Task 3: New Data Pipeline Using Delta Format

Gabriel Lima Barros - 2020006531

Maria Luiza Leão Silva - 2020100953

Creating Delta Format Datalake

```
! python3 /home/gabrielbarros/tp4/task3/create_delta_lake.py delta
```

```
delta
:: loading settings :: url = jar:file:/usr/local/lib/python3.11/dist-packages/pyspark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settir
Ivy Default Cache set to: /home/gabrielbarros/.ivy2/cache
The jars for the packages stored in: /home/gabrielbarros/.ivy2/jars
io.delta#delta-core_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-ad2bc82a-fb7d-4d6e-a291-f560630c2deb;1.0
        confs: [default]
        found io.delta#delta-core_2.12;2.4.0 in central
        found io.delta#delta-storage;2.4.0 in central
        found org.antlr#antlr4-runtime;4.9.3 in central
:: resolution report :: resolve 326ms :: artifacts dl 7ms
        :: modules in use:
        io.delta#delta-core_2.12;2.4.0 from central in [default]
        io.delta#delta-storage;2.4.0 from central in [default]
        org.antlr#antlr4-runtime;4.9.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-ad2bc82a-fb7d-4d6e-a291-f560630c2deb
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/5ms)
25/02/05 23:11:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
25/02/05 23:11:27 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
25/02/05 23:11:39 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted
25/02/05 23:11:42 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
25/02/05 23:11:46 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
25/02/05 23:11:58 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
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:02 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:04 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:12:17 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
25/02/05 23:12:33 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
Processing time for delta data lake: 76.44 seconds
```

Ingest V2

```
! python3 /home/gabrielbarros/tp4/task3/ingest_v_delta.py
```

```
:: loading settings :: url = jar:file:/usr/local/lib/python3.11/dist-packages/pyspark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settir
Ivy Default Cache set to: /home/gabrielbarros/.ivy2/cache
The jars for the packages stored in: /home/gabrielbarros/.ivy2/jars
io.delta#delta-core_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-5521bd06-f926-4f5d-83c9-01dce9f32084;1.0
        confs: [default]
        found io.delta#delta-core_2.12;2.4.0 in central
        found io.delta#delta-storage;2.4.0 in central
        found org.antlr#antlr4-runtime;4.9.3 in central
:: resolution report :: resolve 178ms :: artifacts dl 6ms
```

```
        :: modules in use:
        io.delta#delta-core_2.12;2.4.0 from central in [default]
        io.delta#delta-storage;2.4.0 from central in [default]
        org.antlr#antlr4-runtime;4.9.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-5521bd06-f926-4f5d-83c9-01dce9f32084
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/5ms)
25/02/05 23:12:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
25/02/05 23:12:44 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
25/02/05 23:12:59 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted
25/02/05 23:13:19 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 Update Time: 29.565 seconds
Gold Layer Update Time: 6.840 seconds
Bronze layer storage: 44.65 mb
Silver layer storage: 143.24 mb
Gold layer storage: 17.9 mb
Total storage: 205.78 mb
```

Parquet x Delta Ingest V2 time comparison

Format	Silver Layer (s)	Gold Layer (s)	Total Time (s)
Parquet	25.89	10.29	49.45
Delta	29.56	06.84	36.40

Delta vs. Parquet

Silver Layer Update:

- Delta Lake: 29.56 seconds
- Parquet: 25.89 seconds
- Observation: Delta Lake took slightly longer due to its transactional overhead and the use of the MERGE statement, which ensures data consistency.

Gold Layer Update:

- Delta Lake: 6.84 seconds
- Parquet: 10.29 seconds
- Observation: Delta Lake was faster because its optimized storage and indexing reduced read and write latencies.

Total Time:

- Delta Lake: 36.40 seconds
- Parquet: 49.45 seconds
- Observation: While Delta Lake's write operations had additional overhead in the Silver layer, the optimized reads for the Gold layer resulted in a significantly faster total runtime.

Parquet x Delta space comparison

Format	Bronze Layer (MB)	Silver Layer (MB)	Gold Layer (MB)	Total Storage (MB)
Parquet	44.64	124.02	24.74	193.41
Delta	44.65	143.25	17.09	205.78

Delta vs. Parquet

Bronze Layer:

- Delta Lake: 44.65 MB
- Parquet: 44.64 MB
- Observation: Minimal difference as both formats store raw data.

Silver Layer:

- Delta Lake: 143.25 MB

- Parquet: 124.02 MB
- Observation: Delta Lake requires additional storage for metadata and transaction logs, increasing the Silver layer size.

Gold Layer:

- Delta Lake: 17.09 MB
- Parquet: 24.74 MB
- Observation: Delta Lake's optimized file compaction reduced the storage required for aggregated data.

Total Storage:

- Delta Lake: 205.78 MB
- Parquet: 193.41 MB
- Observation: Delta Lake uses slightly more storage overall due to transaction logs and metadata but compensates with faster reads and write operations.

Update playlist

```
! python3 /home/gabrielbarros/tp4/task3/update_playlist_delta.py
```

```
:: loading settings :: url = jar:file:/usr/local/lib/python3.11/dist-packages/pyspark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settir
Ivy Default Cache set to: /home/gabrielbarros/.ivy2/cache
The jars for the packages stored in: /home/gabrielbarros/.ivy2/jars
io.delta#delta-core_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-c28641bd-60bb-46a3-af5d-4d4e89d602c7;1.0
        confs: [default]
        found io.delta#delta-core_2.12;2.4.0 in central
        found io.delta#delta-storage;2.4.0 in central
        found org.antlr#antlr4-runtime;4.9.3 in central
:: resolution report :: resolve 271ms :: artifacts dl 7ms
        :: modules in use:
        io.delta#delta-core_2.12;2.4.0 from central in [default]
        io.delta#delta-storage;2.4.0 from central in [default]
        org.antlr#antlr4-runtime;4.9.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-c28641bd-60bb-46a3-af5d-4d4e89d602c7
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/13ms)
25/02/05 23:13:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
25/02/05 23:13:35 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
[INFO] Updating datalake/silver Layer for playlist 11992...
25/02/05 23:13:43 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted
[INFO] datalake/silver Layer updated in 15.42 seconds.
[INFO] Updating datalake/gold Layer for playlist 11992...
[INFO] datalake/gold Layer updated in 4.74 seconds.
[INFO] Playlist 11992 update completed in 25.12 seconds.
```

Parquet x Delta playlist update time comparison

Format	Silver Layer (s)	Gold Layer (s)	Total Time (s)
Parquet	13.67	05.56	28.58
Delta	15.42	04.74	25.12

Performance Observations

Silver Layer Update:

- Parquet: 13.67 seconds
- Delta Lake: 15.42 seconds
- Observation: Delta Lake takes slightly longer in the Silver Layer due to its transactional overhead (e.g., managing metadata and maintaining ACID properties).

Gold Layer Update:

- Parquet: 5.56 seconds

- Delta Lake: 4.74 seconds
- Observation: Delta Lake outperforms Parquet in the Gold Layer due to its optimized file structure and indexing.

Total Time:

- Parquet: 28.58 seconds
- Delta Lake: 25.12 seconds
- Observation: Delta Lake completes the entire playlist update pipeline faster by ~12%.

Ingest V3

```
! python3 /home/gabrielbarros/tp4/task3/ingest_v_delta.py -p "/shared/sampled/playlists_v3.json" -t "/shared/sampled/tracks_v3.json"
```

```
:: loading settings :: url = jar:file:/usr/local/lib/python3.11/dist-packages/pyspark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settir
Ivy Default Cache set to: /home/gabrielbarros/.ivy2/cache
The jars for the packages stored in: /home/gabrielbarros/.ivy2/jars
io.delta#delta-core_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-7b7baaed-d086-44f2-9894-e5986ed6792d;1.0
        confs: [default]
        found io.delta#delta-core_2.12;2.4.0 in central
        found io.delta#delta-storage;2.4.0 in central
        found org.antlr#antlr4-runtime;4.9.3 in central
:: resolution report :: resolve 410ms :: artifacts dl 16ms
        :: modules in use:
        io.delta#delta-core_2.12;2.4.0 from central in [default]
        io.delta#delta-storage;2.4.0 from central in [default]
        org.antlr#antlr4-runtime;4.9.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-7b7baaed-d086-44f2-9894-e5986ed6792d
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/7ms)
25/02/05 23:14:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
25/02/05 23:14:02 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
25/02/05 23:14:17 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted
25/02/05 23:14:32 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:14:32 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
25/02/05 23:14:34 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 Update Time: 26.061 seconds
Gold Layer Update Time: 5.930 seconds
Bronze layer storage: 44.65 mb
Silver layer storage: 230.88 mb
Gold layer storage: 19.94 mb
Total storage: 295.47 mb
```

Parquet x Delta Ingest V3 time comparison

Format	Silver Layer (s)	Gold Layer (s)	Total Time (s)
Parquet	15.03	07.55	31.00
Delta	26.06	05.93	31.99

Silver Layer Update:

- Parquet: 15.03 seconds
- Delta Lake: 26.06 seconds
- Observation: Delta Lake takes longer due to overhead from transactional consistency and metadata management, which ensures reliable and accurate updates.

Gold Layer Update:

- Parquet: 7.55 seconds
- Delta Lake: 5.93 seconds
- Observation: Delta Lake outperforms Parquet in the Gold Layer due to optimized querying and file compaction strategies.

Total Time:

- Parquet: 31.00 seconds

- Delta Lake: 31.99 seconds
- Observation: The overall runtime difference is negligible (~3%), with Delta Lake slightly slower due to Silver Layer overhead.

Comece a programar ou gere código com IA.