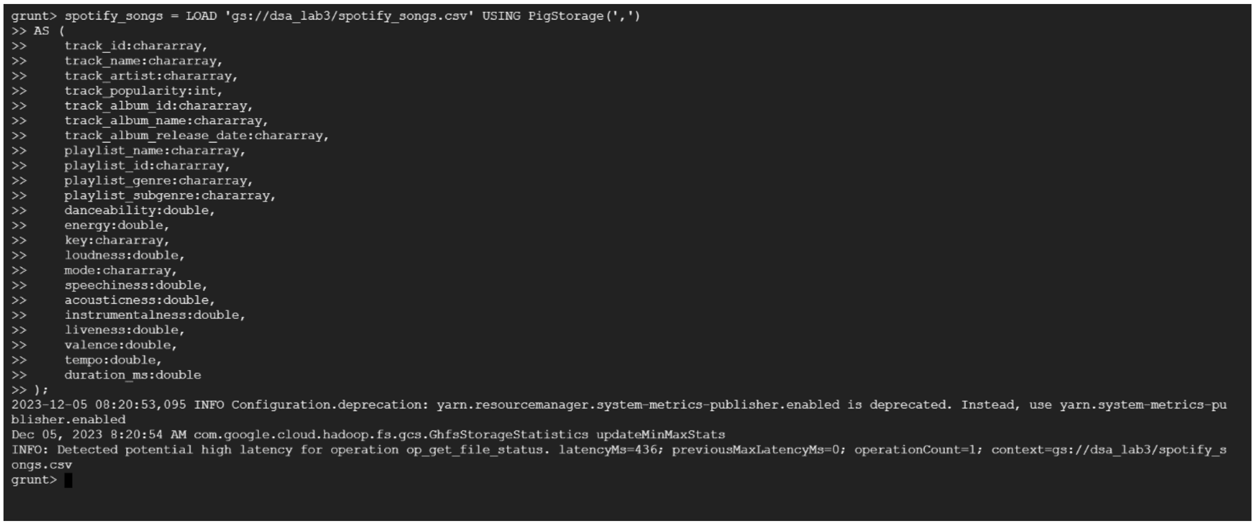
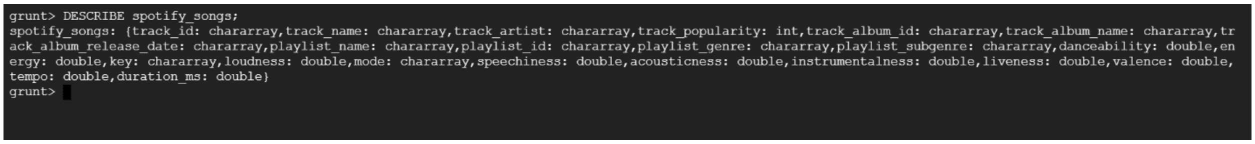
Part 2: Data Manipulation with Apache Pig (5%)

1. Data Loading

Load the dataset into a Pig relation with appropriate data types foreach column.



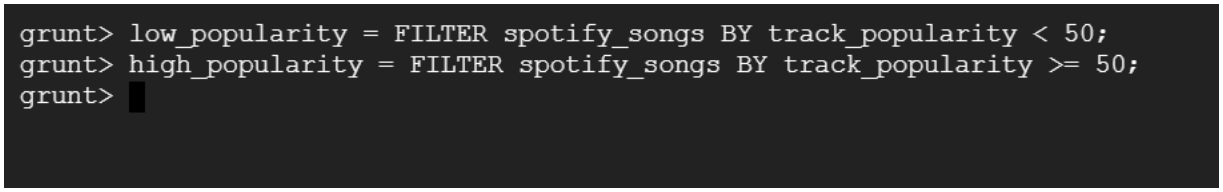
```
grunt> spotify_songs = LOAD 'gs://dsa_lab3/spotify_songs.csv' USING PigStorage(',')
>> AS (
>>     track_id:chararray,
>>     track_name:chararray,
>>     track_artist:chararray,
>>     track_popularity:int,
>>     track_album_id:chararray,
>>     track_album_name:chararray,
>>     track_album_release_date:chararray,
>>     playlist_name:chararray,
>>     playlist_id:chararray,
>>     playlist_genre:chararray,
>>     playlist_subgenre:chararray,
>>     danceability:double,
>>     energy:double,
>>     key:chararray,
>>     loudness:double,
>>     mode:chararray,
>>     speechiness:double,
>>     acousticness:double,
>>     instrumentalness:double,
>>     liveness:double,
>>     valence:double,
>>     tempo:double,
>>     duration_ms:double
>> );
2023-12-05 08:20:53,095 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-pu
blisher.enabled
Dec 05, 2023 8:20:54 AM com.google.cloud.hadoop.fs.gcs.GhfsStorageStatistics updateMinMaxStats
INFO: Detected potential high latency for operation op_get_file_status. latencyMs=436; previousMaxLatencyMs=0; operationCount=1; context=gs://dsa_lab3/spotify_s
ongs.csv
grunt> 
```



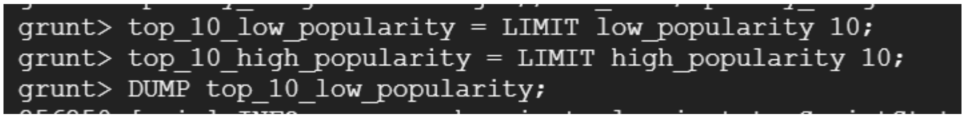
```
grunt> DESCRIBE spotify_songs;
spotify_songs: {track_id: chararray,track_name: chararray,track_artist: chararray,track_popularity: int,track_album_id: chararray,track_album_name: chararray,tr
ack_album_release_date: chararray,playlist_name: chararray,playlist_id: chararray,playlist_genre: chararray,playlist_subgenre: chararray,danceability: double,en
ergy: double,key: chararray,loudness: double,mode: chararray,speechiness: double,acousticness: double,instrumentalness: double,liveness: double,valence: double,
tempo: double,duration_ms: double}
grunt> 
```

2. Data Filtering

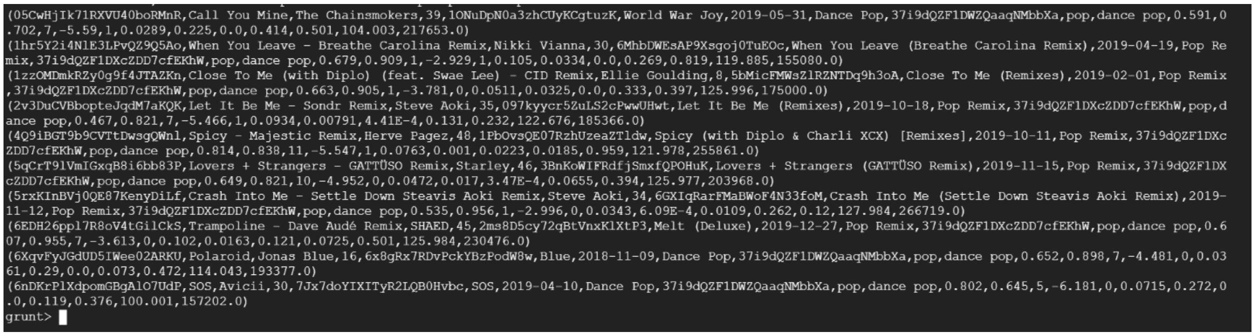
Split data into tracks with a track popularity less than 50(low_popularity) and equal or greater than 50 (high_popularity).



```
grunt> low_popularity = FILTER spotify_songs BY track_popularity < 50;
grunt> high_popularity = FILTER spotify_songs BY track_popularity >= 50;
grunt> 
```



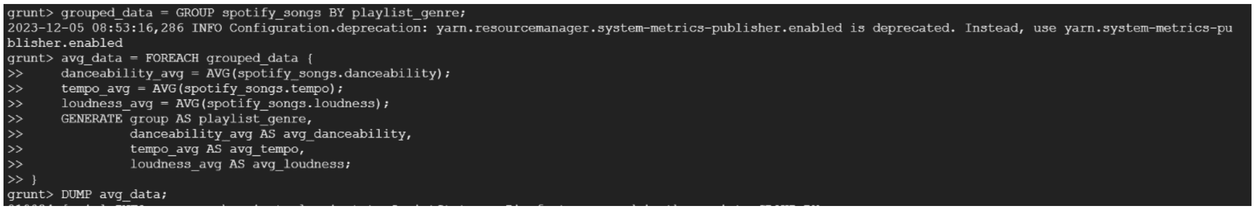
```
grunt> top_10_low_popularity = LIMIT low_popularity 10;
grunt> top_10_high_popularity = LIMIT high_popularity 10;
grunt> DUMP top_10_low_popularity;
```



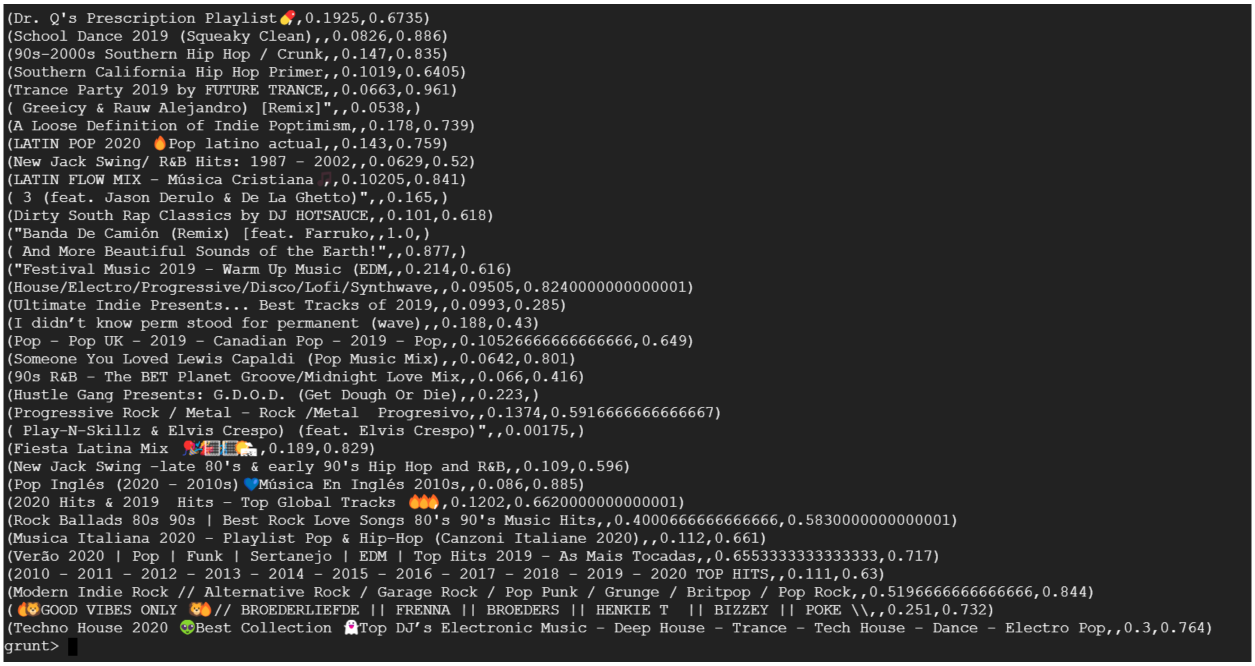
```
(05CwHjIk71RXVU40boRMnR,Call You Mine,The Chainsmokers,39,1ONuDpN0a3zhCUyKCgtuzK,World War Joy,2019-05-31,Dance Pop,37i9dQZF1DWZQaaqNMbbXa,pop,dance pop,0.591,0
.702,7,-5.59,1,0.0289,0.225,0.0,0.414,0.501,104.003,217653.0)
(1hr5Y2i4NlE3LPvQZ9Q5Ao,When You Leave - Breathe Carolina Remix,Nikki Vianna,30,6MhbDWEsAP9Xsgoj0TuEOc,When You Leave (Breathe Carolina Remix),2019-04-19,Pop Re
mix,37i9dQZF1DXcZDD7cfEKhW,pop,dance pop,0.679,0.909,1,-2.929,1,0.105,0.0334,0.0,0.269,0.819,119.885,155080.0)
(1zzOMDmkRZy0g9f4JTAZKn,Close To Me (with Diplo) (feat. Swae Lee) - CID Remix,Ellie Goulding,8,5bMicFMWsZlRZNTDq9h3oA,Close To Me (Remixes),2019-02-01,Pop Remix
,37i9dQZF1DXcZDD7cfEKhW,pop,dance pop,0.663,0.905,1,-3.781,0,0.0511,0.0325,0.0,0.333,0.397,125.996,175000.0)
(2v3DuCVBbopteJqdM7aKQK,Let It Be Me - Sondr Remix,Steve Aoki,35,097kyycr5ZuLS2cPwwUHwt,Let It Be Me (Remixes),2019-10-18,Pop Remix,37i9dQZF1DXcZDD7cfEKhW,pop,d
ance pop,0.467,0.821,7,-5.466,1,0.0934,0.00791,4.41E-4,0.131,0.232,122.676,185366.0)
(4Q9iBGT9b9CVTtDwsgQWnl,Spicy - Majestic Remix,Herve Pagez,48,1PbOvsQE07RzhUzeaZTldw,Spicy (with Diplo & Charli XCX) [Remixes],2019-10-11,Pop Remix,37i9dQZF1DXc
ZDD7cfEKhW,pop,dance pop,0.814,0.838,11,-5.547,1,0.0763,0.001,0.0223,0.0185,0.959,121.978,255861.0)
(5qCrT9lVmIGxqB8i6bb83P,Lovers + Strangers - GATTÜSO Remix,Starley,46,3BnKoWIFRdfjSmxfQPOHuK,Lovers + Strangers (GATTÜSO Remix),2019-11-15,Pop Remix,37i9dQZF1DX
cZDD7cfEKhW,pop,dance pop,0.649,0.821,10,-4.952,0,0.0472,0.017,3.47E-4,0.0655,0.394,125.977,203968.0)
(5rxKInBVj0QE87KenyDiLf,Crash Into Me - Settle Down Steavis Aoki Remix,Steve Aoki,34,6GXIqRarFMaBWoF4N33foM,Crash Into Me (Settle Down Steavis Aoki Remix),2019-
11-12,Pop Remix,37i9dQZF1DXcZDD7cfEKhW,pop,dance pop,0.535,0.956,1,-2.996,0,0.0343,6.09E-4,0.0109,0.262,0.12,127.984,266719.0)
(6EDH26ppl7R8oV4tGilCkS,Trampoline - Dave Audé Remix,SHAED,45,2ms8D5cy72qBtVnxKlXtP3,Melt (Deluxe),2019-12-27,Pop Remix,37i9dQZF1DXcZDD7cfEKhW,pop,dance pop,0.6
07,0.955,7,-3.613,0,0.102,0.0163,0.121,0.0725,0.501,125.984,230476.0)
(6XqvFyJGdUD5IWee02ARKU,Polaroid,Jonas Blue,16,6x8gRx7RDvPckYBzPodW8w,Blue,2018-11-09,Dance Pop,37i9dQZF1DWZQaaqNMbbXa,pop,dance pop,0.652,0.898,7,-4.481,0,0.03
61,0.29,0.0,0.073,0.472,114.043,193377.0)
(6nDKrPlXdpomGBgAlO7UdP,SOS,Avicii,30,7Jx7doYIXITyR2LQB0Hvbc,SOS,2019-04-10,Dance Pop,37i9dQZF1DWZQaaqNMbbXa,pop,dance pop,0.802,0.645,5,-6.181,0,0.0715,0.272,0
.0,0.119,0.376,100.001,157202.0)
grunt> █
```

3. Data Grouping and Aggregation

Group the data by playlist_genre and calculate the averagedanceability, tempo, and loudness for each genre.



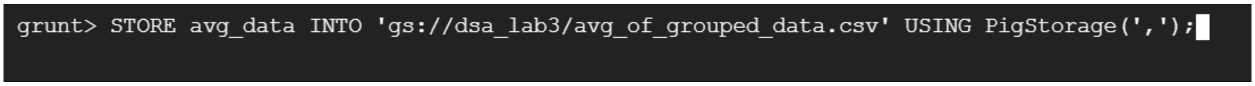
```
grunt> grouped_data = GROUP spotify_songs BY playlist_genre;
2023-12-05 08:53:16,286 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-pu
blisher.enabled
grunt> avg_data = FOREACH grouped_data {
>>     danceability_avg = AVG(spotify_songs.danceability);
>>     tempo_avg = AVG(spotify_songs.tempo);
>>     loudness_avg = AVG(spotify_songs.loudness);
>>     GENERATE group AS playlist_genre,
>>              danceability_avg AS avg_danceability,
>>              tempo_avg AS avg_tempo,
>>              loudness_avg AS avg_loudness;
>> }
grunt> DUMP avg_data;
```



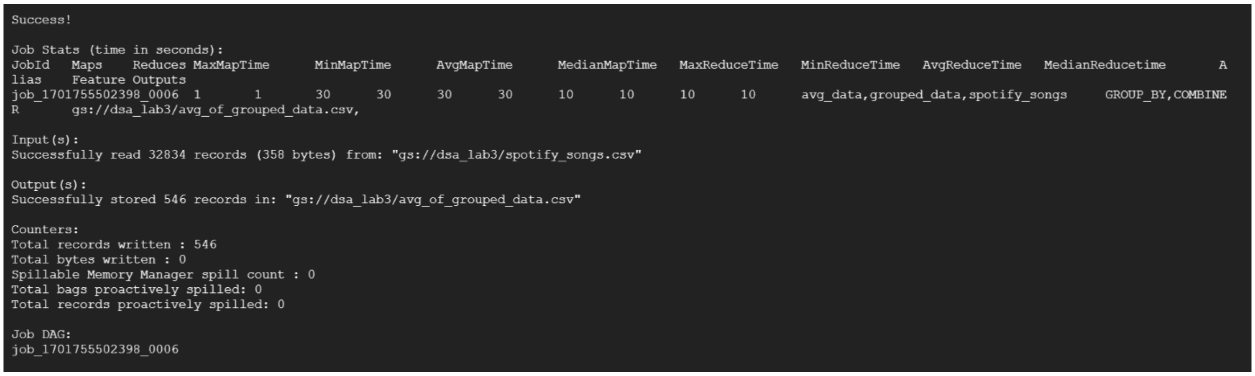
```
(Dr. Q's Prescription Playlist💊,0.1925,0.6735)
(School Dance 2019 (Squeaky Clean),,0.0826,0.886)
(90s-2000s Southern Hip Hop / Crunk,,0.147,0.835)
(Southern California Hip Hop Primer,,0.1019,0.6405)
(Trance Party 2019 by FUTURE TRANCE,,0.0663,0.961)
( Greeicy & Rauw Alejandro) [Remix]",,0.0538,)
(A Loose Definition of Indie Poptimism,,0.178,0.739)
(LATIN POP 2020 🔥Pop latino actual,,0.143,0.759)
(New Jack Swing/ R&B Hits: 1987 - 2002,,0.0629,0.52)
(LATIN FLOW MIX - Música Cristiana 🎵,0.10205,0.841)
( 3 (feat. Jason Derulo & De La Ghetto)",,0.165,)
(Dirty South Rap Classics by DJ HOTSAUCE,,0.101,0.618)
("Banda De Camión (Remix) [feat. Farruko,,1.0,)
( And More Beautiful Sounds of the Earth!",,0.877,)
("Festival Music 2019 - Warm Up Music (EDM,,0.214,0.616)
(House/Electro/Progressive/Disco/Lofi/Synthwave,,0.09505,0.8240000000000001)
(Ultimate Indie Presents... Best Tracks of 2019,,0.0993,0.285)
(I didn’t know perm stood for permanent (wave),,0.188,0.43)
(Pop - Pop UK - 2019 - Canadian Pop - 2019 - Pop,,0.10526666666666666,0.649)
(Someone You Loved Lewis Capaldi (Pop Music Mix),,0.0642,0.801)
(90s R&B - The BET Planet Groove/Midnight Love Mix,,0.066,0.416)
(Hustle Gang Presents: G.D.O.D. (Get Dough Or Die),,0.223,)
(Progressive Rock / Metal - Rock /Metal  Progresivo,,0.1374,0.5916666666666667)
( Play-N-Skillz & Elvis Crespo) (feat. Elvis Crespo)",,0.00175,)
(Fiesta Latina Mix 🎈🎉💃🎹🕺🎹🌞🎫,0.189,0.829)
(New Jack Swing -late 80's & early 90's Hip Hop and R&B,,0.109,0.596)
(Pop Inglés (2020 - 2010s)💙Música En Inglés 2010s,,0.086,0.885)
(2020 Hits & 2019  Hits - Top Global Tracks 🔥🔥🔥,0.1202,0.6620000000000001)
(Rock Ballads 80s 90s | Best Rock Love Songs 80's 90's Music Hits,,0.4000666666666666,0.5830000000000001)
(Musica Italiana 2020 - Playlist Pop & Hip-Hop (Canzoni Italiane 2020),,0.112,0.661)
(Verão 2020 | Pop | Funk | Sertanejo | EDM | Top Hits 2019 - As Mais Tocadas,,0.6553333333333333,0.717)
(2010 - 2011 - 2012 - 2013 - 2014 - 2015 - 2016 - 2017 - 2018 - 2019 - 2020 TOP HITS,,0.111,0.63)
(Modern Indie Rock // Alternative Rock / Garage Rock / Pop Punk / Grunge / Britpop / Pop Rock,,0.5196666666666666,0.844)
( 🔥🦁GOOD VIBES ONLY 🦁🔥// BROEDERLIEFDE || FRENNA || BROEDERS || HENKIE T  || BIZZEY || POKE \\,,0.251,0.732)
(Techno House 2020 👽Best Collection 👻Top DJ’s Electronic Music - Deep House - Trance - Tech House - Dance - Electro Pop,,0.3,0.764)
grunt> █
```

4. Data Exporting

Store the results of the aggregations into files for further analysis or visualization.



```
grunt> STORE avg_data INTO 'gs://dsa_lab3/avg_of_grouped_data.csv' USING PigStorage(',');
```



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        A
lias    Feature Outputs
job_1701755502398_0006  1       1       30      30      30      30      10      10      10      10      avg_data,grouped_data,spotify_songs     GROUP_BY,COMBINE
R       gs://dsa_lab3/avg_of_grouped_data.csv,

Input(s):
Successfully read 32834 records (358 bytes) from: "gs://dsa_lab3/spotify_songs.csv"

Output(s):
Successfully stored 546 records in: "gs://dsa_lab3/avg_of_grouped_data.csv"

Counters:
Total records written : 546
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1701755502398_0006
```