Challenge Overview

NOVEMBER 2023

Bob Foreman
Software Engineering Lead
LexisNexis Risk Solutions

The Challenge!

Music is Life!

The goal of this challenge is to analyze three music datasets across different genres and discover insights in the data, using the HPCC Systems platform.

You will be presented with several challenge questions in different categories. The more questions you answer, the higher your score will be at the end of the day.

The Challenge!

Music is Life!

We have provided three public music datasets for you to query and analyze:

1. **The Music Mozilla Dataset** (musicmoz.org)

MusicMoz is a comprehensive directory of all things music, edited by volunteers. We list, and accept submissions of, music-related reviews, articles, factual information, biographies, and websites. Also known as the Open Music Project.

2. **The Million Song Dataset** (millionsongdataset.com)

The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks.

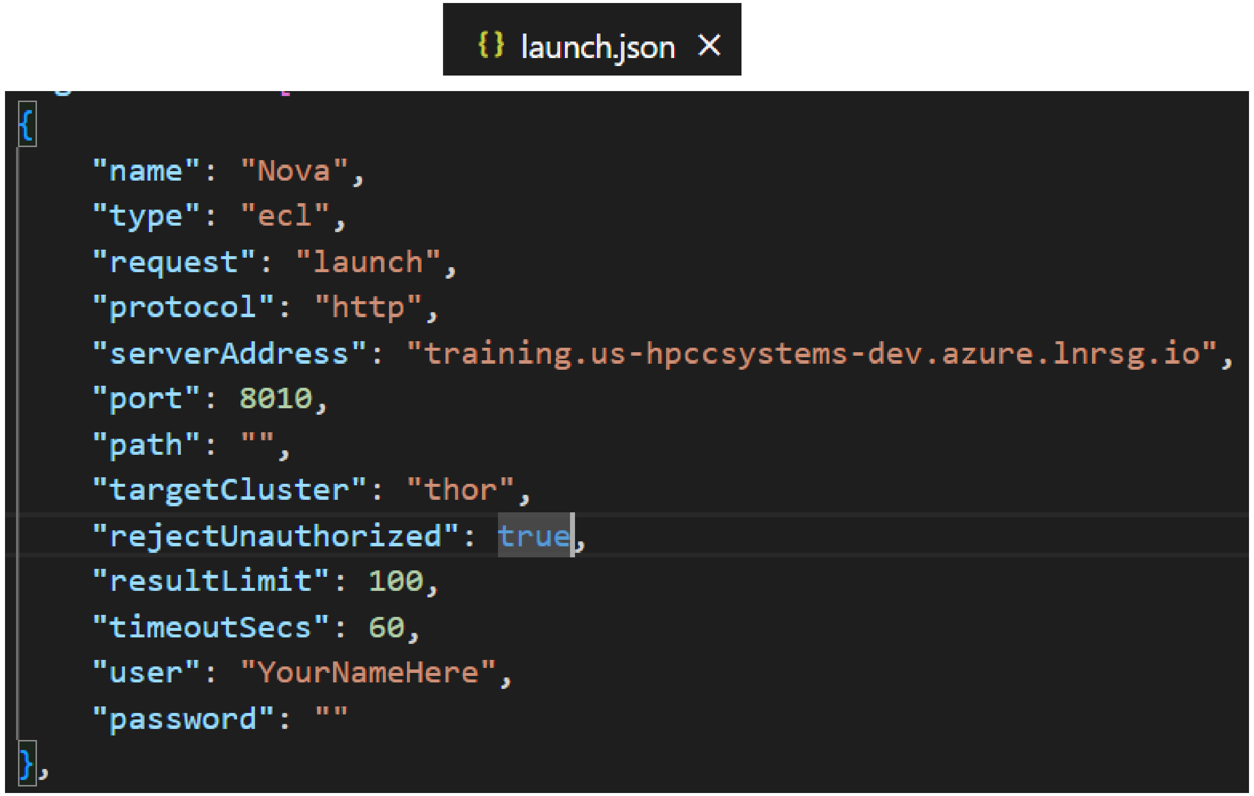
3. **The Spotify Top 2000** (www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset)

This data is extracted from the Spotify playlist - Top 2000s on PlaylistMachinery(@plamere) using Selenium with Python.

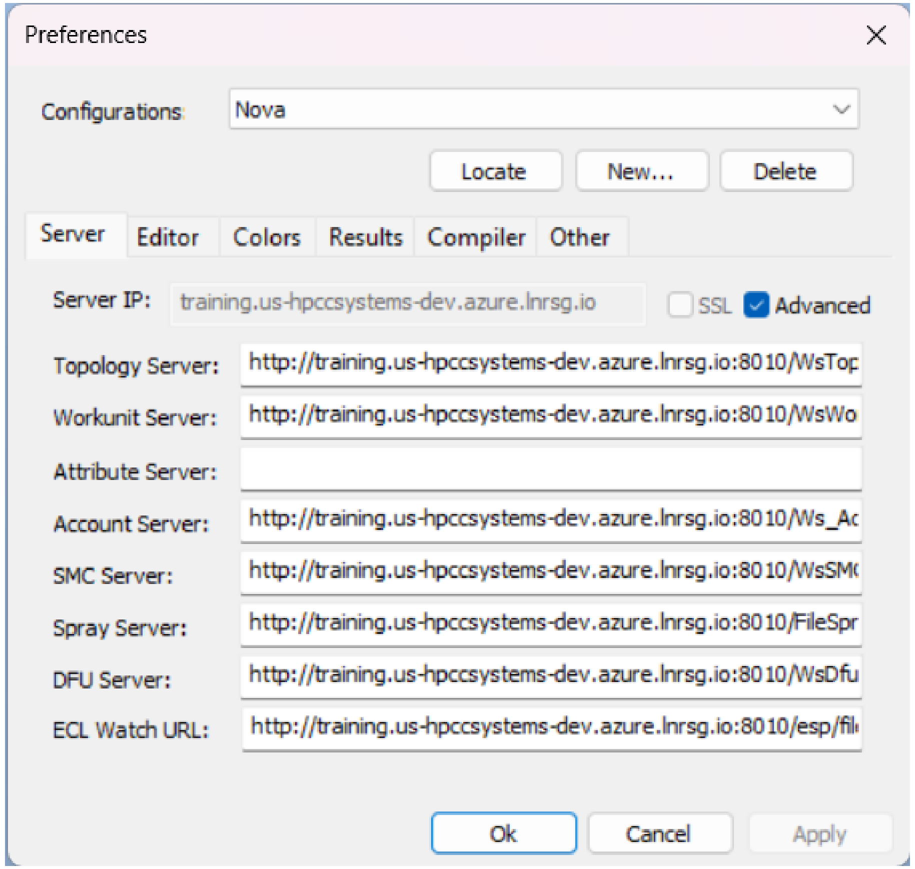
The Playing Field!

HPCC Cluster ECL Watch:

http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/



```
{
    "name": "Nova",
    "type": "ecl",
    "request": "launch",
    "protocol": "http",
    "serverAddress": "training.us-hpccsystems-dev.azure.lnrsg.io",
    "port": 8010,
    "path": "",
    "targetCluster": "thor",
    "rejectUnauthorized": true,
    "resultLimit": 100,
    "timeoutSecs": 60,
    "user": "YourNameHere",
    "password": ""
},
```



Configurations: Nova

Server IP: training.us-hpccsystems-dev.azure.lnrsg.io □ SSL ☑ Advanced

Topology Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/WsTop

Workunit Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/WsWo

Attribute Server:

Account Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/Ws_Ac

SMC Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/WsSM

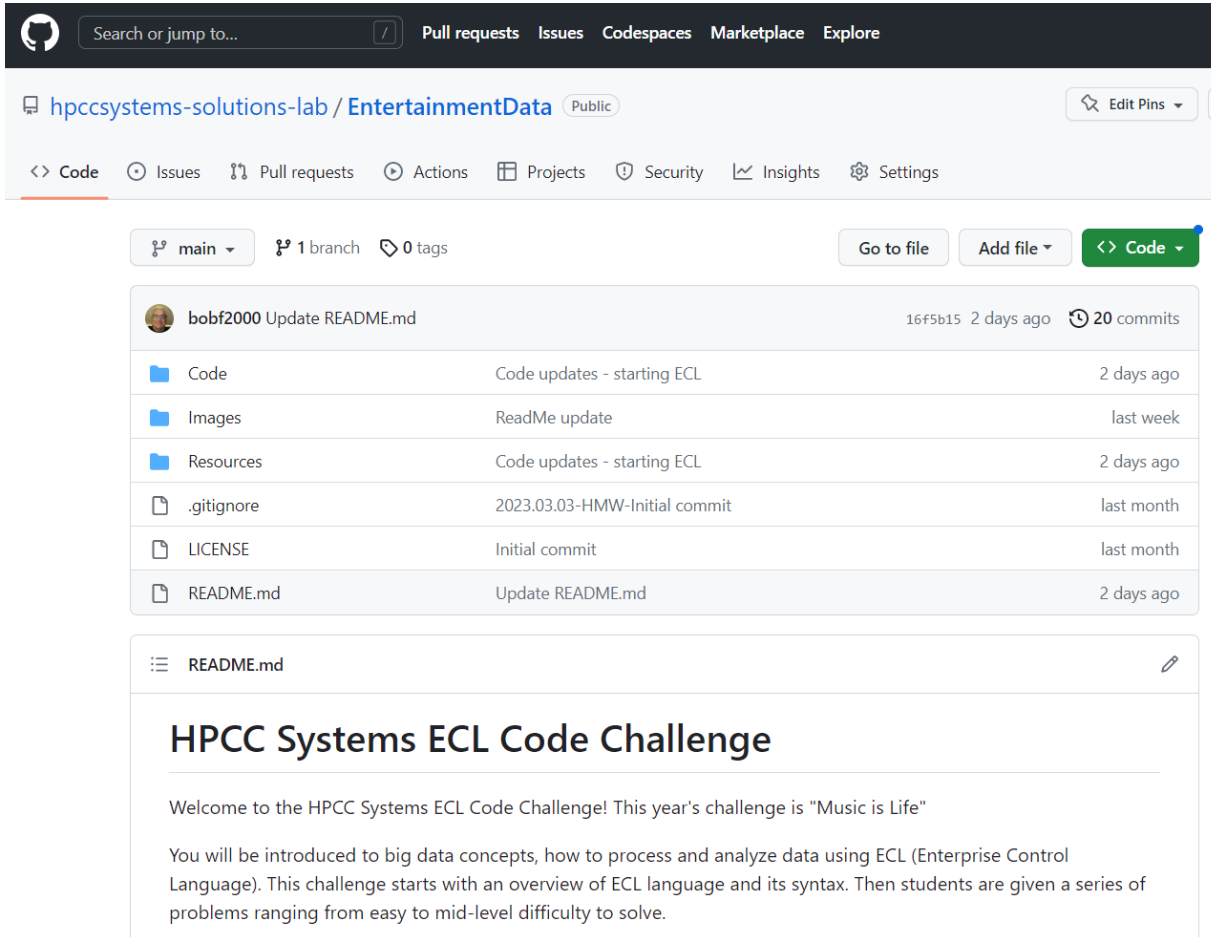
Spray Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/FileSpr

DFU Server: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/WsDfu

ECL Watch URL: http://training.us-hpccsystems-dev.azure.lnrsg.io:8010/esp/fil

The Repo!

https://github.com/hpccsystems-solutions-lab/EntertainmentData



bobf2000 Update README.md 16f5b15 2 days ago 20 commits

Code	Code updates - starting ECL	2 days ago
Images	ReadMe update	last week
Resources	Code updates - starting ECL	2 days ago
.gitignore	2023.03.03-HMW-Initial commit	last month
LICENSE	Initial commit	last month
README.md	Update README.md	2 days ago

README.md

HPCC Systems ECL Code Challenge

Welcome to the HPCC Systems ECL Code Challenge! This year's challenge is "Music is Life"

You will be introduced to big data concepts, how to process and analyze data using ECL (Enterprise Control Language). This challenge starts with an overview of ECL language and its syntax. Then students are given a series of problems ranging from easy to mid-level difficulty to solve.

The Resources!

NSU Code Sharks HPCC Systems Wiki Page:
https://wiki.hpccsystems.com/display/hpcc/Nova+Southeastern+University+-+Code+Sharks

Learn ECL Portal:
https://hpccsystems-solutions-lab.github.io

ECL documentation
https://cdn.hpccsystems.com/releases/CE-Candidate-9.4.4/docs/EN_US/ECLLanguageReference_EN_US-9. 4.4-1.pdf

Visualization document
https://cdn.hpccsystems.com/releases/CE-Candidate-9. 4.4/docs/EN_US/VisualizingECL_EN_US-9. 4.4-1.pdf

Standard Library
https://cdn.hpccsystems.com/releases/CE-Candidate-9.4.4/docs/EN_US/ECLStandardLibraryReference_EN_US-9. 4.4-1.pdf

Machine Learning
https://hpccsystems.com/download/free-modules/machine-learning-library

Music Mozilla (MusicMoz)

Music Mozilla (MusicMOZ)

The dataset you will be working with was extracted and cleaned from the original XML format. ECL supports this format directly, but we wanted to give you a cleaned and extracted (normalized) format by songs (or tracks).

The dataset layout (field information):

name	The artist name behind the release. There are 1276 unique names in the MusicMoz dataset
id	A 16-character unique id for each release. There are a little over 12000 unique releases in this dataset
rtype	Extracted from the original dataset – always “release”
title	Release name.
genre	There are 1000 genre types in MusicMoz (Example, Alternative, Rock, Country, etc.)
releasedate	Release Date in no specific format, generally only year is specified.
disc	This field is not used and is always blank
number	Track number of the release
tracktitle	Name of the track (song)
formats	Wide variety of release formats (over 400)
label	The name of the record company who released the album
catalognumber	Record companies’ catalog number
producers	Comma delimited list of primary producers
coversrc	Web link to Release (Album) Cover art.
guestmusicians	Comma delimited list of guest musicians on the release
description	General free form comments regarding the release.

MusicMoz Challenge Questions:

Category One (MM1):

(A) Count the total records in the dataset

(B) Sort MusicMoz by "name" and display (OUTPUT) the first 50 records(Hint: use CHOOSEN)

(C) Display the first 50 songs by "Rock" genre and then count the total

(C) How many songs were released by “Depeche Mode" beween 1980 and 1989?

(D) How many artists sang the song “My Way"? Display all songs and the total count.

(F) What song(s) in the Music Moz Dataset have the longest "description"?

Category Two (MM2):

(A) How many songs were produced by “U2"? , SORT result by song title, and also display the total count in a separate output.

(B) Count and display all songs where "guest musicians" appeared.

(C) Create a new dataset which only has "TrackTitle", "Title", "Name", and "ReleaseDate"

•Rename the columns to Track, Release, Artist, and Year respectively

•Display the first 50

Category Three (MM3):

(A) Display the number of songs grouped by "Genre", display the first 50 and count your total genres.

(B) What Artist had the most releases between 2001-2010 (releasedate)?

Million Song Dataset (MusicMSD)

Million Song Dataset

The Million Song Dataset (MSD) was first created by a company named Echo Nest(which later was acquired by Spotify in 2014). A lot of the data you will see was used as a basis for creating the Spotify search engine.

In this challenge, the original MSD was cleaned and “slimmed down” for this event.

The data dictionary:

RecID	Unique Record ID
song_id	The original song ID used by Echo Nest, not really used in this challenge
title	song title
year	year song was released
song_hotness	download indicator (0 to 1)
artist_id	original artist id from musicbrainz.org
artist_name	artist name
artist_hotness	overall downloads of artist (0 to 1)
familiarity	search indicator of artist
release_id	Album id where song (title) exists

Million Song Dataset (Continued)

release_name	name of release where song exists
latitude	latitude where the song was recorded
Longitude	Longitude where the song was recorded
Location	where the song was recorded
key	Estimation of the key the song in in by Spotify
key_conf	Confidence of the key estimation
loudness	General loudness of the track relative to -60db
mode	Estimation of mode the song is in by Spotify
mode_conf	Confidence of the mode estimation
duration	Song duration in seconds
start_of_fade_out	Fade out of song in seconds
end_of_fade_in	Fade in to song in seconds
tempo	tempo in beats per minute (BPM)
time_signature	number of beats per bar
time_signature_conf	Confidence of the time signature estimation

Million Song Dataset (Continued)

```
CntBars       Total Bars in the song
AvgBarsConf   //Bars_Analysis
BarsConfDev   //Bars_Analysis
AvgBarsStart  //Bars_Analysis
BarsStartDev  //Bars_Analysis
CntBeats      //Beats_Analysis
AvgBeatsConf  //Beats_Analysis
BeatsConfDev  //Beats_Analysis
AvgBeatsStart //Beats_Analysis
BeatsStartDev //Beats_Analysis
```

A bar is one small segment that holds a number of beats.

Multiple beats make up a bar and multiple bars make up a song.

Beats in a bar is dependent on the time signature of the song.

MSD Challenge Questions:

Category One (MS1):

(A)Reverse sort your dataset by "year", count the total number of records and display only the first 50

(B)Count the total number of songs released in 2010 and display the first 50 results

(C)How many songs were produced by "Prince" in 1982?

(D)Who sang "Into Temptation?"

(E)Sort songs by Artist and Song Title, and output the first 100

(F)What are the hottest songs by year in the Million Song Dataset? Exclude songs with no year value

- Get the datasets' maximum song_hotness value and use it in your output filter.

Category Two (MS2):

(A)Display all songs produced by the artist "Coldplay" that have a "Song Hotness" greater or equal to .75 (>= .75)

- SORT the output by title.
- Also, output the count of the total result

(B)Count all songs whose "Duration" is between 200 AND 250 (inclusive) AND "song_hotness" is not equal to 0 AND "familarity" > .9

(C)Create a new dataset which only has the "Title", "Artist_Name", "Release_Name" and "Year" information.

(D)Calculate Correlation:

- between "song_hotness" AND "artist_hotness“ and between "barsstartdev" AND "beatsstartdev"

MSD Challenge Questions:

Category Three (MS3):

(A)Create a new dataset which only has following conditions

- Column named "Song" that has "Title" values
- Column named "Artist" that has "artist_name" values
- New BOOLEAN Column called isPopular, and it's TRUE is IF "song_hotness" is greater than .80
- New BOOLEAN Column called "IsTooLoud" which is TRUE IF "Loudness" > 0
- Display the first 50
- Result should have 4 columns named "Song", "Artist", "isPopular", and "IsTooLoud"

(B)Display number of songs per "Year" and count total songs released per year

- Result has 2 fields, Year and TotalSongs, verify count is 89

(C)What Artist had the overall hottest songs between 2006-2007?

- Calculate the average "song_hotness" per "Artist_name" for "Year" 2006 and 2007

Spotify 2000 (SpotMusic)

Spotify 2000

Extracted the top 2000 most popular streamed songs using the Spotify API:

```
Data dictionary:
ID:            ID - Record Index
Title:         Name of the Track
Artist:        Name of the Artist
TopGenre:      Genre of the track
Year:          Release Year of the track.
BeatsperMinute: (BPM) The tempo of the song.
Energy:        The energy of a song - the higher the value, the more energtic song.
Danceability:  The higher the value, the easier it is to dance to this song.
Loudness:      The higher the value, the louder the song.
Valence:       The higher the value, the more positive mood for the song.
SongDuration:  The duration of the song.
Acoustic:      The higher the value the more acoustic the song is.
Speechiness:   The higher the value the more spoken words the song contains.
Popularity:    The higher the value the more popular the song is.
```

Spotify 2000 Challenge Questions:

Category One (SP1):

(A) Sort "TopGenre" and count your total music dataset and display the first 50.

(B) Display first 50 songs by "garage rock" genre and then count the total.

(C) Count how many songs were produced by "Prince" in 1984.

(D) Who sang "Into Temptation"?

(E) Sort songs by Artist and title, output the first 100.

(F) Find the least Popular song using "Popularity" field

Category Two (SP2):

(A) Display all songs produced by "Coldplay" Artist AND with a “Popularity” greater or equal to 75 (>= 75) , SORT it by title. Count the result.

(B) Count all songs where "SongDuration" is between 200 AND 250 AND "Speechiness" is above 14. Hint: (SongDuration > 200 AND SongDuration < 250).

(C) Create a new dataset which only has "Artist", "Title" and "Year", Display the first 50. Hint: Create your new layout and use TRANSFORM for new fields. Use PROJECT, to loop through your music dataset

(D) What is the correlation between "Popularity" AND "Liveness"? What is the correlation between "Loudness" AND "Energy"?

Spotify 2000 Challenge Questions:

Category Three (SP3):

(A) Create a new dataset which only has following conditions:

•Column named "Song" that has "Title" values

•New BOOLEAN Column called isPopular, and it's TRUE IF "Popularity" is greater than 80

•New INTEGER Column called "Funkiness" which is "Loudness" + "Danceability"

•Display the first 50

•Hint: Create your new layout and use TRANSFORM for new fields. Use PROJECT, to loop through your music dataset

(B) Display number of songs per "TopGenre", display the first 50 and count your total. Hint: All you need is a TABLE

(C) Calculate average "Danceability" per "Artist" for "Year" 2008 Hint: All you need is a TABLE

pularity" AND "Liveness"? What is the correlation between "Loudness" AND "Energy"?

Bonus Challenge:

Combine the above 3 datasets into a composite dataset with the following format:

```
CombMusicLayout := RECORD
UNSIGNED RECID;
STRING   SongTitle;
STRING   AlbumTitle;
STRING   Artist;
STRING4  ReleaseYear;
END;
```

Remove any duplicate songs, sequence the song records and count the new total.

Final Thoughts

- ✓ Since your solution is the key part to this challenge you can use `#OPTION('obfuscateOutput', TRUE);` at the start of your code to hide it from being viewed on ECL Watchpage. If you decide to use #OPTION make sure to remove if from the WUID that you shared with the judges. When obfuscateOutput set to true, details are removed from the generated workunit, including ECL code, estimates of record size, and number of records.

- ✓ If you want to write the result to a file, make sure the file name starts with your team's name for uniqueness purpose.

- ✓ Make sure the query names are unique and easy to identify. Do not use generic names like test, mentors, or roxie. **We suggest adding your team's name as well.** General names will result in other teams overwriting your files, queries, and results

Let’s Get Started!

Robert.Foreman@lexisnexisrisk.com

HPCC

SYSTEMS