Big Data Analytics Using HADOOP Technology

[1]K. Saraswathi, [2]P. Ganeshbabu, [3]V. Vaneeswari, [4]S. Ranichandra

*Department of computer Applications,Dhanalakshmi Srinivasan College of Arts and Science for Women,,Perambalur , 621 212, Tamilnadu, , India.

Email: saraswathik5511@yahoo.com (*Saraswathi K) Corresponding author: Saraswathi K*

ABSTRACT

Using a network of many computers to address issues requiring enormous volumes of data and computing is made easier with Apache Hadoop, an open-source software collection. The MapReduce programming approach is used to create a software framework for distributed storage and processing of massive data. Big data is a way for storing, disseminating, and analysing large datasets at a high rate of speed. Big data may be organised, unstructured, or semi-structured, making it impossible to use traditional records management procedures to deal with it all. The Hadoop Distributed File System (HDFS) and the MapReduce programming style are at the heart of Apache Hadoop's architecture. Large chunks of data are broken up into smaller chunks and distributed across the nodes in a cluster using Hadoop. The data is subsequently processed in parallel by transferring the packed code to nodes. Because of data locality, nodes can only alter the data they have. A more typical supercomputer design that depends on a parallel file system and distributes computation and data across high-speed networking enables the dataset to be handled quicker and more effectively. This article discusses the technology of big data and its challenges, as well as the answer to the problem that is Hadoop framework and its applications.

Keywords:-Hadoop, HDFS, MapReduce, and YARN are some of the terms used to describe big data.

1. INTRODUCTION TO BIG DATA

Modules in the core Apache Hadoop framework include: the libraries and utilities used by other Hadoop modules may be found here. Commodity machines are used to store data in the Hadoop Distributed File System (HDFS), which has a high aggregate bandwidth throughout the cluster. Hadoop YARN is a framework that manages computing resources in clusters and uses them to schedule programmes for users; it was announced in 2012. For big data

processing, Hadoop MapReduce is a MapReduce programming paradigm implementation. There are a number of additional software packages that can be installed on top of or alongside the Hadoop platform that are called Hadoop's ecosystem, such as Apache Pig, Apache Hive, Apache HBase, Apache Phoenix and Apache Spark; Cloudera Impala and Flume; Apache Sqoop and Oozie; Oozie and Storm; and ZooKeeper. The MapReduce and HDFS components of Apache Hadoop were derived from Google publications on MapReduce and Google File System [1]-[5]

The Hadoop framework is mostly written in Java, with a small amount of native C code and shell scripts for command-line tools. Hadoop Streaming may construct the map and reduce portions of the user's application using any programming language, not only MapReduce Java code. Users of other Hadoop projects have access to more advanced features.

Nowadays, almost everyone carries a smartphone around with them. A single mobile phone user generates around 40 exabytes of data every day in the form of text, calls, messages, photos, recordings, searches, and music. Despite the fact that there are billions of smartphone users, conventional computer systems are incapable of handling the vast amounts of data generated by smartphones. The term "Big Data" is used to describe such a large volume of data. There are millions of people who use Facebook, YouTube, and Google every day, which creates a lot of data. Because of this, big data is created by giant organisations to evaluate vast amounts of organised, unstructured, or semi-structured data [6]-[15].

THE BIG DATA TYPES:

An example of structured data would be data that can be stored, referenced, or controlled according to a predetermined arrangement.

Using relational data as an example, The term "unstructured" refers to data that has no established form or organisation. [6] Data in XML format, for instance. Unstructured and structured data may be mixed and matched to create semi-structured data.

Word, PDF, Text, and Media Logs are all examples.

HADOOP HISTORY

Hadoop's co-founders, Doug Cutting and Mike Cafarella, claim that the Google File System article released in October 2003 was the inspiration for the project. [16] [17] A follow-up study from Google, "MapReduce: Simplifying Data Processing on Large Clusters," was based on this one. [18] After the Apache Nutch project was abandoned in January 2006,

development was transferred to the new Hadoop sub-project. [19] Named after Doug Cutting's son's toy elephant when he worked at Yahoo! [20] For HDFS and MapReduce, around 5,000 and 6,000 lines of code respectively were factored out of Nutch.

As of April 2006, Hadoop 0.1.0 has been released.[21] Owen O'Malley was the first committer to join the project in March 2006.

[22] The project is constantly evolving as a result of the contributions that are being made. [23]

In 2002, Doug Cutting and Mike Cafarella began working on the Apache Nutch project, which led to the creation of Hadoop. In order to construct a search engine that could index 1 billion documents, the Apache Nutch project was launched. Using Nutch, they determined that a system like this would cost nearly half a million dollars in hardware and a monthly operating cost of $30, 000, which is a lot of money. Since the web contains billions of pages, it became clear to them that their project design would not be able to handle the problem. As a result, they were searching for a viable solution that might lower the installation costs and the difficulty of storing and analysing massive datasets.

While researching Google's GFS (Google File System) distributed file system in 2003, they came upon a document describing the architecture of Google's distributed file system. Now that they've read this article, they've come to recognise that it can help them handle the issue of storing enormous files created by web crawling and indexing operations. This study, on the other hand, was just a partial answer.

That year, Google issued a new document on how to handle these massive datasets using MapReduce: Now Doug Cutting and Mike Cafarella have another half-solution for their Nutch project thanks to this research. GFS and MapReduce were only white papers at Google at the time. There were two strategies that Google did not use. As Lucene (a free and open-source information retrieval software library, initially created in Java by Doug Cutting in 1999) showed him, open-source is a terrific method to get the technology into the hands of as many people as possible. So, with the help of Mike Cafarella, he began incorporating Google's open-source methodologies (GFS and MapReduce) in the Apache Nutch project.

At the beginning of 2009, Hadoop successfully proved the ability to sort one petabyte (PB) of data within 17 hours in order to handle billions of queries and index million web pages. And Doug Cutting departed Yahoo to join Cloudera in order to take on the task of bringing Hadoop to a broader audience. Apache Hadoop version 1.0 was published by the Apache Software Foundation in December 2011. Then, in August 2013, the latest version, 2.0.6, was released.

Apache Hadoop 3.0 was published in December of 2017 and is the most recent version available.

NEEDS

Before I go into Hadoop, I'd want to discuss big data, which is a fascinating topic. Because, as you may know, there's a lot of data out there. It's hard to imagine how much information can be gleaned from just social media, search engine inquiries, and emails. Unstructured data (pictures, videos) and structured data (excel records) are being created at an ever-increasing pace in today's world. In the past, there was a lot less data, and it was organised. There were no more than two computers needed for the storing and processing of that data, making it simpler. However, as the amount of data grows, so does the difficulty of storing and processing it. The term "Big data" refers to data that cannot be stored, processed, or analysed in typical databases. Then, how do we store and handle this information? Hadoop is our tool of choice for this task. Hadoop is a system for storing, processing, and analysing massive amounts of data.

Hadoop solves the problems that conventional systems encounter when it comes to storing and processing large amounts of data. There are a variety of units inside Hadoop that can store and handle large amounts of data. Data is stored in a distributed manner using Hadoop HDFS. Data is split down into smaller parts and stored on several computers in this environment. Using Hadoop MapReduce, numerous computers work together to process large amounts of data in a parallel method. This solution is simpler and quicker to implement. Hadoop's resource allocation component is called YARN (Hadoop YARN).

FEATURES

Hadoop Features and Design Principles - Goal

In this lesson, we'll cover the features, characteristics, and design concepts of the Hadoop platform. This feratures of hadoop blog will also address the underlying assumptions on which Hadoop was developed. Start by learning about the features of Hadoop and its design concepts.

Hadoop's features

Apache MapReduce and YARN are the most popular and capable big data processing engines in the world. Hadoop offers the world's most dependable storage layer (HDFS) and a batch processing engine (MapReduce). In this part of Hadoop features, we'll focus on the following significant aspects of Hadoop:

Open Source

This project, Apache Hadoop, is free and open-source software. It implies that the code may

be altered to meet the needs of the company.

Processing Done in Parts

Data is handled in parallel on a cluster of nodes because HDFS stores data in a distributed fashion throughout the cluster.

Tolerance for Errors

This is one of the most crucial aspects of Hadoop's architecture. In Hadoop, by default, three copies of each block are kept throughout the cluster, although this may be altered if necessary. This means that if a node goes down, data on that node may be simply retrieved from other nodes using this feature. The system automatically recovers from failures of nodes or jobs. This is how Hadoop is able to withstand failures.

Reliability

Data is reliably kept on the machine cluster despite machine failures due to replication of data in the cluster. This characteristic of Hadoop ensures that even if your system fails, your data will still be accessible.

Intuitive Accessibility

In the event of a hardware failure, the data is readily available and accessible. If a computer or a few pieces of hardware go down, the data will be accessible through a different route.

Scalability

New hardware may be added to the nodes of Hadoop with ease. In addition, this feature of Hadoop allows horizontal scalability, which means more nodes may be added on the go without any downtime.

Economic

Apache Hadoop is a low-cost solution since it operates on commodity hardware in a cluster. It does not need the use of a specialist machine. Hadoop also saves a lot of money since it is so simple to add more nodes on the fly. As a result, if your needs grow, you can simply add more nodes without any downtime or extensive preparation.

It's simple to use.

The framework handles everything, so there's no need for a client to deal with distributed computing. As a result, this functionality of Hadoop is straightforward to implement.

The location of the data

This is one of Hadoop's distinctive properties, which has enabled it to manage Big Data with ease. Hadoop adheres to the idea of data localization, which argues that computing should be moved near the data rather than the other way around. Instead of sending data to where an

algorithm is submitted and then processing it, a MapReduce algorithm is transported to the cluster and then processed.

What sets Hadoop apart from other data management solutions are these characteristics. Here are several Hadoop Assumptions that must be taken into account before utilising Hadoop.

CHALLENGES OF BIG DATA

Data volume is referred to as "volume." From megabytes and gigabytes to petabytes, the data may be measured.

The data is too huge because of the variety. Many formats and types of documents are available; they may be organised or unstructured and include anything from audio and video to log files and other data.

velocity provides information about the rate at which data is being processed. Time-critical data is coming in at a rapid pace.

There is a tremendous amount of value in Big data. Big data relies on the ability to store a large number of values in a database, which is critical for businesses and IT systems.

There is a lot of noise, biases, and abnormalities when dealing with a large amount of information.

Challenges in Data Analysis

What if the amount of data becomes so huge and long that it isn't considered the best approach to deal with it?

Is there any way the information may be put to good use?

Is every piece of data eager to be mined?

What is the best way to identify the most important data points?

The question is whether or not all of the necessary data has been cached.

Manpower, Personnel, and Human Capital

Organizations and young individuals with various new skill sets will be drawn to Big Data because of its increasing period. To be successful, these skills should not only include technical ones, but also include analytical, interpretative, and creative ones. Tutoring initiatives should be undertaken by the organisations in order to develop these skills in people. It's also important for universities to educate their students about Big Data so that they can produce a highly skilled workforce.

There are many different kinds of data.

Unstructured data encompasses a wide range of records, from social media interactions to taped meetings, to PDF files, fax transfers, and emails. Managing unstructured data is a large

and expensive challenge. Unstructured data cannot be converted into structured data, and this is also not possible. It's common for structured data to be arranged in a highly automated and manageable way. It seems to be well integrated with the database, however the data is completely unstructured and fresh.

HADOO PHYSICAL SPECIFICATIONS

Economical: Product equipment is used in Hadoop (like your PC, computer). The expense of assuming responsibility for a Hadoop-based activity is modest. Hadoop environments are easier to maintain and more cost-effective to run. In addition, Hadoop is free software, therefore there is no licence fee.

Integration with cloud-based services: Hadoop offers built-in capability that integrates effortlessly with cloud-based services. In other words, if you're running Hadoop on a cloud, you don't have to worry about the scalability issue since you can order more hardware and expand your system in minutes whenever necessary.

Hadoop's ability to handle a broad variety of information makes it a flexible tool. Hadoop can store and process any sort of data, whether it is established, semi-established, or unstructured records, as we have previously discussed "Variety."

For example, if a computer goes down, any of the other machines will assume the responsibility and perform in a fault-tolerant and dependable manner. There are inherent fault tolerance features in the Hadoop system, therefore it is very dependable.

Resource management is handled by the YARN layer of Hadoop known as Yet Another Resource Negotiator (YARN). Resource management and activity scheduling/monitoring features are separated into distinct daemons in YARN. There may be a single global Resource Manager and Application Master in YARN, which is uniform across all applications. An Application may be either a single job or a DAG of several tasks. There are two daemons in the YARN framework: the Resource Manager and the Node Manager. In order to keep the system running, the Resource Manager acts as an arbitrator between all of the competing applications.

YARN is made up of the following components:

1) The person in charge of managing the company's resources (one according to cluster)

2) The Master of Applications (one per application)

There are three types of node managers: (one consistent with node)

The person in charge of managing the company's resources

The Resource Manager is responsible for allocating the cluster's resources and keeping track of how much resources each node supervisor contributes. It's made up of two major parts:

It does not monitor or measure the popularity of the programmes it allocates resources to, but it does schedule resources according to the requirements of, various walking programmes.

Monitoring and resuming application masters in the event of a failure are two responsibilities of an Application Manager.

The Application Master works with the scheduler to obtain the required resources and controls the useful resource requirements of discrete apps. In order to perform and monitor responsibilities, it connects to the node supervisor

To communicate the state of a node to the resource management, the Node Manager watches ongoing tasks and delivers signals (or heartbeats). It also keeps track of how much resources each container is using. .

BIG DATA AND HADOOP IN DIFFERENT DOMAINS

Here, we'll examine how Hadoop is helping organisations address their challenges and where Hadoop applications are being used.

Finance and Banking

A variety of challenging scenarios confront the banking and finance industry, including card fraud, tick analytics, audit trail archiving and business credit risk reporting. For early detection of security fraud and alternate visibility, Hadoop is being used by the team at IBM. For pre-trade decision-making analytics, they employ Hadoop to process and evaluate customer data for improved insights, among other things.

Media, communication, and leisure

Finding trends in real-time media use, leveraging social media, and phone content are just some of the issues that the media, entertainment, and conversation sectors confront. These companies use Hadoop to better understand their customers' data and develop content for their target audiences. Wimbledon Championships, for example, uses a massive amount of data to gently complete assumption examinations in real time for customers during tennis events.

Those who work in the medical field

Unstructured data, such as patient records and illness case histories, may be analysed using Hadoop in the healthcare industry. They are able to effectively treat patients based only on past case histories because of this. Precautions may be taken and treatments can be made accessible to a particular area by recognising the sickness that is prevalent there. Public health records and Google Maps are two tools used by the University of Florida to show data that helps researchers more quickly determine where chronic illnesses are spreading[16-25].

Big data is used extensively in the education industry. With almost 26,000 students, the University of Tasmania has implemented a Learning Management System (LMS) that keeps track of how much time each student spends on various websites and their overall progress over time.

Government

A vast amount of data is being generated by a variety of government programmes. Faster treatment responses are one of the goals of the FDA's use of Big Data to find and examine the types of food-associated disorders.

CONCLUSION

A general summary of big data, including its properties and benefits, is provided here. A variety of challenging scenarios confront the banking and finance industry, including card fraud, tick analytics, audit trail archiving and business credit risk reporting. For early detection of security fraud and alternate visibility, Hadoop is being used by the team at IBM. Hadoop, a technique for processing large amounts of data, has also been discussed in the study. Big data is used extensively in the education industry.

With almost 26,000 students, the University of Tasmania has implemented a Learning Management System (LMS) that keeps track of how much time each student spends on various websites and their overall progress over time. Hadoop technology and its features and modules have been discussed in this article. Big data and Hadoop technologies are explained in detail in the paper's applications.

REFERENCES

[1] SIMPLILEARN (2019, DECEMBER 10), BIG DATA IN 5 MINUTES | WHAT IS BIG DATA?| INTRODUCTION TO BIG DATA

[2] |BIG DATA EXPLAINED RETRIEVED FROM

[3] Https://Youtu.Be/Bayrobl7tye

[4] https://www.tutorialspoint.com/hadoo p/hadoop_big_data_overview.htm

[5] Kaur, I. (2016). Navneet kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur,“Research Paper on Big Data and Hadoop”. International Journal of Computer Science and Technology, 7940.

[6] Bobade, V. B. (2016). Survey paper on big data and Hadoop. International

[7] Research Journal of Engineering and Technology (IRJET), 3(01).

[8] https://www.guru99.com/what-is-big- data.html

[9] Talha, M., Elmarzouqi, N., & Abou El Kalam, A. (2020). Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution. J. Ubiquitous Syst. Pervasive Networks, 12(1), 9-15.

[10] https://www.datamation.com/big- data/big-data-pros-and-cons.html

[11] https://magnimindacademy.com/what- are-the-advantages-and-disadvantages-of-big-data/

[12] S.Kannadhasan and R.Nagarajan, Development of an H-Shaped Antenna with FR4 for 1-10GHz Wireless Communications, Textile Research Journal, DOI: 10.1177/00405175211003167 journals.sagepub.com/home/trj, March 21, 2021, Volume 91, Issue 15-16, August 2021, Sage Publishing

[13] S.Kannadhasan and R,Nagarajan, Performance Improvement of H-Shaped Antenna With Zener Diode for Textile Applications, The Journal of the Textile Institute, Taylor & Francis Group, DOI: 10.1080/00405000.2021.1944523

[14] Bhosale, H. S., & Gadekar, D. P. (2014). A review paper on big data and hadoop. International Journal of Scientific and Research Publications, 4(10), 1-7.

[15] Tamilselvi, K., Sumithra, V., &Dhanapriyadharsini, M. K. (2018). Big Data Analytics Using Hadoop Technology.

[16] Chandrika, V. S., Thalib, M. M., Karthick, A., Sathyamurthy, R., Manokar, A. M., Subramaniam, U., & Stalin, B. (2021). Performance assessment of free standing and building integrated grid connected photovoltaic system for southern part of India. Building Services Engineering Research and Technology, 42(2), 237-248.

[17] Naveenkumar, R., Ravichandran, M., Mohanavel, V., Karthick, A., Aswin, L. S. R. L., Priyanka, S. S. H., ... & Kumar, S. P. (2022). Review on phase change materials for solar energy storage applications. Environmental Science and Pollution Research, 29(7), 9491-9532.

[18] Mohan Kumar, A., Rajasekar, R., Manoj Kumar, P., Parameshwaran, R., Karthick, A., Mohanavel, V., ... & Muhibbullah, M. (2021). Investigation of drilling process parameters of Palmyra based composite. Advances in Materials Science and Engineering, 2021.

[19] Muthuraman, U., Shankar, R., Nassa, V. K., Karthick, A., Malla, C., Kumar, A., ... & Bharani, M. (2021). Energy and economic analysis of curved, straight, and

spiral flow flat-plate solar water collector. International Journal of Photoenergy, 2021.

[20] Karthick, A., Kalidasa Murugavel, K., Sudalaiyandi, K., & Muthu Manokar, A. (2020). Building integrated photovoltaic modules and the integration of phase change materials for equatorial applications. Building Services Engineering Research and Technology, 41(5), 634-652.

[21] Sathish, T., Mohanavel, V., Ansari, K., Saravanan, R., Karthick, A., Afzal, A., ... & Saleel, C. A. (2021). Synthesis and characterization of mechanical properties and wire cut EDM process parameters analysis in AZ61 magnesium alloy+ B4C+ SiC. Materials, 14(13), 3689.

[22] Singh, D., Chaudhary, R., & Karthick, A. (2021). Review on the progress of building-applied/integrated photovoltaic system. Environmental Science and Pollution Research, 28(35), 47689-47724.

[23] Stalin, B., Ravichandran, M., Sudha, G. T., Karthick, A., Prakash, K. S., Asirdason, A. B., & Saravanan, S. (2021). Effect of titanium diboride ceramic particles on mechanical and wear behaviour of Cu-10 wt% W alloy composites processed by P/M route. Vacuum, 184, 109895.

[24] Naveenkumar, R., Ravichandran, M., Stalin, B., Ghosh, A., Karthick, A., Aswin, L. S. R. L., ... & Kumar, S. K. (2021). Comprehensive review on various parameters that influence the performance of parabolic trough collector. Environmental Science and Pollution Research, 28(18), 22310-22333.

[25] Kumar, P. M., Saravanakumar, R., Karthick, A., & Mohanavel, V. (2022). Artificial neural network-based output power prediction of grid-connected semitransparent photovoltaic system. Environmental Science and Pollution Research, 29(7), 10173-10182.