MAKERERE UNIVERSITY
COLLEGE OF COMPUTING & INFORMATION SCIENCES
SEMESTER 1 FINAL EXAM 2021/2022
ADVANCED TOPICS IN COMPUTER SCIENCE
HADOOP
WENAREEBA INNOCENT | REG NO : 2021/HD05/2315U | MCS 710

1.0 Abstract

Big Data analytics has become a necessary tool for the industry for processing, storing and analyzing large data volumes. Large volumes of data may now be examined more easily thanks to distributed computing frameworks like Hadoop which have become popular because of its API performance and availability and traditional data systems cannot handle this kind of processing. This research explores the impact of Hadoop in BigData, the technology's benefits and drawbacks, applications, and future predictions on how Hadoop will be adopted by the Big Data world.
Keywords: Hadoop, Big data

2.0 Introduction

For decades, the primary goal of computing has been to enhance the computational capability of a single machine and to develop supercomputers with incredible processing speed and memory. Traditional computing was entirely process-based, requiring a lot of complicated processing on very tiny amounts of data and current systems have to deal with big data processing. With the rise of many platforms like social media , search engines and streaming , big data is being created , and hadoop has the capability of handling large data volumes and providing significant processing and analytical capabilities, which has made it very popular. This paper describes how Hadoop is being utilized in big data in the current world and how the future adopts it.

3.0 Applications of the technology

(i) Finance sectors, Finance companies utilize Hadoop for fraud detection, such as recognizing fraud trends, preventing unauthorized users, and patterns, which helps firms decrease risks in their systems.

(ii) Security and Law Enforcement, Security agencies use Hadoop in detecting cyber crimes and preventing them. Police departments also use these Big Data hadoop techniques to predict criminal activities.

(iii) Government sectors, Governments' use Hadoop to analyze country , city data , and state data for easy planning and development. It's utilized in traffic data analysis which improves transportation in the city.

(iv) Healthcare sectors , In the health-care industry, Hadoop plays a critical role in improving public health by analyzing massive data volumes from lab findings , medical devices and

imaging reports which helps healthcare providers treat patients more effectively and enhance public health using the information gleaned from the analysis.

4.0 Advantages of Hadoop

(i) Cost-effective , Hadoop stores its data in commodity hardware clusters. There are few machines needed when storing this data since there is a reduced data volume.

(ii) Performance , Its data processing system that can handle large amounts of data quickly. It separates jobs into many nodes, which run in parallel, resulting in improved performance.

(iii) Fault-Tolerant ,In Hadoop, the parity blocks and the remaining data blocks can be used to recover the data block that was damaged by a node failure.

(iv) Open source , The code is always freely available and developers or data scientists can manipulate it to suit their needs which makes Hadoop readily available for all users.

(v) Scalable , Horizontal scaling is the principle hadoop works on , it allows nodes to be appended to the clusters of hadoop , and this makes the framework more scalable.

5.0 Disadvantages of hadoop

(i) Issue With Small Files , With programs that handle a large number of small files, Hadoop performs really badly.The Namenode becomes overburdened as a result of the enormous number of little files, making Hadoop difficult to use.

(ii) Vulnerable , Since it is written in a programming language , java and java is a common language , it means it's easily attacked by cyber criminal personnel and it's vulnerable to threats .

(iii) Processing Overhead , Hadoop handles reading of data from the disk and then writes it to the disk , while dealing with terabytes and petabytes of data, and this makes read/write (rw) operations extremely expensive.

(iv) Supports Only Batch Processing, Hadoop is based on a batch processing engine that is inefficient when dealing with streams.It cannot produce output with little delay in real time. Only the data saved in the files before processing is what Hadoop works with.

6.0 How technology works
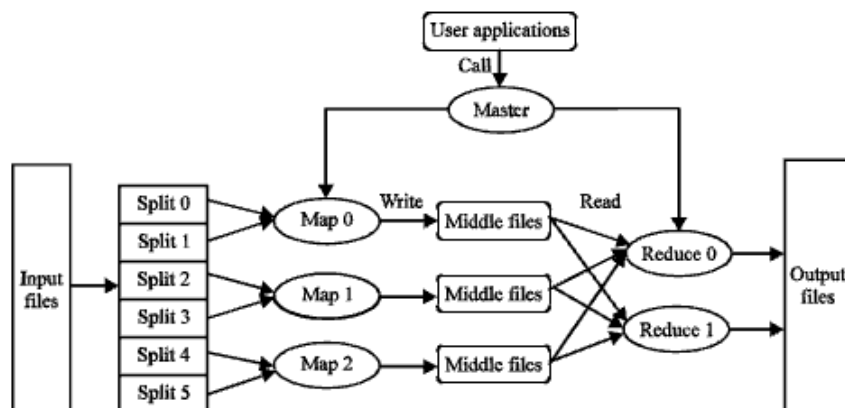
7.0 Future of Hadoop

Today, Big Data is having an impact on the IT sector. Sensor-enabled machinery, mobile devices, cloud computing, and satellites generate vast amounts of data that help various firms enhance their making decisions and push business to be better. The technology has the capacity to transform how governments, businesses,and individuals produce discoveries, it is expected to transform how everyone lives their daily lives.

8.0 Conclusion

Managing data is the big issue. As a result of the massive volume of generated data throughout the data creation process, the concept of big data has taken a prominent position in the scene, Hadoop is the technology being used for data management. Hadoop can manage massive amounts of data at a low cost, and it can also handle these massive amounts of data at a very quick processing speed and it creates copies incase of system failure.Data volume , data source format , speed , security and privacy, cost, connectivity, and data sharing are all aspects that influence the performance of scheduling strategies and Hadoop is continuously being improved to make it better and be able to address these issues.

10.0 References

[1] International Journal of Advanced Research in Science and Engineering , Volume no.07, Special Issue No.02, February 2018, Research paper on Big Data and Hadoop Map reduce , Real time Scheduling

[2] International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015 ISSN 2229-5518 , Advantages of Hadoop

[3] Top 12 Real Time Big Data Hadoop Applications ,
https://data-flair.training/blogs/hadoop-applications/

[4] Features & Applications of Hadoop,
https://www.upgrad.com/blog/features-applications-of-hadoop/

[5] Journal of Scientific Research and Reports , November 2020, DOI:10.9734/JSRR/2020/v26i930310 , Using Hadoop Technology to Overcome Big Data Problems by Choosing Proposed Cost-efficient Scheduler Algorithm for Heterogeneous Hadoop System (BD3)

[6] International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016