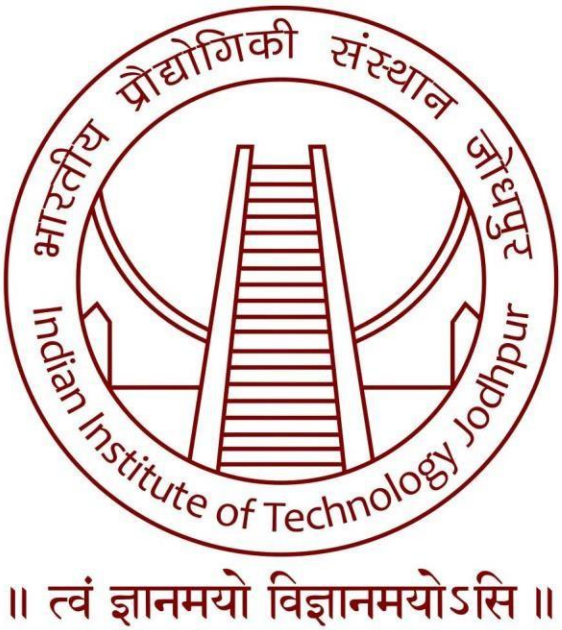
Basic Statistics with Hadoop Cloud Computing



Indian Institute of Technology, Jodhpur

Special Thanks to Prof. Pradip Sasmal for guidance and support

Vibha Sharma(G23AI1050)
Pooja P C(G23AI1057)
Assignment1|Assignment2|Assignment3|Project
Advance Data Engineering in cloud-components
Github link: https://github.com/G23AI1050/data-engineering-project.git

Abstract

This project demonstrates the use of Hadoop for performing basic statistical analysis within a cloud computing environment. By leveraging Hadoop's distributed computing power and cloud infrastructure, the project aims to efficiently process large datasets and compute fundamental statistical measures such as mean, median, mode, standard deviation, and variance. The report covers the setup of the cloud environment, Hadoop configuration, implementation of MapReduce jobs for statistical analysis, and evaluation of performance.

Introduction

- **Hadoop Overview:** Hadoop is an open-source framework designed for distributed storage and processing of large datasets. It consists of Hadoop Distributed File System (HDFS) for storage and MapReduce for data processing.
- **Cloud Computing:** Cloud computing provides scalable and on-demand computing resources over the internet, allowing users to run applications and store data in a virtual environment. Major cloud providers include Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure.
- **Basic Statistics:** Basic statistics involves summarizing and analyzing data to uncover patterns and trends. Key measures include mean (average), median (middle value), mode (most frequent value), standard deviation (measure of variability), and variance (measure of dispersion).
- **Goal:** To illustrate how Hadoop can be utilized to perform basic statistical analysis on large datasets in a cloud computing environment.
- **Expected Outcomes:** Efficiently calculate statistical measures using Hadoop’s MapReduce framework, demonstrating the benefits of distributed processing in a cloud setup.
- **Tasks:** Setting up a cloud environment, configuring Hadoop, implementing MapReduce jobs for statistical analysis, and evaluating performance.
- **Outcomes:** Provide a comprehensive analysis of the statistical measures and assess the performance and scalability of the Hadoop solution.

Literature Review

Hadoop Overview

- **Architecture:** Hadoop's architecture includes HDFS for storing data across distributed nodes and MapReduce for processing data. YARN (Yet Another Resource Negotiator) manages resources and job scheduling.
- **Components:**
- **HDFS:** Stores large files by breaking them into blocks and replicating them across multiple nodes.
- **MapReduce:** Processes data in parallel by dividing tasks into Map and Reduce phases.

Cloud Computing

- **Benefits:** Scalability, flexibility, cost-effectiveness, and access to powerful computing resources on demand.
- **Platforms:** Major cloud platforms like AWS, GCP, and Azure offer services for running Hadoop clusters and managing big data applications.

Methodology

Dataset

- **Source:** The dataset can be sourced from publicly available repositories such as Kaggle or UCI Machine Learning Repository.
- **Characteristics:** Include information on dataset size, structure, and relevant attributes.

Environment Setup

- **Cloud Provider:** Choose a cloud provider (e.g., AWS, GCP, Azure) and set up a virtual environment.
- **Hadoop Cluster:** Launch and configure a Hadoop cluster within the chosen cloud platform, specifying the number of nodes and resources required.

Data Processing with Hadoop

- **Data Ingestion:** Use Hadoop's HDFS to store and manage the dataset. Perform initial data cleaning and preprocessing.
- **MapReduce Jobs:** Develop MapReduce jobs to perform the following statistical computations:
- **Mean Calculation:** Implement a MapReduce job to compute the mean by summing all values and dividing by the count.
- **Median Calculation:** Implement a job to sort the values and find the median.
- **Mode Calculation:** Use a job to count occurrences of each value and identify the mode.
- **Standard Deviation and Variance:** Implement jobs to calculate variance and standard deviation based on the mean and individual value differences.

Implementation

Cloud Setup

- **Provider Selection:** Choose a cloud provider and create an account.
- **Cluster Configuration:** Set up a Hadoop cluster by selecting appropriate instance types, configuring networking, and ensuring security settings.

Hadoop Configuration

- **Cluster Setup:** Install and configure Hadoop on the cloud instances. Set up HDFS, MapReduce, and YARN.
- **Data Upload:** Load the dataset into HDFS using command-line tools or Hadoop APIs.

MapReduce Jobs

- **Code Implementation:** Write Java or Python code for MapReduce jobs to calculate statistical measures.
- **Execution:** Submit the MapReduce jobs to the Hadoop cluster using command-line tools or web interfaces.

Performance Optimization

- **Tuning Parameters:** Adjust Hadoop configuration parameters to optimize performance, such as the number of mappers and reducers, memory allocation, and data replication settings.
- **Monitoring:** Use cloud provider and Hadoop monitoring tools to track performance metrics and resource usage.

Results

Statistical Outputs

- **Mean:** Present the computed mean value from the MapReduce job.
- **Median:** Display the median value and discuss the method used for its calculation.
- **Mode:** Show the mode value and explain the frequency analysis process.
- **Standard Deviation and Variance:** Provide the calculated values and interpret their significance.

Performance Metrics

- **Processing Time:** Report the time taken for MapReduce jobs to complete.
- **Resource Utilization:** Discuss the resource usage, including CPU, memory, and storage.

Scalability

- **Analysis:** Evaluate how well the Hadoop solution scales with increasing dataset sizes and the impact on performance.

Analysis of Results

- **Interpretation:** Discuss the significance of the statistical measures obtained and their relevance to the dataset.
- **Comparison:** Compare the Hadoop-based approach with traditional statistical analysis methods in terms of efficiency and scalability.
- **Method Comparison:** Compare the performance and accuracy of Hadoop-based analysis with other methods, such as standalone statistical software or databases.

Challenges

- **Data Processing Issues:** Address any difficulties encountered during data processing, such as data inconsistencies or format issues.
- **Cloud Setup Complexities:** Discuss challenges faced in setting up and configuring the cloud environment and Hadoop cluster.

Future Work

- **Improvements:** Suggest improvements for the Hadoop setup or analysis techniques.
- **Further Research:** Propose areas for further research, such as exploring advanced statistical methods or integrating other big data tools.