**HADOOP CLUSTER IN MACHINE LEARNING**

**Developed By**

**Name of the Students**

1. **Saptarshi Ghosh**
2. **Swapnil Sarkar**
3. **Mayuraksha Sikdar**
4. **Anushka Ghosh**

**Under the Supervision of**

**MRS. ANANYA BANERJEE**

**Assistant Professor, Department of Computer Science & Engineering**

Date: 15th November, 2019

**Hadoop Cluster in Machine Learning**

**A Dissertation Submitted in partial fulfillment for the Degree of Bachelor of Technology (B. TECH), 7th Semester in Computer Science & Engineering**

**Submitted By**

|  |  |  |
| --- | --- | --- |
| **Name** | **University Registration Number** | **University Roll Number** |
| Saptarshi Ghosh | 161270110091 of 2016-2017 | 430116010088 |
| Swapnil Sarkar | 161270110119 of 2016-2017 | 430116010117 |
| Mayuraksha Sikdar | 161270110050 of 2016-2017 | 430116010048 |
| Anushka Ghosh | 161270110016 of 2016-2017 | 430116020016 |

**Under the Supervision of**

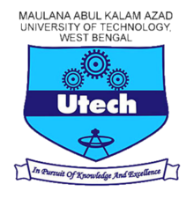
**MRS. ANANYA BANERJEE**

**Assistant Professor,**

**Department of Computer Science & Engineering**



**Narula Institute of Technology**



**Maulana Abul Kalam Azad University of Technology**

**(November, 2019)**

**CERTIFICATE OF ORIGINALITY**

The project entitled “Hadoop Cluster in Machine Learning” has been carried out by ourselves in partial fulfillment of the degree of Bachelor of Technology in Computer Science & Engineering of Narula Institute of Technology, Agarpara, Kolkata under Maulana Abul Kalam Azad University of Technology during the academic year 2016-2020

While developing this project no unfair means or illegal copies of software etc. have been used and neither any part of this project nor any documentation have been submitted elsewhere or copied as far in our knowledge.

Signature

Name: SAPTARSHI GHOSH

University Roll No.: 430116010088

University Registration No.: 161270110091 of 2016-2017

Signature

Name: SWAPNIL SARKAR

University Roll No.: 430116010117

University Registration No.: 161270110119 of 2016-2017

Signature

Name: MAYURAKSHA SIKDAR

University Roll No.: 430116010048

University Registration No.: 161270110050 of 2016-2017

Signature

Name: ANUSHKA GHOSH

University Roll No.: 430116010016

University Registration No.: 161270110016 of 2016-2017

**CERTIFICATE OF APPROVAL**

This is to certify that the project entitled “HADOOP CLUSTER IN MACHINE LEARNING” has been carried out by SAPTARSHI GHOSH, SWAPNIL SARKAR, MAYURAKSHA SIKDAR, ANUSHKA GHOSH, under my supervision in partial fulfillment for the degree of Bachelor of Technology (B.TECH) in Computer Science & Engineering of Narula Institute of Technology, Agarpara affiliated to Maulana Abul Kalam Azad University of Technology during the academic year 2016-2020

It is understood that by this approval the undersigned do not necessarily endorse any of the statements made or opinion expressed therein but approves it only for the purpose for which it is submitted.



-------------------------- -----------------------------

Mrs. Ananya Banerjee (External Examiner)

Assistant Professor,

Computer Science & Engineering

--------------------------------------------------

Mr. Jayanta Pal

Head of the Department

Computer Science & Engineering

**ACKNOWLEDGEMENT**

We owe our deep gratitude to our project guide Mrs. Ananya Banerjee (Asst. Professor of Department of Computer Science & Engineering, Narula Institute of Technology), who took a keen interest in our project work and guided us, by providing all the necessary information for developing a good system that became the base model for our work and also in overcoming the various challenges that we faced throughout the project.

**CONTENTS**

Abstracts

I. Introduction

II. Motivation

III. Survey

IV. Broad observations

V. Objectives

VI. Working Procedures/Proposed Approaches/Implementation

6.1. Working Procedure

6.2. Proposed Approach for ML

6.3. Proposed Approach for Distributed System

6.4. Implementation

VII. Result Evaluation/Discussion

7.1. Expected Result

7.2. Challenges

IX. Scope for future work

X. Conclusion

References

**ABSTRACT**

This research deals with distribution of data, across a distributed system architecture, and creation of a machine learning model for benchmarking Bengali Handwritten Numeric Text. [12] [14]

It is often seen that a machine learning model or cluster working on a single machine takes up considerable amount of resource and time when the amount of data is significantly high.[2] A distributed system overcomes these shortcomings and enhances speed of processing, using horizontal scaling instead of vertical scaling. This distribution is achieved using Hadoop Cluster here. [1]

**I. INTRODUCTION**

**Machine learning** (**ML**) has become a primary mechanism for distilling structured, semi-structured or unstructured information and knowledge from raw data, turning them into automatic predictions and actionable hypotheses for diverse applications, such as: analyzing social networks; reasoning about customer behaviour; interpreting texts, images, and videos; identifying disease and treatment paths; driving vehicles without the need for a human; and tracking anomalous activity for cybersecurity, among others. [14]

**Artificial Neural Networks** (**ANN**) or **connectionist systems** are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. [17] [20]

**Convolutional Neural Network** (CNN, or ConvNet) in deep learning [8], is a subset of Artificial Neural Network, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. They have applications in image and video recognition, recommender systems, image classification, medical image analysis, and natural language processing.[21]

**Unstructured data** (or **unstructured information**) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

**II. MOTIVATION**

Machine Learning has become an integral part of any futuristic project work and thus a subject of interest for undergraduates. The increasing number of projects has given rise to another problem – the allocation of resources and processing power of a system. However, to conduct a research of considerable significance, many people don’t have access to a powerful system. But it is relatively easier to get access to a number of less powerful machines. This is the motivation of our work. Hadoop Cluster provides a pragmatic solution to the distributed system architecture problems. [7] [9]

**III. SURVEY**

More than 80% of all potentially useful business information is unstructured data, in kind of sensor readings, console logs and so on. The large number and complexity of unstructured data opens up many new possibilities for the analyst. Text mining and natural language processing are two techniques with their methods for knowledge discovery from textual context in documents. An approach to organize complex unstructured data and to retrieve necessary information in an efficient way has already been tested. Unstructured data is heterogeneous and variable in nature and comes in many formats, including text, document, image, video and more. Unstructured data is growing faster than structured data. As a new, relatively untapped source of insight, unstructured data analytics can reveal important interrelationships that were previously difficult or impossible to determine. New technologies are emerging to make unstructured data analytics possible and cost-efficient. The new approach redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources. It utilizes easily scalable “shared nothing” architecture, distributed processing frameworks, and non-relational and parallel relational databases.

Traditionally, Machine Learning models has been trained and tested through data stored in structured format in databases and excel files (.csv). But due to the huge amount of data generation and requirement of real-time analysis, the data source from a single machine is creating a bottleneck of data input to the model. [3] [10]

This project aims to address this situation, with a concept of data distribution across machines, so, that the data feed to the model is improved and bottleneck can be reduced. [4] [5] [6]

**IV. BROAD OBSERVATIONS**

It is seen that as data quantity is increasing, machine learning model computation is becoming a heavy task for the system. Also, the possibility of data corruption, system failure, link break, among others are increasing the counter productivity of the system.

Huge data feed to the model from data storage, like, relational database, excel file (.csv), is also a traditional way, which acts as a bottleneck in a system.

**V. OBJECTIVES**

* To create an optical character recognition (OCR) for Bengali handwritten numeric text.
* Security Feature: fault tolerant feature where data is processed in one node and data is replicated in another note in the cluster. So, when a failure occurs in the one node or data is corrupted, copy of data is available in another node. [11][23][24]
* Autonomy: Because of sharing data by means of data distribution each site is able to retain a degree of control over data that are stored locally. [11][23][24]
* In distributed system there is a global database administrator responsible for the entire system. A part of global data base administrator responsibilities is delegated to local data base administrator for each site. Depending upon the design of distributed database each local database administrator may have different degree of local autonomy. [11][23][24]
* Availability: If one site fails in a distributed system, the remaining sites may be able to continue operating. Thus, a failure of a site doesn't necessarily imply the shutdown of the System. [11][23][24]
* Real-time analysis: Distributed structure can process small data sets to petabytes of data in real-time. [11][23][24]

**VI. WORKING PROCEDURES/PROPOSED APPROACHES/IMPLEMENTATION**

**6.1 WORKING PROCEDURE**

This project is based on the concept of applying Machine Learning to data stored across distributed systems.

For Machine Learning analysis, a model for unstructured data (images) [13] classification is made with Convolution Neural Network (CNN) approach, which uses Google’s TensorFlow as backend and Keras (neural network library). [20][25]

For the purpose of storing huge amounts of unstructured data (images) in a distributed computing environment, a special type of computational cluster is designed, that, we call as Hadoop Clusters. A Hadoop cluster is a collection of independent components connected through a dedicated network to work as a single centralized data processing resource. The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. [15]

**6.2 PROPOSED APPROACH FOR ML**

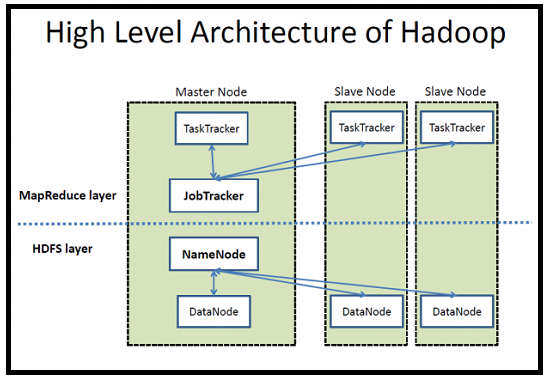
One of the key topics of Natural Language Processing (NLP) is optical character recognition (OCR). NLP of Bengali numeric text is done using Google’s TensorFlow framework as a backend along with Keras (neural network library). NumtaDB [12], a dataset consisting of more than 85,000 images of hand-written Bengali digits, is used for training machine learning model for optical handwriting recognition. [14]

**6.3 PROPOSED APPROACH FOR DISTRIBUTED SYSTEM**

Unstructured images data are stored in Hadoop Distributed File System across multiple nodes. This data is accessed by the model for the purpose of prediction. Since, Hadoop provides for real time analysis of data (up to Peta-Byte), the computational power for input of data to the model is reduced significantly. Also, the notion of data corruption is reduced, as data stored is replicated. Failure of the system is unlikely, as in the event of a node failure another node provide backup.[15] [22] [23]

**Machine Learning Algorithm**

**→**



**Unstructured**

**Dataset**

**(images)**

**Result**

**→**

**→**

Figure: Overview of the system [15]

**6.4 IMPLEMENTATION**

**6.4.1 DATASET DESCRIPTION**

NumtaDB, is a Bangla handwritten digit imagesdataset classification. Dataset contains 85,000+ of Bengali digits which can be used for creating OCR in Bengali language. The dataset is a combination of six datasets that were gathered from different sources and at different times. The sources are labeled from 'a' to 'f'. The training and testing sets have separate subsets depending on the source of the data. All the datasets have been partitioned into training and testing sets so that handwriting from the same subject/contributor is not present in both. [12] [14] [19]

**6.4.1 TOOLS TECHNOLOGY REQUIRED**

* **Anaconda** — Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications, that aims to simplify package management and deployment.
* **Jupyter Notebook** — The Jupyter Notebook is an incredibly powerful tool for interactively developing and presenting data science projects. It comes installed with anaconda.
* **TensorFlow** — TensorFlow is an open-source software library for dataflow programming across a range of tasks. [20]
* **Keras** — Keras is an open source neural network library written in Python. [21]
* **CNN** — Convolution Neural network, a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery.
* **Hadoop —** Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage. [15]
* **Other Python Library —** Python libraries like matplotlib, pandas, numpy to name a few.

**VII. RESULT EVALUATION/DISCUSSION**

**7.1 Expected Result**

* Optical character recognition (OCR) system that is capable of recognizing Bengali numeric text. [18]
* Reduction of computation power with distributed data processing
* Increase in flexibility, scalability, and performance

**7.2 Challenges**

* Creation of multi-node cluster architecture of Distributed System using Hadoop framework. [15]
* Limitation of specification on a standard personal use machine. [20]

**VIII. SCOPE FOR FUTURE WORK**

Distributed computing provides a lot of advantages for the huge amount of data that is generated. Data stored in a single machine can have a lot of disadvantages, like, corruption, failure, etc.

Real-time processing of data is also required, without major hardware upgrade. This gives the way for improvement in automation of IoT enabled devices. [11]

Using the structure of distributed system established in the project, analysis of different types of unstructured data using different machine learning algorithms can be achieved.

**IX. CONCLUSION**

Apart from the obvious advantages of increased amount of processing power, we also see increased resource allocation efficiency which benefits any kind of machine learning model and allows programmers to decrease processing time while maximizing the benefits. Hadoop is a very good ecosystem that allows us to implement the distributed system efficiently [15]. The results prove that the idea is feasible and pragmatic and delivers on its promise.

**REFERENCES**

1. T.K.Das and P.Mohan Kumar: BIG Data Analytics: A Framework for Unstructured Data Analysis

2. Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis: https://dl.acm.org/citation.cfm?id=3320060

3. Parallel and Distributed Deep Learning: https://stanford.edu/~rezab/classes/cme323 /S16/projects\_reports/hedge\_usmani.pdf

4. IBM Distributed Deep Learning (DDL): https://dataplatform.cloud.ibm.com/docs /content/wsj/analyze-data/ml\_dlaas\_ibm\_ddl.html

5. Distributed Deep Learning Framework based on Shared Memory for Fast Deep Neural Network Training: https://ieeexplore.ieee.org/document/8539420

6. Distributed Deep Learning: <https://www.uis.no/getfile.php/13458394/Forskning/> Vedlegg/10%20IKT/deep\_learning\_intro.pdf

7. Large Scale Distributed Deep Networks: <https://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf>

8. Intro to Distributed Deep Learning Systems: <https://medium.com/@Petuum/intro-to-distributed-deep-learning-systems-a2e45c6b8e7>

9. Parallel and Distributed Deep Learning: A Survey: <https://towardsdatascience.com> /parallel-and-distributed-deep-learning-a-survey-97137ff94e4c

10. Parallel and Distributed Deep Learning: A Survey: https://towardsdatascience.com/ parallel-and-distributed-deep-learning-a-survey-97137ff94e4c

11. Educba : https://www.educba.com/cloud-computing-vs-hadoop/

12. NumtaDB Assembled Bengali Handwritten Digits: https://www.researchgate.net/publicat ion/325643173NumtaDB\_-\_Assembled\_Bengali\_Handwritten\_Digits

13. Science Direct: https://www.sciencedirect.com/science/article/pii/S2352340917301117

14. Kaggle: https://www.kaggle.com/BengaliAI/numta/data

15. Hadoop Architecture: https://www.dezyre.com/article/hadoop-architecture-explained-what-it-is-and-why-it-matters/317

16. SuperDataScience: https://www.superdatascience.com/pages/machine-learning

17. StackAbuse: https://stackabuse.com/image-recognition-in-python-with-tensorflow-and-keras/

18. Github: https://github.com/Geektrovert/DeepNumta/blob/master/DeepNumtaCNN.ipynb

19. ArXiv org: https://arxiv.org/abs/1806.02452

20. TensorFlow: https://www.tensorflow.org/tutorials/images/transfer\_learning\_with\_hub

21. Implementing Convolutional Neural Networks for Image Classification and Facial Recognition Using Tensorflow v1.0: https://rstudio-pubs-static.s3.amazonaws.com/279 500\_57cad8c546724b40ad3e90692f716ae4.html

22. Analyze Video Data Using Hadoop: https://www.tothenew.com/blog/how-to-manage-and -analyze-video-data-using-hadoop/

23. About Distributed Computing Environments: https://www.c-sharpcorner.com/article/ getting-detail-about-distributed-computing-environment/

24. Distributed Systems | Characteristics | Advantages | Disadvantages : http://www.imfrosty.com/2014/11/distributed-system.html

25. From raw images to real-time predictions with Deep Learning : https://towardsdatascience.com/from-raw-images-to-real-time-predictions-with-deep-learning-ddbbda1be0e4