Big Data Machine Learning And Hadoop

Aliyu Bello

***Abstract*—Big Data applications have grown increasingly essential in recent years due to the exponential growth in data capture and generation. By empowering process optimization, enhancing insight discovery, and improving decision making, the big data revolution offers to redefine how we live, work, and think.This paper provides a review of recent Big Data technology developments. It shows a description of Big Data techniques, Machine learning, and Hadoop and their components. Research was made on both, followed by experiments and how effective they are on Big Data. The findings indicated how the techniques and their components played a significant role through the years on Big Data.**

***Index Terms*—Big Data, Machine Learning, Hadoop.**

I. Introduction

THE Big Data refers to the massive amount of data that is generated in this modern age. From engineering sciences to social networks, business, research, and security, large-scale data sets are collected and evaluated in a variety of fields. Digital data, which is created by a wide range of digital devices, is rising at a staggering rate. The amount of data generated and saved in the digital world has increased dramatically in a short period of time. As a result, the rapid growth of data has caused numerous challenges. Many Big Data models, frameworks, and new technologies have been developed as a result of various Big Data projects around the world to provide additional storage capacity, parallel processing, and real-time analysis of various sources[1]. Big Data is often defined by the well-known 5V features; Volume, Velocity, Variety, Veracity, Value.

Machine learning techniques have been applied in a number of vast and complicated data-intensive sectors during the last decade, such as medicine, astronomy, biology, and so on since they present potential solutions for extracting the information stored in the data. To discover the global optima, machine-learning-based models are known to have a high computational cost. As a result of the learning job in an extensive dataset, the number of hidden nodes inside the network will grow considerably, resulting in an overall increase in computing complexity. Organizations may work more efficiently and gain a competitive advantage by utilizing big data and machine learning.

Hadoop is an open-source system that provides the distributed processing of massive data volumes using basic programming across clusters of machines. It has outstanding scalability, allowing us to scale from a single server to hundreds of servers based on our demands. Hadoop, unlike older techniques, does not copy the entire remote data into memory to do computations. Hadoop, on the other hand, performs jobs that store data.

Many Big Data models, frameworks, and new technologies have emerged as a result of various Big Data projects around the world, allowing for increased storage capacity, parallel processing, and real-time analysis of various sources. New data privacy and security solutions have also been developed. In this paper, we provide research findings of modern Big Data technologies. We classify and compare them based on their intended use, benefits, limitations, and features.

Aliyu Bello, March, 2022

II. Big Data And Machine Learning

Self-learning algorithms are at the core of machine learning, and they develop by continuously improving at their assigned task. Machine learning algorithms can produce outcomes in the settings of pattern recognition and predictive modeling when properly structured and given data. As the volume of training datasets expands, machine-learning algorithms become increasingly effective. When big data and machine learning are combined, we gain two benefits: we can keep up with the continuous volume of data due to the algorithms, and the volume and variety of the same data stimulate the algorithms and help them improve. By applying machine learning to Big data, the expectation is to see a well-structured and analyzed result that can help with predictive modelings, such as hidden patterns and analytics[2].

Machine learning algorithms analyze incoming information and identify patterns, which are then converted into useful insights that may be applied to company operations. The algorithms were then utilized to automate parts of the decision-making process.

A. Examples of machine learning in Big Data

- To classify crime data into categories, use an unsupervised machine learning technique; to cluster the crime data into risky, average, and safe regions, use the K-means clustering algorithm.
- To determine whether a specific place is risky or safe, use supervised machine learning techniques and the decision tree classification algorithm.

B. Machine learning Subdomains

In general, supervised learning, unsupervised learning, and reinforcement learning are the three subdomains of Machine Learning. Below is a table showing the application of machine learning in various data techniques.

Learning types	Data processing tasks	Distinction norm
Supervised learning	Classification/Regression/Estimation	Computational classifiers
		Statistical classifiers
		Connectionist classifiers
Unsupervised learning	Clustering/Prediction	Parametric
		Nonparametric
Reinforcement learning	Decision-making	Model-free
		Model-based

Fig. 1. Caption

Supervised learning necessitates the use of labeled data with inputs and outputs. Unsupervised learning, in contrast to supervised learning, does not require labeled training data, and the environment just gives inputs without the intended targets. Reinforcement learning allows you to learn from the feedback you get from your interactions with the outside environment. Many theory mechanisms and application services for dealing with data problems have been offered based on these three key learning paradigms. Subsection text here.

1) Deep Learning: The capacity of traditional machine-learning approaches and feature engineering algorithms to process natural data in its raw form is limited [3]. Deep Learning, on the other hand, is more powerful in resolving data analytical and learning challenges in large data sets. In reality, it aids in the extraction of sophisticated data representations from vast volumes of unstructured and uncategorized raw data automatically.

C. Challanges Of Machine Learning

Designing scalable and flexible computational architectures for machine learning, as well as the capacity to grasp data features before applying machine learning algorithms and tools, are all common machine learning issues.

- A machine learning algorithm that has been trained on a certain labeled dataset may not be applicable for another dataset, and the classification may not be consistent across datasets.
- A machine learning approach is built for a single learning task. It is ineffective for today's many learning tasks and knowledge transfer requirements of Big data analytics[4].
- Since a machine learning approach is typically trained on a limited number of class types, a constantly evolving dataset with a wide variety of class types will result in incorrect classification results.

The usage of the Big Data concept to categorize the machine learning challenges allows for the construction of cause-and-effect relationships for each issue. Furthermore, the establishment of explicit relationships between approaches and issues allows for a more comprehensive knowledge of machine learning with Big Data.

III. Big Data And Hadoop

Hadoop is a well-known Big Data technology with a significant community of users. It was created to prevent the low performance and complexity that come with utilizing existing technologies to process and analyze Big Data[5]. Hadoop's key benefit is its ability to process enormous data sets quickly, thanks to its parallel clusters and distributed file system.
Hadoop is more than a single application; it's a platform with a number of interconnected components that allow for distributed data storage and processing. The Hadoop ecosystem is made up of these components. Some of these are vital components that form the framework's base, while others are additional components that add to Hadoop's functionality.

A. Key Components Of Hadoop

1) Hadoop Distributed File System (HDFS): Hadoop's distributed file system is maintained by Hadoop's pillar, HDFS. It enables the storage and replication of data across several servers. It can accommodate hundreds of nodes in a cluster and offers cost-effective and dependable storage. It can store large amounts of both structured and unstructured data. It is more scalable, reliable, and distributed in Hadoop's framework context.

2) YARN: YARN is an acronym for Yet Another Resource Negotiator. It organizes and arranges resources, as well as determining what should happen in each data node. In comparison to MapReduce, it has higher scalability, parallelism, and improved resource management. It provides Big Data analytical applications with operating system features. It consists of multiple parts, including a resource manager, a node manager, and an application master. YARN Ressource Manager has been incorporated into Hadoop's design.

3) MapReduce: MapReduce is a programming model and its implementation in one framework. It is one of the first and most important steps in the development of a new generation of Big Data management and analytics software. To produce tuples, the Map function first groups, filters, and sorts numerous data sets in concurrent pairs, then the reduce function takes the output of a map as input and combines the data tuples into a smaller set. The reduction job is always carried out after the map job, as the term MapReduce implies. For Big Data applications, MapReduce has a unique advantage.

B. Challanges of Hadoop

Hadoop is meant to distribute massive volumes of data over a cluster of servers efficiently. And it was proven that Hadoop solves Big Data challenges such as Volume, Variety, Velocity, and Values, but it was found to be fraught with difficulties, some of which are listed below.

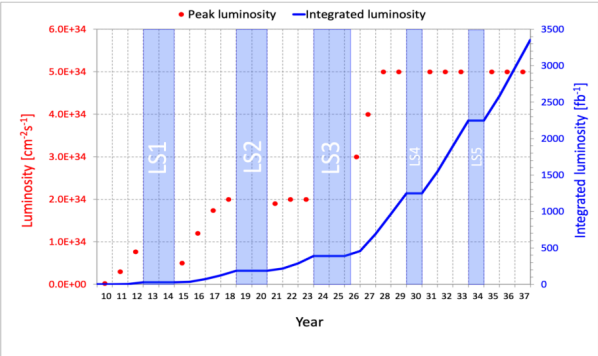
- Hadoop is more than just batch analytics and offline storage. Ingestion in real-time and in batches necessitates a close integration of numerous components.
- Finding the Root Cause of Problems is Difficult. End-to-end testing is either impractical or impossible to automate.

- A large number of small files will necessitate a large number of disc I/Os, resulting in more time being spent on I/O rather than data processing. [6] Also, each file in a MapReduce architecture is handled by a separate mapper; we must utilize an equivalent number of mappers, which results in significant overhead and lowers overall system performance.

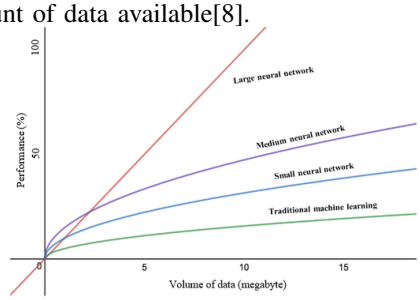
IV. Experiments And Discussion

A. Experiment of Machine Learning on Big data

The Large Hadron Collider (LHC) is the world's largest and most powerful particle collider. Two high-energy particle beams are traveling at nearly light speed. Processing and storing the massive amount of data collected at the LHC is a significant problem. During Run 2, the total amount of data flowed was around 25 GB/s[7].Due to the massive amount of data, a factor of ten increase in particle rate is projected during the High luminosity LHC era. During this time, machine learning and big data tools will be critical.



While machine learning techniques are being used in a variety of industries due to their extraordinary capacity to identify complicated relationships in vast datasets, the tightening of data ownership and privacy restrictions is making it more difficult to apply them to a limited number of data types. The graph above illustrates how machine learning algorithms scale with the amount of data available[8].



B. Experiment of Hadoop on Big Data

The hadoop process goes through phases. After the developer has finished storing and processing the data, it is ready for report generation. Prior to that, we must ensure that the processed data is accurate and that the data has been loaded and processed correctly. Testing in Hadoop includes Unit Testing, Regression Testing, System Testing, and Performance Testing, among many other types. Test Reports are summaries of everything done so far to complete the testing process. All of the planning, scripting, and execution of test cases, as well as the results we obtained, are all documented in the form of Test Reports.

V. Approaches to improve Big Data techniques

The challenges of Big Data include not just the need to overcome the 5V characteristics but also developing tools for data capture, transformation, integration, and modeling. Other critical challenges in Big Data analysis include privacy, security, governance, and ethical considerations.
An approach to data reduction for learning from Big Data sets by integrating stacking, rotation, and agent population learning techniques. An approach to data reduction for learning from Big Data sets integrates stacking, rotation, and agent population learning techniques. According to I. Czarnowski and P. Jedrzejowicz (2018), the method is based on the classifier ensemble paradigm, which ensures heterogeneity by stacking ensembles utilizing rotation-based approaches. Data reduction in an instance, as well as feature dimensions, have been used to reduce the data's dimensionality[9].
Organizations should also consider purchasing artificial intelligence/machine learning-powered knowledge analytics solutions. [10]Professionals who aren't data scientists but have a rudimentary understanding of the subject frequently use these Big Data Tools. This step assists businesses in saving a significant amount of money when it comes to recruiting.

VI. Conclusion

This paper gives an overview of machine learning and Hadoop in the context of Big Data. It has given an overview of machine learning approaches and addressed how these techniques address the many issues that have been found. Machine learning is at its core for its ability to study from data and give data-driven insights, actions, and predictions. Hadoop is the platform for managing Big Data, solving the problem, and making it useful for analytics. To build next-generation Big Data infrastructures, more work is needed in various areas, including data organisation, domain-specific tools, and platform tools.To build next-generation Big Data infrastructures, more work is needed in various areas, including data organization, domain-specific tools, and platform tools. Big Data allows for big research, analysis, and study, which leads to big changes to improve the lives of people and solve the world's mysteries.

VII. References

[1] B.M. Purcell. Big Data using cloud computing Holy Family Univ. J. Technol. Res. (2013) from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.8238rep=re
[2] Junfei Qiu, Qihui Wu, Guoru Ding. A survey of machine learning for big data Processing (2016). Retrived from https://link.springer.com/content/pdf/10.1186/s13634-016-0355-x.pdf
[3] Ahmed OussousFatima-Zahra BenjellounAyoub Ait Lahcen Big Data technologies: A survey (2019) https://reader.elsevier.com/reader/sd/pii/S1319157817300034?token=B3E1west-1originCreation=20220322181421

[4] Wang, Lidong; Alexander, Cheryl Ann. Machine Learning in Big Data. (2016) https://www.proquest.com/openview/f30188aaa2216f6d0cb73b75ff64c975/1?pq-origsite=gscholarcbl=5363605

[5] Ahmed OussousFatima-Zahra BenjellounAyoub Ait Lahcen Big Data technologies: A survey (2019) https://reader.elsevier.com/reader/sd/pii/S1319157817300034?token=B3E1A2DD908820822B22E12A19FA84EC0055595445C287C303F5F
west-1originCreation=20220322181421

[6] Kumar Chinnakali. The top 12 Apache Hadoop challenges. (Nov 23, 2015). Retrived from https://bigdata-madesimple.com/the-top-12-apache-hadoop-challenges/

Alfredo Castaneda. Overview of ML and Big data tools at HEP experiments (2019) https://indico.cern.ch/event/687651/contributions/3427643/attachments/1849430/3036967/LHC2019$_M LBigData_f inal.pdf$

[8] Tobore Igbe. Machine learning techniques scale(2019) https://www.researchgate.net/figure/How-machine-learning-techniques-scale-with-amount-of-data$_f ig2_3$35107259

[9] Piotr Jedrzejowicz, Kuo-Ming Chao. Overcoming "Big Data" Barriers in Machine Learning Techniques for the Real-Life Applications (2018). Retrieved from. https://www.hindawi.com/journals/complexity/2018/1234390/?utm$_s ource =$ $google utm_m edium = cpc utm_c ampaign = HDW_M RKT_G BL_S UB_A DW O_P AI_D Y NA_J OU R_{X P} J_G ROUP3 gclid =$ $EAIaIQobChM I_b Tvy4nk9gIVVZBoCR1PggzKEAMY ASAAEgLnqf D_B wE$

[10] Chandan Gaur. Top 6 Big Data Challenges and Solutions to Overcome (December 2020) Retrieved from https://www.xenonstack.com/insights/big-data-challenges