BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

(DEPARTMENT OF MANAGEMENT)

MASTERS OF BUSINESS ADMINISTRATION IN BUSINESS ANALYTICS (2024–2026)

MPBA G517 – BIG DATA ANALYTICS

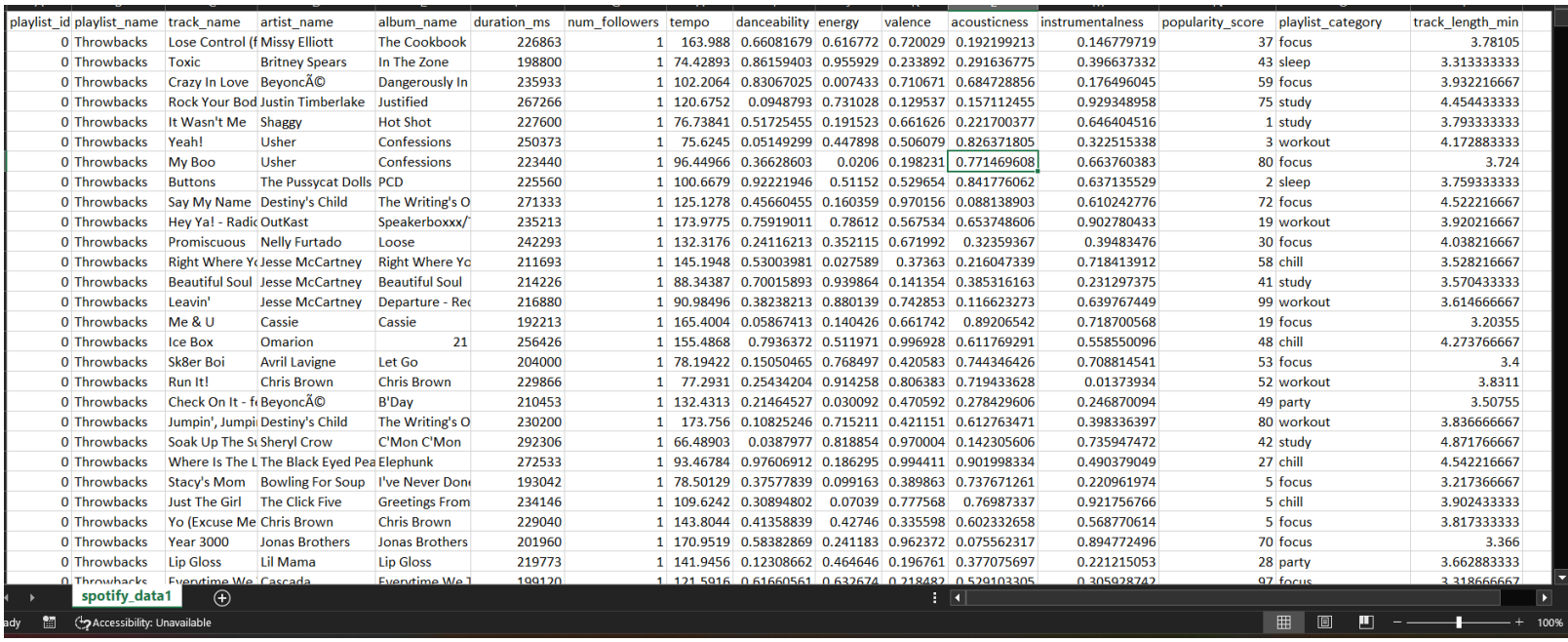
Submitted by Group – 13

2024H1540822P	PARAGEE SHARMA
2024H1540845P	MANKARAN SINGH BHATIA

GUIDED BY

PROF. ASHUTOSH VYAS

Decoding Global Music Trends: A Big Data Analysis of Spotify Audio Features



playlist_id	playlist_name	track_name	artist_name	album_name	duration_ms	num_followers	tempo	danceability	energy	valence	acousticness	instrumentalness	popularity_score	playlist_category	track_length_min
0	Throwbacks	Lose Control (f	Missy Elliott	The Cookbook	226863	1	163.988	0.66081679	0.616772	0.720029	0.192199213	0.146779719	37	focus	3.78105
0	Throwbacks	Toxic	Britney Spears	In The Zone	198800	1	74.42893	0.86159403	0.955929	0.233892	0.291636775	0.396637332	43	sleep	3.313333333
0	Throwbacks	Crazy In Love	BeyoncÃ©	Dangerously In	235933	1	102.2064	0.83067025	0.007433	0.710671	0.684728856	0.176496045	59	focus	3.932216667
0	Throwbacks	Rock Your Bod	Justin Timberlake	Justified	267266	1	120.6752	0.0948793	0.731028	0.129537	0.157112455	0.929348958	75	study	4.454433333
0	Throwbacks	It Wasn't Me	Shaggy	Hot Shot	227600	1	76.73841	0.51725455	0.191523	0.661626	0.221700377	0.646404516	1	study	3.793333333
0	Throwbacks	Yeah!	Usher	Confessions	250373	1	75.6245	0.05149299	0.447898	0.506079	0.826371805	0.322515338	3	workout	4.172883333
0	Throwbacks	My Boo	Usher	Confessions	223440	1	96.44966	0.36628603	0.0206	0.198231	0.771469608	0.663760383	80	focus	3.724
0	Throwbacks	Buttons	The Pussycat Dolls	PCD	225560	1	100.6679	0.92221946	0.51152	0.529654	0.841776062	0.637135529	2	sleep	3.759333333
0	Throwbacks	Say My Name	Destiny's Child	The Writing's O	271333	1	125.1278	0.45660455	0.160359	0.970156	0.088138903	0.610242776	72	focus	4.522216667
0	Throwbacks	Hey Ya! - Radic	OutKast	Speakerboxxx/	235213	1	173.9775	0.75919011	0.78612	0.567534	0.653748606	0.902780433	19	workout	3.920216667
0	Throwbacks	Promiscuous	Nelly Furtado	Loose	242293	1	132.3176	0.24116213	0.352115	0.671992	0.32359367	0.39483476	30	focus	4.038216667
0	Throwbacks	Right Where Yo	Jesse McCartney	Right Where Yo	211693	1	145.1948	0.53003981	0.027589	0.37363	0.216047339	0.718413912	58	chill	3.528216667
0	Throwbacks	Beautiful Soul	Jesse McCartney	Beautiful Soul	214226	1	88.34387	0.70015893	0.939864	0.141354	0.385316163	0.231297375	41	study	3.570433333
0	Throwbacks	Leavin'	Jesse McCartney	Departure - Rec	216880	1	90.98496	0.38238213	0.880139	0.742853	0.116623273	0.639767449	99	workout	3.614666667
0	Throwbacks	Me & U	Cassie	Cassie	192213	1	165.4004	0.05867413	0.140426	0.661742	0.89206542	0.718700568	19	focus	3.20355
0	Throwbacks	Ice Box	Omarion	21	256426	1	155.4868	0.7936372	0.511971	0.996928	0.611769291	0.558550096	48	chill	4.273766667
0	Throwbacks	Sk8er Boi	Avril Lavigne	Let Go	204000	1	78.19422	0.15050465	0.768497	0.420583	0.744346426	0.708814541	53	focus	3.4
0	Throwbacks	Run It!	Chris Brown	Chris Brown	229866	1	77.2931	0.25434204	0.914258	0.806383	0.719433628	0.01373934	52	workout	3.8311
0	Throwbacks	Check On It - fe	BeyoncÃ©	B'Day	210453	1	132.4313	0.21464527	0.030092	0.470592	0.278429606	0.246870094	49	party	3.50755
0	Throwbacks	Jumpin', Jumpi	Destiny's Child	The Writing's O	230200	1	173.756	0.10825246	0.715211	0.421151	0.612763471	0.398336397	80	workout	3.836666667
0	Throwbacks	Soak Up The Su	Sheryl Crow	C'Mon C'Mon	292306	1	66.48903	0.0387977	0.818854	0.970004	0.142305606	0.735947472	42	study	4.871766667
0	Throwbacks	Where Is The L	The Black Eyed Pea	Elephunk	272533	1	93.46784	0.97606912	0.186295	0.994411	0.901998334	0.490379049	27	chill	4.542216667
0	Throwbacks	Stacy's Mom	Bowling For Soup	I've Never Done	193042	1	78.50129	0.37577839	0.099163	0.389863	0.737671261	0.220961974	5	focus	3.217366667
0	Throwbacks	Just The Girl	The Click Five	Greetings From	234146	1	109.6242	0.30894802	0.07039	0.777568	0.76987337	0.921756766	5	chill	3.902433333
0	Throwbacks	Yo (Excuse Me	Chris Brown	Chris Brown	229040	1	143.8044	0.41358839	0.42746	0.335598	0.602332658	0.568770614	5	focus	3.817333333
0	Throwbacks	Year 3000	Jonas Brothers	Jonas Brothers	201960	1	170.9519	0.58382869	0.241183	0.962372	0.075562317	0.894772496	70	focus	3.366
0	Throwbacks	Lip Gloss	Lil Mama	Lip Gloss	219773	1	141.9456	0.12308662	0.464646	0.196761	0.377075697	0.221215053	28	party	3.662883333
0	Throwbacks	Everytime We	Cascada	Everytime We T	199120	1	121.5916	0.61660561	0.632674	0.218482	0.529103305	0.305928742	97	focus	3.318666667

The Spotify Million Playlist Dataset Challenge consists of a dataset and evaluation to enable research in music recommendations. The dataset contains 1,000,000 playlists, including playlist titles and track titles, created by users on the Spotify platform between January 2010 and October 2017.

We are using Spotify's million-playlist dataset to find out what drives global music trends — what makes songs popular, which moods dominate playlists, and how music features have changed over time.

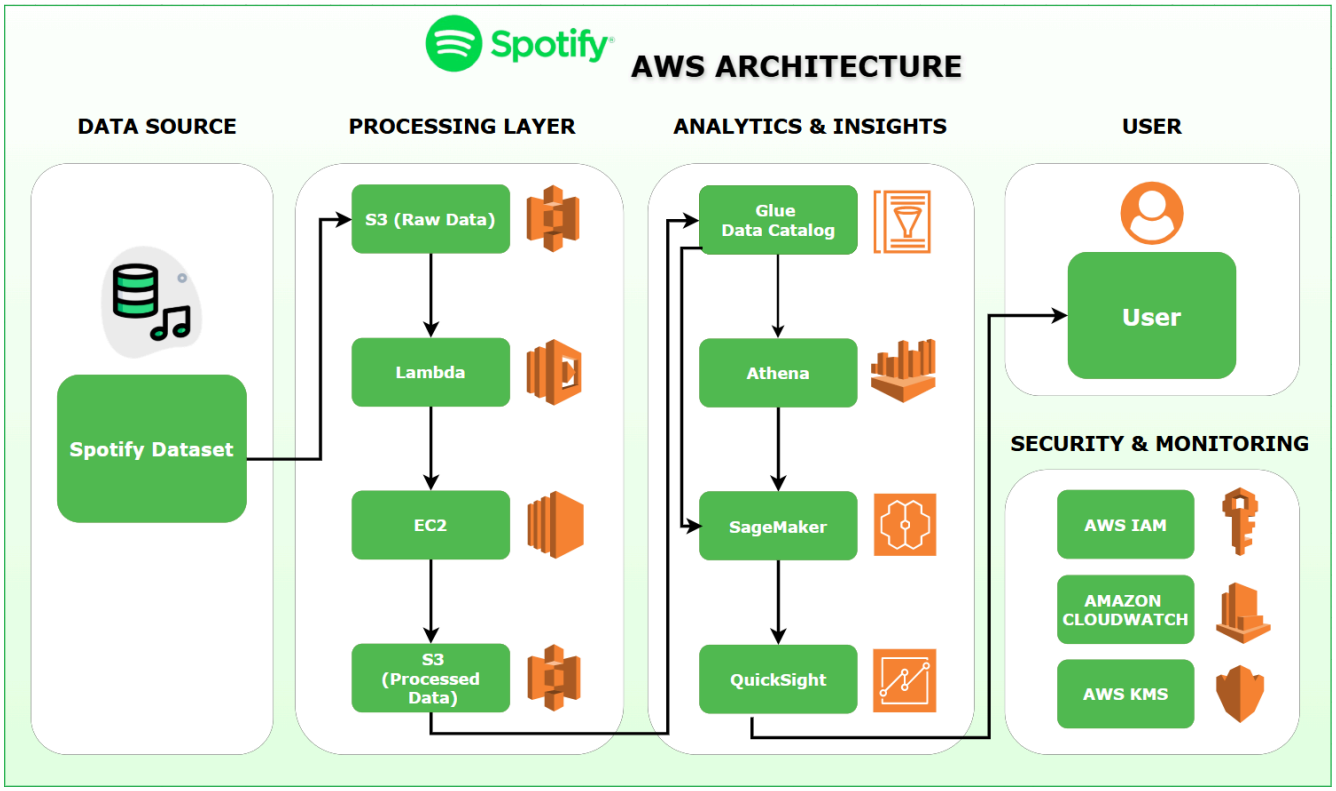
Since the data is very large (12.3 GB), we're using AWS to store, clean, analyze, and visualize it.

S3 stores it, Lambda and EC2 clean and process it, Glue and Athena help us query it, and SageMaker + QuickSight give us predictions and dashboards.

In the future, we can even train models to predict song popularity or recommend songs based on mood and energy.

Summary Table

Feature	Meaning	Range	Example (High Value)	Example (Low Value)
Tempo	Speed of the song	BPM (e.g. 60–180)	Fast dance track	Slow ballad
Danceability	How easy it is to dance to	0–1	Uptown Funk	Fix You
Energy	How loud/intense it feels	0–1	Blinding Lights	Let It Be
Valence	Mood positivity	0–1	Happy	Creep
Acousticness	How acoustic it is	0–1	Blackbird	Titanium
Instrumentalness	Whether it's instrumental	0–1	Movie score	Pop song



🎯 Project Title

Decoding Global Music Trends: A Big Data Analysis of Spotify Audio Features

🧠 Problem Statement (in short)

Spotify has millions of playlists and tracks — each with details like tempo, energy, danceability, mood (valence), popularity, and more.
Our goal is to **analyze global music trends** — to understand *what makes songs popular*, *how genres evolve*, and *how mood or energy levels differ across playlists and years.*

🏗️ Step-by-Step Process Flow (AWS Big Data Pipeline)

1. Data Source

- We start with a **12.3 GB Spotify dataset (CSV)** — that's *big data* because it's too large for normal processing on a laptop.
- It contains millions of playlists, each with song-level features (energy, danceability, valence, etc.).

2. Data Ingestion (S3 – Raw Layer)

- The raw dataset is uploaded to **Amazon S3 (Simple Storage Service)** — this is our **data lake**.
- S3 stores the raw CSV safely and at scale.

3. Data Processing (ETL Layer)

- **AWS Lambda** automatically triggers when new data arrives.
    - It performs light cleaning (removing nulls, converting formats).
- **Amazon EC2** (Elastic Compute Cloud) handles **heavy processing** — e.g., joining large tables, transforming data, and generating new metrics like *average danceability per genre*.
- Cleaned data is stored back into **S3 (Processed Data Bucket)**.

4. Data Catalog & Query (Glue + Athena)

- **AWS Glue Data Catalog** organizes and classifies the data (like a smart index).
- **Amazon Athena** allows running SQL queries directly on the S3 data — no need to move it anywhere.

  Example queries:
    - "Which playlists have the highest average energy?"
    - "Top 10 most popular artists in each region?"

- “Average danceability of songs from 2010–2020?”

5. Machine Learning & Insights (SageMaker + QuickSight)

- **Amazon SageMaker** will be used later to build ML models such as:
  - 🎵 **Popularity Prediction Model** – predicts a song’s popularity score using its features.
  - 🎶 **Recommendation Model** – suggests similar tracks based on mood and energy.
  - 🎤 **Trend Forecasting** – predicts emerging genres or sound patterns.
- **Amazon QuickSight** is used for creating interactive dashboards showing:
  - Music mood trends by year
  - Most energetic or chill playlists
  - Global vs. regional listening preferences
  - Correlation between *danceability* and *popularity*

6. User Access & Visualization

- The processed results and visual dashboards are shared with users (analysts, marketing teams, or Spotify itself) for decisions and insights.

7. Security & Monitoring

- **AWS IAM** – manages access rights and ensures only authorized users handle the data.
- **AWS CloudWatch** – monitors data pipeline performance.
- **AWS KMS** – encrypts data for security.

📊 What Insights Can Be Generated

1. **Music Preferences Across Categories**

- Compare “focus”, “workout”, “chill”, and “party” playlists.
- Example: “Workout playlists have the highest energy; chill playlists have the lowest valence (mood).”

2. **Trends Over Time**
   - How have song features (tempo, danceability) changed from 2010–2017?
3. **What Makes Songs Popular**
   - Identify patterns between popularity and song attributes like tempo, danceability, or valence.
4. **Artist & Album-Level Insights**
   - Which artists consistently produce high-energy tracks?
   - Which albums dominate particular playlists?
5. **User & Region-Based Preferences** *(if data includes region/user info)*
   - Find out if certain moods or tempos are more common in specific countries or cultures.

🤖 Future Machine Learning Possibilities

Goal	ML Technique	Example Output
Predict if a new song will be a hit	**Regression / XGBoost**	Predict popularity score
Recommend songs similar to a track	**Clustering / K-Means**	“You may also like…”
Classify playlists by mood	**Classification (SVM, Random Forest)**	Tag playlist as focus, chill, workout
Forecast music trends	**Time Series (ARIMA, LSTM)**	Predict mood/energy changes by year

---

💡 Why This Is a Good Dataset

✅ **Large & Diverse:** 1 million playlists from across years, genres, moods — gives true “big data” scale.
✅ **Rich Audio Features:** Attributes like tempo, energy, valence, danceability — perfect for analytics & ML.
✅ **Real-World Relevance:** Connects data analysis with how people *actually* consume music globally.
✅ **Scalable on AWS:** Easily handled using distributed services (S3, Athena, Glue, etc.).
✅ **End-to-End Potential:** Covers entire data life cycle — from raw data to ML-driven insights.
