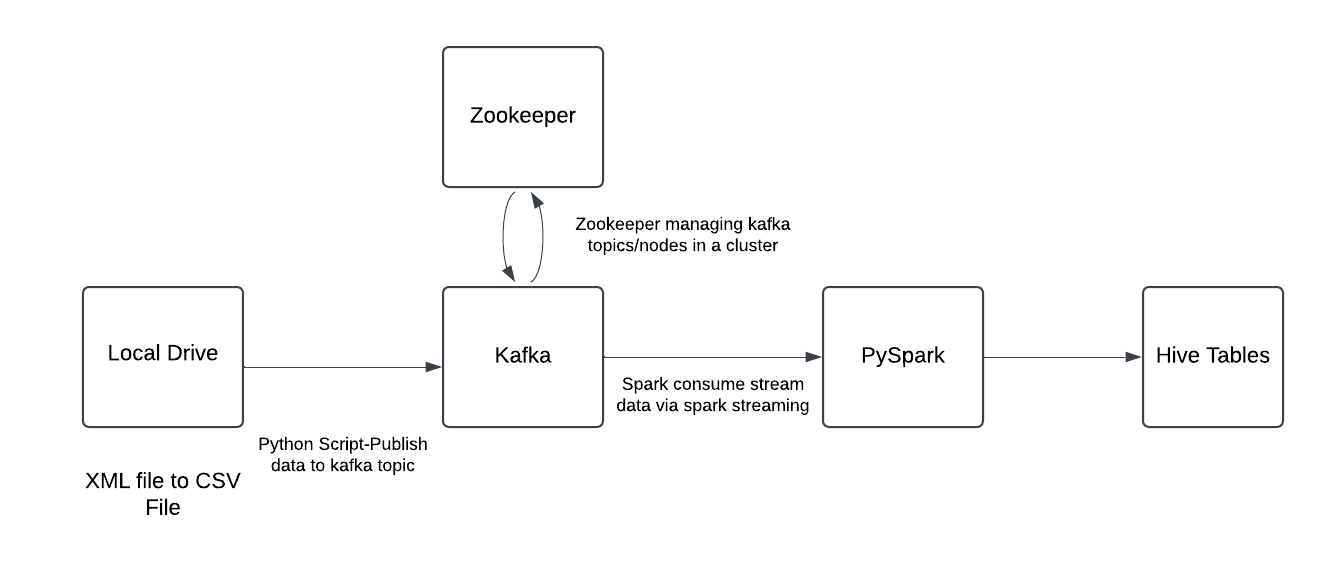
**Final capstone project:Air Quality Analysis and Prediction**

**Overview:**

This project involves analyzing air quality data from Indian cities to predict air quality trends. By creating a scalable data pipeline, we will process using data engineering tools like Python,Kafka,Pyspark,Hive and analyse to get some insights from those data.



**Task Execution Order:**

1. Data Extraction:

Downloaded XML file from India air quality index set and converted XML file to CSV file using Python Script.

1. Data Ingestion:

Using Kafka producers to send data to topics and stored raw data in MySQL.

1. Data Processing:

Using Pyspark for streaming data from kafka topic and made some spark jobs for data cleaning and transformation

1. Loaded in Hive Table:

* After loading into the hive table and performing some queries to get some data insights via charts.

**Steps Followed:**

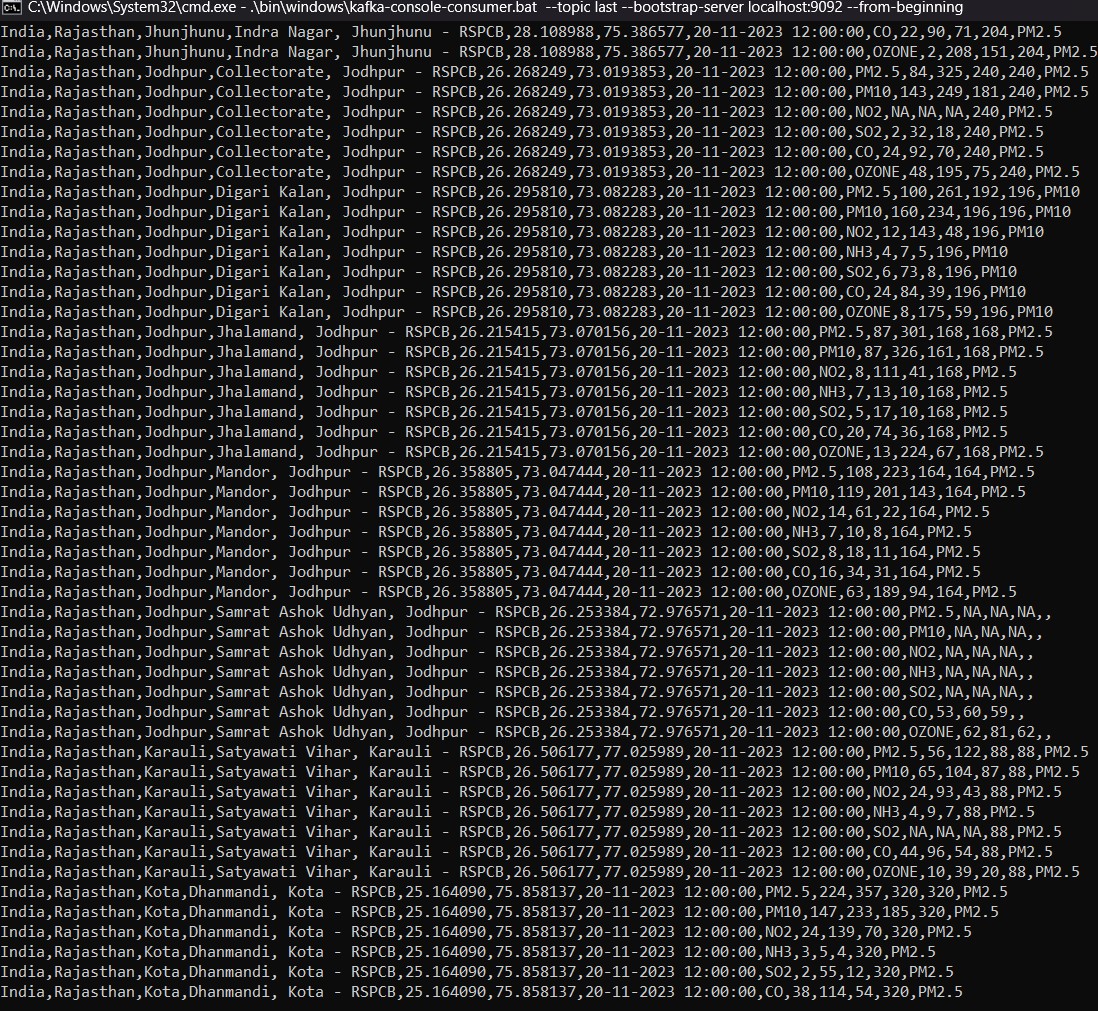
First Downloaded the XML file from India Air Quality set page and converted the XML file to CSV File.

I wanted to stream data from files,i used kafka for that

* Start Zookeeper
* Start Kafka Server
* Created Kafka topic

Kafka is all set and running and published the csv file to kafka topic using kafka producer.

The data is streaming via respective topics,



I want to capture this data and do some transformations and load it on tables(Hive).

used spark-Structured-Streaming (Spark-Consumer) to capture it , for this used Pyspark(python Library).



Records are available in Hive tables and performed some SQL queries to see some insights

Total 3 Python Jobs:

→ One for Kafka-Producer (JOB1)

→ One for spark-Streaming consumer (JOB2)

→ one for Sql queries and insights (JOB3)

