Assignment: Real-Time Flight Data Analytics Pipeline

Objective: Develop a scalable pipeline to ingest, process, and analyze live flight position data using Apache Kafka, ClickHouse, and Apache Airflow.

Data Source: Utilize the Flightradar24 API endpoint: `https://fr24api.flightradar24.com/api/live/flight-positions/light`

Tasks:

1. **Data Ingestion with Apache Kafka:**
    - **API Integration:** Develop a Python script to fetch real-time flight data from the Flightradar24 API at regular intervals (e.g., every 10 seconds). Ensure adherence to the API's authentication and rate-limiting policies.
    - **Kafka Producer:** Configure a Kafka producer to publish the fetched flight data to a Kafka topic named `flight_positions`.
    - **Data Schema:** Define a consistent schema for the flight data, including attributes such as flight ID, latitude, longitude, altitude, speed, and timestamp.
2. **Real-Time Data Processing with ClickHouse:**
    - **Kafka Integration:** Set up ClickHouse to consume data directly from the `flight_positions` Kafka topic using the Kafka engine.
    - **Table Schema:** Design an appropriate table schema in ClickHouse to store the ingested flight data efficiently.
    - **Data Transformation:** Implement materialized views in ClickHouse to aggregate and transform the raw flight data for analytical queries.
3. **Workflow Orchestration with Apache Airflow:**
    - **DAG Creation:** Develop Airflow DAGs to automate tasks such as:
        - Fetching data from the Flightradar24 API.
        - Publishing data to the Kafka topic.
        - Triggering ClickHouse data ingestion and transformation processes.
    - **Monitoring:** Set up Airflow alerts to monitor the pipeline's health and performance, notifying stakeholders of any failures or significant delays.

4. **Data Analysis:**

    - **Analytical Queries:** Use ClickHouse to perform queries such as identifying the busiest airspaces, tracking specific flights in real-time, and analyzing flight patterns over time. Share the performance of such queries and try different indexing/merge trees to test whether performance could be improved or not.

Deliverables:

- **Codebase:** Python scripts for data fetching and Kafka production, ClickHouse table schemas and materialized views, and Airflow DAG definitions.

- **Documentation:** Comprehensive documentation detailing the setup process, configurations, and usage instructions.

- **Video Demonstration:** A video demonstrating the steps and entire working state of the assignment

Evaluation Criteria:

- **Functionality:** The pipeline should ingest, process, and analyze real-time flight data effectively.

- **Scalability:** Design considerations for handling increased data loads.

- **Reliability:** Robustness of the pipeline, including error handling and recovery mechanisms.

- **Documentation:** Clarity and completeness of the provided documentation.

Note: Python code with API attached in the same folder, API will expire by 10th April.