M.Sc. in Data Science

Large Scale Data Management

Assignment 2 - End-to-End Data Streaming

Phevos A. Margonis - f3352317

Instructor: Liakos Panagiotis

March 5, 2024

Abstract

This paper explores the integration of Apache Spark and Apache Cassandra within a Structured Streaming Spark process to handle Kafka message consumption and persistence in a simulated real-world environment. Utilizing a virtual machine for Spark and Docker containers for Kafka and Cassandra, the study outlines a comprehensive methodology involving data generation, streaming, processing, and persistence. Key components include the use of Apache Kafka for real-time data streaming, Spark for data processing, and Cassandra NoSQL database as a persistence sink, demonstrating a scalable and fault-tolerant system for real-time data analytics.

Contents

1 Introduction

In this project, we utilize the Apache Spark framework and the Apache Cassandra NoSQL database to create a Structured Streaming Spark process. This process consumes Kafka messages and uses Cassandra as a sink to persist information. We simulate this real-world environment using a virtual machine with the Spark framework installed locally, and the Kafka-broker and Cassandra NoSQL on two separate Docker containers.

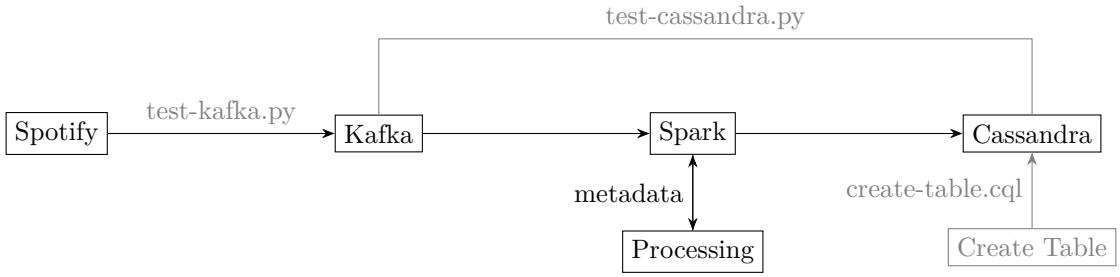

Figure 1: Workflow overview

2 Documentation

2.1 Kafka-broker

The first part of this Structured Streaming pipeline involves the generation and delivery of streamed data to a Kafka-broker. Apache Kafka, a distributed streaming platform, is employed for building real-time data pipelines and streaming applications. It facilitates the publishing, subscription to, storage, and processing of streams of records in a fault-tolerant and scalable manner.

This scenario simulates a case where Spotify users' listening activities are logged. A preparatory step involves installing the pandas Python library on the VM. A Python script (Listing 6) then creates a list of 10 random users using the `Faker` library, with the name of the author added as an eleventh user for testing purposes. Spotify songs are loaded using pandas, and a random song is assigned to each user. The users' name, the song they listened to, and the request time are packed into JSON objects and sent to the Kafka-broker via the `AIOKafkaProducer` library. An asynchronous loop sends bursts of data for all users every 10 seconds for demonstration purposes.

2.2 Structured Spark Streaming

The second part employs the Apache Spark framework to receive, process, and persist the messages. Spark, a unified analytics engine, is known for its speed, ease of use, and general-purpose computing capabilities. Using a python script (Listing 7) a SparkSession is initiated to receive the messages, and a DataFrame containing song metadata is loaded, striped of double quotes and cached to speed up operations. The streamed data are decoded from JSON to string, string and long-type for the entries name, song and timestamp respectively, and are stored in a separate DataFrame. A LeftJoin operation combines user activity with the metadata. A writeStream operation appends the processed data to a Cassandra NoSQL table, with options set for efficiency and fault-tolerance. The `trigger` option creates batches of streamed data every 30 seconds, employing the `update` output mode for load-balancing. Crucially, the `checkpointLocation` option safeguards the state of operations, providing fault tolerance in the event a node (in this context, a Docker container) fails. Upon such failure, Spark consults the checkpoint file to prioritize processing of uncompleted messages before proceeding to newer batches.

2.3 Cassandra NoSQL Database

Finally, the Apache Cassandra NoSQL database acts as a sink to persist the streamed information. Cassandra, a scalable and distributed NoSQL database, ensures high availability with no single point of failure.

Initially, a `.cql` script (Listing 8) defines the schema for storing data. This script starts by removing any existing `Keyspace` named *Spotify*, followed by the creation of a new `Keyspace` with a simple strategy class and a replication factor of 1, suitable for local experimentation. The

table's schema is then established, detailing the necessary fields and their corresponding data types to accommodate the streamed data. Notably, the timestamp is read as a `LongType` integer representing milliseconds, but it is stored in the database as a `timestamp` type and converted to a date format upon querying to improve readability.

Selecting an appropriate primary key is essential in Cassandra for efficient data distribution across nodes, which facilitates streamlined querying, retrieval, and storage. For this project, aimed at analyzing users' listening patterns, the primary key consists of the user's name as the partition key and the timestamp as the clustering key. Additionally, a sample of 50 persisted lines, along with various `cql` queries and their outcomes are documented in (Appendix A).

3 Walk-through

Listing 1: Steps for Reproducing the Experiment

```
vagrant up
vagrant ssh
python3 -m pip install pandas

# Restart Exited Containers
docker start $(docker ps -a -q -f status=exited)

# Tab 2:
cd /vagrant
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0,
↪  com.datastax.spark:spark-cassandra-connector_2.12:3.0.0 test-cassandra.py

# Tab 3:
cd /vagrant
docker cp create_table.cql cassandra:/
docker exec -it cassandra bash
cqlsh -f create_table.cql
cqlsh
use spotify
select * from spotify.records
truncate records

# Tab 1:
cd /vagrant
python3 test-kafka.py

# Tab 3:
exit
nodetool flush spotify records
nodetool tablehistograms spotify records
```

Percentile	Read Latency (micros)	Write Latency (micros)	SSTables	Partition Size (bytes)	Cell Count
50%	0.00	0.00	0.00	924	103
75%	0.00	0.00	0.00	924	103
95%	0.00	0.00	0.00	1109	103
98%	0.00	0.00	0.00	3973	446
99%	0.00	0.00	0.00	3973	446
Min	0.00	0.00	0.00	643	87
Max	0.00	0.00	0.00	3973	446

Table 1: Nodetool's Spotify/Records Histograms

A Appendix - Outputs

Listing 2: Sample of 50 persisted lines in the Cassandra table.

```
Michael Hill,2024-03-02 16:49:03.540+0000,0.0437,Podme bratia do Betlehema,2007-11-30,Tublatanka,0.365,252520,0.897,0,9,0.167,-6.515,1,Podme bratia do
↪  Betlehema,0.116,140.011,0.137
Michael Hill,2024-03-02 16:49:13.596+0000,0.746,Waterdicht,2023-10-20,Hannah Mae,0.689,171932,0.328,0,0,0.141,-10.667,1,Waterdicht,0.0396,119.104,0.531
Michael Hill,2024-03-02 16:49:23.658+0000,0.51,NewJeans 'Super Shy',2023-07-07,NewJeans,0.807,108986,0.721,0.00216,4,0.115,-6.211,0,New
↪  Jeans,0.0508,134.01401,0.532
Michael Hill,2024-03-02 16:49:33.733+0000,0.276,   ,2023-01-26,DOROFEEVA,0.766,169586,0.613,0,4,0.0852,-6.257,0,   ,0.0417,120.945,0.142
Michael Hill,2024-03-02 16:49:43.834+0000,0.0112,Christmas,2023-10-20,"Cher, Stevie Wonder",0.362,155360,0.818,0,0,0.164,-4.274,1,What Christmas Means To Me
↪  (with Stevie Wonder),0.0634,168.271,0.71
Michael Hill,2024-03-02 16:49:53.930+0000,0.251,shhhhhhh..,2023-10-07,"WEAN, tlinh",0.739,230675,0.537,0,2,0.0998,-7.241,0,shhhhhhh..,0.0394,137.97,0.379
Jonathan Riley,2024-03-02 17:09:44.202+0000,0.235,Mon coeur avait raison,2015-08-28,GIMS,0.779,237093,0.852,0,2,0.212,-3.353,1,Est-ce que tu m'aimes ? - Pilule
↪  bleue,0.0651,119.892,0.601
Jonathan Riley,2024-03-02 17:09:54.332+0000,0.47   ,    -           |        2023,-10-25,M6,0.583,242600,0.863,0.000203,10,0.301,-4.521,1   ,    -           |
↪          0.0481,,127.566,0.612
Jonathan Riley,2024-03-02 17:10:04.433+0000,0.157,Mesh Shayfenhom ( Coke Studio Egypt 2023 ),2023-08-03,"Hassan El Shafei, Bosy, Double
↪  Zuksh",0.741,222760,0.905,1.13e-06,7,0.189,-3.336,0,Mesh Shayfenhom ( Coke Studio Egypt 2023 ),0.0442,124.008,0.668
Jonathan Riley,2024-03-02 17:10:14.526+0000,0.337,The Twilight Saga: Breaking Dawn - Part 1 (Original Motion Picture Soundtrack),2011-11-04,Bruno
↪  Mars,0.576,257720,0.835,0,2,0.082,-6.826,1,It Will Rain,0.0486,150.017,0.476
Jonathan Riley,2024-03-02 17:10:24.623+0000,0.0694,moonfantom,2021-12-17,Hiro,0.824,206653,0.416,0.00374,2,0.095,-9.471,1,       ,0.132,105.034,0.445
Jonathan Riley,2024-03-02 17:10:34.724+0000,0.0163,Leo,2024-01-05,Shubh,0.749,143200,0.571,8.41e-06,1,0.136,-9.111,0,Safety Off,0.0504,90.014,0.329
Austin Williams,2024-03-02 16:49:03.527+0000,0.0823,Ăra Ărata,2023-04-07,nublu,0.843,179812,0.738,0.0505,6,0.141,-6.066,0,Ăra Ărata,0.0612,113.972,0.535
Austin Williams,2024-03-02 16:49:13.590+0000,0.00961,1989 (Taylor's Version),2023-10-26,Taylor Swift,0.761,212600,0.607,2.2e-05,7,0.367,-4.83,1,Welcome To New
↪  York (Taylor's Version),0.0312,116.98,0.674
Austin Williams,2024-03-02 16:49:23.647+0000,0.382,DESHPERADO,2022-10-25,"DESH, Azahriah",0.833,140480,0.722,9.9e-05,5,0.0986,-7.579,0,Papa,0.0748,104.017,0.425
Austin Williams,2024-03-02 16:49:33.713+0000,0.417,"Stockholm, Sweden",2023-11-17,Yasin,0.748,160000,0.588,0,5,0.0983,-8.157,0,"Stockholm,
↪  Sweden",0.41,144.23399,0.801
Austin Williams,2024-03-02 16:49:43.809+0000,0.107,27,2023-11-24,ElGrandeToto,0.479,211906,0.842,1.9e-05,4,0.0968,-5.815,0,27,0.0426,113.955,0.0815
Austin Williams,2024-03-02 16:49:53.900+0000,0.531,París como Hakimi,2024-01-12,Morad,0.835,185864,0.699,0,6,0.331,-5.53,0,París como
↪  Hakimi,0.0717,133.02901,0.905
Michael Randall,2024-03-02 17:20:56.459+0000,0.207,Que La Choque,2023-10-01,Rochy RD,0.71,125004,0.544,0,5,0.147,-6.55,1,Que La Choque,0.298,119.285,0.912
Michael Randall,2024-03-02 17:21:06.521+0000,0.174,Görmem Böylesini,2023-08-04,"Sefo, Simge",0.833,192446,0.645,4.1e-05,4,0.247,-4.585,0,Görmem
↪  Böylesini,0.102,93.981,0.587
Michael Randall,2024-03-02 17:21:16.567+0000,0.0267,Neonlys,2014-03-17,Ukendt Kunstner,0.796,187013,0.801,0,11,0.0965,-4.673,1,Neonlys,0.0316,124.905,0.71
Michael Randall,2024-03-02 17:21:26.622+0000,0.346,Boy Boy,2023-06-21,"Yaisel LM, Hansel El De La H",0.643,134565,0.578,0.00249,10,0.149,-3.318,1,Boy
↪  Boy,0.461,152.58501,0.71
Michael Randall,2024-03-02 17:21:36.679+0000,0.053,SZÍNVAK,2022-10-26,VALMAR,0.611,198278,0.767,0,10,0.288,-4.695,0,SZÍNVAK,0.038,133.88499,0.803
Michael Randall,2024-03-02 17:21:46.747+0000,0.0825,RÄPPÄRI,2023-11-03,ibe,0.755,209677,0.561,0.000173,2,0.0937,-5.694,1,FAMOUS,0.0413,93,0.311
Caitlin Ortiz,2024-03-02 17:09:44.199+0000,0.025,Talk That Talk,2011-11-18,"Rihanna, Calvin Harris",0.734,215226,0.766,0.00138,1,0.108,-4.485,1,We Found
↪  Love,0.0383,127.986,0.6
Caitlin Ortiz,2024-03-02 17:09:54.321+0000,0.783,Heading South,2019-09-30,Zach Bryan,0.68,171692,0.246,0,4,0.106,-14.112,1,Heading South,0.058,110.23,0.388
Caitlin Ortiz,2024-03-02 17:10:04.427+0000,0.107,MARATON (From \"Haita De Acţiune\" The Movie),2023-03-10,M.G.L.,0.724,158297,0.708,0,4,0.153,-5.868,0,MARATON -
↪  From \"Haita De Acţiune\" The Movie,0.0424,93.985,0.19
Caitlin Ortiz,2024-03-02 17:10:14.520+0000,0.131,                ,2023-04-25,         ,0.718,131666,0.682,0,1,0.484,-6.411,0,            ,0.0329,89.98,0.711
Caitlin Ortiz,2024-03-02 17:10:24.614+0000,0.00627,Swimming Pools (Drank),2012-01-01,Kendrick Lamar,0.577,247800,0.447,0.00022,8,0.0899,-5.892,1,Swimming Pools
↪  (Drank),0.277,74.236,0.189
Caitlin Ortiz,2024-03-02 17:10:34.709+0000,0.544,Petals to Thorns,2023-05-26,d4vd,0.571,242484,0.458,9.3e-05,4,0.123,-9.283,1,Here With
↪  Me,0.0258,132.02499,0.299
Theresa Thompson,2024-03-02 17:09:44.178+0000,0.108,Le cose cambiano,2023-12-01,Massimo Pericolo,0.627,168263,0.636,0.0002,8,0.114,-7.602,0,Ancora
↪  Qua,0.165,75.021,0.409
Theresa Thompson,2024-03-02 17:09:54.309+0000,0.116,   ,2023-04-28,Keung To,0.425,198000,0.397,7.37e-06,6,0.0808,-10.623,1,   ,0.0403,198.058,0.311
Theresa Thompson,2024-03-02 17:10:04.414+0000,0.407,Chica Brasileña (Remix),2023-12-22,"El Super Hobby, Tomi Narbondo,
↪  Laguna",0.714,156193,0.902,0,9,0.12,-5.05,0,Chica Brasileña - Remix,0.0718,96.925,0.919
Theresa Thompson,2024-03-02 17:10:14.494+0000,0.152,Hautajaiset,2024-01-04,"Gettomasa,
↪  Sexmane",0.666,203430,0.705,0,10,0.111,-3.245,0,Hautajaiset,0.0515,174.145,0.706
Theresa Thompson,2024-03-02 17:10:24.604+0000,0.912,Papa Meri Jaan (From \"ANIMAL\"),2023-11-14,"Sonu Nigam, Harshavardhan Rameshwar, Raj
↪  Shekhar",0.247,321549,0.197,0.000723,4,0.123,-11.504,0,Papa Meri Jaan (From \"ANIMAL\"),0.0417,57.18,0.209
Theresa Thompson,2024-03-02 17:10:34.690+0000,0.416,Citi+,2023-12-18,YANGHONGWON,0.818,151666,0.463,0,0,0.107,-8.572,1,Citi+,0.0433,103.977,0.382
Diane Rivas DVM,2024-03-02 17:09:44.196+0000,0.873,Kau Rumahku,2022-01-10,raissa anggiani,0.35,275250,0.323,1.48e-06,1,0.103,-8.982,1,Kau
↪  Rumahku,0.0369,169.513,0.278
Diane Rivas DVM,2024-03-02 17:09:54.311+0000,0.32,nadie sabe lo que va a pasar mañana,2023-10-13,Bad Bunny,0.914,225654,0.698,8.76e-06,1,0.083,-5.228,1,NO ME
↪  QUIERO CASAR,0.115,99.011,0.757
Diane Rivas DVM,2024-03-02 17:10:04.419+0000,0.587,99%,2023-03-02,"RPT MCK, Trung Trần",0.598,156303,0.489,0,11,0.119,-8.884,0,Chìm Sâu,0.0486,204.179,0.502
Diane Rivas DVM,2024-03-02 17:10:14.516+0000,0.0809,TÄYDELLINEN AJOITUS,2023-05-05,"SHRTY, Joalin",0.819,154913,0.733,2.6e-06,8,0.166,-4.164,1,Pariisin
↪  kevät,0.0351,115.972,0.587
Diane Rivas DVM,2024-03-02 17:10:24.614+0000,0.0141,BARBATULA BARBATULA,2023-12-15,"gyuris, Grasa",0.691,177692,0.651,0,1,0.142,-7.543,0,10 PINK
↪  BB,0.346,155.925,0.326
Diane Rivas DVM,2024-03-02 17:10:34.708+0000,0.22,   ,2023-04-20,Panther Chan,0.713,216000,0.896,0,6,0.0859,-3.37,0,   ,0.0502,115.004,0.789
Daniel Schultz,2024-03-02 17:20:56.453+0000,0.353,Chambre 140 (Part.1),2024-01-18,PLK,0.902,144853,0.539,3.6e-05,4,0.0967,-8.891,0,Flash,0.384,137.021,0.455
Daniel Schultz,2024-03-02 17:21:06.513+0000,0.166,Coast to Costi,2023-11-16,Costi,0.806,160002,0.615,0.00504,8,0.496,-6.636,0,DNA,0.0479,119.988,0.425
Daniel Schultz,2024-03-02 17:21:16.561+0000,0.0537,Rúntarinn,2023-08-03,Steindi Jr.,0.759,165812,0.775,1.79e-06,7,0.141,-5.957,1,Rúntarinn,0.16,154.87399,0.596
Daniel Schultz,2024-03-02 17:21:26.618+0000,0.175,B.,2023-09-01,Anıl Emre Daldal,0.479,287226,0.515,0.0162,11,0.135,-9.84,0,B.,0.0349,139.942,0.233
Daniel Schultz,2024-03-02 17:21:36.673+0000,0.073,Nah,2023-06-12,Anson Lo   ,0.644,233188,0.855,0,5,0.132,-3.426,0,Nah,0.248,155.92999,0.67
Daniel Schultz,2024-03-02 17:21:46.740+0000,0.83,Daylight,2023-04-14,David Kushner,0.508,212953,0.43,0.000441,2,0.093,-9.475,0,Daylight,0.0335,130.09,0.324
Suzanne James,2024-03-02 17:09:44.200+0000,0.797  ,        (By Tamar Yahalomy & Yonatan Kalimi),2023-01-25,"Avi Aburomi,
↪  Mor",0.522,184675,0.428,0,6,0.11,-7.764,1  ,        (By Tamar Yahalomy & Yonatan Kalimi),0.0358,154.088,0.368
Suzanne James,2024-03-02 17:09:54.322+0000,0.265,One Thing At A Time,2023-03-03,Morgan Wallen,0.732,157477,0.839,0,9,0.602,-5.007,1,You
↪  Proof,0.0345,119.724,0.629
Suzanne James,2024-03-02 17:10:04.429+0000,0.0437,Ladrón de Amor,2020-04-24,Don Medardo y sus Players Mauricio
↪  Luzuriaga,0.659,203600,0.787,0.000462,7,0.329,-5.934,1,Ladrón de Amor,0.05,109.981,0.961
```

A.1 Queries

Listing 3: Query 1

```
cqlsh:spotify>
SELECT * FROM records WHERE name = 'Phevos Margonis';
```

```
 name           | timestamp                       | acousticness | album_name                          | album_release_date | artists                                    | danceability | durat
----------------+---------------------------------+--------------+-------------------------------------+--------------------+--------------------------------------------+--------------+------
 Phevos Margonis | 2024-03-02 16:49:03.546000+0000 |       0.0877 |                       AA LANGUAGE 2 |         2023-10-27 |                 Aarne, MAYOT, Huzzy Buzzy |        0.636 |
 Phevos Margonis | 2024-03-02 16:49:13.604000+0000 |      0.00824 |    Dubula (feat. DJ Latimmy) [Remake] |         2023-11-03 |   Harrycane, Master KG, Eemoh, DJ Latimmy |        0.765 |
 Phevos Margonis | 2024-03-02 16:49:23.660000+0000 |        0.354 |   Sinterklaasliedjes om mee te zingen |         2020-07-28 | Alles Kids, Sinterklaasliedjes Alles Kids |        0.955 |
 Phevos Margonis | 2024-03-02 16:49:33.750000+0000 |        0.859 |                               ROSAS |         2022-05-11 |                           Kappa Jotta, MUN |          0.8 |
 Phevos Margonis | 2024-03-02 16:49:43.859000+0000 |        0.315 |                            Shall We |         2023-08-04 |              Percy, 4ourYou, GENA DESOUZA |         0.57 |
 Phevos Margonis | 2024-03-02 16:49:53.946000+0000 |       0.0179 |          By the Way (Deluxe Edition) |         2002-07-09 |                     Red Hot Chili Peppers |        0.618 |
 Phevos Margonis | 2024-03-02 17:03:23.935000+0000 |       0.0455 |         Mitä se maksaa (feat. TUULI) |         2023-12-08 |                       Alina Burnet, TUULI |        0.784 |
 Phevos Margonis | 2024-03-02 17:03:34.348000+0000 |        0.492 |                               Tenet |         2023-05-26 |                                 DJ Louder |        0.643 |
 Phevos Margonis | 2024-03-02 17:03:44.471000+0000 |         0.82 |                       FRXV (Ao Vivo) |         2023-12-15 |                                Filipe Ret |        0.436 |
 Phevos Margonis | 2024-03-02 17:03:54.618000+0000 |        0.808 |        ECLIPSE (Original Soundtrack) |         2023-11-13 |                             Motive, Pango |        0.464 |
 Phevos Margonis | 2024-03-02 17:04:04.755000+0000 |        0.588 |                                     |         2022-02-11 |                   Full Trunk, Jimbo J |        0.671 |      21
 Phevos Margonis | 2024-03-02 17:04:14.839000+0000 |        0.808 |        ECLIPSE (Original Soundtrack) |         2023-11-13 |                             Motive, Pango |        0.464 |
 Phevos Margonis | 2024-03-02 17:09:44.224000+0000 |       0.0406 |                            Ayrı Gitme |         2023-08-18 |                              Aleyna Tilki |        0.731 |
 Phevos Margonis | 2024-03-02 17:09:54.333000+0000 |        0.616 |                      Haciendo Lo Mío |         2022-07-15 |                 Luis Mexia, Grupo Firme |         0.74 |
 Phevos Margonis | 2024-03-02 17:10:04.433000+0000 |        0.727 |                                null |               null |                             Frank Sinatra |        0.512 |
 Phevos Margonis | 2024-03-02 17:10:14.526000+0000 |        0.384 |                 La Divina Commedia |         2023-06-02 |                                    Tedua |        0.616 |
 Phevos Margonis | 2024-03-02 17:10:24.631000+0000 |        0.048 | M' Manc (con Geolier & Sfera Ebbasta) |         2020-06-11 |          Shablo, Geolier, Sfera Ebbasta |        0.703 |
 Phevos Margonis | 2024-03-02 17:10:34.730000+0000 |        0.111 |                           GIPSY TRAP |         2023-11-01 |                 Lava, ROPEX LATERIO |         0.82 |
 Phevos Margonis | 2024-03-02 17:20:56.459000+0000 |        0.126 |                      Bun Venit Acasa |         2023-09-15 |                 Mgk666, Nutu, ROBERTO |        0.765 |
 Phevos Margonis | 2024-03-02 17:21:06.522000+0000 |        0.312 |                             Khayaal |         2023-03-16 |                       Talwiinder, NDS |        0.626 |
 Phevos Margonis | 2024-03-02 17:21:16.572000+0000 |       0.0976 |          1989 (Taylor's Version) |         2023-10-26 |                          Taylor Swift |        0.737 |
 Phevos Margonis | 2024-03-02 17:21:26.626000+0000 |        0.018 |             Hunting High and Low |         1985-06-01 |                                 a-ha |        0.573 |
 Phevos Margonis | 2024-03-02 17:21:36.682000+0000 |        0.256 |                   Sister Bethina |         2006-01-01 |                              Mgarimbe |        0.683 |
 Phevos Margonis | 2024-03-02 17:21:46.749000+0000 |        0.607 |              x (Wembley Edition) |         2013-01-01 |                             Ed Sheeran |        0.614 |

(24 rows)
```

Listing 4: Query 2

```
cqlsh:spotify>
SELECT * FROM spotify.records
WHERE name = 'Phevos Margonis'
AND timestamp >= '2024-03-02 16:49:00.000000+0000'
AND timestamp < '2024-03-02 16:50:00.000000+0000';
```

```
 name           | timestamp                       | acousticness | album_name                          | album_release_date | artists                                    | danceability | duratio
----------------+---------------------------------+--------------+-------------------------------------+--------------------+--------------------------------------------+--------------+--------
 Phevos Margonis | 2024-03-02 16:49:03.546000+0000 |       0.0877 |                       AA LANGUAGE 2 |         2023-10-27 |                 Aarne, MAYOT, Huzzy Buzzy |        0.636 |     15
 Phevos Margonis | 2024-03-02 16:49:13.604000+0000 |      0.00824 |    Dubula (feat. DJ Latimmy) [Remake] |         2023-11-03 |   Harrycane, Master KG, Eemoh, DJ Latimmy |        0.765 |     30
 Phevos Margonis | 2024-03-02 16:49:23.660000+0000 |        0.354 |   Sinterklaasliedjes om mee te zingen |         2020-07-28 | Alles Kids, Sinterklaasliedjes Alles Kids |        0.955 |      4
 Phevos Margonis | 2024-03-02 16:49:33.750000+0000 |        0.859 |                               ROSAS |         2022-05-11 |                           Kappa Jotta, MUN |          0.8 |     17
 Phevos Margonis | 2024-03-02 16:49:43.859000+0000 |        0.315 |                            Shall We |         2023-08-04 |              Percy, 4ourYou, GENA DESOUZA |         0.57 |     19
 Phevos Margonis | 2024-03-02 16:49:53.946000+0000 |       0.0179 |          By the Way (Deluxe Edition) |         2002-07-09 |                     Red Hot Chili Peppers |        0.618 |     26
```

Listing 5: Query 3

```
cqlsh:spotify>
SELECT avg(danceability)
FROM spotify.records
WHERE name = 'Phevos Margonis'
AND timestamp >= '2024-03-02 16:49:00.000000+0000'
AND timestamp < '2024-03-02 16:50:00.000000+0000';

 system.avg(danceability)
--------------------------
                    0.724

(1 rows)
```

B Appendix - Python Scripts

Listing 6: Python script to stream data to a Kafka-Broker.

```
import json
import asyncio
from aiokafka import AIOKafkaProducer
import pandas as pd
import time
from datetime import datetime, timezone
from faker import Faker

# Create a Faker instance
fake = Faker()
topic = 'test'
# Read Songs table
df = pd.read_csv('spotify-songs.csv')
df.dropna(subset=['name'], inplace=True)

# Create a list of users listening to Spotify
users = [fake.name() for _ in range(10)]
users.append('Phevos Margonis')

def serializer(value):
    '''Convert data to JSON before streaming'''
    return json.dumps(value).encode()

async def produce():
    '''Simulate data stream'''
    global users

    producer = AIOKafkaProducer(
        bootstrap_servers='localhost:29092',
        value_serializer=serializer,
        compression_type="gzip")

    # Start the streaming to Kafka
    await producer.start()

    # Stream simulated data
    for _,name in enumerate(users):
        # song = df[df['name'].str.contains('From
        ↪  "')]['name'].sample(n=1).iloc[0] # DEBUG escaped ""
        song = df.name.sample().iloc[0]
        now_utc = datetime.now(timezone.utc)
        timestamp = int(now_utc.timestamp() * 1000)
        data = {"name": name, "song": song, "timestamp": timestamp} # Create a
        ↪  data entry
        await producer.send(topic, data)

    await producer.stop()

loop = asyncio.get_event_loop()

# Send 10 batches, one every 10 seconds, for each user
for _ in range(10):
    result = loop.run_until_complete(produce())
    time.sleep(10)
```

Listing 7: Receive, Processes and Persists the messages read from Kafka-Broker.

```
from pyspark.sql import SparkSession
from pyspark.sql.types import (
    StructType,
    StructField,
    LongType,
    IntegerType,
    FloatType,
    StringType,
)
from pyspark.sql.functions import split, from_json, col

# Schema to Reconstruct String from JSON
songSchema = StructType(
    [
        StructField("name", StringType(), False),
        StructField("timestamp", LongType(), False),
        StructField("song", StringType(), False),
    ]
)

# Initialize Spark Session
spark = (
    SparkSession.builder.appName("SSKafka")
    .config("spark.jars.packages",
    ↪  "org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0")
    .getOrCreate()
)
spark.sparkContext.setLogLevel("ERROR")

# Read and Cache local CSV for Augmentation. Escape double quotes
localDF = (
    spark.read.option("header", True)
    .option("quote", '"')
    .option("escape", '"')
    .csv("spotify-songs.csv")
)
# Rename column name to song, for easier JOIN
localDF = localDF.withColumnRenamed("name", "song")
# Cash the result for repeated computations
localDF.cache()

# Read Data from Kafka Stream
df = (
    spark.readStream.format("kafka")
    .option("kafka.bootstrap.servers", "localhost:29092")
    .option("subscribe", "test")
    .option("startingOffsets", "latest")
    .load()
)
# Convert JSON to Sting
sdf = (
    df.selectExpr("CAST(value AS STRING)")
    .select(from_json(col("value"), songSchema).alias("data"))
    .select("data.*")
)
# JOIN Song metadata from localDF to Streamed df ON column 'song'
sdf = sdf.join(localDF, "song", "leftOuter")
```

```
def writeToCassandra(writeDF, _):
    ↪ writeDF.write.format("org.apache.spark.sql.cassandra").mode("append").options(
        table="records", keyspace="spotify"
    ).save()


result = None
while result is None:
    try:
        # connect
        result = (
            sdf.writeStream.option("spark.cassandra.connection.host",
            ↪ "localhost:9042")
            .foreachBatch(writeToCassandra)
            .outputMode("update")  # Because we are streaming
            .trigger(
                processingTime="30 seconds"
            )  # Process a batch of stream data every 30sec
            .option("checkpointLocation", "chk-point-dir")  # For Fault-Tolerance
            .start()
            .awaitTermination()
        )
    except:
        pass
```

Listing 8: Details about the Cassandra data model.

```
DROP KEYSPACE IF EXISTS spotify;

CREATE KEYSPACE IF NOT EXISTS spotify
WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};

CREATE TABLE IF NOT EXISTS spotify.records (
    name text,
    timestamp timestamp,
    song text,
    artists text,
    duration_ms text,
    album_name text,
    album_release_date date,
    danceability float,
    energy float,
    key tinyint,
    loudness float,
    mode tinyint,
    speechiness float,
    acousticness float,
    instrumentalness float,
    liveness float,
    valence float,
    tempo float,
    PRIMARY KEY ((name), timestamp)
) ;
```