Production Setup

This document outlines the production setup for a data pipeline focusing on data ingestion, processing, and updating queries based on new data. The cloud-based services used here are from Amazon Web Services (AWS), but similar services are available on other cloud platforms.

Components and Process

1. **Data Ingestion**
    - Tool: AWS Lambda (for triggering) + Amazon S3 (for storage)
    - Process:
        - AWS Lambda: Triggers the data ingestion process when new data is available.
        - Amazon S3: Stores the incoming raw data files.
2. **Data Processing and Storage**
    - Tool: AWS Glue (for ETL) + Amazon S3 (for processed data storage)
    - Process:
        - AWS Glue: Runs ETL jobs to clean, transform, and prepare data for analysis.
        - Amazon S3: Stores the processed data in a structured format, ready for querying by DuckDB.
3. **Query Execution and Analysis**
    - Tool: DuckDB (hosted on an EC2 instance or Lambda function for periodic processing)
    - Process:
        - EC2 Instance / Lambda Fuction: Runs scheduled DuckDB queries to update analytics .
4. **Orchestration and Scheduling**
    - Tool: Apache Airflow (for orchestration) + AWS CloudWatch (for monitoring)
    - Process:
        - Airflow: Manages the data pipeline, scheduling tasks and handling dependencies.
        - AWS CloudWatch: Monitors the pipeline and triggers alerts for any issues.
5. **Data Access and Visualization**
    - Tool: Tableau for data visualization
    - Process:
        - Data Visualization: Connects to the processed data in S3 and presents it in dashboards and reports.