Kafka Setup Description

Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications. It plays a central role in this project by facilitating the continuous flow of air quality data from a simulated source (producer) to the data processing and prediction engine (consumer). This section outlines the end-to-end setup of Kafka for the real-time air quality monitoring pipeline.

1. **Prerequisites**
   - Download Java Development Kit (JDK)
     - Install JDK 8 or higher: Download from the Oracle website or use OpenJDK.
     - Set JAVA_HOME Environment Variable:
       - Right-click on This PC or My Computer and select Properties.
       - Click on Advanced system settings.
       - Click on Environment Variables.
       - Under System variables, click New:
         - Variable name: JAVA_HOME
         - Variable value: Path to your JDK installation (e.g., C:\ProgramFiles\Java\jdk-17.0.1)
       - Edit the Path variable:
         - Add %JAVA_HOME%\bin to the list.
   - Create kafka folder
     - Go to C:\
     - Create new folder and name it as "kafka"
     -
2. **Download Apache Kafka**
   - Visit the Apache Kafka Downloads page. (https://kafka.apache.org/downloads)
   - Download the latest binary release (e.g., kafka_2.13-4.0.0.tgz).
   - Extract the downloaded .tgz file by using 7-Zip or a similar tool.
   - Double-click into the extracted folder (e.g., folder name: kafka_2.13-4.0.0)
   - Double-click into the inside single folder (e.g., single folder name: kafka_2.13-4.0.0)
   - Move all the inside folder to the kafka directory created in step 1 (e.g., C:\kafka)

3. **Generate Cluster ID**
   - Open Command Prompt (CMD) in administrator mode by
     - go to window
     - in search box type cmd
     - once cmd program appear, right click and click "Run as administrator"
   - Change current directory to C:\kafka by the following command:
     - cd C:\kafka
   - Generate Cluster ID by the following command:
     - ./bin/kafka-storage.sh random-uuid
   - Format Log Directories by the following command:
     - ./bin/kafka-storage.sh format -t <CLUSTER_ID> -c fig/kraft/server.properties
     - replace CLUSTER_ID by the cluster id from the previous step

- e.g. ./bin/kafka-storage.sh format -t c9a5c1a2-d6f2-4c11-8f7f-db0015e8ab3 -c config/kraft/server.properties

4. **Configuration on broker.properties**
   - Go to C:\kafka\config
   - Open file server.properties with text editor (name of the file will only server)
   - Change the file according to the following:
     - node.id=0
     - process.roles=broker,controller
     - log.dirs=C:/tmp/kraft-combined-logs
     - listeners=PLAINTEXT://localhost:9092,CONTROLLER://localhost:9093
     - advertised.listeners=PLAINTEXT://localhost:9092
     - controller.listener.names=CONTROLLER
     - inter.broker.listener.name=PLAINTEXT
     - controller.quorum.voters=0@localhost:9093
     - auto.create.topics.enable=true
     - num.network.threads=3
     - num.io.threads=8

5. **Create tmp folder for kraft-combined-logs**
   - Go to C:\
   - Create folder name tmp

6. **Start Kafka in KRaft mode**
   - Open CMD in administrator mode
   - Open kafka in KRaft mode from the following command:
     - .\bin\windows\kafka-server-start.bat .\config\broker.properties
   - If you see "[KafkaServer id=0] started" on the CMD then it means the system fully running Kafka 4.0.0 in KRaft mode

7. **Create kafka Topic**
   - Open CMD in
   - Create kafka topic from the following command:
     - .\bin\windows\kafka-topics.bat --create --topic <TopicName> --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1
     - e.g. .\bin\windows\kafka-topics.bat --create --topic uci_air_quality_data --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1