Project 4: Data Lake Modeling

Gabriel Lima Barros - 2020006531

Maria Luiza Leão Silva - 2020100953

```
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import when,col


spark = SparkSession.builder \
    .appName("spotify-datalake") \
    .config("spark.jars.packages", "io.delta:delta-core_2.13:2.0.0") \
    .config("spark.sql.extensions",
"io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog",
"org.apache.spark.sql.delta.catalog.DeltaCatalog") \
    .config("spark.executor.instances", "2") \
    .config("spark.executor.cores", "2") \
    .config("spark.executor.memory", "1024M") \
    .getOrCreate()

spark.sparkContext.setLogLevel("WARN")
sc = spark.sparkContext
```

Task 2: Data Pipeline

Ingest v2

```
! python3 /home/mariasilva/tp4/merge_new_info.py
```

```
/shared/sampled/playlists_v2.json /shared/sampled/tracks_v2.json
:: loading settings :: url = jar:file:/opt/spark-3.4.2-bin-hadoop3-
scala2.13/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/
ivysettings.xml
Ivy Default Cache set to: /home/mariasilva/.ivy2/cache
The jars for the packages stored in: /home/mariasilva/.ivy2/jars
io.delta#delta-core_2.13 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-
12b769b0-dda4-4fde-8d2b-5a8972dfeea8;1.0
      confs: [default]
      found io.delta#delta-core_2.13;2.0.0 in central
      found io.delta#delta-storage;2.0.0 in central
      found org.antlr#antlr4-runtime;4.8 in central
      found org.codehaus.jackson#jackson-core-asl;1.9.13 in central
```

```
:: resolution report :: resolve 376ms :: artifacts dl 8ms
      :: modules in use:
      io.delta#delta-core_2.13;2.0.0 from central in [default]
      io.delta#delta-storage;2.0.0 from central in [default]
      org.antlr#antlr4-runtime;4.8 from central in [default]
      org.codehaus.jackson#jackson-core-asl;1.9.13 from central in
[default]
      -----------------------------------------------------------------
----
      |                  |            modules            ||   artifacts   
|
      |       conf       | number| search|dwnlded|evicted|| number|
dwnlded|
      -----------------------------------------------------------------
----
      |      default     |   4   |   0   |   0   |   0   ||   4   |   0   
|
      -----------------------------------------------------------------
----
:: retrieving :: org.apache.spark#spark-submit-parent-12b769b0-dda4-
4fde-8d2b-5a8972dfeea8
      confs: [default]
      0 artifacts copied, 4 already retrieved (0kB/6ms)
25/02/04 22:52:28 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where
applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
25/02/04 22:52:29 WARN Utils: Service 'SparkUI' could not bind on port
4060. Attempting port 4061.
------------------------

Total: 49.458
------------------------
Gold layer: 10.297
------------------------
Silver layer: 25.898
------------------------
Bronze layer storage: 44.64 mb
Silver layer storage: 99.6 mb
Gold layer storage: 20.56 mb
Total storage: 164.79 mb
```

Apply the update to playlist 11992

```
# Modifing db
sample_playlist_v2 = '/shared/sampled/playlists_v2.json'
sample_tracks_v2 = '/shared/sampled/tracks_v2.json'
```

```
playlist_v2_df = spark.read.json(sample_playlist_v2)
playlist_v2_df = playlist_v2_df.withColumn(
    "name",
    when(playlist_v2_df["pid"] == 11992, "GYM WORKOUT")  # Novo nome
    .otherwise(playlist_v2_df["name"])  # Mantém os outros valores inalterados
).withColumn(
    "collaborative",
    when(playlist_v2_df["pid"] == 11992, "True")  # Atualiza para True
    .otherwise(playlist_v2_df["collaborative"])
)
playlist_v2_df.filter(playlist_v2_df["pid"] == 11992).show()
playlist_v2_df.write.format("json").mode("overwrite").save("/home/mariasilva/tp4/modified_playlists_v2.json")
```

```
+-------------+-----------+-----------+-----+
|collaborative|description|       name|  pid|
+-------------+-----------+-----------+-----+
|         True|       null|GYM WORKOUT|11992|
+-------------+-----------+-----------+-----+
```

```
! python3 /home/mariasilva/tp4/merge_new_info.py -p "/home/mariasilva/tp4/modified_playlists_v2.json" -t "/shared/sampled/tracks_v2.json"
```

```
/home/mariasilva/tp4/modified_playlists_v2.json
/shared/sampled/tracks_v2.json
:: loading settings :: url = jar:file:/opt/spark-3.4.2-bin-hadoop3-scala2.13/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/mariasilva/.ivy2/cache
The jars for the packages stored in: /home/mariasilva/.ivy2/jars
io.delta#delta-core_2.13 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-98330d67-bd16-4902-b0cb-524c29c6a0bb;1.0
      confs: [default]
      found io.delta#delta-core_2.13;2.0.0 in central
      found io.delta#delta-storage;2.0.0 in central
      found org.antlr#antlr4-runtime;4.8 in central
      found org.codehaus.jackson#jackson-core-asl;1.9.13 in central
:: resolution report :: resolve 182ms :: artifacts dl 7ms
      :: modules in use:
      io.delta#delta-core_2.13;2.0.0 from central in [default]
      io.delta#delta-storage;2.0.0 from central in [default]
      org.antlr#antlr4-runtime;4.8 from central in [default]
      org.codehaus.jackson#jackson-core-asl;1.9.13 from central in [default]
      ---------------------------------------------------------------------
      |                  |            modules            ||   artifacts
|
      |       conf       | number| search|dwnlded|evicted|| number|
dwnlded|
      -----------------------------------------------------------------
----
      |      default     |   4   |   0   |   0   |   0   ||   4   |   0
|
      -----------------------------------------------------------------
----
:: retrieving :: org.apache.spark#spark-submit-parent-98330d67-bd16-
4902-b0cb-524c29c6a0bb
      confs: [default]
      0 artifacts copied, 4 already retrieved (0kB/5ms)
25/02/04 23:01:13 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where
applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
25/02/04 23:01:14 WARN Utils: Service 'SparkUI' could not bind on port
4060. Attempting port 4061.
------------------------

Total: 28.583
------------------------
Gold layer: 5.565
------------------------
Silver layer: 13.670
------------------------
Bronze layer storage: 44.64 mb
Silver layer storage: 99.62 mb
Gold layer storage: 20.56 mb
Total storage: 164.81 mb
```

```python
# Caminho do arquivo Parquet
silver_playlist_path = "/home/mariasilva/datalake/silver"

# Lendo os dados existentes
df =
spark.read.format("parquet").load(f"{silver_playlist_path}/playlists")

print("Silver layer:")
df.filter(df["playlist_id"] == 11992).show()

gold_playlist_path = "/home/mariasilva/datalake/gold"

# Lendo os dados existentes
df =
spark.read.format("parquet").load(f"{gold_playlist_path}/playlists")
```

```
print("Gold layer:")
df.filter(df["playlist_id"] == 11992).show()
```

```
Silver layer:
+-----------+-----------+-----------+-------------+
|playlist_id|       name|description|collaborative|
+-----------+-----------+-----------+-------------+
|      11992|GYM WORKOUT|       null|         True|
+-----------+-----------+-----------+-------------+

Gold layer:
+-----------+----------+-----------+----------+-----------------
+-----------+-----------+-------------+
|playlist_id|num_tracks|num_artists|num_albums|total_duration_ms|
name|description|collaborative|
+-----------+----------+-----------+----------+-----------------
+-----------+-----------+-------------+
|      11992|        16|         16|        16|          3158997|GYM
WORKOUT|       null|         True|
+-----------+----------+-----------+----------+-----------------
+-----------+-----------+-------------+
```

Ingest v3

```
! python3 /home/mariasilva/tp4/merge_new_info.py -p
"/shared/sampled/playlists_v3.json" -t
"/shared/sampled/tracks_v3.json"
```

```
/shared/sampled/playlists_v3.json /shared/sampled/tracks_v3.json
:: loading settings :: url = jar:file:/opt/spark-3.4.2-bin-hadoop3-
scala2.13/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/
ivysettings.xml
Ivy Default Cache set to: /home/mariasilva/.ivy2/cache
The jars for the packages stored in: /home/mariasilva/.ivy2/jars
io.delta#delta-core_2.13 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-
29f6dc0e-cc16-4841-833d-061439d4b937;1.0
	confs: [default]
	found io.delta#delta-core_2.13;2.0.0 in central
	found io.delta#delta-storage;2.0.0 in central
	found org.antlr#antlr4-runtime;4.8 in central
	found org.codehaus.jackson#jackson-core-asl;1.9.13 in central
:: resolution report :: resolve 228ms :: artifacts dl 6ms
	:: modules in use:
	io.delta#delta-core_2.13;2.0.0 from central in [default]
	io.delta#delta-storage;2.0.0 from central in [default]
	org.antlr#antlr4-runtime;4.8 from central in [default]
	org.codehaus.jackson#jackson-core-asl;1.9.13 from central in
```

```
[default]
      -----------------------------------------------------------------
----
      |                  |            modules            ||   artifacts
|
      |       conf       | number| search|dwnlded|evicted|| number|
dwnlded|
      -----------------------------------------------------------------
----
      |      default     |   4   |   0   |   0   |   0   ||   4   |   0
|
      -----------------------------------------------------------------
----
:: retrieving :: org.apache.spark#spark-submit-parent-29f6dc0e-cc16-
4841-833d-061439d4b937
      confs: [default]
      0 artifacts copied, 4 already retrieved (0kB/6ms)
25/02/04 23:02:08 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where
applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
25/02/04 23:02:09 WARN Utils: Service 'SparkUI' could not bind on port
4060. Attempting port 4061.
------------------------

Total: 31.003
------------------------
Gold layer: 7.557
------------------------
Silver layer: 15.031
------------------------
Bronze layer storage: 44.64 mb
Silver layer storage: 124.02 mb
Gold layer storage: 24.74 mb
Total storage: 193.41 mb
```

Challenges and Limitations

While processing the new data samples, the following difficulties were encountered:

1. Incremental Updates:

Since Parquet does not support native row-level updates, modifying individual records requires reading and rewriting the entire dataset.

1. Duplicate Management:

Ensuring that duplicate records were not introduced while merging v2 and v3 data with v1 required careful handling of unique identifiers (track_uri, playlist_id).

1. Storage Efficiency:

While Parquet is more efficient than JSON, handling large-scale updates can still introduce overhead due to file rewrites.