1.prepare_for_incremental_load SQL ∨
File Edit View Run Help   Last edit was 7 minutes ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ∨   🗓 Schedule   Share   ∧

Cmd 1

Drop all the tables

Cmd 2

```sql
DROP DATABASE IF EXISTS f1_processed CASCADE;
```

OK

Command took 1.44 minutes -- by avinashchinta@my.unt.edu at 7/14/2023, 11:09:59 PM on dbkcourse_cluster

Cmd 3

```sql
CREATE DATABASE IF NOT EXISTS f1_processed
LOCATION "abfss://processed@erformula1dl.dfs.core.windows.net/";
```

OK

Command took 0.44 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:09:59 PM on dbkcourse_cluster

Cmd 4

```sql
DROP DATABASE IF EXISTS f1_presentation CASCADE;
```

OK

Command took 44.64 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:09:59 PM on dbkcourse_cluster

Cmd 5

SQL ▶▾ ∨ − ✕

```sql
CREATE DATABASE IF NOT EXISTS f1_presentation
LOCATION "abfss://presentation@erformula1dl.dfs.core.windows.net/";
```

OK

Command took 0.17 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:09:59 PM on dbkcourse_cluster

Cmd 6

New

Workspace

Repos

Recents

Data

Compute

Workflows

p_data_source
testing

p_file_date
2021-03-21

```
circuits_final_df = add_ingestion_date(circuits_renamed_df)
```

▸ 🗔 circuits_final_df: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string … 9 more fields]

Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:17:31 PM on dbkcourse_cluster

Cmd 18

Step 5 - Write data to datalake as parquet

Cmd 19

```
circuits_final_df.write.mode("overwrite").format("delta").saveAsTable("f1_processed.circuits")
```

▸ (6) Spark Jobs

Command took 3.81 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:17:31 PM on dbkcourse_cluster

Cmd 20

```sql
%sql
SELECT * FROM f1_processed.circuits;
```

▸ (2) Spark Jobs

▸ 🗔 _sqldf: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string … 9 more fields]

Table ⌄   +

	circuit_id	circuit_ref	name	location	country	latitude	longitude	altitude	data_source	file_date	ingestion_date
1	1	albert_park	Albert Park Grand Prix Circuit	Melbourne	Australia	-37.8497	144.968	10	testing	2021-03-21	2023-07-15T04:17:34.359+0000
2	2	sepang	Sepang International Circuit	Kuala Lumpur	Malaysia	2.76083	101.738	18	testing	2021-03-21	2023-07-15T04:17:34.359+0000
3	3	bahrain	Bahrain International Circuit	Sakhir	Bahrain	26.0325	50.5106	7	testing	2021-03-21	2023-07-15T04:17:34.359+0000
4	4	catalunya	Circuit de Barcelona-Catalunya	Montmeló	Spain	41.57	2.26111	109	testing	2021-03-21	2023-07-15T04:17:34.359+0000
5	5	istanbul	Istanbul Park	Istanbul	Turkey	40.9517	29.405	130	testing	2021-03-21	2023-07-15T04:17:34.359+0000
6	6	monaco	Circuit de Monaco	Monte-Carlo	Monaco	43.7347	7.42056	7	testing	2021-03-21	2023-07-15T04:17:34.359+0000
7	7	villeneuve	Circuit Gilles Villeneuve	Montreal	Canada	45.5	-73.5228	13	testing	2021-03-21	2023-07-15T04:17:34.359+0000

⬇ 77 rows | 0.79 seconds runtime   Refreshed now

ℹ SQL cell result stored as PySpark data frame _sqldf . Learn more

Marketplace NEW

Partner Connect

1/4 Tasks Completed

Enable new UI NEW

Menu options

processed ···
Container

×

Upload + Add Directory ↻ Refresh | ↶ Rename 🗑 Delete ⇄ Change tier ⟋ Acquire lease ⟋ Break lease 🗨 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed

Search blobs by prefix (case-sensitive)   ⬤◯ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐ 📁 circuits						-	···

Search
«

🔲 Overview
⟋ Diagnose and solve problems
👥 Access Control (IAM)

Settings

🔗 Shared access tokens
👥 Manage ACL
🔑 Access policy
📊 Properties
ⓘ Metadata

p_data_source

p_file_date

2021-03-21

Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:20:05 PM on dbkcourse_cluster

Cmd 16

Write the output to processed container in parquet format

Cmd 17

```
races_selected_df.write.mode("overwrite").partitionBy('race_year').format("delta").saveAsTable("f1_processed.races")
```

▶ (5) Spark Jobs

Command took 11.02 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:20:05 PM on dbkcourse_cluster

Cmd 18

```
%sql
SELECT * FROM f1_processed.races;
```

▶ (3) Spark Jobs

▶ ▤ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, race_year: integer ... 5 more fields]

Table ∨   +

	race_id	race_year	round	circuit_id	name	ingestion_date	race_timestamp
1	1053	2021	2	21	Emilia Romagna Grand Prix	2023-07-15T04:20:10.336+0000	2021-04-18T13:00:00.000+0000
2	1052	2021	1	3	Bahrain Grand Prix	2023-07-15T04:20:10.336+0000	2021-03-28T15:00:00.000+0000
3	1051	2021	21	1	Australian Grand Prix	2023-07-15T04:20:10.336+0000	2021-11-21T06:00:00.000+0000
4	1054	2021	3	20	TBC	2023-07-15T04:20:10.336+0000	null
5	1055	2021	4	4	Spanish Grand Prix	2023-07-15T04:20:10.336+0000	2021-05-09T13:00:00.000+0000
6	1056	2021	5	6	Monaco Grand Prix	2023-07-15T04:20:10.336+0000	2021-05-23T13:00:00.000+0000
7	1057	2021	6	73	Azerbaijan Grand Prix	2023-07-15T04:20:10.336+0000	2021-06-06T12:00:00.000+0000

⤓ 1,058 rows | 2.04 seconds runtime   Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 2.04 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:20:05 PM on dbkcourse_cluster

▶ Run all   ● dbkcourse_cluster ∨   📅 Schedule   Share ∧

p_data_source
[   ]

p_file_date
[2021-03-21   ]

Step 4 Write output to parquet file

Cmd 17

```python
constructor_final_df.write.mode("overwrite").format("delta").saveAsTable("f1_processed.constructors")
```

▶ (6) Spark Jobs

Command took 3.62 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:21:12 PM on dbkcourse_cluster

Cmd 18

```sql
%sql
SELECT * FROM f1_processed.constructors;
```

▶ (2) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [constructor_id: integer, constructor_ref: string … 5 more fields]

Table ∨ +

	constructor_id	constructor_ref	name	nationality	data_source	file_date	ingestion_date
1	1	mclaren	McLaren	British		2021-03-21	2023-07-15T04:21:16.212+0000
2	2	bmw_sauber	BMW Sauber	German		2021-03-21	2023-07-15T04:21:16.212+0000
3	3	williams	Williams	British		2021-03-21	2023-07-15T04:21:16.212+0000
4	4	renault	Renault	French		2021-03-21	2023-07-15T04:21:16.212+0000
5	5	toro_rosso	Toro Rosso	Italian		2021-03-21	2023-07-15T04:21:16.212+0000
6	6	ferrari	Ferrari	Italian		2021-03-21	2023-07-15T04:21:16.212+0000
7	7	toyota	Toyota	Japanese		2021-03-21	2023-07-15T04:21:16.212+0000

⬇ 211 rows | 0.74 seconds runtime   Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 0.74 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:21:12 PM on dbkcourse_cluster

4.ingest_drivers_file Python ✓
File Edit View Run Help   Last edit was 3 days ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ✓   🗓 Schedule   Share

⊕ New
🔲 Workspace
🗂 Repos
🕐 Recents
🔷 Data
🔗 Compute
🗒 Workflows

p_data_source

p_file_date
2021-03-21

Step 4 - Write to output to processed container in parquet format

Cmd 18

```
drivers_final_df.write.mode("overwrite").format("delta").saveAsTable("f1_processed.drivers")
```

▶ (6) Spark Jobs

Command took 3.71 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:21:49 PM on dbkcourse_cluster

Cmd 19

```
%sql
SELECT * FROM f1_processed.drivers
```

▶ (2) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [driver_id: integer, driver_ref: string ... 8 more fields]

Table ✓   +

	driver_id	driver_ref	number	code	name	dob	nationality	ingestion_date	data_source	file_date
1	1	hamilton	44	HAM	Lewis Hamilton	1985-01-07	British	2023-07-15T04:21:55.742+0000		2021-03-21
2	2	heidfeld	null	HEI	Nick Heidfeld	1977-05-10	German	2023-07-15T04:21:55.742+0000		2021-03-21
3	3	rosberg	6	ROS	Nico Rosberg	1985-06-27	German	2023-07-15T04:21:55.742+0000		2021-03-21
4	4	alonso	14	ALO	Fernando Alonso	1981-07-29	Spanish	2023-07-15T04:21:55.742+0000		2021-03-21
5	5	kovalainen	null	KOV	Heikki Kovalainen	1981-10-19	Finnish	2023-07-15T04:21:55.742+0000		2021-03-21
6	6	nakajima	null	NAK	Kazuki Nakajima	1985-01-11	Japanese	2023-07-15T04:21:55.742+0000		2021-03-21
7	7	bourdais	null	BOU	Sébastien Bourdais	1979-02-28	French	2023-07-15T04:21:55.742+0000		2021-03-21

⤓   853 rows | 0.79 seconds runtime   Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 0.79 seconds -- by avinashchinta@my.unt.edu at 7/14/2023, 11:21:49 PM on dbkcourse_cluster

Cmd 20

🏪 Marketplace NEW
🔗 Partner Connect
1/4 Tasks Completed
◉ Enable new UI NEW

processed ···
Container

Search «

- Overview
- Diagnose and solve problems
- Access Control (IAM)

Settings

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

⌃ Upload   + Add Directory   ↻ Refresh | ↰ Rename   🗑 Delete   ⇄ Change tier   🖉 Acquire lease   🖉 Break lease   🖉 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed

Search blobs by prefix (case-sensitive)	● Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐ 📁 circuits						-	···
☐ 📁 constructors						-	···
☐ 📁 drivers						-	···
☐ 📁 races						-	···

common_functions Python ▾

File   Edit   View   Run   Help   Last edit was 3 days ago   Provide feedback

▶ Run all   ● Connect ▾   📅 Schedule   Share   ^

⊕ New

▮ Workspace

Repos

Recents

Data

Compute

Workflows

```python
        column_value_list = [row[column_name] for row in df_row_list]
        return column_value_list
```

Cmd 5

```python
    def merge_delta_data(input_df, db_name, table_name, folder_path, merge_condition, partition_column):
        spark.conf.set("spark.databricks.optimizer.dynamicPartitionPruning","true")

        from delta.tables import DeltaTable
        if (spark._jsparkSession.catalog().tableExists(f"{db_name}.{table_name}")):
            deltaTable = DeltaTable.forPath(spark, f"{folder_path}/{table_name}")
            deltaTable.alias("tgt").merge(
                input_df.alias("src"),
                merge_condition) \
            .whenMatchedUpdateAll()\
            .whenNotMatchedInsertAll()\
            .execute()
        else:
            input_df.write.mode("overwrite").partitionBy(partition_column).format("delta").saveAsTable(f"{db_name}.{table_name}")
```

Cmd 6

Shift+Enter to run
Shift+Ctrl+Enter to run selected text

5.ingest_results_file Python ▾

File   Edit   View   Run   Help   Last edit was 1 minute ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ▾   📅 Schedule   Share   ^

p_data_source

Ergast

p_file_date

2021-03-21

✏ ⚙ 📌

```
merge_condition = "tgt.result_id = src.result_id AND tgt.race_id = src.race_id"
merge_delta_data(results_deduped_df, 'f1_processed', 'results', processed_folder_path, merge_condition, 'race_id')
```

▶ (6) Spark Jobs

Command took 1.92 minutes -- by avinashchinta@my.unt.edu at 7/15/2023, 12:37:13 AM on dbkcourse_cluster

Cmd 26

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.61 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:37:13 AM on dbkcourse_cluster

Cmd 27

SQL   ▶▾ 📊 ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.results
GROUP BY race_id
ORDER BY race_id DESC;
```

▶ (3) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1047	20
2	1046	20
3	1045	20
4	1044	20
5	1043	20
6	1042	20
7	1041	20

⬇ 1,035 rows | 1.00 second runtime   Refreshed now

▶ Run all   ● dbkcourse_cluster ▾   🗓 Schedule   Share   ⌃

p_data_source
Ergast

p_file_date
2021-03-28

```python
merge_condition = "tgt.result_id = src.result_id AND tgt.race_id = src.race_id"
merge_delta_data(results_deduped_df, 'f1_processed', 'results', processed_folder_path, merge_condition, 'race_id')
```

▶ (8) Spark Jobs

Command took 19.66 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:44:46 AM on dbkcourse_cluster

Cmd 26

```python
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:44:46 AM on dbkcourse_cluster

Cmd 27

SQL   ▶▾ ⌊ⅈⅈ ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.results
GROUP BY race_id
ORDER BY race_id DESC;
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1052	20
2	1047	20
3	1046	20
4	1045	20
5	1044	20
6	1043	20
7	1042	20

⤓ 1,036 rows | 1.68 seconds runtime   Refreshed now

⊕ New

■ Workspace

📁 Repos

🕐 Recents

⛁ Data

🏠 Compute

☰ Workflows

p_data_source
Ergast

p_file_date
2021-04-18

✎ ⚙ 📌

```
        merge_condition = "tgt.result_id = src.result_id AND tgt.race_id = src.race_id"
        merge_delta_data(results_deduped_df, 'f1_processed', 'results', processed_folder_path, merge_condition, 'race_id')
```

▸ (8) Spark Jobs

Command took 16.90 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:46:47 AM on dbkcourse_cluster

Cmd 26

```
        dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:46:47 AM on dbkcourse_cluster

Cmd 27

SQL ▶▾ 📊 ∨ − ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.results
GROUP BY race_id
ORDER BY race_id DESC;
```

▸ (5) Spark Jobs

▸ ⊞ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1053	20
2	1052	20
3	1047	20
4	1046	20
5	1045	20
6	1044	20
7	1043	20

⤓ 1,037 rows | 1.50 seconds runtime   Refreshed now

💬
⚖
🕐
{×}
ılıı

📕 Marketplace NEW

⚙ Partner Connect

1/4 Tasks Completed

👁 Enable new UI NEW

☐ Menu options

processed ...
Container

- Overview
- Diagnose and solve problems
- Access Control (IAM)

Settings

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

↑ Upload   + Add Directory   ↻ Refresh | ↶ Rename   🗑 Delete   ⇄ Change tier   🔑 Acquire lease   🔑 Break lease   📷 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / results

| Search blobs by prefix (case-sensitive) | | (●) Show deleted objects |

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐	📁 [..]							•••
☐	📁 _delta_log					-		•••
☐	📁 race_id=1					-		•••
☐	📁 race_id=10					-		•••
☐	📁 race_id=100					-		•••
☐	📁 race_id=1000					-		•••
☐	📁 race_id=1001					-		•••
☐	📁 race_id=1002					-		•••
☐	📁 race_id=1003					-		•••
☐	📁 race_id=1004					-		•••
☐	📁 race_id=1005					-		•••
☐	📁 race_id=1006					-		•••
☐	📁 race_id=1007					-		•••
☐	📁 race_id=1008					-		•••
☐	📁 race_id=1009					-		•••
☐	📁 race_id=101					-		•••

▶ Run all   ● dbkcourse_cluster ∨   🗓 Schedule   Share   ∧

p_data_source

p_file_date
2021-03-21

Cmd 16

```
merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.stop = src.stop AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'pit_stops', processed_folder_path, merge_condition, 'race_id')
```

▶ (5) Spark Jobs

Command took 20.84 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:51:11 AM on dbkcourse_cluster

Cmd 17

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.59 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:51:11 AM on dbkcourse_cluster

Cmd 18

SQL ▶▾ ⬛ ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.pit_stops
GROUP BY race_id
ORDER BY race_id DESC;
```

▶ (3) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1047	23
2	1046	39
3	1045	57
4	1044	38
5	1043	30
6	1042	25
7	1041	33

▶ Run all   ● dbkcourse_cluster ✓   📅 Schedule   Share   ^

p_data_source   p_file_date
[   ]   [2021-03-28   ]

Cmd 16

```
merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.stop = src.stop AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'pit_stops', processed_folder_path, merge_condition, 'race_id')
```

▸ (7) Spark Jobs

Command took 4.34 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:53:24 AM on dbkcourse_cluster

Cmd 17

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:53:24 AM on dbkcourse_cluster

Cmd 18

```
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.pit_stops
GROUP BY race_id
ORDER BY race_id DESC;
```

▸ (5) Spark Jobs

▸ ▤ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ✓   +

	race_id ▲	count(1) ▲
1	1052	40
2	1047	23
3	1046	39
4	1045	57
5	1044	38
6	1043	30
7	1042	25

New

Workspace

Repos

Recents

Data

Compute

Workflows

Marketplace **NEW**

Partner Connect

1/4 Tasks Completed

Enable new UI **NEW**

Menu options

6.ingest_pit_stops_file Python ⌄
File Edit View Run Help   Last edit was 2 minutes ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ⌄   🗓 Schedule   **Share**   ⌃

p_data_source

p_file_date
2021-04-18

Cmd 16

```
merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.stop = src.stop AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'pit_stops', processed_folder_path, merge_condition, 'race_id')
```

▶ (7) Spark Jobs

Command took 4.37 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:54:07 AM on dbkcourse_cluster

Cmd 17

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:54:07 AM on dbkcourse_cluster

Cmd 18

SQL ▶⌄ 📊 ⌄ ━ ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.pit_stops
GROUP BY race_id
ORDER BY race_id DESC;
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ⌄   +

	race_id ▲	count(1) ▲
1	1053	56
2	1052	40
3	1047	23
4	1046	39
5	1045	57
6	1044	38
7	1043	30

processed ...
Container

×

Upload   + Add Directory   ↻ Refresh | ↶ Rename   🗑 Delete   ⇄ Change tier   🖊 Acquire lease   🖊 Break lease   📷 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / pit_stops

| Search blobs by prefix (case-sensitive) | ⚫ Show deleted objects |

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐	📁 [..]							...
☐	📁 _delta_log						-	...
☐	📁 race_id=1000						-	...
☐	📁 race_id=1001						-	...
☐	📁 race_id=1002						-	...
☐	📁 race_id=1003						-	...
☐	📁 race_id=1004						-	...
☐	📁 race_id=1005						-	...
☐	📁 race_id=1006						-	...
☐	📁 race_id=1007						-	...
☐	📁 race_id=1008						-	...
☐	📁 race_id=1009						-	...
☐	📁 race_id=1010						-	...
☐	📁 race_id=1011						-	...
☐	📁 race_id=1012						-	...
☐	📁 race_id=1013						-	...

⊕ New

▮ Workspace

🔀 Repos

🕐 Recents

🔠 Data

🔗 Compute

☰ Workflows

p_data_source

p_file_date
2021-03-21

🖉 ⚙ 📌

```
Cmd 16
```

```python
    merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.lap = src.lap AND tgt.race_id = src.race_id"
    merge_delta_data(final_df, 'f1_processed', 'lap_times', processed_folder_path, merge_condition, 'race_id')
```

▶ (5) Spark Jobs

Command took 19.80 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:57:58 AM on dbkcourse_cluster

```
Cmd 17
```

```python
    dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.61 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:57:58 AM on dbkcourse_cluster

```
Cmd 18
```

SQL ▶▾ 📊 ⌄ ━ ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.lap_times
GROUP BY race_id
ORDER BY race_id DESC
```

▶ (4) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ⌄   +

	race_id ▲	count(1) ▲
1	1047	1043
2	1046	1531
3	1045	1016
4	1044	1076
5	1043	1128
6	1042	1288
7	1041	1017

⬇ 454 rows | 1.06 seconds runtime

Refreshed now

🏬 Marketplace   NEW

🔢 Partner Connect

1/4 Tasks Completed

👁 Enable new UI   NEW

▥ Menu options

7.ingest_lap_times_file Python ∨

File Edit View Run Help   Last edit was 1 minute ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ∨   📅 Schedule   **Share**   ⌃

⊕ New

🔲 Workspace

📚 Repos

🕐 Recents

🔺 Data

🏠 Compute

≅ Workflows

p_data_source

p_file_date

2021-03-28

📝 ⚙ 📌

Cmd 16

```
merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.lap = src.lap AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'lap_times', processed_folder_path, merge_condition, 'race_id')
```

▶ (7) Spark Jobs

Command took 7.55 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:59:55 AM on dbkcourse_cluster

Cmd 17

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 12:59:55 AM on dbkcourse_cluster

Cmd 18

SQL ▶▼ 📊 ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.lap_times
GROUP BY race_id
ORDER BY race_id DESC
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1052	1026
2	1047	1043
3	1046	1531
4	1045	1016
5	1044	1076
6	1043	1128
7	1042	1288

↓   455 rows | 1.31 seconds runtime   Refreshed now

7.ingest_lap_times_file Python ▼
File Edit View Run Help   Last edit was 2 minutes ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ▼   📅 Schedule   Share

New
Workspace
Repos
Recents
Data
Compute
Workflows

p_data_source

p_file_date
2021-04-18

Cmd 16

```
merge_condition = "tgt.race_id = src.race_id AND tgt.driver_id = src.driver_id AND tgt.lap = src.lap AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'lap_times', processed_folder_path, merge_condition, 'race_id')
```

▶ (7) Spark Jobs

Command took 6.67 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:00:34 AM on dbkcourse_cluster

Cmd 17

```
dbutils.notebook.exit("Success")
```

Notebook exited: Success

Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:00:34 AM on dbkcourse_cluster

Cmd 18

SQL ▶▼ ⎍ ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.lap_times
GROUP BY race_id
ORDER BY race_id DESC
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1053	1124
2	1052	1026
3	1047	1043
4	1046	1531
5	1045	1016
6	1044	1076
7	1043	1128

⬇ 456 rows | 1.23 seconds runtime   Refreshed now

Marketplace NEW
Partner Connect
1/4 Tasks Completed

Enable new UI NEW
Menu options

processed ···
Container

✕

↑ Upload   ＋ Add Directory   ↻ Refresh | ↱ Rename   🗑 Delete   ⇄ Change tier   🖉 Acquire lease   🖉 Break lease   🗨 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / lap_times

Search blobs by prefix (case-sensitive)		⬤ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐ 📁 [..]							···
☐ 📁 _delta_log						-	···
☐ 📁 race_id=1						-	···
☐ 📁 race_id=10						-	···
☐ 📁 race_id=100						-	···
☐ 📁 race_id=1000						-	···
☐ 📁 race_id=1001						-	···
☐ 📁 race_id=1002						-	···
☐ 📁 race_id=1003						-	···
☐ 📁 race_id=1004						-	···
☐ 📁 race_id=1005						-	···
☐ 📁 race_id=1006						-	···
☐ 📁 race_id=1007						-	···
☐ 📁 race_id=1008						-	···
☐ 📁 race_id=1009						-	···
☐ 📁 race_id=101						-	···

8.ingest_qualifying_file Python ˅
File   Edit   View   Run   Help   Last edit was 2 minutes ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ˅   📅 Schedule   Share   ⌃

⊕ New

▮ Workspace

⎘ Repos

🕐 Recents

⚘ Data

🏠 Compute

⋮≡ Workflows

p_data_source

p_file_date

2021-03-21

Cmd 16

```python
merge_condition = "tgt.qualify_id = src.qualify_id AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'qualifying', processed_folder_path, merge_condition, 'race_id')
```

▸ (5) Spark Jobs

Command took 23.70 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:03:29 AM on dbkcourse_cluster

Cmd 17

SQL ▶˅ ⊡ ˅ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.qualifying
GROUP BY race_id
ORDER BY race_id DESC
```

▸ (4) Spark Jobs

▸ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ˅   +

	race_id ▲	count(1) ▲
1	1047	20
2	1046	20
3	1045	20
4	1044	20
5	1043	20
6	1042	20
7	1041	20

⬇ 404 rows | 1.02 seconds runtime   Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 1.02 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:03:29 AM on dbkcourse_cluster

Cmd 18

🏪 Marketplace   NEW

⊡ Partner Connect

1/4 Tasks Completed

◉ Enable new UI   NEW

▢ Menu options

8.ingest_qualifying_file Python ∨
File Edit View Run Help Last edit was 1 minute ago Provide feedback

▶ Run all ● dbkcourse_cluster ∨ 📅 Schedule Share ⌃

- New
- **Workspace**
- Repos
- Recents
- Data
- Compute
- Workflows

p_data_source

p_file_date

2021-03-28

Cmd 16

```
merge_condition = "tgt.qualify_id = src.qualify_id AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'qualifying', processed_folder_path, merge_condition, 'race_id')
```

▶ (7) Spark Jobs

Command took 5.61 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:04:34 AM on dbkcourse_cluster

Cmd 17

SQL ▶▼ ⌊⊞⌋ ∨ — ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.qualifying
GROUP BY race_id
ORDER BY race_id DESC
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨ +

	race_id ▲	count(1) ▲	
1	1052	20	
2	1047	20	
3	1046	20	
4	1045	20	
5	1044	20	
6	1043	20	
7	1042	20	

⬇ 405 rows | 1.30 seconds runtime Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 1.30 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:04:34 AM on dbkcourse_cluster

Cmd 18

- 🏪 Marketplace NEW
- 🔗 Partner Connect
- 1/4 Tasks Completed
- 👁 Enable new UI NEW
- ⊞ Menu options

New

Workspace

Repos

Recents

Data

Compute

Workflows

8.ingest_qualifying_file Python ▾
File Edit View Run Help   Last edit was 2 minutes ago   Provide feedback

▶ Run all   ● dbkcourse_cluster ▾   📅 Schedule   Share

p_data_source

p_file_date
2021-04-18

Cmd 16

```
merge_condition = "tgt.qualify_id = src.qualify_id AND tgt.race_id = src.race_id"
merge_delta_data(final_df, 'f1_processed', 'qualifying', processed_folder_path, merge_condition, 'race_id')
```

▶ (7) Spark Jobs

Command took 5.11 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:05:06 AM on dbkcourse_cluster

Cmd 17

SQL ▶▾ �julⅼ ∨ − ✕

```sql
%sql
SELECT race_id, COUNT(1)
FROM f1_processed.qualifying
GROUP BY race_id
ORDER BY race_id DESC
```

▶ (5) Spark Jobs

▶ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [race_id: integer, count(1): long]

Table ∨   +

	race_id ▲	count(1) ▲
1	1053	20
2	1052	20
3	1047	20
4	1046	20
5	1045	20
6	1044	20
7	1043	20

⬇ 406 rows | 1.24 seconds runtime   Refreshed now

ⓘ SQL cell result stored as PySpark data frame _sqldf . Learn more

Command took 1.24 seconds -- by avinashchinta@my.unt.edu at 7/15/2023, 1:05:06 AM on dbkcourse_cluster

Cmd 18

Marketplace NEW

Partner Connect

1/4 Tasks Completed

Enable new UI NEW

processed ...
Container

⬆ Upload   ＋ Add Directory   🔄 Refresh   │   🗘 Rename   🗑 Delete   ⇄ Change tier   🖉 Acquire lease   🖉 Break lease   👥 Give feedback

Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / qualifying

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	📁 [..]							•••
□	📁 _delta_log						-	•••
□	📁 race_id=1						-	•••
□	📁 race_id=10						-	•••
□	📁 race_id=100						-	•••
□	📁 race_id=1000						-	•••
□	📁 race_id=1001						-	•••
□	📁 race_id=1002						-	•••
□	📁 race_id=1003						-	•••
□	📁 race_id=1004						-	•••
□	📁 race_id=1005						-	•••
□	📁 race_id=1006						-	•••
□	📁 race_id=1007						-	•••
□	📁 race_id=1008						-	•••
□	📁 race_id=1009						-	•••
□	📁 race_id=101						-	•••

Search blobs by prefix (case-sensitive)   ⬤ Show deleted objects