DATA PIPELINE WITH AIRFLOW & AWS GLUE

Problem Description
A music streaming service requires a real-time data pipeline to process and analyze user streaming behavior. Unlike batch processing, this pipeline must handle data arriving at unpredictable intervals, ensuring timely computation of key metrics. The pipeline should integrate data from multiple sources, process it efficiently, and store the results in DynamoDB for consumption by downstream applications.

The incoming streaming data is stored in Amazon S3 in batch files that arrive at irregular intervals. The processing logic should validate, transform, and compute metrics on the data before making it available for real-time business intelligence and application consumption.

Objective:

- Build a data pipeline that ingests streaming data from S3 at unpredictable intervals.
- Use Apache Airflow for orchestration and AWS Glue for transformation.
- Perform necessary validations and transformations using PySpark and Python Shell jobs.
- Compute key **daily** KPIs such as:
    - **Daily Genre-Level KPIs:**
        - Listen Count: Total number of times tracks in a genre have been played in a day.
        - Unique Listeners: Distinct users who streamed a track in a given genre per day.
        - Total Listening Time: The cumulative listening time for tracks in a genre per day.
        - Average Listening Time per User: The mean listening duration per user per day.
        - **Top 3 Songs per Genre per Day:** The most played songs in each genre daily.
        - **Top 5 Genres per Day:** The five most popular genres based on listen count per day.
- Store processed data in Amazon DynamoDB for fast lookups by downstream applications.

User Stories:

1. As a data engineer, I want to ingest streaming data from S3 and process it using an automated pipeline.

2. As a data engineer, I want to validate incoming datasets to ensure all required columns exist before processing.
3. As a data engineer, I want to transform raw streaming data into meaningful metrics efficiently using AWS Glue.
4. As a data engineer, I want to store processed data in DynamoDB for fast access by downstream applications.
5. As a business analyst, I want to query the processed data in DynamoDB to gain insights into user behavior and song performance.

Deliverables:

1. **ETL Pipeline Implementation:**
    - An Apache Airflow DAG that orchestrates data ingestion, validation, transformation, and storage.
    - AWS Glue PySpark and Python Shell jobs for data transformation and ingestion.
2. **Data Validation Module:**
    - Automated checks to verify the presence of required columns in the incoming data.
3. **Transformation & KPI Computation:**
    - Glue job to compute daily genre-based streaming metrics.
4. **DynamoDB Data Ingestion Module:**
    - Glue job to reshape and insert transformed metrics into DynamoDB tables.
5. **File Archival Process:**
    - Airflow DAG task to move processed files to an archive directory in S3.
6. **Logging & Error Handling:**
    - Detailed logging and error handling mechanisms for troubleshooting.
7. **Documentation:**
    - Step-by-step documentation on setting up and running the pipeline.
    - Sample queries for retrieving insights from DynamoDB.

Evaluation Criteria:

- Proper implementation of the Airflow DAG for orchestration using MWAA.
- Efficient usage of AWS Glue PySpark and Python Shell jobs for transformation.
- Robust validation and error handling mechanisms.

- Optimization of DynamoDB storage for fast lookups.
- Clear and structured documentation for usability and troubleshooting.