Project 4: Data Lake Modeling

Gabriel Lima Barros - 2020006531

Maria Luiza Leão Silva - 2020100953

```
import pyspark
from pyspark.sql import SparkSession

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
```

```
:: 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-
bc7eac20-d982-48d6-b4ac-16774be9f93e;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 230ms :: 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-bc7eac20-d982-
48d6-b4ac-16774be9f93e
      confs: [default]
      0 artifacts copied, 4 already retrieved (0kB/6ms)
25/02/04 22:26:13 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 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4040. Attempting port 4041.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4041. Attempting port 4042.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4042. Attempting port 4043.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4043. Attempting port 4044.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4044. Attempting port 4045.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4045. Attempting port 4046.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4046. Attempting port 4047.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4047. Attempting port 4048.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4048. Attempting port 4049.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4049. Attempting port 4050.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4050. Attempting port 4051.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4051. Attempting port 4052.
25/02/04 22:26:14 WARN Utils: Service 'SparkUI' could not bind on port
4052. Attempting port 4053.
```

Task 1: Data Modeling

Running implementation

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

```
:: 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-
bcdba4b5-271e-4078-aea6-920d21c2a6ea;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 259ms :: 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-bcdba4b5-271e-
4078-aea6-920d21c2a6ea
     confs: [default]
     0 artifacts copied, 4 already retrieved (0kB/6ms)
25/02/04 22:38:56 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 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4040. Attempting port 4041.
```

```
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4041. Attempting port 4042.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4042. Attempting port 4043.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4043. Attempting port 4044.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4044. Attempting port 4045.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4045. Attempting port 4046.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4046. Attempting port 4047.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4047. Attempting port 4048.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4048. Attempting port 4049.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4049. Attempting port 4050.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4050. Attempting port 4051.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4051. Attempting port 4052.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4052. Attempting port 4053.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4053. Attempting port 4054.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4054. Attempting port 4055.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4055. Attempting port 4056.
25/02/04 22:38:57 WARN Utils: Service 'SparkUI' could not bind on port
4056. Attempting port 4057.
25/02/04 22:39:06 WARN MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
25/02/04 22:39:08 WARN MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
25/02/04 22:39:15 WARN MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
```

```
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:39:18 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:39:31 WARN MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
25/02/04 22:39:41 WARN MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
------------------------

Silver layer: Songs: 2.2422
------------------------
Silver layer: Albums: 2.8438
------------------------
Silver layer: Artists: 1.3319
------------------------
Silver layer: Playlists: 0.3993
------------------------
Silver layer: Playlists Tracks: 1.5118
------------------------
Gold layer: Playlists: 27.1944
------------------------
Gold layer: Playlists Tracks: 8.3291
------------------------
```

Task 1A: Tables for the Silver and Gold Layers

```python
# Paths
bronze_path = "/home/mariasilva/datalake/bronze"
silver_path = "/home/mariasilva/datalake/silver"
gold_path = "/home/mariasilva/datalake/gold"
# Read layers
bronze_playlists_df = spark.read.parquet(f"{bronze_path}/playlists")
bronze_tracks_df = spark.read.parquet(f"{bronze_path}/tracks")
silver_songs_df = spark.read.parquet(f"{silver_path}/songs")
silver_albums_df = spark.read.parquet(f"{silver_path}/albums")
silver_artists_df = spark.read.parquet(f"{silver_path}/artists")
silver_playlists_df = spark.read.parquet(f"{silver_path}/playlists")
silver_playlist_tracks_df =
spark.read.parquet(f"{silver_path}/playlist_tracks")
gold_playlists_df = spark.read.parquet(f"{gold_path}/playlists")
gold_playlist_tracks_df =
spark.read.parquet(f"{gold_path}/playlist_tracks")
```

Bronze layer

This layer contains raw data ingested directly from the source without transformations. The files are stored as-is, ensuring historical traceability.

Playlists

Raw playlist data from playlists_v1.json

```
bronze_playlists_df.printSchema()
```

```
root
 |-- collaborative: string (nullable = true)
 |-- description: string (nullable = true)
 |-- name: string (nullable = true)
 |-- pid: long (nullable = true)
```

```
bronze_playlists_df.show(10)
```

```
+-------------+-----------+------------------+----+
|collaborative|description|              name| pid|
+-------------+-----------+------------------+----+
|        false|       null|       Winter 2014|  26|
|        false|       null|            groovy|  29|
|        false|       null|             KILLA| 964|
|        false|       null|       Country mix|1677|
|        false|       null|            Disney|1806|
|        false|       null|         Beep Boop|2040|
|         true|       null| Spring Break 2015|2214|
|        false|       null|oldies but goodies|2250|
|        false|       null|           Why Not|2453|
|        false|       null|               idk|2509|
+-------------+-----------+------------------+----+
only showing top 10 rows
```

Tracks

tracks_bronze: Raw track data from tracks_v1.json

```
bronze_tracks_df.printSchema()
```

```
root
 |-- album_name: string (nullable = true)
 |-- album_uri: string (nullable = true)
 |-- artist_name: string (nullable = true)
 |-- artist_uri: string (nullable = true)
 |-- duration_ms: long (nullable = true)
 |-- pid: long (nullable = true)
 |-- pos: long (nullable = true)
 |-- track_name: string (nullable = true)
 |-- track_uri: string (nullable = true)
```

```
bronze_tracks_df.show(10)
```

```
+--------------------+--------------------+-----------
+--------------------+-----------+-----+---+--------------------
+--------------------+
|          album_name|           album_uri|artist_name|
artist_uri|duration_ms|  pid|pos|          track_name|
track_uri|
+--------------------+--------------------+-----------
+--------------------+-----------+-----+---+--------------------
+--------------------+
|       Teenage Dream|spotify:album:06S...| Katy Perry|
spotify:artist:6j...|     230527|14382| 53|Last Friday Night...|
spotify:track:3oH...|
|       Teenage Dream|spotify:album:06S...| Katy Perry|
spotify:artist:6j...|     233685|14382| 54|California Gurls ...|
spotify:track:3f7...|
|     The Eminem Show|spotify:album:1ft...|     Eminem|
spotify:artist:7d...|     297893|14382| 55|    'Till I Collapse|
spotify:track:6yr...|
|           Woah Stop|spotify:album:5WY...|       98kb|
spotify:artist:7f...|     150768|14382| 56|           Woah Stop|
spotify:track:0sc...|
|            Oxymoron|spotify:album:7Et...|ScHoolboy Q|
spotify:artist:5I...|     278066|14382| 57|              Studio|
spotify:track:29g...|
|       Blank Face LP|spotify:album:0Yb...|ScHoolboy Q|
spotify:artist:5I...|     285346|14382| 58|     Tookie Knows II|
spotify:track:3mV...|
|          Chill Bill|spotify:album:5qB...|  Rob $tone|
spotify:artist:2h...|     177184|14382| 59|          Chill Bill|
spotify:track:5uD...|
|    Gangsta Memorial|spotify:album:63e...|     Eazy-E|
spotify:artist:7B...|     332733|14382| 60|Real Muthaphuckki...|
spotify:track:53B...|
|            StéLouse|spotify:album:60B...|   StéLouse|
spotify:artist:6k...|     246146|14391|  1|        Been So Long|
spotify:track:5r7...|
|Chasing Colors (f...|spotify:album:4Lp...| Marshmello|
spotify:artist:64...|     195200|14391|  2|Chasing Colors (f...|
spotify:track:1Vx...|
+--------------------+--------------------+-----------
+--------------------+-----------+-----+---+--------------------
+--------------------+
only showing top 10 rows
```

Silver layer

The Silver layer restructures the raw data, ensuring consistency and efficiency for analytical queries.

Song information table

```
silver_songs_df.printSchema()

root
 |-- track_name: string (nullable = true)
 |-- track_uri: string (nullable = true)
 |-- duration_ms: long (nullable = true)
 |-- album_uri: string (nullable = true)
 |-- artist_uri: string (nullable = true)


silver_songs_df.show(10)

+--------------------+--------------------+-----------
+--------------------+--------------------+
|          track_name|           track_uri|duration_ms|
album_uri|          artist_uri|
+--------------------+--------------------+-----------
+--------------------+--------------------+
|Last Friday Night...|spotify:track:3oH...|     230527|
spotify:album:06S...|spotify:artist:6j...|
|California Gurls ...|spotify:track:3f7...|     233685|
spotify:album:06S...|spotify:artist:6j...|
|    'Till I Collapse|spotify:track:6yr...|     297893|
spotify:album:1ft...|spotify:artist:7d...|
|           Woah Stop|spotify:track:0sc...|     150768|
spotify:album:5WY...|spotify:artist:7f...|
|              Studio|spotify:track:29g...|     278066|
spotify:album:7Et...|spotify:artist:5I...|
|     Tookie Knows II|spotify:track:3mV...|     285346|
spotify:album:0Yb...|spotify:artist:5I...|
|          Chill Bill|spotify:track:5uD...|     177184|
spotify:album:5qB...|spotify:artist:2h...|
|Real Muthaphuckki...|spotify:track:53B...|     332733|
spotify:album:63e...|spotify:artist:7B...|
|        Been So Long|spotify:track:5r7...|     246146|
spotify:album:60B...|spotify:artist:6k...|
|Chasing Colors (f...|spotify:track:1Vx...|     195200|
spotify:album:4Lp...|spotify:artist:64...|
+--------------------+--------------------+-----------
+--------------------+--------------------+
only showing top 10 rows
```

Album information table

```
silver_albums_df.printSchema()

root
 |-- album_uri: string (nullable = true)
 |-- album_name: string (nullable = true)
 |-- artist_uri: string (nullable = true)


silver_albums_df.show(10)

+--------------------+--------------------+--------------------+
|           album_uri|          album_name|          artist_uri|
+--------------------+--------------------+--------------------+
|spotify:album:0P0...|   Lost On The River|spotify:artist:2o...|
|spotify:album:6mU...|       Back In Black|spotify:artist:71...|
|spotify:album:6zV...|         Bag Raiders|spotify:artist:6f...|
|spotify:album:6rl...|             Trouble|spotify:artist:0z...|
|spotify:album:2UL...|Mickey Mouse Oper...|spotify:artist:3c...|
|spotify:album:7vL...|    Folk Hop N' Roll|spotify:artist:3w...|
|spotify:album:12p...|Mac and Devin Go ...|spotify:artist:7h...|
|spotify:album:085...|          Jimmy Choo|spotify:artist:6P...|
|spotify:album:17j...|       March Madness|spotify:artist:1R...|
|spotify:album:6Dx...|Blake Shelton's B...|spotify:artist:1U...|
+--------------------+--------------------+--------------------+
only showing top 10 rows
```

Artist information table

```
silver_artists_df.printSchema()

root
 |-- artist_uri: string (nullable = true)
 |-- artist_name: string (nullable = true)


silver_artists_df.show(10)

+--------------------+-----------------+
|          artist_uri|      artist_name|
+--------------------+-----------------+
|spotify:artist:20...|        Dom Dolla|
|spotify:artist:0d...| Barenaked Ladies|
|spotify:artist:1a...|Chance The Rapper|
|spotify:artist:6p...|       Chase Rice|
|spotify:artist:1G...|  Christophe Beck|
|spotify:artist:1x...|     Welshly Arms|
|spotify:artist:51...|    The Perishers|
|spotify:artist:3p...|       Tim Legend|
|spotify:artist:7m...|    Flux Pavilion|
|spotify:artist:6x...|    Anna Kendrick|
+--------------------+-----------------+
only showing top 10 rows
```

Playlist information table

```
silver_playlists_df.printSchema()

root
 |-- playlist_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- collaborative: string (nullable = true)
 |-- description: string (nullable = true)


silver_playlists_df.show(10)

+-----------+------------------+-------------+-----------+
|playlist_id|              name|collaborative|description|
+-----------+------------------+-------------+-----------+
|         26|       Winter 2014|        false|       null|
|         29|            groovy|        false|       null|
|        964|             KILLA|        false|       null|
|       1677|       Country mix|        false|       null|
|       1806|            Disney|        false|       null|
|       2040|         Beep Boop|        false|       null|
|       2214| Spring Break 2015|         true|       null|
|       2250|oldies but goodies|        false|       null|
|       2453|           Why Not|        false|       null|
|       2509|               idk|        false|       null|
+-----------+------------------+-------------+-----------+
only showing top 10 rows
```

Playlist tracks information table

```
silver_playlist_tracks_df.printSchema()

root
 |-- playlist_id: long (nullable = true)
 |-- track_uri: string (nullable = true)
 |-- album_uri: string (nullable = true)
 |-- artist_uri: string (nullable = true)
 |-- pos: long (nullable = true)
 |-- duration_ms: long (nullable = true)


silver_playlist_tracks_df.show(10)

+-----------+--------------------+--------------------
+--------------------+---+-----------+
|playlist_id|           track_uri|           album_uri|
artist_uri|pos|duration_ms|
+-----------+--------------------+--------------------
+--------------------+---+-----------+
|      14382|spotify:track:3oH...|spotify:album:06S...|
spotify:artist:6j...| 53|     230527|
|      14382|spotify:track:3f7...|spotify:album:06S...|
spotify:artist:6j...| 54|     233685|
|      14382|spotify:track:6yr...|spotify:album:1ft...|
spotify:artist:7d...| 55|     297893|
|      14382|spotify:track:0sc...|spotify:album:5WY...|
spotify:artist:7f...| 56|     150768|
|      14382|spotify:track:29g...|spotify:album:7Et...|
spotify:artist:5I...| 57|     278066|
|      14382|spotify:track:3mV...|spotify:album:0Yb...|
spotify:artist:5I...| 58|     285346|
|      14382|spotify:track:5uD...|spotify:album:5qB...|
spotify:artist:2h...| 59|     177184|
|      14382|spotify:track:53B...|spotify:album:63e...|
spotify:artist:7B...| 60|     332733|
|      14391|spotify:track:5r7...|spotify:album:60B...|
spotify:artist:6k...|  1|     246146|
|      14391|spotify:track:1Vx...|spotify:album:4Lp...|
spotify:artist:64...|  2|     195200|
+-----------+--------------------+--------------------
+--------------------+---+-----------+
only showing top 10 rows
```

Gold layer

The Gold layer consists of de-normalized tables optimized for reporting and analytical queries.

Playlist information aggregated table

```
gold_playlists_df.printSchema()
```

```
root
 |-- playlist_id: long (nullable = true)
 |-- num_tracks: long (nullable = true)
 |-- num_artists: long (nullable = true)
 |-- num_albums: long (nullable = true)
 |-- total_duration_ms: long (nullable = true)
 |-- name: string (nullable = true)
 |-- collaborative: string (nullable = true)
 |-- description: string (nullable = true)
```

```
gold_playlists_df.show()
```

```
+-----------+----------+-----------+----------+-----------------
+----------------+-------------+-----------+
|playlist_id|num_tracks|num_artists|num_albums|total_duration_ms|
name|collaborative|description|
+-----------+----------+-----------+----------+-----------------
+----------------+-------------+-----------+
|        592|        35|         20|        30|           9378900|
2015|        false|       null|
|       1250|         8|          8|         8|           1887520|
Electric|        false|       null|
|       2572|        12|         11|        12|           2697512|Once
Upon A Time|        false|       null|
|       4389|         9|          5|         8|           2886891|
Gospel songs|        false|       null|
|      11183|        29|         22|        28|           6059965|
□□|        false|       null|
|      11290|        52|         29|        44|          11259128|
country|        false|       null|
|      11823|        36|         28|        34|           7353941|
old country|        false|       null|
|      12100|        22|         22|        22|           4619810|
FrEe sPiRiT|        false|       null|
|      12548|         8|          6|         8|           2033908|
Tattoo |        false|       null|
|       1334|        28|         28|        28|           6926442|
my|        false|       null|
|       3756|        19|         18|        19|           4436744|
Reign|        false|       null|
|       4080|        25|         14|        21|           5742027|
explicit|        false|       null|
|       4315|        22|         21|        21|           6074952|
Pool party|        false|       null|
|      11058|        20|         12|        18|           5179770|
Sophomore|        false|       null|
|       3137|        27|         25|        26|           5999836|
Running|        false|       null|
|       1287|         9|          6|         9|           1947850|
Spanish Music|        false|       null|
|      12871|         9|          7|         9|           2950151|
Fml|        false|       null|
|      13015|        21|         17|        19|           5462260|
fiesta|        false|       null|
|       2755|        15|         11|        15|           3078553|
Country rock|        false|       null|
|      11275|         7|          6|         6|           1851721|
Yep|        false|       null|
+-----------+----------+-----------+----------+-----------------
+---------------+------------+----------+
only showing top 20 rows
```

Playlist tracks aggregated table

```
gold_playlist_tracks_df.printSchema()
```

```
root
 |-- playlist_id: long (nullable = true)
 |-- pos: long (nullable = true)
 |-- track_name: string (nullable = true)
 |-- album_name: string (nullable = true)
 |-- artist_name: string (nullable = true)


gold_playlist_tracks_df.show(10)

+-----------+---+--------------------+--------------------
+---------------+
|playlist_id|pos|          track_name|          album_name|
artist_name|
+-----------+---+--------------------+--------------------
+---------------+
|      41586|  5|    Missing Missouri|Original Album Cl...|     Sara
Evans|
|      14617|  9|           Come Thru|           Come Thru|
TYuS|
|     127174|  5|               Blame|             Rapture|
Tropics|
|     149739|  4|              Illume|       International| Lust For
Youth|
|      39711|  3|    Heart-Shaped Box|    Heart-Shaped Box|      Dead
Sara|
|      14245| 33| You Saved Me (Live)|Isla Vista Worshi...|     Ryan
Ellis|
|     162467| 10|The New National ...|    Selfish Machines|Pierce The
Veil|
|      13899| 38|       Girar O Mundo|     Buddha-Bar XVII|  Pattern
Drama|
|      13899| 38|       Girar O Mundo|     Buddha-Bar XVII|  Pattern
Drama|
|      13899| 38|       Girar O Mundo|     Buddha-Bar XVII|  Pattern
Drama|
+-----------+---+--------------------+--------------------
+---------------+
only showing top 10 rows
```

Data transformation

The following transformations occur between layers:

- Bronze → Silver:

1. Extract relevant columns

2. Remove duplicates

3. Normalize relationships (splitting playlists, tracks, albums, and artists into separate tables)

- Silver → Gold:

1. Compute summary statistics (total duration, number of tracks, albums, artists)

2. Join tables to create denormalized datasets for fast querying

Task 1B: Evaluate Parquet Performance

JSON is flexible but inefficient for large-scale data processing due to lack of compression and indexing. In this task, we evaluate the performance of JSON vs. Parquet for time efficiency.

```
# Running with json
! python3 /home/mariasilva/tp4/create_data_lake.py -f 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-
a20d8a37-f3e6-4922-859d-fa467ac611c9;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 184ms :: 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-a20d8a37-f3e6-
4922-859d-fa467ac611c9
```

```
      confs: [default]
      0 artifacts copied, 4 already retrieved (0kB/5ms)
25/02/04 22:29:44 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 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4040. Attempting port 4041.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4041. Attempting port 4042.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4042. Attempting port 4043.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4043. Attempting port 4044.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4044. Attempting port 4045.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4045. Attempting port 4046.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4046. Attempting port 4047.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4047. Attempting port 4048.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4048. Attempting port 4049.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4049. Attempting port 4050.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4050. Attempting port 4051.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4051. Attempting port 4052.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4052. Attempting port 4053.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4053. Attempting port 4054.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4054. Attempting port 4055.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4055. Attempting port 4056.
25/02/04 22:29:45 WARN Utils: Service 'SparkUI' could not bind on port
4056. Attempting port 4057.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
```

```
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:58 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
25/02/04 22:29:59 WARN RowBasedKeyValueBatch: Calling spill() on
RowBasedKeyValueBatch. Will not spill but return 0.
------------------------

Silver layer: Songs: 1.4743
------------------------
Silver layer: Albums: 1.8414
------------------------
Silver layer: Artists: 1.0657
------------------------
Silver layer: Playlists: 0.3566
------------------------
Silver layer: Playlists Tracks: 0.9414
------------------------
Gold layer: Playlists: 23.0245
------------------------
Gold layer: Playlists Tracks: 5.6793
------------------------
```

Results

-	Json	Parquet
Silver layer Songs	1.4743	1.3657
Silver layer Albums	1.8414	1.5082
Silver layer Artists	1.0657	0.7723
Silver layer Playlists	0.3566	0.2795
Silver layer Playlists Tracks	0.9414	0.9134
Gold layer Playlists	23.0245	20.6193
Gold layer Playlists Tracks	5.6793	4.8391

Conclusion

By implementing a structured data modeling approach in the Data Lake using the Medallion Architecture, we ensure efficient data retrieval while maintaining raw data integrity. The Silver and Gold layers provide structured, optimized data ready for analytical queries. Our performance evaluation confirms that Parquet is a superior format compared to JSON, offering significant improvements in storage and query performance.