Spotify Top 50 India ETL Pipeline on AWS

Overview: This project demonstrates how to build a scalable, serverless ETL pipeline for extracting, processing, and analyzing Spotify playlist data using AWS services. The focus is on automating data extraction from the Spotify API, transforming it, and enabling analysis through Amazon Athena.

About Spotify API: Spotify's Web API allows access to metadata about artists, albums, and tracks. In this project, we specifically use the *Top 50 - India* playlist to extract song, artist, and album data for analysis.

Services Used:

- **Amazon S3:** Used for storing raw and transformed JSON/CSV files.
- **AWS Lambda:** Automates the ETL process. One function extracts data from the Spotify API and loads it into S3 (raw), and another function transforms it into structured CSV format (transformed).
- **AWS Glue Crawler:** Crawls the S3 data and updates the AWS Glue Data Catalog with schema metadata.
- **AWS CloudWatch:** Triggers the Lambda function on a scheduled basis (weekly) for automated extraction.
- **AWS Athena:** Performs SQL-based analysis on the structured data stored in S3.
- **AWS Glue Data Catalog:** Central metadata repository accessed by Athena for querying.

Tools & Libraries:

- `pandas`: Data manipulation
- `boto3`: AWS SDK for Python
- `spotipy`: Python client for Spotify Web API
- `numpy`: Numerical processing

Install packages via pip:

```
pip install pandas boto3 spotipy numpy
```

Execution Flow:

1. **Spotify API Connection:**
   - Connects to Spotify's Web API using `spotipy` to extract playlist data.
2. **CloudWatch Scheduler:**
   - Triggers Lambda every week to automate the extraction process.
3. **Lambda Function (Extraction):**
   - Extracts playlist data (songs, albums, artists) from Spotify API.
   - Stores raw JSON in the `raw/` folder in S3.
4. **Lambda Function (Transformation):**
   - Reads raw data from S3.

    - Cleans and structures the data (e.g., deduplication, formatting).
    - Saves cleaned CSVs into the `transformed/` folder in S3.
5. **Glue Crawler:**
    - Crawls both raw and transformed data in S3.
    - Updates Glue Data Catalog with schema and partition info.
6. **Athena Queries:**
    - Queries structured data using SQL.
    - Enables insights like top artists, song durations, release trends, etc.

Key Benefits:

- Serverless architecture (no infrastructure management)
- Fully automated pipeline
- Real-time access to updated playlist data
- Scalable and cost-effective
- Easy SQL-based analysis using Athena

Use Cases:

- Track popular songs in India over time
- Analyze artists' presence across playlists
- Generate playlist-based reports or dashboards

Future Enhancements:

- Add visualization layer using QuickSight or Power BI
- Integrate more playlists or genres
- Enable real-time streaming with AWS Kinesis