Large Scale Data Management
Athens University of Economics and Business
M.Sc. in Data Science

Programming Project II

Theofanis Nitsos – p3352325

Part I

The part_1.py file is the script for the 1st part of the project that produces messages to the Kafka topic.

A custom JSON encoder is used to overcome some issues that came up during the execution of this script. Then, a loop limited to 1000 counts is used to generate 1000 songs, person name and current timestamp combinations with a 1 second delay. 10 random names are generated using the Faker library and the list is appended with the name 'Mr. Fanis Nitsos', while the song names are extracted from the provided csv file.

Part II

The part_2.py file is the script for the 2nd part of the project. This script effectively sets up a continuous stream processing job using Spark, consuming data from Kafka, processing it, and persisting the results into a Cassandra database.

In detail this script receives JSON information from Kafka (defined in Part I) in the schema of person name, timestamp and song name. It initializes a Spark session and reads the data from the provided CSV. The information from the CSV is joined with the information from Kafka based on the song name and are streamed to Cassandra with a 30 second interval.

Cassandra

The Cassandra table is created using the variable person name (personname) as partition key and the timestamp (listenattime) as clustering key to facilitate aggregations for a particular person and hour. The commands for Cassandra are included in the cassandra.txt file.

```
CREATE TABLE spotify.fanis_records_2 (
   personname TEXT,
   listenattime TIMESTAMP,
   name TEXT,
   song TEXT,
   artists TEXT,
   duration_ms BIGINT,
   album_name TEXT,
   album_release_date DATE,
   danceability FLOAT,
   energy FLOAT,
   key INT,
   loudness FLOAT,
   mode INT,
   speechiness FLOAT,
   acousticness FLOAT,
   instrumentalness FLOAT,
   liveness FLOAT,
   valence FLOAT,
   tempo FLOAT,
   PRIMARY KEY (personname, listenattime)
);
```

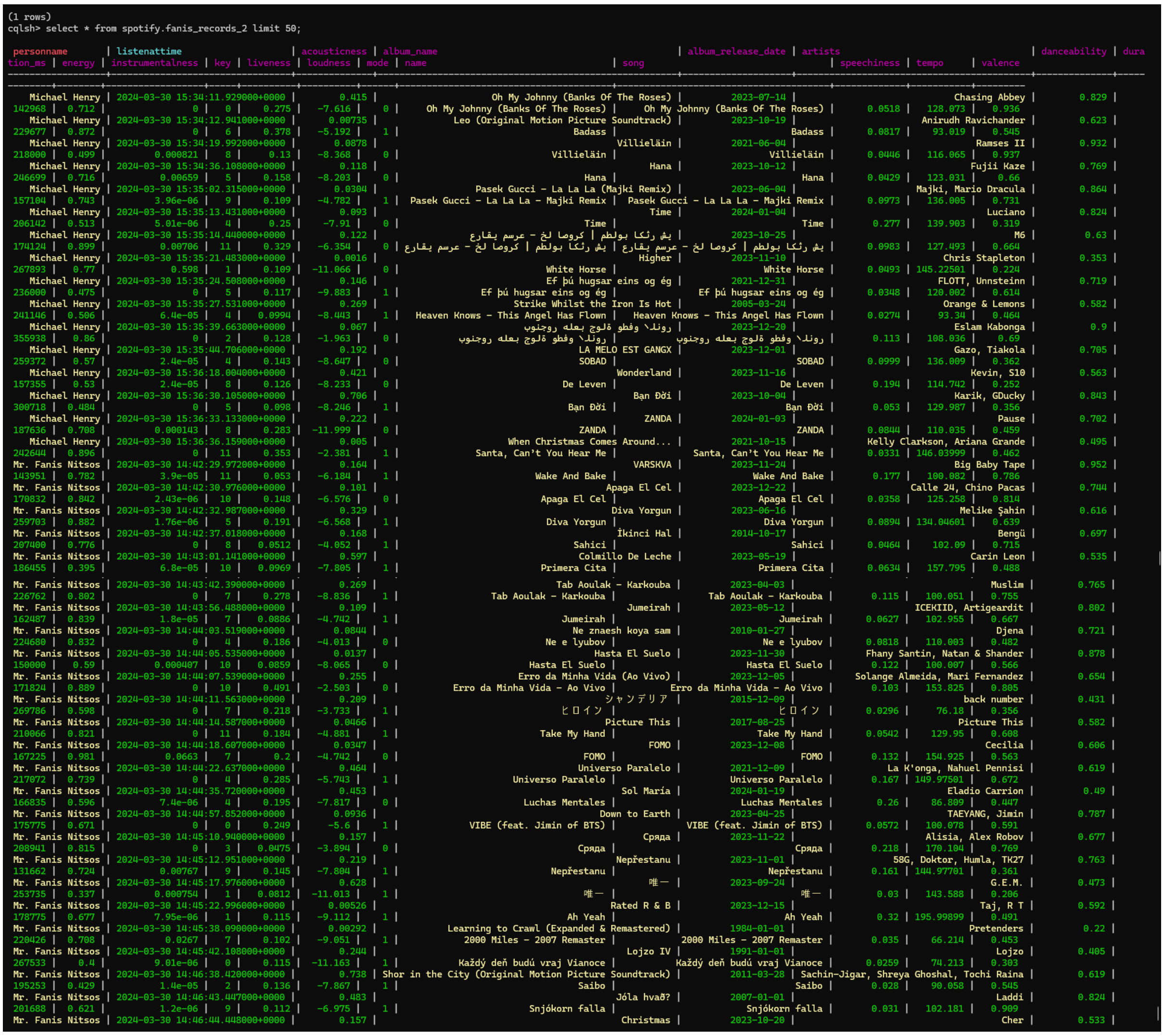
Subsequently 2 queries are executed, using Cassandra’s features to our advantage; filtering the data based on the partition and clustering keys.

1st query calculates the average danceability for the user Mr. Fanis Nitsos between two timestamps

```
select avg(danceability) from spotify.fanis_records_2 where personname='Mr.
Fanis Nitsos' and listenattime >= '2024-03-28 21:20:00' and listenattime <=
'2024-03-28 21:25:00';
```

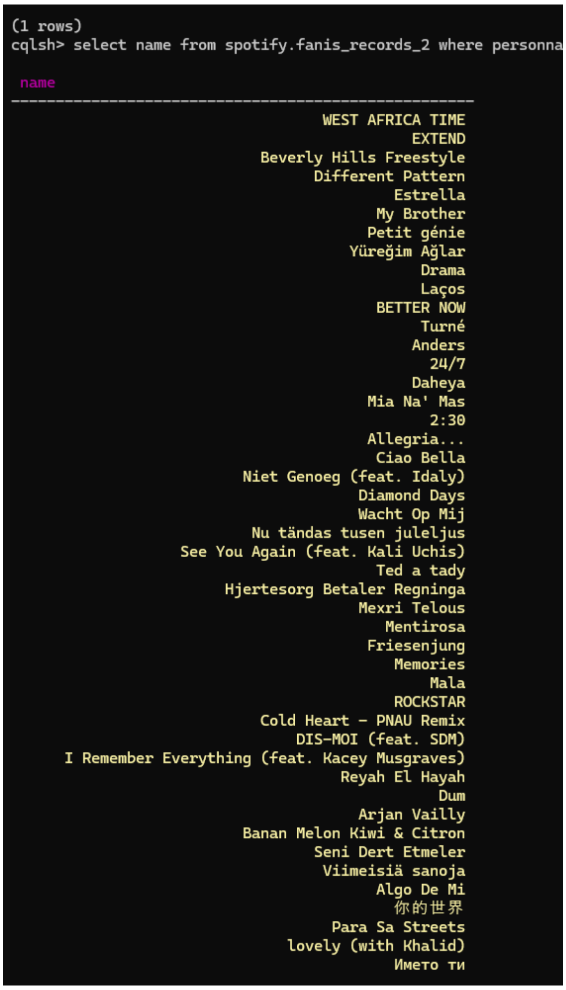
2nd query retrieves the songs player for the user Mr. Fanis Nitsos between two timestamps

```
select name from spotify.fanis_records_2 where personname='Mr. Fanis Nitsos'
and listenattime >= '2024-03-28 21:20:00' and listenattime <= '2024-03-28
21:25:00';
```

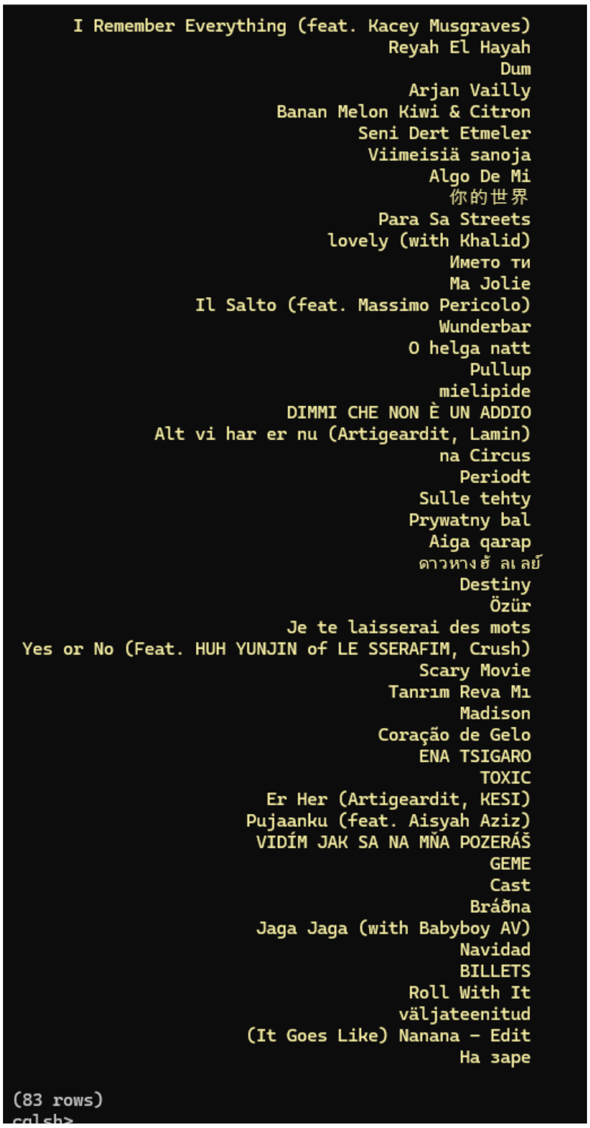


```
(1 rows)
cqlsh> select * from spotify.fanis_records_2 limit 50;

 personname        | listenattime                    | acousticness | album_name                                              | album_release_date | artists                                 | danceability | dura
tion_ms | energy | instrumentalness | key | liveness | loudness | mode | name                                     | song                                     | speechiness | tempo     | valence
-------------------+---------------------------------+--------------+---------------------------------------------------------+--------------------+-----------------------------------------+--------------+-----
--------+--------+------------------+-----+----------+----------+------+------------------------------------------+------------------------------------------+-------------+-----------+---------
      Michael Henry | 2024-03-30 15:34:11.929000+0000 |        0.415 |                          Oh My Johnny (Banks Of The Roses) |         2023-07-14 |                          Chasing Abbey |        0.829 |
 142968 |  0.712 |                0 |   0 |    0.275 |   -7.616 |    0 |        Oh My Johnny (Banks Of The Roses) |        Oh My Johnny (Banks Of The Roses) |      0.0518 |   128.073 |   0.936
      Michael Henry | 2024-03-30 15:34:12.941000+0000 |      0.00735 |                 Leo (Original Motion Picture Soundtrack) |         2023-10-19 |                    Anirudh Ravichander |        0.623 |
 229677 |  0.872 |                0 |   6 |    0.378 |   -5.192 |    1 |                                   Badass |                                   Badass |      0.0817 |    93.019 |   0.545
      Michael Henry | 2024-03-30 15:34:19.992000+0000 |       0.0878 |                                               Villieläin |         2021-06-04 |                              Ramses II |        0.932 |
 218000 |  0.499 |         0.000821 |   8 |     0.13 |   -8.368 |    0 |                               Villieläin |                               Villieläin |      0.0446 |   116.065 |   0.937
      Michael Henry | 2024-03-30 15:34:36.108000+0000 |        0.118 |                                                     Hana |         2023-10-12 |                             Fujii Kaze |        0.769 |
 246699 |  0.716 |          0.00659 |   5 |    0.158 |   -8.203 |    0 |                                     Hana |                                     Hana |      0.0429 |   123.031 |    0.66
      Michael Henry | 2024-03-30 15:35:02.315000+0000 |       0.0304 |                         Pasek Gucci - La La La (Majki Remix) |         2023-06-04 |                   Majki, Mario Dracula |        0.864 |
 157104 |  0.743 |         3.96e-06 |   9 |    0.109 |   -4.782 |    1 |   Pasek Gucci - La La La - Majki Remix |   Pasek Gucci - La La La - Majki Remix |      0.0973 |   136.005 |   0.731
      Michael Henry | 2024-03-30 15:35:13.431000+0000 |        0.093 |                                                     Time |         2024-01-04 |                                Luciano |        0.824 |
 206142 |  0.513 |         5.01e-06 |   4 |     0.25 |    -7.91 |    0 |                                     Time |                                     Time |       0.277 |   139.903 |   0.319
      Michael Henry | 2024-03-30 15:35:14.440000+0000 |        0.122 |                  يش رثكا بولطم | كروصا لخ - عرسم يقارع |         2023-10-25 |                                     M6 |         0.63 |
 174124 |  0.899 |          0.00706 |  11 |    0.329 |   -6.354 |    0 | يش رثكا بولطم | كروصا لخ - عرسم يقارع | يش رثكا بولطم | كروصا لخ - عرسم يقارع |      0.0983 |   127.493 |   0.664
      Michael Henry | 2024-03-30 15:35:21.483000+0000 |       0.0016 |                                                   Higher |         2023-11-10 |                       Chris Stapleton |        0.353 |
 267893 |   0.77 |            0.598 |   1 |    0.109 |  -11.066 |    0 |                              White Horse |                              White Horse |      0.0493 | 145.22501 |   0.224
      Michael Henry | 2024-03-30 15:35:24.508000+0000 |        0.146 |                                 Ef þú hugsar eins og ég |         2021-12-31 |                       FLOTT, Unnsteinn |        0.719 |
 236000 |  0.475 |                0 |   5 |    0.117 |   -9.883 |    1 |                  Ef þú hugsar eins og ég |                  Ef þú hugsar eins og ég |      0.0348 |   120.002 |   0.614
      Michael Henry | 2024-03-30 15:35:27.531000+0000 |        0.269 |                           Strike Whilst the Iron Is Hot |         2005-03-24 |                        Orange & Lemons |        0.582 |
 241146 |  0.506 |          6.4e-05 |   4 |   0.0994 |   -8.443 |    1 |   Heaven Knows - This Angel Has Flown |   Heaven Knows - This Angel Has Flown |      0.0274 |     93.34 |   0.464
      Michael Henry | 2024-03-30 15:35:39.663000+0000 |        0.067 |                                         ةلوج بعله روجنوب |         2023-12-20 |                          Eslam Kabonga |          0.9 |
 355938 |   0.86 |                0 |   2 |    0.128 |   -1.963 |    0 |             رونلا\ وفطو ةلوج بعله روجنوب |             رونلا\ وفطو ةلوج بعله روجنوب |       0.113 |   108.036 |    0.69
      Michael Henry | 2024-03-30 15:35:44.706000+0000 |        0.192 |                                      LA MELO EST GANGX |         2023-12-01 |                          Gazo, Tiakola |        0.705 |
 259372 |   0.57 |          2.4e-05 |   4 |    0.143 |   -8.647 |    0 |                                    SOBAD |                                    SOBAD |      0.0999 |   136.009 |   0.362
      Michael Henry | 2024-03-30 15:36:18.004000+0000 |        0.421 |                                               Wonderland |         2023-11-16 |                             Kevin, S10 |        0.563 |
 157355 |   0.53 |          2.4e-05 |   8 |    0.126 |   -8.233 |    0 |                                 De Leven |                                 De Leven |       0.194 |   114.742 |   0.252
      Michael Henry | 2024-03-30 15:36:30.105000+0000 |        0.706 |                                                  Bạn Đời |         2023-10-04 |                          Karik, GDucky |        0.843 |
 300718 |  0.484 |                0 |   5 |    0.098 |   -8.246 |    1 |                                  Bạn Đời |                                  Bạn Đời |       0.053 |   129.987 |   0.356
      Michael Henry | 2024-03-30 15:36:33.133000+0000 |        0.222 |                                                    ZANDA |         2024-01-03 |                                  Pause |        0.702 |
 187636 |  0.708 |         0.000143 |   8 |    0.283 |  -11.999 |    0 |                                    ZANDA |                                    ZANDA |      0.0844 |   110.035 |   0.459
      Michael Henry | 2024-03-30 15:36:36.159000+0000 |        0.005 |                               When Christmas Comes Around... |         2021-10-15 |         Kelly Clarkson, Ariana Grande |        0.495 |
 242644 |  0.896 |                0 |  11 |    0.353 |   -2.381 |    1 |                Santa, Can’t You Hear Me |                Santa, Can’t You Hear Me |      0.0331 | 146.03999 |   0.462
   Mr. Fanis Nitsos | 2024-03-30 14:42:29.972000+0000 |        0.164 |                                                  VARSKVA |         2023-11-24 |                          Big Baby Tape |        0.952 |
 143951 |  0.782 |          3.9e-05 |  11 |    0.053 |   -6.184 |    1 |                            Wake And Bake |                            Wake And Bake |       0.177 |   100.082 |   0.786
   Mr. Fanis Nitsos | 2024-03-30 14:42:30.976000+0000 |        0.101 |                                             Apaga El Cel |         2023-12-22 |                 Calle 24, Chino Pacas |        0.744 |
 170832 |  0.842 |         2.43e-06 |  10 |    0.148 |   -6.576 |    0 |                             Apaga El Cel |                             Apaga El Cel |      0.0358 |   125.258 |   0.814
   Mr. Fanis Nitsos | 2024-03-30 14:42:32.987000+0000 |        0.329 |                                              Diva Yorgun |         2023-06-16 |                           Melike Şahin |        0.616 |
 259703 |  0.882 |         1.76e-06 |   5 |    0.191 |   -6.568 |    1 |                              Diva Yorgun |                              Diva Yorgun |      0.0894 | 134.04601 |   0.639
   Mr. Fanis Nitsos | 2024-03-30 14:42:37.018000+0000 |        0.168 |                                              İkinci Hal |         2014-10-17 |                                  Bengü |        0.697 |
 207400 |  0.776 |                0 |   8 |   0.0512 |   -4.052 |    1 |                                   Sahici |                                   Sahici |      0.0464 |    102.09 |   0.715
   Mr. Fanis Nitsos | 2024-03-30 14:43:01.141000+0000 |        0.597 |                                       Colmillo De Leche |         2023-05-19 |                             Carin Leon |        0.535 |
 186455 |  0.395 |          6.8e-05 |  10 |   0.0969 |   -7.805 |    1 |                             Primera Cita |                             Primera Cita |      0.0634 |   157.795 |   0.488

   Mr. Fanis Nitsos | 2024-03-30 14:43:42.390000+0000 |        0.269 |                                   Tab Aoulak - Karkouba |         2023-04-03 |                                 Muslim |        0.765 |
 226762 |  0.802 |                0 |   7 |    0.278 |   -8.836 |    1 |                    Tab Aoulak - Karkouba |                    Tab Aoulak - Karkouba |       0.115 |   100.051 |   0.755
   Mr. Fanis Nitsos | 2024-03-30 14:43:56.488000+0000 |        0.109 |                                                 Jumeirah |         2023-05-12 |                  ICEKIID, Artigeardit |        0.802 |
 162487 |  0.839 |          1.8e-05 |   7 |   0.0886 |   -4.742 |    1 |                                 Jumeirah |                                 Jumeirah |      0.0627 |   102.955 |   0.667
   Mr. Fanis Nitsos | 2024-03-30 14:44:03.519000+0000 |       0.0844 |                                       Ne znaesh koya sam |         2010-01-27 |                                  Djena |        0.721 |
 224680 |  0.832 |                0 |   4 |    0.186 |   -4.013 |    0 |                              Ne e lyubov |                              Ne e lyubov |      0.0818 |   110.003 |   0.482
   Mr. Fanis Nitsos | 2024-03-30 14:44:05.535000+0000 |       0.0137 |                                          Hasta El Suelo |         2023-11-30 |          Fhany Santin, Natan & Shander |        0.878 |
 150000 |   0.59 |         0.000407 |  10 |   0.0859 |   -8.065 |    0 |                           Hasta El Suelo |                           Hasta El Suelo |       0.122 |   100.007 |   0.566
   Mr. Fanis Nitsos | 2024-03-30 14:44:07.539000+0000 |        0.255 |                              Erro da Minha Vida (Ao Vivo) |         2023-12-05 |      Solange Almeida, Mari Fernandez |        0.654 |
 171824 |  0.889 |                0 |  10 |    0.491 |   -2.503 |    0 |             Erro da Minha Vida - Ao Vivo |             Erro da Minha Vida - Ao Vivo |       0.103 |   153.825 |   0.805
   Mr. Fanis Nitsos | 2024-03-30 14:44:11.563000+0000 |        0.209 |                                                シャンデリア |         2015-12-09 |                            back number |        0.431 |
 269786 |  0.598 |                0 |   7 |    0.218 |   -3.733 |    1 |                                  ヒロイン |                                  ヒロイン |      0.0296 |     76.18 |   0.356
   Mr. Fanis Nitsos | 2024-03-30 14:44:14.587000+0000 |       0.0466 |                                             Picture This |         2017-08-25 |                           Picture This |        0.582 |
 210066 |  0.821 |                0 |  11 |    0.184 |   -4.881 |    1 |                             Take My Hand |                             Take My Hand |      0.0542 |    129.95 |   0.608
   Mr. Fanis Nitsos | 2024-03-30 14:44:18.607000+0000 |       0.0347 |                                                     FOMO |         2023-12-08 |                                Cecilia |        0.606 |
 167225 |  0.981 |           0.0663 |   7 |      0.2 |   -4.742 |    0 |                                     FOMO |                                     FOMO |       0.132 |   154.925 |   0.563
   Mr. Fanis Nitsos | 2024-03-30 14:44:22.637000+0000 |        0.464 |                                        Universo Paralelo |         2021-12-09 |          La K'onga, Nahuel Pennisi |        0.619 |
 217072 |  0.739 |                0 |   4 |    0.285 |   -5.743 |    1 |                        Universo Paralelo |                        Universo Paralelo |       0.167 | 149.97501 |   0.672
   Mr. Fanis Nitsos | 2024-03-30 14:44:35.720000+0000 |        0.453 |                                                Sol María |         2024-01-19 |                         Eladio Carrion |         0.49 |
 166835 |  0.596 |          7.4e-06 |   4 |    0.195 |   -7.817 |    0 |                          Luchas Mentales |                          Luchas Mentales |        0.26 |    86.809 |   0.447
   Mr. Fanis Nitsos | 2024-03-30 14:44:57.852000+0000 |       0.0936 |                                            Down to Earth |         2023-04-25 |                          TAEYANG, Jimin |        0.787 |
 175775 |  0.671 |                0 |   0 |    0.249 |     -5.6 |    1 |               VIBE (feat. Jimin of BTS) |               VIBE (feat. Jimin of BTS) |      0.0572 |   100.078 |   0.591
   Mr. Fanis Nitsos | 2024-03-30 14:45:10.940000+0000 |        0.157 |                                                    Сряда |         2023-11-22 |                    Alisia, Alex Robov |        0.677 |
 208941 |  0.815 |                0 |   3 |   0.0475 |   -3.894 |    0 |                                    Сряда |                                    Сряда |       0.218 |   170.104 |   0.769
   Mr. Fanis Nitsos | 2024-03-30 14:45:12.951000+0000 |        0.219 |                                               Nepřestanu |         2023-11-01 |              58G, Doktor, Humla, TK27 |        0.763 |
 131662 |  0.724 |          0.00767 |   9 |    0.145 |   -7.804 |    1 |                               Nepřestanu |                               Nepřestanu |       0.161 | 144.97701 |   0.361
   Mr. Fanis Nitsos | 2024-03-30 14:45:17.976000+0000 |        0.628 |                                                     唯一 |         2023-09-24 |                                 G.E.M. |        0.473 |
 253735 |  0.337 |         0.000754 |   1 |   0.0812 |  -11.013 |    1 |                                     唯一 |                                     唯一 |        0.03 |   143.588 |   0.206
   Mr. Fanis Nitsos | 2024-03-30 14:45:22.996000+0000 |      0.00526 |                                              Rated R & B |         2023-12-15 |                              Taj, R T |        0.592 |
 178775 |  0.677 |         7.95e-06 |   1 |    0.115 |   -9.112 |    1 |                                  Ah Yeah |                                  Ah Yeah |        0.32 | 195.99899 |   0.491
   Mr. Fanis Nitsos | 2024-03-30 14:45:38.090000+0000 |      0.00292 |             Learning to Crawl (Expanded & Remastered) |         1984-01-01 |                             Pretenders |         0.22 |
 220426 |  0.708 |           0.0267 |   7 |    0.102 |   -9.051 |    1 |               2000 Miles - 2007 Remaster |               2000 Miles - 2007 Remaster |       0.035 |    66.214 |   0.453
   Mr. Fanis Nitsos | 2024-03-30 14:45:42.108000+0000 |        0.244 |                                                 Lojzo IV |         1991-01-01 |                                  Lojzo |        0.405 |
 267533 |    0.4 |         9.01e-06 |   0 |    0.115 |  -11.163 |    1 |               Každý deň budú vraj Vianoce |               Každý deň budú vraj Vianoce |      0.0259 |    74.213 |   0.303
   Mr. Fanis Nitsos | 2024-03-30 14:46:38.420000+0000 |        0.738 | Shor in the City (Original Motion Picture Soundtrack) |         2011-03-28 | Sachin-Jigar, Shreya Ghoshal, Tochi Raina |        0.619 |
 195253 |  0.429 |          1.4e-05 |   2 |    0.136 |   -7.867 |    1 |                                    Saibo |                                    Saibo |       0.028 |    90.058 |   0.545
   Mr. Fanis Nitsos | 2024-03-30 14:46:43.447000+0000 |        0.483 |                                               Jóla hvað? |         2007-01-01 |                                  Laddi |        0.824 |
 201688 |  0.621 |          1.2e-06 |   9 |    0.112 |   -6.975 |    1 |                           Snjókorn falla |                           Snjókorn falla |       0.031 |   102.181 |   0.909
   Mr. Fanis Nitsos | 2024-03-30 14:46:44.448000+0000 |        0.157 |                                                Christmas |         2023-10-20 |                                   Cher |        0.533 |
```

Figure 1 Output of the 1st Cassandra Query & 50 persisted lines of the Cassandra table



```
(1 rows)
cqlsh> select name from spotify.fanis_records_2 where personna

 name
----------------------------------------------------
                                    WEST AFRICA TIME
                                              EXTEND
                               Beverly Hills Freestyle
                                   Different Pattern
                                            Estrella
                                          My Brother
                                         Petit génie
                                        Yüreğim Ağlar
                                               Drama
                                               Laços
                                          BETTER NOW
                                               Turné
                                              Anders
                                                24/7
                                              Daheya
                                         Mia Na' Mas
                                                2:30
                                         Allegria...
                                          Ciao Bella
                             Niet Genoeg (feat. Idaly)
                                        Diamond Days
                                        Wacht Op Mij
                                 Nu tändas tusen juleljus
                         See You Again (feat. Kali Uchis)
                                          Ted a tady
                              Hjertesorg Betaler Regninga
                                        Mexri Telous
                                           Mentirosa
                                         Friesenjung
                                            Memories
                                                Mala
                                            ROCKSTAR
                               Cold Heart - PNAU Remix
                                  DIS-MOI (feat. SDM)
            I Remember Everything (feat. Kacey Musgraves)
                                        Reyah El Hayah
                                                 Dum
                                        Arjan Vailly
                             Banan Melon Kiwi & Citron
                                   Seni Dert Etmeler
                                     Viimeisiä sanoja
                                         Algo De Mi
                                              你的世界
                                     Para Sa Streets
                                 lovely (with Khalid)
                                           Името ти
```



```
      I Remember Everything (feat. Kacey Musgraves)
                                     Reyah El Hayah
                                                Dum
                                        Arjan Vailly
                           Banan Melon Kiwi & Citron
                                   Seni Dert Etmeler
                                    Viimeisiä sanoja
                                         Algo De Mi
                                            你的世界
                                     Para Sa Streets
                                lovely (with Khalid)
                                           Името ти
                                           Ma Jolie
                   Il Salto (feat. Massimo Pericolo)
                                          Wunderbar
                                        O helga natt
                                             Pullup
                                          mielipide
                             DIMMI CHE NON È UN ADDIO
              Alt vi har er nu (Artigeardit, Lamin)
                                          na Circus
                                            Periodt
                                        Sulle tehty
                                       Prywatny bal
                                         Aiga qarap
                                      ดาวหางฮัลเลย์
                                            Destiny
                                              Özür
                           Je te laisserai des mots
 Yes or No (Feat. HUH YUNJIN of LE SSERAFIM, Crush)
                                        Scary Movie
                                     Tanrım Reva Mı
                                            Madison
                                     Coração de Gelo
                                        ENA TSIGARO
                                              TOXIC
                            Er Her (Artigeardit, KESI)
                           Pujaanku (feat. Aisyah Aziz)
                            VIDÍM JAK SA NA MŇA POZERÁŠ
                                               GEME
                                               Cast
                                             Bráðna
                        Jaga Jaga (with Babyboy AV)
                                            Navidad
                                            BILLETS
                                       Roll With It
                                      väljateenitud
                        (It Goes Like) Nanana - Edit
                                            Ha зape

(83 rows)
cqlsh>
```

Output of the 2nd Cassandra query