Phase 3 Documentation: Development Part 1

Project: Big Data Analysis

Phase 3 Submission by: Arun Kumar
Year: 3rd Year, Computer Science and Engineering (CSE)

Table of Contents

1. [Introduction]
2. [Database Setup]
3. [Data Ingestion]
4. [Data Transformation]
5. [Initial Analysis]
  - [Data Exploration]
  - [Machine Learning (Basic)]
6. [Conclusion]

1. Introduction

This document represents Phase 3 of the "Big Data Analysis" project, submitted by Arun Kumar, a 3rd-year student of Computer Science and Engineering (CSE). Phase 3, titled "Development Part 1," focuses on setting up the database, ingesting data, performing data transformation, and initiating initial data analysis.

2. Database Setup

Objective: Begin building the big data analysis solution by setting up the IBM Cloud Databases.

What to Do:
- The IBM Cloud Database is preconfigured for the project, with connection details and necessary permissions.
- Ensure that the database is provisioned and configured to accommodate the selected datasets, such as "rainfall in India 1901-2015."

3. Data Ingestion

Objective: Describe how you'll ingest the selected datasets into the databases.

What to Do:
- Create a data ingestion process to transfer the "rainfall in India 1901-2015" dataset into the IBM Cloud Database.
- Implement the data loading procedure, which can include batch uploads, streaming, or other methods based on dataset size and structure.

4. Data Transformation

Objective: Explain any data preprocessing or transformation steps required to prepare the data for analysis.

What to Do:
- Perform data preprocessing tasks to ensure the data is in a suitable format for analysis. These tasks may include handling missing values, scaling, normalizing, and converting data types.
- Document all data transformation steps for reproducibility and clarity.

5. Initial Analysis

Objective: Provide details about the initial data analysis steps, including running basic queries and scripts.

Data Exploration

What to Do:
- Utilize Python libraries for data analysis, such as Pandas, Matplotlib, and Seaborn, to explore the dataset.
- Perform preliminary analysis, including summary statistics, histograms, scatter plots, and correlation analysis.

Machine Learning (Basic)

What to Do:
- Optionally, apply basic machine learning techniques for analysis. In this example, Linear Regression is used for basic regression analysis, and you may explore other algorithms like classification, clustering, or advanced regression.
- Evaluate the performance of the machine learning models using relevant metrics (e.g., Mean Squared Error for regression).

6. Conclusion

Phase 3, "Development Part 1," sets the foundation for the big data analysis project by configuring the database, ingesting data, and initiating initial data analysis. The completion of this phase facilitates further exploration and in-depth analysis in subsequent project phases.