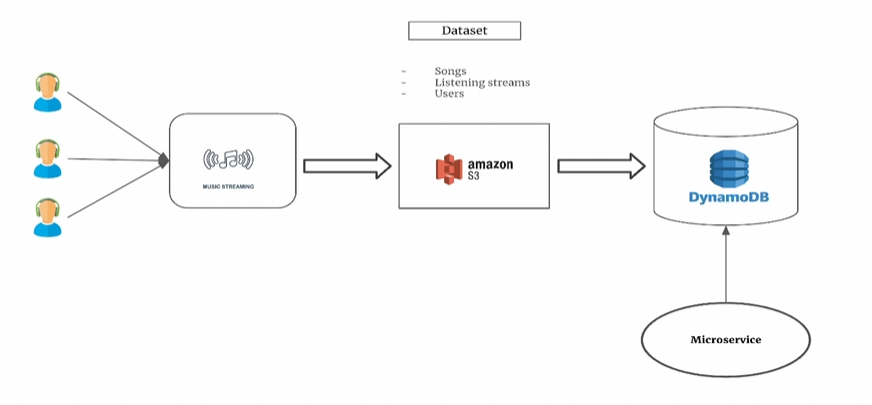
Distributed Music Streams Processing Using Airflow & DynamoDB

# Project Overview

This ETL pipeline processes music streaming data using AWS services. Airflow orchestrates workflow execution, Spark transforms raw data, and DynamoDB stores the processed results for efficient access.

# Architecture

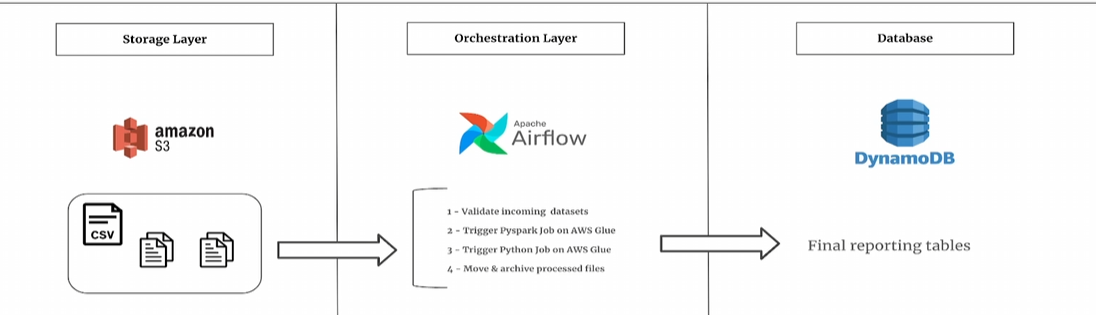
The solution follows this architecture:



# Data Pipeline Steps

1. \*\*Validate Data\*\*: Ensure input files exist in S3.  
2. \*\*Spark Transformations (AWS Glue)\*\*: Process streams & compute KPIs.  
3. \*\*Python Ingestion Job\*\*: Write transformed data to DynamoDB.  
4. \*\*Archive Processed Files\*\*: Move data post-processing.

# Workflow



# Tools & Technologies

- \*\*Amazon S3\*\*: Data storage.  
- \*\*Apache Airflow\*\*: Orchestration.  
- \*\*AWS Glue (Spark & Python)\*\*: Data transformation & ingestion.  
- \*\*Amazon DynamoDB\*\*: Fast-access NoSQL database.

# Conclusion

This workflow ensures efficient, scalable ETL processing of music streams. Future enhancements could introduce event-driven triggers and real-time analytics.