Northeastern University
College *of* Engineering

Data Warehousing & Integration
IE 6750
FALL 2024

[Formula 1 Data Pipeline]

Milestone 6

Group 10

Avani Kala

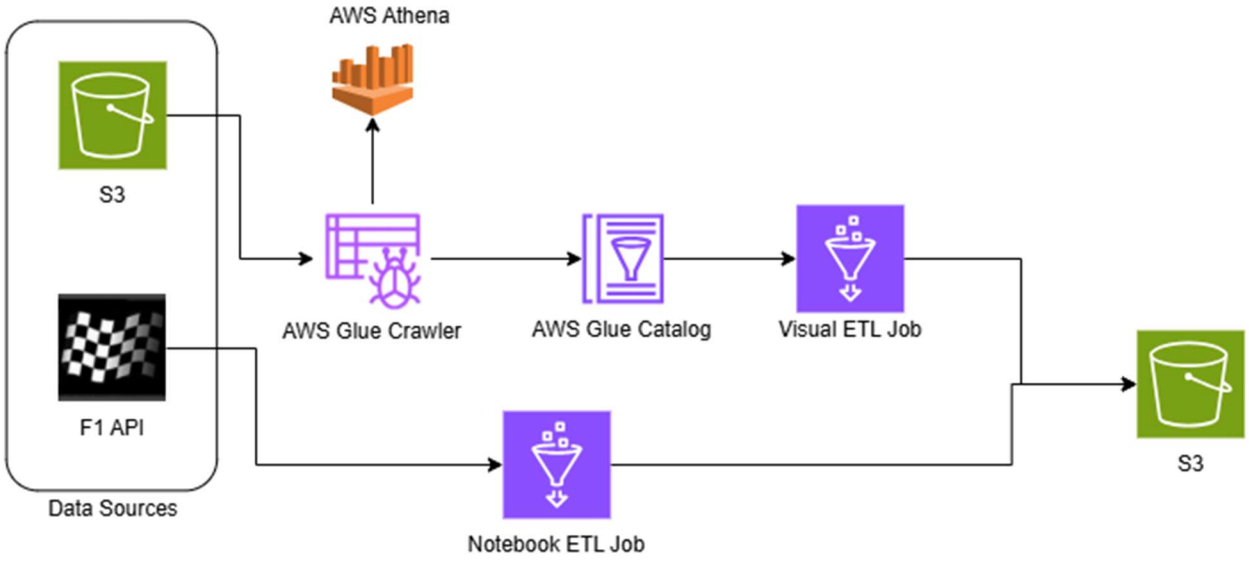
Sri Sai Prabhath Reddy Gudipalli

gudipalli.s@northeastern.edu

kala.a@northeastern.edu

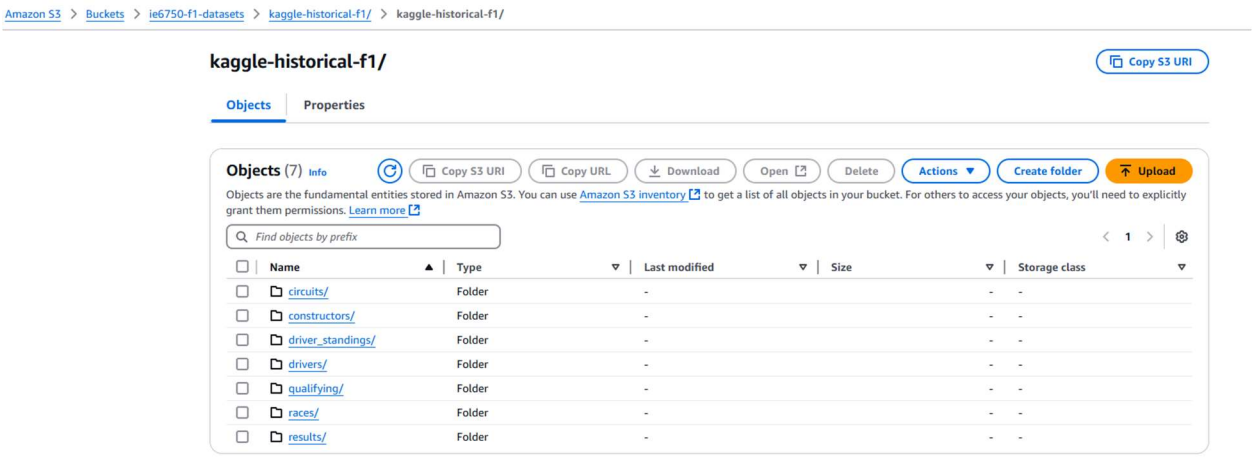
Submission Date: 11/28/2024

Architecture diagram of data pipeline:

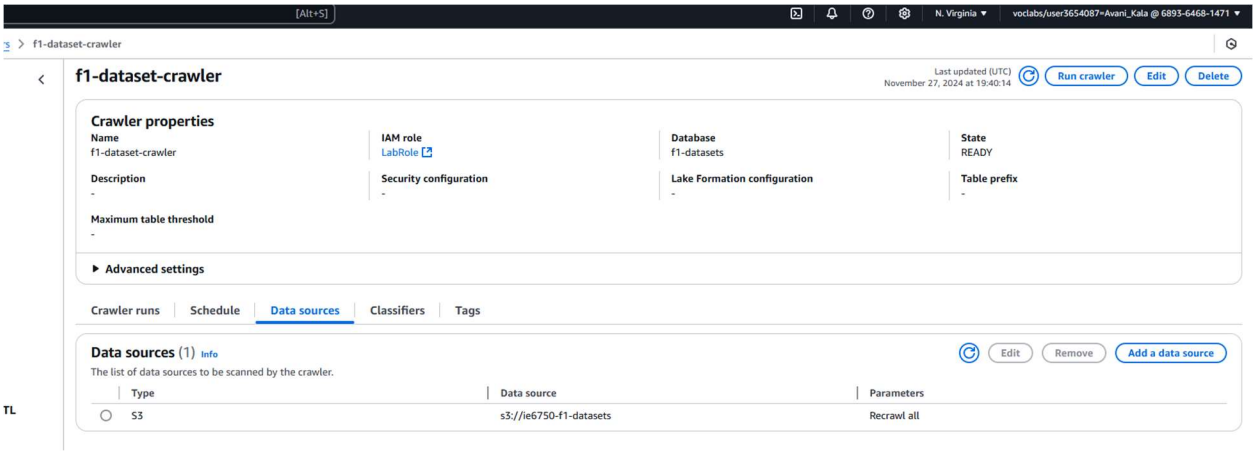


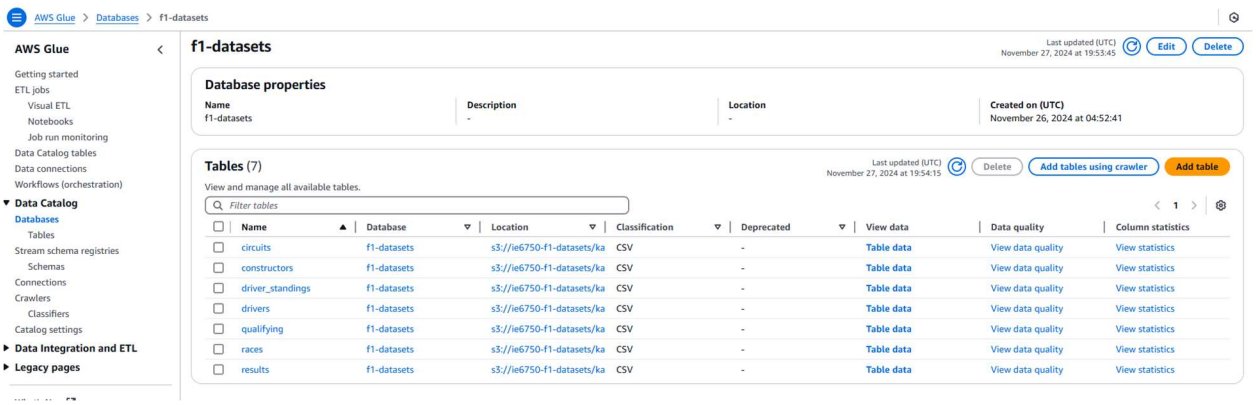
Steps involved in building data pipeline:

1. Loading csv files in AWS S3 bucket:
   Loaded the csv data into S3 bucket by manual upload. Separate folders are created for each file.

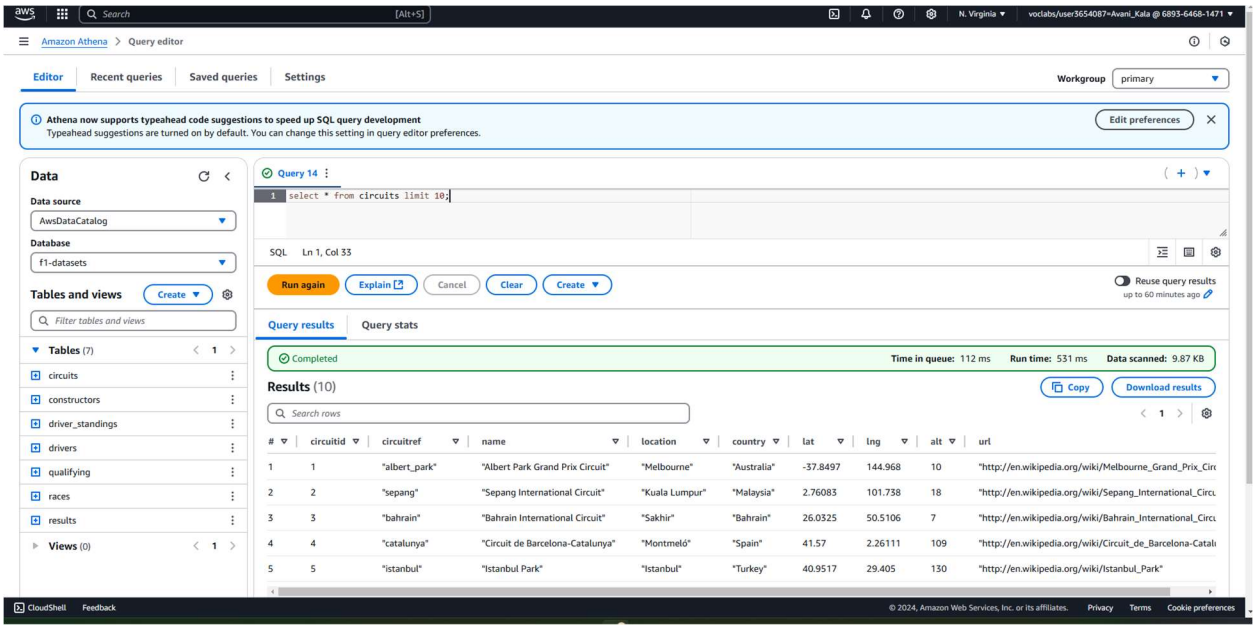


2. Created AWS Glue Crawler and a new database in Glue Data Catalog. Once the crawler successfully reads the data from S3, the data was loaded in form of tables in database in Glue catalog:



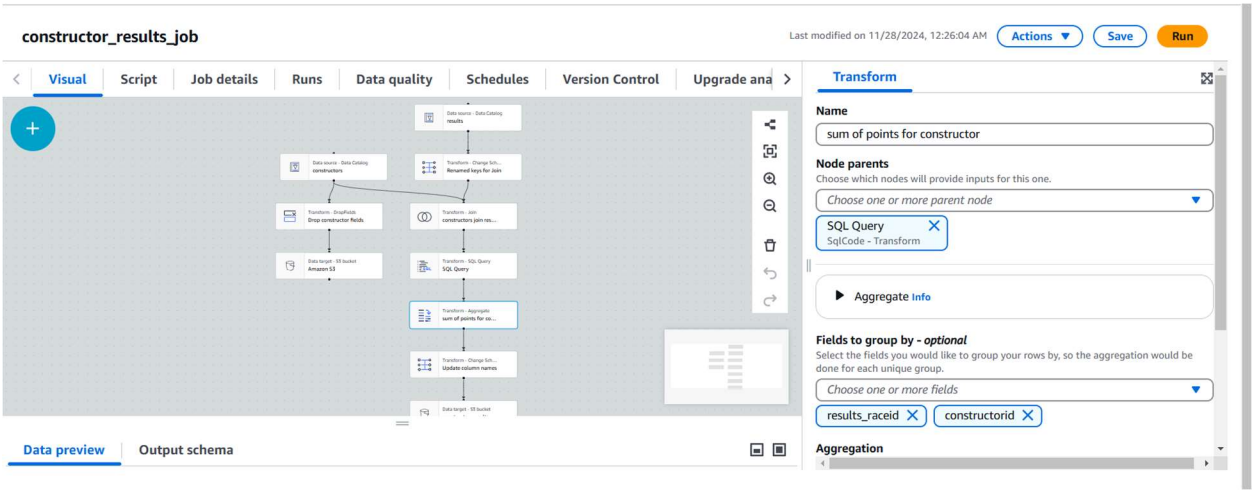


3. To read the data from Catalog database tables, used AWS Athena to run the SQL queries:

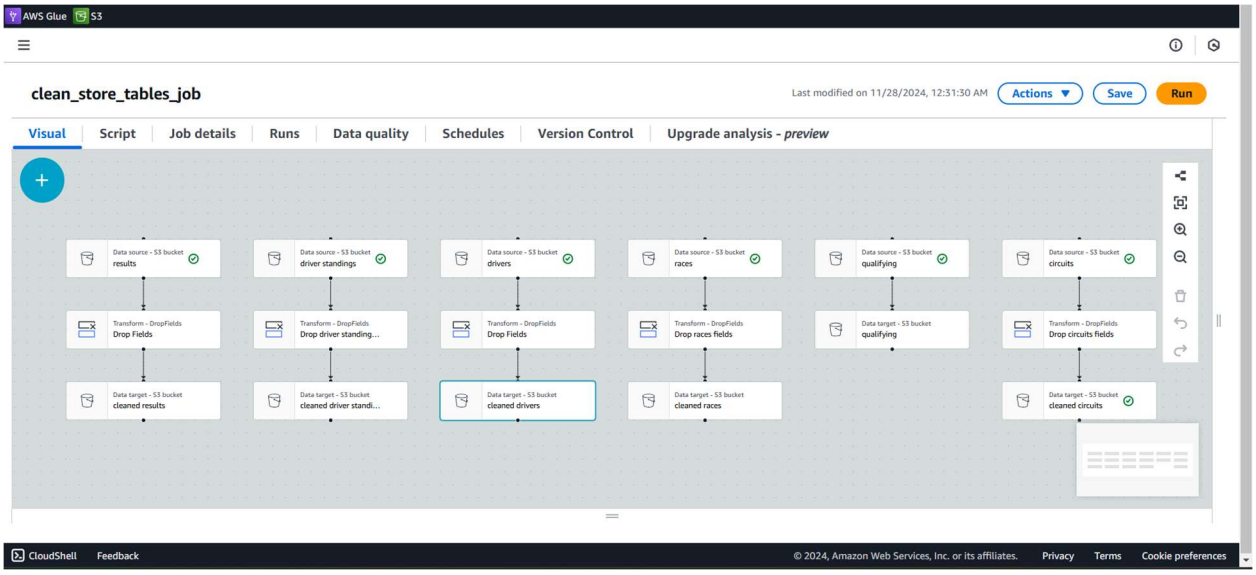


4. Created 2 jobs using Visual ETL to clean and transform the data and store the generated data into S3 bucket in csv format:

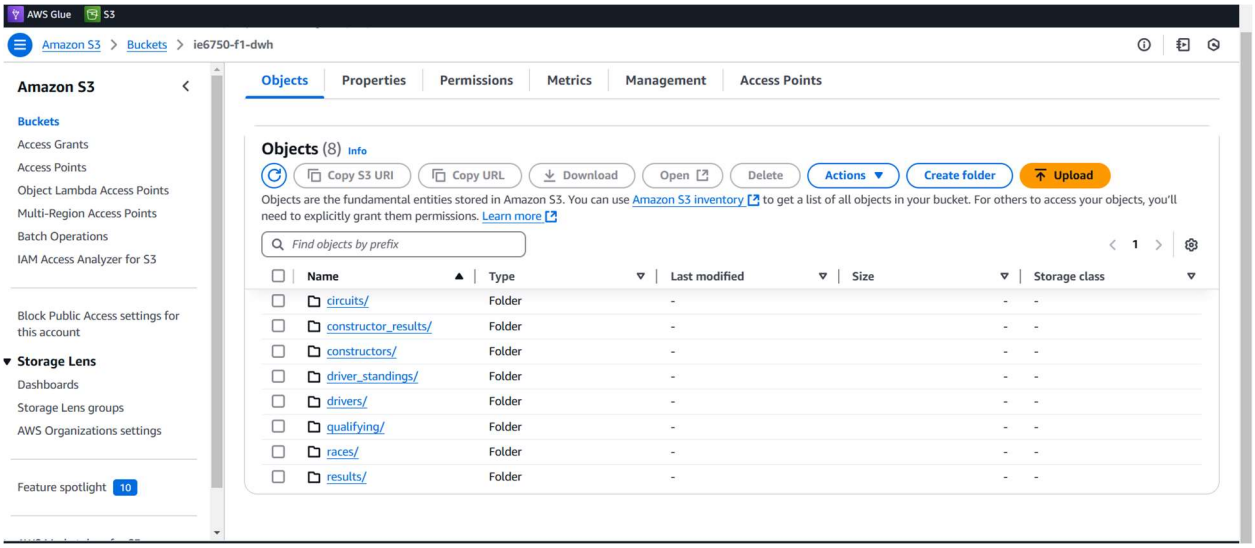
    1. Joined two input tables to find the sum(points) for each constructor for all of its drivers



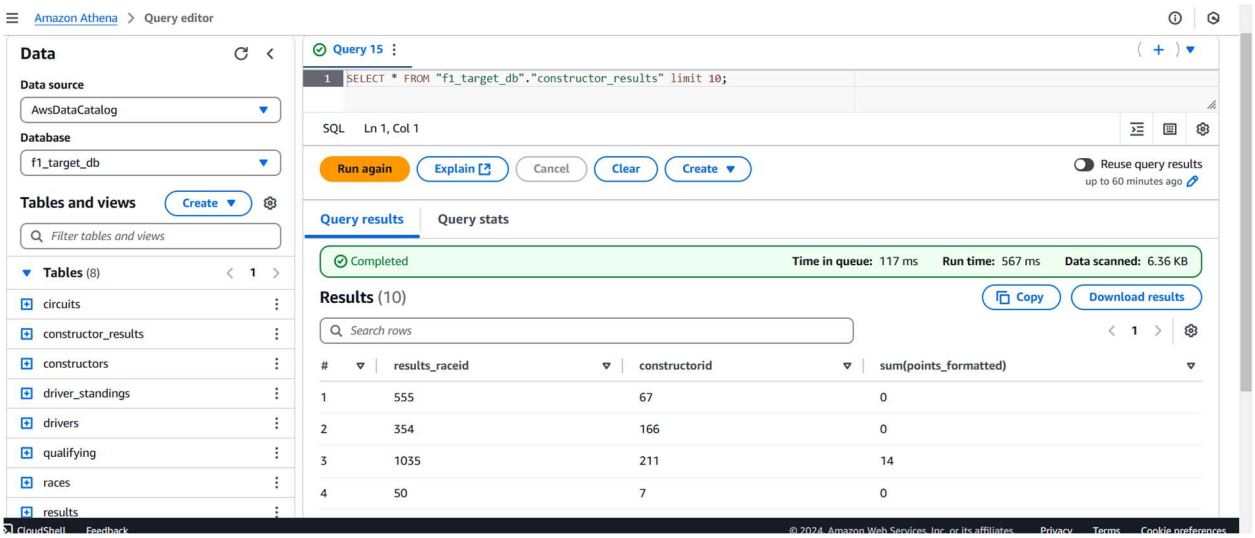
    2. Cleaned and kept the attributes required for analysis:



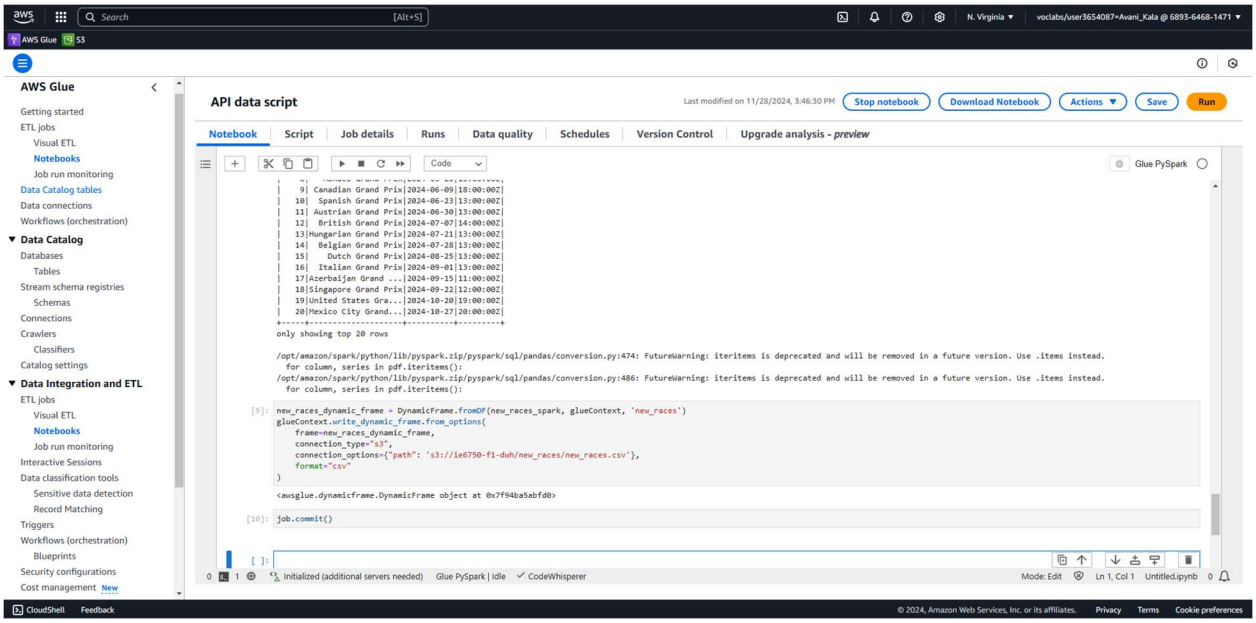
5. ETL Job output loaded in S3 bucket:

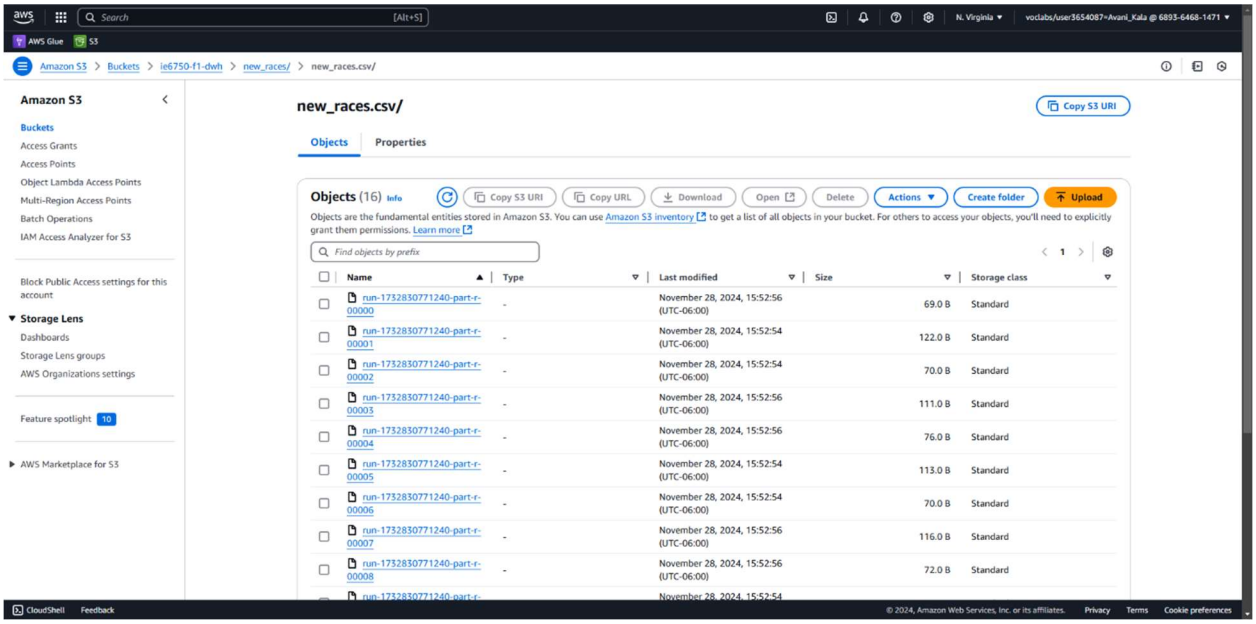


6. Querying the target data in Athena:



7. Created Notebook for accessing data from second data source (Ergast F1 API) to see for new races and store the new data if any in S3:


AWS Glue
API data script
Last modified on 11/28/2024, 3:46:30 PM
Stop notebook
Download Notebook
Actions
Save
Run
Notebook
Script
Job details
Runs
Data quality
Schedules
Version Control
Upgrade analysis - preview
Glue PySpark
only showing top 20 rows
new_races_dynamic_frame = DynamicFrame.fromDF(new_races_spark, glueContext, 'new_races')
glueContext.write_dynamic_frame.from_options(
frame=new_races_dynamic_frame,
connection_type="s3",
connection_options={"path": "s3://ie6750-f1-dwh/new_races/new_races.csv'},
format="csv"
)
<awsglue.dynamicframe.DynamicFrame object at 0x7f94ba5abfd0>
job.commit()
Initialized (additional servers needed)
Glue PySpark | Idle
CodeWhisperer
Mode: Edit
Ln 1, Col 1
Untitled.ipynb
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon S3
Buckets
ie6750-f1-dwh
new_races/
new_races.csv/
Copy S3 URI
Objects
Properties
Objects (16)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
Storage class
run-1732830771240-part-r-00000 - November 28, 2024, 15:52:56 (UTC-06:00) 69.0 B Standard
run-1732830771240-part-r-00001 - November 28, 2024, 15:52:54 (UTC-06:00) 122.0 B Standard
run-1732830771240-part-r-00002 - November 28, 2024, 15:52:54 (UTC-06:00) 70.0 B Standard
run-1732830771240-part-r-00003 - November 28, 2024, 15:52:56 (UTC-06:00) 111.0 B Standard
run-1732830771240-part-r-00004 - November 28, 2024, 15:52:56 (UTC-06:00) 76.0 B Standard
run-1732830771240-part-r-00005 - November 28, 2024, 15:52:54 (UTC-06:00) 113.0 B Standard
run-1732830771240-part-r-00006 - November 28, 2024, 15:52:54 (UTC-06:00) 70.0 B Standard
run-1732830771240-part-r-00007 - November 28, 2024, 15:52:56 (UTC-06:00) 116.0 B Standard
run-1732830771240-part-r-00008 - November 28, 2024, 15:52:56 (UTC-06:00) 72.0 B Standard
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences