Big Data Analytics

A Project Submitted

By

Batch 2

Under the Guidance of

Kunal

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA BANGALORE

Phase 1 Documentation

Project: Big Data Analytics – Rainfall Analysis

Phase 1 Submission by: Batch 2
Year: 4th Year, Information Science and Engineering (ISE)
Institute: BMS Institute of Technology and Management, Yelahanka, Bangalore

---

1. Introduction to Data Acquisition

In the first phase of our project, the primary focus was on acquiring and preparing rainfall data for further analysis and modeling. The dataset used is the publicly available "Rainfall in India: 1901–2015" dataset, which contains monthly rainfall statistics across various states and districts of India.

Objectives of Phase 1

- Load and inspect the rainfall dataset.
- Perform initial exploration and statistical summary.
- Establish a database connection for data storage.
- Prepare SQL schema for structured storage.

---

2. Dataset Overview

- Source: Government of India Open Data Platform
- Duration: 1901 to 2015
- Features:
  - Year and Month
  - State and District names
  - Rainfall measurements in millimeters (mm)

Python

```
# Code snippet to load the dataset

import pandas as pd
rainfall_data = pd.read_csv('rainfall_india_1901_2015.csv')
print(rainfall_data.head())
```

This helped us understand the structure of the data and identify missing values, outliers, or inconsistencies in columns such as 'Rainfall'.

3. Database Setup

To store and manage large datasets efficiently, we used IBM DB2 Cloud Database. A secure connection was established using the ibm_db Python package.

python

```
import ibm_db
db2_conn = ibm_db.connect("DATABASE=...;UID=...;PWD=...;", "", "")
```

This connection allows us to perform SQL-based operations on the dataset, enabling integration with large-scale analytics workflows.

---

4. Data Exploration and Descriptive Statistics

Basic statistical analysis was performed to summarize the dataset using mean and standard deviation.

python

```
mean_rainfall = rainfall_data['Rainfall'].mean()
std_dev_rainfall = rainfall_data['Rainfall'].std()
print(f"Mean Rainfall: {mean_rainfall}")
print(f"Standard Deviation of Rainfall: {std_dev_rainfall}")
```

Insights Gathered:

- The average rainfall varies significantly between regions and months.
- Standard deviation highlighted regions with erratic rainfall patterns.

---

5. SQL Table Creation for Rainfall Data

To enable structured storage in the DB2 database, a SQL schema was created.

Sql

```
CREATE TABLE RainfallData (
    Year INT,
    Month INT,
    State VARCHAR(255),
    District VARCHAR(255),
    Rainfall FLOAT
);
```

This schema allows indexing and filtering based on year, location, and rainfall for advanced querying in later phases.

python

```
stmt = ibm_db.exec_immediate(db2_conn, create_table_sql)
```

6.Summary of Phase 1

Task	Status
Dataset Acquired	✅ Completed
Data Loading and Inspection	✅ Completed
DB2 Database Setup	✅ Completed
Statistical Summary	✅ Completed
SQL Schema Creation	✅ Completed

7. Tools and Technologies Used

- Python: Data analysis, DB connection
- Pandas: Data manipulation and cleaning
- IBM DB2: Cloud-based data storage
- SQL: Structured data operations