# Informatica Challenge

**Welcome to the Informatica challenge.**

You are provided with **Music** details dataset that contains information about tracks available on the music platform. Each record represents a single track and includes a variety of attributes that describe different aspects of the track, such as its metadata, audio features, and popularity.

As a BigData Engineer, your task involves cleaning the data, analyzing the data using Informatica PowerCenter.

**Data Preparation:**

- **Oracle SQL Setup:**
- Log in to Oracle SQL Developer in **Admin** connection using the credentials:
    1. **Username:** system
    2. **Password:** Admin

Create a table named **Music** with the below provided structure.

Load music.**csv** into the **Music** table.

You are given data set is in the ***"~/Desktop/Project/kickoffs-informatica-music/"***

In this task you are going to do Cleaning the data and two analytics tasks. So, create the tables in oracle before performing the operations.

**Music** : CREATE TABLE Music (slno int,track_id VARCHAR(50),artists VARCHAR(100),album_name VARCHAR(100),track_name VARCHAR(150),popularity int,duration_ms int,explicit VARCHAR(50),danceability float,energy int,track_key float,loudness int,track_mode float,speechiness float,acousticness varchar(50),instrumentalness varchar(50),liveness float,valence float,tempo float,time_signature int,track_genre VARCHAR(50));

**Music_Cleaned_Data** : CREATE TABLE Music_Cleaned_Data (track_id VARCHAR(50),artists VARCHAR(100),album_name VARCHAR(100),track_name VARCHAR(100),popularity int,duration_ms int,explicit VARCHAR(50),danceability float,energy int,track_key float,loudness int,track_mode float,speechiness float,valence float,tempo float,time_signature int,track_genre VARCHAR(50));

**popular_track_analysis :** CREATE TABLE popular_track_analysis(popularity_category varchar(250), total_tracks int, avg_dancebility float, avg_energy float);

**alt_rock_genre :** create table alt_rock_genre(track_id varchar(250), artists varchar(200), album_name varchar(200), track_name varchar(200), popularity float, track_genre varchar(250), duration_m float)

**ambient_genre :** create table ambient_genre(track_id varchar(250), artists varchar(200), album_name varchar(200), track_name varchar(200), popularity float, track_genre varchar(250), duration_m float)

**alternative_genre :** create table alternative_genre(track_id varchar(250), artists varchar(200), album_name varchar(200), track_name varchar(200), popularity float, track_genre varchar(250), duration_m float)

**Data Cleaning:**

- **Mapping Name:** Map_Cleaned_Data
- **Workflow Name:** Workflow_Cleaned_Data
- **Session Name:** Session_Cleaned_Data
- **Target Table:** Music_Cleaned_Data

**Operations:**

- Import table **Music** from oracle as source
- There are some unwanted columns in the given dataset, to clean the dataset drop columns **slno, acousticness, instrumentalness, liveness**
- After cleaning load, the data into **Music_Cleaned_Data** target table (For columns check sample output), create the target table in Oracle SQL

- After completing mapping, in the workflow manager.
- We are going to connect to repository "**Repo_etlclass**", so double click on the folder to connect.

**Sample Output:**

**track_id,artists ,album_name,track_name,popularity ,duration_ms ,explicit ,danceability ,energy ,track_key ,loudness ,track_mode ,speechiness ,valence ,tempo ,time_signature ,track_genre**

2K7xn816oNHJZ0aVqdQsha,The Neighbourhood,Hard To Imagine Neighbourhood Ever Changing,Softcore,86,206280,False,575,1,9,-6,0,0.03,0.37,93.986, 4,alt-rock

**NOTE: Music_Cleaned_Data table data is used as source for the below every tasks.**

**Analysis Tasks:**

## Task 1: Popular Track Analysis

- **Mapping Name:** Map_Popular_Track_Analysis
- **Workflow Name:** Workflow_Popular_Track_Analysis
- **Session Name:** Session_Popular_Track_Analysis
- **Target Table:** Popular_Track_Analysis

**Problem Statement:** Identify and analyze the most popular tracks in the dataset.

**Operations :**

- Filter out tracks with `**popularity**` greather than and equals to 70.
- Create a new column named `**POPULARITY_CATEGORY**`:
  - If `popularity` is between 70 and 80 (inclusive), categorize as 'High'.
  - If `popularity` is between 81 and 90(inclusive), categorize as 'Very High'.
  - If `popularity` is between 91 and 100 (inclusive), categorize as 'Top'.
- Group the tracks by `POPULARITY_CATEGORY` and calculate the total number of tracks in each popularity_category.
- Calculate the average `danceability`, `energy` for each `POPULARITY_CATEGORY` and store values in new columns **AVG_DANCEABILITY**, **AVG_ENERGY** respectively.
- Load the above data into the **TotalBalance_Group** target table (For columns check sample output), create the target table in Oracle SQL

**Sample Output:**

| POPULARITY_CATEGORY | TOTAL_TRACKS | AVG_DANCEABILITY | AVG_ENERGY |
|---|---|---|---|
| High | 42 | 0.54 | 1 |
| Top | 102 | 0.02 | 0.06 |

## Task 2: **Track Genre Analysis**

- **Mapping Name:** Map_Track_Genre_Analysis
- **Workflow Name:** Workflow_ Track_Genre_Analysis
- **Session Name:** Session_ Track_Genre_Analysis
- **Target Tables:** alt_rock_genre, alternative_genre, ambient_genre

**Operations:**

- Duration_MS has time in millisecounds so change the value to minutes and store in new column Duration_M(Formula : (millisecound/1000)/60)

- Round the Duration_M ith 2 decimal points.
- Using Router Transformation filter the Track_Genre and store output in respective tables
    1. TRACK_GENRE='alt-rock'
    2. TRACK_GENRE='ambient'
    3. TRACK_GENRE='alternativ

- Drop the unnecessary columns, kindly check the sample output.

- Load data into the **alt_rock_genre, alternative_genre, ambient_genre** target tables (For columns check sample output), create the target table in Oracle SQL

**Sample output:**

**alt-rock table :**

track_id , artists , album_name , track_name , popularity , track_genre , duration_m
2K7xn816oNHJZ0aVqdQsha,The Neighbourhood,Hard To Imagine The Neighbourhood Ever Changing,Softcore,86,alt-rock

**ambient table :**
track_id , artists , album_name , track_name , popularity , track_genre , duration_m
2K7xn816oNHJZ0aVqdQsha,The Neighbourhood,Hard To Imagine The Neighbourhood Ever Changing,Softcore,86,ambient

**alternative_genre table :**
track_id , artists , album_name , track_name , popularity , track_genre , duration_m
2K7xn816oNHJZ0aVqdQsha,The Neighbourhood,Hard To Imagine The Neighbourhood Ever Changing,Softcore,86, alternative_genre

Once you complete the challenge, make sure your output data is loaded into the respective target table in Oracle SQL

After completing the challenge, run **sample_test.ps1 file to** get the sample score

Click on the **submit** button to validate your solution