Scalable Music Data Analysis Using the Million Song Dataset

Team Name: The Hardy Hadoopers

Teammates:
Alec Pippas (awp251)
Bhavya Matam (bm3792)
Naman Vashishta (nv2375)
Saja Alsulami (sfa7673)
Zhuyuan Wang (zw4759)

1. Introduction

Music recommendation, playlist curation, and genre classification are critical challenges in the modern music industry, where data volumes are massive and diverse. This project analyzes large-scale music data to uncover patterns and groupings within songs using unsupervised techniques and scalable data processing tools.

We will use the **Million Song Dataset (MSD)**, a large-scale, industry-grade dataset, to explore clustering techniques that allow us to identify similar songs and regression techniques that allow us to identify which musical metadata (e.g. genre, year) and audio features (e.g. tempo, loudness) are important to predicting song popularity.
This project aims to find patterns and groupings within songs using unsupervised techniques and scalable data processing tools.

2. Dataset Overview

The Million Song Dataset is a **280 GB collection of metadata and audio features** for one million popular contemporary music tracks from 1920s- 2010. The dataset is created by Bertin-Mahieux et al.(2011). The dataset was published in the 12th International Society for Music Information Retrieval Conference (ISMIR 2011). Notably, the dataset consists of text-based metadata and features rather than raw audio files.

2.1 Core Features

Each track includes more than **50 fields**, such as:

- **High-level audio features**:
  - danceability, energy, tempo, loudness, key, mode, duration, time_signature, year, etc.
- **Time-series musical structure**:
  - bars_start, beats_confidence, segments_pitches (935×12), segments_timbre, tatums_start, etc.
- **Metadata**:
  - artist_name, title, release, song_hotttnesss, track_id, artist_terms, etc.

These fields are extracted via **The Echo Nest Analyze API** and are well-suited for machine learning tasks.

3. Project Goals

We aim to:

- Explore **unsupervised clustering** of songs based on audio features
- Develop **interpretable visualizations** of music similarity
- Analyze **musical trends and structures** across time and feature space

This project emphasizes **scalable big data analysis**, using Spark and dimensionality reduction to handle high-dimensional, large-volume input efficiently.

4. Methodology

4.1 Data Handling

- Load and process the dataset using **PySpark**
- Select relevant acoustic features and clean incomplete records
- Normalize data for clustering
- Preprocess dataset for multi-class classification, using one-hot encoding for song genres

4.2 Dimensionality Reduction

- Aggregate time-series vectors (e.g., timbre, pitch) into fixed-length feature representations
- Apply **PCA** to remove noise and retain key variance
- Use **t-SNE** or **UMAP** for non-linear 2D/3D projection
- Enable better visualization and separation of clusters
- Enhance clustering performance by reducing redundant dimensions

4.3 Clustering

We will implement three unsupervised learning algorithms and compare their results:

- **K-Means** for baseline partitioning
- **DBSCAN** for density-aware clusters
- **Hierarchical Clustering** to explore nested groups

4.4 Visualization

- 2D/3D scatter plots of song clusters
- Annotate points with metadata (e.g., song name, year)
- Compare clusters against known tags or trends

5. Tools & Technologies

- **PySpark**, **Spark MLlib** – scalable processing
- **Scikit-learn**, **UMAP-learn** – modeling
- **Matplotlib**, **Seaborn**, **Plotly** – visualization
- **Google Colab**, **AWS**, or **HDFS** – compute infrastructure

6. Optional Tasks

Depending on timing and progress, we may also address the following topics:

- Identify audio features and metadata that are important for driving song popularity (number of plays)
- Tracking temporal changes in musical features (e.g., average tempo per year)
- Using available genre labels or tags to evaluate clusters of similar songs
- Predict song genre or mood from features
- Analyze user taste clusters using the Taste Profile subset

7. Complementary Datasets

We may use additional datasets to enrich our analysis with supervised labels, genre mapping, or user engagement/popularity metrics. The MSD integrates community-contributed datasets for extended analysis:

- **Cover Songs**: SecondHandSongs dataset.
- **Lyrics**: musiXmatch dataset.
- **Tags & Similarity**: Last.fm dataset.
- **User Behavior**: Taste Profile Subset (user play counts).
- **Genre Labels**: Tagtraum and Top MAGD.

8. Million Song Dataset Citation:

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
Million Song Dataset, official website by Thierry Bertin-Mahieux,

Available at: millionsongdataset.com