Technical Solution Document: ETL Pipeline for Music Streaming Data Analysis

1. Introduction

1.1 Purpose

This document outlines the technical design and implementation details of an **ETL pipeline** for processing and analyzing music streaming data. The pipeline orchestrates data ingestion, validation, transformation, and storage using **Apache Airflow, AWS Glue, S3, and DynamoDB**.

1.2 Scope

- Ingest streaming data from **Amazon S3** at unpredictable intervals.
- Validate incoming data to ensure compliance with schema requirements.
- Transform raw data into meaningful **daily KPIs** using AWS Glue.
- Store processed results in **Amazon DynamoDB** for downstream consumption.
- Implement error handling, logging, and file archival mechanisms.

1.3 Definitions, Acronyms, and Abbreviations

Acronym	Definition
ETL	Extract, Transform, Load
DAG	Directed Acyclic Graph
S3	Amazon Simple Storage Service
AWS Glue	Amazon's managed ETL service
DynamoDB	NoSQL database service from AWS
KPI	Key Performance Indicator

1.4 References

- Project Codebase (Airflow DAG, Glue Scripts)
- AWS Documentation for S3, Glue, DynamoDB
- Best practices for **data pipelines** and **ETL optimization**

1.5 Overview

This document details the pipeline's **architecture, data flow, validation mechanisms, transformations, storage strategy**, and **deployment considerations**.

---

2. Architectural Representation

2.1 Description of Architectural Style

The pipeline follows a **data lake architecture** with **event-driven** processing:

- **Orchestration**: Apache Airflow DAG
- **Storage**: Amazon S3 (raw, processed, and archived data)
- **Processing**: AWS Glue for transformations
- **Storage & Querying**: DynamoDB for quick lookups

2.2 Architectural Goals and Constraints

- **Scalability**: Handle large volumes of streaming data efficiently.
- **Fault Tolerance**: Automated failure handling and retries.
- **Data Integrity**: Ensure valid schema and data consistency.
- **Real-time Insights**: Compute KPIs for immediate analytics.

---

3. Use-Case View

3.1 Use-Case Model

Actors:

1. **Airflow DAG**: Triggers and orchestrates the pipeline.
2. **AWS Glue**: Performs data transformations and KPI calculations.
3. **DynamoDB**: Stores processed KPIs for fast access.
4. **S3 Storage**: Stores raw, processed, and archived files.

Use-Cases

1. **Streaming Data Ingestion** – Detect new files in S3 and trigger processing.
2. **Data Validation** – Check schema and content validity.

3. **Transformation & KPI Computation** – Convert raw logs to daily insights.
4. **Data Storage & Archival** – Store KPIs in DynamoDB and archive files in S3.
5. **Failure Handling** – Detect and manage errors automatically.

---

4. Logical View

4.1 Major Components

- **Airflow DAG (music_streaming_etl2)**
    - Sensors & Operators: S3KeySensor, PythonOperator, GlueJobOperator
- **AWS Glue Transformation (streaming_transformation.py)**
    - Validates and merges data
    - Computes **Daily Genre-Based KPIs**
- **DynamoDB Storage**
    - Stores processed KPIs for fast retrieval.

---

5. Process View

5.1 Concurrent Processes & Synchronization

- Airflow **triggers DAG** when new files arrive in **S3**.
- **S3KeySensor** detects files and initiates processing.
- **GlueJobOperator** runs the transformation job.
- The pipeline waits for **previous steps to complete** before moving forward.

---

6. Deployment View

6.1 Mapping of Software Components

Component	Deployment Target
Airflow DAG	Managed Workflow for Apache Airflow (MWAA)

Component	Deployment Target
Glue Jobs	AWS Glue (PySpark)
S3 Storage	Amazon S3 (raw, processed, archive)
KPI Storage	DynamoDB

---

7. Implementation View

7.1 Development Environment

- **Python 3.9**
- **Apache Airflow 2.x**
- **AWS SDK (Boto3)**
- **PySpark (for Glue jobs)**

7.2 Configuration Management

- **Environment Variables in Airflow (project-two-vars)**
- **Versioning of DAG & Glue Scripts**

---

8. Data View (Optional)

8.1 Database Schema (DynamoDB)

Attribute	Type	Description
genre::date	String (Partition Key)	Unique genre identifier and date
metric-type	String (Sort Key)	The type of metric in an item
top_5_genres	Array	Top five genres in an array
listen_count	Number	Total plays per genre per day
unique_listeners	Number	Unique users per genre per day
top_3_songs	Array	Names of top three songs

8.2 Data Integrity & Security

- **IAM Roles & Policies**: Secure access control.
- **Data Encryption**: S3 & DynamoDB encryption.

---

9. Size and Performance

9.1 Performance Benchmarks

- **S3KeySensor** checks every **30minutes**.
- Glue transformations complete within **2-5 minutes**.

9.2 Scalability Constraints

- Glue processing time depends on dataset size.
- DynamoDB throughput must be tuned for query performance.

---

10. Quality Attributes

- **Reliability**: Retries failed jobs using Airflow.
- **Maintainability**: Modular DAG and well-documented code.
- **Security**: Uses IAM roles for data access.
- **Scalability**: AWS Glue auto-scales based on demand.

---

11. Appendices

11.1 Glossary

Term	Definition
DAG	Directed Acyclic Graph in Airflow
ETL	Extract, Transform, Load

11.3 Revision History

Version	Date	Description	Author
1.0	2025-03-20	Initial Draft	Marzuk