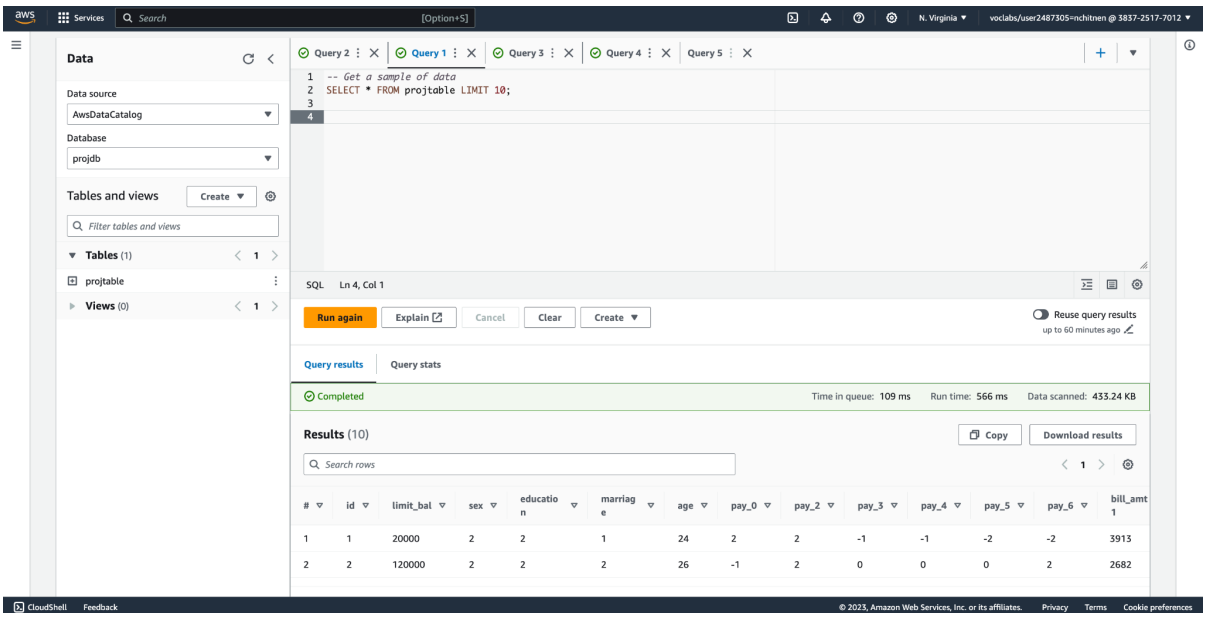
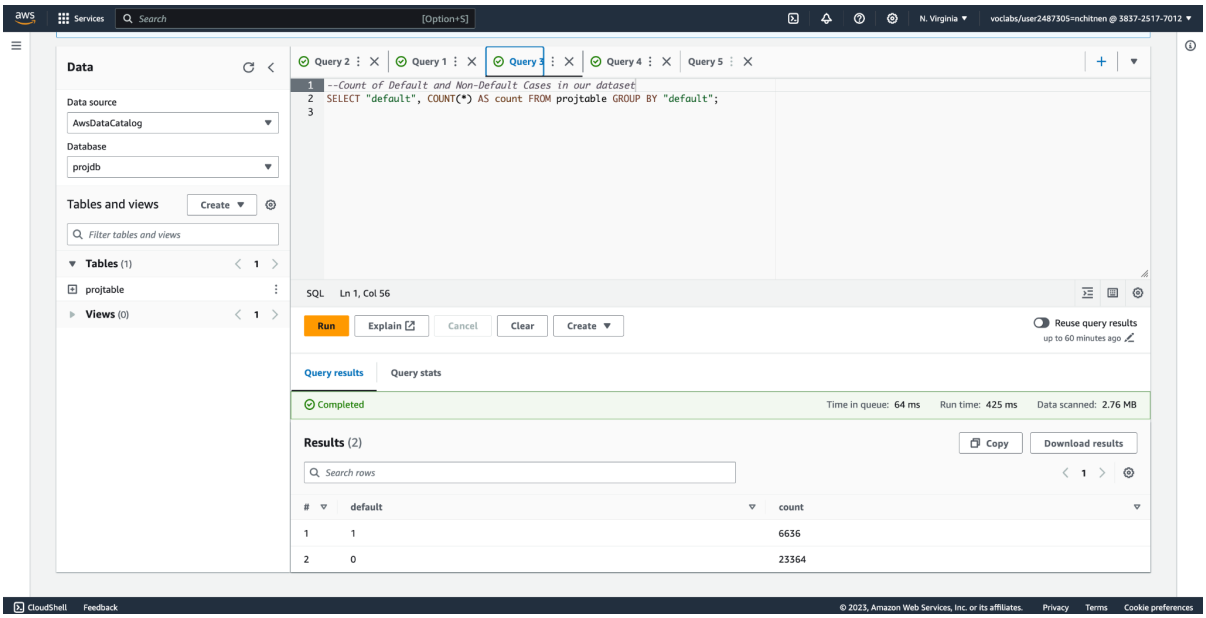
We completed various essential tasks related to the management and exploration of data on AWS Initially, we uploaded the default credit card dataset from the UCI archives "http://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients" to an AWS S3 bucket, utilizing Amazon S3 as a highly scalable and dependable storage solution for our datasets. With the data securely stored in S3, we utilized AWS Athena for data exploration. Athena, a serverless query service, allowed us to directly analyze data in S3 using SQL queries, offering a user-friendly approach for ad-hoc analysis without the necessity of overseeing a traditional database.

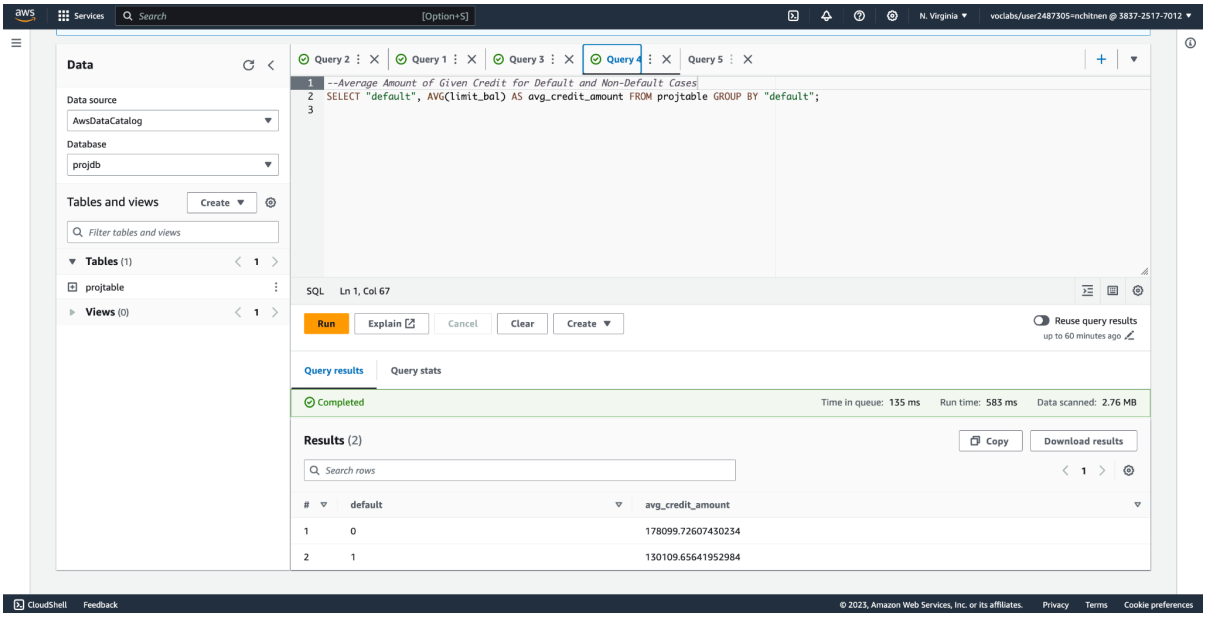
Furthermore, we experimented with AWS Glue, a fully managed ETL service. Specifically, we employed the Glue crawler to automatically discover and catalog metadata from our raw dataset, aiding in the organization and structuring of the data. Following this, we established a database and tables in the AWS Glue Data Catalog, enabling us to efficiently query and analyze the data. These tables seamlessly integrate into the broader AWS ecosystem, facilitating advanced analytics through smooth interaction with other services.

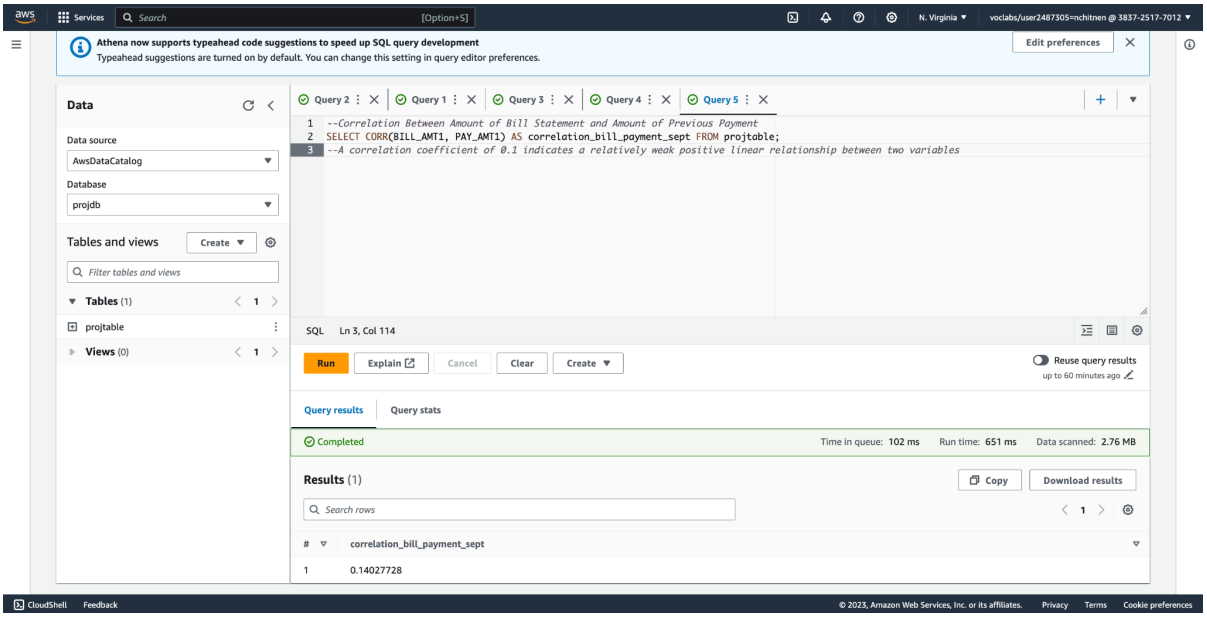
Through querying the tables within the database, we extracted valuable insights from the dataset. This comprehensive workflow, incorporating S3 for storage, Athena for query-driven exploration, and Glue for ETL and cataloging, constitutes a robust data pipeline on AWS. The integration of these services streamlines scalable and efficient data processing, empowering us to uncover meaningful information from the dataset and make well-informed decisions based on the insights derived.

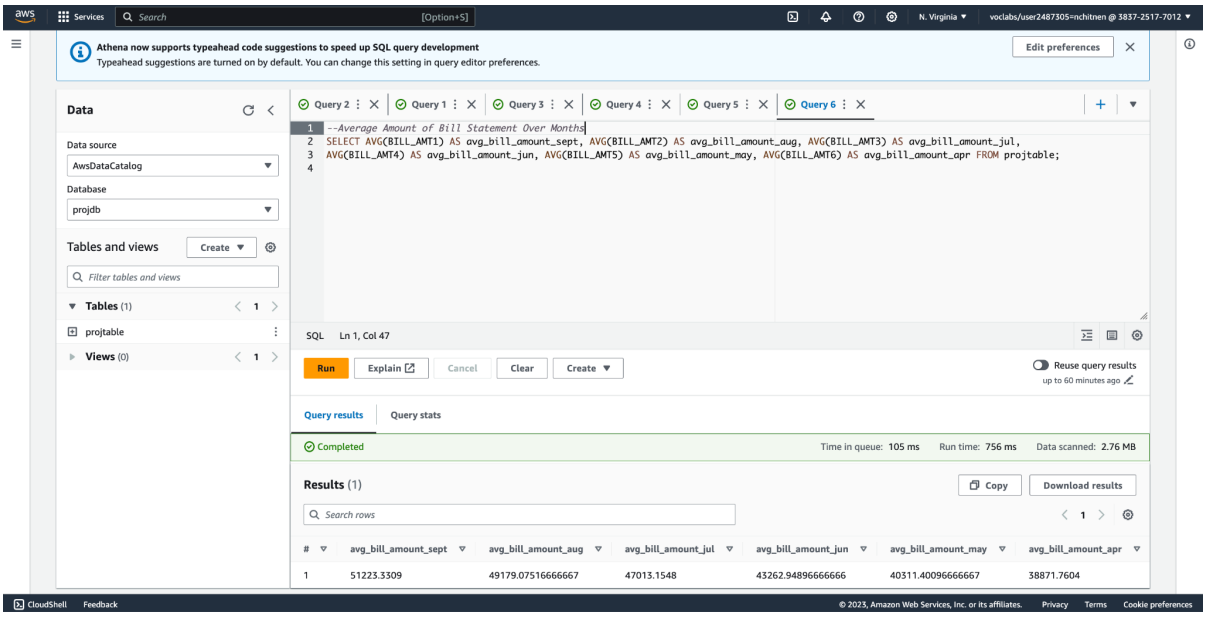
Below are screenshots of the queries and outputs:




Services
Search
[Option+S]
N. Virginia
voclabs/user2487305=nchitnen @ 3837-2517-7012
Data
Data source
AwsDataCatalog
Database
projdb
Tables and views
Create
Filter tables and views
Tables (1)
projtable
Views (0)
Query 2
Query 1
Query 3
Query 4
Query 5
1 --Count of Default and Non-Default Cases in our dataset
2 SELECT "default", COUNT(*) AS count FROM projtable GROUP BY "default";
3
SQL Ln 1, Col 56
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 64 ms
Run time: 425 ms
Data scanned: 2.76 MB
Results (2)
Copy
Download results
Search rows
default count
1 1 6636
2 0 23364
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Services
Search
[Option+S]
N. Virginia
voclabs/user2487305=nchitnen @ 3837-2517-7012
Data
Data source
AwsDataCatalog
Database
projdb
Tables and views
Create
Filter tables and views
Tables (1)
projtable
Views (0)
Query 2
Query 1
Query 3
Query 4
Query 5
1 --Average Amount of Given Credit for Default and Non-Default Cases
2 SELECT "default", AVG(limit_bal) AS avg_credit_amount FROM projtable GROUP BY "default";
3
SQL Ln 1, Col 67
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 135 ms
Run time: 583 ms
Data scanned: 2.76 MB
Results (2)
Copy
Download results
Search rows
default avg_credit_amount
1 0 178099.72607430234
2 1 130109.65641952984
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.
Edit preferences
Data
Data source
AwsDataCatalog
Database
projdb
Tables and views
Create
Filter tables and views
Tables (1)
projtable
Views (0)
Query 2
Query 1
Query 3
Query 4
Query 5
1 --Correlation Between Amount of Bill Statement and Amount of Previous Payment
2 SELECT CORR(BILL_AMT1, PAY_AMT1) AS correlation_bill_payment_sept FROM projtable;
3 --A correlation coefficient of 0.1 indicates a relatively weak positive linear relationship between two variables
SQL Ln 3, Col 114
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 102 ms
Run time: 651 ms
Data scanned: 2.76 MB
Results (1)
Copy
Download results
Search rows
correlation_bill_payment_sept
1 0.14027728
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.
Edit preferences
Data
Data source
AwsDataCatalog
Database
projdb
Tables and views
Create
Filter tables and views
Tables (1)
projtable
Views (0)
Query 2
Query 1
Query 3
Query 4
Query 5
Query 6
1 --Average Amount of Bill Statement Over Months
2 SELECT AVG(BILL_AMT1) AS avg_bill_amount_sept, AVG(BILL_AMT2) AS avg_bill_amount_aug, AVG(BILL_AMT3) AS avg_bill_amount_jul,
3 AVG(BILL_AMT4) AS avg_bill_amount_jun, AVG(BILL_AMT5) AS avg_bill_amount_may, AVG(BILL_AMT6) AS avg_bill_amount_apr FROM projtable;
4
SQL Ln 1, Col 47
Run
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 105 ms
Run time: 756 ms
Data scanned: 2.76 MB
Results (1)
Copy
Download results
Search rows
avg_bill_amount_sept avg_bill_amount_aug avg_bill_amount_jul avg_bill_amount_jun avg_bill_amount_may avg_bill_amount_apr
1 51223.3309 49179.07516666667 47013.1548 43262.94896666666 40311.40096666667 38871.7604
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

we extended our efforts by working on creating an AWS pipeline: