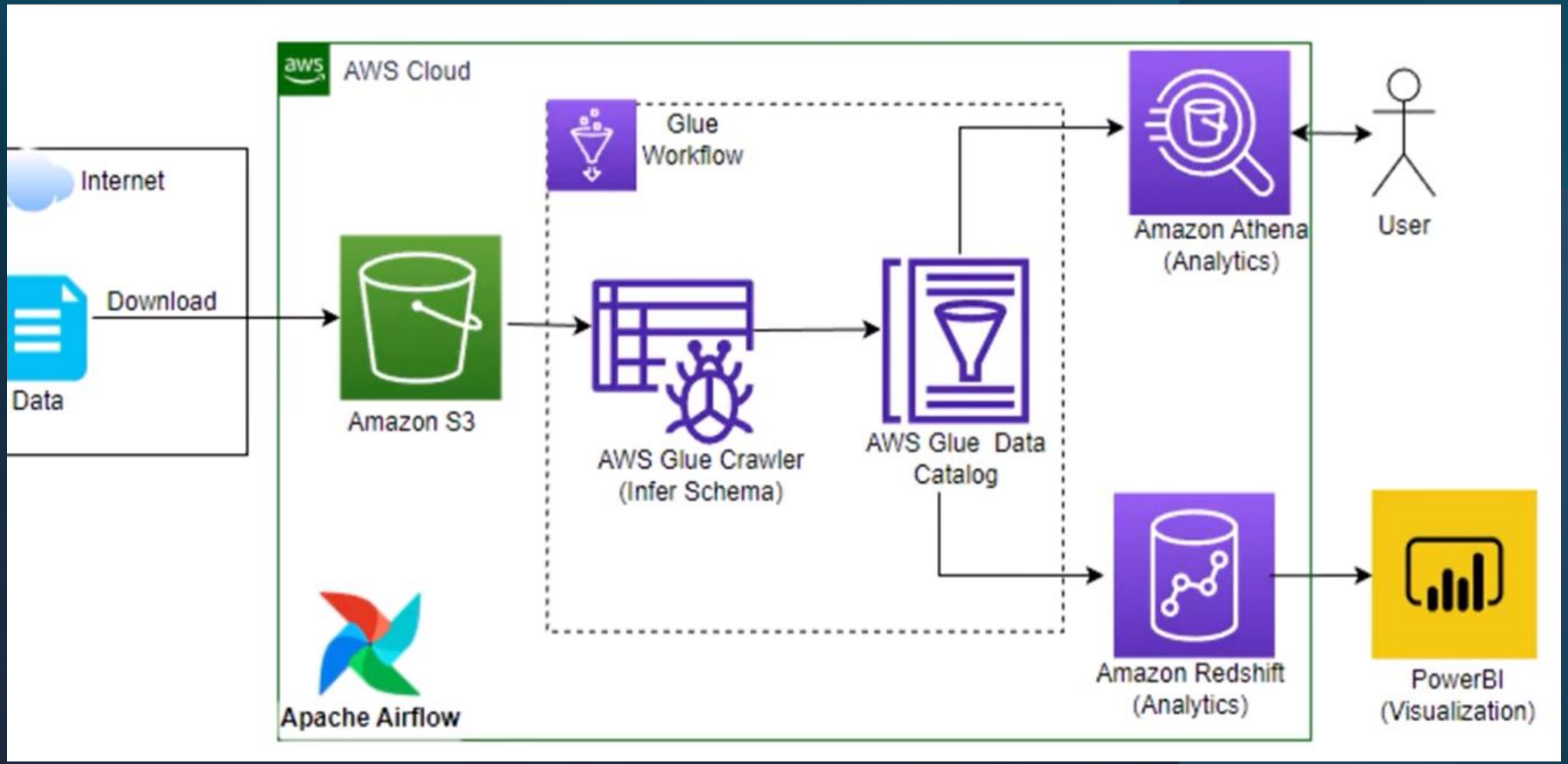

AWS Cloud
Internet
Download
Data
Amazon S3
Glue
Workflow
AWS Glue Crawler
(Infer Schema)
AWS Glue Data
Catalog
Amazon Athena
(Analytics)
User
Amazon Redshift
(Analytics)
PowerBI
(Visualization)
Apache Airflow

Project Architecture

- **Data Ingestion:** *Data is downloaded from the internet and stored in an Amazon S3 bucket.*

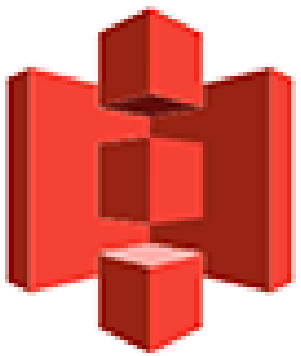
- **Data Cataloging**: *AWS Glue Crawler is used to infer the schema and create a data catalog from the raw data stored in S3.*

- **Data Querying:** *Amazon Athena is employed to run SQL queries on the data stored.*

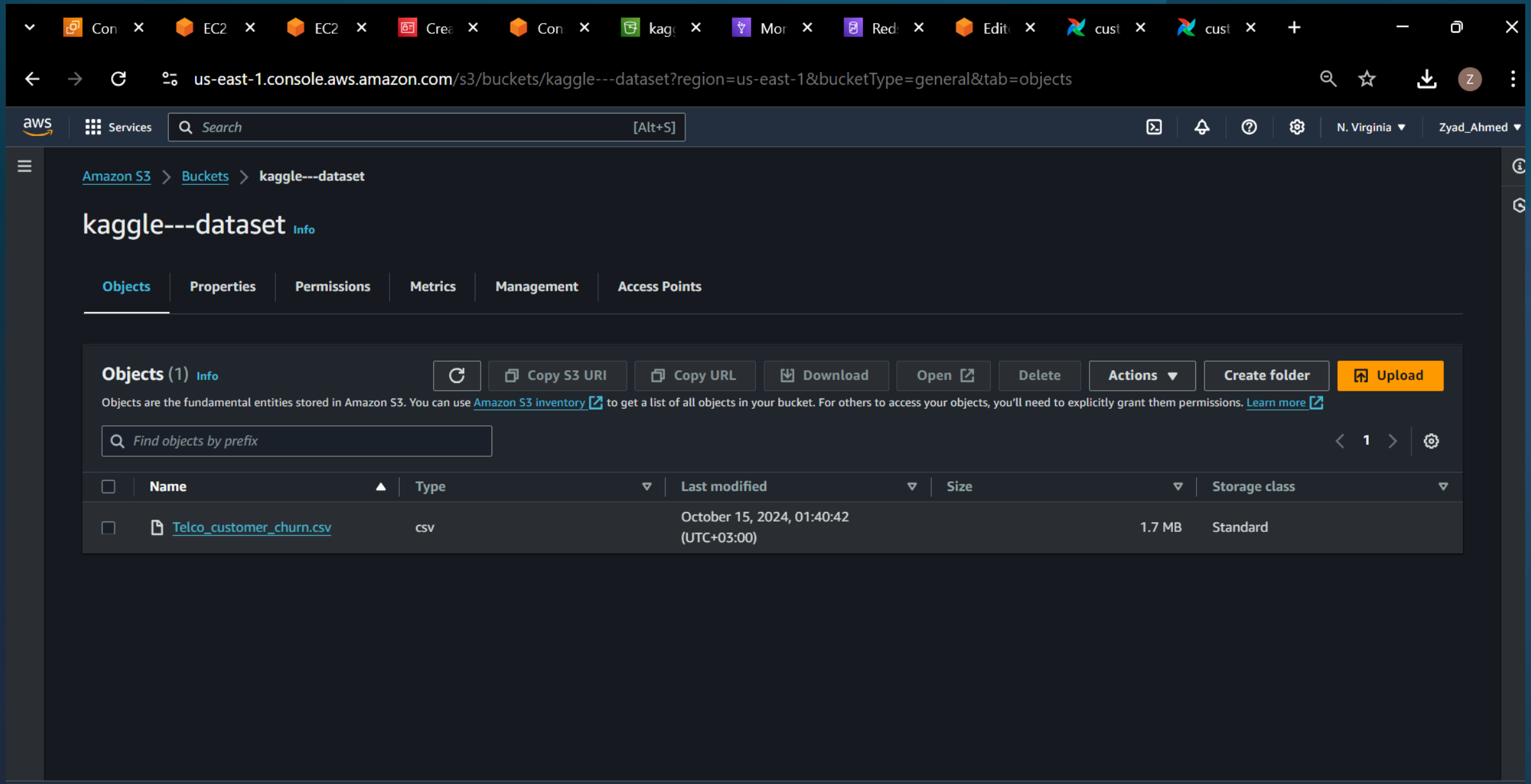
- **Data Loading:** *AWS Glue ETL jobs load the cataloged data into an Amazon Redshift data warehouse.*

- **Data Visualization:** *Power BI is connected to the Redshift cluster for visualization and reporting.*

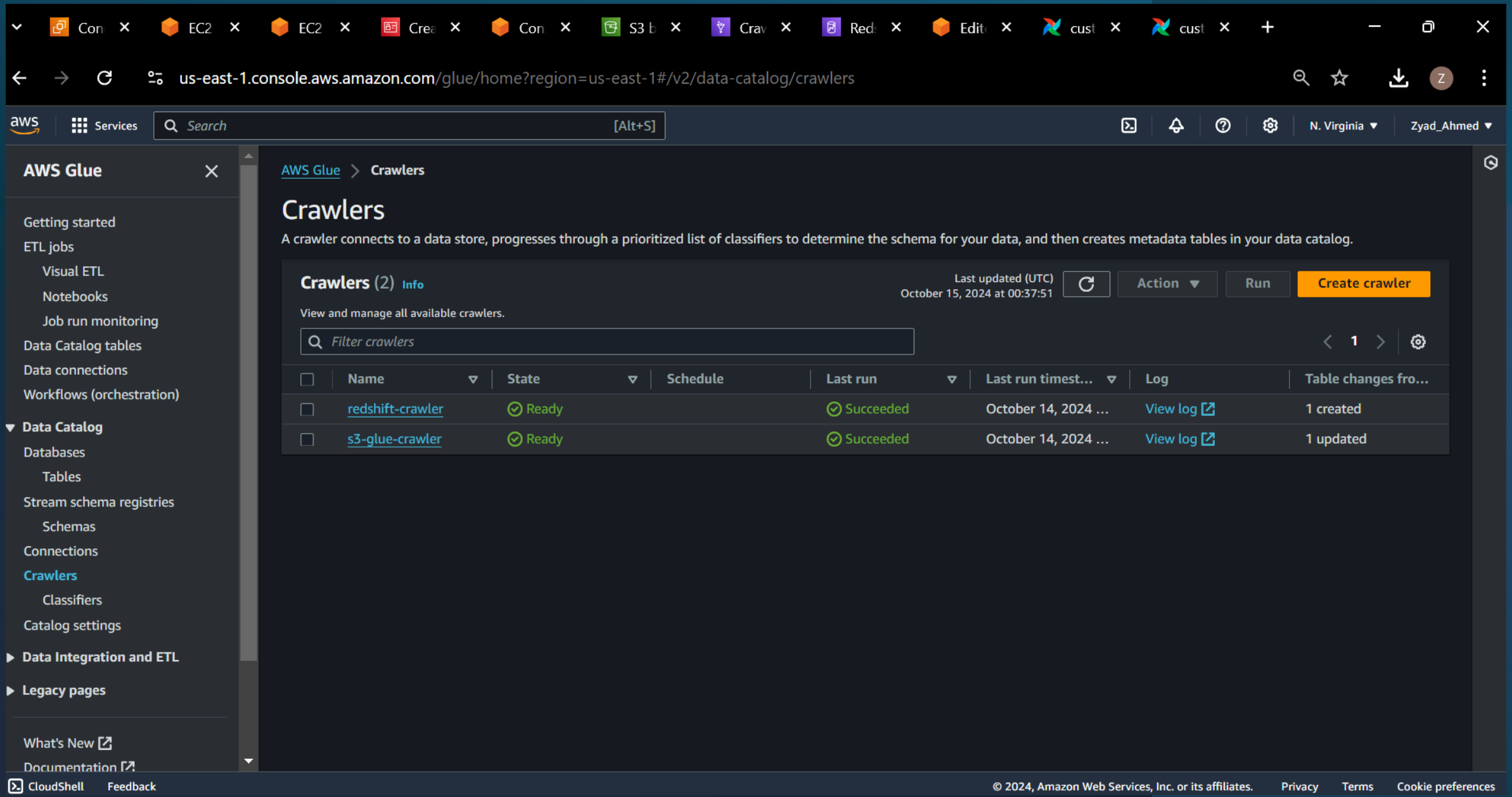
- **Automation:** *Apache Airflow is used to automate the entire ETL process.*

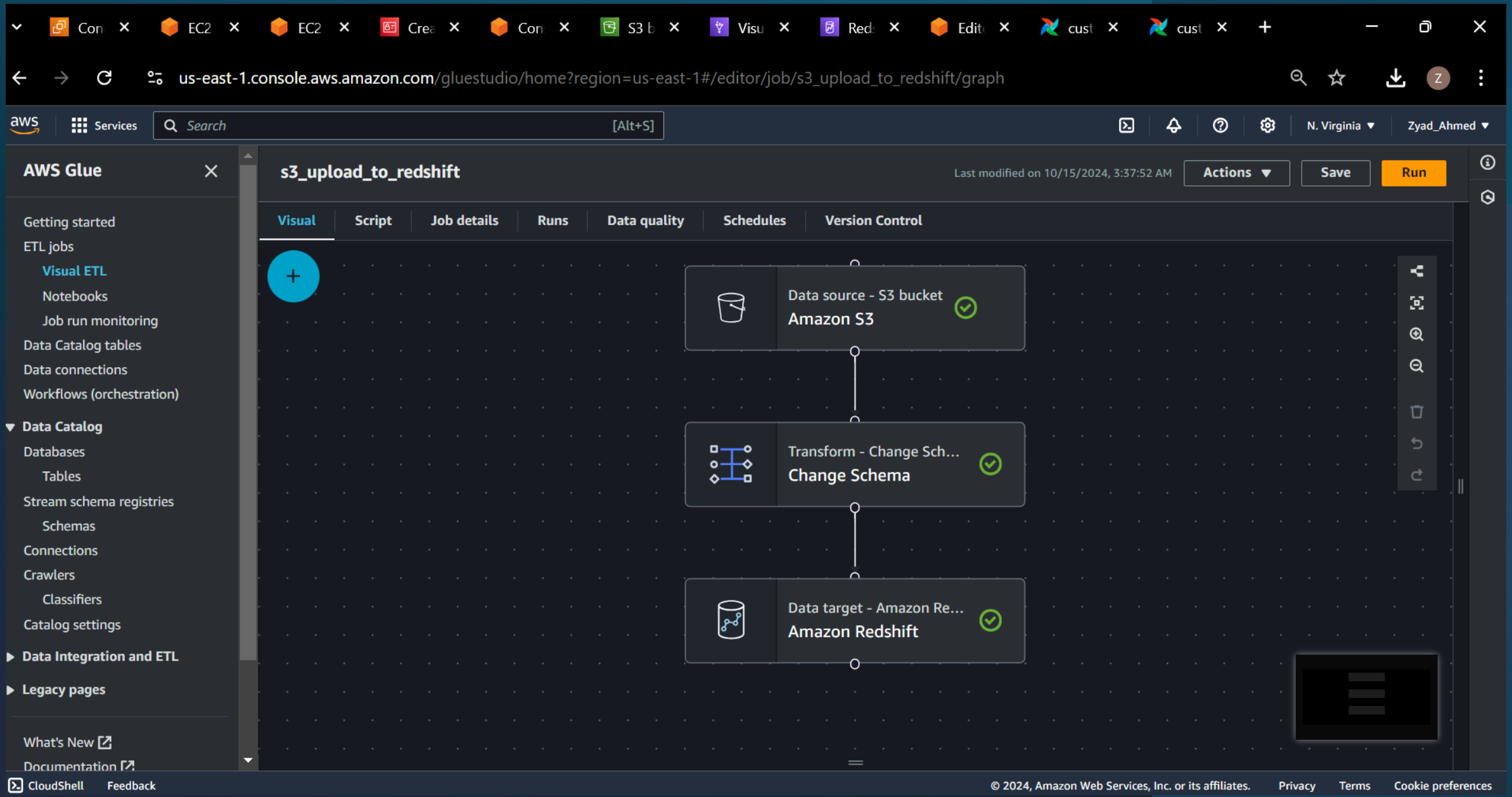


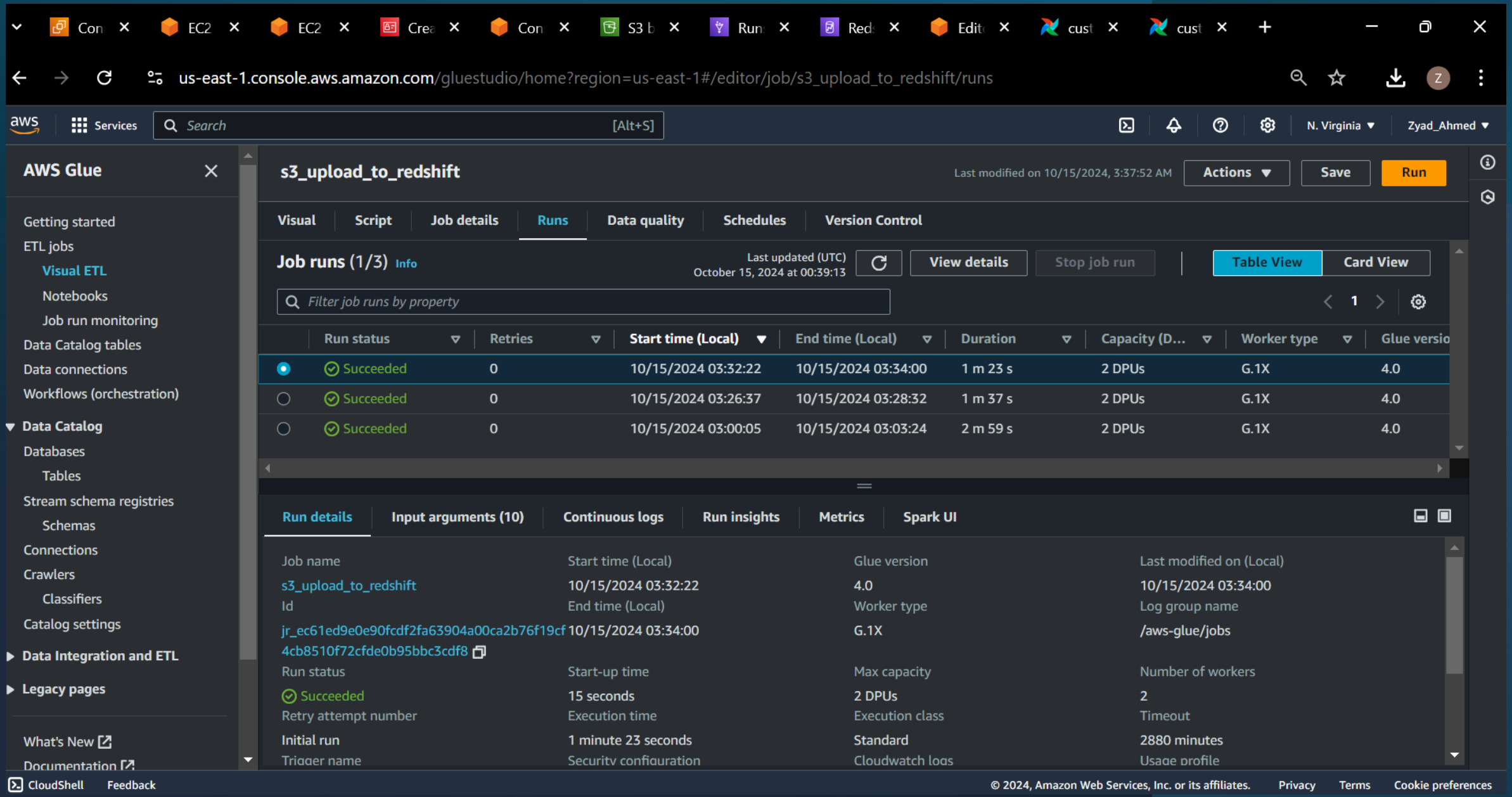
amazon
S3


us-east-1.console.aws.amazon.com/s3/buckets/kaggle---dataset?region=us-east-1&bucketType=general&tab=objects
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
Amazon S3 > Buckets > kaggle---dataset
kaggle---dataset Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (1) Info
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
1
Name
Type
Last modified
Size
Storage class
Telco_customer_churn.csv
csv
October 15, 2024, 01:40:42 (UTC+03:00)
1.7 MB
Standard
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

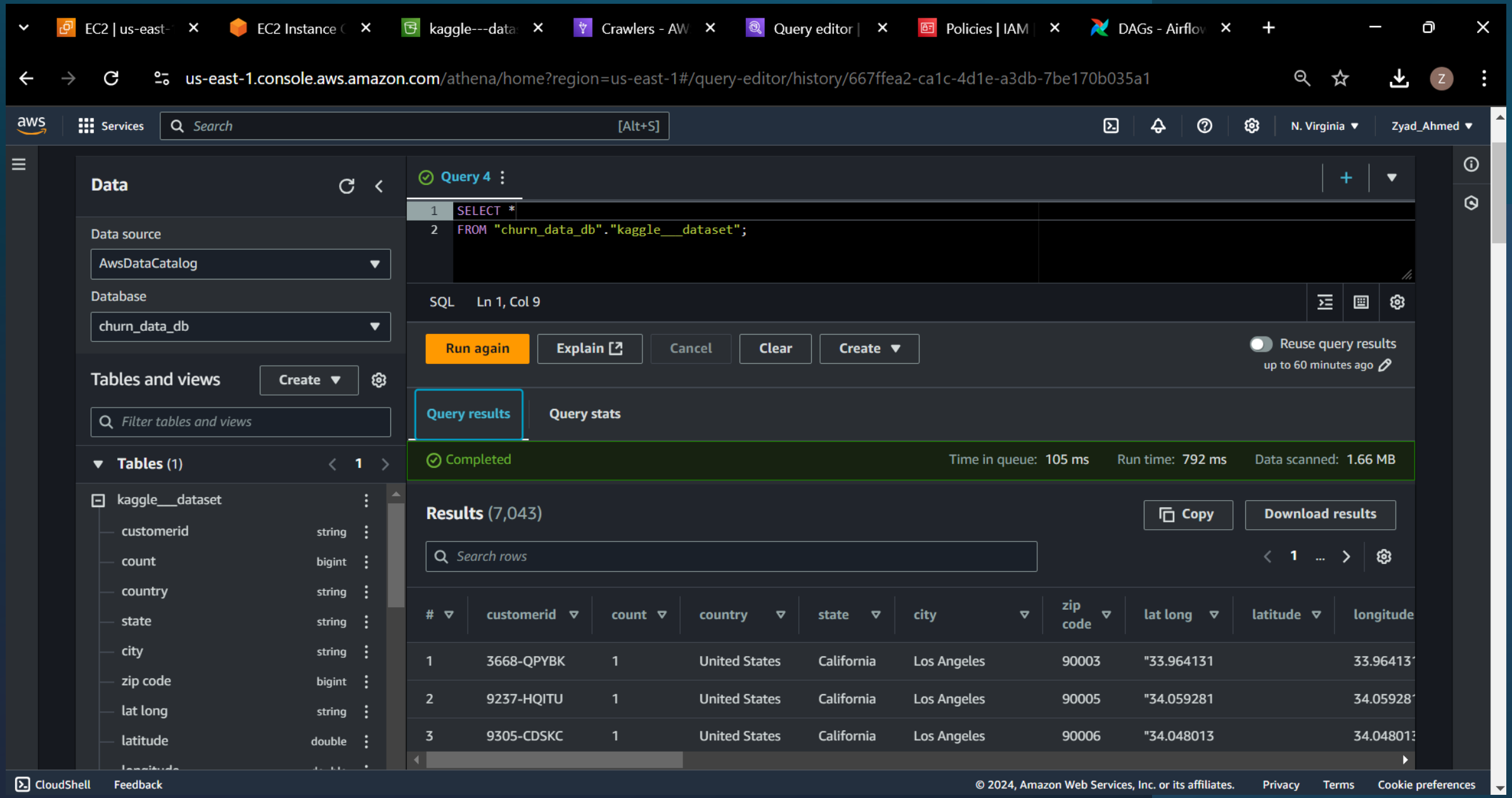

AWS Glue

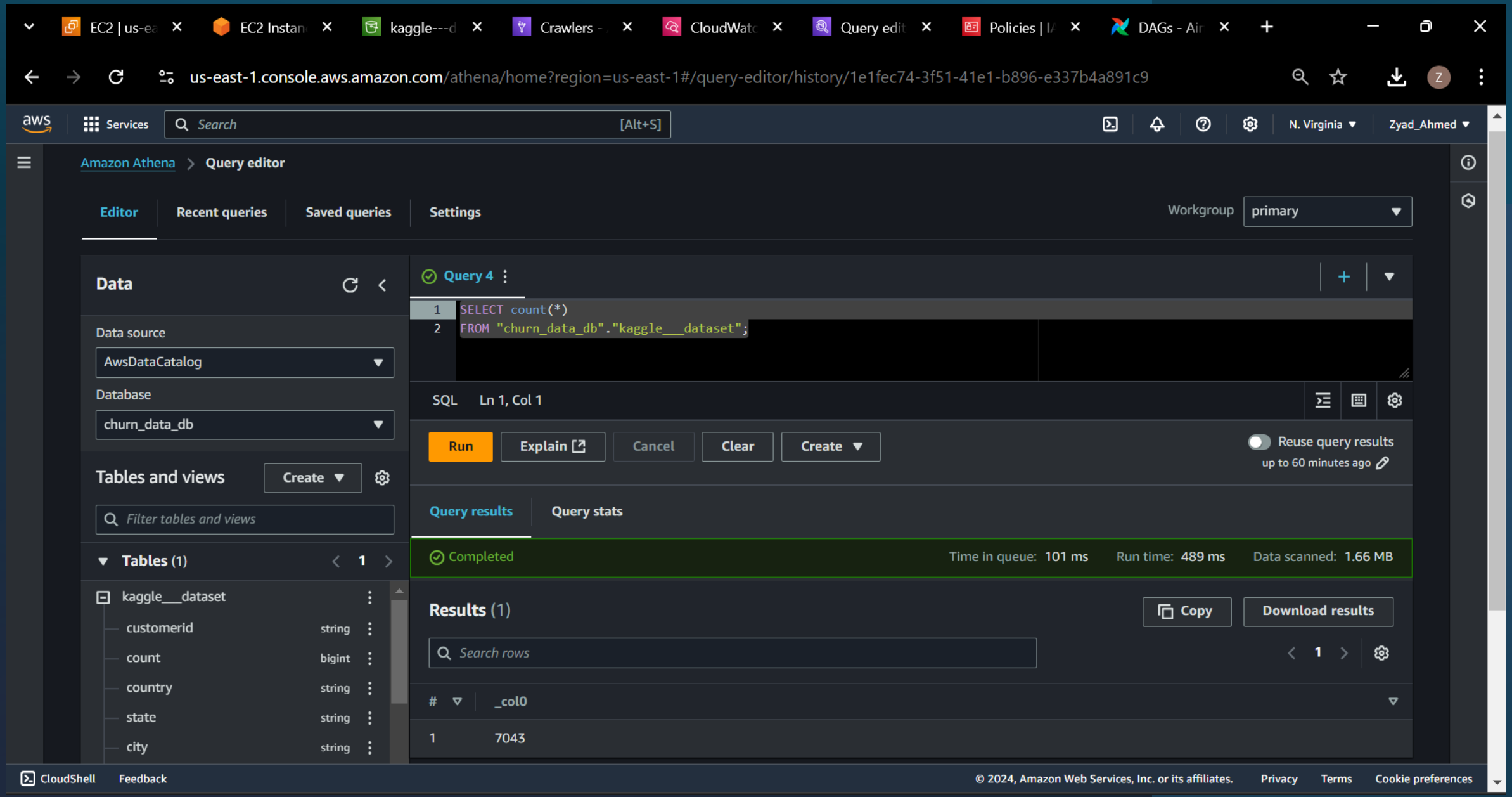

us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Glue > Crawlers
Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.
Crawlers (2) Info
Last updated (UTC)
October 15, 2024 at 00:37:51
Action
Run
Create crawler
View and manage all available crawlers.
Filter crawlers
1
Name
State
Schedule
Last run
Last run timest...
Log
Table changes fro...
redshift-crawler
Ready
Succeeded
October 14, 2024 ...
View log
1 created
s3-glue-crawler
Ready
Succeeded
October 14, 2024 ...
View log
1 updated
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/s3_upload_to_redshift/graph
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
s3_upload_to_redshift
Last modified on 10/15/2024, 3:37:52 AM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - S3 bucket
Amazon S3
Transform - Change Sch...
Change Schema
Data target - Amazon Re...
Amazon Redshift
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

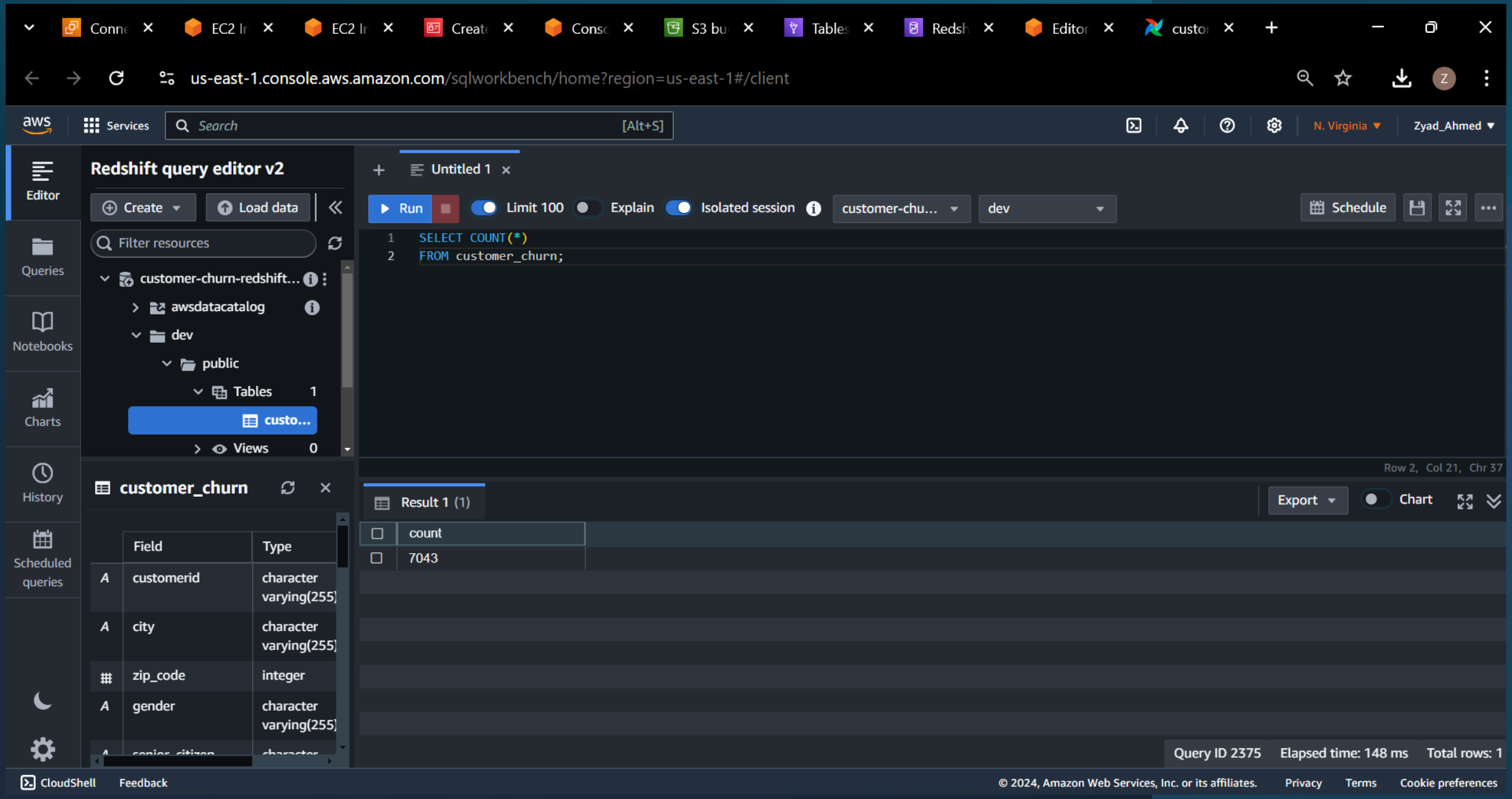

us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/s3_upload_to_redshift/runs
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
s3_upload_to_redshift
Last modified on 10/15/2024, 3:37:52 AM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Job runs (1/3) Info
Last updated (UTC)
October 15, 2024 at 00:39:13
View details
Stop job run
Table View
Card View
Filter job runs by property
1
Run status
Retries
Start time (Local)
End time (Local)
Duration
Capacity (D...
Worker type
Glue versio
Succeeded 0 10/15/2024 03:32:22 10/15/2024 03:34:00 1 m 23 s 2 DPUs G.1X 4.0
Succeeded 0 10/15/2024 03:26:37 10/15/2024 03:28:32 1 m 37 s 2 DPUs G.1X 4.0
Succeeded 0 10/15/2024 03:00:05 10/15/2024 03:03:24 2 m 59 s 2 DPUs G.1X 4.0
Run details
Input arguments (10)
Continuous logs
Run insights
Metrics
Spark UI
Job name
s3_upload_to_redshift
Start time (Local)
10/15/2024 03:32:22
Glue version
4.0
Last modified on (Local)
10/15/2024 03:34:00
Id
jr_ec61ed9e0e90fcdf2fa63904a00ca2b76f19cf4cb8510f72cfde0b95bbc3cdf8
End time (Local)
10/15/2024 03:34:00
Worker type
G.1X
Log group name
/aws-glue/jobs
Run status
Succeeded
Start-up time
15 seconds
Max capacity
2 DPUs
Number of workers
2
Retry attempt number
Initial run
Execution time
1 minute 23 seconds
Execution class
Standard
Timeout
2880 minutes
Trigger name
Security configuration
Cloudwatch logs
Usage profile
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Amazon Athena

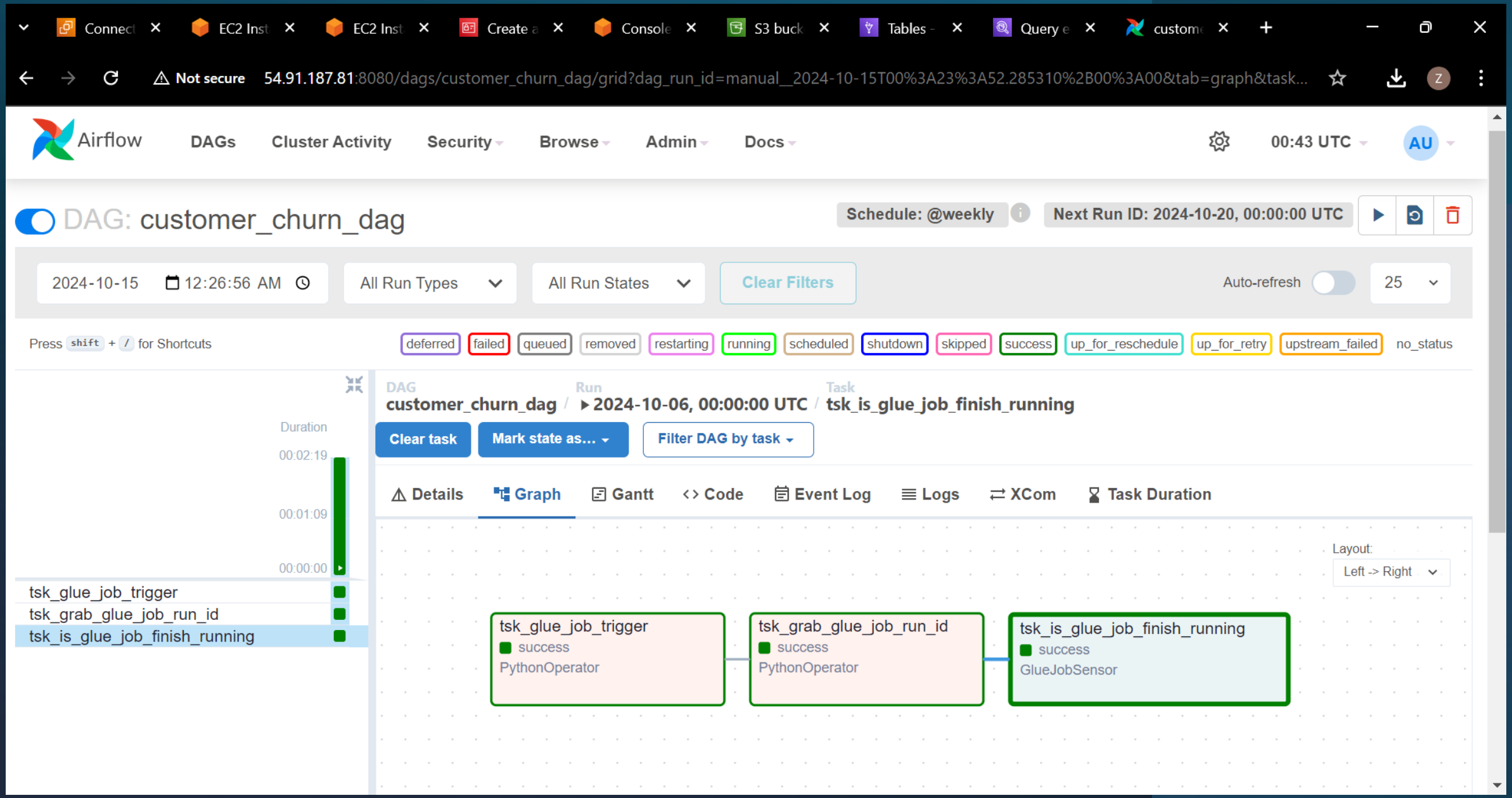

EC2 | us-east-
EC2 Instance
kaggle---data
Crawlers - AW
Query editor |
Policies | IAM
DAGs - Airflo
us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/667ffea2-ca1c-4d1e-a3db-7be170b035a1
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
Data
Data source
AwsDataCatalog
Database
churn_data_db
Tables and views
Create
Filter tables and views
Tables (1)
kaggle___dataset
customerid string
count bigint
country string
state string
city string
zip code bigint
lat long string
latitude double
Query 4
SELECT *
FROM "churn_data_db"."kaggle___dataset";
SQL Ln 1, Col 9
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 105 ms
Run time: 792 ms
Data scanned: 1.66 MB
Results (7,043)
Copy
Download results
Search rows
customerid count country state city zip code lat long latitude longitude
1 3668-QPYBK 1 United States California Los Angeles 90003 "33.964131 33.96413
2 9237-HQITU 1 United States California Los Angeles 90005 "34.059281 34.05928
3 9305-CDSKC 1 United States California Los Angeles 90006 "34.048013 34.04801
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/1e1fec74-3f51-41e1-b896-e337b4a891c9
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
primary
Data
Data source
AwsDataCatalog
Database
churn_data_db
Tables and views
Create
Filter tables and views
Tables (1)
kaggle___dataset
customerid string
count bigint
country string
state string
city string
Query 4
SELECT count(*)
FROM "churn_data_db"."kaggle___dataset";
SQL Ln 1, Col 1
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 101 ms
Run time: 489 ms
Data scanned: 1.66 MB
Results (1)
Copy
Download results
Search rows
_col0
1 7043
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

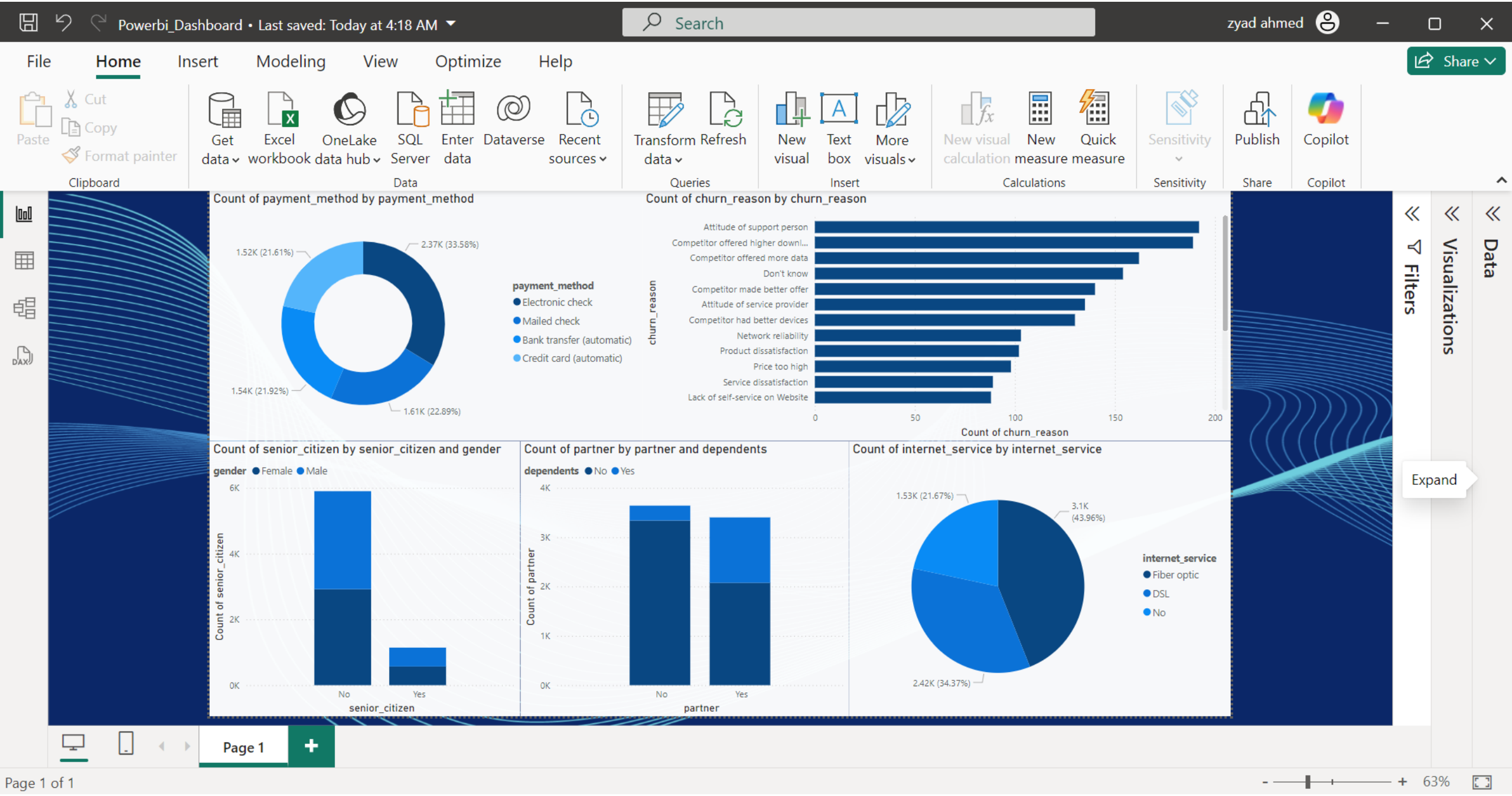
Amazon Redshift


us-east-1.console.aws.amazon.com/sqlworkbench/home?region=us-east-1#/client
aws
Services
Search
[Alt+S]
N. Virginia
Zyad_Ahmed
Redshift query editor v2
Editor
Queries
Notebooks
Charts
History
Scheduled queries
Create
Load data
Filter resources
customer-churn-redshift...
awsdatacatalog
dev
public
Tables 1
custo...
Views 0
Untitled 1
Run
Limit 100
Explain
Isolated session
customer-chu...
dev
Schedule
1 SELECT COUNT(*)
2 FROM customer_churn;
Row 2, Col 21, Chr 37
customer_churn
Field
Type
customerid
character varying(255)
city
character varying(255)
zip_code
integer
gender
character varying(255)
Result 1 (1)
Export
Chart
count
7043
Query ID 2375
Elapsed time: 148 ms
Total rows: 1
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Apache
Airflow


Not secure 54.91.187.81:8080/dags/customer_churn_dag/grid?dag_run_id=manual__2024-10-15T00%3A23%3A52.285310%2B00%3A00&tab=graph&task...
Airflow
DAGs
Cluster Activity
Security
Browse
Admin
Docs
00:43 UTC
AU
DAG: customer_churn_dag
Schedule: @weekly
Next Run ID: 2024-10-20, 00:00:00 UTC
2024-10-15
12:26:56 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:02:19
00:01:09
00:00:00
tsk_glue_job_trigger
tsk_grab_glue_job_run_id
tsk_is_glue_job_finish_running
DAG
customer_churn_dag
Run
2024-10-06, 00:00:00 UTC
Task
tsk_is_glue_job_finish_running
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Event Log
Logs
XCom
Task Duration
Layout:
Left -> Right
tsk_glue_job_trigger
success
PythonOperator
tsk_grab_glue_job_run_id
success
PythonOperator
tsk_is_glue_job_finish_running
success
GlueJobSensor

Power BI


Powerbi_Dashboard • Last saved: Today at 4:18 AM
Search
zyad ahmed
File
Home
Insert
Modeling
View
Optimize
Help
Share
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New visual calculation
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Copilot
Copilot
Count of payment_method by payment_method
1.52K (21.61%)
2.37K (33.58%)
1.54K (21.92%)
1.61K (22.89%)
payment_method
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
Count of churn_reason by churn_reason
Attitude of support person
Competitor offered higher downl...
Competitor offered more data
Don't know
Competitor made better offer
Attitude of service provider
Competitor had better devices
Network reliability
Product dissatisfaction
Price too high
Service dissatisfaction
Lack of self-service on Website
churn_reason
0
50
100
150
200
Count of churn_reason
Count of senior_citizen by senior_citizen and gender
gender Female Male
6K
4K
2K
0K
Count of senior_citizen
No
Yes
senior_citizen
Count of partner by partner and dependents
dependents No Yes
4K
3K
2K
1K
0K
Count of partner
No
Yes
partner
Count of internet_service by internet_service
1.53K (21.67%)
3.1K (43.96%)
2.42K (34.37%)
internet_service
Fiber optic
DSL
No
Filters
Visualizations
Data
Expand
Page 1
Page 1 of 1
63%