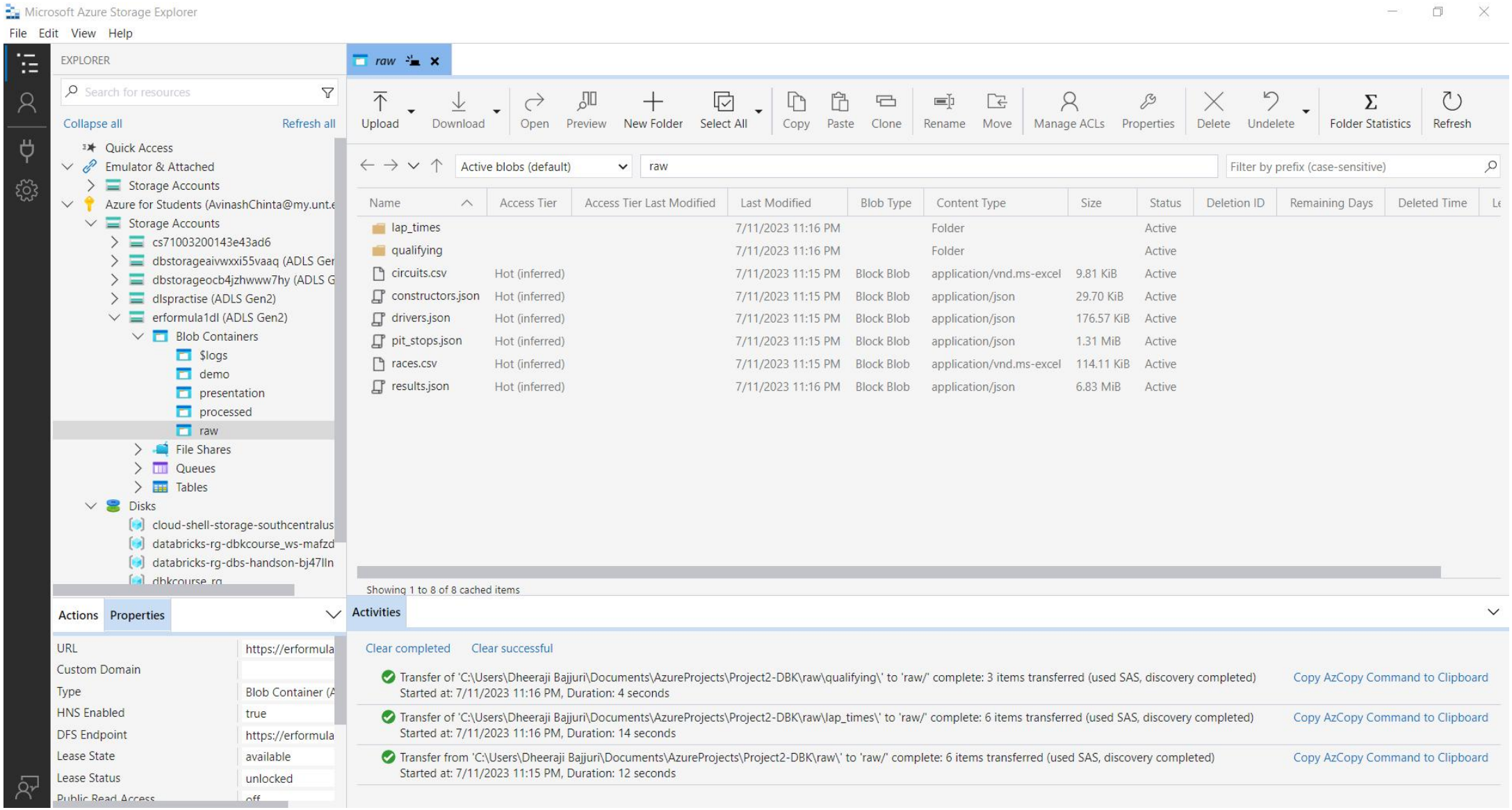
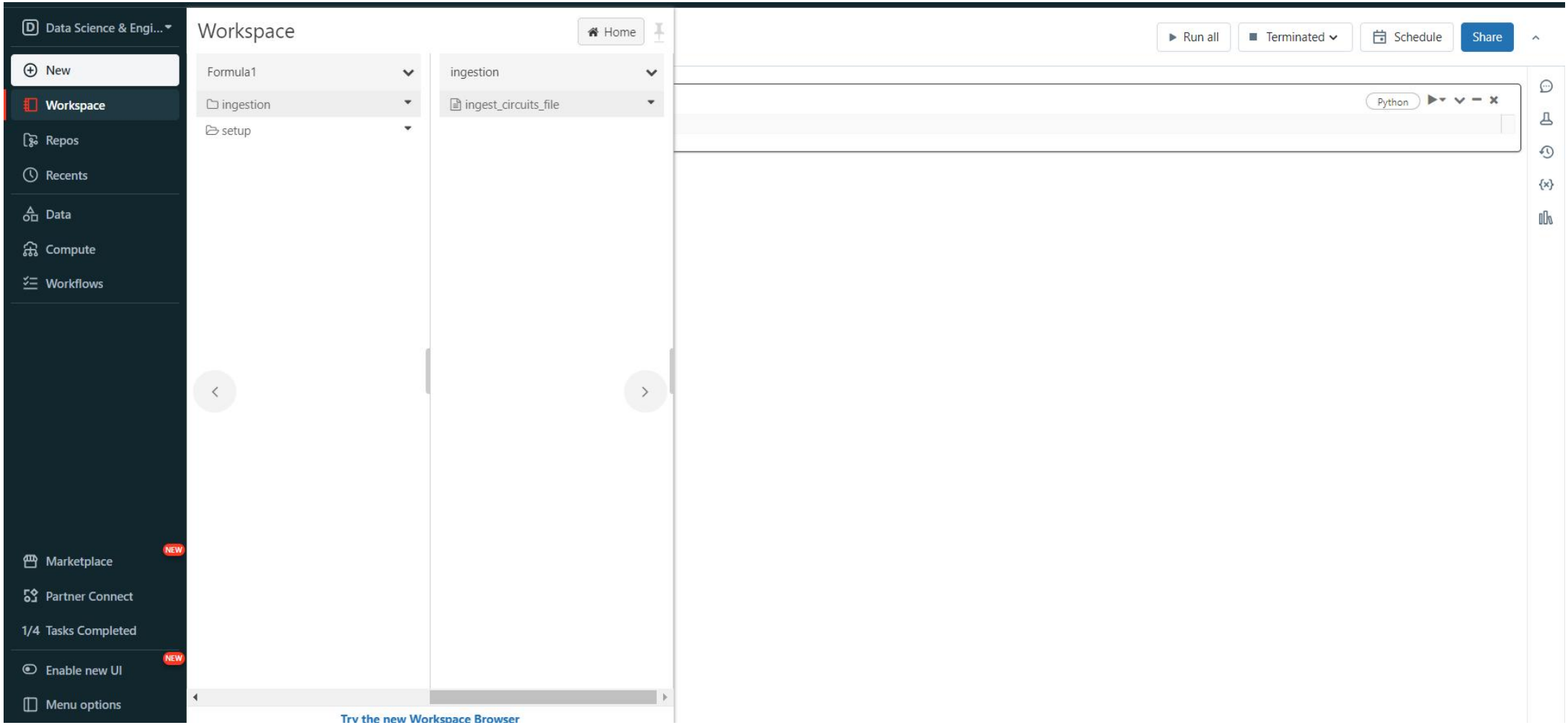
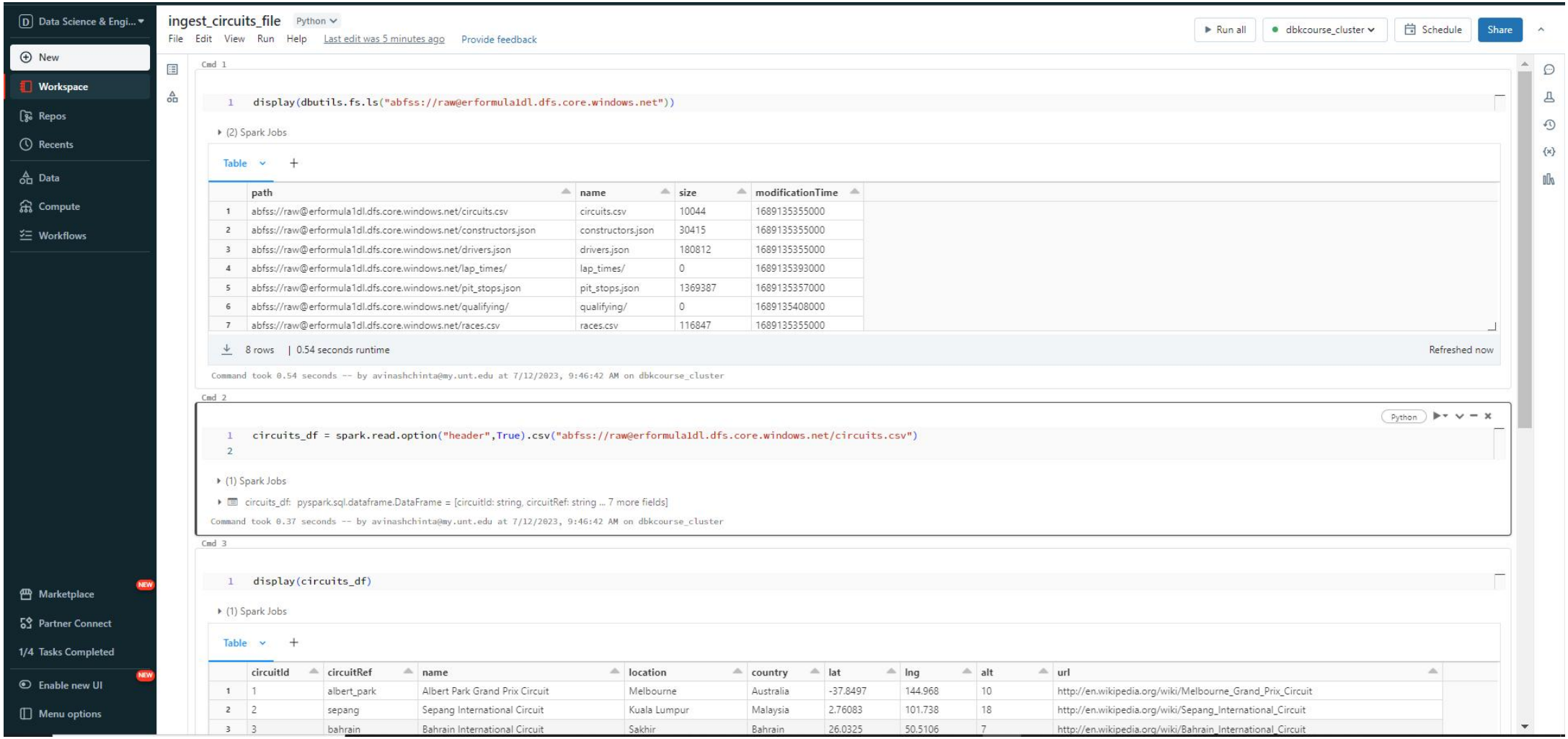
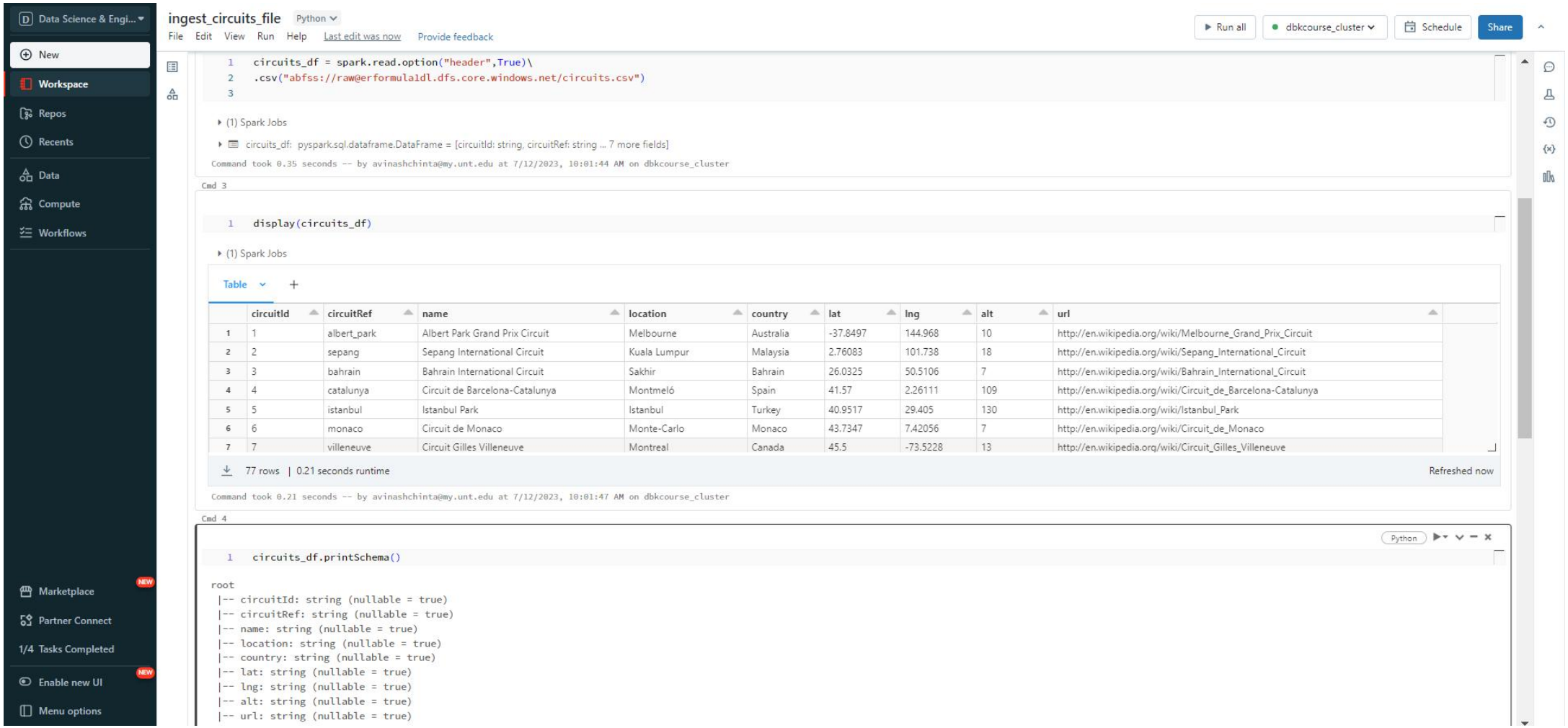
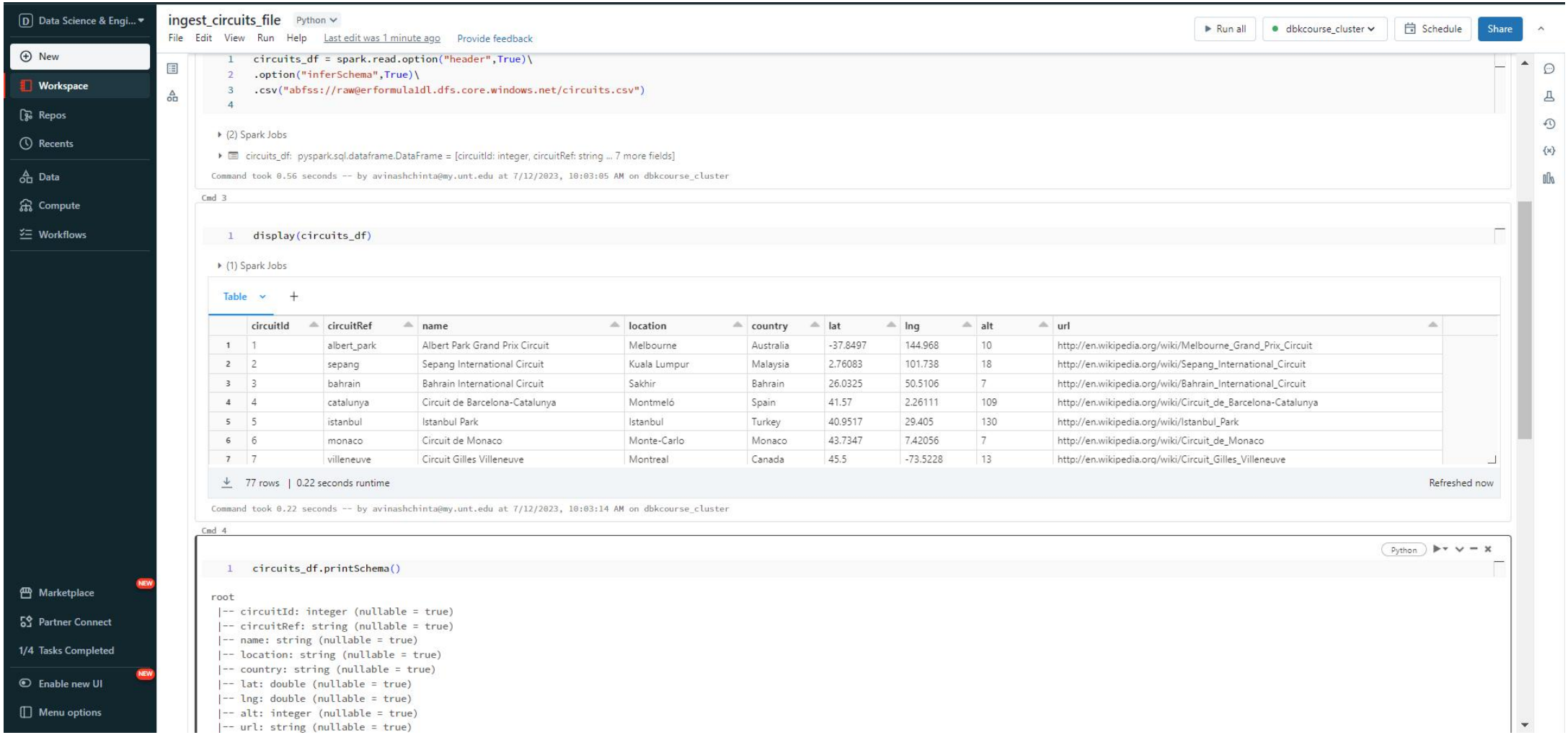
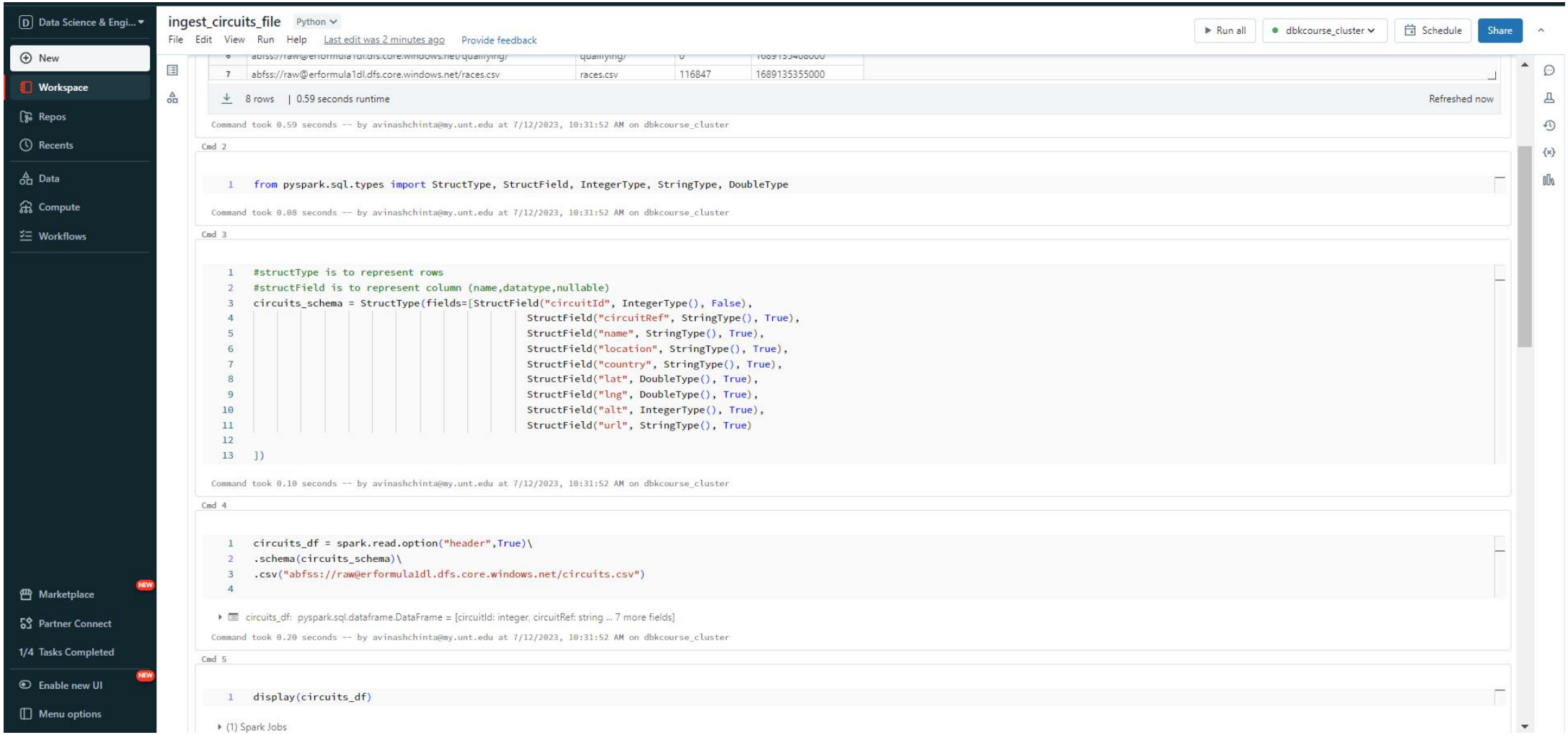

Microsoft Azure Storage Explorer
File Edit View Help
EXPLORER
Search for resources
Collapse all
Refresh all
Quick Access
Emulator & Attached
Storage Accounts
Azure for Students (AvinashChinta@my.unt.e
Storage Accounts
cs71003200143e43ad6
dbstorageaivwxxi55vaaq (ADLS Ger
dbstorageocb4jzhwww7hy (ADLS G
dlspractise (ADLS Gen2)
erformula1dl (ADLS Gen2)
Blob Containers
$logs
demo
presentation
processed
raw
File Shares
Queues
Tables
Disks
cloud-shell-storage-southcentralus
databricks-rg-dbkcourse_ws-mafzd
databricks-rg-dbs-handson-bj47lln
dbkcourse_rg
raw
Upload
Download
Open
Preview
New Folder
Select All
Copy
Paste
Clone
Rename
Move
Manage ACLs
Properties
Delete
Undelete
Folder Statistics
Refresh
Active blobs (default)
raw
Filter by prefix (case-sensitive)
Name
Access Tier
Access Tier Last Modified
Last Modified
Blob Type
Content Type
Size
Status
Deletion ID
Remaining Days
Deleted Time
lap_times 7/11/2023 11:16 PM Folder Active
qualifying 7/11/2023 11:16 PM Folder Active
circuits.csv Hot (inferred) 7/11/2023 11:15 PM Block Blob application/vnd.ms-excel 9.81 KiB Active
constructors.json Hot (inferred) 7/11/2023 11:15 PM Block Blob application/json 29.70 KiB Active
drivers.json Hot (inferred) 7/11/2023 11:15 PM Block Blob application/json 176.57 KiB Active
pit_stops.json Hot (inferred) 7/11/2023 11:15 PM Block Blob application/json 1.31 MiB Active
races.csv Hot (inferred) 7/11/2023 11:15 PM Block Blob application/vnd.ms-excel 114.11 KiB Active
results.json Hot (inferred) 7/11/2023 11:16 PM Block Blob application/json 6.83 MiB Active
Showing 1 to 8 of 8 cached items
Actions
Properties
URL https://erformula
Custom Domain
Type Blob Container (A
HNS Enabled true
DFS Endpoint https://erformula
Lease State available
Lease Status unlocked
Public Read Access off
Activities
Clear completed
Clear successful
Transfer of 'C:\Users\Dheeraji Bajjuri\Documents\AzureProjects\Project2-DBK\raw\qualifying\' to 'raw/' complete: 3 items transferred (used SAS, discovery completed)
Started at: 7/11/2023 11:16 PM, Duration: 4 seconds
Copy AzCopy Command to Clipboard
Transfer of 'C:\Users\Dheeraji Bajjuri\Documents\AzureProjects\Project2-DBK\raw\lap_times\' to 'raw/' complete: 6 items transferred (used SAS, discovery completed)
Started at: 7/11/2023 11:16 PM, Duration: 14 seconds
Copy AzCopy Command to Clipboard
Transfer from 'C:\Users\Dheeraji Bajjuri\Documents\AzureProjects\Project2-DBK\raw\' to 'raw/' complete: 6 items transferred (used SAS, discovery completed)
Started at: 7/11/2023 11:15 PM, Duration: 12 seconds
Copy AzCopy Command to Clipboard

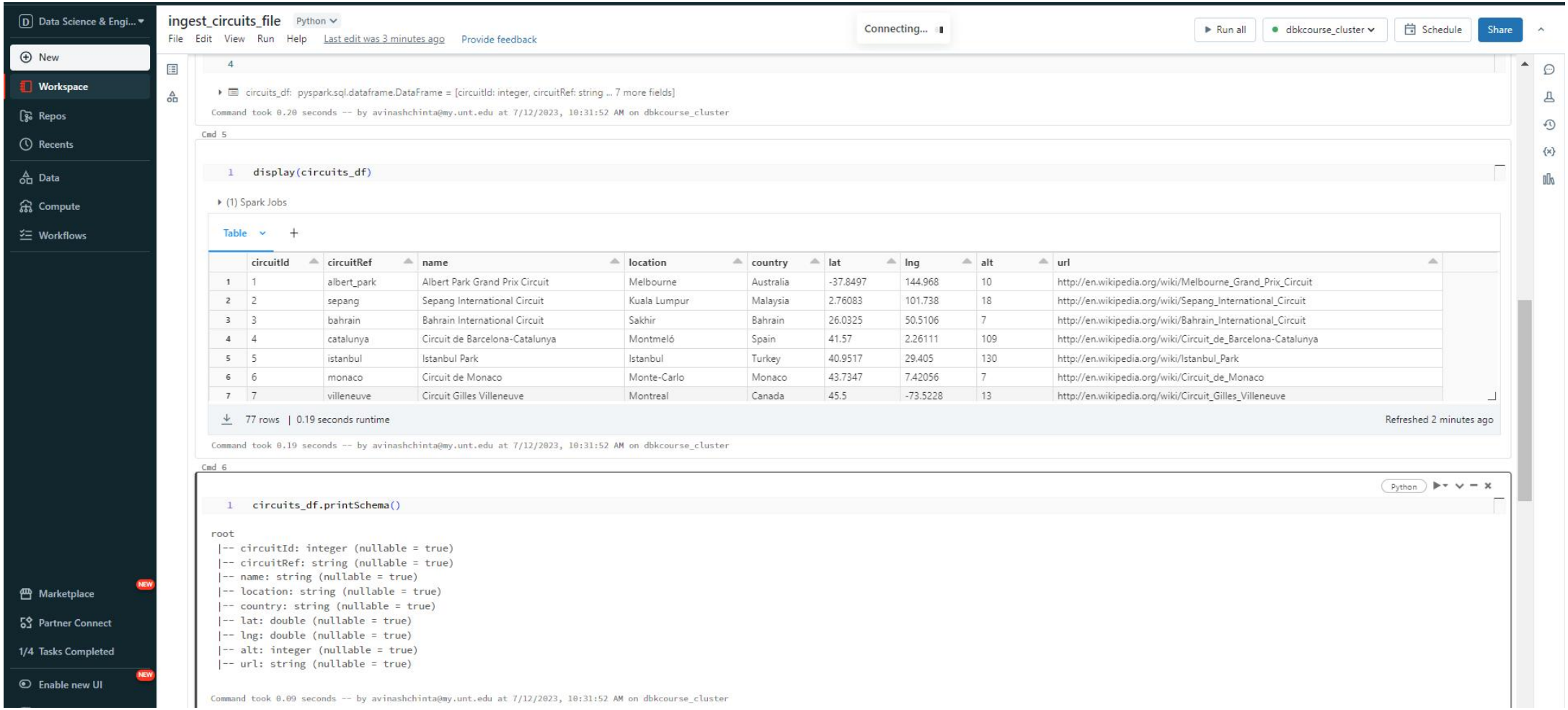

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
Partner Connect
1/4 Tasks Completed
Enable new UI
NEW
Menu options
Workspace
Home
Formula1
ingestion
setup
ingestion
ingest_circuits_file
Try the new Workspace Browser
Run all
Terminated
Schedule
Share
Python

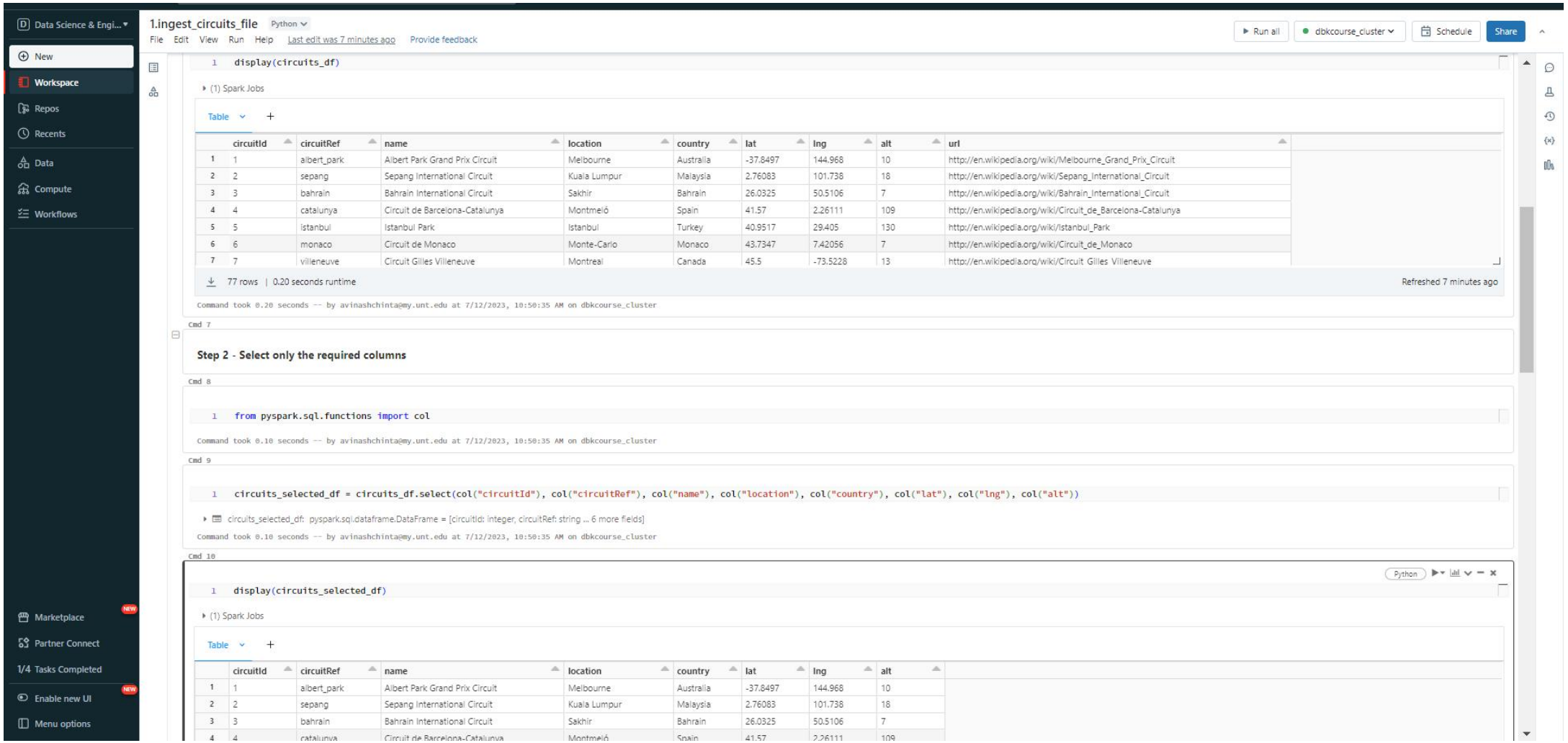

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
ingest_circuits_file
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 1
display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path name size modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 0.54 seconds runtime
Refreshed now
Command took 0.54 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 9:46:42 AM on dbkcourse_cluster
Cmd 2
Python
circuits_df = spark.read.option("header",True).csv("abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv")
(1) Spark Jobs
circuits_df: pyspark.sql.dataframe.DataFrame = [circuitId: string, circuitRef: string ... 7 more fields]
Command took 0.37 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 9:46:42 AM on dbkcourse_cluster
Cmd 3
display(circuits_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt url
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 http://en.wikipedia.org/wiki/Sepang_International_Circuit
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 http://en.wikipedia.org/wiki/Bahrain_International_Circuit

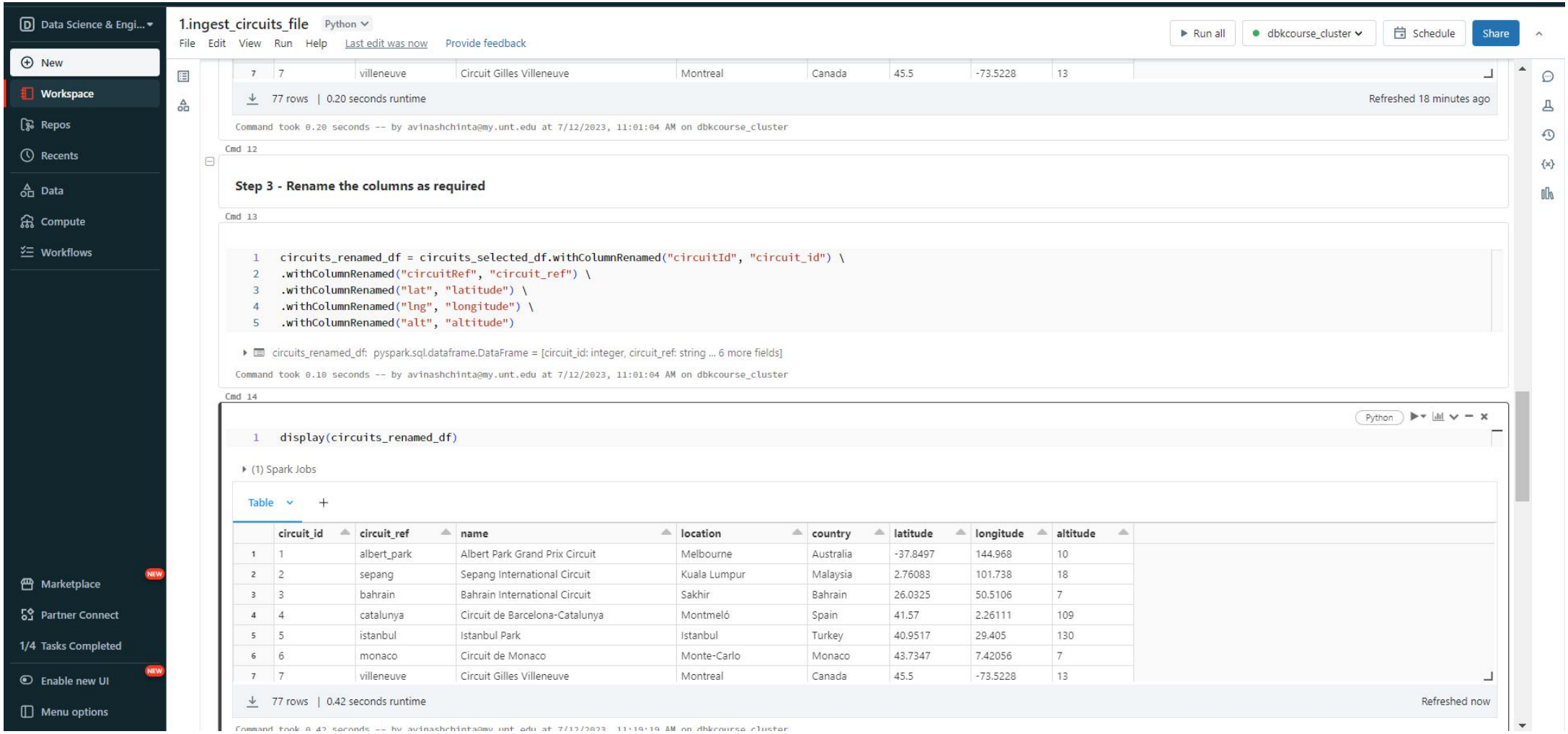

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
ingest_circuits_file
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
1 circuits_df = spark.read.option("header",True)\
2 .csv("abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv")
3
(1) Spark Jobs
circuits_df: pyspark.sql.dataframe.DataFrame = [circuitId: string, circuitRef: string ... 7 more fields]
Command took 0.35 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:01:44 AM on dbkcourse_cluster
Cmd 3
1 display(circuits_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt url
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 http://en.wikipedia.org/wiki/Sepang_International_Circuit
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 http://en.wikipedia.org/wiki/Bahrain_International_Circuit
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 http://en.wikipedia.org/wiki/Circuit_de_Barcelona-Catalunya
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 http://en.wikipedia.org/wiki/Istanbul_Park
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 http://en.wikipedia.org/wiki/Circuit_de_Monaco
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 http://en.wikipedia.org/wiki/Circuit_Gilles_Villeneuve
77 rows | 0.21 seconds runtime
Refreshed now
Command took 0.21 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:01:47 AM on dbkcourse_cluster
Cmd 4
Python
1 circuits_df.printSchema()
root
|-- circuitId: string (nullable = true)
|-- circuitRef: string (nullable = true)
|-- name: string (nullable = true)
|-- location: string (nullable = true)
|-- country: string (nullable = true)
|-- lat: string (nullable = true)
|-- lng: string (nullable = true)
|-- alt: string (nullable = true)
|-- url: string (nullable = true)

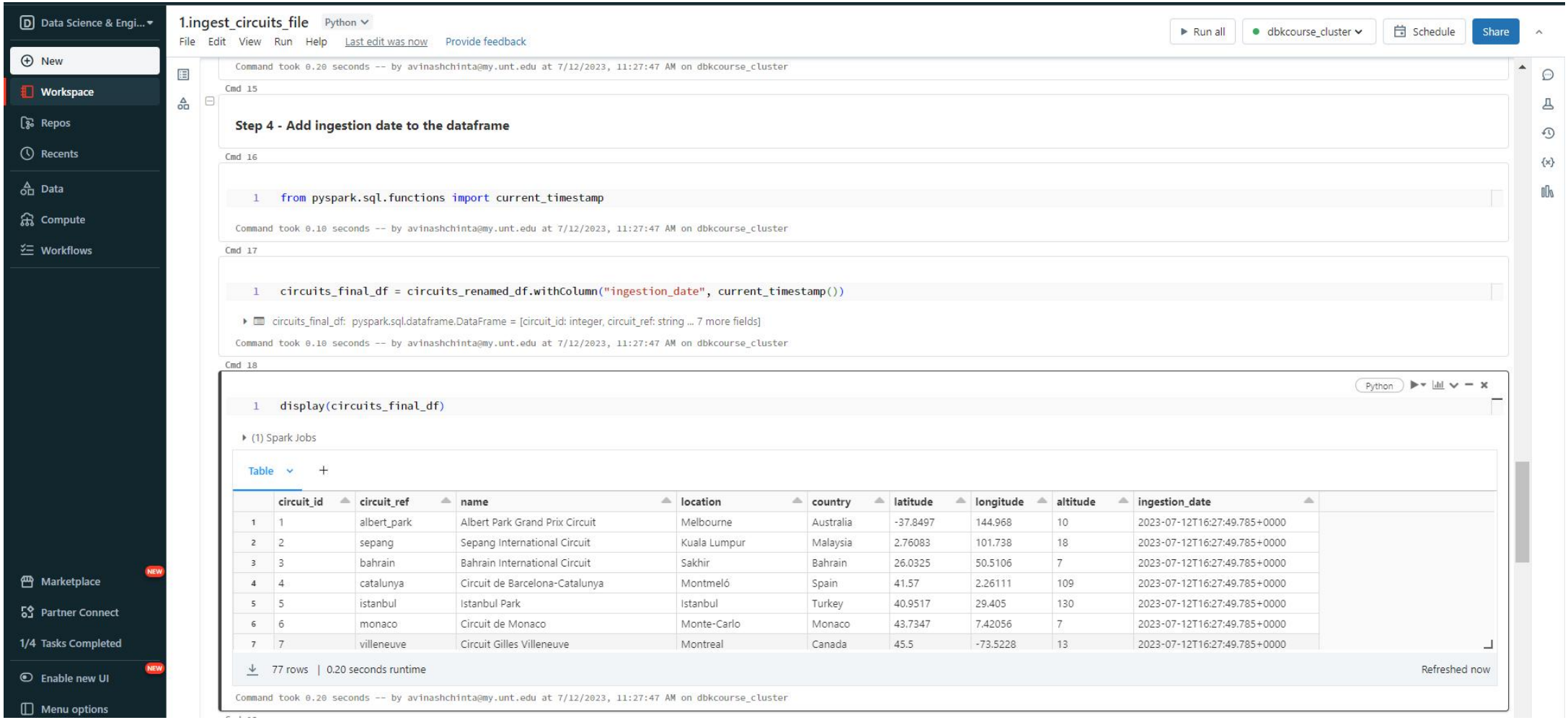

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
ingest_circuits_file
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
1 circuits_df = spark.read.option("header",True)\
2 .option("inferSchema",True)\
3 .csv("abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv")
4
(2) Spark Jobs
circuits_df: pyspark.sql.dataframe.DataFrame = [circuitId: integer, circuitRef: string ... 7 more fields]
Command took 0.56 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:03:05 AM on dbkcourse_cluster
Cmd 3
1 display(circuits_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt url
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 http://en.wikipedia.org/wiki/Sepang_International_Circuit
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 http://en.wikipedia.org/wiki/Bahrain_International_Circuit
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 http://en.wikipedia.org/wiki/Circuit_de_Barcelona-Catalunya
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 http://en.wikipedia.org/wiki/Istanbul_Park
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 http://en.wikipedia.org/wiki/Circuit_de_Monaco
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 http://en.wikipedia.org/wiki/Circuit_Gilles_Villeneuve
77 rows | 0.22 seconds runtime
Refreshed now
Command took 0.22 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:03:14 AM on dbkcourse_cluster
Cmd 4
Python
1 circuits_df.printSchema()
root
|-- circuitId: integer (nullable = true)
|-- circuitRef: string (nullable = true)
|-- name: string (nullable = true)
|-- location: string (nullable = true)
|-- country: string (nullable = true)
|-- lat: double (nullable = true)
|-- lng: double (nullable = true)
|-- alt: integer (nullable = true)
|-- url: string (nullable = true)

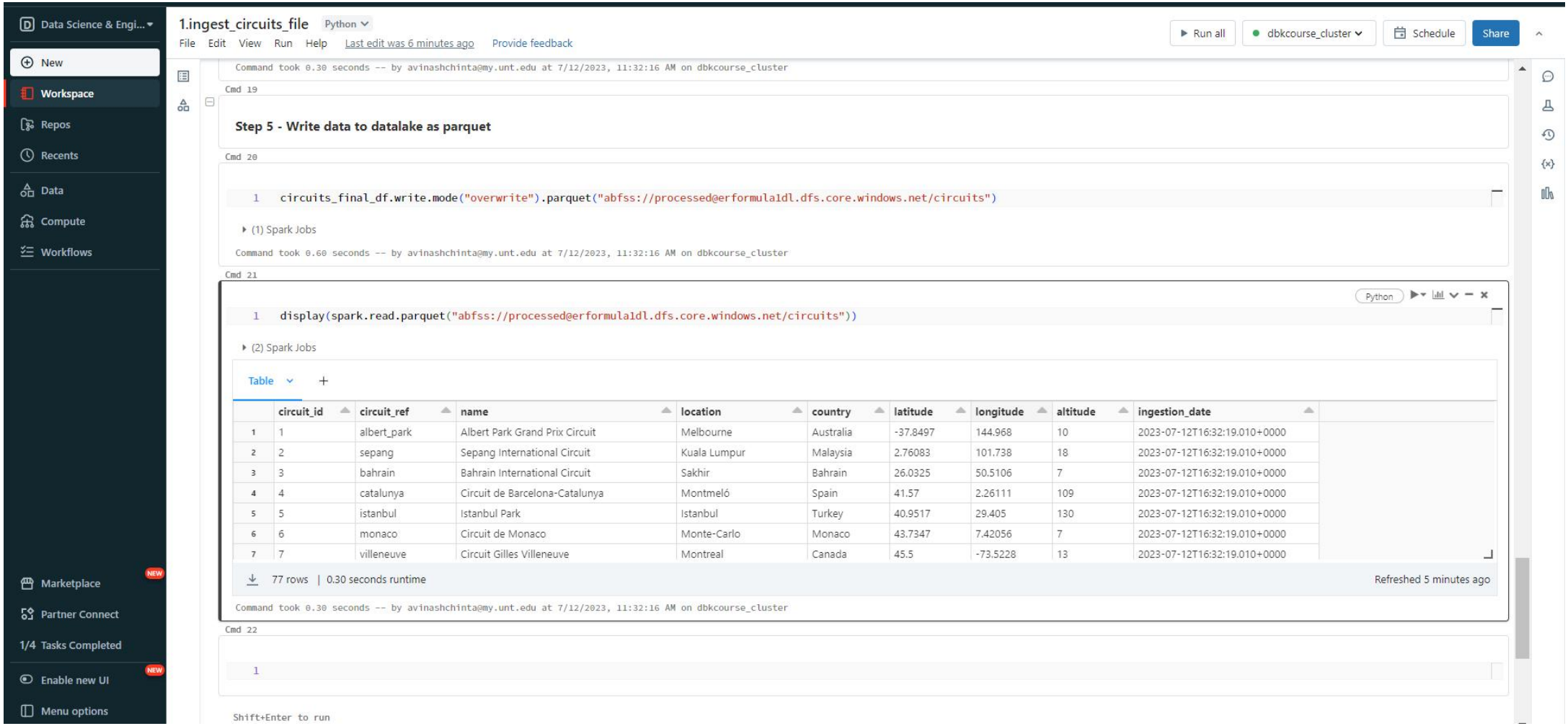

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
ingest_circuits_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 0.59 seconds runtime
Refreshed now
Command took 0.59 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 2
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 3
1 #structType is to represent rows
2 #structField is to represent column (name,datatype,nullable)
3 circuits_schema = StructType(fields=[StructField("circuitId", IntegerType(), False),
4 StructField("circuitRef", StringType(), True),
5 StructField("name", StringType(), True),
6 StructField("location", StringType(), True),
7 StructField("country", StringType(), True),
8 StructField("lat", DoubleType(), True),
9 StructField("lng", DoubleType(), True),
10 StructField("alt", IntegerType(), True),
11 StructField("url", StringType(), True)
12
13])
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 4
1 circuits_df = spark.read.option("header",True)\
2 .schema(circuits_schema)\
3 .csv("abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv")
4
circuits_df: pyspark.sql.dataframe.DataFrame = [circuitId: integer, circuitRef: string ... 7 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 5
1 display(circuits_df)
(1) Spark Jobs

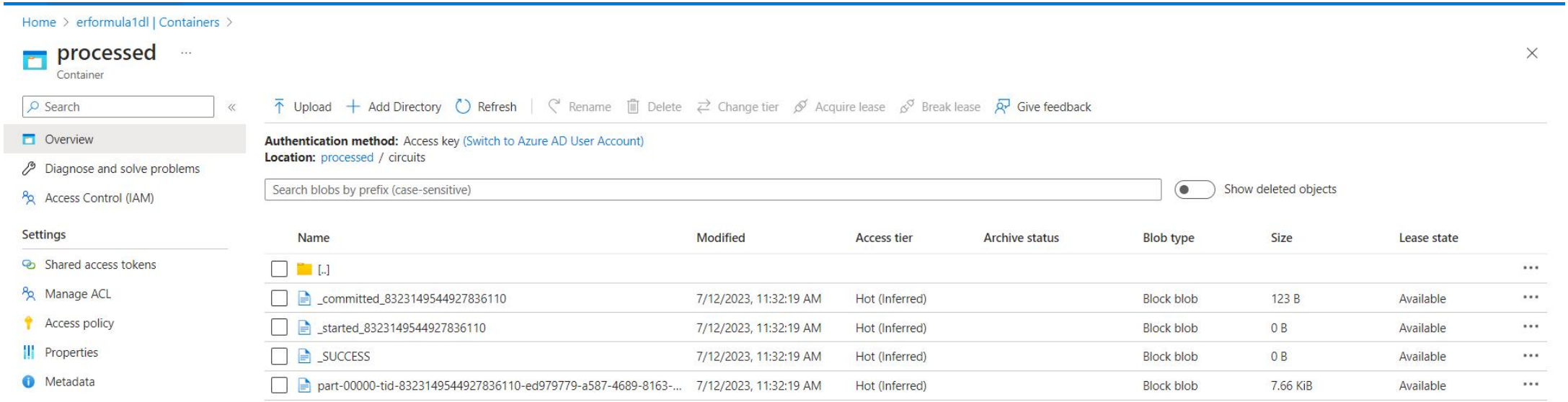

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
ingest_circuits_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Connecting...
Run all
dbkcourse_cluster
Schedule
Share
4
circuits_df: pyspark.sql.dataframe.DataFrame = [circuitId: integer, circuitRef: string ... 7 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 5
1 display(circuits_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt url
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 http://en.wikipedia.org/wiki/Sepang_International_Circuit
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 http://en.wikipedia.org/wiki/Bahrain_International_Circuit
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 http://en.wikipedia.org/wiki/Circuit_de_Barcelona-Catalunya
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 http://en.wikipedia.org/wiki/Istanbul_Park
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 http://en.wikipedia.org/wiki/Circuit_de_Monaco
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 http://en.wikipedia.org/wiki/Circuit_Gilles_Villeneuve
77 rows | 0.19 seconds runtime
Refreshed 2 minutes ago
Command took 0.19 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster
Cmd 6
Python
1 circuits_df.printSchema()
root
|-- circuitId: integer (nullable = true)
|-- circuitRef: string (nullable = true)
|-- name: string (nullable = true)
|-- location: string (nullable = true)
|-- country: string (nullable = true)
|-- lat: double (nullable = true)
|-- lng: double (nullable = true)
|-- alt: integer (nullable = true)
|-- url: string (nullable = true)
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:31:52 AM on dbkcourse_cluster

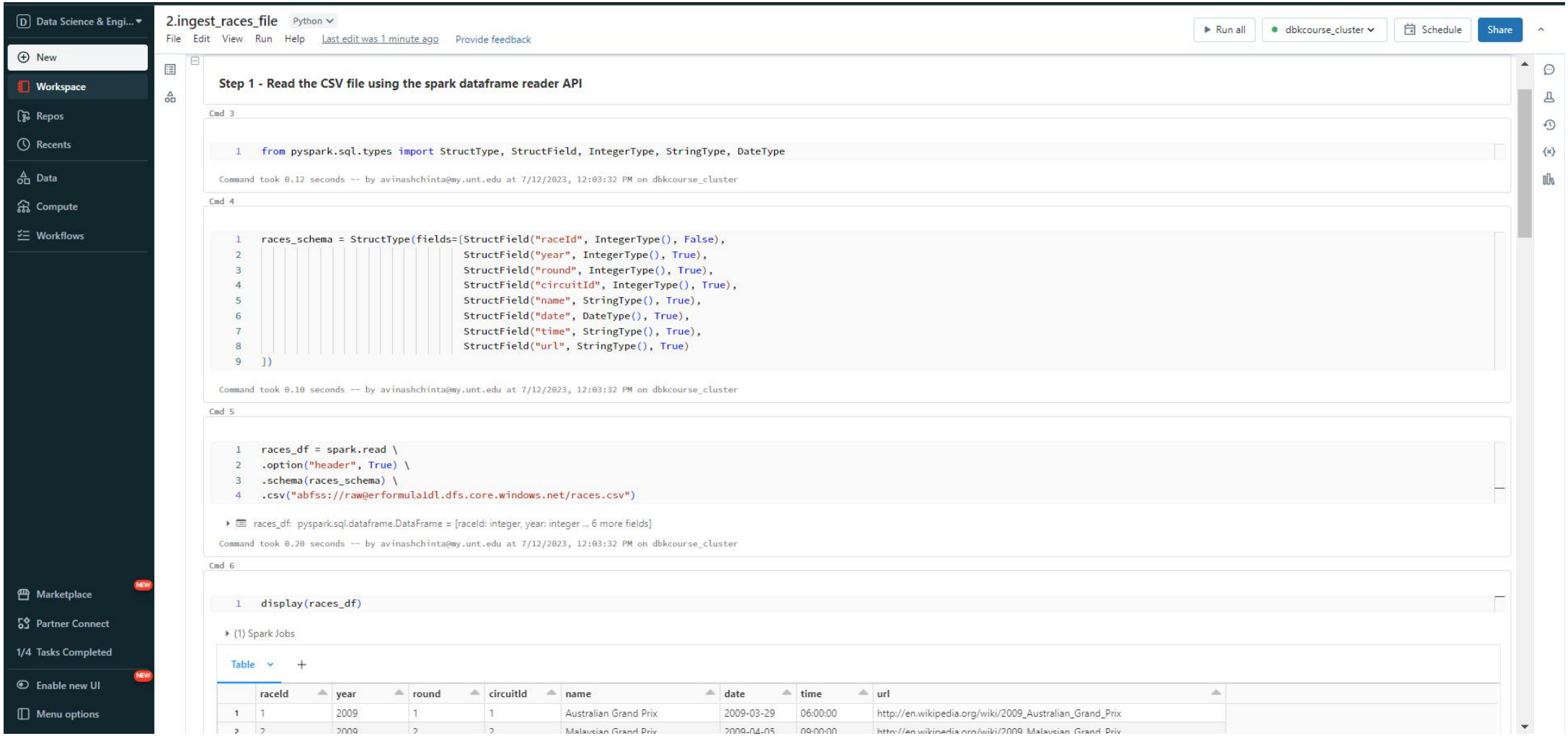

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
1 display(circuits_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt url
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 http://en.wikipedia.org/wiki/Sepang_International_Circuit
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 http://en.wikipedia.org/wiki/Bahrain_International_Circuit
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 http://en.wikipedia.org/wiki/Circuit_de_Barcelona-Catalunya
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 http://en.wikipedia.org/wiki/Istanbul_Park
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 http://en.wikipedia.org/wiki/Circuit_de_Monaco
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 http://en.wikipedia.org/wiki/Circuit_Gilles_Villeneuve
77 rows | 0.20 seconds runtime
Refreshed 7 minutes ago
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:50:35 AM on dbkcourse_cluster
Cmd 7
Step 2 - Select only the required columns
Cmd 8
1 from pyspark.sql.functions import col
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:50:35 AM on dbkcourse_cluster
Cmd 9
1 circuits_selected_df = circuits_df.select(col("circuitId"), col("circuitRef"), col("name"), col("location"), col("country"), col("lat"), col("lng"), col("alt"))
circuits_selected_df: pyspark.sql.dataframe.DataFrame = [circuitId: integer, circuitRef: string ... 6 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 10:50:35 AM on dbkcourse_cluster
Cmd 10
Python
1 display(circuits_selected_df)
(1) Spark Jobs
Table
circuitId circuitRef name location country lat lng alt
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109

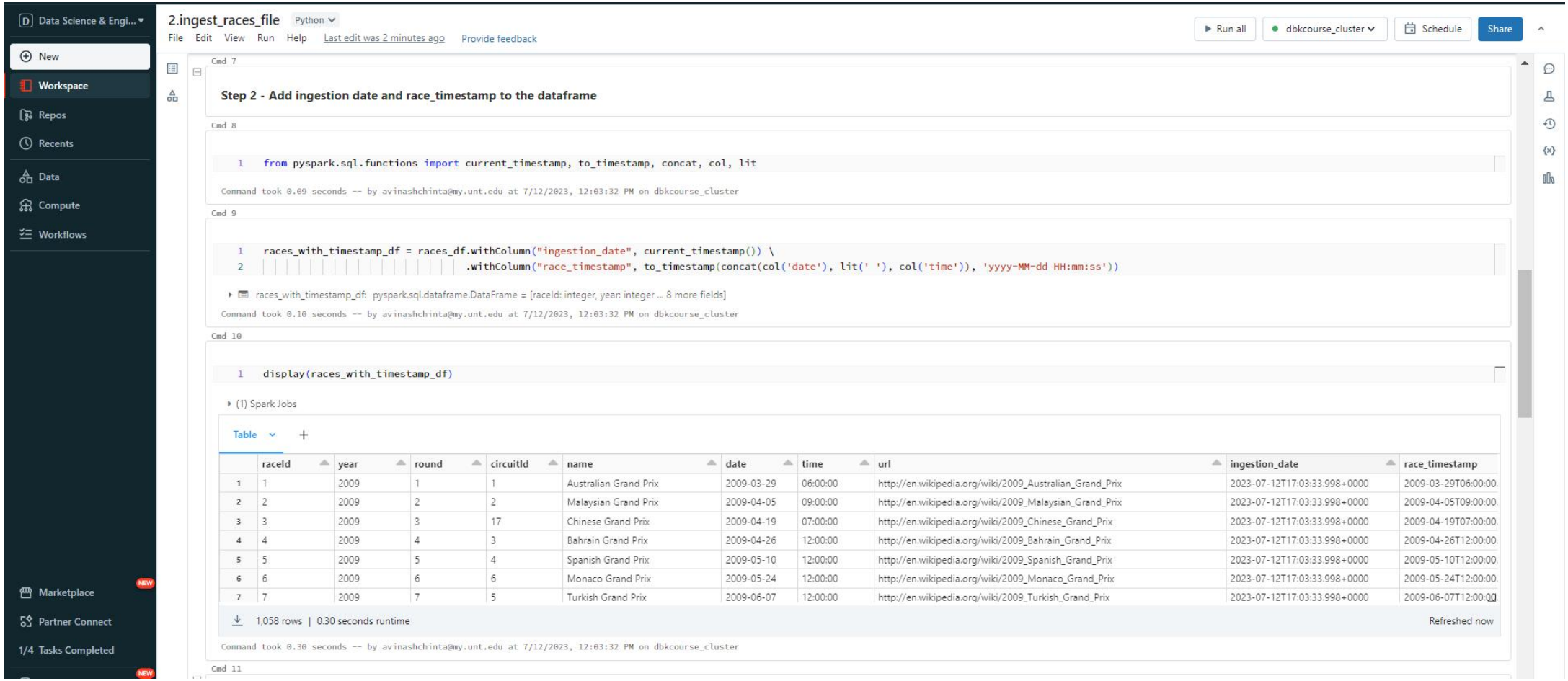

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was now
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13
77 rows | 0.20 seconds runtime
Refreshed 18 minutes ago
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:01:04 AM on dbkcourse_cluster
Cmd 12
Step 3 - Rename the columns as required
Cmd 13
circuits_renamed_df = circuits_selected_df.withColumnRenamed("circuitId", "circuit_id") \
.withColumnRenamed("circuitRef", "circuit_ref") \
.withColumnRenamed("lat", "latitude") \
.withColumnRenamed("lng", "longitude") \
.withColumnRenamed("alt", "altitude")
circuits_renamed_df: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string ... 6 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:01:04 AM on dbkcourse_cluster
Cmd 14
Python
display(circuits_renamed_df)
(1) Spark Jobs
Table
circuit_id circuit_ref name location country latitude longitude altitude
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13
77 rows | 0.42 seconds runtime
Refreshed now

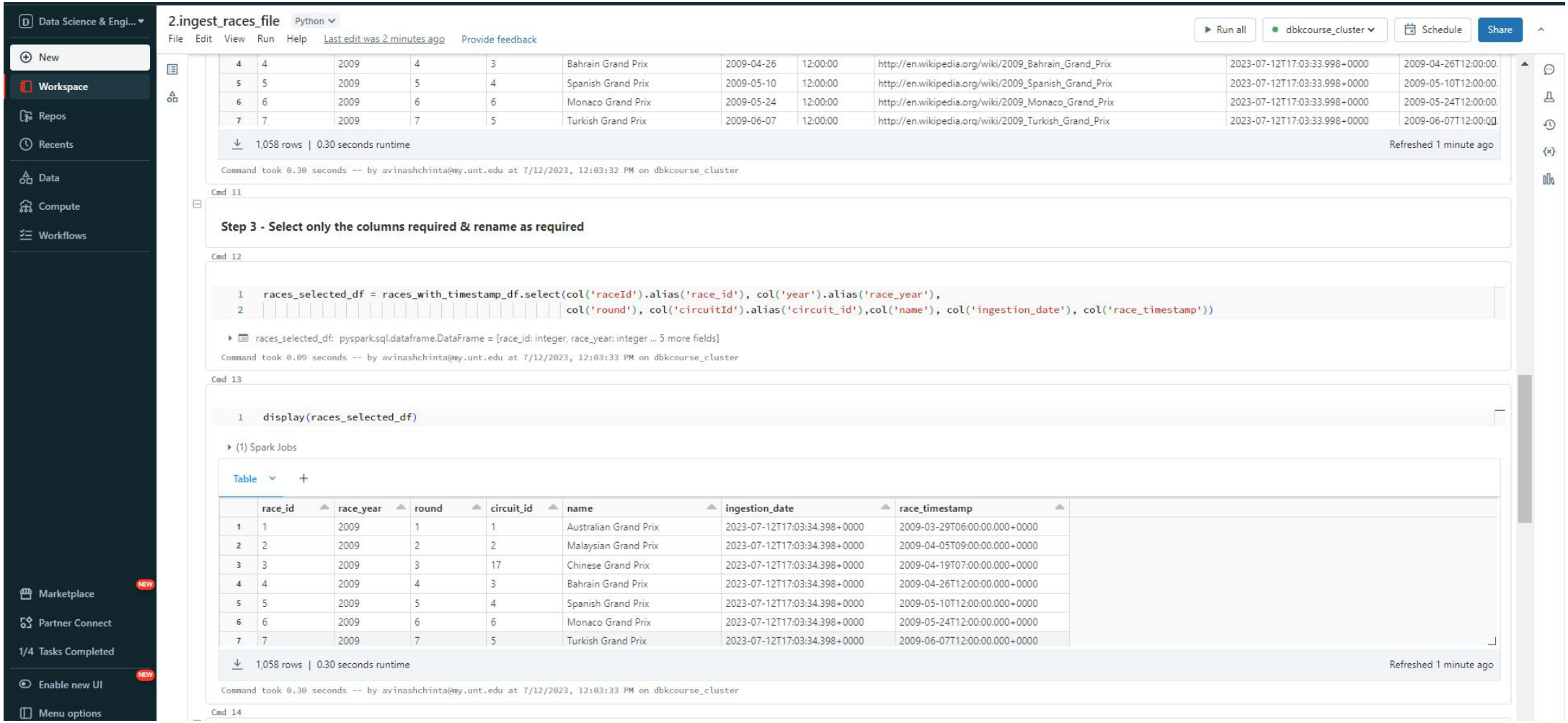

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was now
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:27:47 AM on dbkcourse_cluster
Cmd 15
Step 4 - Add ingestion date to the dataframe
Cmd 16
1 from pyspark.sql.functions import current_timestamp
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:27:47 AM on dbkcourse_cluster
Cmd 17
1 circuits_final_df = circuits_renamed_df.withColumn("ingestion_date", current_timestamp())
circuits_final_df: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string ... 7 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:27:47 AM on dbkcourse_cluster
Cmd 18
Python
1 display(circuits_final_df)
(1) Spark Jobs
Table
circuit_id circuit_ref name location country latitude longitude altitude ingestion_date
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 2023-07-12T16:27:49.785+0000
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 2023-07-12T16:27:49.785+0000
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 2023-07-12T16:27:49.785+0000
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 2023-07-12T16:27:49.785+0000
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 2023-07-12T16:27:49.785+0000
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 2023-07-12T16:27:49.785+0000
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 2023-07-12T16:27:49.785+0000
77 rows | 0.20 seconds runtime
Refreshed now
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:27:47 AM on dbkcourse_cluster

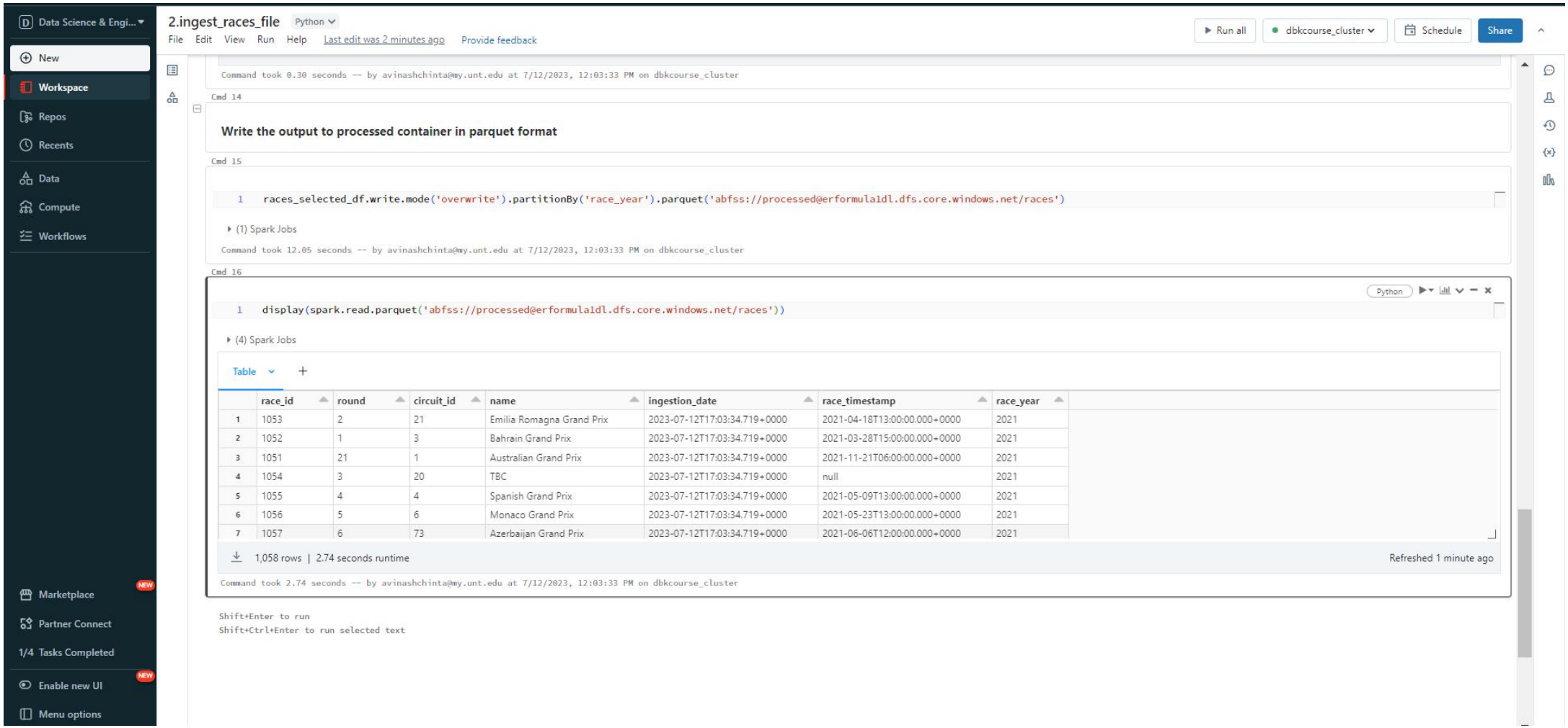

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:32:16 AM on dbkcourse_cluster
Cmd 19
Step 5 - Write data to datalake as parquet
Cmd 20
1 circuits_final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/circuits")
(1) Spark Jobs
Command took 0.60 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:32:16 AM on dbkcourse_cluster
Cmd 21
Python
1 display(spark.read.parquet("abfss://processed@erformula1dl.dfs.core.windows.net/circuits"))
(2) Spark Jobs
Table
circuit_id circuit_ref name location country latitude longitude altitude ingestion_date
1 1 albert_park Albert Park Grand Prix Circuit Melbourne Australia -37.8497 144.968 10 2023-07-12T16:32:19.010+0000
2 2 sepang Sepang International Circuit Kuala Lumpur Malaysia 2.76083 101.738 18 2023-07-12T16:32:19.010+0000
3 3 bahrain Bahrain International Circuit Sakhir Bahrain 26.0325 50.5106 7 2023-07-12T16:32:19.010+0000
4 4 catalunya Circuit de Barcelona-Catalunya Montmeló Spain 41.57 2.26111 109 2023-07-12T16:32:19.010+0000
5 5 istanbul Istanbul Park Istanbul Turkey 40.9517 29.405 130 2023-07-12T16:32:19.010+0000
6 6 monaco Circuit de Monaco Monte-Carlo Monaco 43.7347 7.42056 7 2023-07-12T16:32:19.010+0000
7 7 villeneuve Circuit Gilles Villeneuve Montreal Canada 45.5 -73.5228 13 2023-07-12T16:32:19.010+0000
77 rows | 0.30 seconds runtime
Refreshed 5 minutes ago
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 11:32:16 AM on dbkcourse_cluster
Cmd 22
1
Shift+Enter to run

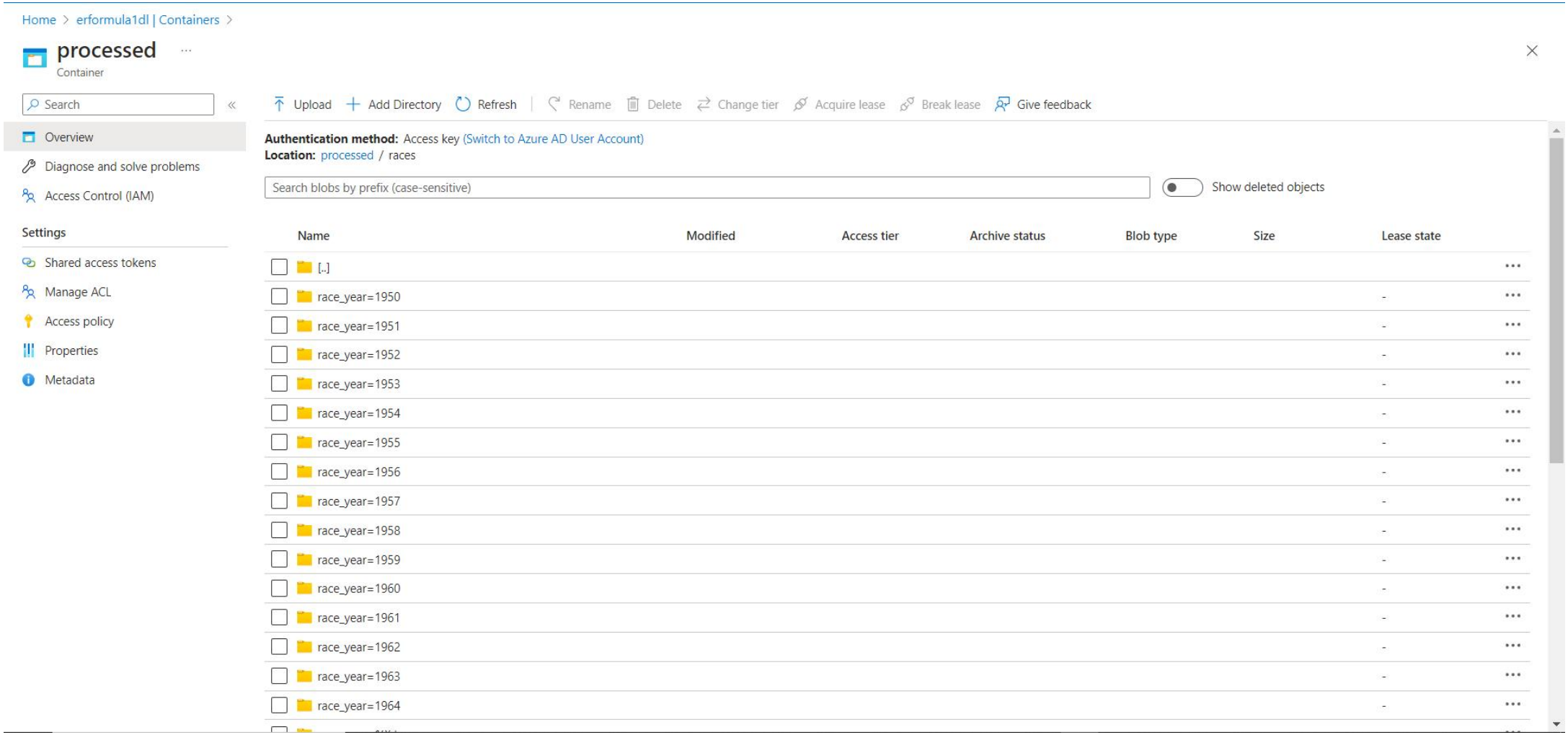

Home > erformula1dl | Containers >
processed
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / circuits
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
_committed_8323149544927836110
7/12/2023, 11:32:19 AM
Hot (Inferred)
Block blob
123 B
Available
_started_8323149544927836110
7/12/2023, 11:32:19 AM
Hot (Inferred)
Block blob
0 B
Available
_SUCCESS
7/12/2023, 11:32:19 AM
Hot (Inferred)
Block blob
0 B
Available
part-00000-tid-8323149544927836110-ed979779-a587-4689-8163-...
7/12/2023, 11:32:19 AM
Hot (Inferred)
Block blob
7.66 KiB
Available

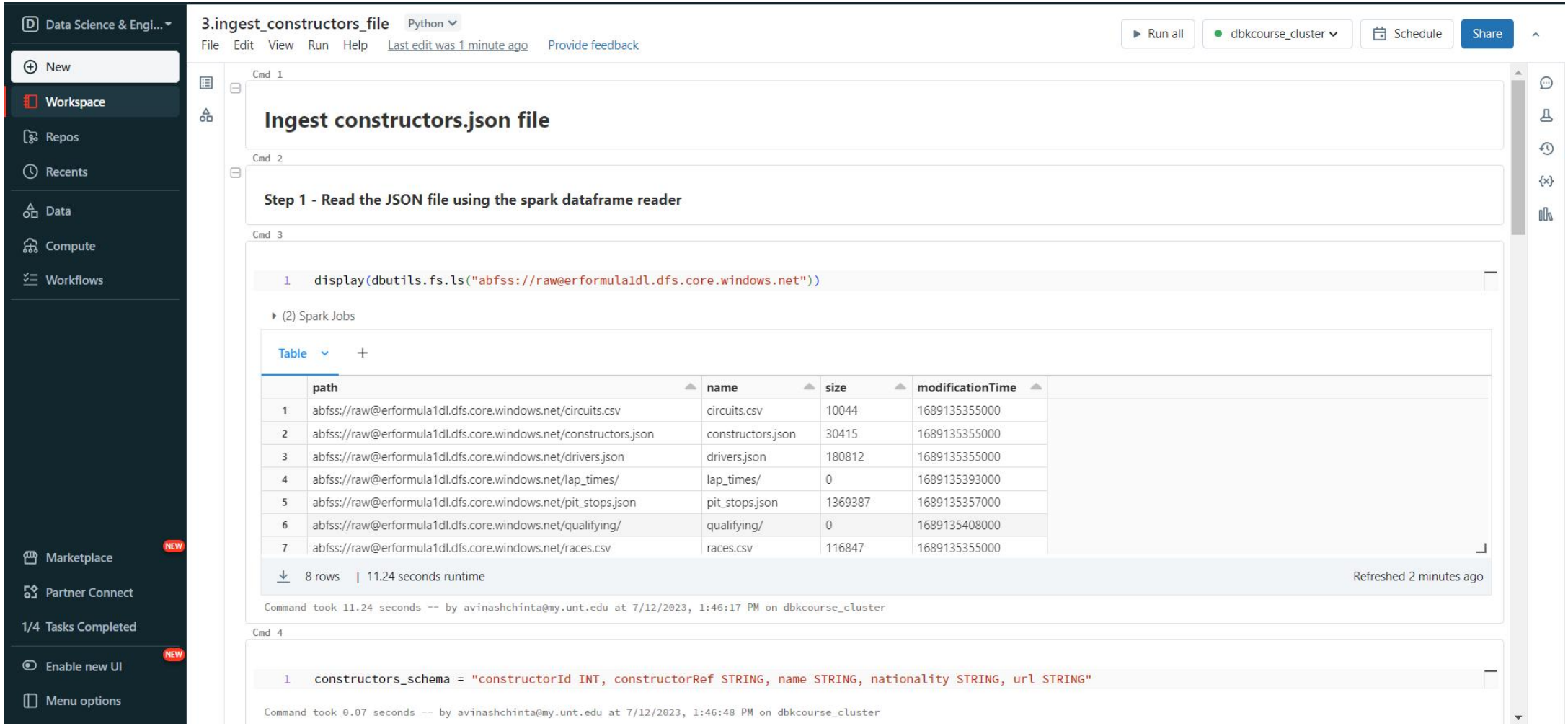

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
2.ingest_races_file
Python
File Edit View Run Help Last edit was 1 minute ago Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 1 - Read the CSV file using the spark dataframe reader API
Cmd 3
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
Command took 0.12 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 4
races_schema = StructType(fields=[StructField("raceId", IntegerType(), False),
StructField("year", IntegerType(), True),
StructField("round", IntegerType(), True),
StructField("circuitId", IntegerType(), True),
StructField("name", StringType(), True),
StructField("date", DateType(), True),
StructField("time", StringType(), True),
StructField("url", StringType(), True)
])
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 5
races_df = spark.read \
.option("header", True) \
.schema(races_schema) \
.csv("abfss://raw@erformula1dl.dfs.core.windows.net/races.csv")
races_df: pyspark.sql.dataframe.DataFrame = [raceId: integer, year: integer ... 6 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 6
display(races_df)
(1) Spark Jobs
Table
raceId year round circuitId name date time url
1 1 2009 1 1 Australian Grand Prix 2009-03-29 06:00:00 http://en.wikipedia.org/wiki/2009_Australian_Grand_Prix
2 2 2009 2 2 Malaysian Grand Prix 2009-04-05 09:00:00 http://en.wikipedia.org/wiki/2009_Malaysian_Grand_Prix

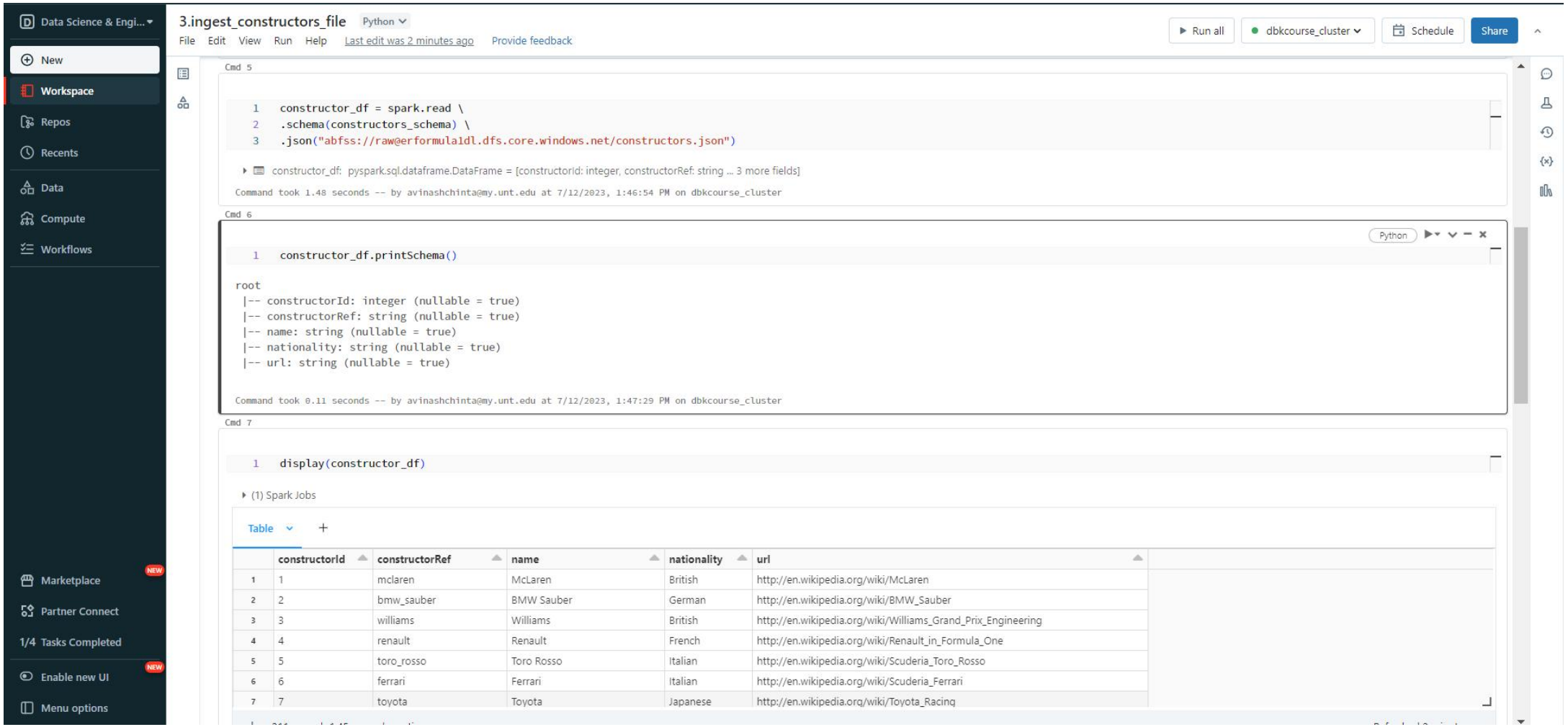

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
2.ingest_races_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 7
Step 2 - Add ingestion date and race_timestamp to the dataframe
Cmd 8
from pyspark.sql.functions import current_timestamp, to_timestamp, concat, col, lit
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 9
races_with_timestamp_df = races_df.withColumn("ingestion_date", current_timestamp()) \
.withColumn("race_timestamp", to_timestamp(concat(col('date'), lit(' '), col('time')), 'yyyy-MM-dd HH:mm:ss'))
races_with_timestamp_df: pyspark.sql.dataframe.DataFrame = [raceId: integer, year: integer ... 8 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 10
display(races_with_timestamp_df)
(1) Spark Jobs
Table
raceId year round circuitId name date time url ingestion_date race_timestamp
1 1 2009 1 1 Australian Grand Prix 2009-03-29 06:00:00 http://en.wikipedia.org/wiki/2009_Australian_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-03-29T06:00:00.
2 2 2009 2 2 Malaysian Grand Prix 2009-04-05 09:00:00 http://en.wikipedia.org/wiki/2009_Malaysian_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-04-05T09:00:00.
3 3 2009 3 17 Chinese Grand Prix 2009-04-19 07:00:00 http://en.wikipedia.org/wiki/2009_Chinese_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-04-19T07:00:00.
4 4 2009 4 3 Bahrain Grand Prix 2009-04-26 12:00:00 http://en.wikipedia.org/wiki/2009_Bahrain_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-04-26T12:00:00.
5 5 2009 5 4 Spanish Grand Prix 2009-05-10 12:00:00 http://en.wikipedia.org/wiki/2009_Spanish_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-05-10T12:00:00.
6 6 2009 6 6 Monaco Grand Prix 2009-05-24 12:00:00 http://en.wikipedia.org/wiki/2009_Monaco_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-05-24T12:00:00.
7 7 2009 7 5 Turkish Grand Prix 2009-06-07 12:00:00 http://en.wikipedia.org/wiki/2009_Turkish_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-06-07T12:00:00
1,058 rows | 0.30 seconds runtime
Refreshed now
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 11

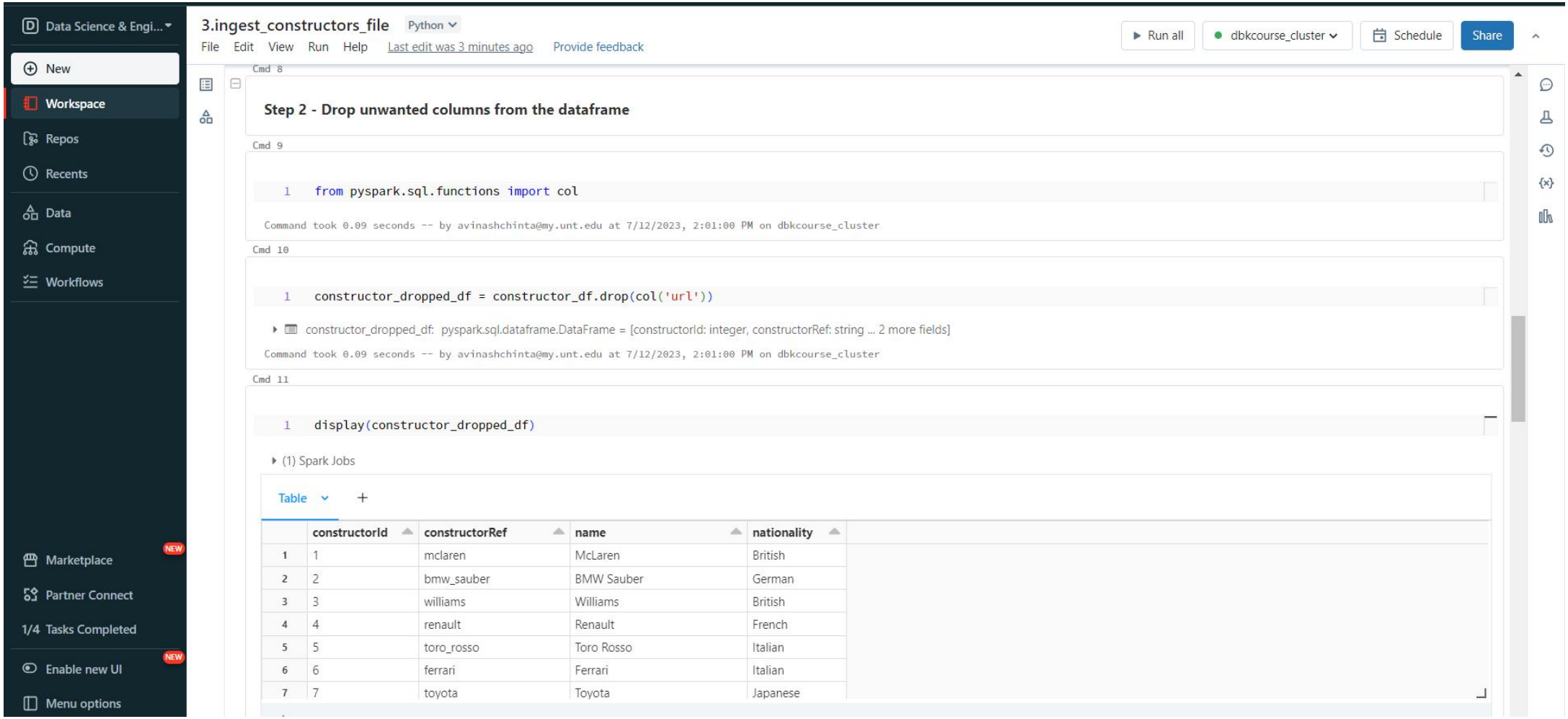

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
2.ingest_races_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
4 4 2009 4 3 Bahrain Grand Prix 2009-04-26 12:00:00 http://en.wikipedia.org/wiki/2009_Bahrain_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-04-26T12:00:00.
5 5 2009 5 4 Spanish Grand Prix 2009-05-10 12:00:00 http://en.wikipedia.org/wiki/2009_Spanish_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-05-10T12:00:00.
6 6 2009 6 6 Monaco Grand Prix 2009-05-24 12:00:00 http://en.wikipedia.org/wiki/2009_Monaco_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-05-24T12:00:00.
7 7 2009 7 5 Turkish Grand Prix 2009-06-07 12:00:00 http://en.wikipedia.org/wiki/2009_Turkish_Grand_Prix 2023-07-12T17:03:33.998+0000 2009-06-07T12:00:00.
1,058 rows | 0.30 seconds runtime
Refreshed 1 minute ago
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:32 PM on dbkcourse_cluster
Cmd 11
Step 3 - Select only the columns required & rename as required
Cmd 12
races_selected_df = races_with_timestamp_df.select(col('raceId').alias('race_id'), col('year').alias('race_year'),
col('round'), col('circuitId').alias('circuit_id'),col('name'), col('ingestion_date'), col('race_timestamp'))
races_selected_df: pyspark.sql.dataframe.DataFrame = [race_id: integer, race_year: integer ... 5 more fields]
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:33 PM on dbkcourse_cluster
Cmd 13
display(races_selected_df)
(1) Spark Jobs
Table
race_id race_year round circuit_id name ingestion_date race_timestamp
1 1 2009 1 1 Australian Grand Prix 2023-07-12T17:03:34.398+0000 2009-03-29T06:00:00.000+0000
2 2 2009 2 2 Malaysian Grand Prix 2023-07-12T17:03:34.398+0000 2009-04-05T09:00:00.000+0000
3 3 2009 3 17 Chinese Grand Prix 2023-07-12T17:03:34.398+0000 2009-04-19T07:00:00.000+0000
4 4 2009 4 3 Bahrain Grand Prix 2023-07-12T17:03:34.398+0000 2009-04-26T12:00:00.000+0000
5 5 2009 5 4 Spanish Grand Prix 2023-07-12T17:03:34.398+0000 2009-05-10T12:00:00.000+0000
6 6 2009 6 6 Monaco Grand Prix 2023-07-12T17:03:34.398+0000 2009-05-24T12:00:00.000+0000
7 7 2009 7 5 Turkish Grand Prix 2023-07-12T17:03:34.398+0000 2009-06-07T12:00:00.000+0000
1,058 rows | 0.30 seconds runtime
Refreshed 1 minute ago
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:33 PM on dbkcourse_cluster
Cmd 14

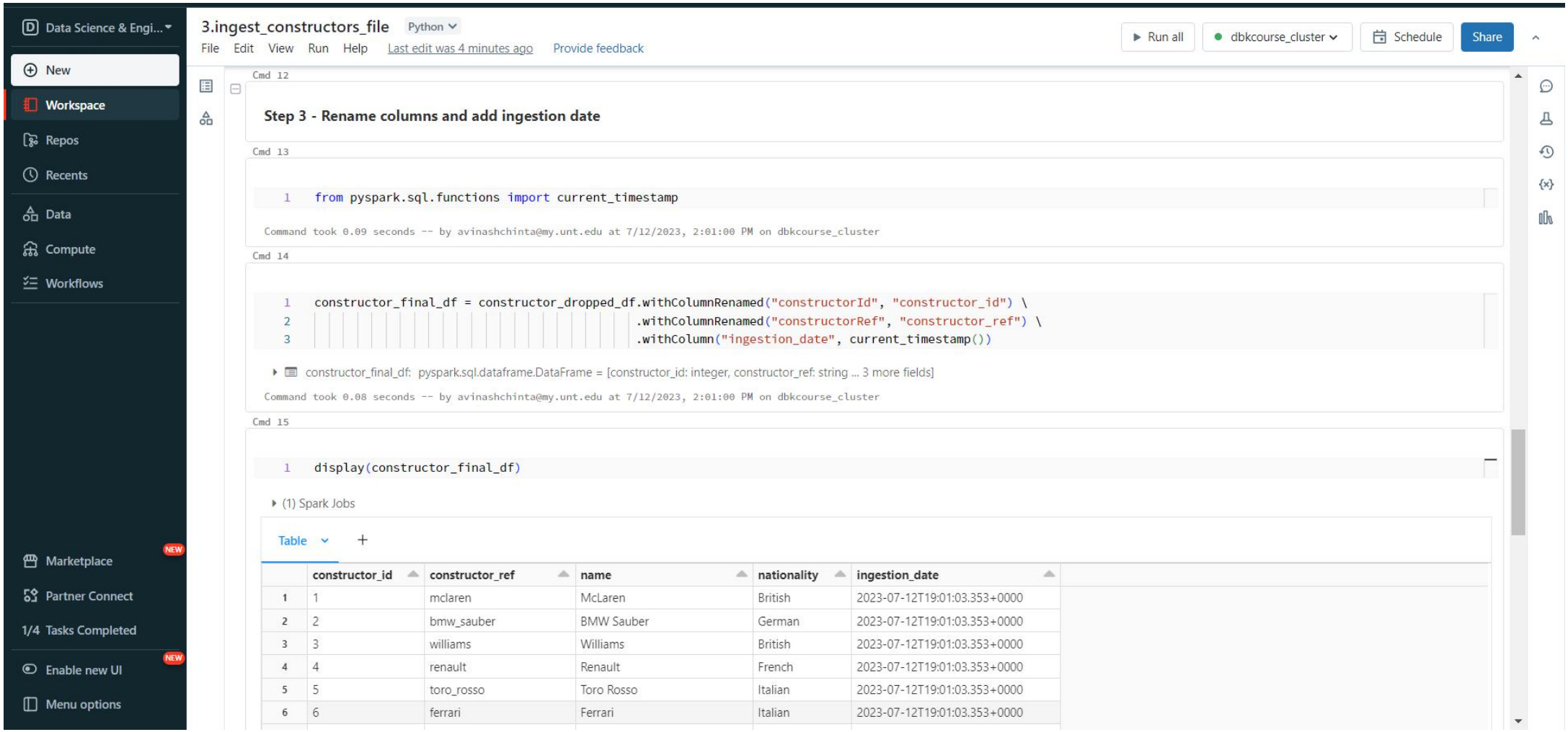

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
2.ingest_races_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:33 PM on dbkcourse_cluster
Cmd 14
Write the output to processed container in parquet format
Cmd 15
1 races_selected_df.write.mode('overwrite').partitionBy('race_year').parquet('abfss://processed@erformula1dl.dfs.core.windows.net/races')
(1) Spark Jobs
Command took 12.05 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:33 PM on dbkcourse_cluster
Cmd 16
Python
1 display(spark.read.parquet('abfss://processed@erformula1dl.dfs.core.windows.net/races'))
(4) Spark Jobs
Table
race_id round circuit_id name ingestion_date race_timestamp race_year
1 1053 2 21 Emilia Romagna Grand Prix 2023-07-12T17:03:34.719+0000 2021-04-18T13:00:00.000+0000 2021
2 1052 1 3 Bahrain Grand Prix 2023-07-12T17:03:34.719+0000 2021-03-28T15:00:00.000+0000 2021
3 1051 21 1 Australian Grand Prix 2023-07-12T17:03:34.719+0000 2021-11-21T06:00:00.000+0000 2021
4 1054 3 20 TBC 2023-07-12T17:03:34.719+0000 null 2021
5 1055 4 4 Spanish Grand Prix 2023-07-12T17:03:34.719+0000 2021-05-09T13:00:00.000+0000 2021
6 1056 5 6 Monaco Grand Prix 2023-07-12T17:03:34.719+0000 2021-05-23T13:00:00.000+0000 2021
7 1057 6 73 Azerbaijan Grand Prix 2023-07-12T17:03:34.719+0000 2021-06-06T12:00:00.000+0000 2021
1,058 rows | 2.74 seconds runtime
Refreshed 1 minute ago
Command took 2.74 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 12:03:33 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

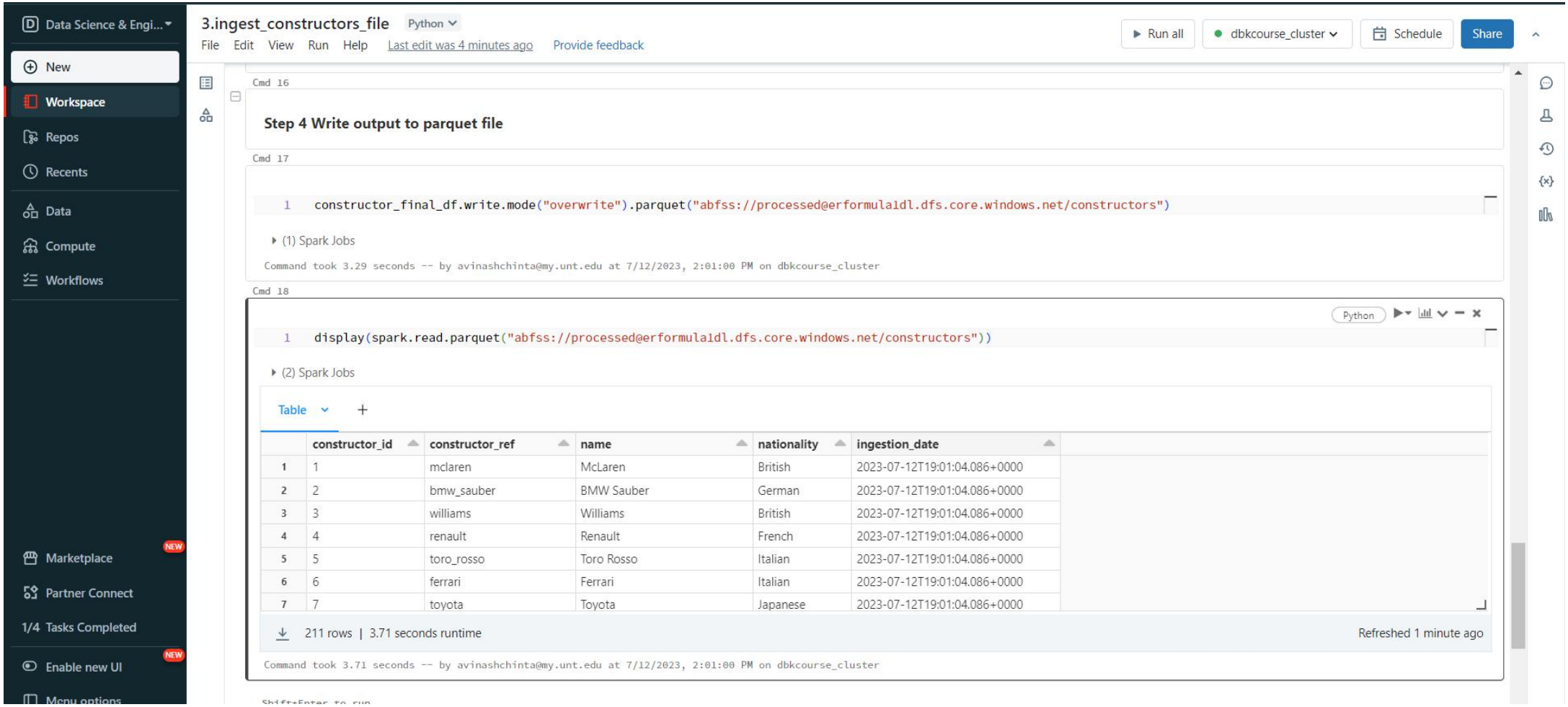

Home > erformula1dl | Containers >
processed
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / races
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
race_year=1950
race_year=1951
race_year=1952
race_year=1953
race_year=1954
race_year=1955
race_year=1956
race_year=1957
race_year=1958
race_year=1959
race_year=1960
race_year=1961
race_year=1962
race_year=1963
race_year=1964

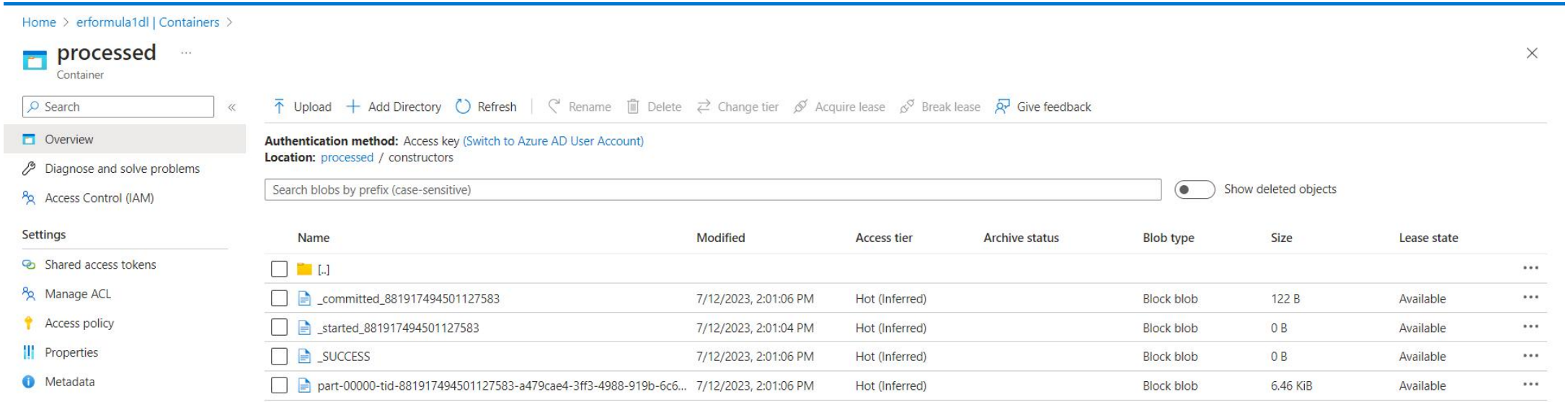

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
3.ingest_constructors_file
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 1
Ingest constructors.json file
Cmd 2
Step 1 - Read the JSON file using the spark dataframe reader
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path
name
size
modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 11.24 seconds runtime
Refreshed 2 minutes ago
Command took 11.24 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 1:46:17 PM on dbkcourse_cluster
Cmd 4
1 constructors_schema = "constructorId INT, constructorRef STRING, name STRING, nationality STRING, url STRING"
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 1:46:48 PM on dbkcourse_cluster

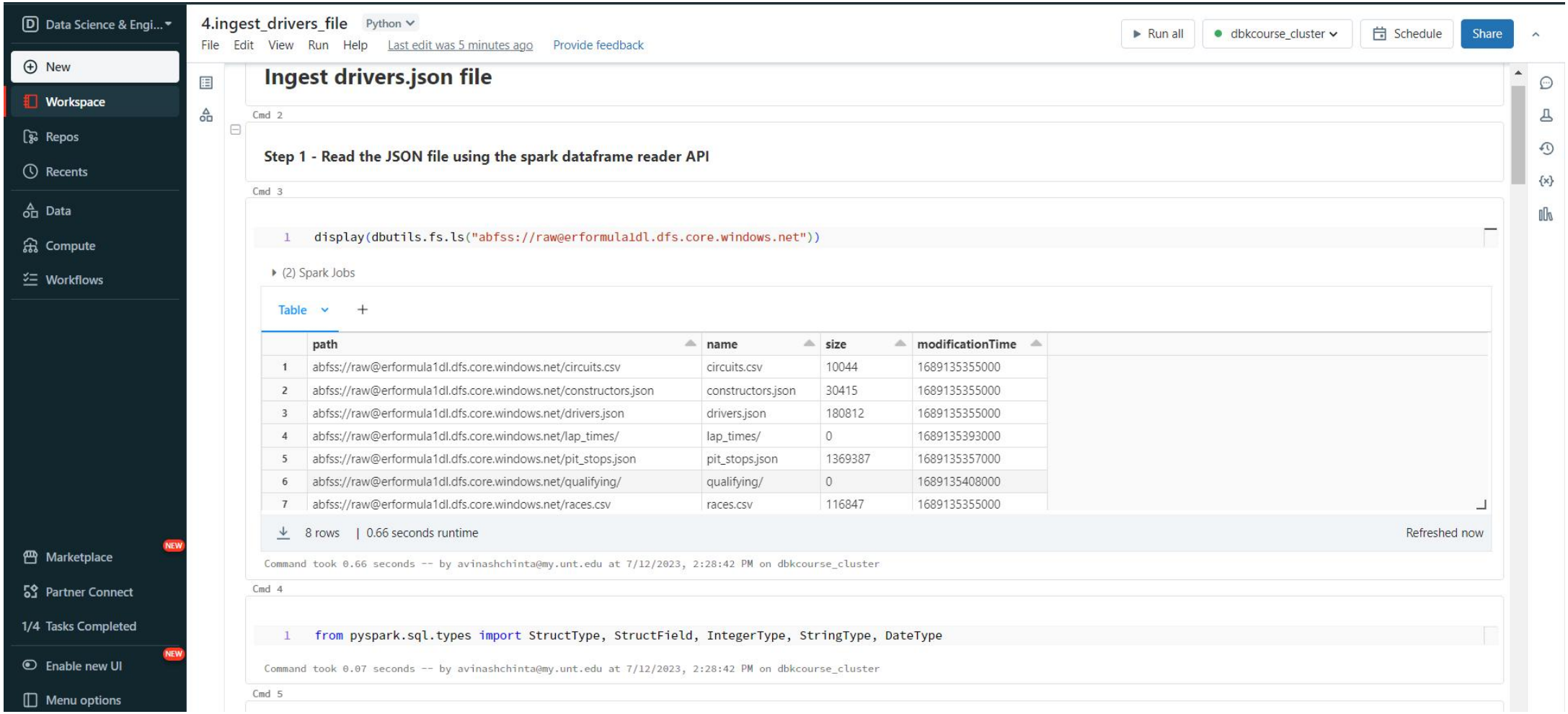

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
3.ingest_constructors_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 5
1 constructor_df = spark.read \
2 .schema(constructors_schema) \
3 .json("abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json")
constructor_df: pyspark.sql.dataframe.DataFrame = [constructorId: integer, constructorRef: string ... 3 more fields]
Command took 1.48 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 1:46:54 PM on dbkcourse_cluster
Cmd 6
Python
1 constructor_df.printSchema()
root
|-- constructorId: integer (nullable = true)
|-- constructorRef: string (nullable = true)
|-- name: string (nullable = true)
|-- nationality: string (nullable = true)
|-- url: string (nullable = true)
Command took 0.11 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 1:47:29 PM on dbkcourse_cluster
Cmd 7
1 display(constructor_df)
(1) Spark Jobs
Table
constructorId constructorRef name nationality url
1 1 mclaren McLaren British http://en.wikipedia.org/wiki/McLaren
2 2 bmw_sauber BMW Sauber German http://en.wikipedia.org/wiki/BMW_Sauber
3 3 williams Williams British http://en.wikipedia.org/wiki/Williams_Grand_Prix_Engineering
4 4 renault Renault French http://en.wikipedia.org/wiki/Renault_in_Formula_One
5 5 toro_rosso Toro Rosso Italian http://en.wikipedia.org/wiki/Scuderia_Toro_Rosso
6 6 ferrari Ferrari Italian http://en.wikipedia.org/wiki/Scuderia_Ferrari
7 7 toyota Toyota Japanese http://en.wikipedia.org/wiki/Toyota_Racing

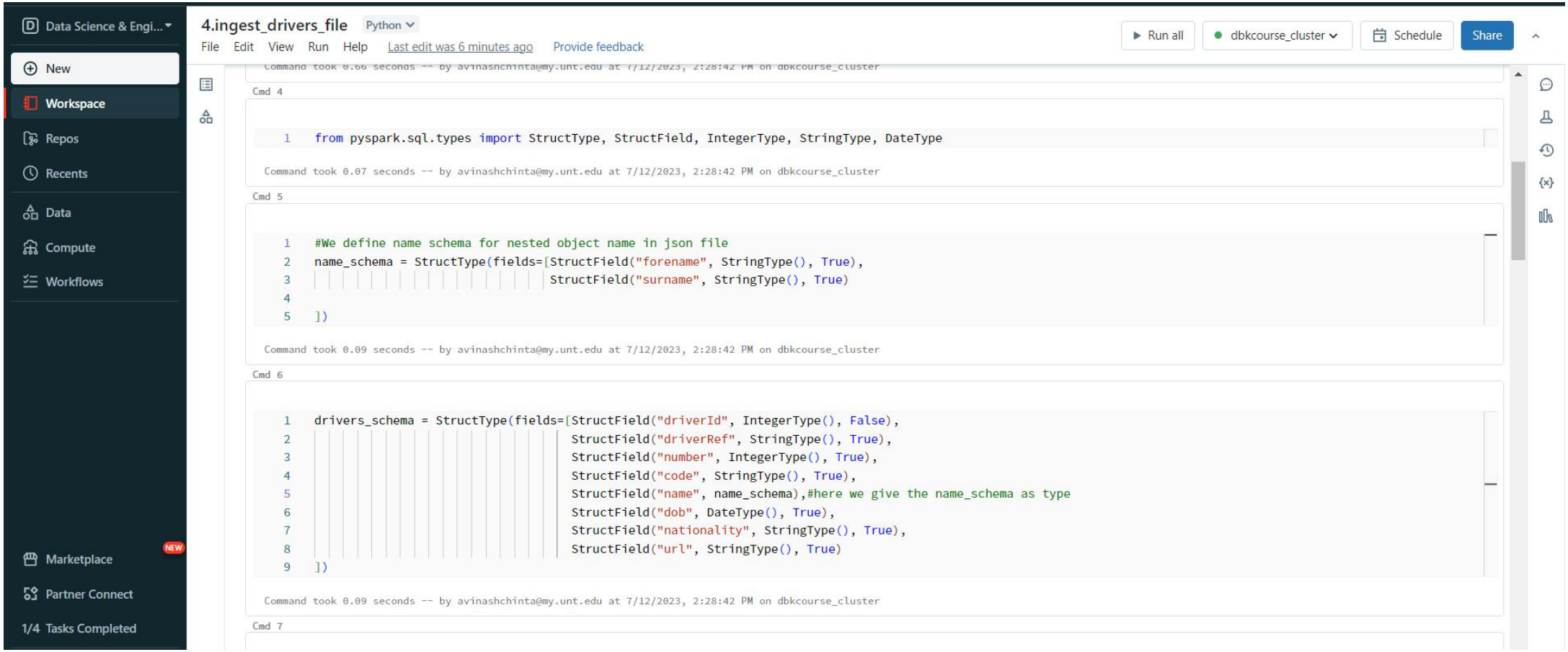

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
3.ingest_constructors_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 8
Step 2 - Drop unwanted columns from the dataframe
Cmd 9
1 from pyspark.sql.functions import col
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster
Cmd 10
1 constructor_dropped_df = constructor_df.drop(col('url'))
constructor_dropped_df: pyspark.sql.dataframe.DataFrame = [constructorId: integer, constructorRef: string ... 2 more fields]
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster
Cmd 11
1 display(constructor_dropped_df)
(1) Spark Jobs
Table
constructorId constructorRef name nationality
1 1 mclaren McLaren British
2 2 bmw_sauber BMW Sauber German
3 3 williams Williams British
4 4 renault Renault French
5 5 toro_rosso Toro Rosso Italian
6 6 ferrari Ferrari Italian
7 7 toyota Toyota Japanese

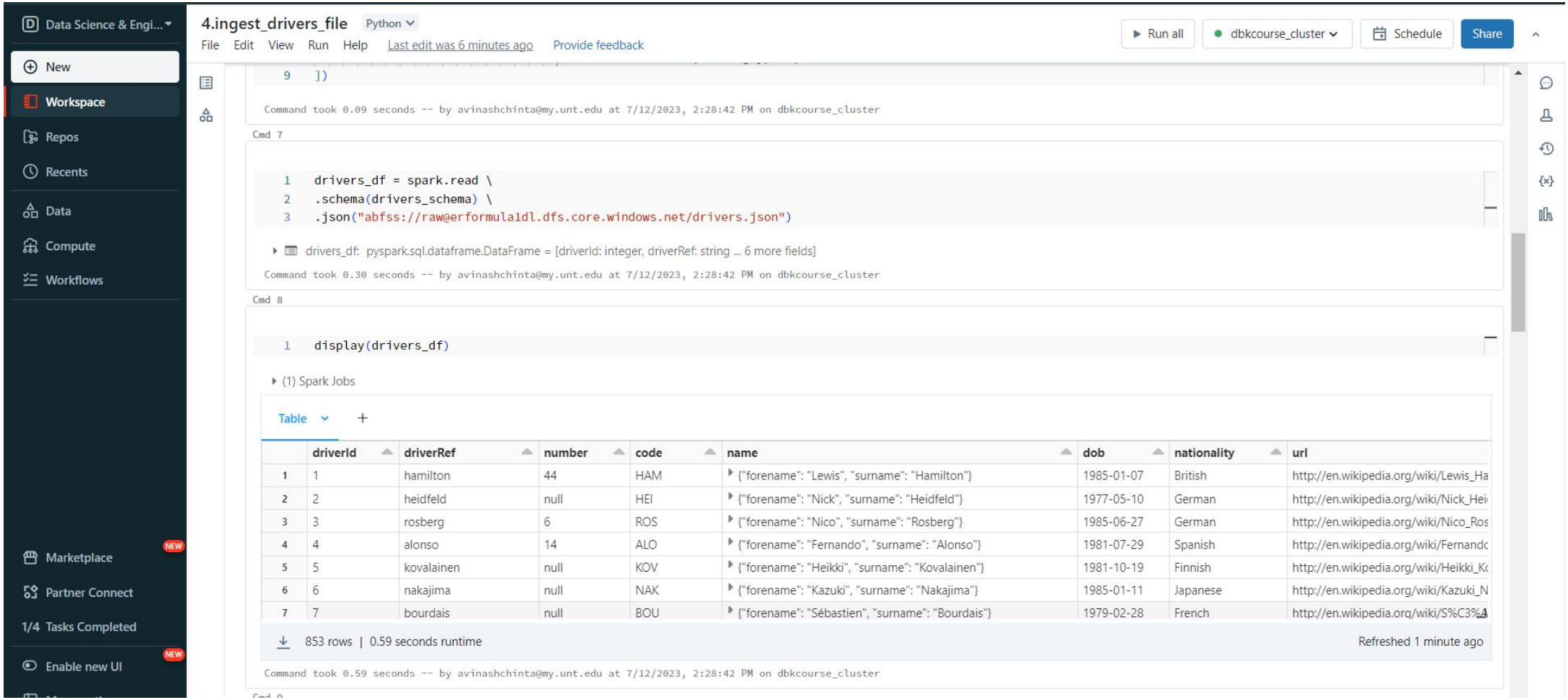

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
3.ingest_constructors_file
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 12
Step 3 - Rename columns and add ingestion date
Cmd 13
from pyspark.sql.functions import current_timestamp
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster
Cmd 14
constructor_final_df = constructor_dropped_df.withColumnRenamed("constructorId", "constructor_id") \
.withColumnRenamed("constructorRef", "constructor_ref") \
.withColumn("ingestion_date", current_timestamp())
constructor_final_df: pyspark.sql.dataframe.DataFrame = [constructor_id: integer, constructor_ref: string ... 3 more fields]
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster
Cmd 15
display(constructor_final_df)
(1) Spark Jobs
Table
constructor_id
constructor_ref
name
nationality
ingestion_date
1 1 mclaren McLaren British 2023-07-12T19:01:03.353+0000
2 2 bmw_sauber BMW Sauber German 2023-07-12T19:01:03.353+0000
3 3 williams Williams British 2023-07-12T19:01:03.353+0000
4 4 renault Renault French 2023-07-12T19:01:03.353+0000
5 5 toro_rosso Toro Rosso Italian 2023-07-12T19:01:03.353+0000
6 6 ferrari Ferrari Italian 2023-07-12T19:01:03.353+0000

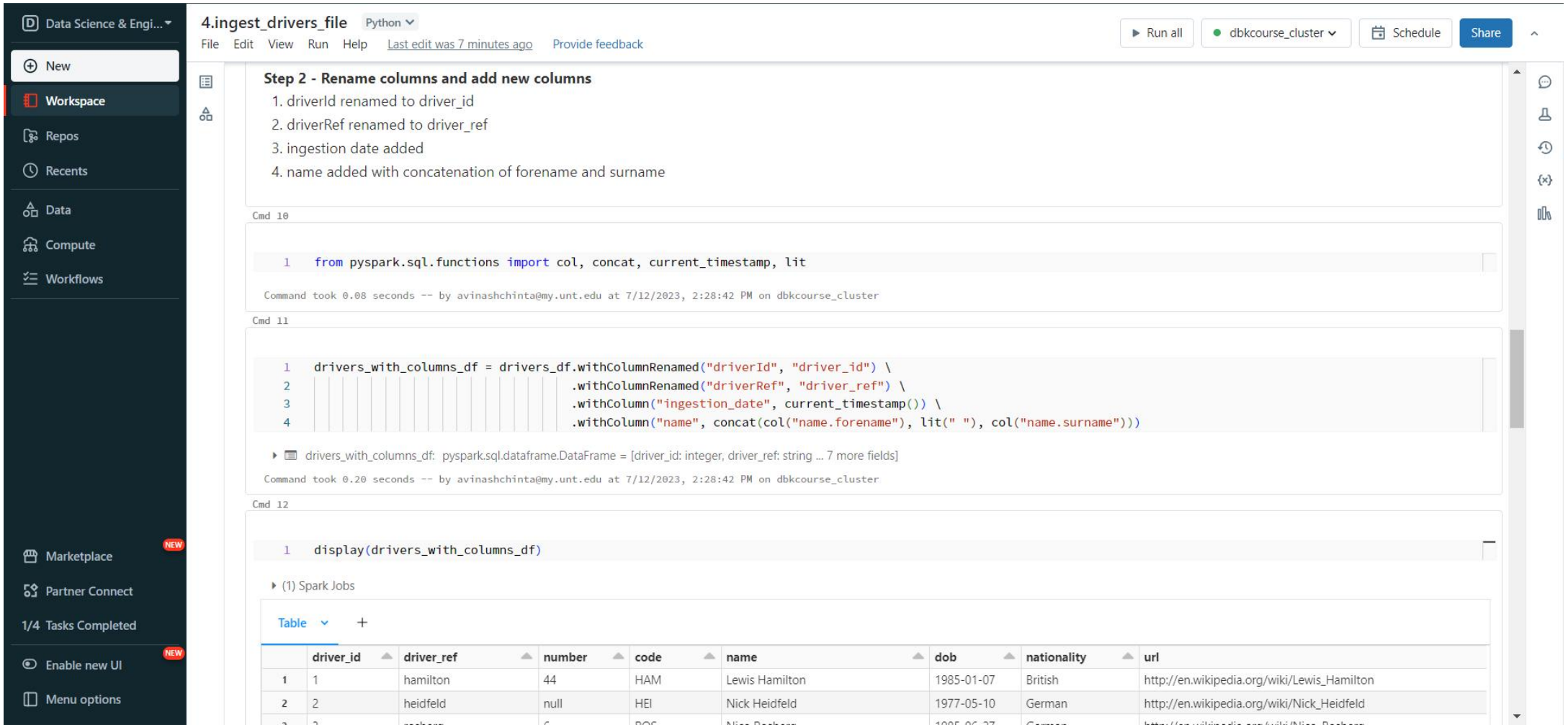

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
3.ingest_constructors_file
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 16
Step 4 Write output to parquet file
Cmd 17
1 constructor_final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/constructors")
(1) Spark Jobs
Command took 3.29 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster
Cmd 18
Python
1 display(spark.read.parquet("abfss://processed@erformula1dl.dfs.core.windows.net/constructors"))
(2) Spark Jobs
Table
constructor_id constructor_ref name nationality ingestion_date
1 1 mclaren McLaren British 2023-07-12T19:01:04.086+0000
2 2 bmw_sauber BMW Sauber German 2023-07-12T19:01:04.086+0000
3 3 williams Williams British 2023-07-12T19:01:04.086+0000
4 4 renault Renault French 2023-07-12T19:01:04.086+0000
5 5 toro_rosso Toro Rosso Italian 2023-07-12T19:01:04.086+0000
6 6 ferrari Ferrari Italian 2023-07-12T19:01:04.086+0000
7 7 toyota Toyota Japanese 2023-07-12T19:01:04.086+0000
211 rows | 3.71 seconds runtime
Refreshed 1 minute ago
Command took 3.71 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:01:00 PM on dbkcourse_cluster

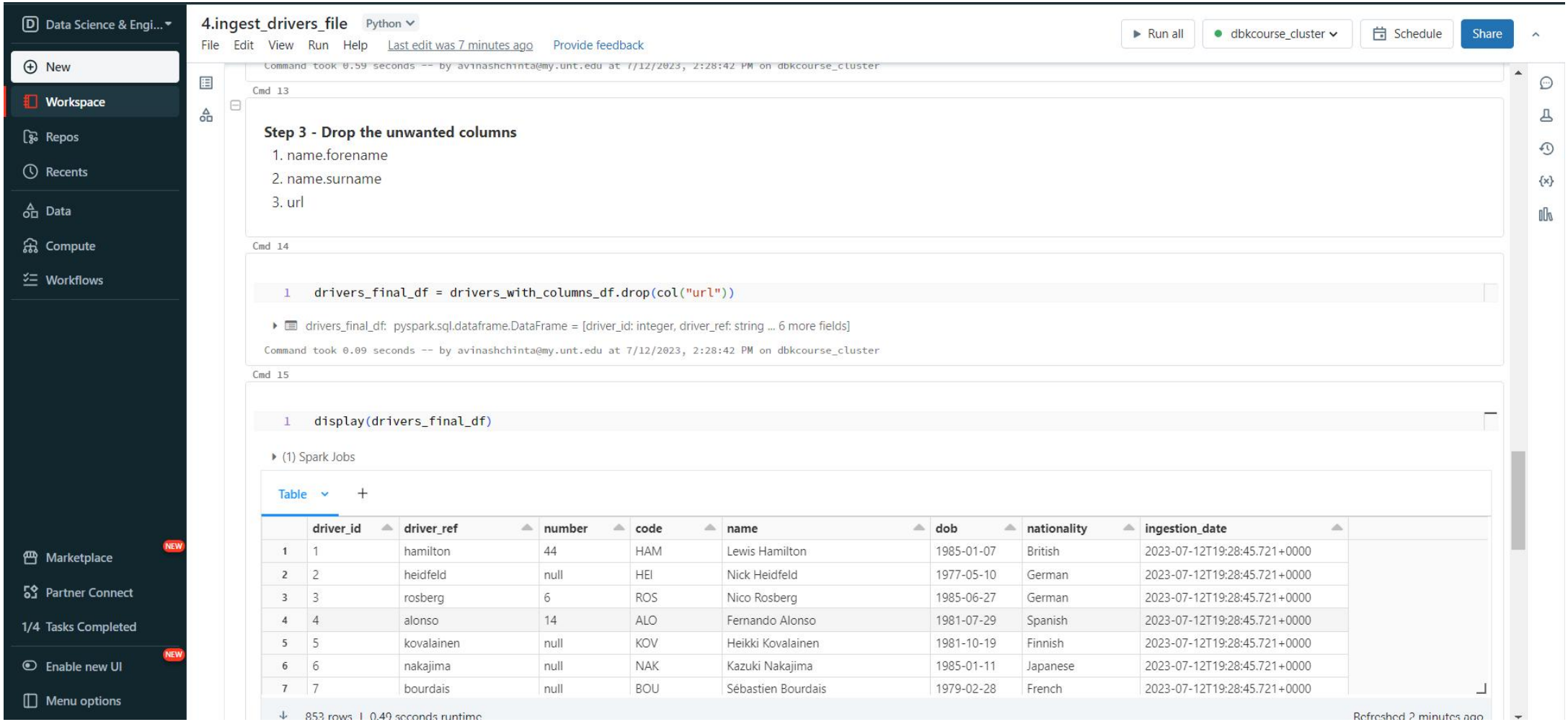

Home > erformula1dl | Containers >
processed
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / constructors
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
_committed_881917494501127583
7/12/2023, 2:01:06 PM
Hot (Inferred)
Block blob
122 B
Available
_started_881917494501127583
7/12/2023, 2:01:04 PM
Hot (Inferred)
Block blob
0 B
Available
_SUCCESS
7/12/2023, 2:01:06 PM
Hot (Inferred)
Block blob
0 B
Available
part-00000-tid-881917494501127583-a479cae4-3ff3-4988-919b-6c6...
7/12/2023, 2:01:06 PM
Hot (Inferred)
Block blob
6.46 KiB
Available

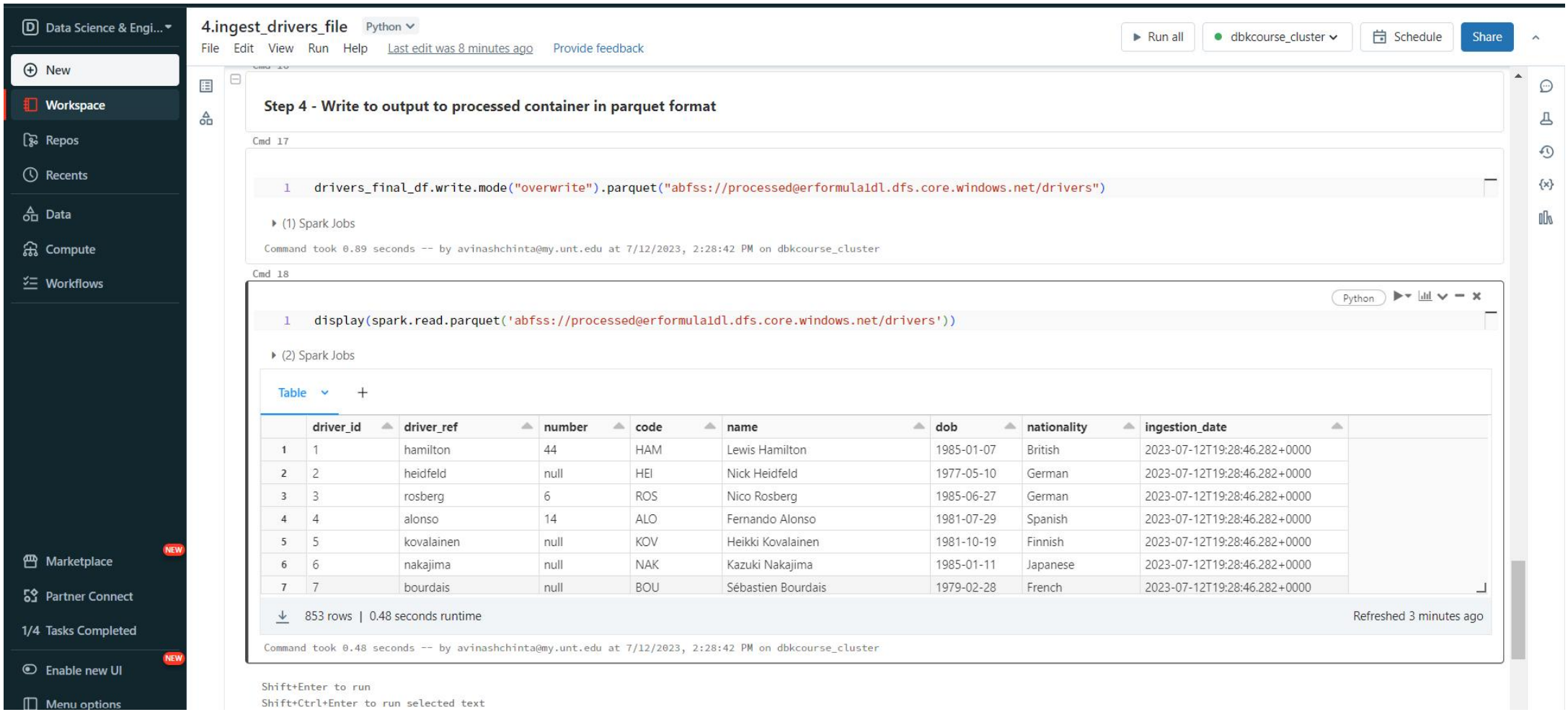

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Ingest drivers.json file
Cmd 2
Step 1 - Read the JSON file using the spark dataframe reader API
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path
name
size
modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 0.66 seconds runtime
Refreshed now
Command took 0.66 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 4
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 5

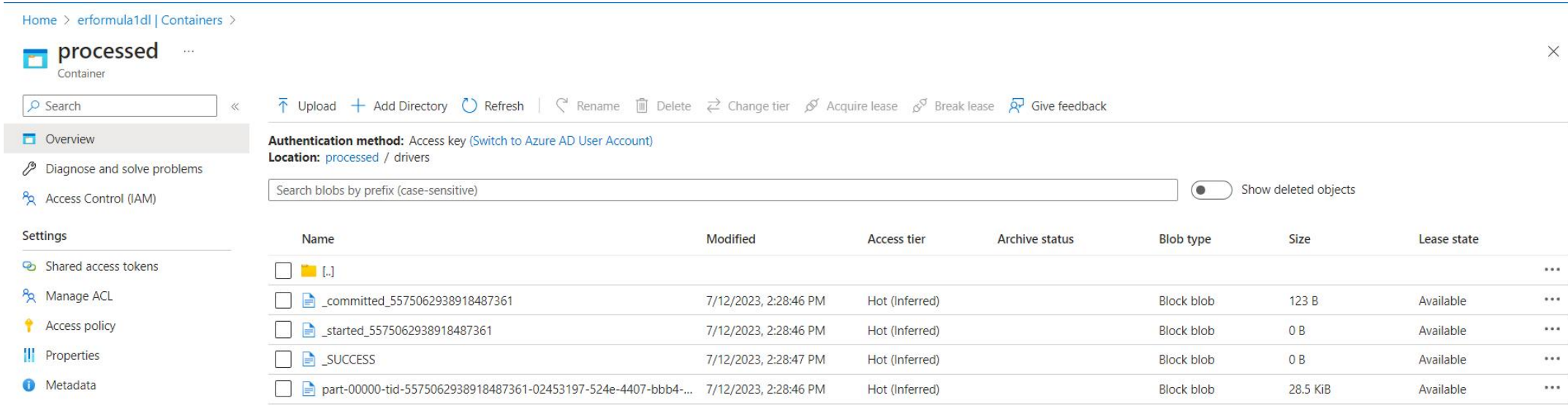

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
Partner Connect
1/4 Tasks Completed
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.66 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 4
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DateType
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 5
1 #We define name schema for nested object name in json file
2 name_schema = StructType(fields=[StructField("forename", StringType(), True),
3 StructField("surname", StringType(), True)
4
5])
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 6
1 drivers_schema = StructType(fields=[StructField("driverId", IntegerType(), False),
2 StructField("driverRef", StringType(), True),
3 StructField("number", IntegerType(), True),
4 StructField("code", StringType(), True),
5 StructField("name", name_schema),#here we give the name_schema as type
6 StructField("dob", DateType(), True),
7 StructField("nationality", StringType(), True),
8 StructField("url", StringType(), True)
9])
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 7

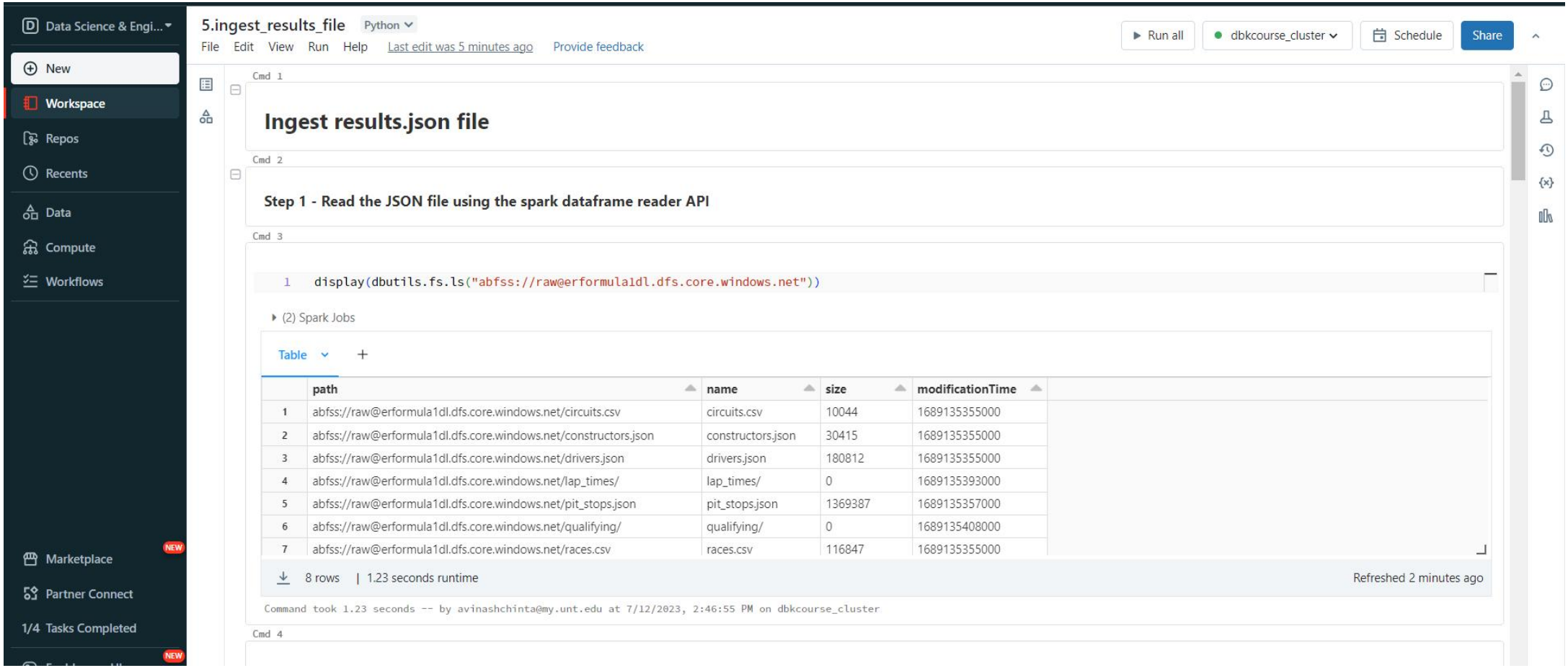

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
])
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 7
drivers_df = spark.read \
.schema(drivers_schema) \
.json("abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json")
drivers_df: pyspark.sql.dataframe.DataFrame = [driverId: integer, driverRef: string ... 6 more fields]
Command took 0.30 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 8
display(drivers_df)
(1) Spark Jobs
Table
driverId driverRef number code name dob nationality url
1 1 hamilton 44 HAM {"forename": "Lewis", "surname": "Hamilton"} 1985-01-07 British http://en.wikipedia.org/wiki/Lewis_Ha
2 2 heidfeld null HEI {"forename": "Nick", "surname": "Heidfeld"} 1977-05-10 German http://en.wikipedia.org/wiki/Nick_Hei
3 3 rosberg 6 ROS {"forename": "Nico", "surname": "Rosberg"} 1985-06-27 German http://en.wikipedia.org/wiki/Nico_Ros
4 4 alonso 14 ALO {"forename": "Fernando", "surname": "Alonso"} 1981-07-29 Spanish http://en.wikipedia.org/wiki/Fernandc
5 5 kovalainen null KOV {"forename": "Heikki", "surname": "Kovalainen"} 1981-10-19 Finnish http://en.wikipedia.org/wiki/Heikki_Kc
6 6 nakajima null NAK {"forename": "Kazuki", "surname": "Nakajima"} 1985-01-11 Japanese http://en.wikipedia.org/wiki/Kazuki_N
7 7 bourdais null BOU {"forename": "Sébastien", "surname": "Bourdais"} 1979-02-28 French http://en.wikipedia.org/wiki/S%C3%A
853 rows | 0.59 seconds runtime
Refreshed 1 minute ago
Command took 0.59 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster

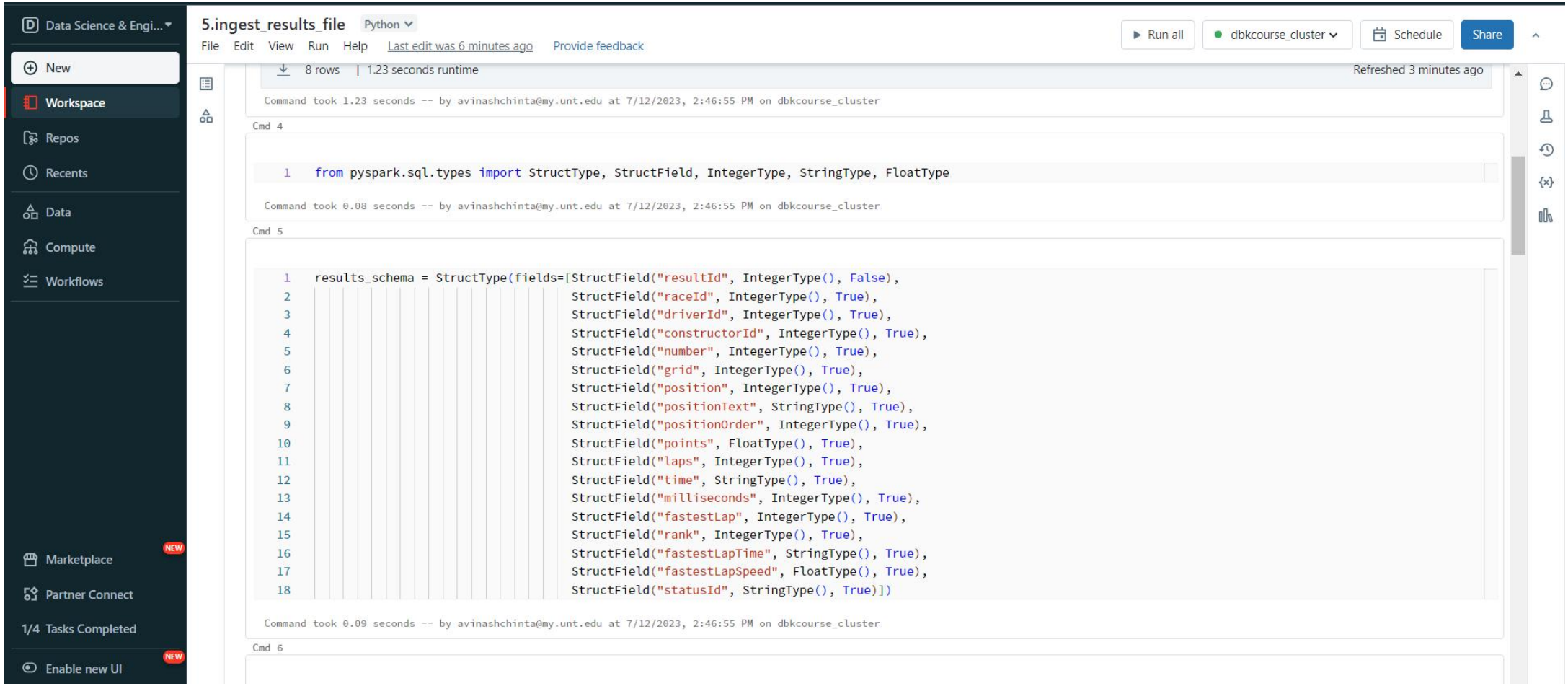

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 2 - Rename columns and add new columns
1. driverId renamed to driver_id
2. driverRef renamed to driver_ref
3. ingestion date added
4. name added with concatenation of forename and surname
Cmd 10
1 from pyspark.sql.functions import col, concat, current_timestamp, lit
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 11
1 drivers_with_columns_df = drivers_df.withColumnRenamed("driverId", "driver_id") \
2 .withColumnRenamed("driverRef", "driver_ref") \
3 .withColumn("ingestion_date", current_timestamp()) \
4 .withColumn("name", concat(col("name.forename"), lit(" "), col("name.surname")))
drivers_with_columns_df: pyspark.sql.dataframe.DataFrame = [driver_id: integer, driver_ref: string ... 7 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 12
1 display(drivers_with_columns_df)
(1) Spark Jobs
Table
driver_id driver_ref number code name dob nationality url
1 1 hamilton 44 HAM Lewis Hamilton 1985-01-07 British http://en.wikipedia.org/wiki/Lewis_Hamilton
2 2 heidfeld null HEI Nick Heidfeld 1977-05-10 German http://en.wikipedia.org/wiki/Nick_Heidfeld

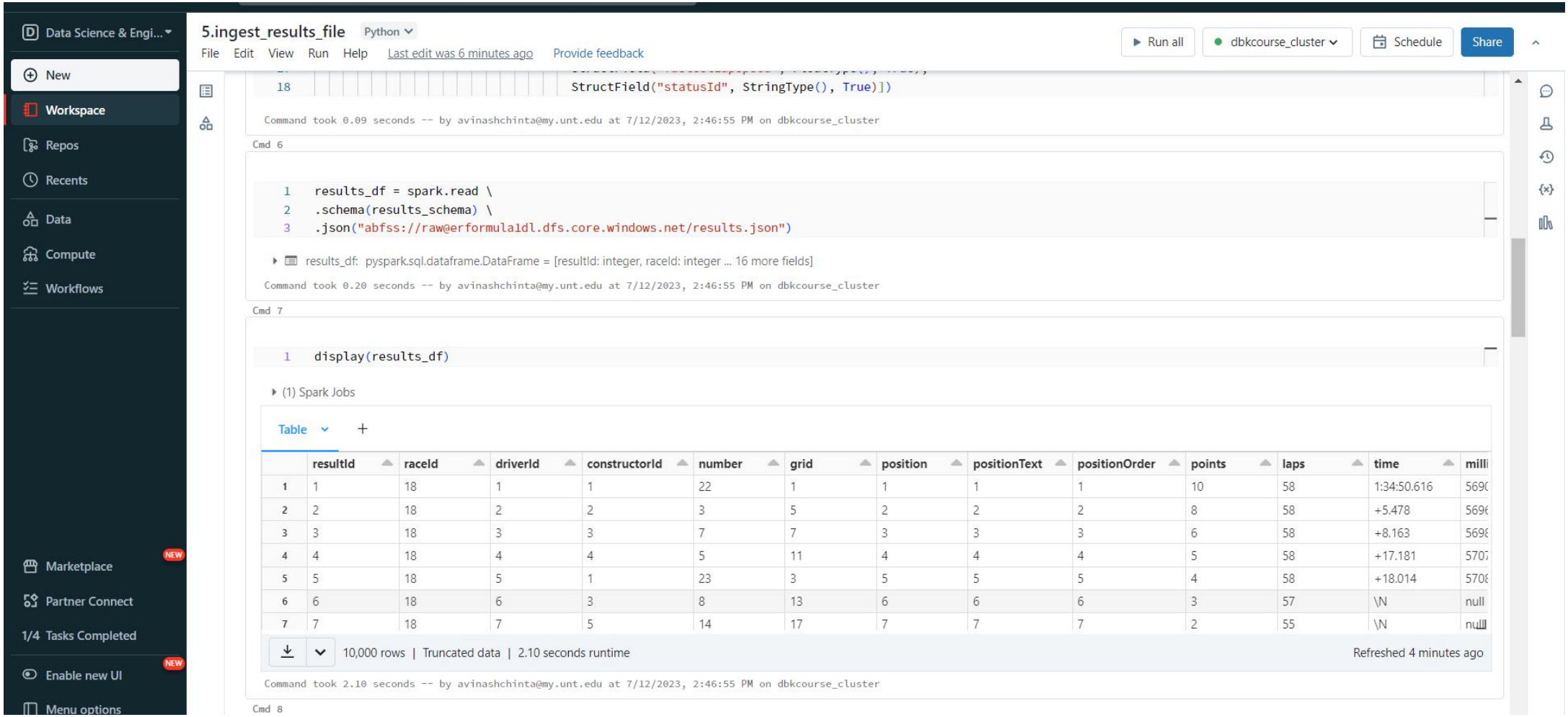

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.59 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 13
Step 3 - Drop the unwanted columns
1. name.forename
2. name.surname
3. url
Cmd 14
1 drivers_final_df = drivers_with_columns_df.drop(col("url"))
drivers_final_df: pyspark.sql.dataframe.DataFrame = [driver_id: integer, driver_ref: string ... 6 more fields]
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 15
1 display(drivers_final_df)
(1) Spark Jobs
Table
driver_id driver_ref number code name dob nationality ingestion_date
1 1 hamilton 44 HAM Lewis Hamilton 1985-01-07 British 2023-07-12T19:28:45.721+0000
2 2 heidfeld null HEI Nick Heidfeld 1977-05-10 German 2023-07-12T19:28:45.721+0000
3 3 rosberg 6 ROS Nico Rosberg 1985-06-27 German 2023-07-12T19:28:45.721+0000
4 4 alonso 14 ALO Fernando Alonso 1981-07-29 Spanish 2023-07-12T19:28:45.721+0000
5 5 kovalainen null KOV Heikki Kovalainen 1981-10-19 Finnish 2023-07-12T19:28:45.721+0000
6 6 nakajima null NAK Kazuki Nakajima 1985-01-11 Japanese 2023-07-12T19:28:45.721+0000
7 7 bourdais null BOU Sébastien Bourdais 1979-02-28 French 2023-07-12T19:28:45.721+0000
853 rows | 0.49 seconds runtime
Refreshed 2 minutes ago

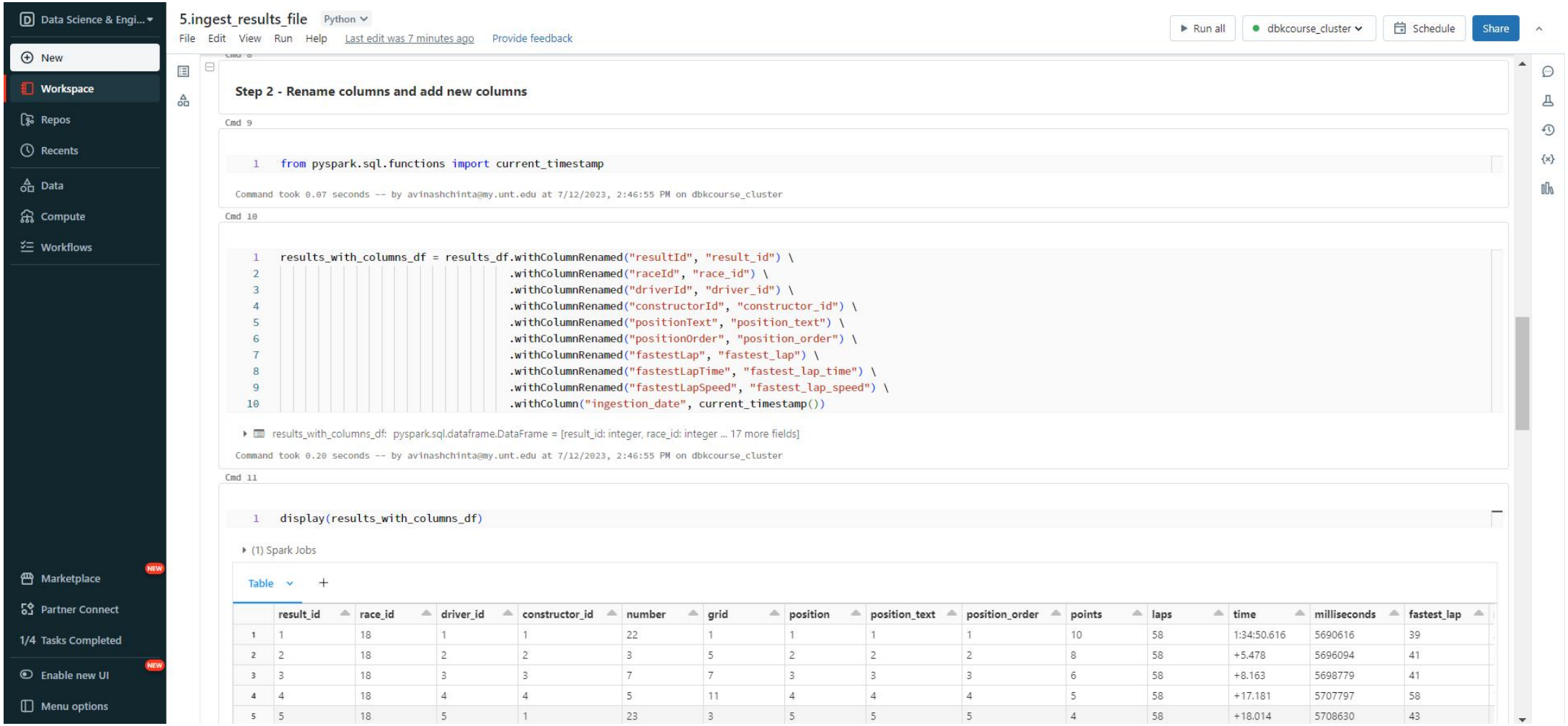

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
4.ingest_drivers_file
Python
File Edit View Run Help
Last edit was 8 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 4 - Write to output to processed container in parquet format
Cmd 17
1 drivers_final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/drivers")
(1) Spark Jobs
Command took 0.89 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Cmd 18
Python
1 display(spark.read.parquet('abfss://processed@erformula1dl.dfs.core.windows.net/drivers'))
(2) Spark Jobs
Table
driver_id driver_ref number code name dob nationality ingestion_date
1 1 hamilton 44 HAM Lewis Hamilton 1985-01-07 British 2023-07-12T19:28:46.282+0000
2 2 heidfeld null HEI Nick Heidfeld 1977-05-10 German 2023-07-12T19:28:46.282+0000
3 3 rosberg 6 ROS Nico Rosberg 1985-06-27 German 2023-07-12T19:28:46.282+0000
4 4 alonso 14 ALO Fernando Alonso 1981-07-29 Spanish 2023-07-12T19:28:46.282+0000
5 5 kovalainen null KOV Heikki Kovalainen 1981-10-19 Finnish 2023-07-12T19:28:46.282+0000
6 6 nakajima null NAK Kazuki Nakajima 1985-01-11 Japanese 2023-07-12T19:28:46.282+0000
7 7 bourdais null BOU Sébastien Bourdais 1979-02-28 French 2023-07-12T19:28:46.282+0000
853 rows | 0.48 seconds runtime
Refreshed 3 minutes ago
Command took 0.48 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:28:42 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

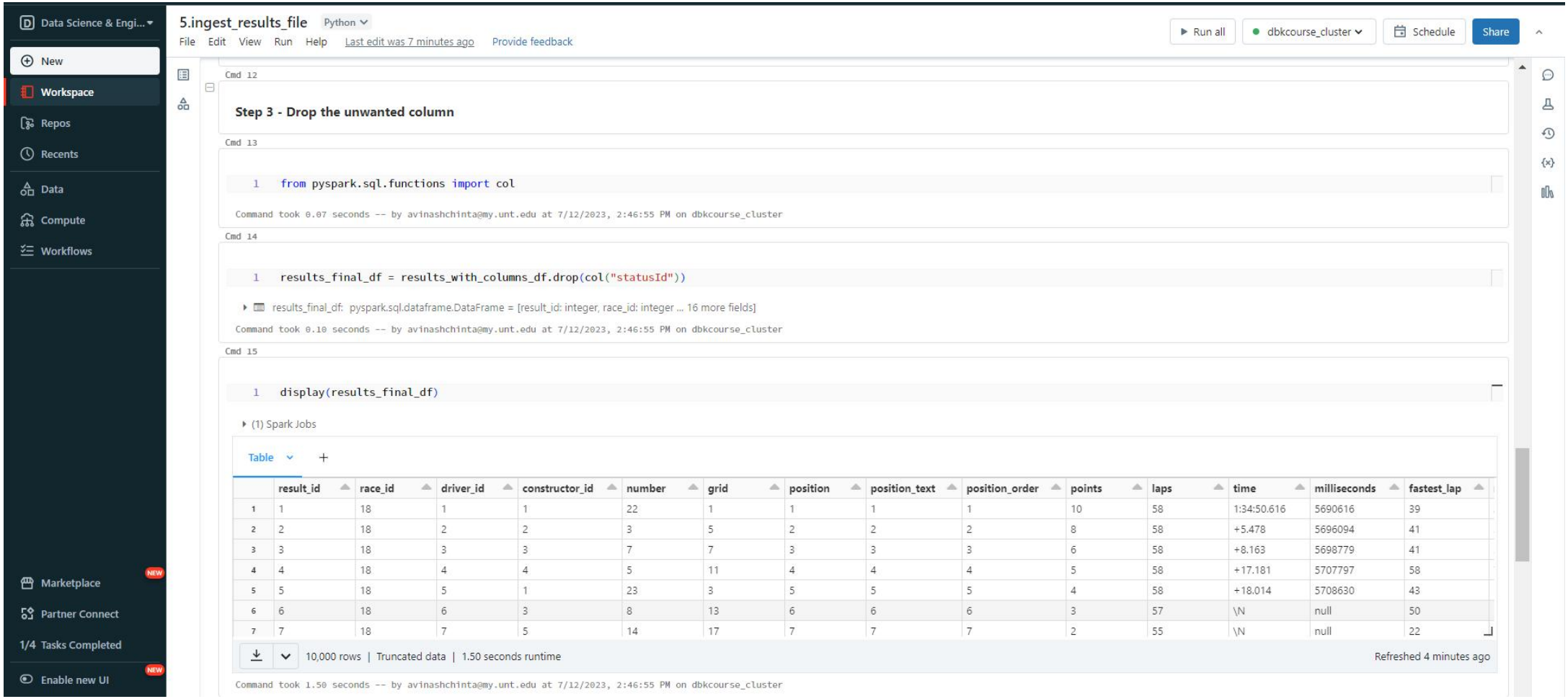

Home > erformula1dl | Containers >
processed
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / drivers
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
_committed_5575062938918487361
7/12/2023, 2:28:46 PM
Hot (Inferred)
Block blob
123 B
Available
_started_5575062938918487361
7/12/2023, 2:28:46 PM
Hot (Inferred)
Block blob
0 B
Available
_SUCCESS
7/12/2023, 2:28:47 PM
Hot (Inferred)
Block blob
0 B
Available
part-00000-tid-5575062938918487361-02453197-524e-4407-bbb4-...
7/12/2023, 2:28:46 PM
Hot (Inferred)
Block blob
28.5 KiB
Available

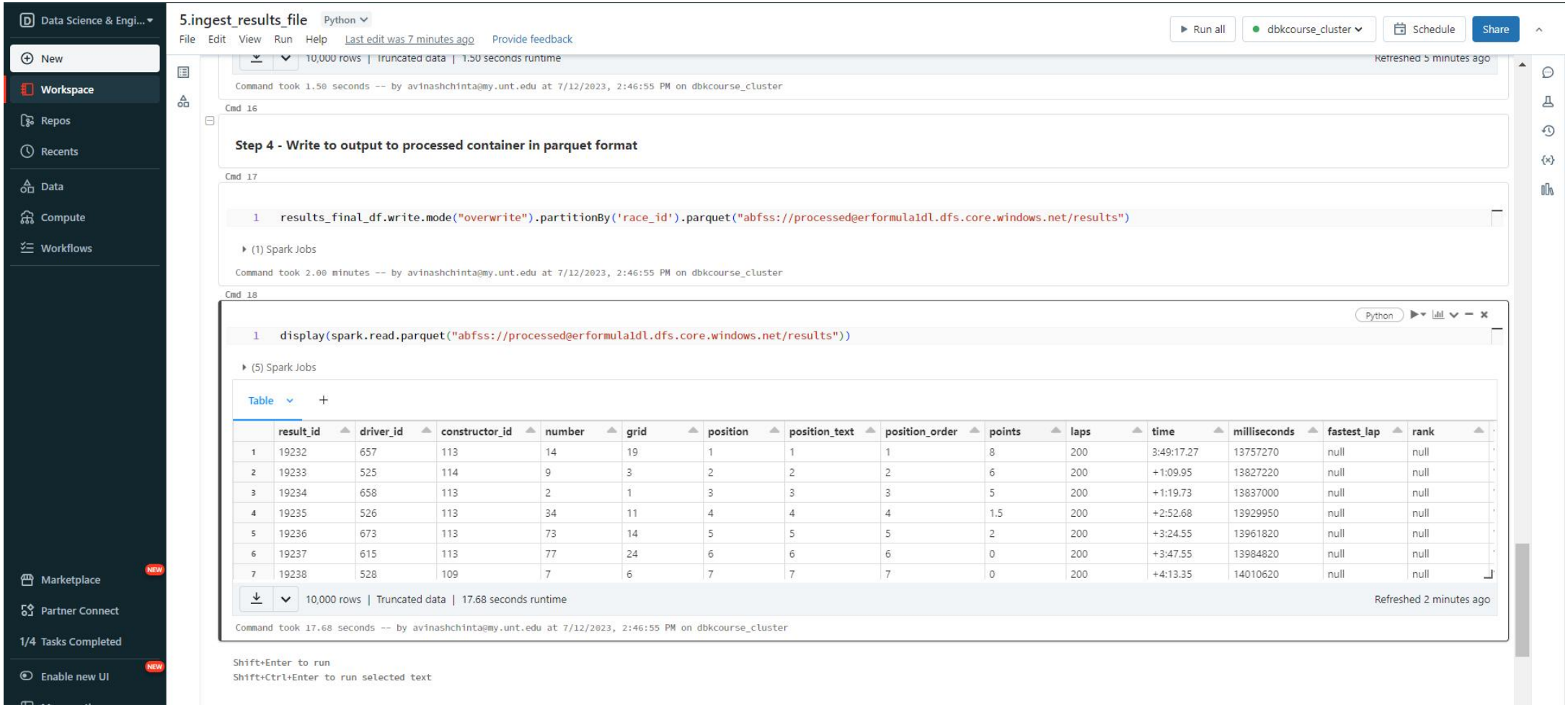

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 1
Ingest results.json file
Cmd 2
Step 1 - Read the JSON file using the spark dataframe reader API
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path
name
size
modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 1.23 seconds runtime
Refreshed 2 minutes ago
Command took 1.23 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 4


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
8 rows | 1.23 seconds runtime
Refreshed 3 minutes ago
Command took 1.23 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 4
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, FloatType
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 5
1 results_schema = StructType(fields=[StructField("resultId", IntegerType(), False),
2 StructField("raceId", IntegerType(), True),
3 StructField("driverId", IntegerType(), True),
4 StructField("constructorId", IntegerType(), True),
5 StructField("number", IntegerType(), True),
6 StructField("grid", IntegerType(), True),
7 StructField("position", IntegerType(), True),
8 StructField("positionText", StringType(), True),
9 StructField("positionOrder", IntegerType(), True),
10 StructField("points", FloatType(), True),
11 StructField("laps", IntegerType(), True),
12 StructField("time", StringType(), True),
13 StructField("milliseconds", IntegerType(), True),
14 StructField("fastestLap", IntegerType(), True),
15 StructField("rank", IntegerType(), True),
16 StructField("fastestLapTime", StringType(), True),
17 StructField("fastestLapSpeed", FloatType(), True),
18 StructField("statusId", StringType(), True)])
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 6


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
StructField("statusId", StringType(), True)])
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 6
results_df = spark.read \
.schema(results_schema) \
.json("abfss://raw@erformula1dl.dfs.core.windows.net/results.json")
results_df: pyspark.sql.dataframe.DataFrame = [resultId: integer, raceId: integer ... 16 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 7
display(results_df)
(1) Spark Jobs
Table
resultId raceId driverId constructorId number grid position positionText positionOrder points laps time milli
1 1 18 1 1 22 1 1 1 1 10 58 1:34:50.616 569(
2 2 18 2 2 3 5 2 2 2 8 58 +5.478 569(
3 3 18 3 3 7 7 3 3 3 6 58 +8.163 569(
4 4 18 4 4 5 11 4 4 4 5 58 +17.181 570
5 5 18 5 1 23 3 5 5 5 4 58 +18.014 570(
6 6 18 6 3 8 13 6 6 6 3 57 \N null
7 7 18 7 5 14 17 7 7 7 2 55 \N null
10,000 rows | Truncated data | 2.10 seconds runtime
Refreshed 4 minutes ago
Command took 2.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 8


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 2 - Rename columns and add new columns
Cmd 9
from pyspark.sql.functions import current_timestamp
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 10
results_with_columns_df = results_df.withColumnRenamed("resultId", "result_id") \
.withColumnRenamed("raceId", "race_id") \
.withColumnRenamed("driverId", "driver_id") \
.withColumnRenamed("constructorId", "constructor_id") \
.withColumnRenamed("positionText", "position_text") \
.withColumnRenamed("positionOrder", "position_order") \
.withColumnRenamed("fastestLap", "fastest_lap") \
.withColumnRenamed("fastestLapTime", "fastest_lap_time") \
.withColumnRenamed("fastestLapSpeed", "fastest_lap_speed") \
.withColumn("ingestion_date", current_timestamp())
results_with_columns_df: pyspark.sql.dataframe.DataFrame = [result_id: integer, race_id: integer ... 17 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 11
display(results_with_columns_df)
(1) Spark Jobs
Table
result_id race_id driver_id constructor_id number grid position position_text position_order points laps time milliseconds fastest_lap
1 1 18 1 1 22 1 1 1 1 10 58 1:34:50.616 5690616 39
2 2 18 2 2 3 5 2 2 2 8 58 +5.478 5696094 41
3 3 18 3 3 7 7 3 3 3 6 58 +8.163 5698779 41
4 4 18 4 4 5 11 4 4 4 5 58 +17.181 5707797 58
5 5 18 5 1 23 3 5 5 5 4 58 +18.014 5708630 43


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 12
Step 3 - Drop the unwanted column
Cmd 13
1 from pyspark.sql.functions import col
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 14
1 results_final_df = results_with_columns_df.drop(col("statusId"))
results_final_df: pyspark.sql.dataframe.DataFrame = [result_id: integer, race_id: integer ... 16 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 15
1 display(results_final_df)
(1) Spark Jobs
Table
result_id race_id driver_id constructor_id number grid position position_text position_order points laps time milliseconds fastest_lap
1 1 18 1 1 22 1 1 1 1 10 58 1:34:50.616 5690616 39
2 2 18 2 2 3 5 2 2 2 8 58 +5.478 5696094 41
3 3 18 3 3 7 7 3 3 3 6 58 +8.163 5698779 41
4 4 18 4 4 5 11 4 4 4 5 58 +17.181 5707797 58
5 5 18 5 1 23 3 5 5 5 4 58 +18.014 5708630 43
6 6 18 6 3 8 13 6 6 6 3 57 \N null 50
7 7 18 7 5 14 17 7 7 7 2 55 \N null 22
10,000 rows | Truncated data | 1.50 seconds runtime
Refreshed 4 minutes ago
Command took 1.50 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
5.ingest_results_file
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
10,000 rows | Truncated data | 1.50 seconds runtime
Refreshed 5 minutes ago
Command took 1.50 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 16
Step 4 - Write to output to processed container in parquet format
Cmd 17
1 results_final_df.write.mode("overwrite").partitionBy('race_id').parquet("abfss://processed@erformula1dl.dfs.core.windows.net/results")
(1) Spark Jobs
Command took 2.00 minutes -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Cmd 18
Python
1 display(spark.read.parquet("abfss://processed@erformula1dl.dfs.core.windows.net/results"))
(5) Spark Jobs
Table
result_id driver_id constructor_id number grid position position_text position_order points laps time milliseconds fastest_lap rank
1 19232 657 113 14 19 1 1 1 8 200 3:49:17.27 13757270 null null
2 19233 525 114 9 3 2 2 2 6 200 +1:09.95 13827220 null null
3 19234 658 113 2 1 3 3 3 5 200 +1:19.73 13837000 null null
4 19235 526 113 34 11 4 4 4 1.5 200 +2:52.68 13929950 null null
5 19236 673 113 73 14 5 5 5 2 200 +3:24.55 13961820 null null
6 19237 615 113 77 24 6 6 6 0 200 +3:47.55 13984820 null null
7 19238 528 109 7 6 7 7 7 0 200 +4:13.35 14010620 null null
10,000 rows | Truncated data | 17.68 seconds runtime
Refreshed 2 minutes ago
Command took 17.68 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 2:46:55 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

Home > erformula1dl | Containers >

processed

Container

Search

Overview
Diagnose and solve problems
Access Control (IAM)

Settings

Shared access tokens
Manage ACL
Access policy
Properties
Metadata

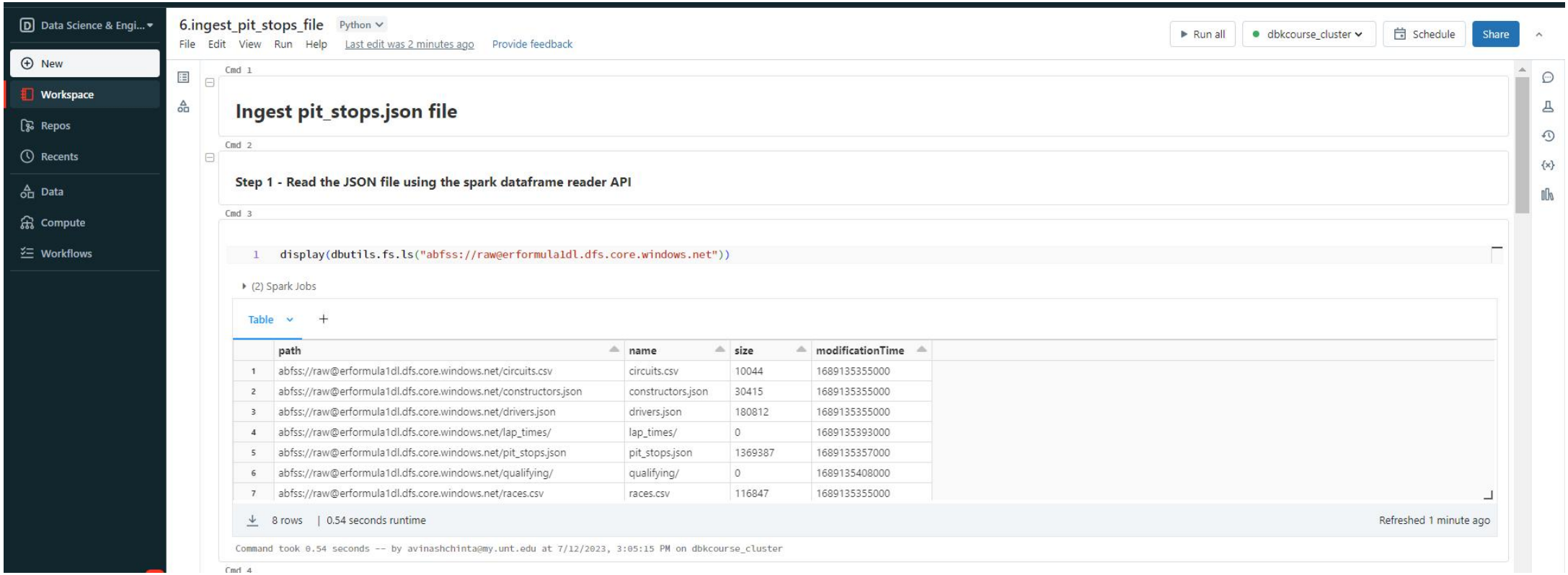
Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

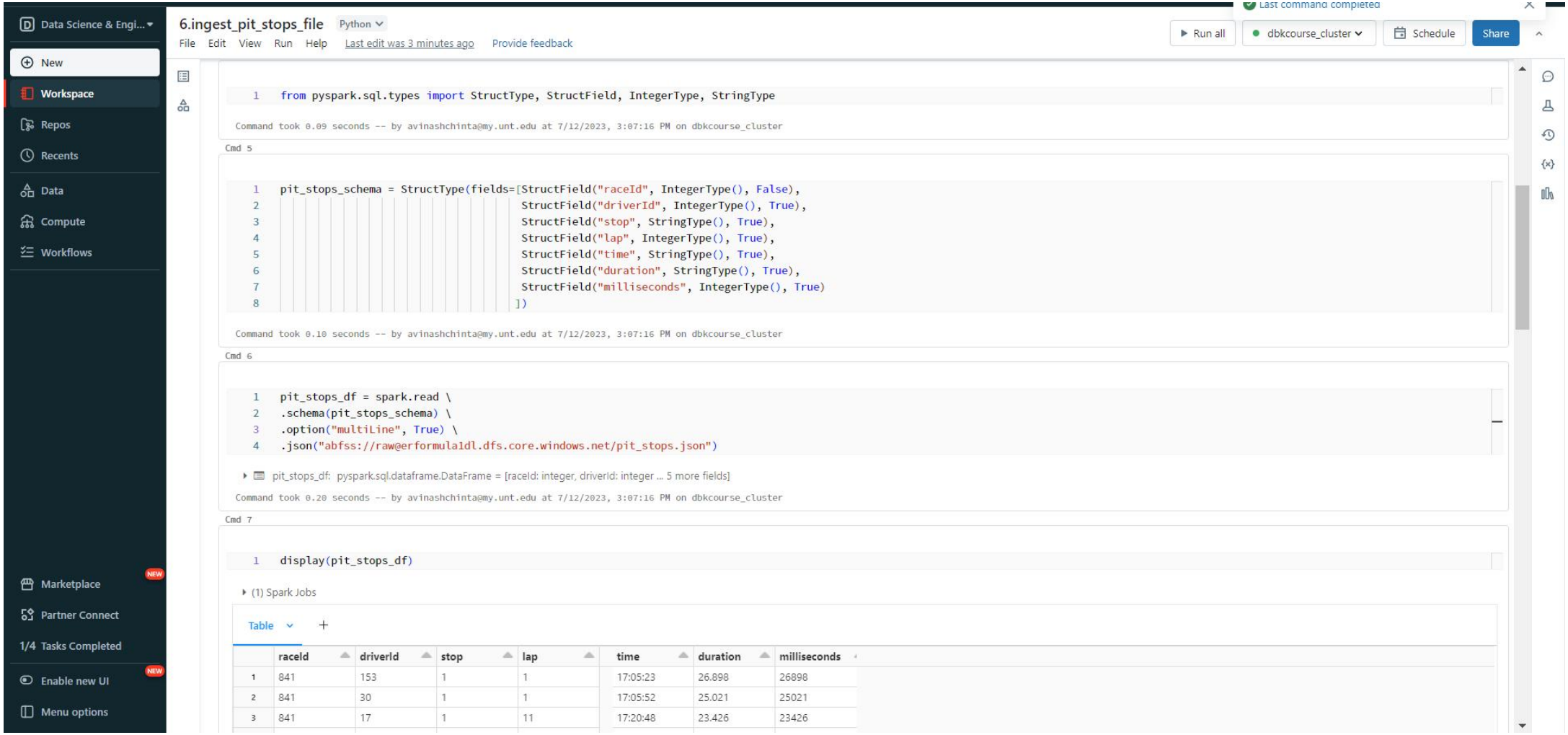
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / results

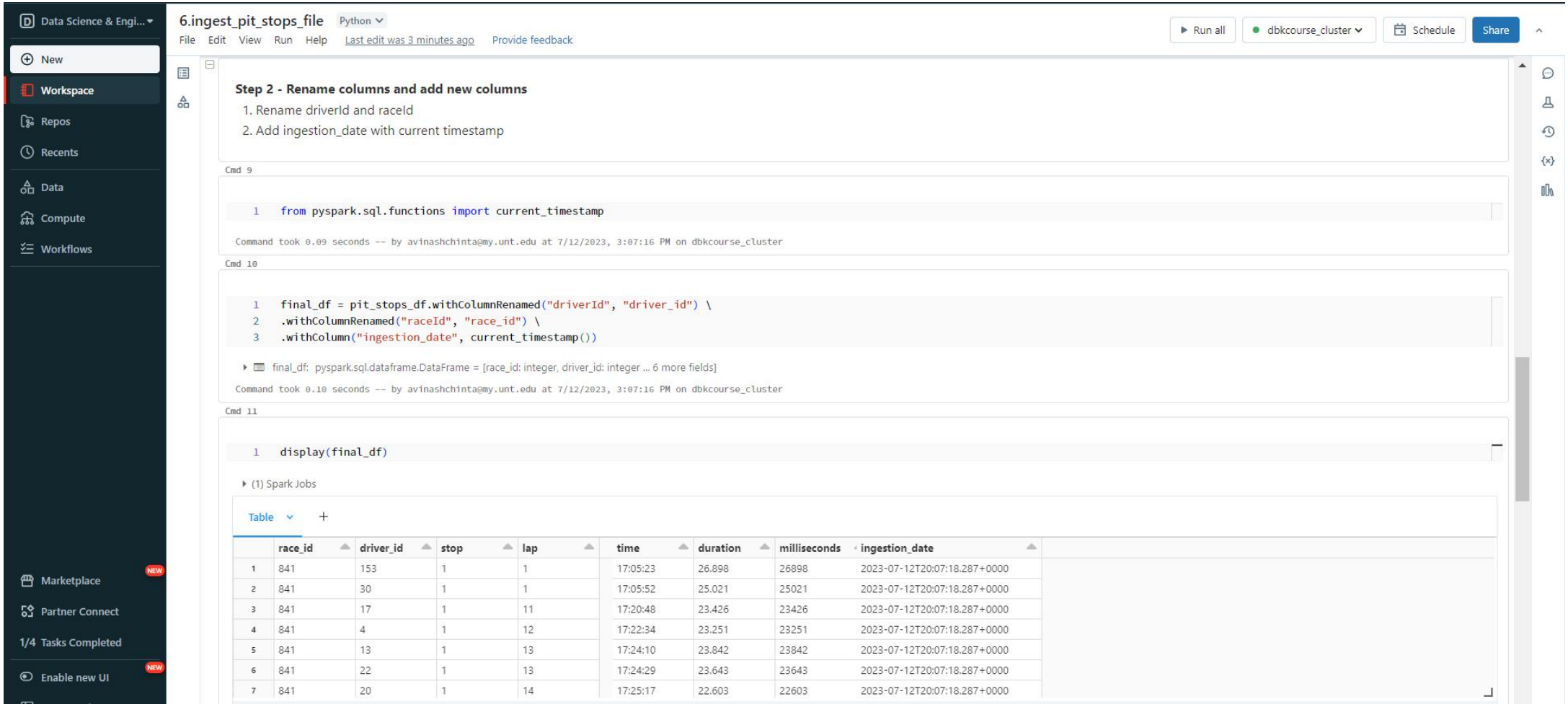
Search blobs by prefix (case-sensitive)

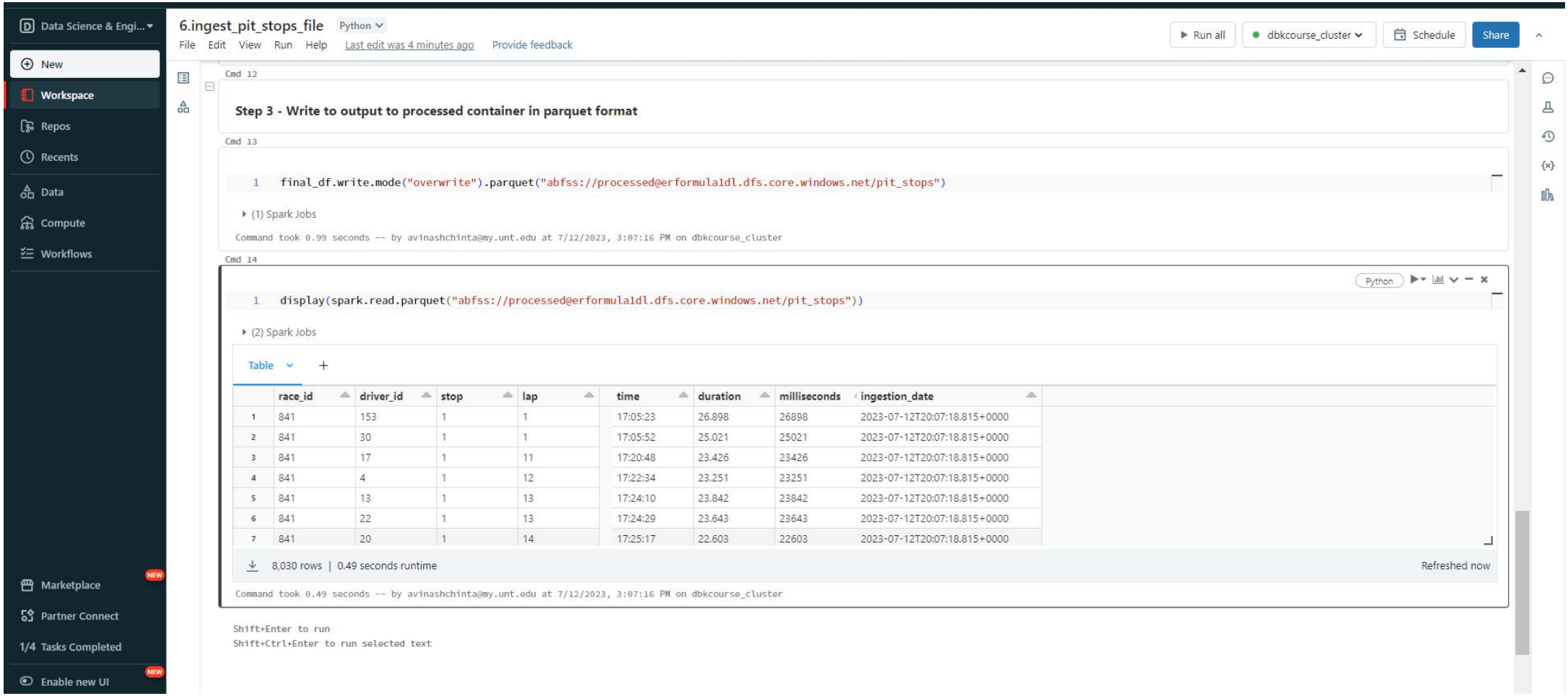
Show deleted objects

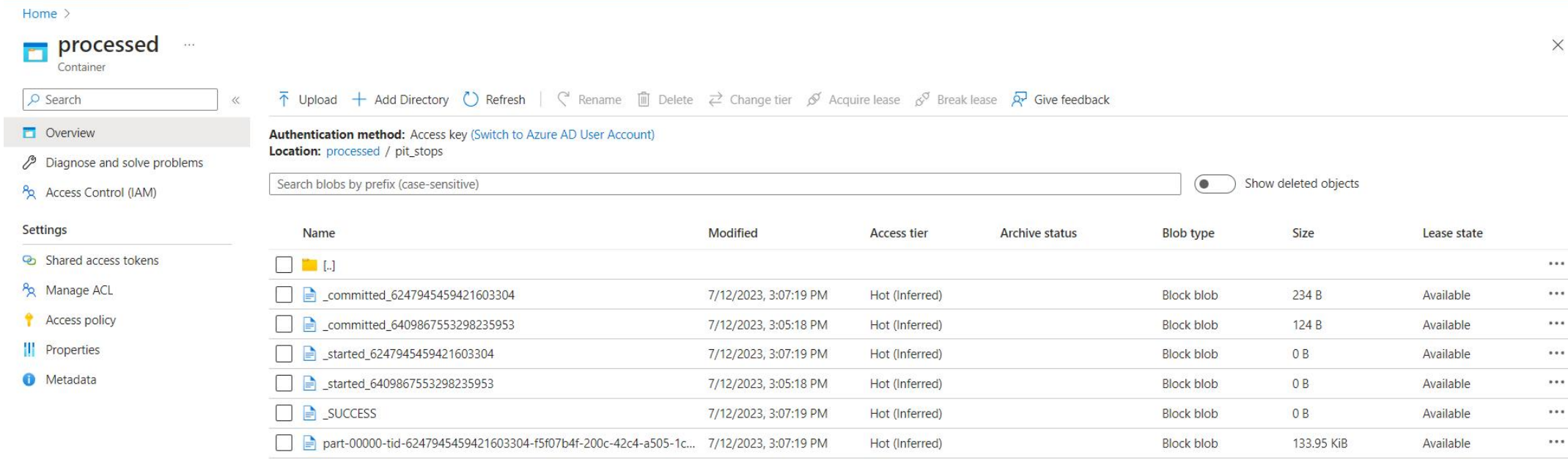
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
race_id=1						-
race_id=10						-
race_id=100						-
race_id=1000						-
race_id=1001						-
race_id=1002						-
race_id=1003						-
race_id=1004						-
race_id=1005						-
race_id=1006						-
race_id=1007						-
race_id=1008						-
race_id=1009						-
race_id=101						-
race_id=1010						-

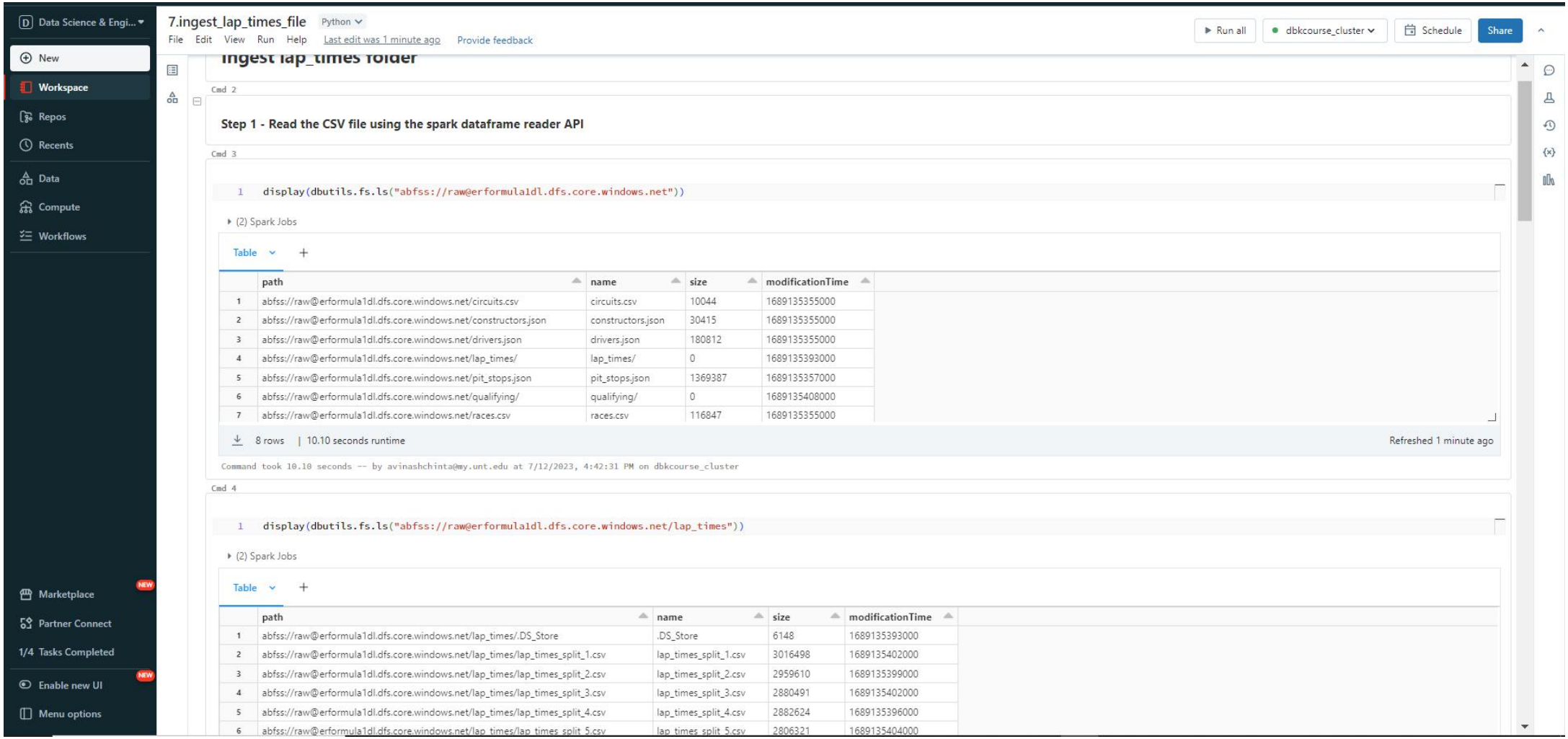

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
6.ingest_pit_stops_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 1
Ingest pit_stops.json file
Cmd 2
Step 1 - Read the JSON file using the spark dataframe reader API
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path
name
size
modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 0.54 seconds runtime
Refreshed 1 minute ago
Command took 0.54 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:05:15 PM on dbkcourse_cluster
Cmd 4


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
6.ingest_pit_stops_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Last command completed
Run all
dbkcourse_cluster
Schedule
Share
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 5
1 pit_stops_schema = StructType(fields=[StructField("raceId", IntegerType(), False),
2 StructField("driverId", IntegerType(), True),
3 StructField("stop", StringType(), True),
4 StructField("lap", IntegerType(), True),
5 StructField("time", StringType(), True),
6 StructField("duration", StringType(), True),
7 StructField("milliseconds", IntegerType(), True)
8])
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 6
1 pit_stops_df = spark.read \
2 .schema(pit_stops_schema) \
3 .option("multiLine", True) \
4 .json("abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json")
pit_stops_df: pyspark.sql.dataframe.DataFrame = [raceId: integer, driverId: integer ... 5 more fields]
Command took 0.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 7
1 display(pit_stops_df)
(1) Spark Jobs
Table
raceId driverId stop lap time duration milliseconds
1 841 153 1 1 17:05:23 26.898 26898
2 841 30 1 1 17:05:52 25.021 25021
3 841 17 1 11 17:20:48 23.426 23426


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
6.ingest_pit_stops_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 2 - Rename columns and add new columns
1. Rename driverId and raceId
2. Add ingestion_date with current timestamp
Cmd 9
from pyspark.sql.functions import current_timestamp
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 10
final_df = pit_stops_df.withColumnRenamed("driverId", "driver_id") \
.withColumnRenamed("raceId", "race_id") \
.withColumn("ingestion_date", current_timestamp())
final_df: pyspark.sql.dataframe.DataFrame = [race_id: integer, driver_id: integer ... 6 more fields]
Command took 0.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 11
display(final_df)
(1) Spark Jobs
Table
race_id driver_id stop lap time duration milliseconds ingestion_date
1 841 153 1 1 17:05:23 26.898 26898 2023-07-12T20:07:18.287+0000
2 841 30 1 1 17:05:52 25.021 25021 2023-07-12T20:07:18.287+0000
3 841 17 1 11 17:20:48 23.426 23426 2023-07-12T20:07:18.287+0000
4 841 4 1 12 17:22:34 23.251 23251 2023-07-12T20:07:18.287+0000
5 841 13 1 13 17:24:10 23.842 23842 2023-07-12T20:07:18.287+0000
6 841 22 1 13 17:24:29 23.643 23643 2023-07-12T20:07:18.287+0000
7 841 20 1 14 17:25:17 22.603 22603 2023-07-12T20:07:18.287+0000


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
6.ingest_pit_stops_file
Python
File Edit View Run Help
Last edit was 4 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 12
Step 3 - Write to output to processed container in parquet format
Cmd 13
1 final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/pit_stops")
(1) Spark Jobs
Command took 0.99 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Cmd 14
Python
1 display(spark.read.parquet("abfss://processed@erformula1dl.dfs.core.windows.net/pit_stops"))
(2) Spark Jobs
Table
race_id driver_id stop lap time duration milliseconds ingestion_date
1 841 153 1 1 17:05:23 26.898 26898 2023-07-12T20:07:18.815+0000
2 841 30 1 1 17:05:52 25.021 25021 2023-07-12T20:07:18.815+0000
3 841 17 1 11 17:20:48 23.426 23426 2023-07-12T20:07:18.815+0000
4 841 4 1 12 17:22:34 23.251 23251 2023-07-12T20:07:18.815+0000
5 841 13 1 13 17:24:10 23.842 23842 2023-07-12T20:07:18.815+0000
6 841 22 1 13 17:24:29 23.643 23643 2023-07-12T20:07:18.815+0000
7 841 20 1 14 17:25:17 22.603 22603 2023-07-12T20:07:18.815+0000
8,030 rows | 0.49 seconds runtime
Refreshed now
Command took 0.49 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 3:07:16 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text


Home >
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 / pit_stops
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
_committed_6247945459421603304 7/12/2023, 3:07:19 PM Hot (Inferred) Block blob 234 B Available
_committed_6409867553298235953 7/12/2023, 3:05:18 PM Hot (Inferred) Block blob 124 B Available
_started_6247945459421603304 7/12/2023, 3:07:19 PM Hot (Inferred) Block blob 0 B Available
_started_6409867553298235953 7/12/2023, 3:05:18 PM Hot (Inferred) Block blob 0 B Available
_SUCCESS 7/12/2023, 3:07:19 PM Hot (Inferred) Block blob 0 B Available
part-00000-tid-6247945459421603304-f5f07b4f-200c-42c4-a505-1c... 7/12/2023, 3:07:19 PM Hot (Inferred) Block blob 133.95 KiB Available


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
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
Ingest lap_times folder
Cmd 2
Step 1 - Read the CSV file using the spark dataframe reader API
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path name size modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 10.10 seconds runtime
Refreshed 1 minute ago
Command took 10.10 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster
Cmd 4
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net/lap_times"))
(2) Spark Jobs
Table
path name size modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/.DS_Store .DS_Store 6148 1689135393000
2 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/lap_times_split_1.csv lap_times_split_1.csv 3016498 1689135402000
3 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/lap_times_split_2.csv lap_times_split_2.csv 2959610 1689135399000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/lap_times_split_3.csv lap_times_split_3.csv 2880491 1689135402000
5 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/lap_times_split_4.csv lap_times_split_4.csv 2882624 1689135396000
6 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/lap_times_split_5.csv lap_times_split_5.csv 2806321 1689135404000

Data Science & Engi...

New

Workspace

Repos

Recents

Data

Compute

Workflows

Marketplace

Partner Connect

1/4 Tasks Completed

Enable new UI

Menu options

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

```
1  from pyspark.sql.types import StructType, StructField, IntegerType, StringType
```

Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster

Cmd 6

```
1  lap_times_schema = StructType(fields=[StructField("raceId", IntegerType(), False),
2                                        StructField("driverId", IntegerType(), True),
3                                        StructField("lap", IntegerType(), True),
4                                        StructField("position", IntegerType(), True),
5                                        StructField("time", StringType(), True),
6                                        StructField("milliseconds", IntegerType(), True)
7                                       ])
```

Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster

Cmd 7

```
1  lap_times_df = spark.read \
2  .schema(lap_times_schema) \
3  .csv("abfss://raw@erformula1dl.dfs.core.windows.net/lap_times")
```

▸ lap_times_df: pyspark.sql.dataframe.DataFrame = [raceId: integer, driverId: integer ... 4 more fields]

Command took 1.18 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster

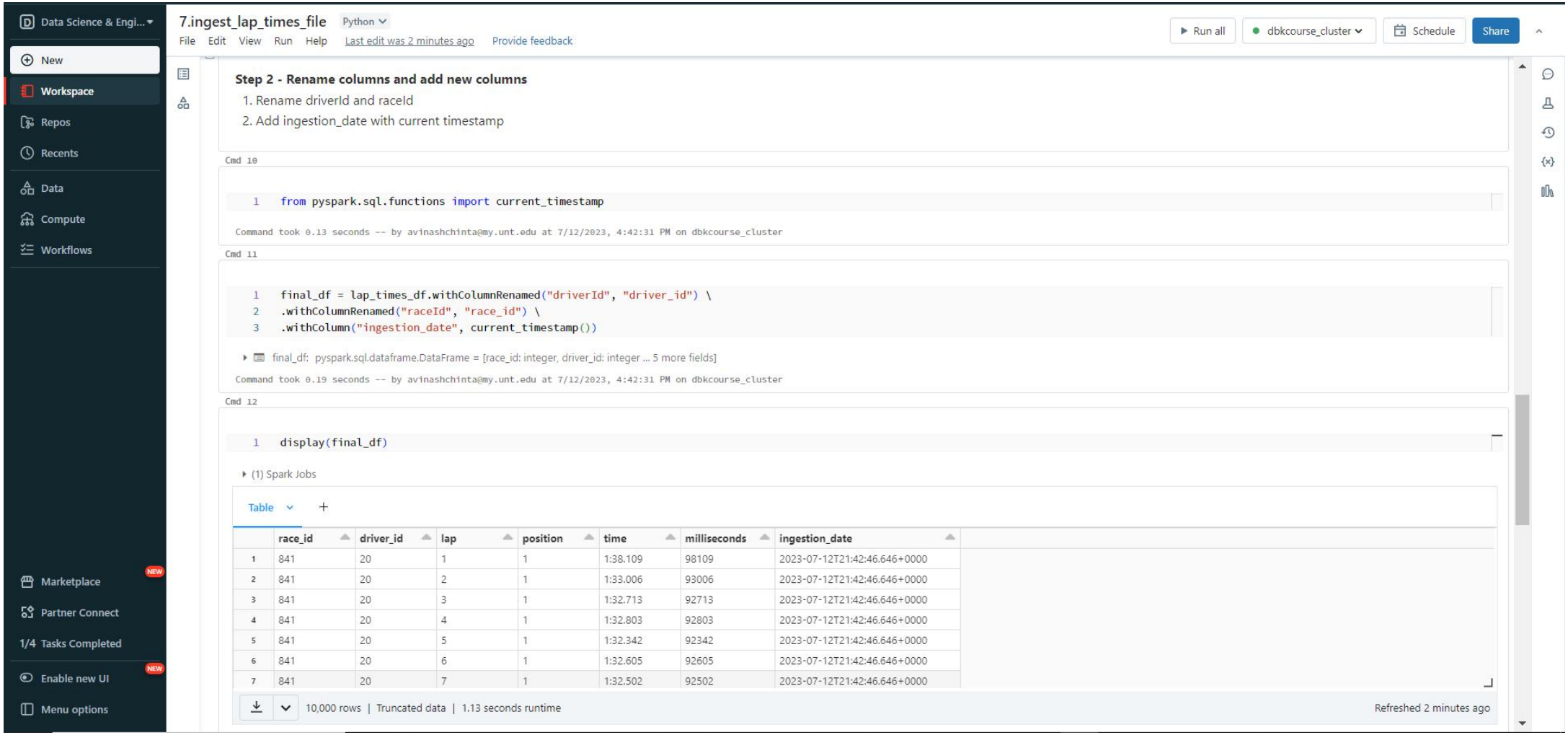
Cmd 8

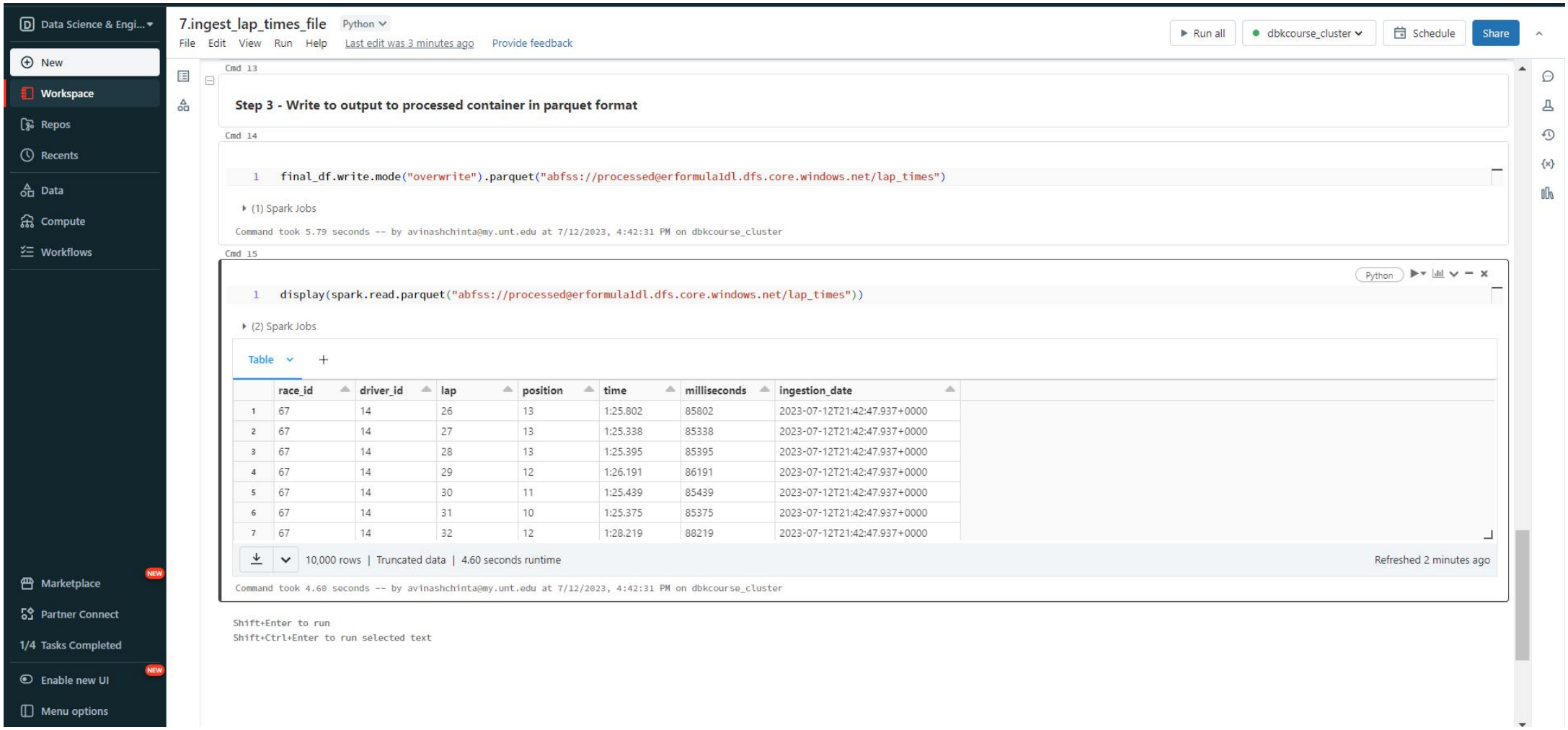
```
1  display(lap_times_df)
```

▸ (1) Spark Jobs

Table +

	raceId	driverId	lap	position	time	milliseconds
1	841	20	1	1	1:38.109	98109
2	841	20	2	1	1:33.006	93006
3	841	20	3	1	1:32.713	92713
4	841	20	4	1	1:32.803	92803
5	841	20	5	1	1:32.342	92342
6	841	20	6	1	1:32.605	92605


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
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
Step 2 - Rename columns and add new columns
1. Rename driverId and raceId
2. Add ingestion_date with current timestamp
Cmd 10
1 from pyspark.sql.functions import current_timestamp
Command took 0.13 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster
Cmd 11
1 final_df = lap_times_df.withColumnRenamed("driverId", "driver_id") \
2 .withColumnRenamed("raceId", "race_id") \
3 .withColumn("ingestion_date", current_timestamp())
final_df: pyspark.sql.dataframe.DataFrame = [race_id: integer, driver_id: integer ... 5 more fields]
Command took 0.19 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster
Cmd 12
1 display(final_df)
(1) Spark Jobs
Table
race_id driver_id lap position time milliseconds ingestion_date
1 841 20 1 1 1:38.109 98109 2023-07-12T21:42:46.646+0000
2 841 20 2 1 1:33.006 93006 2023-07-12T21:42:46.646+0000
3 841 20 3 1 1:32.713 92713 2023-07-12T21:42:46.646+0000
4 841 20 4 1 1:32.803 92803 2023-07-12T21:42:46.646+0000
5 841 20 5 1 1:32.342 92342 2023-07-12T21:42:46.646+0000
6 841 20 6 1 1:32.605 92605 2023-07-12T21:42:46.646+0000
7 841 20 7 1 1:32.502 92502 2023-07-12T21:42:46.646+0000
10,000 rows | Truncated data | 1.13 seconds runtime
Refreshed 2 minutes ago


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
7.ingest_lap_times_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Cmd 13
Step 3 - Write to output to processed container in parquet format
Cmd 14
1 final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/lap_times")
(1) Spark Jobs
Command took 5.79 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster
Cmd 15
Python
1 display(spark.read.parquet("abfss://processed@erformula1dl.dfs.core.windows.net/lap_times"))
(2) Spark Jobs
Table
race_id driver_id lap position time milliseconds ingestion_date
1 67 14 26 13 1:25.802 85802 2023-07-12T21:42:47.937+0000
2 67 14 27 13 1:25.338 85338 2023-07-12T21:42:47.937+0000
3 67 14 28 13 1:25.395 85395 2023-07-12T21:42:47.937+0000
4 67 14 29 12 1:26.191 86191 2023-07-12T21:42:47.937+0000
5 67 14 30 11 1:25.439 85439 2023-07-12T21:42:47.937+0000
6 67 14 31 10 1:25.375 85375 2023-07-12T21:42:47.937+0000
7 67 14 32 12 1:28.219 88219 2023-07-12T21:42:47.937+0000
10,000 rows | Truncated data | 4.60 seconds runtime
Refreshed 2 minutes ago
Command took 4.60 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:42:31 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

Home >

processed

Container

Search

Overview
Diagnose and solve problems
Access Control (IAM)

Settings

Shared access tokens
Manage ACL
Access policy
Properties
Metadata

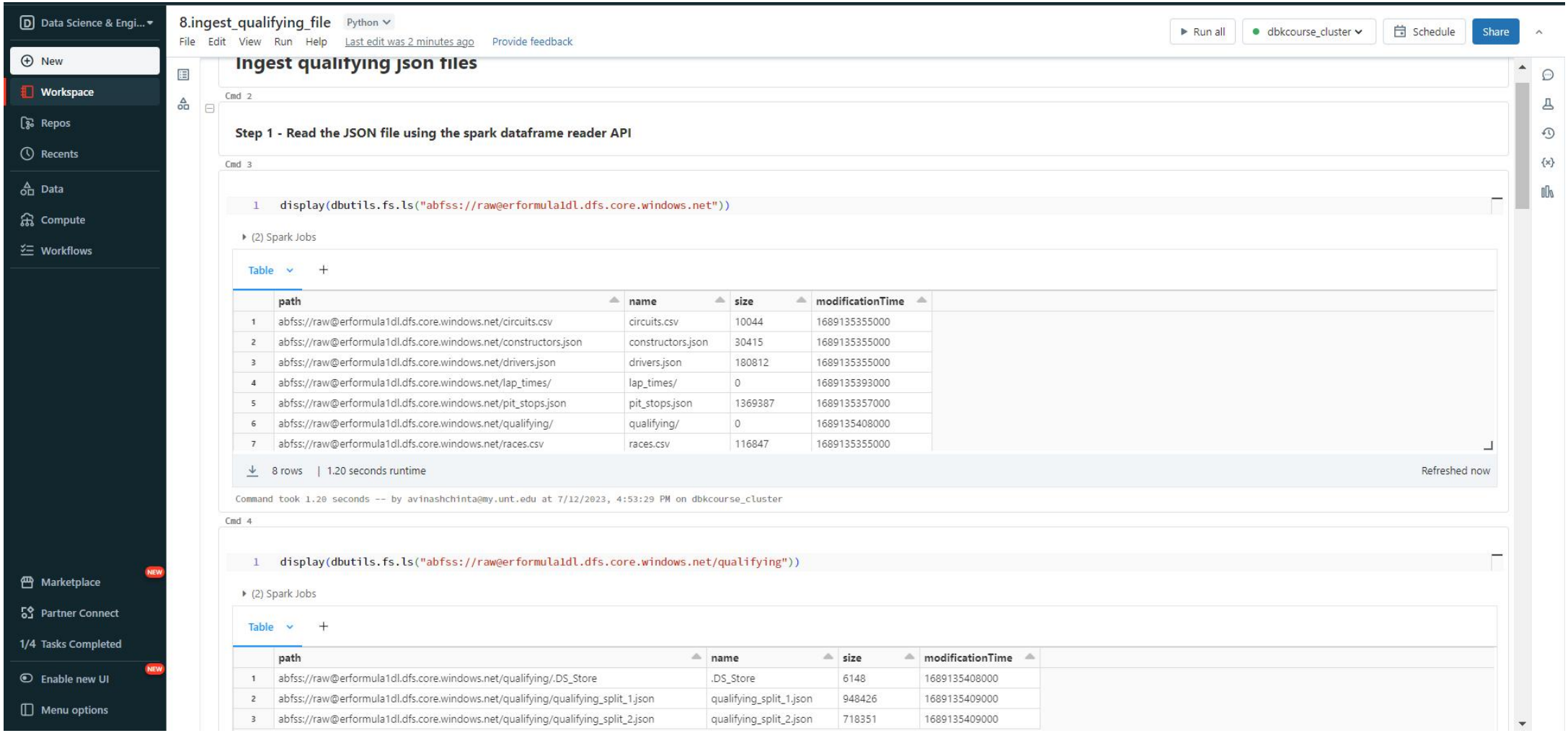
Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

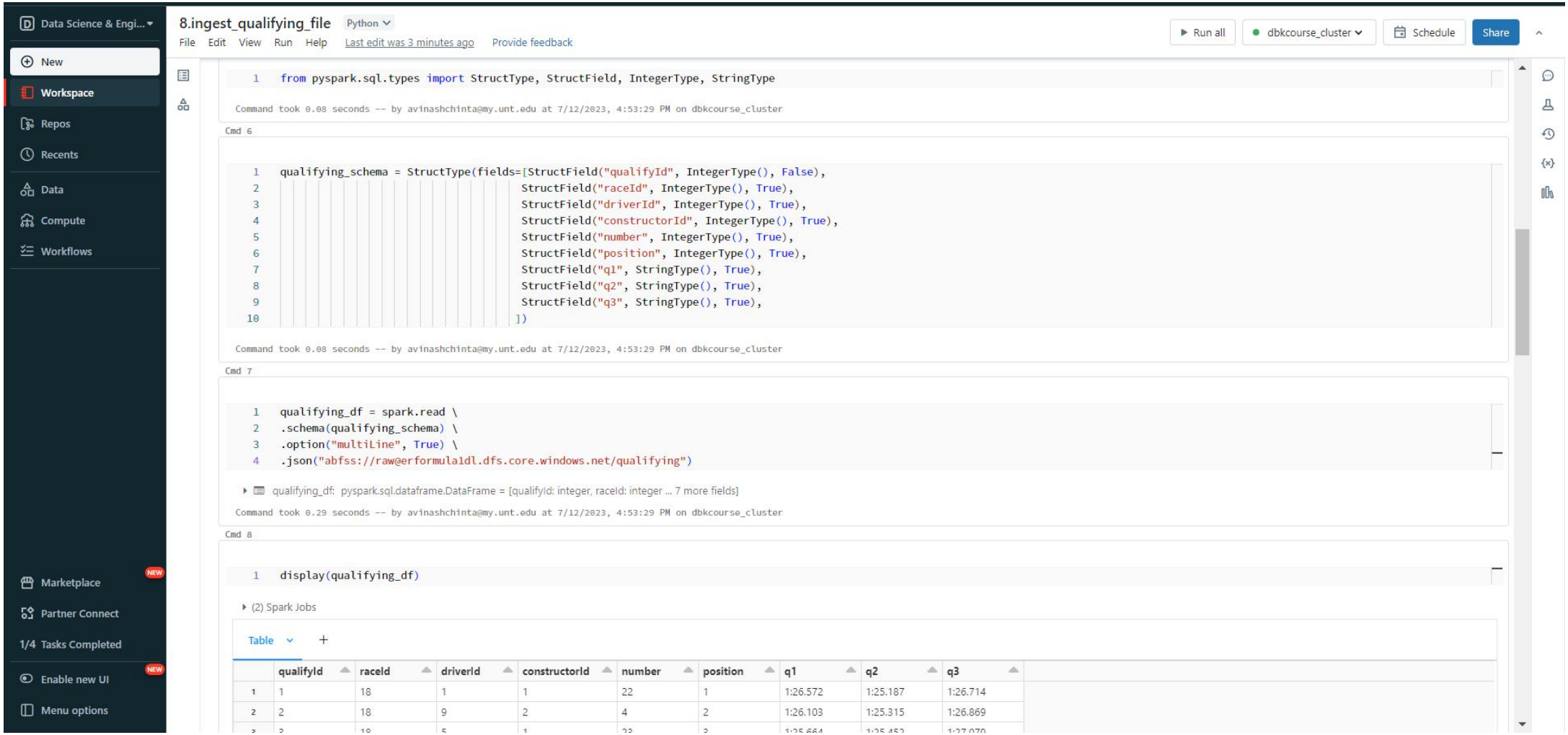
Authentication method: Access key (Switch to Azure AD User Account)
Location: processed / lap_times

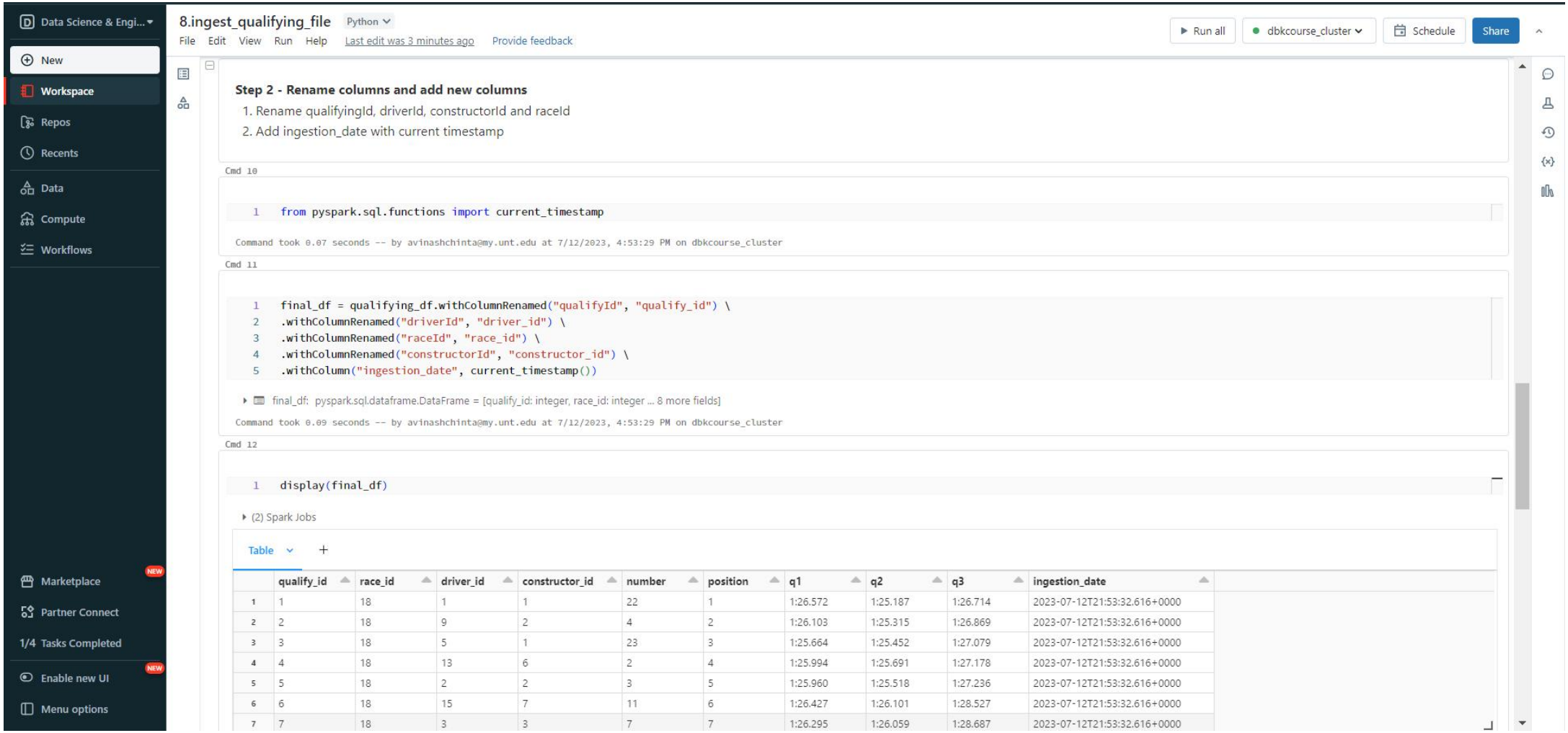
Search blobs by prefix (case-sensitive)

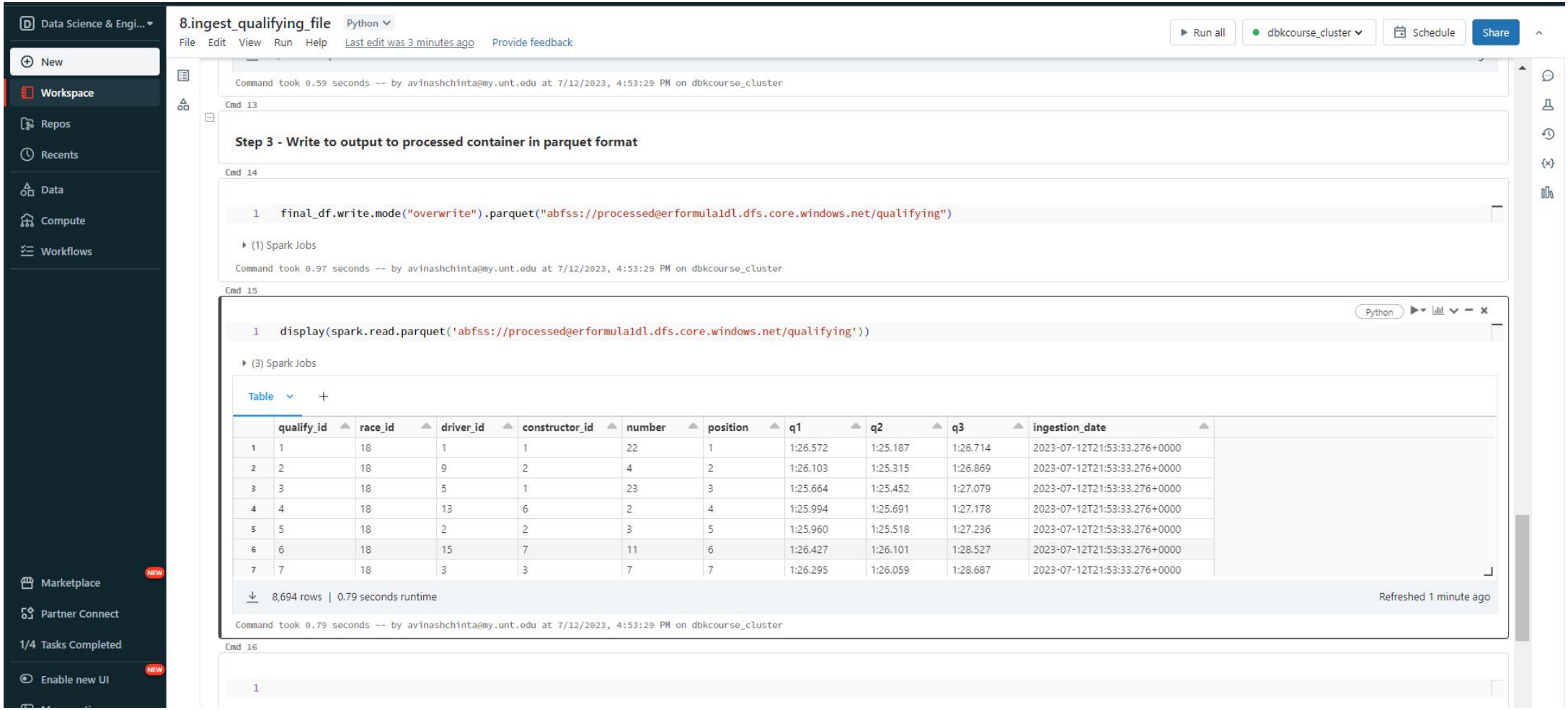
Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
_committed_2582663847247224090	7/12/2023, 4:42:53 PM	Hot (Inferred)		Block blob	519 B	Available
_started_2582663847247224090	7/12/2023, 4:42:48 PM	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	7/12/2023, 4:42:53 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-2582663847247224090-70bfa940-a283-4b5e-a307-a...	7/12/2023, 4:42:52 PM	Hot (Inferred)		Block blob	853.67 KiB	Available
part-00001-tid-2582663847247224090-70bfa940-a283-4b5e-a307-a...	7/12/2023, 4:42:52 PM	Hot (Inferred)		Block blob	833.48 KiB	Available
part-00002-tid-2582663847247224090-70bfa940-a283-4b5e-a307-a...	7/12/2023, 4:42:52 PM	Hot (Inferred)		Block blob	856.23 KiB	Available
part-00003-tid-2582663847247224090-70bfa940-a283-4b5e-a307-a...	7/12/2023, 4:42:52 PM	Hot (Inferred)		Block blob	805.47 KiB	Available
part-00004-tid-2582663847247224090-70bfa940-a283-4b5e-a307-a...	7/12/2023, 4:42:53 PM	Hot (Inferred)		Block blob	791.22 KiB	Available


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
8.ingest_qualifying_file
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Ingest qualifying json files
Cmd 2
Step 1 - Read the JSON file using the spark dataframe reader API
Cmd 3
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net"))
(2) Spark Jobs
Table
path name size modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/circuits.csv circuits.csv 10044 1689135355000
2 abfss://raw@erformula1dl.dfs.core.windows.net/constructors.json constructors.json 30415 1689135355000
3 abfss://raw@erformula1dl.dfs.core.windows.net/drivers.json drivers.json 180812 1689135355000
4 abfss://raw@erformula1dl.dfs.core.windows.net/lap_times/ lap_times/ 0 1689135393000
5 abfss://raw@erformula1dl.dfs.core.windows.net/pit_stops.json pit_stops.json 1369387 1689135357000
6 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/ qualifying/ 0 1689135408000
7 abfss://raw@erformula1dl.dfs.core.windows.net/races.csv races.csv 116847 1689135355000
8 rows | 1.20 seconds runtime
Refreshed now
Command took 1.20 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 4
1 display(dbutils.fs.ls("abfss://raw@erformula1dl.dfs.core.windows.net/qualifying"))
(2) Spark Jobs
Table
path name size modificationTime
1 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/.DS_Store .DS_Store 6148 1689135408000
2 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/qualifying_split_1.json qualifying_split_1.json 948426 1689135409000
3 abfss://raw@erformula1dl.dfs.core.windows.net/qualifying/qualifying_split_2.json qualifying_split_2.json 718351 1689135409000


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
8.ingest_qualifying_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
from pyspark.sql.types import StructType, StructField, IntegerType, StringType
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 6
qualifying_schema = StructType(fields=[StructField("qualifyId", IntegerType(), False),
StructField("raceId", IntegerType(), True),
StructField("driverId", IntegerType(), True),
StructField("constructorId", IntegerType(), True),
StructField("number", IntegerType(), True),
StructField("position", IntegerType(), True),
StructField("q1", StringType(), True),
StructField("q2", StringType(), True),
StructField("q3", StringType(), True),
])
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 7
qualifying_df = spark.read \
.schema(qualifying_schema) \
.option("multiLine", True) \
.json("abfss://raw@erformula1dl.dfs.core.windows.net/qualifying")
qualifying_df: pyspark.sql.dataframe.DataFrame = [qualifyId: integer, raceId: integer ... 7 more fields]
Command took 0.29 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 8
display(qualifying_df)
(2) Spark Jobs
Table
qualifyId raceId driverId constructorId number position q1 q2 q3
1 1 18 1 1 22 1 1:26.572 1:25.187 1:26.714
2 2 18 9 2 4 2 1:26.103 1:25.315 1:26.869


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
8.ingest_qualifying_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Step 2 - Rename columns and add new columns
1. Rename qualifyingId, driverId, constructorId and raceId
2. Add ingestion_date with current timestamp
Cmd 10
from pyspark.sql.functions import current_timestamp
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 11
final_df = qualifying_df.withColumnRenamed("qualifyId", "qualify_id") \
.withColumnRenamed("driverId", "driver_id") \
.withColumnRenamed("raceId", "race_id") \
.withColumnRenamed("constructorId", "constructor_id") \
.withColumn("ingestion_date", current_timestamp())
final_df: pyspark.sql.dataframe.DataFrame = [qualify_id: integer, race_id: integer ... 8 more fields]
Command took 0.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 12
display(final_df)
(2) Spark Jobs
Table
qualify_id race_id driver_id constructor_id number position q1 q2 q3 ingestion_date
1 1 18 1 1 22 1 1:26.572 1:25.187 1:26.714 2023-07-12T21:53:32.616+0000
2 2 18 9 2 4 2 1:26.103 1:25.315 1:26.869 2023-07-12T21:53:32.616+0000
3 3 18 5 1 23 3 1:25.664 1:25.452 1:27.079 2023-07-12T21:53:32.616+0000
4 4 18 13 6 2 4 1:25.994 1:25.691 1:27.178 2023-07-12T21:53:32.616+0000
5 5 18 2 2 3 5 1:25.960 1:25.518 1:27.236 2023-07-12T21:53:32.616+0000
6 6 18 15 7 11 6 1:26.427 1:26.101 1:28.527 2023-07-12T21:53:32.616+0000
7 7 18 3 3 7 7 1:26.295 1:26.059 1:28.687 2023-07-12T21:53:32.616+0000


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
8.ingest_qualifying_file
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
Command took 0.59 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 13
Step 3 - Write to output to processed container in parquet format
Cmd 14
1 final_df.write.mode("overwrite").parquet("abfss://processed@erformula1dl.dfs.core.windows.net/qualifying")
(1) Spark Jobs
Command took 0.97 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 15
Python
1 display(spark.read.parquet('abfss://processed@erformula1dl.dfs.core.windows.net/qualifying'))
(3) Spark Jobs
Table
qualify_id race_id driver_id constructor_id number position q1 q2 q3 ingestion_date
1 1 18 1 1 22 1 1:26.572 1:25.187 1:26.714 2023-07-12T21:53:33.276+0000
2 2 18 9 2 4 2 1:26.103 1:25.315 1:26.869 2023-07-12T21:53:33.276+0000
3 3 18 5 1 23 3 1:25.664 1:25.452 1:27.079 2023-07-12T21:53:33.276+0000
4 4 18 13 6 2 4 1:25.994 1:25.691 1:27.178 2023-07-12T21:53:33.276+0000
5 5 18 2 2 3 5 1:25.960 1:25.518 1:27.236 2023-07-12T21:53:33.276+0000
6 6 18 15 7 11 6 1:26.427 1:26.101 1:28.527 2023-07-12T21:53:33.276+0000
7 7 18 3 3 7 7 1:26.295 1:26.059 1:28.687 2023-07-12T21:53:33.276+0000
8,694 rows | 0.79 seconds runtime
Refreshed 1 minute ago
Command took 0.79 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 4:53:29 PM on dbkcourse_cluster
Cmd 16
1

Home >

processed

Container

Search

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

Settings

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

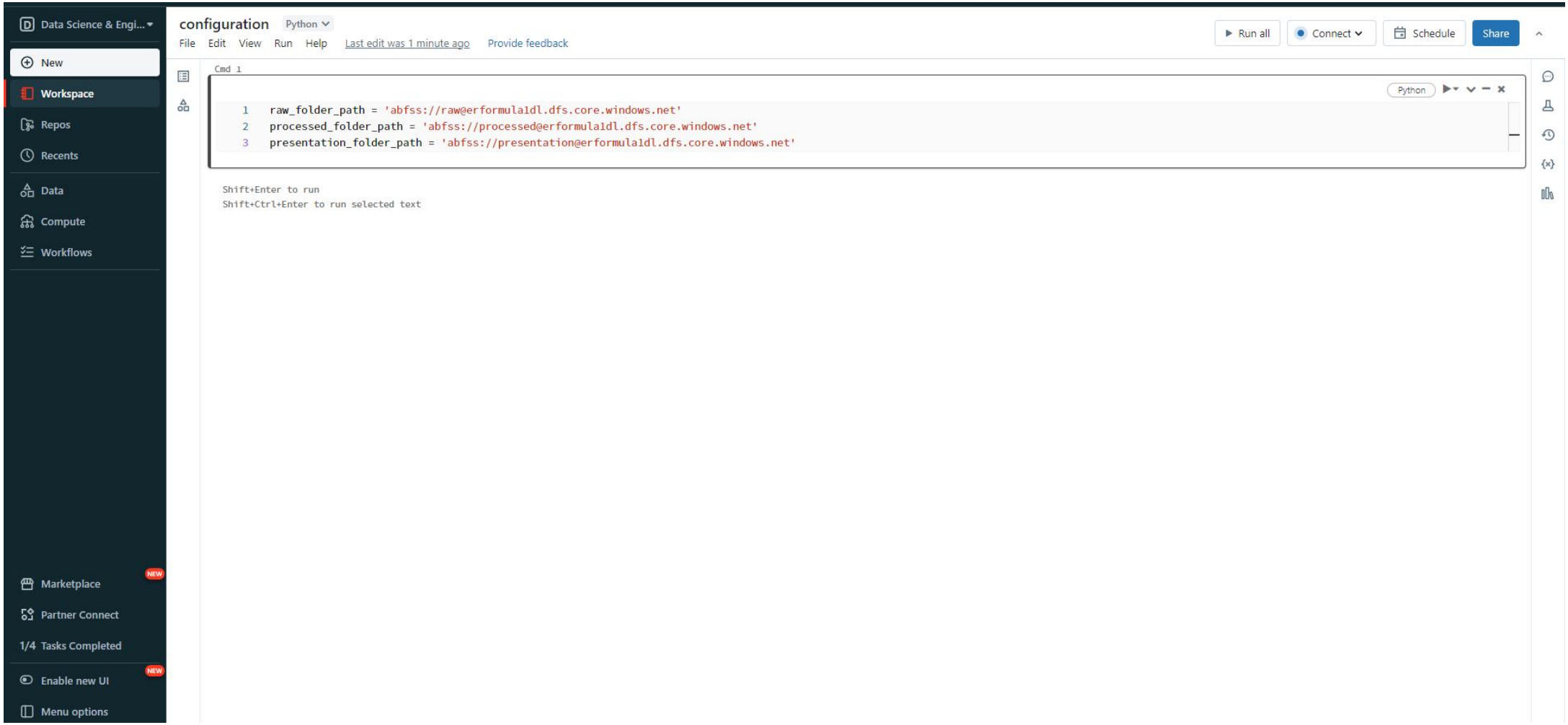
Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

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

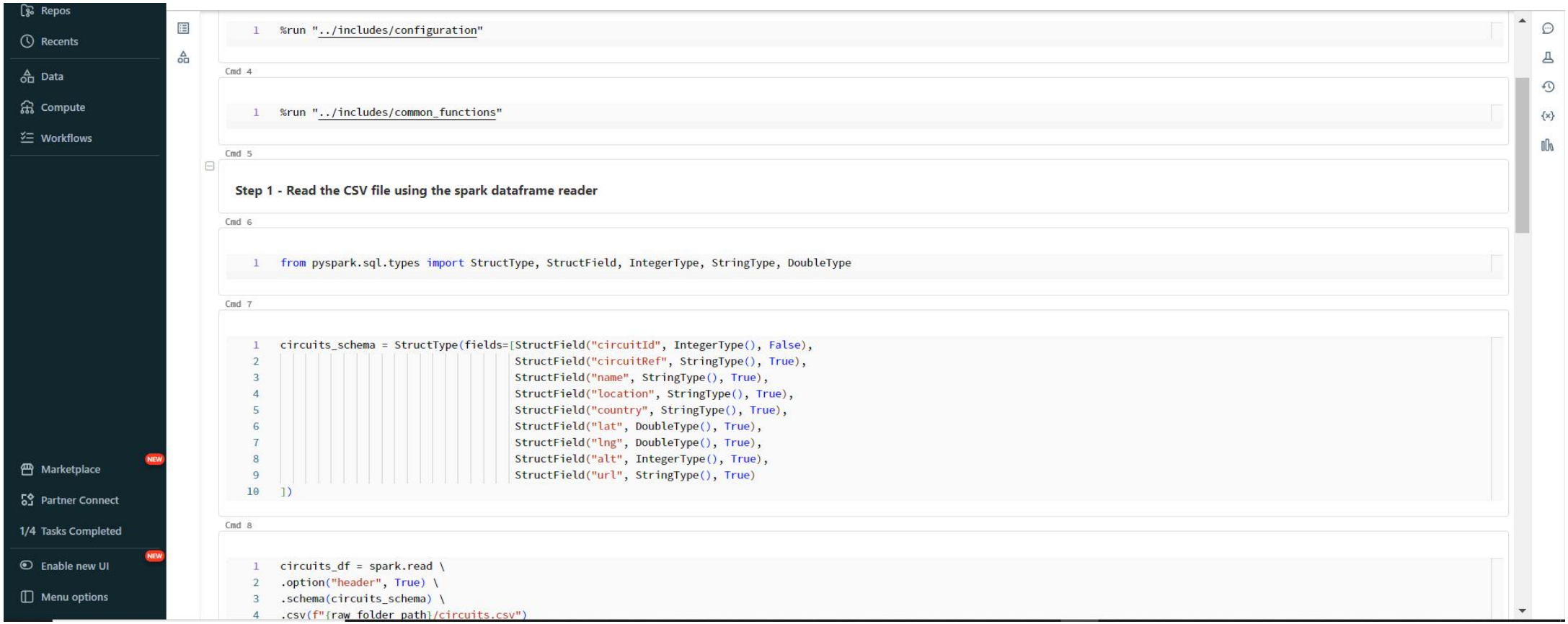
Search blobs by prefix (case-sensitive)

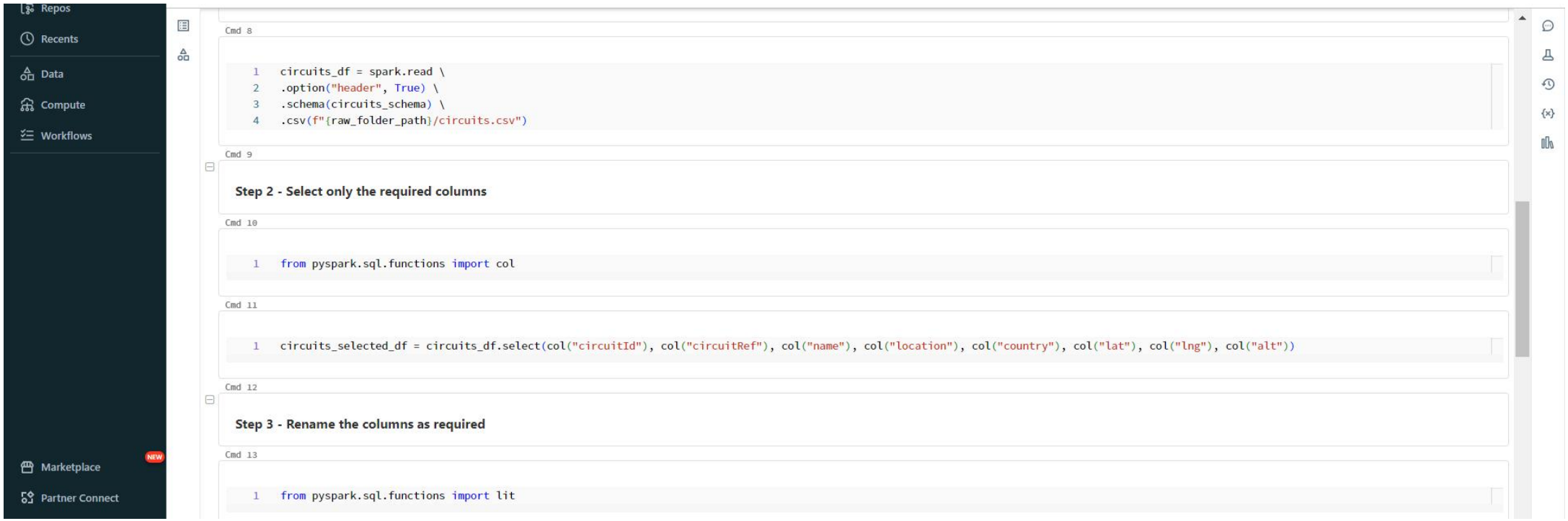
Show deleted objects

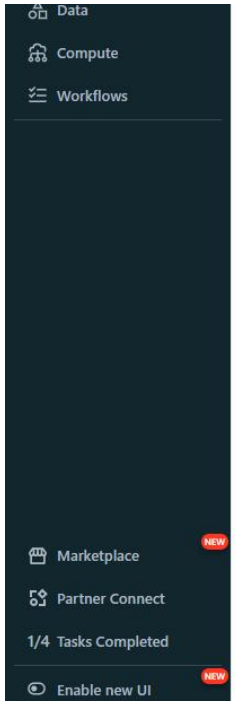
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
_committed_5797577338261052253	7/12/2023, 4:53:33 PM	Hot (Inferred)		Block blob	222 B	Available
_started_5797577338261052253	7/12/2023, 4:53:33 PM	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	7/12/2023, 4:53:34 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-5797577338261052253-04a328cf-e1da-4e00-8f9d-b...	7/12/2023, 4:53:33 PM	Hot (Inferred)		Block blob	81.03 KiB	Available
part-00001-tid-5797577338261052253-04a328cf-e1da-4e00-8f9d-b...	7/12/2023, 4:53:33 PM	Hot (Inferred)		Block blob	75.27 KiB	Available


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
Partner Connect
1/4 Tasks Completed
Enable new UI
NEW
Menu options
configuration
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Run all
Connect
Schedule
Share
Cmd 1
Python
1 raw_folder_path = 'abfss://raw@erformula1dl.dfs.core.windows.net'
2 processed_folder_path = 'abfss://processed@erformula1dl.dfs.core.windows.net'
3 presentation_folder_path = 'abfss://presentation@erformula1dl.dfs.core.windows.net'
Shift+Enter to run
Shift+Ctrl+Enter to run selected text


Data Science & Engi...
New
Workspace
Repos
Recents
Data
common_functions
Python
File Edit View Run Help
Last edit was 17 minutes ago
Provide feedback
Run all
Connect
Schedule
Share
Cmd 1
1 from pyspark.sql.functions import current_timestamp
2 def add_ingestion_date(input_df):
3 output_df = input_df.withColumn("ingestion_date", current_timestamp())
4 return output_df
Cmd 2


Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
Partner Connect
1/4 Tasks Completed
Enable new UI
NEW
Menu options
1 %run "../includes/configuration"
Cmd 4
1 %run "../includes/common_functions"
Cmd 5
Step 1 - Read the CSV file using the spark dataframe reader
Cmd 6
1 from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
Cmd 7
1 circuits_schema = StructType(fields=[StructField("circuitId", IntegerType(), False),
2 StructField("circuitRef", StringType(), True),
3 StructField("name", StringType(), True),
4 StructField("location", StringType(), True),
5 StructField("country", StringType(), True),
6 StructField("lat", DoubleType(), True),
7 StructField("lng", DoubleType(), True),
8 StructField("alt", IntegerType(), True),
9 StructField("url", StringType(), True)
10])
Cmd 8
1 circuits_df = spark.read \
2 .option("header", True) \
3 .schema(circuits_schema) \
4 .csv(f"{raw_folder_path}/circuits.csv")


Repos
Recents
Data
Compute
Workflows
Marketplace
NEW
Partner Connect
Cmd 8
1 circuits_df = spark.read \
2 .option("header", True) \
3 .schema(circuits_schema) \
4 .csv(f"{raw_folder_path}/circuits.csv")
Cmd 9
Step 2 - Select only the required columns
Cmd 10
1 from pyspark.sql.functions import col
Cmd 11
1 circuits_selected_df = circuits_df.select(col("circuitId"), col("circuitRef"), col("name"), col("location"), col("country"), col("lat"), col("lng"), col("alt"))
Cmd 12
Step 3 - Rename the columns as required
Cmd 13
1 from pyspark.sql.functions import lit




```
    .withColumnRenamed("alt", "altitude") \
    .withColumn("data_source", lit(v_data_source))
```

Cmd 15

Step 4 - Add ingestion date to the dataframe

Cmd 16

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

Cmd 17

Step 5 - Write data to datalake as parquet

Cmd 18

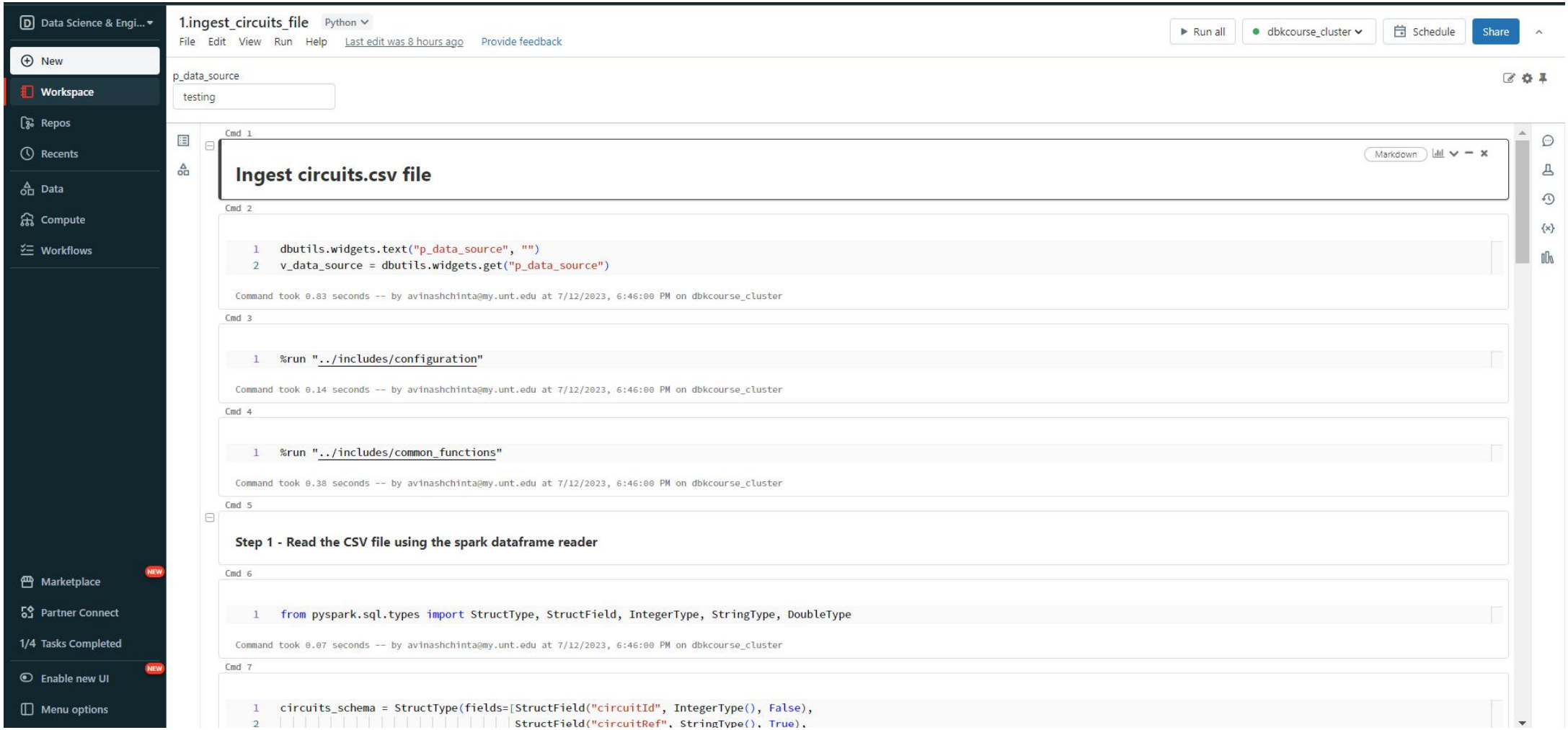
```
circuits_final_df.write.mode("overwrite").parquet(f"{processed_folder_path}/circuits")
```

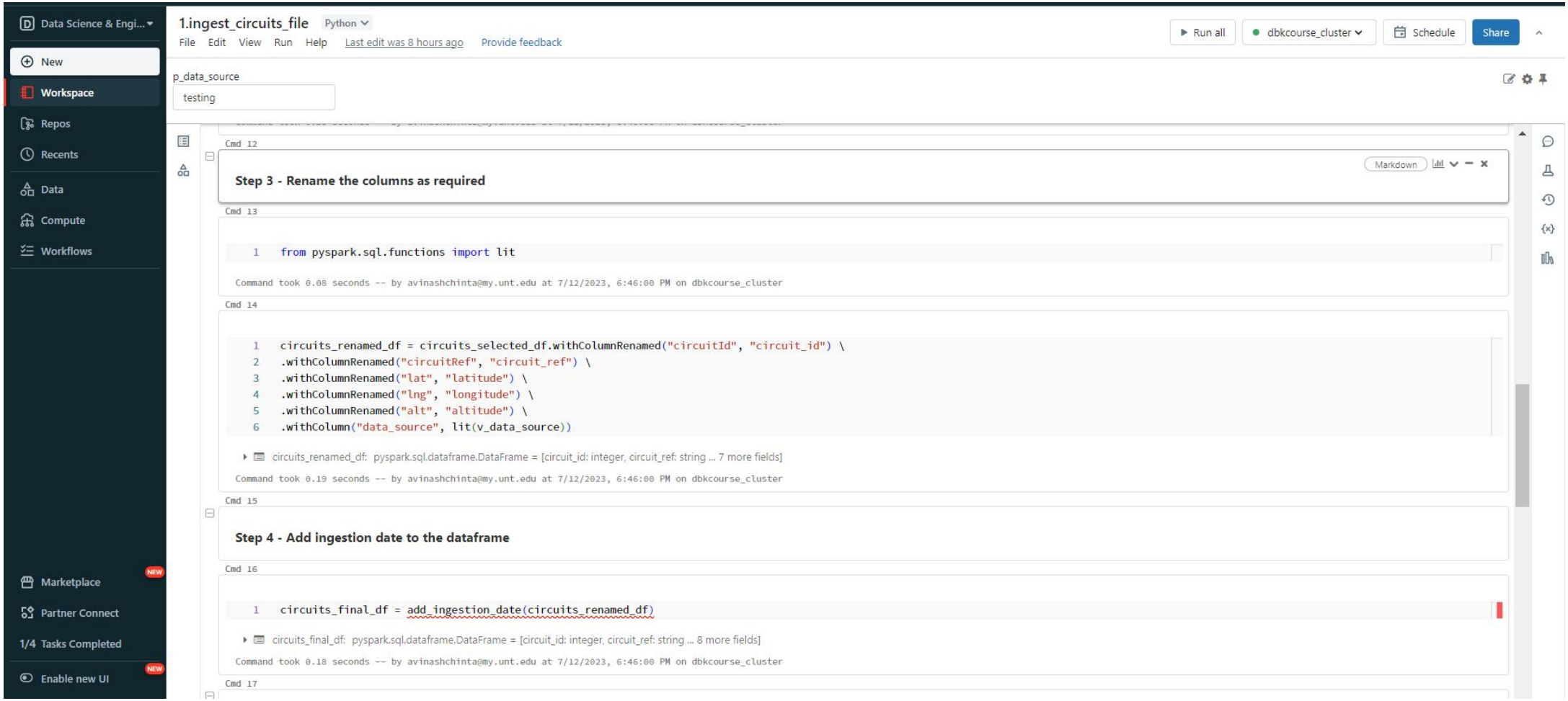
Cmd 19

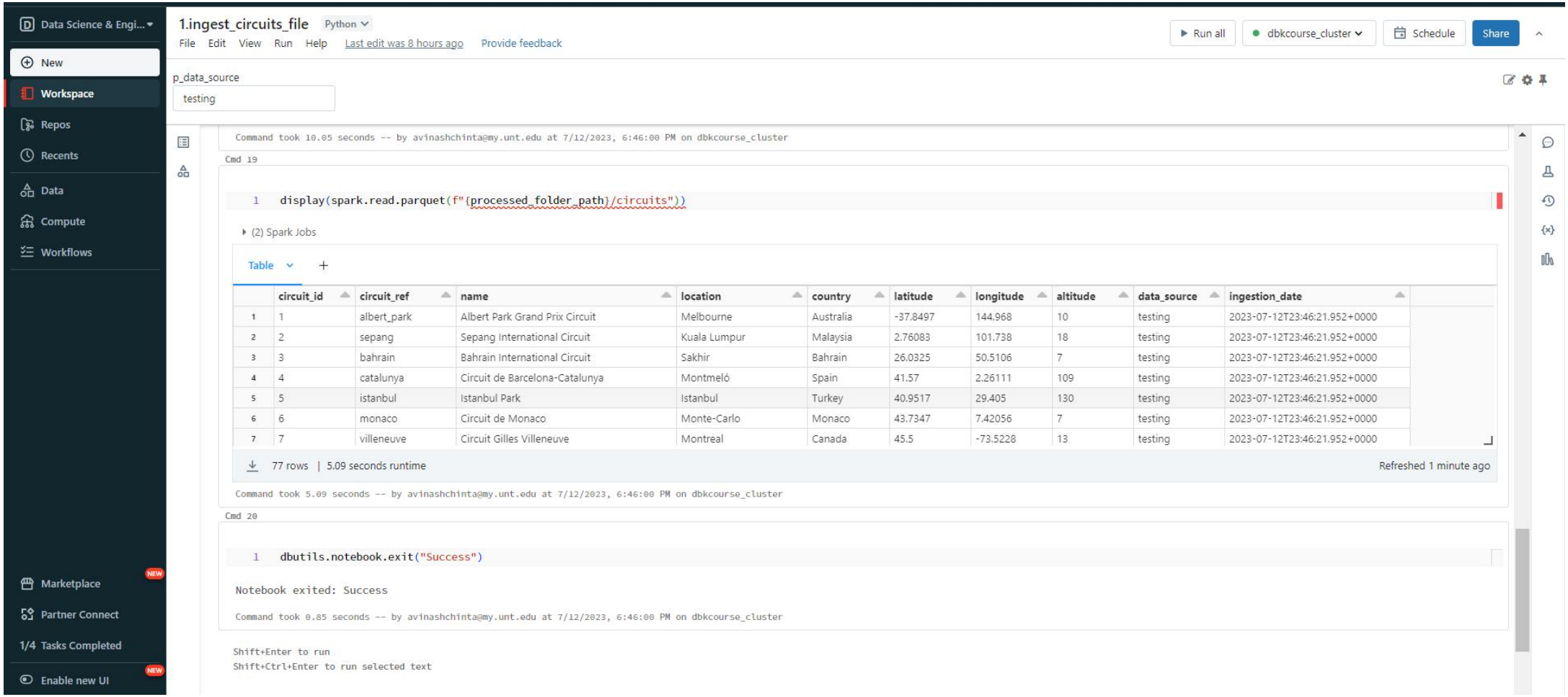
```
display(spark.read.parquet(f"{processed_folder_path}/circuits"))
```

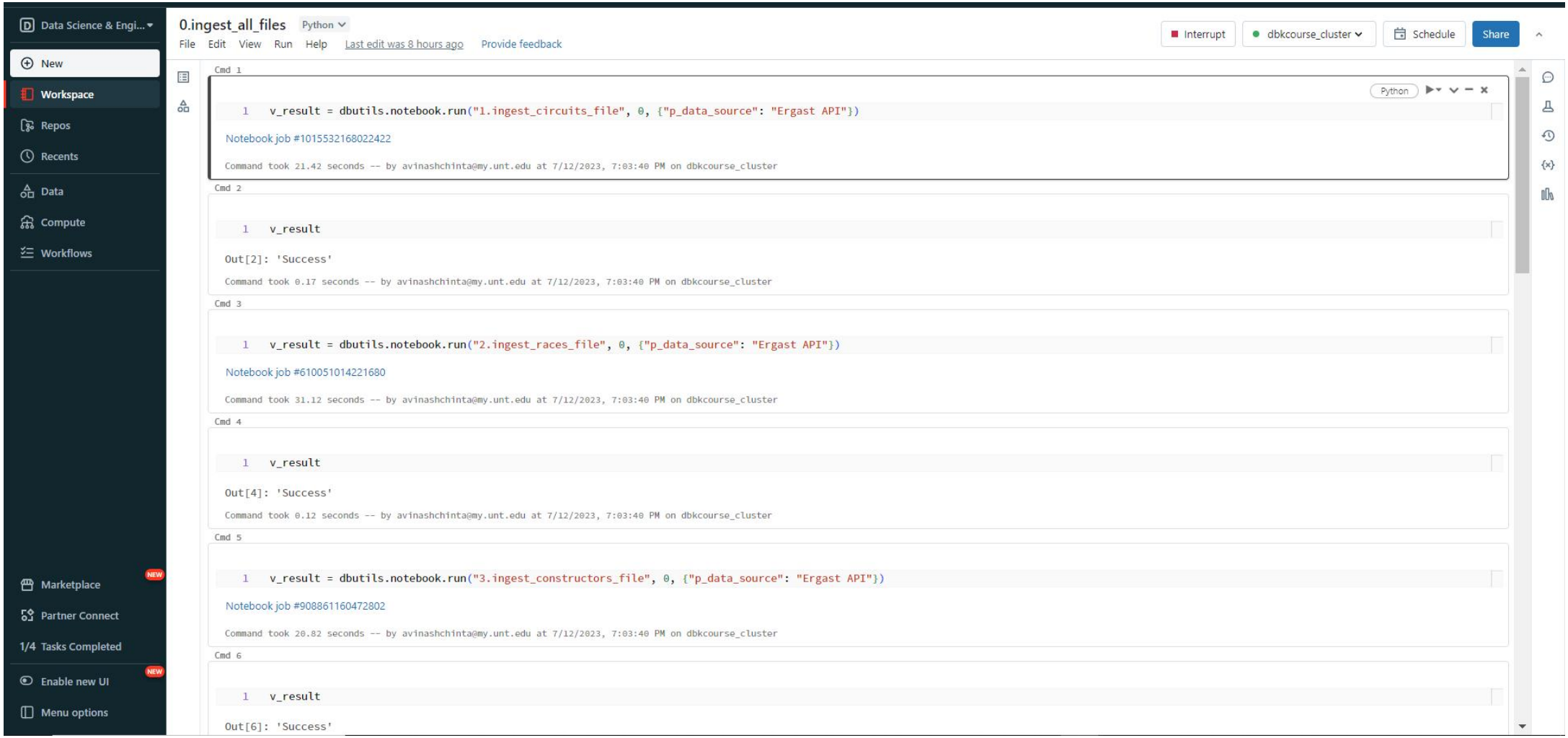
Cmd 20

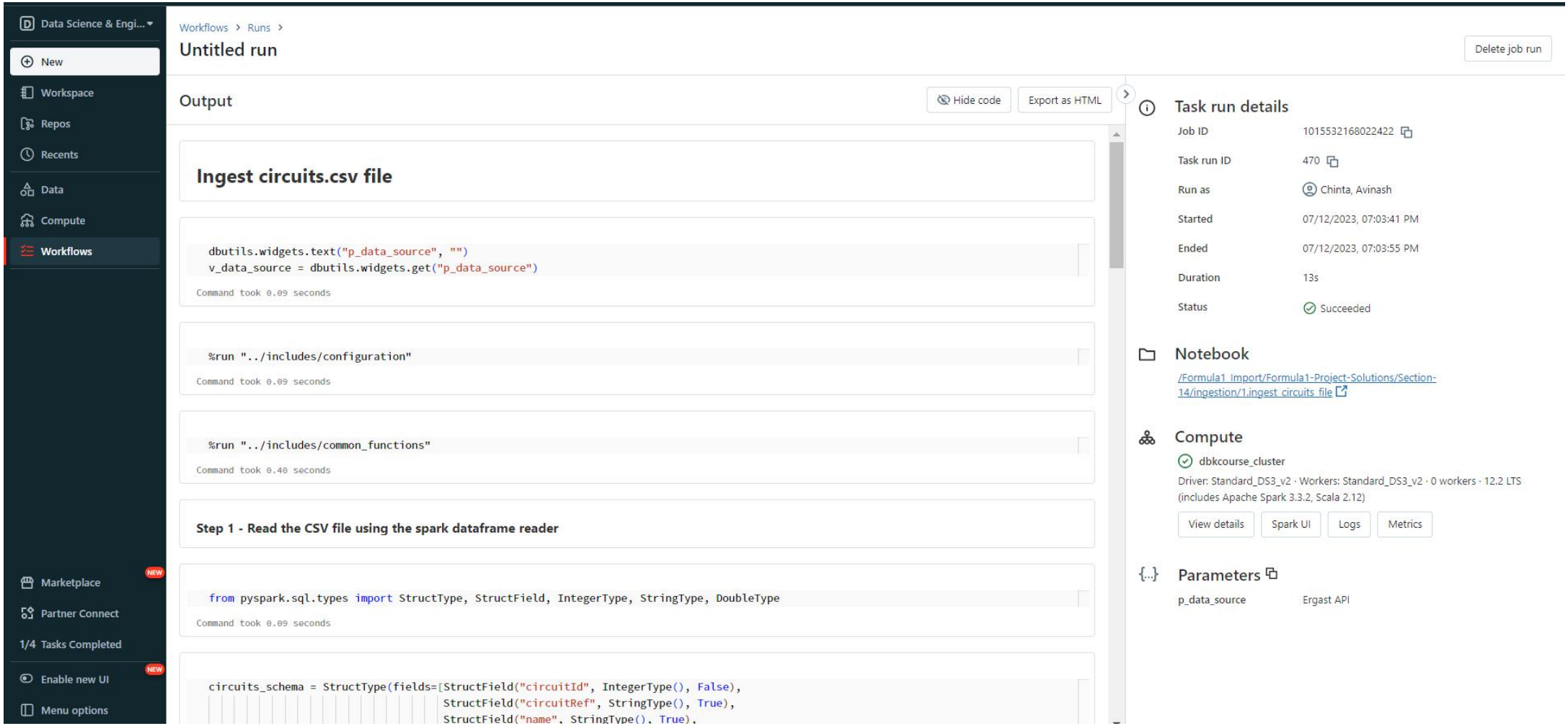
```
dbutils.notebook.exit("Success")
```

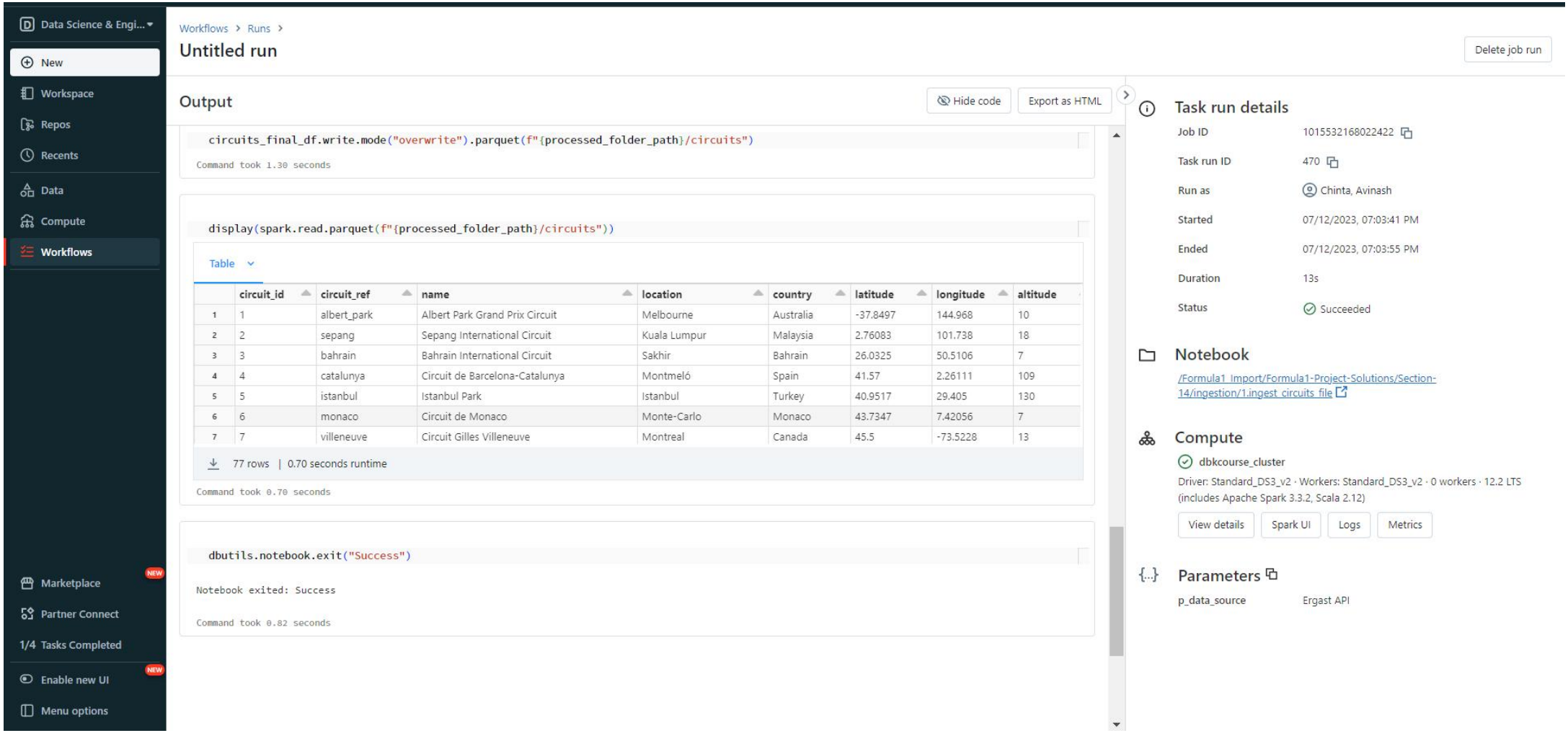

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was 8 hours ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
p_data_source
testing
Cmd 1
Markdown
Ingest circuits.csv file
Cmd 2
dbutils.widgets.text("p_data_source", "")
v_data_source = dbutils.widgets.get("p_data_source")
Command took 0.83 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 3
%run "../includes/configuration"
Command took 0.14 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 4
%run "../includes/common_functions"
Command took 0.38 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 5
Step 1 - Read the CSV file using the spark dataframe reader
Cmd 6
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
Command took 0.07 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 7
circuits_schema = StructType(fields=[StructField("circuitId", IntegerType(), False),
StructField("circuitRef", StringType(), True),

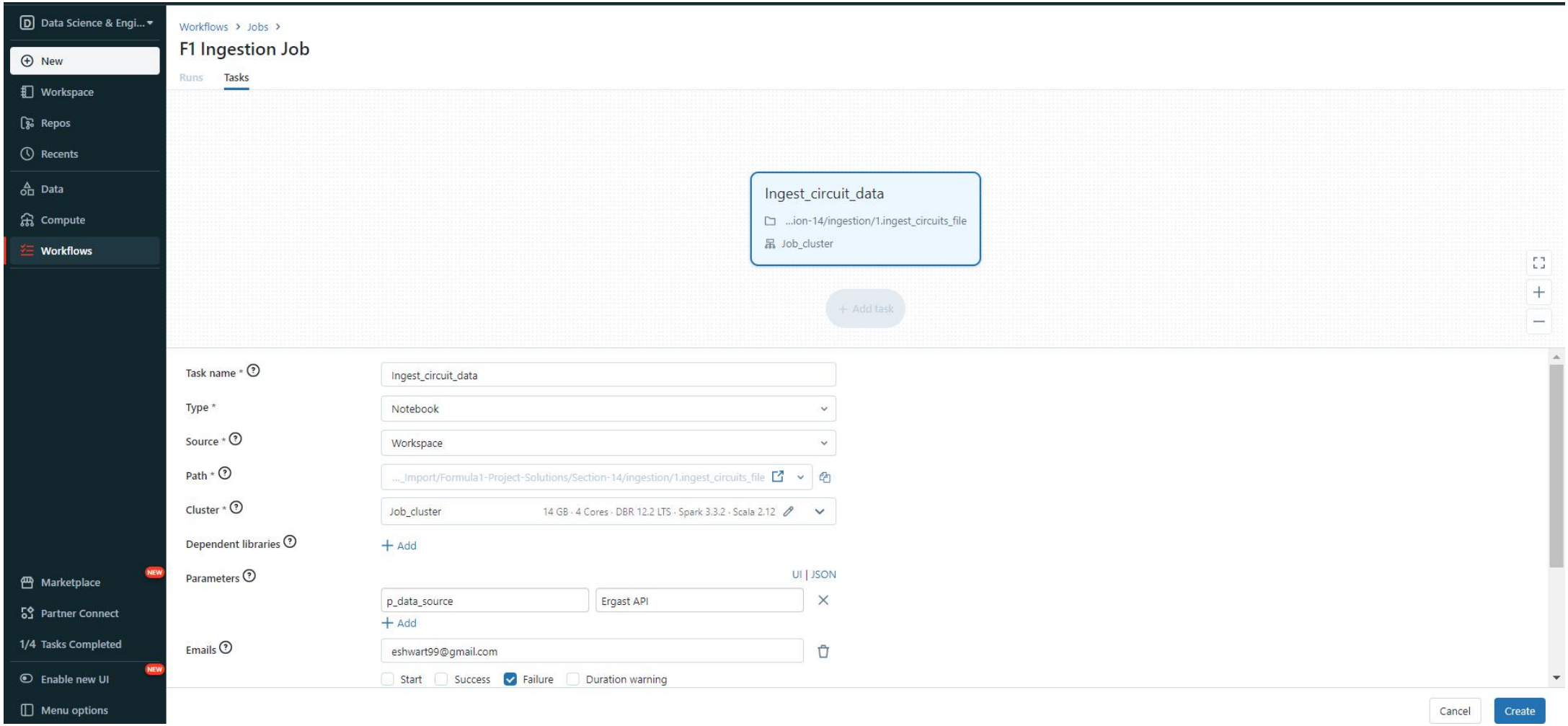

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was 8 hours ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
p_data_source
testing
Cmd 12
Markdown
Step 3 - Rename the columns as required
Cmd 13
from pyspark.sql.functions import lit
Command took 0.08 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 14
circuits_renamed_df = circuits_selected_df.withColumnRenamed("circuitId", "circuit_id") \
.withColumnRenamed("circuitRef", "circuit_ref") \
.withColumnRenamed("lat", "latitude") \
.withColumnRenamed("lng", "longitude") \
.withColumnRenamed("alt", "altitude") \
.withColumn("data_source", lit(v_data_source))
circuits_renamed_df: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string ... 7 more fields]
Command took 0.19 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 15
Step 4 - Add ingestion date to the dataframe
Cmd 16
circuits_final_df = add_ingestion_date(circuits_renamed_df)
circuits_final_df: pyspark.sql.dataframe.DataFrame = [circuit_id: integer, circuit_ref: string ... 8 more fields]
Command took 0.18 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 17

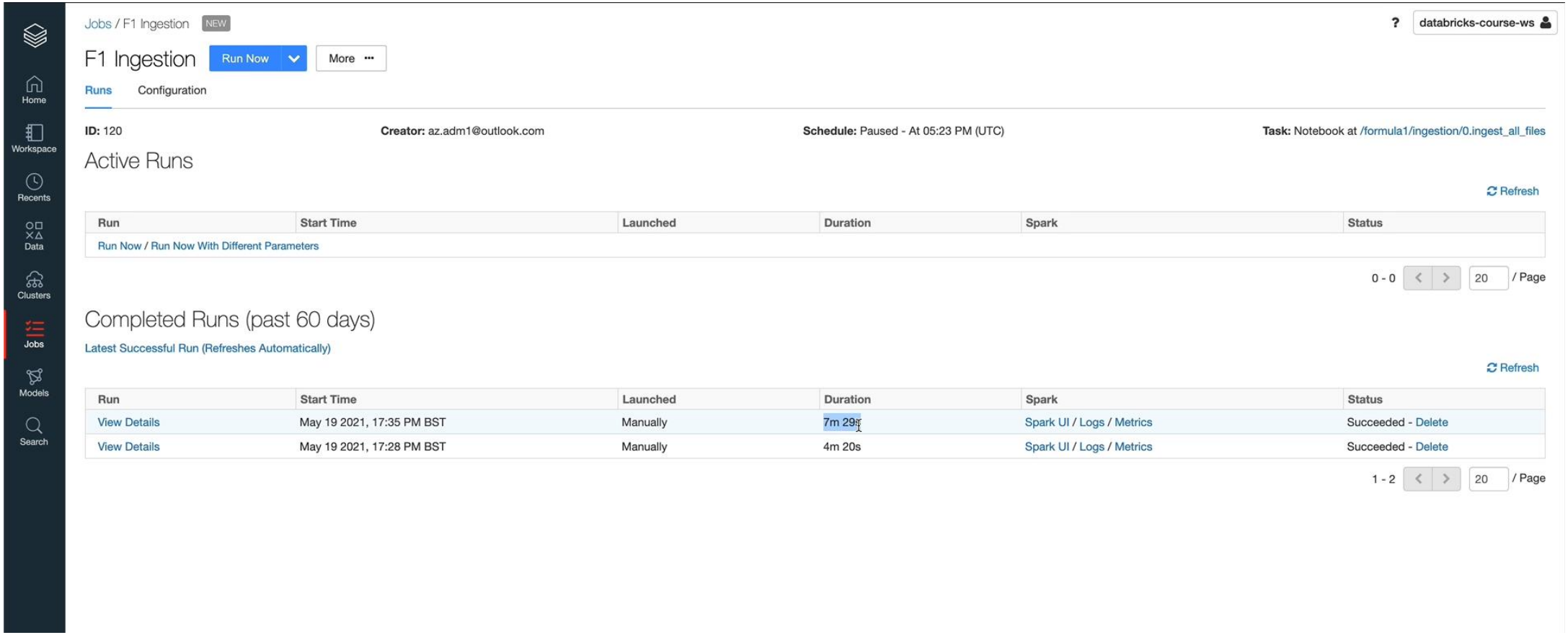

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
1.ingest_circuits_file
Python
File Edit View Run Help
Last edit was 8 hours ago
Provide feedback
Run all
dbkcourse_cluster
Schedule
Share
p_data_source
testing
Command took 10.05 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 19
1 display(spark.read.parquet(f"{processed_folder_path}/circuits"))
(2) Spark Jobs
Table
circuit_id | circuit_ref | name | location | country | latitude | longitude | altitude | data_source | ingestion_date
1 | 1 | albert_park | Albert Park Grand Prix Circuit | Melbourne | Australia | -37.8497 | 144.968 | 10 | testing | 2023-07-12T23:46:21.952+0000
2 | 2 | sepang | Sepang International Circuit | Kuala Lumpur | Malaysia | 2.76083 | 101.738 | 18 | testing | 2023-07-12T23:46:21.952+0000
3 | 3 | bahrain | Bahrain International Circuit | Sakhir | Bahrain | 26.0325 | 50.5106 | 7 | testing | 2023-07-12T23:46:21.952+0000
4 | 4 | catalunya | Circuit de Barcelona-Catalunya | Montmeló | Spain | 41.57 | 2.26111 | 109 | testing | 2023-07-12T23:46:21.952+0000
5 | 5 | istanbul | Istanbul Park | Istanbul | Turkey | 40.9517 | 29.405 | 130 | testing | 2023-07-12T23:46:21.952+0000
6 | 6 | monaco | Circuit de Monaco | Monte-Carlo | Monaco | 43.7347 | 7.42056 | 7 | testing | 2023-07-12T23:46:21.952+0000
7 | 7 | villeneuve | Circuit Gilles Villeneuve | Montreal | Canada | 45.5 | -73.5228 | 13 | testing | 2023-07-12T23:46:21.952+0000
77 rows | 5.09 seconds runtime
Refreshed 1 minute ago
Command took 5.09 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Cmd 20
1 dbutils.notebook.exit("Success")
Notebook exited: Success
Command took 0.85 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 6:46:00 PM on dbkcourse_cluster
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

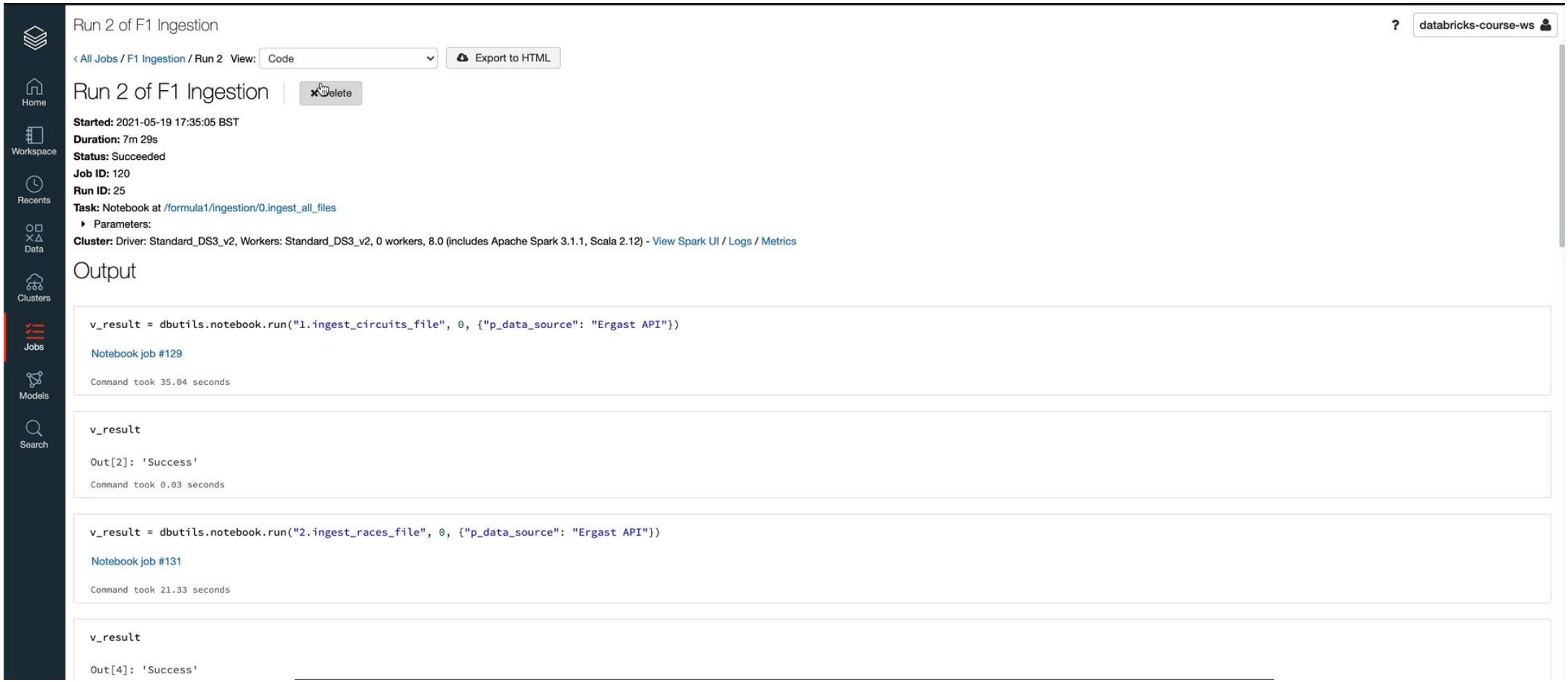

Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
0.ingest_all_files
Python
File Edit View Run Help
Last edit was 8 hours ago
Provide feedback
Interrupt
dbkcourse_cluster
Schedule
Share
Cmd 1
1 v_result = dbutils.notebook.run("1.ingest_circuits_file", 0, {"p_data_source": "Ergast API"})
Notebook job #1015532168022422
Command took 21.42 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 7:03:40 PM on dbkcourse_cluster
Cmd 2
1 v_result
Out[2]: 'Success'
Command took 0.17 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 7:03:40 PM on dbkcourse_cluster
Cmd 3
1 v_result = dbutils.notebook.run("2.ingest_races_file", 0, {"p_data_source": "Ergast API"})
Notebook job #610051014221680
Command took 31.12 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 7:03:40 PM on dbkcourse_cluster
Cmd 4
1 v_result
Out[4]: 'Success'
Command took 0.12 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 7:03:40 PM on dbkcourse_cluster
Cmd 5
1 v_result = dbutils.notebook.run("3.ingest_constructors_file", 0, {"p_data_source": "Ergast API"})
Notebook job #908861160472802
Command took 20.82 seconds -- by avinashchinta@my.unt.edu at 7/12/2023, 7:03:40 PM on dbkcourse_cluster
Cmd 6
1 v_result
Out[6]: 'Success'


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
Workflows > Runs >
Untitled run
Delete job run
Output
Hide code
Export as HTML
Ingest circuits.csv file
dbutils.widgets.text("p_data_source", "")
v_data_source = dbutils.widgets.get("p_data_source")
Command took 0.09 seconds
%run "../includes/configuration"
Command took 0.09 seconds
%run "../includes/common_functions"
Command took 0.40 seconds
Step 1 - Read the CSV file using the spark dataframe reader
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
Command took 0.09 seconds
circuits_schema = StructType(fields=[StructField("circuitId", IntegerType(), False),
StructField("circuitRef", StringType(), True),
StructField("name", StringType(), True),
Task run details
Job ID 1015532168022422
Task run ID 470
Run as Chinta, Avinash
Started 07/12/2023, 07:03:41 PM
Ended 07/12/2023, 07:03:55 PM
Duration 13s
Status Succeeded
Notebook
/Formula1_Import/Formula1-Project-Solutions/Section-14/ingestion/1.ingest_circuits_file
Compute
dbkcourse_cluster
Driver: Standard_DS3_v2 · Workers: Standard_DS3_v2 · 0 workers · 12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)
View details
Spark UI
Logs
Metrics
Parameters
p_data_source Ergast API


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
Workflows > Runs >
Untitled run
Delete job run
Output
Hide code
Export as HTML
circuits_final_df.write.mode("overwrite").parquet(f"{processed_folder_path}/circuits")
Command took 1.30 seconds
display(spark.read.parquet(f"{processed_folder_path}/circuits"))
Table
circuit_id | circuit_ref | name | location | country | latitude | longitude | altitude
1 | 1 | albert_park | Albert Park Grand Prix Circuit | Melbourne | Australia | -37.8497 | 144.968 | 10
2 | 2 | sepang | Sepang International Circuit | Kuala Lumpur | Malaysia | 2.76083 | 101.738 | 18
3 | 3 | bahrain | Bahrain International Circuit | Sakhir | Bahrain | 26.0325 | 50.5106 | 7
4 | 4 | catalunya | Circuit de Barcelona-Catalunya | Montmeló | Spain | 41.57 | 2.26111 | 109
5 | 5 | istanbul | Istanbul Park | Istanbul | Turkey | 40.9517 | 29.405 | 130
6 | 6 | monaco | Circuit de Monaco | Monte-Carlo | Monaco | 43.7347 | 7.42056 | 7
7 | 7 | villeneuve | Circuit Gilles Villeneuve | Montreal | Canada | 45.5 | -73.5228 | 13
77 rows | 0.70 seconds runtime
Command took 0.70 seconds
dbutils.notebook.exit("Success")
Notebook exited: Success
Command took 0.82 seconds
Task run details
Job ID 1015532168022422
Task run ID 470
Run as Chinta, Avinash
Started 07/12/2023, 07:03:41 PM
Ended 07/12/2023, 07:03:55 PM
Duration 13s
Status Succeeded
Notebook
/Formula1_Import/Formula1-Project-Solutions/Section-14/ingestion/1.ingest_circuits_file
Compute
dbkcourse_cluster
Driver: Standard_DS3_v2 · Workers: Standard_DS3_v2 · 0 workers · 12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)
View details
Spark UI
Logs
Metrics
Parameters
p_data_source Ergast API


Data Science & Engi...
New
Workspace
Repos
Recents
Data
Compute
Workflows
Marketplace
Partner Connect
1/4 Tasks Completed
Enable new UI
Menu options
Workflows > Jobs >
F1 Ingestion Job
Runs
Tasks
Ingest_circuit_data
...ion-14/ingestion/1.ingest_circuits_file
Job_cluster
+ Add task
Task name *
Ingest_circuit_data
Type *
Notebook
Source *
Workspace
Path *
.../Import/Formula1-Project-Solutions/Section-14/ingestion/1.ingest_circuits_file
Cluster *
Job_cluster
14 GB · 4 Cores · DBR 12.2 LTS · Spark 3.3.2 · Scala 2.12
Dependent libraries
+ Add
Parameters
UI | JSON
p_data_source
Ergast API
+ Add
Emails
eshwart99@gmail.com
Start
Success
Failure
Duration warning
Cancel
Create


Jobs / F1 Ingestion
NEW
databricks-course-ws
F1 Ingestion
Run Now
More
Runs
Configuration
ID: 120
Creator: az.adm1@outlook.com
Schedule: Paused - At 05:23 PM (UTC)
Task: Notebook at /formula1/ingestion/0.ingest_all_files
Active Runs
Refresh
Run
Start Time
Launched
Duration
Spark
Status
Run Now / Run Now With Different Parameters
0 - 0
20
/ Page
Completed Runs (past 60 days)
Latest Successful Run (Refreshes Automatically)
Refresh
Run
Start Time
Launched
Duration
Spark
Status
View Details
May 19 2021, 17:35 PM BST
Manually
7m 29s
Spark UI / Logs / Metrics
Succeeded - Delete
View Details
May 19 2021, 17:28 PM BST
Manually
4m 20s
Spark UI / Logs / Metrics
Succeeded - Delete
1 - 2
20
/ Page
Home
Workspace
Recents
Data
Clusters
Jobs
Models
Search


Run 2 of F1 Ingestion
databricks-course-ws
< All Jobs / F1 Ingestion / Run 2 View: Code
Export to HTML
Run 2 of F1 Ingestion
Delete
Started: 2021-05-19 17:35:05 BST
Duration: 7m 29s
Status: Succeeded
Job ID: 120
Run ID: 25
Task: Notebook at /formula1/ingestion/0.ingest_all_files
Parameters:
Cluster: Driver: Standard_DS3_v2, Workers: Standard_DS3_v2, 0 workers, 8.0 (includes Apache Spark 3.1.1, Scala 2.12) - View Spark UI / Logs / Metrics
Output
v_result = dbutils.notebook.run("1.ingest_circuits_file", 0, {"p_data_source": "Ergast API"})
Notebook job #129
Command took 35.04 seconds
v_result
Out[2]: 'Success'
Command took 0.03 seconds
v_result = dbutils.notebook.run("2.ingest_races_file", 0, {"p_data_source": "Ergast API"})
Notebook job #131
Command took 21.33 seconds
v_result
Out[4]: 'Success'
Home
Workspace
Recents
Data
Clusters
Jobs
Models
Search