blog/what-is-hadoop-in-big-data-analytics

[4] https://www.techopedia.com/definition/27745/big-data

[5] https://www.geeksforgeeks.org/hadoop-an-introduction/

[6] https://www.databricks.com/glossary/hadoop