

Programming Project II

Course: Large Scale Data Management

Student: Vasileios Dimopoulos

Date: 8/3/2024

Part 1 Python Script:

```
import json
import asyncio
import csv
import random
from datetime import datetime
from aiokafka import AIOKafkaProducer

from faker import Faker

# Create a Faker instance
fake = Faker()

topic = "test"

def serializer(value):
    return json.dumps(value).encode()

def read_random_songs(csv_file, num_songs):
    with open(csv_file, "r", newline="", encoding="utf-8") as file:
        reader = csv.reader(file)
        songs = list(reader)
        first_columns = [row[0] for row in songs]
        random_songs = random.sample(first_columns, num_songs)

    return random_songs

async def produce():
    producer = AIOKafkaProducer(
        bootstrap_servers="localhost:29092",
        value_serializer=serializer,
        compression_type="gzip",
    )

    csv_file_path = "spotify-songs.csv"
```

```

names = []
for i in range(10):
    name = fake.name()
    names.append(name)
names.append("Vasilis Dimopoulos")
k = 0
await producer.start()
for j in range(20):
    user_data = []
    for i in range(11):
        num_random_songs = 11
        songs = read_random_songs(csv_file_path, num_random_songs)
        current_time = datetime.now()
        formatted_time = current_time.strftime("%Y-%m-%d %H:%M:%S")
        data = {
            "id": k,
            "user_name": names[i],
            "song": songs[i],
            "time": formatted_time,
        }
        user_data.append(data)
        k += 1
    for data in user_data:
        await producer.send(topic, data)
        await asyncio.sleep(30)
    await producer.stop()

loop = asyncio.get_event_loop()
result = loop.run_until_complete(produce())

```

Part 2 Python Script:

```

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

songSchema = StructType(
    [

```

```

        StructField("id", IntegerType(), False),
        StructField("user_name", StringType(), False),
        StructField("song", StringType(), False),
        StructField("time", StringType(), False),
    ]
)

csvSchema = StructType(
    [
        StructField("song", StringType(), False),
        StructField("artists", StringType(), False),
        StructField("duration_ms", IntegerType(), False),
        StructField("album_name", StringType(), False),
        StructField("album_release_date", StringType(), False),
        StructField("danceability", FloatType(), False),
        StructField("energy", FloatType(), False),
        StructField("key", IntegerType(), False),
        StructField("loudness", FloatType(), False),
        StructField("mode", IntegerType(), False),
        StructField("speechiness", FloatType(), False),
        StructField("acousticness", FloatType(), False),
        StructField("instrumentalness", FloatType(), False),
        StructField("liveness", FloatType(), False),
        StructField("valence", FloatType(), False),
        StructField("tempo", FloatType(), False),
    ]
)

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")

df = (
    spark.readStream.format("kafka")
        .option("kafka.bootstrap.servers", "localhost:29092")
        .option("subscribe", "test")
        .option("startingOffsets", "latest")
        .load()
)

sdf = (
    df.selectExpr("CAST(value AS STRING)")
        .select(from_json(col("value"), songSchema).alias("data"))
        .select("data.*")
)

```

```

)

sdf = sdf.withColumn("year", col("time").cast("string").substr(1,
4).cast("int"))
sdf = sdf.withColumn("month", col("time").cast("string").substr(6,
2).cast("int"))
sdf = sdf.withColumn("day", col("time").cast("string").substr(9,
2).cast("int"))
sdf = sdf.withColumn("hour", col("time").cast("string").substr(12,
2).cast("int"))
sdf = sdf.withColumn("minute", col("time").cast("string").substr(15,
2).cast("int"))
sdf = sdf.withColumn("second", col("time").cast("string").substr(18,
2).cast("int"))
sdf = sdf.drop("time")

csv_df = spark.read.csv("spotify-songs.csv", header=True,
schema=csvSchema)
enriched_df = sdf.join(csv_df, "song", "left_outer")

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 = (
            enriched_df.writeStream.option(
                "spark.cassandra.connection.host", "localhost:9042"
            )
            .foreachBatch(writeToCassandra)
            .outputMode("update")
            .start()
            .awaitTermination()
        )
    except:
        pass

```

Details about your Cassandra data model

Schema:

```
CREATE TABLE spotify.records (  
  user_name text,  
  year int,  
  month int,  
  day int,  
  hour int,  
  song text,  
  acousticness float,  
  album_name text,  
  album_release_date text,  
  artists text,  
  danceability float,  
  duration_ms int,  
  energy float,  
  id int,  
  instrumentalness float,  
  key int,  
  liveness float,  
  loudness float,  
  minute int,  
  mode int,  
  second int,  
  speechiness float,  
  tempo float,  
  valence float,  
  PRIMARY KEY (user_name, year, month, day, hour, song)  
) WITH CLUSTERING ORDER BY (year ASC, month ASC, day ASC, hour ASC, song ASC)  
  AND additional_write_policy = '99p'  
  AND bloom_filter_fp_chance = 0.01  
  AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}  
  AND cdc = false  
  AND comment = ''  
  AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}  
  AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}  
  AND memtable = 'default'  
  AND crc_check_chance = 1.0  
  AND default_time_to_live = 0  
  AND extensions = {}  
  AND gc_grace_seconds = 864000  
  AND max_index_interval = 2048  
  AND memtable_flush_period_in_ms = 0  
  AND min_index_interval = 128  
  AND read_repair = 'BLOCKING'  
  AND speculative_retry = '99p';  
cqlsh> █
```

Since we only had only one Cassandra Block (Docker Container) we didn't set any partitioning keys. Just primary keys were enough for the requirements of the assignment's queries.

A sample of persisted lines (around 50) of the Cassandra table

```
user_name      | year | month | day | hour | song                      | acousticness |
album_name     |      |        |     |      | album_release_date | artists
| danceability | duration_ms | energy | id | instrumentalness | key | liveness | loudness |
minute | mode | second | speechiness | tempo | valence

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+

Stephanie Anderson | 2024 | 3 | 8 | 17 | 7 SÜNDEN | 0.387 |
7 SÜNDEN | 2024-01-07 | Tream, treamiboi
| 0.682 | 188038 | 0.829 | 67 |

0 | 1 | 0.165 | -5.932 | 30 | 0 | 11 | 0.363 | 168.15601 | 0.537

Stephanie Anderson | 2024 | 3 | 8 | 17 | El Cañonazo | 0.483 |
Super Éxitos Maracaibo 15 | 1997-09-01 |
Maracaibo 15 | 0.696 | 181226 | 0.843 | 34 |

0 | 1 | 0.101 | -9.929 | 27 | 1 | 57 | 0.0863 | 112.954 | 0.924

Stephanie Anderson | 2024 | 3 | 8 | 17 | El Viejo del Sombrerón | 0.275 |
Greatest Cumbia Classics Of Colombia, Vol. 1 | 2015-07-01 |
La Sonora Dinamita, La India Meliyará | 0.776 | 252186 | 0.791 | 45 |

1.6e-05 | 5 | 0.143 | -6.551 | 28 | 0 | 44 | 0.0415 | 107.954 | 0.961

Stephanie Anderson | 2024 | 3 | 8 | 17 | Feeling Myself - Roc Boyz Remix | 0.202 |
Feeling Myself (Roc Boyz Remix) | 2023-08-10 |
23, Roc Boyz | 0.801 | 182769 | 0.566 | 56 |

0 | 1 | 0.0999 | -7.022 | 29 | 0 | 29 | 0.234 | 130.09599 | 0.642

Stephanie Anderson | 2024 | 3 | 8 | 17 | Home To You (This Christmas) | 0.684 |
Home To You (This Christmas) | 2021-11-05 |
Sigrid | 0.616 | 225817 | 0.321 | 78 |

0 | 8 | 0.148 | -5.505 | 30 | 1 | 56 | 0.0313 | 139.233 | 0.287

Stephanie Anderson | 2024 | 3 | 8 | 17 | Oceaan | 0.866 |
The Singles Collection | 2013-03-15 |
Racoon | 0.593 | 164757 | 0.313 | 89 |

0 | 9 | 0.108 | -8.961 | 31 | 1 | 38 | 0.0383 | 95.904 | 0.219

Dr. Sean Gardner | 2024 | 3 | 8 | 17 | Back Die Bokke | 0.00572 |
Back Die Bokke | 2019-08-09 | Early B, Justin
Vega | 0.928 | 139876 | 0.818 | 41 |
```

0 11 0.0959 -4.121 28 0 6 0.0718 126.014 0.916	
Dr. Sean Gardner 2024 3 8 17 Carta al Universo 0.299	
Carta al Universo 2023-11-10	Alexa
Sotelo 0.763 142500 0.671 74	
0.000101 11 0.238 -5.469 30 1 21 0.164 100.124 0.368	
Dr. Sean Gardner 2024 3 8 17 Nonsense 0.0268	
emails i can't send 2022-07-15	Sabrina
Carpenter 0.74 163648 0.697 96	
0 8 0.224 -4.912 31 1 45 0.034 138.992 0.732	
Dr. Sean Gardner 2024 3 8 17 YO AK 0.125	
FERXXOCALIPSIS 2023-12-01	Feid,
Jhay P 0.778 205066 0.705 52	
1.26e-06 1 0.0811 -4.565 28 0 54 0.0649 90.964 0.591	
Dr. Sean Gardner 2024 3 8 17 ZAZA PART. 2 0.14	
24 2023-12-22	La Fève 0.711
153468 0.461 63	
1e-06 4 0.191 -11.041 29 0 36 0.0451 127.015 0.0843	
Dr. Sean Gardner 2024 3 8 17 kel kel :) 0.149	
AYAULYM 2023-11-22	Ayau
0.932 197500 0.588 85	
0.571 4 0.112 -7.935 31 0 3 0.0603 121.988 0.813	
Andrea Bell 2024 3 8 17 El Super Junte Rkt 0.102	
El Super Junte Rkt 2023-11-30 Gusty dj, Salastkbron, Callejero Fino, Alejo Isakk, R	
Jota, L-Gante, Lolo OG, DobleP, Lauty Gram 0.849 383369 0.794 93	
0 1 0.0813 -2.595 31 0 42 0.127 105.002 0.567	
Andrea Bell 2024 3 8 17 Ferrari 0.0127	
Ferrari 2022-04-01	James Hype, Miggy Dela
Rosa 0.847 186661 0.69 82	
6e-05 1 0.0526 -7.877 30 0 59 0.0493 125.004 0.692	
Andrea Bell 2024 3 8 17 Frate 0.121	
Club Dogo 2024-01-12	Club Dogo
0.679 163621 0.831 38	
0 3 0.229 -3.989 28 1 3 0.312 74.861 0.787	
Andrea Bell 2024 3 8 17 Gugguvaktin 0.0623	
PBT 2023-05-05	PATRi!K
0.894 125000 0.776 49	
0.367 0 0.257 -8.252 28 0 50 0.14 150.01401 0.493	

Andrea Bell 2024 3 8 17	Skaala 0.0905	
Kävi Miten Kävi 2023-06-01		Isac Elliot
0.623 141735 0.729 60		
8.8e-05 7 0.0906 -7.659 29 0 33	0.0299 146.96899 0.766	
Andrea Bell 2024 3 8 17	Tata 0.771	
Afro Fado 2023-11-24		Slow J
0.571 159751 0.718 71		
0.0298 5 0.104 -4.836 30 1 16	0.425 92.023 0.491	
Laura Camacho 2024 3 8 17	BHAGWADHARI 0.282	
BHAGWADHARI 2022-04-04		Bucks
Boy 0.759 198233 0.501 53		
0 3 0.595 -5.675 28 0 55	0.165 77.518 0.405	
Laura Camacho 2024 3 8 17	JPN II 0.135	
Euphorie 2023-11-10		Furelise
0.779 134117 0.512 64		
0 10 0.419 -9.645 29 1 38	0.237 135.85699 0.343	
Laura Camacho 2024 3 8 17	Lost 0.0272	
channel ORANGE 2012-07-10		Frank
Ocean 0.913 234093 0.603 97		
0.000503 8 0.167 -4.892 31 1 46	0.226 123.061 0.497	
Laura Camacho 2024 3 8 17	Superjung 0.178	
Tage vor 2000 2023-12-21		
Pashanim 0.721 169586 0.448 86		
0.0881 7 0.108 -9.595 31 1 4	0.337 169.871 0.436	
Laura Camacho 2024 3 8 17	Turné 0.422	
Turné 2023-03-17		Bausa
0.547 174986 0.57 42		
9.1e-05 3 0.319 -7.673 28 0 8	0.0528 139.728 0.347	
Laura Camacho 2024 3 8 17	Пьяный дождь 0.578	
Малый повзрослел, Ч. 2 2017-10-27		
Max Korzh 0.704 195532 0.332 75		
0 9 0.101 -8.24 30 1 23	0.0525 91.091 0.352	
Jasmine Rice 2024 3 8 17	ERROR 403 0.0506	
ERROR 403 2023-11-24		Lu de la Tower,
Corona 0.512 149187 0.94 77		
0 11 0.306 -4.573 30 0 54	0.0624 174.013 0.827	

Jasmine Rice | 2024 | 3 | 8 | 17 | Even for a moment | 0.758 |
 Even for a moment | 2023-10-19 | Sung Si
 Kyung, Naul | 0.441 | 293160 | 0.455 | 66 |
 9e-05 | 2 | 0.233 | -6.608 | 30 | 1 | 10 | 0.0283 | 84.534 | 0.265
 Jasmine Rice | 2024 | 3 | 8 | 17 | Malja | 0.451 |
 Malja | 2022-07-29 | Rasmus Gozzi Finland, Levi-
 Äijä | 0.461 | 162093 | 0.671 | 33 |
 0 | 3 | 0.104 | -8.997 | 27 | 0 | 56 | 0.263 | 163.729 | 0.65
 Jasmine Rice | 2024 | 3 | 8 | 17 | Milzīte Ilzīte | 0.335 |
 Milzīte Ilzīte | 2022-04-11 | Kapteinis Reinis
 | 0.961 | 180010 | 0.473 | 44 |
 1.5e-05 | 7 | 0.0884 | -10.733 | 28 | 1 | 42 | 0.0581 | 102.011 | 0.964
 Jasmine Rice | 2024 | 3 | 8 | 17 | Toxic | 0.0249 |
 In The Zone | 2003-11-13 | Britney Spears
 | 0.774 | 198800 | 0.838 | 55 |
 0.025 | 5 | 0.242 | -3.914 | 29 | 0 | 28 | 0.114 | 143.03999 | 0.924
 Jasmine Rice | 2024 | 3 | 8 | 17 | smugryger | 0.162 |
 smugryger | 2023-03-31 | andreas odbjerg
 | 0.841 | 163254 | 0.541 | 88 |
 2.6e-05 | 8 | 0.0568 | -7.567 | 31 | 0 | 37 | 0.0425 | 140.039 | 0.813
 Vasilis Dimopoulos | 2024 | 3 | 8 | 17 | Abl Mawsalek | 0.749 |
 Abl Mawsalek | 2021-08-07 | Muslim -
 98 | 0.323 | 278000 | 0.438 | مُسْلِم |
 1.1e-06 | 4 | 0.106 | -10.903 | 31 | 0 | 47 | 0.0406 | 115.627 | 0.329
 Vasilis Dimopoulos | 2024 | 3 | 8 | 17 | FREESTYLE | 0.25 |
 FREESTYLE | 2023-11-07 | FLY LO |
 0.858 | 137400 | 0.622 | 76 |
 3e-05 | 10 | 0.102 | -11.018 | 30 | 0 | 23 | 0.197 | 100.003 | 0.686
 Vasilis Dimopoulos | 2024 | 3 | 8 | 17 | GMFU (w/ 6arelyhuman) | 0.00105 |
 GMFU (w/ 6arelyhuman) | 2023-07-26 |
 Odetari, 6arelyhuman | 0.735 | 127741 | 0.778 | 65 |
 0.025 | 11 | 0.0923 | -3.23 | 29 | 1 | 39 | 0.0487 | 126.091 | 0.278
 Vasilis Dimopoulos | 2024 | 3 | 8 | 17 | Goeie Man | 0.509 |
 Wonderland | 2023-11-16 | Kevin |
 0.672 | 149368 | 0.514 | 87 |
 8.2e-06 | 2 | 0.112 | -10.904 | 31 | 0 | 5 | 0.36 | 90.023 | 0.652

Vasilis Dimopoulos 2024 3 8 17	Važiuoju Kur Širdis 0.187
Verkiu Raudonų Vyšnių Sode 2024-01-15	
GJan 0.722 225000 0.615 54	
4.71e-06 2 0.369 -5.154 28 0 57	0.038 125.002 0.785
Vasilis Dimopoulos 2024 3 8 17	Ела, Хабиби 0.143
Ела, Хабиби 2023-07-01	Emilia, Galin
0.818 171400 0.79 43	
2.39e-06 11 0.239 -2.019 28 0 10	0.1 149.992 0.758
Vincent Anderson 2024 3 8 17	Evil Receive 0.361
Presido La Pluto 2023-11-10	
Shallipopi 0.845 231157 0.686 90	
0 10 0.127 -5.573 31 0 39	0.131 107.99 0.699
Vincent Anderson 2024 3 8 17	La ziguezon 0.488
Chic & Swell 1982-01-01	La Bottine
Souriante 0.896 222693 0.635 68	
0 1 0.103 -12.339 30 1 12	0.148 104.843 0.929
Vincent Anderson 2024 3 8 17	Latyo és haze 0.0732
Focus 2023-01-13	Kain, Ekhoë, Pogány Induló
0.823 153650 0.531 46	
0 1 0.128 -8.994 28 1 46	0.143 124.928 0.413
Vincent Anderson 2024 3 8 17	Meuda 0.508
Mélo 2022-05-27	Tiakola
0.724 152546 0.593 57	
0 9 0.196 -6.498 29 1 31	0.0766 141.901 0.624
Vincent Anderson 2024 3 8 17	Soy Feo Pero Rico 0.00927
De Amor y Vacilón 2014-07-13	La
Combo Tortuga 0.738 187506 0.847 35	
0 2 0.325 -4.408 27 1 59	0.0812 111.962 0.826
Vincent Anderson 2024 3 8 17	Ya No Somos Ni Seremos 0.437
Ya No Somos Ni Seremos 2022-02-18	
Christian Nodal 0.588 185722 0.452 79	
0 7 0.344 -4.75 30 1 56	0.0268 139.953 0.734
Lance Wilkinson 2024 3 8 17	Beast & Peace 0.408
Blessed 2023-06-30	Mohbad
0.551 145188 0.638 47	
1.77e-06 11 0.66 -12.126 28 1 47	0.305 85.443 0.378

Lance Wilkinson | 2024 | 3 | 8 | 17 | Dej bůh štěstí | 0.802 | Vánoční
koledy, Vol. 1 (Pokoj lidem dobré vůle) | 2013-11-04 |
Dětský sbor Camerata | 0.327 | 67666 | 0.243 | 91 |

0 | 5 | 0.384 | -10.076 | 31 | 1 | 40 | 0.0306 | 87.891 | 0.53

Lance Wilkinson | 2024 | 3 | 8 | 17 | Overdose | 0.0343 |
劇場 | 2023-12-20 | natori |
0.728 | 194106 | 0.696 | 80 |

2.2e-05 | 4 | 0.228 | -6.215 | 30 | 0 | 57 | 0.0324 | 118.027 | 0.836

Lance Wilkinson | 2024 | 3 | 8 | 17 | PARIWO | 0.262 |
Blessed | 2023-06-30 | Mohbad, Bella
Shmurda | 0.771 | 120439 | 0.636 | 58 |

0.000551 | 7 | 0.153 | -7.731 | 29 | 0 | 32 | 0.0632 | 106.997 | 0.764

Lance Wilkinson | 2024 | 3 | 8 | 17 | Take on Me - Slowed | 0.966 |
Take on Me (Slowed) | 2022-06-25 |
m3gan | 0.279 | 200288 | 0.0664 | 69 |

0.971 | 8 | 0.149 | -19.479 | 30 | 0 | 13 | 0.0425 | 118.089 | 0.2

Lance Wilkinson | 2024 | 3 | 8 | 17 | Фломастер | 0.0366 |
POX VAWĚ 0.5 | 2023-11-17 | OG Buda,
MAYOT | 0.665 | 104000 | 0.696 | 36 |

0.000321 | 9 | 0.0931 | -6.892 | 28 | 0 | 0 | 0.0416 | 150.02699 | 0.534

James Mills | 2024 | 3 | 8 | 17 | 33 | 0.0424 |
Back In Town | 2024-01-05 | KERZA |
0.86 | 154303 | 0.576 | 39 |

0.000136 | 0 | 0.176 | -5.655 | 28 | 0 | 4 | 0.0406 | 140.00301 | 0.739

James Mills | 2024 | 3 | 8 | 17 | Dying 2 Live | 0.0344 |
VARSKVA | 2023-11-24 | Big Baby Tape |
0.817 | 161684 | 0.839 | 72 |

0 | 0 | 0.1 | -4.871 | 30 | 1 | 18 | 0.0635 | 94.951 | 0.452

Two CQL queries and their results in your database about your own name and a particular hour that generate the average danceability of the songs that you've listened during this hour, and the names of the songs, respectively.

1.

```
cqlsh> SELECT AVG(danceability) AS average_danceability FROM spotify.records WHERE user_name = 'Vasilis Dimopoulos' AND year = 2024 AND month = 3 AND day = 8 AND hour = 17;

average_danceability
-----
0.707167
```

2.

```
cqlsh> SELECT song FROM spotify.records WHERE user_name = 'Vasilis Dimopoulos' AND year = 2024 AND month = 3 AND day = 8 AND hour = 17;

song
-----
Ab1 Mawsalek
FREESTYLE
GMFU (w/ Barelyhuman)
Goeie Man
Važiuoju Kur Širdis
Ела, Хабиби

(6 rows)
```