Course: Large Scale Data Management

Professor: Panagiotis Liakos

Student name: Vasileios Ilias Drouzas

Student ID: f3352301

2nd Assignment

In this project, we will use the Apache Spark framework and the Apache Cassandra NoSQL database. The scope is to create a Structured Streaming Spark process that consumes Kafka messages and uses Cassandra as a sink to persist information.

PART I:

Here we will create using Python and Kafka a stream of songs listened by a group of people. The code for this part is on "*kafka_spotify_song_data.py*" and includes explanatory comments for most code segments. Some key aspects:

- We use the faker library to create names for 10 people.
- A random interval of up to 60 seconds is preserved and after that, a random person is picked and is offered with a song.
- We keep the person, the song (both as random choices, as we stated earlier) and the current time. We used the pytz library to get the Greek local time (to install, run *sudo apt-get install python3-tz*).
- The requested output is printed on the screen, but it is also saved in a file called 'spotify-song-data.txt', which preserves all the logs.

- The script is designed to be executed infinitely until the user requests termination. To terminate the program, use ‘Ctrl + C’ and it will terminate in at most 10 seconds.

Note: the file ‘*spotify-song-data.txt*’ is initially filled with some first logs from when I run the script. If you want to run the script, you may want to delete all the logs to get the new ones before running.

PART II:

Here, we will develop a pyspark script that receives, processes and persists the messages produced by the first part of the project. The code for this part is on ‘*cassandra-spark-streaming-example.py’* and includes explanatory comments for most code segments. Some key aspects:

- We define at first the *songSchema*, with columns ‘person’, ‘time’ and ‘song’, which were defined in the first part and the *csvSchema*, which consists of all the columns in the csv file.
- We start a Sparksession (*SSKafka*) with the necessary configurations and read the Kafka stream on the variable ‘*df*’.
- *sdf*: Selects the value column from the Kafka stream, converts it to a string, and then applies schema "songSchema" to parse the JSON data.
- After these, we read the CSV data and do a left join between the initial data and the CSV data on the name of the song and save into *joined_df*.
- We store the *csv_df* in cache to optimize query times. We cannot do the same for *joined_df*, because caching is not supported for streaming dataframes in Spark.
- We print the schema to the console and the streaming query logs.

- Finally, we write the data to Cassandra with the *writeToCassandra* function. The persistence happens every 30 seconds (as defined in the *trigger* part).

Before running the code, we need to make sure we have defined the records table in Cassandra. The commands to create the keyspace and the table are the following:

```
CREATE KEYSPACE spotify WITH replication = {'class':'SimpleStrategy', 'replication_factor' : 1};
CREATE TABLE IF NOT EXISTS records (
   id INT,
   person TEXT,
   time TIMESTAMP,
   name TEXT,
   artists TEXT,
   duration_ms BIGINT,
   album_name TEXT,
   album_release_date TEXT,
   danceability FLOAT,
   energy FLOAT,
   key INT,
   loudness FLOAT,
   mode INT,
   speechiness FLOAT,
   acousticness FLOAT,
   instrumentalness FLOAT,
   liveness FLOAT,
   valence FLOAT,
   tempo FLOAT,
   song TEXT,
   PRIMARY KEY (person, time)
);
```

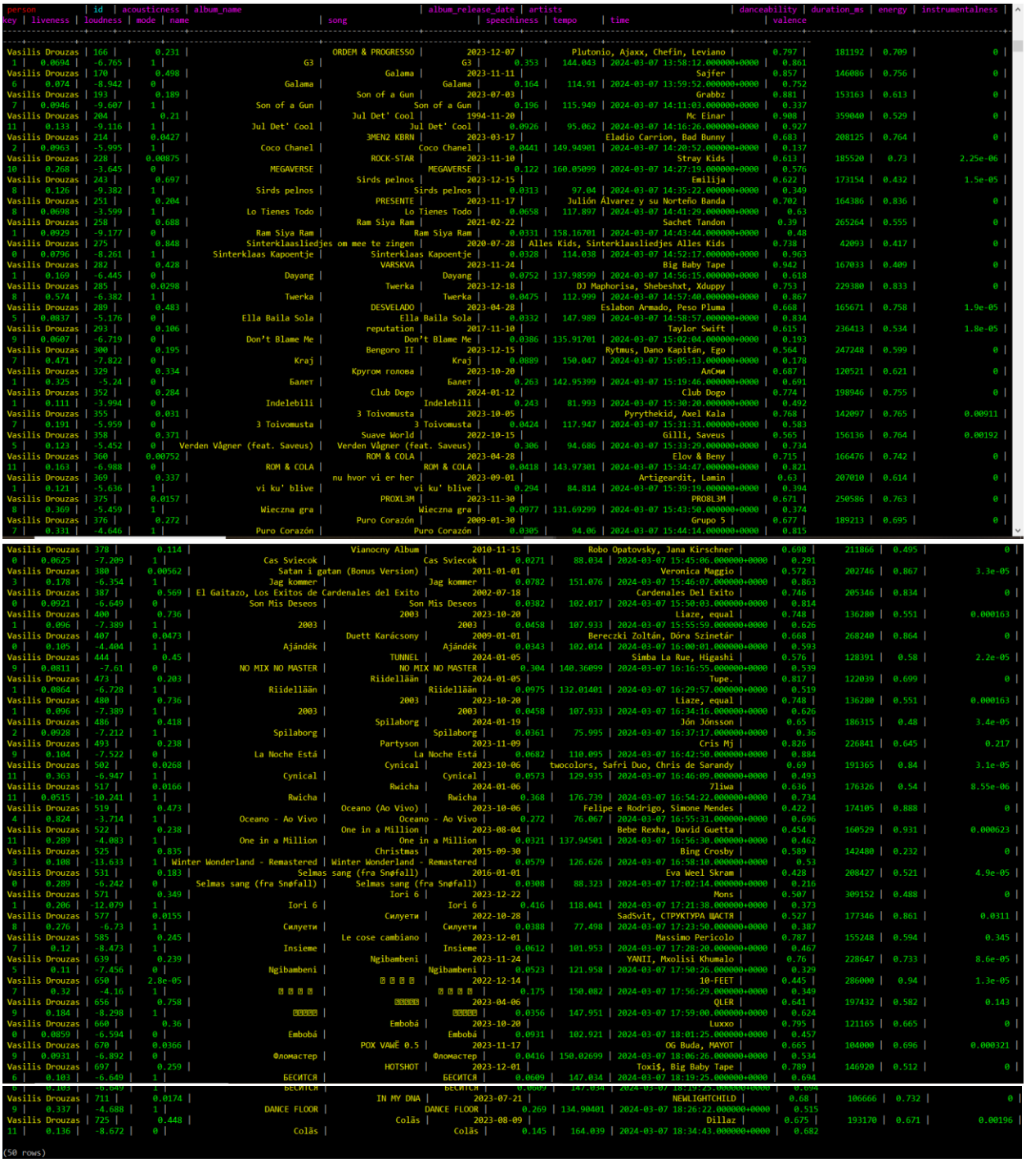
Regarding the Cassandra data model, remark that:

- Cassandra organizes data into keyspaces, where a keyspace holds related tables. In our case, we use a keyspace named ‘spotify’.
- Within the ‘spotify’ keyspace, we define a table called ‘records’.
- The table *records* consists of the columns ‘person’ (the person hearing the song), ‘song’, ‘time’ (timestamp when the person is hearing the song) and all the columns from the *spotify-songs.csv* (name, artists, duration_ms, album_name, album_release_date, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo).
- In Cassandra, schema evolution is handled by allowing columns to be added or modified over time without disrupting existing data. In our case, since Spark is writing to Cassandra in ‘append’ mode, new data is being added to the existing table.
- The queries are optimized for aggregations that a particular person listen on a specific time, since we include the time column as a clustering key and person as the partition key. This means that all rows with the same “person” are stored in the same partition. Within each partition, data will be sorted based on the “time” column. This way, Cassandra locates all the entries for every person quickly and efficiently.

Below we demonstrate a sample of 50 persisted lines of the Cassandra table (the outputs are also included in the report). Note that they are generated with the query

SELECT * FROM spotify.records LIMIT 50;

So we expect one person to be monopolizing our interest here, since we are partitioning the table by the *person* column and with the condition that he/she has appeared at least 50 times (which we expect to be true):



person	id	acousticness	album_name	album_release_date	artists	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	name	song	speechiness	tempo	time	valence
Vasilis Drouzas	166	0.231	ORDEM & PROGRESSO	2023-12-07	Plutonio, Ajaxx, Chefin, Leviano	0.797	181192	0.709	0	1	0.0694	-6.765	1	G3	G3	0.353	144.043	2024-03-07 13:58:12.000000+0000	0.861
Vasilis Drouzas	170	0.498	Galama	2023-11-11	Sajfer	0.857	146086	0.756	0	6	0.074	-8.942	0	Galama	Galama	0.164	114.91	2024-03-07 13:59:52.000000+0000	0.752
Vasilis Drouzas	193	0.189	Son of a Gun	2023-07-03	Grabbz	0.881	153163	0.613	0	7	0.0946	-9.607	1	Son of a Gun	Son of a Gun	0.196	115.949	2024-03-07 14:11:03.000000+0000	0.337
Vasilis Drouzas	204	0.21	Jul Det' Cool	1994-11-20	Mc Einar	0.908	359040	0.529	0	11	0.133	-9.116	1	Jul Det' Cool	Jul Det' Cool	0.0926	95.062	2024-03-07 14:16:26.000000+0000	0.927
Vasilis Drouzas	214	0.0427	3MEN2 KBRN	2023-03-17	Eladio Carrion, Bad Bunny	0.683	208125	0.764	0	2	0.0963	-5.995	1	Coco Chanel	Coco Chanel	0.0441	149.94901	2024-03-07 14:20:52.000000+0000	0.137
Vasilis Drouzas	228	0.00875	ROCK-STAR	2023-11-10	Stray Kids	0.613	185520	0.73	2.25e-06	10	0.268	-3.645	0	MEGAVERSE	MEGAVERSE	0.122	160.05099	2024-03-07 14:27:19.000000+0000	0.576
Vasilis Drouzas	243	0.697	Sirds pelnos	2023-12-15	Emilija	0.622	173154	0.432	1.5e-05	8	0.126	-9.382	1	Sirds pelnos	Sirds pelnos	0.0313	97.04	2024-03-07 14:35:22.000000+0000	0.349
Vasilis Drouzas	251	0.204	PRESENTE	2023-11-17	Julión Álvarez y su Norteño Banda	0.702	164386	0.836	0	8	0.0698	-3.599	1	Lo Tienes Todo	Lo Tienes Todo	0.0658	117.897	2024-03-07 14:41:29.000000+0000	0.63
Vasilis Drouzas	258	0.688	Ram Siya Ram	2021-02-22	Sachet Tandon	0.39	265264	0.555	0	1	0.0929	-9.177	0	Ram Siya Ram	Ram Siya Ram	0.0331	158.16701	2024-03-07 14:43:44.000000+0000	0.48
Vasilis Drouzas	275	0.848	Sinterklaasliedjes om mee te zingen	2020-07-28	Alles Kids, Sinterklaasliedjes Alles Kids	0.738	42093	0.417	0	0	0.0796	-8.261	1	Sinterklaas Kapoentje	Sinterklaas Kapoentje	0.0328	114.038	2024-03-07 14:52:17.000000+0000	0.963
Vasilis Drouzas	282	0.428	VARSKVA	2023-11-24	Big Baby Tape	0.942	167033	0.409	0	1	0.169	-6.445	0	Dayang	Dayang	0.0752	137.98599	2024-03-07 14:56:15.000000+0000	0.618
Vasilis Drouzas	285	0.0298	Twerka	2023-12-18	DJ Maphorisa, Shebeshxt, Xduppy	0.753	229380	0.833	0	8	0.574	-6.382	1	Twerka	Twerka	0.0475	112.999	2024-03-07 14:57:40.000000+0000	0.867
Vasilis Drouzas	289	0.483	DESVELADO	2023-04-28	Eslabon Armado, Peso Pluma	0.668	165671	0.758	1.9e-05	5	0.0837	-5.176	0	Ella Baila Sola	Ella Baila Sola	0.0332	147.989	2024-03-07 14:58:57.000000+0000	0.834
Vasilis Drouzas	293	0.106	reputation	2017-11-10	Taylor Swift	0.615	236413	0.534	1.8e-05	9	0.0607	-6.719	0	Don't Blame Me	Don't Blame Me	0.0386	135.91701	2024-03-07 15:02:04.000000+0000	0.193
Vasilis Drouzas	300	0.195	Bengoro II	2023-12-15	Rytmus, Dano Kapitán, Ego	0.564	247248	0.599	0	7	0.471	-7.822	0	Kraj	Kraj	0.0889	150.047	2024-03-07 15:05:13.000000+0000	0.178
Vasilis Drouzas	329	0.334	Крупом голова	2023-10-20	АлСми	0.687	120521	0.621	0	1	0.325	-5.24	0	Балет	Балет	0.263	142.95399	2024-03-07 15:19:46.000000+0000	0.691
Vasilis Drouzas	352	0.284	Club Dogo	2024-01-12	Club Dogo	0.774	198946	0.755	0	1	0.111	-3.994	0	Indelebili	Indelebili	0.243	81.993	2024-03-07 15:30:20.000000+0000	0.492
Vasilis Drouzas	355	0.031	3 Toivomusta	2023-10-05	Pyrythekid, Axel Kala	0.768	142097	0.765	0.00911	7	0.191	-5.959	0	3 Toivomusta	3 Toivomusta	0.0424	117.947	2024-03-07 15:31:31.000000+0000	0.583
Vasilis Drouzas	358	0.371	Suave World	2022-10-15	Gilli, Saveus	0.565	156136	0.764	0.00192	5	0.123	-5.452	0	Verden Vågner (feat. Saveus)	Verden Vågner (feat. Saveus)	0.306	94.686	2024-03-07 15:33:29.000000+0000	0.734
Vasilis Drouzas	360	0.00752	ROM & COLA	2023-04-28	Elov & Beny	0.715	166476	0.742	0	11	0.163	-6.988	0	ROM & COLA	ROM & COLA	0.0418	143.97301	2024-03-07 15:34:47.000000+0000	0.821
Vasilis Drouzas	369	0.337	nu hvor vi er her	2023-09-01	Artigeardit, Lamin	0.63	207010	0.614	0	1	0.121	-5.636	1	vi ku' blive	vi ku' blive	0.294	84.814	2024-03-07 15:39:19.000000+0000	0.394
Vasilis Drouzas	375	0.0157	PROXL3M	2023-11-30	PRO8L3M	0.671	250586	0.763	0	8	0.369	-5.459	1	Wieczna gra	Wieczna gra	0.0977	131.69299	2024-03-07 15:43:59.000000+0000	0.374
Vasilis Drouzas	376	0.272	Puro Corazón	2009-01-30	Grupo 5	0.677	189213	0.695	0	7	0.331	-4.646	1	Puro Corazón	Puro Corazón	0.0305	94.06	2024-03-07 15:44:14.000000+0000	0.815
Vasilis Drouzas	378	0.114	Vianocny Album	2010-11-15	Robo Opatovsky, Jana Kirschner	0.698	211866	0.495	0	0	0.0625	-7.209	1	Cas Sviecok	Cas Sviecok	0.0271	88.034	2024-03-07 15:45:06.000000+0000	0.291
Vasilis Drouzas	380	0.00562	Satan i gatan (Bonus Version)	2011-01-01	Veronica Maggio	0.572	202746	0.867	3.3e-05	3	0.178	-6.354	1	Jag kommer	Jag kommer	0.0782	151.076	2024-03-07 15:46:07.000000+0000	0.863
Vasilis Drouzas	387	0.569	El Gaitazo, Los Exitos de Cardenales del Exito	2002-07-18	Cardenales Del Exito	0.746	205346	0.834	0	0	0.0921	-6.649	0	Son Mis Deseos	Son Mis Deseos	0.0382	102.017	2024-03-07 15:50:03.000000+0000	0.814
Vasilis Drouzas	400	0.736	2003	2023-10-20	Liaze, equal	0.748	136280	0.551	0.000163	1	0.096	-7.389	1	2003	2003	0.0458	107.933	2024-03-07 15:55:59.000000+0000	0.626
Vasilis Drouzas	407	0.0473	Duett Karácsony	2009-01-01	Bereczki Zoltán, Dóra Szinetár	0.668	268240	0.864	0	0	0.105	-4.404	1	Ajándék	Ajándék	0.0343	102.014	2024-03-07 16:00:01.000000+0000	0.593
Vasilis Drouzas	444	0.45	TUNNEL	2024-01-05	Simba La Rue, Higashi	0.576	128391	0.58	2.2e-05	9	0.0811	-7.61	0	NO MIX NO MASTER	NO MIX NO MASTER	0.304	140.36099	2024-03-07 16:16:55.000000+0000	0.539
Vasilis Drouzas	473	0.203	Riidellään	2024-01-05	Tupe.	0.817	122039	0.699	0	1	0.0864	-6.728	1	Riidellään	Riidellään	0.0975	132.01401	2024-03-07 16:29:57.000000+0000	0.519
Vasilis Drouzas	480	0.736	2003	2023-10-20	Liaze, equal	0.748	136280	0.551	0.000163	1	0.096	-7.389	1	2003	2003	0.0458	107.933	2024-03-07 16:34:16.000000+0000	0.626
Vasilis Drouzas	486	0.418	Spilaborg	2024-01-19	Jón Jónsson	0.65	186315	0.48	3.4e-05	2	0.0928	-7.212	1	Spilaborg	Spilaborg	0.0361	75.995	2024-03-07 16:37:17.000000+0000	0.36
Vasilis Drouzas	493	0.238	Partyson	2023-11-09	Cris Mj	0.826	226841	0.645	0.217	9	0.104	-7.522	0	La Noche Está	La Noche Está	0.0682	110.095	2024-03-07 16:42:50.000000+0000	0.884
Vasilis Drouzas	502	0.0268	Cynical	2023-10-06	twocolors, Safri Duo, Chris de Sarandy	0.69	191365	0.84	3.1e-05	11	0.363	-6.947	1	Cynical	Cynical	0.0573	129.935	2024-03-07 16:46:09.000000+0000	0.493
Vasilis Drouzas	517	0.0166	Rwicha	2024-01-06	71iwa	0.636	176326	0.54	8.55e-06	11	0.0515	-10.241	1	Rwicha	Rwicha	0.368	176.739	2024-03-07 16:54:22.000000+0000	0.734
Vasilis Drouzas	519	0.473	Oceano (Ao Vivo)	2023-10-06	Felipe e Rodrigo, Simone Mendes	0.422	174105	0.888	0	4	0.824	-3.714	1	Oceano - Ao Vivo	Oceano - Ao Vivo	0.272	76.067	2024-03-07 16:55:31.000000+0000	0.696
Vasilis Drouzas	522	0.238	One in a Million	2023-08-04	Bebe Rexha, David Guetta	0.454	160529	0.931	0.000623	11	0.289	-4.083	1	One in a Million	One in a Million	0.0321	137.94501	2024-03-07 16:56:30.000000+0000	0.462
Vasilis Drouzas	525	0.835	Christmas	2015-09-30	Bing Crosby	0.589	142480	0.232	0	1	0.108	-13.633	1	Winter Wonderland - Remastered	Winter Wonderland - Remastered	0.0579	126.626	2024-03-07 16:58:10.000000+0000	0.53
Vasilis Drouzas	531	0.183	Selmas sang (fra Snøfall)	2016-01-01	Eva Weel Skram	0.428	208427	0.521	4.9e-05	0	0.289	-6.242	0	Selmas sang (fra Snøfall)	Selmas sang (fra Snøfall)	0.0308	88.323	2024-03-07 17:02:14.000000+0000	0.216
Vasilis Drouzas	571	0.349	Iori 6	2023-12-22	Mons	0.507	309152	0.488	0	1	0.206	-12.079	1	Iori 6	Iori 6	0.416	118.041	2024-03-07 17:21:38.000000+0000	0.373
Vasilis Drouzas	577	0.0155	Силуети	2022-10-28	SadSvit, СТРУКТУРА ШАСТЯ	0.527	177346	0.861	0.0311	8	0.276	-6.73	1	Силуети	Силуети	0.0388	77.498	2024-03-07 17:23:50.000000+0000	0.387
Vasilis Drouzas	585	0.245	Le cose cambiano	2023-12-01	Massimo Pericolo	0.787	155248	0.594	0.345	7	0.12	-8.473	1	Insieme	Insieme	0.0612	101.953	2024-03-07 17:28:20.000000+0000	0.467
Vasilis Drouzas	639	0.239	Ngibambeni	2023-11-24	YANII, Mxolisi Khumalo	0.76	228647	0.733	8.6e-05	5	0.11	-7.456	0	Ngibambeni	Ngibambeni	0.0523	121.958	2024-03-07 17:50:26.000000+0000	0.329
Vasilis Drouzas	650	2.8e-05	🅱 🅱 🅱 🅱	2022-12-14	10-FEET	0.445	286000	0.94	1.3e-05	7	0.32	-4.16	1	🅱 🅱 🅱 🅱	🅱 🅱 🅱 🅱	0.175	150.082	2024-03-07 17:56:29.000000+0000	0.349
Vasilis Drouzas	656	0.758	🅱🅱🅱🅱	2023-04-06	QLER	0.641	197432	0.582	0.143	9	0.184	-8.298	1	🅱🅱🅱🅱	🅱🅱🅱🅱	0.0356	147.951	2024-03-07 17:59:00.000000+0000	0.624
Vasilis Drouzas	660	0.36	Embobá	2023-10-20	Luxxo	0.795	121165	0.665	0	0	0.0859	-6.594	0	Embobá	Embobá	0.0931	102.921	2024-03-07 18:01:25.000000+0000	0.457
Vasilis Drouzas	670	0.0366	POX VAWË 0.5	2023-11-17	OG Buda, MAYOT	0.665	104000	0.696	0.000321	9	0.0931	-6.892	0	@ломастер	@ломастер	0.0416	150.02699	2024-03-07 18:06:26.000000+0000	0.534
Vasilis Drouzas	697	0.259	HOTSHOT	2023-12-01	Toxi$, Big Baby Tape	0.789	146920	0.512	0	6	0.103	-6.649	1	БЕСИТСЯ	БЕСИТСЯ	0.0609	147.034	2024-03-07 18:19:25.000000+0000	0.694
Vasilis Drouzas	711	0.0174	IN MY DNA	2023-07-21	NEWLIGHTCHILD	0.68	106666	0.732	0	9	0.337	-4.688	1	DANCE FLOOR	DANCE FLOOR	0.269	134.90401	2024-03-07 18:26:22.000000+0000	0.515
Vasilis Drouzas	725	0.448	Colãs	2023-08-09	Dillaz	0.675	193170	0.671	0.00196	11	0.136	-8.672	0	Colãs	Colãs	0.145	164.039	2024-03-07 18:34:43.000000+0000	0.682

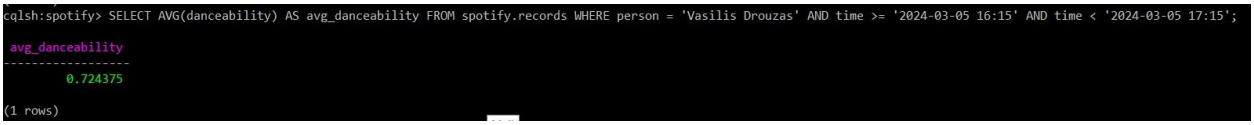
(50 rows)

Indeed, we confirm our predictions! Time to move on to more interesting queries…

Here we show two CQL queries and their results in the database about the author's own name and a particular hour. The first generates the average danceability of the songs that have been listened to during this hour, and the second generates the names of the songs.

1) Average danceability

```
SELECT AVG(danceability) AS avg_danceability
FROM spotify.records
WHERE person = 'Vasilis Drouzas'
AND time >= '2024-03-05 16:15'
AND time < '2024-03-05 17:15' ;
```



2) Names of the songs

```
SELECT name
FROM spotify.records
WHERE person = 'Vasilis Drouzas'
AND time >= '2024-03-05 16:15'
AND time < '2024-03-05 17:15' ;
```