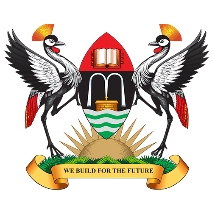
**MAKERERE UNIVERSITY  
COLLEGE OF COMPUTING & INFORMATION SCIENCES**SEMESTER 1 FINAL EXAM 2021/2022  
ADVANCED TOPICS IN COMPUTER SCIENCE  
**HADOOP** WENAREEBA INNOCENT | REG NO : 2021/HD05/2315U | MCS 710

**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 data processing. This paper explores the impact of big data, Hadoop, 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, Map Reduce***

**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.

**5.0 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,** Hadoop is used by the US National Security Agency to prevent terrorist attacks and detect and prevent cyber-attacks. Police departments also employ Big Data techniques to apprehend criminals and even predict criminal activities.  
**(iii) Government sectors,** Hadoop is used by the government to analyze massive volumes of data for the development of the country, states, and cities , it is also used 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.

**6.0 Advantages of Hadoop**

### 1. Varied Data Sources

**Hadoop accepts a variety of data. Data can come from a range of sources like email conversation, social media etc. and can be of structured or unstructured form. Hadoop can derive value from diverse data. Hadoop can accept data in a text file, XML file, images, CSV files etc.**

### 2. Cost-effective

**Hadoop is an economical solution as it uses a cluster of commodity hardware to store data. Commodity hardware is cheap machines hence the cost of adding nodes to the framework is not much high. In Hadoop 3.0 we have only 50% of storage overhead as opposed to 200% in Hadoop2.x. This requires less machine to store data as the redundant data decreased significantly.**

### 3. Performance

**Hadoop with its distributed processing and distributed storage architecture processes huge amounts of data with high speed. Hadoop even defeated supercomputer the fastest machine in 2008. It divides the input data file into a number of blocks and stores data in these blocks over several nodes. It also divides the task that user submits into various sub-tasks which assign to these worker nodes containing required data and these sub-task run in parallel thereby improving the performance.**

### 4. Fault-Tolerant

**In Hadoop 3.0 fault tolerance is provided by erasure coding. For example, 6 data blocks produce 3 parity blocks by using erasure coding technique, so HDFS stores a total of these 9 blocks. In event of failure of any node the data block affected can be recovered by using these parity blocks and the remaining data blocks.**

### 5. Highly Available

**In Hadoop 2.x,** [HDFS architecture](https://data-flair.training/blogs/hadoop-hdfs-architecture/) **has a single active NameNode and a single Standby NameNode, so if a NameNode goes down then we have standby NameNode to count on. But Hadoop 3.0 supports multiple standby NameNode making the system even more highly available as it can continue functioning in case if two or more NameNodes crashes.**

### 6. Low Network Traffic

**In Hadoop, each job submitted by the user is split into a number of independent sub-tasks and these sub-tasks are assigned to the data nodes thereby moving a small amount of code to data rather than moving huge data to code which leads to low network traffic.**

### 8. Open Source

**Hadoop is an open source technology i.e. its source code is freely available. We can modify the source code to suit a specific requirement.**

### 9. Scalable

**Hadoop works on the principle of horizontal scalability i.e. we need to add the entire machine to the cluster of nodes and not change the configuration of a machine like adding RAM, disk and so on which is known as vertical scalability. Nodes can be added to** [Hadoop cluster](https://data-flair.training/blogs/hadoop-cluster/) **on the fly making it a scalable framework.**

**7.0 Disadvantages of hadoop**

### 1. Issue With Small Files

**Hadoop is suitable for a small number of large files but when it comes to the application which deals with a large number of small files, Hadoop fails here. A small file is nothing but a file which is significantly smaller than Hadoop’s block size which can be either 128MB or 256MB by default. These large number of small files overload the Namenode as it stores namespace for the system and makes it difficult for Hadoop to function.**

### 2. Vulnerable By Nature

**Hadoop is written in** [Java which is a widely used programming language](https://data-flair.training/blogs/java-tutorial/) **hence it is easily exploited by cyber criminals which makes Hadoop vulnerable to security breaches.**

### 3. Processing Overhead

**In Hadoop, the data is read from the disk and written to the disk which makes read/write operations very expensive when we are dealing with tera and petabytes of data. Hadoop cannot do in-memory calculations hence it incurs processing overhead.**

### 4. Supports Only Batch Processing

**At the core, Hadoop has a batch processing engine which is not efficient in stream processing. It cannot produce output in real-time with low latency. It only works on data which we collect and store in a file in advance before processing.**

### 6. Security

[For security](https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.6.5/bk_security/content/ch_hdp-security-guide-overview.html)**, Hadoop uses Kerberos authentication which is hard to manage. It is missing encryption at storage and network levels which are a major point of concern.**

**So, this was all about Hadoop Pros and Cons. Hope you liked our explanation.**

**8.0 Future of Hadoop**

**9.0 Conclusion**

**10.0 References**

[1]Jonal of Big data research , open access, A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBenc , 2020, N. Ahmed, Andre L. C. Barczak, Teo Susnjak, and Mohammed A. Rashid

[2] 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

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

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

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