PROJECT 2

WEATHER DATA ANALYTICS

MEMBERS:

SIDDAVATAM LAKSHMI PRANATHI

VARSHA VARDINI

SARAVANA KUMAR

SUMIT SANJAY JADHAV

INTRODUCTION

A **Weather Data Analytics System** in **Data Engineering** is a **data pipeline-based architecture** that collects, processes, stores, and makes weather data usable for **analytics, reporting, and machine learning**. It focuses on the **ETL/ELT flow**, **data architecture**, **scalability**, and **real-time/batch ingestion**.

PROBLEM STATEMENT:

The goal of this project is to design a modern data pipeline for processing historical weather data using cloud-native tools, and to build machine learning models for predicting weather parameters like temperature.

- **ETL Problem:**
- ○ Ingest data from public weather APIs (e.g., NOAA) and satellite feeds using Azure Data Factory.
- ○ Store and transform using Databricks streaming (Autoloader + Delta).
- **ML Problem:**
- ○ Train regression models to predict rainfall, temperature, or humidity. ○ Optimize models using Hyperopt + MLflow in Databricks.

ETL Workflow (Engineering Pipeline)

Step 1: Create Azure Resources

Step 2: Ingest Data from NOAA/Satellite APIs

Step 3: Stream Data Using Databricks Autoloader

ML Workflow (Data Science Pipeline)

Step 4: Load Weather Data from Delta Lake

Step 5: Train Regression Models

Step 6: Optimize with Hyperopt

Step 7: Deploy Model (Optional)

Architecture

Kaggle CSVs → Azure Data Factory → ADLS Gen2

↓

Databricks Streaming (AutoLoader)

↓

Bronze → Silver → Gold Delta Tables

↓

ML Training & Evaluation

↓

Azure DevOps Pipelines

ETL Pipeline (Bronze → Silver → Gold)

1. Bronze Layer: Raw Ingestion

- Used Azure Data Factory with 7 *Copy Data* activities (one per CSV)
- Stored raw data into ADLS container: bronzelayer/weather_raw
- Used Databricks Autoloader to read .csv files and store as Delta Tables in /bronze

2. Silver Layer: Transform

- Unpivoted city-wide columns into long format (location, value)
- Joined humidity and temperature on datetime, location
- Stored result to /silverlayer/

3. Gold Layer: ML Features

- Loaded and unpivoted pressure.csv, wind_speed.csv
- Joined with silver to form final feature set
- Stored to /goldlayer/

Machine Learning:

Model Objective

Predict **temperature** using features: humidity, pressure, wind_speed

Models Trained

- **Linear Regression**
- **Random Forest Regressor** (best performing)

Optimization

- Used **Hyperopt** to tune:
  - n_estimators: [50, 100, 200]
  - max_depth: [5, 10, 15]

Evaluation Metric

- **RMSE (Root Mean Squared Error)**

CONCLUSION

The Weather Data Analytics System is a cloud-based ETL pipeline designed to collect, process, and analyze weather data efficiently. It uses Azure Data Factory to ingest data from public APIs and satellite feeds into Azure Data Lake Storage. Databricks with Auto Loader processes the data in real time, transforming it into clean, structured formats stored in Delta Lake. The system supports scalable, fault-tolerant data processing and is ideal for generating insights through Power BI or training models in Azure ML. Overall, it provides a simple yet powerful solution for managing large-scale weather data.