NARRA LOKESH REDDY
2211cs010415
S-6(52)
Big Data Analytics

Project Report

Air Quality Analysis using PySpark

1.Overview:

- This project involved an exploratory data analysis (EDA) of air quality data using the Apache Spark framework via PySpark.
- The primary goal was to process, clean, and analyze a real-time air quality dataset to derive insights on pollutant distribution, geographical pollution levels, and categorize the Air Quality Index (AQI).
- The analysis utilized PySpark's DataFrame API for transformations and aggregations across a large dataset.
- **Key Findings**: The city with the highest average pollution was Chikkaballapur (88.5), and the most commonly monitored pollutant was Carbon Monoxide (CO).

2.Introduction:

- **Project Title**: Air Quality Analysis
- **Problem Statement**: The project aims to analyze the concentrations of various pollutants across different monitoring stations to understand current air quality status, identify highly polluted regions, and establish a foundational categorization model (AQI) for future prediction and policy recommendations.
- **Data Source**: Real-time Air Quality Index for various locations (sourced from Data.gov.in)
- **Tool/Technology**: PySpark on a Jupyter Notebook environment.

3.Data Understanding and Cleaning:

- **3.1 Data Loading and Schema**
- **File Used**: air_quality_data.csv
- **Initial Data Count**: Not explicitly calculated in the provided output, but operations are run on the DataFrame df.
- **Sample Columns/Data Types (Initial Schema)**:
    - country: string
    - state: string
    - city: string
    - pollutant_id: string
    - pollutant_min: string
    - pollutant_max: string
    - pollutant_avg: string
- **3.2 Data Cleaning (Handling Missing Values)**
- **Action 1**: String values 'NA' were replaced with the null value None.
- **Action 2**: Rows with missing values (None/NaN) in the crucial calculation column pollutant_avg were dropped.
- **Cleaned DataFrame Name**: df_clean

4.Exploratory Data Analysis:

- **4.1 Monitoring Network Summary**
- **Number of Unique Monitoring Stations (Post-Cleaning)**: 470
- **4.2 Pollutant Frequency Analysis**
- **Analysis Performed**: Count of records per unique pollutant_id.
- **Top 3 Most Common Pollutants Monitored**:
    - CO: 445
    - OZONE: 436
    - NO2: 431
- **4.3 Pollutant Value Distribution**
- **Analysis Performed**: Calculated Min, Max, and Average (pollutant_avg) values for each pollutant_id.
- **Overall Average Pollutant Levels (Descending Order)**:
    - PM10: 55.75
    - PM2.5: 42.61

    - CO: 28.93
    - NO2: 20.74
    - OZONE: 19.58
    - SO2: 13.25
    - NH3: 4.53
- **4.4 Geographical Analysis (State and City)**
- **Top 3 States by Average Pollutant Level**:
    - Himachal Pradesh: 44.71
    - Jharkhand: 40.0
    - Delhi: 33.92
- **Top 3 Cities by Average Pollution (Overall)**:
    - Chikkaballapur: 88.5
    - Gummidipoondi: 67.6
    - Mandikhera: 56.75

5.Air Quality Index(AQI) Classification:

- **5.1 AQI Model Definition (Rule-Based Classification)**
- A new column, AQI_Category, was created based on the pollutant_avg value using a when() clause[24].
- **5.2 Critical Pollutant Identification (Filter-based)**
- **Filter Condition**: Selected records where pollutant_avg $>$ 50[26].
- **Top 3 Critical Records (Example Output)**[27]:
    - Bihar, Samastipur, PM2.5: 83
    - Bihar, Rajgir, PM10: 87
    - Andhra_Pradesh, Chittoor, PM2.5: 85
- **Example AQI Classification Results (Top 5 Rows)**[28]:
    - Bihar, Gaya, SO2, 11: Good
    - Bihar, Gaya, PM2.5, 31: Moderate
    - Bihar, Gaya, CO, 40: Moderate
    - Bihar, Gaya, OZONE, 20: Good
    - Bihar, Hajipur, NH3, 5: Good

Conclusion:

The PySpark analysis successfully cleaned and aggregated a large-scale air quality dataset to reveal significant pollution hotspots like Chikkaballapur (Avg. 88.5) and highly monitored pollutants like CO. The preliminary AQI classification into 'Good', 'Moderate', 'Poor', and 'Severe' categories provides a clear assessment of air quality status across different regions.