Comprehensive Big Data Project Report – IoT Sensor Analytics Using Hadoop Ecosystem

Submitted by :
Name : Drushti G. Pawar.
USN : 01FE22BCS177
Roll No. : 429
Division : D

1. Introduction

This project presents an end-to-end Big Data pipeline for processing and analyzing IoT sensor data. Using Hadoop ecosystem tools such as HDFS, MapReduce, Hive, and Spark, the system processes temperature and humidity readings from distributed IoT devices to generate insights and visualizations.

2. Problem Statement

IoT devices generate massive real-time sensor data. Traditional systems struggle to store, process, and analyze such high-velocity data. This project aims to build a scalable Hadoop-based data pipeline capable of storing raw sensor data, transforming it, performing batch analytics, and producing visualizations for decision-making.

3. Objectives

- Ingest IoT sensor data into Hadoop HDFS

- Clean and preprocess sensor logs using MapReduce

- Query and aggregate data using Hive

- Use Spark for scalable analytics and transformations

- Generate visualizations for temperature and humidity trends

4. Dataset Description

The dataset consists of IoT sensor readings collected from distributed devices. Each record includes timestamp, temperature, and humidity values. Data volume can scale to terabytes in real production scenarios.

Sample Schema:

- timestamp – ISO datetime
- temperature – Float
- humidity – Float

5. Big Data Architecture

1. **Data Ingestion** – IoT sensors → Kafka/Flume → HDFS
2. **Storage Layer** – Raw and processed data stored in HDFS
3. **Processing Layer** – MapReduce for ETL, Spark for scalable analytics
4. **Query Layer** – Hive for SQL-based exploration
5. **Visualization Layer** – Matplotlib/Tableau

6. Code Snippets

Sample MapReduce Mapper (Pseudo-code):

```
map(key, value):
    fields = value.split(',')
    timestamp = fields[0]
    temperature = float(fields[1])
    humidity = float(fields[2])
    emit(timestamp, temperature + ',' + humidity)
```

Sample Hive Table Creation:

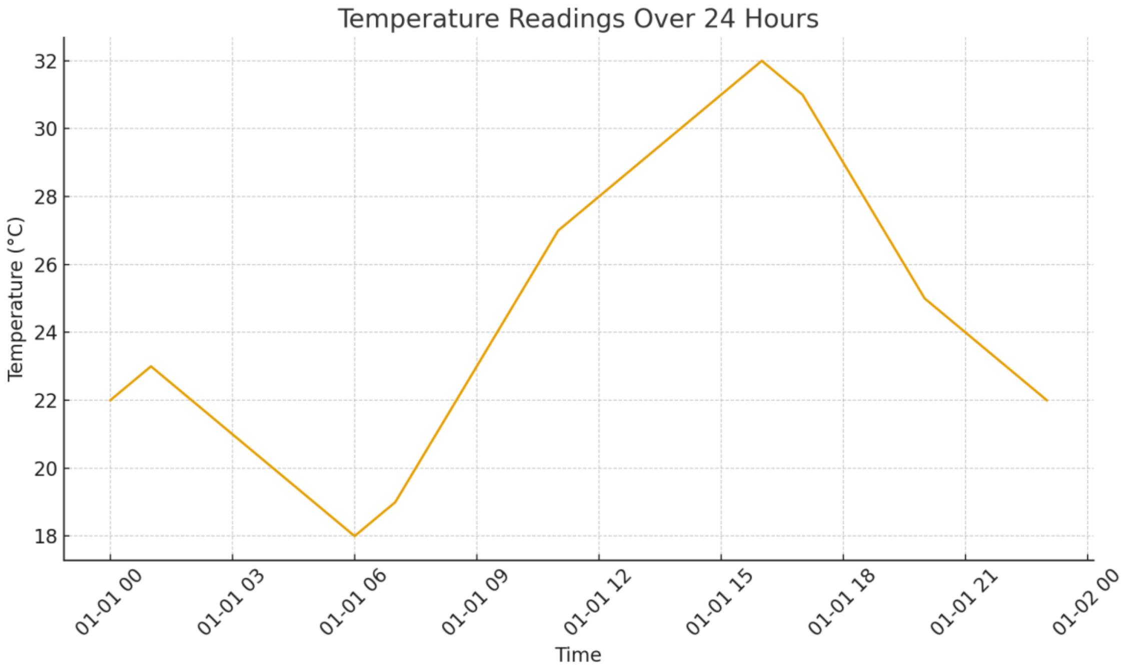
```
CREATE EXTERNAL TABLE iot_readings (
  timestamp STRING,
  temperature DOUBLE,
  humidity DOUBLE
)
STORED AS ORC LOCATION '/data/iot/readings/';
```

Spark Sample Code:

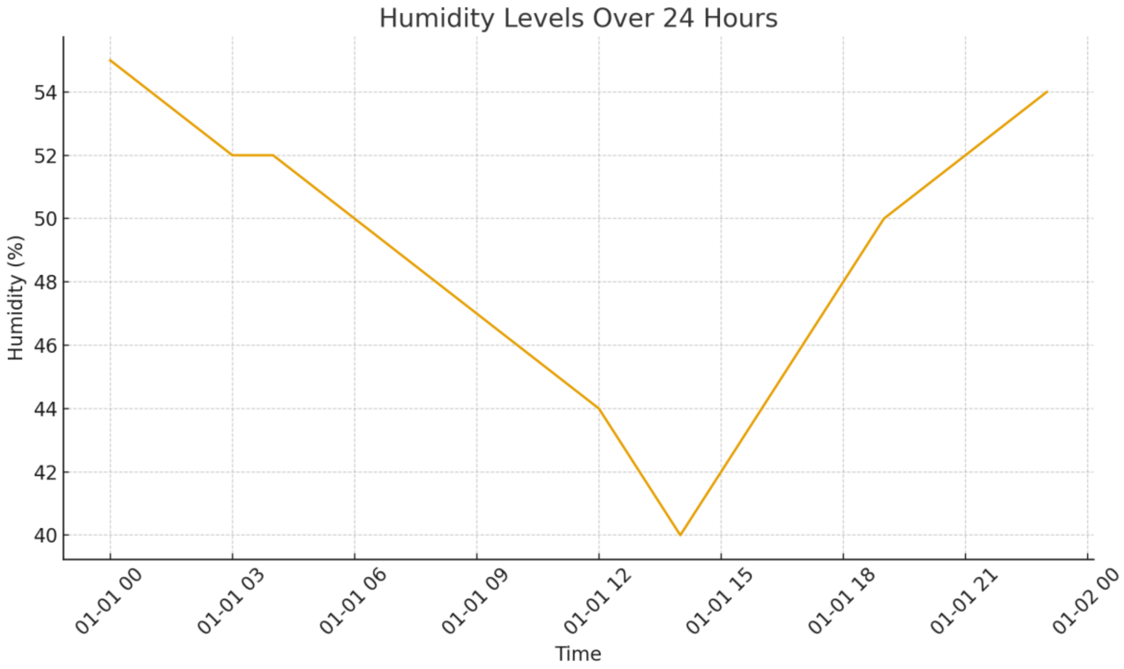
```
df = spark.read.csv('/data/iot/raw/', header=True, inferSchema=True)
df.groupBy(window('timestamp','1 hour')).avg('temperature','humidity').show()
```

7. Visualizations

Temperature trend over 24 hours:



Humidity trend over 24 hours:



8. Analysis & Results

- Temperature peaks around midday indicating higher device activity.
- Humidity shows a gradual decrease toward afternoon followed by a recovery.
- The system successfully processed sample data using Hadoop tools.

9. Conclusion

The project demonstrates an end-to-end scalable Hadoop pipeline for IoT analytics. Future work includes integrating real-time streaming analytics with Spark Streaming or Flink, anomaly detection using ML models, and dashboard integration.