Advanced Data Engineering in Cloud

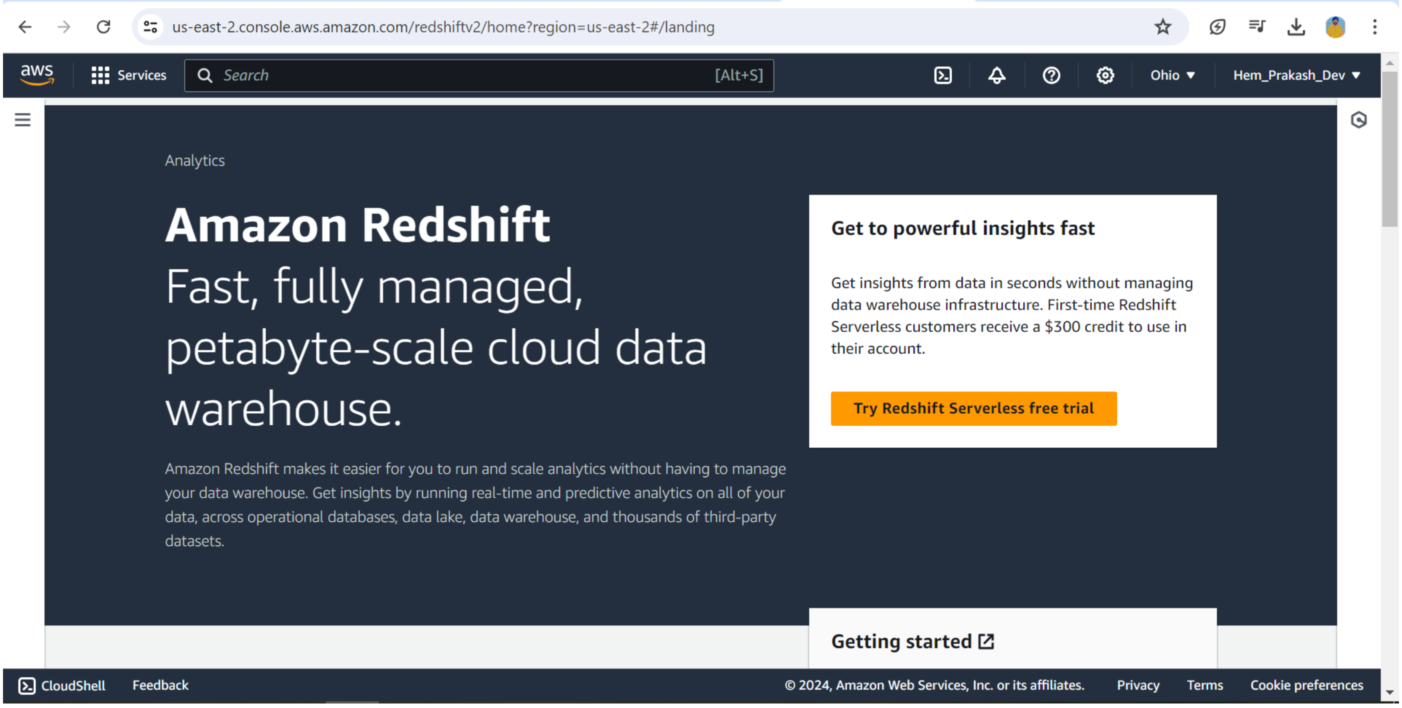
ASSIGNMENT-3

Hem Prakash Dev **Roll-G23AI1054**

1. Configure AWS Redshift or Amazon Athena to aggregate the processed data and enable fast querying and analysis.

Configure AWS Redshift

Create a Redshift Cluster:



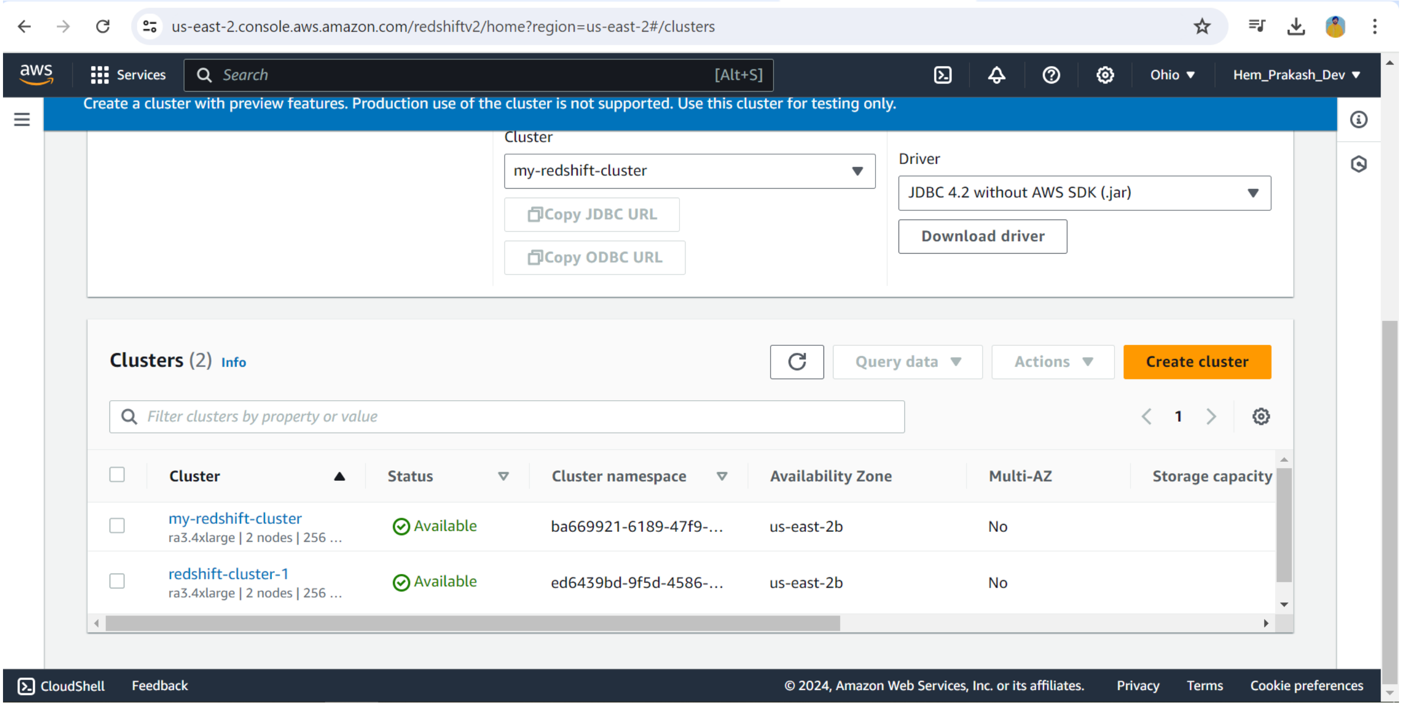
Amazon Redshift
Fast, fully managed, petabyte-scale cloud data warehouse.

Amazon Redshift makes it easier for you to run and scale analytics without having to manage your data warehouse. Get insights by running real-time and predictive analytics on all of your data, across operational databases, data lake, data warehouse, and thousands of third-party datasets.

Get to powerful insights fast

Get insights from data in seconds without managing data warehouse infrastructure. First-time Redshift Serverless customers receive a $300 credit to use in their account.

© 2024, Amazon Web Services, Inc. or its affiliates.



Create a cluster with preview features. Production use of the cluster is not supported. Use this cluster for testing only.

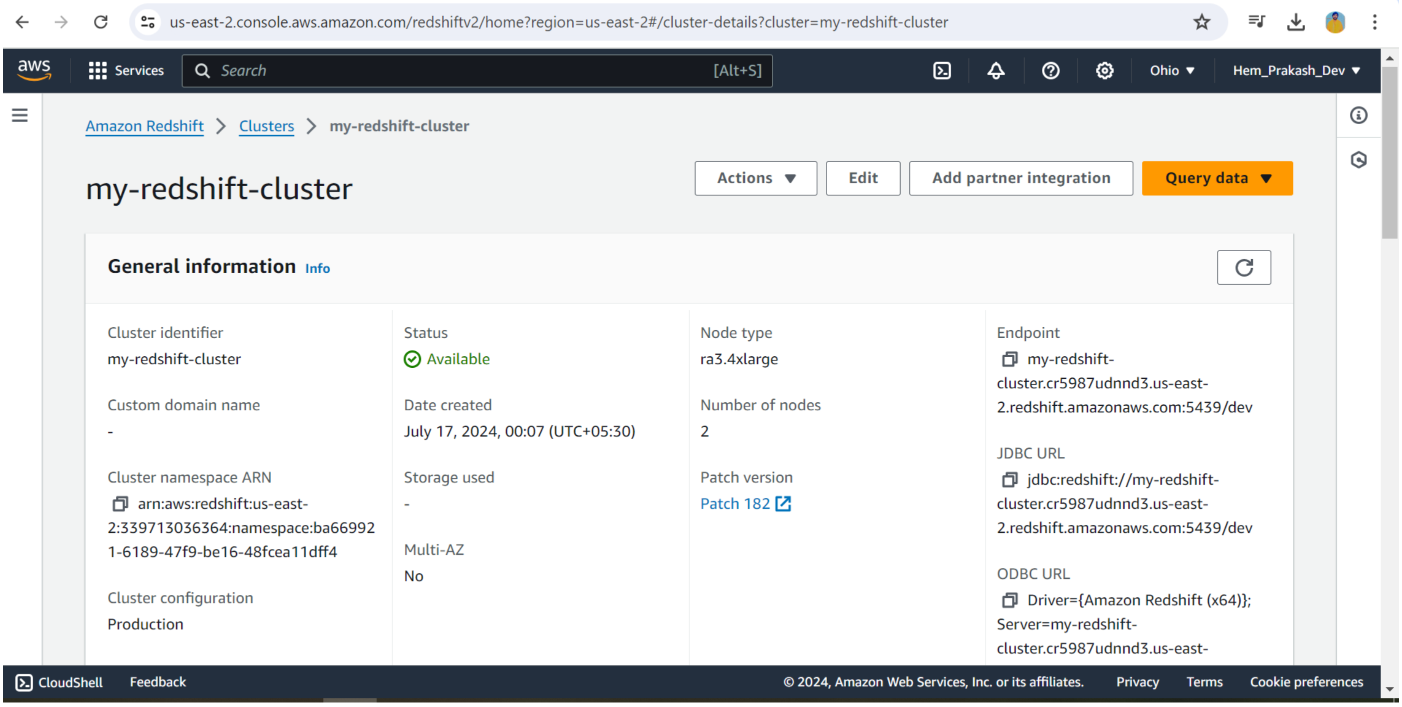
Cluster: my-redshift-cluster

Driver: JDBC 4.2 without AWS SDK (.jar)

Clusters (2) Info

□	Cluster ▲	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity
□	my-redshift-cluster ra3.4xlarge | 2 nodes | 256 …	✓ Available	ba669921-6189-47f9-…	us-east-2b	No	
□	redshift-cluster-1 ra3.4xlarge | 2 nodes | 256 …	✓ Available	ed6439bd-9f5d-4586-…	us-east-2b	No	

© 2024, Amazon Web Services, Inc. or its affiliates.


us-east-2.console.aws.amazon.com/redshiftv2/home?region=us-east-2#/cluster-details?cluster=my-redshift-cluster
aws
Services
Search
[Alt+S]
Ohio
Hem_Prakash_Dev
Amazon Redshift > Clusters > my-redshift-cluster
Actions
Edit
Add partner integration
Query data
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

my-redshift-cluster

General information Info

Cluster identifier
my-redshift-cluster

Custom domain name
-

Cluster namespace ARN
arn:aws:redshift:us-east-2:339713036364:namespace:ba669921-6189-47f9-be16-48fcea11dff4

Cluster configuration
Production

Status
Available

Date created
July 17, 2024, 00:07 (UTC+05:30)

Storage used
-

Multi-AZ
No

Node type
ra3.4xlarge

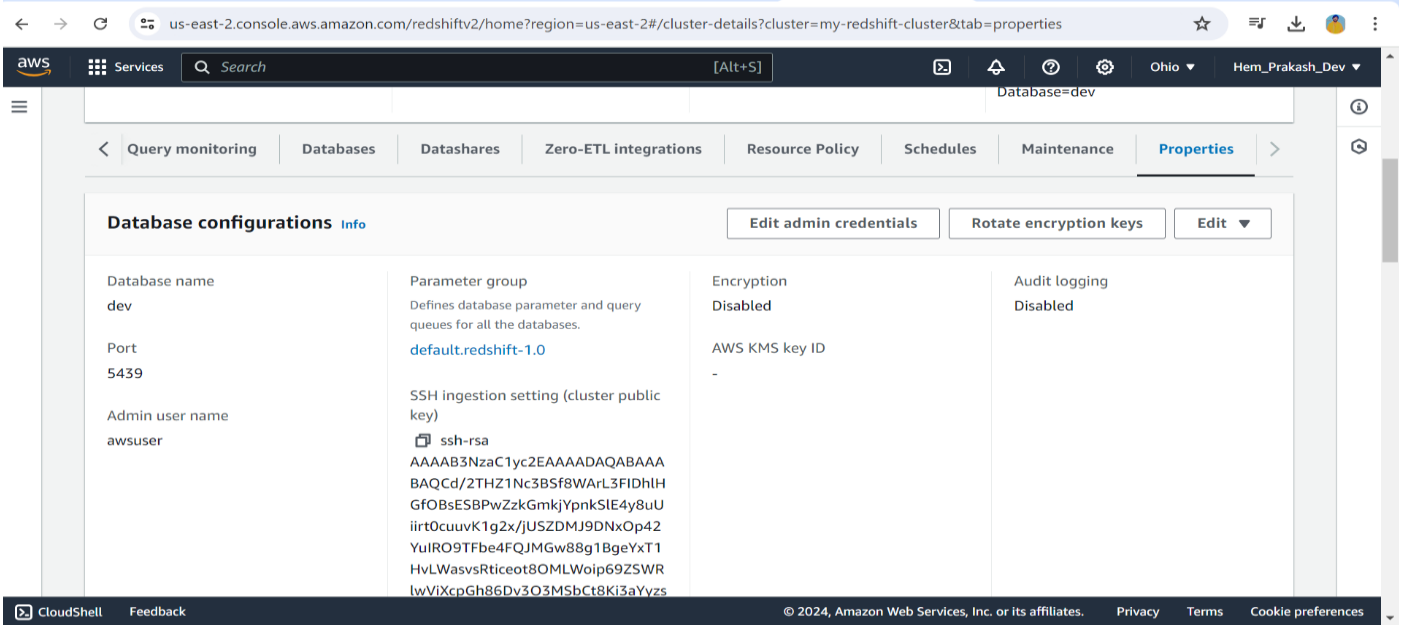
Number of nodes
2

Patch version
Patch 182

Endpoint
my-redshift-cluster.cr5987udnnd3.us-east-2.redshift.amazonaws.com:5439/dev

JDBC URL
jdbc:redshift://my-redshift-cluster.cr5987udnnd3.us-east-2.redshift.amazonaws.com:5439/dev

ODBC URL
Driver={Amazon Redshift (x64)}; Server=my-redshift-cluster.cr5987udnnd3.us-east-



Database configurations Info

Database name
dev

Port
5439

Admin user name
awsuser

Parameter group
Defines database parameter and query queues for all the databases.
default.redshift-1.0

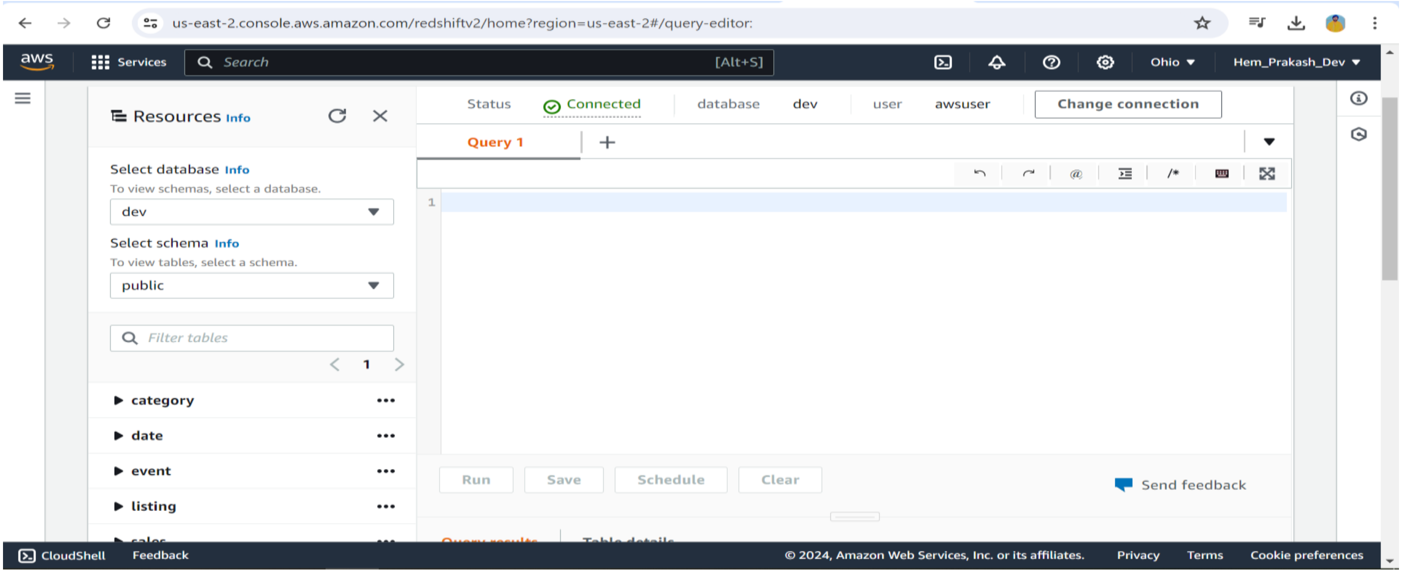
SSH ingestion setting (cluster public key)
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQCd/2THZ1Nc3BSf8WArL3FIDhlHGfOBsESBPwZzkGmkjYpnkSlE4y8uUiirtOcuuvK1g2x/jUSZDMJ9DNxOp42YuIRO9TFbe4FQJMGw88g1BgeYxT1HvLWasvsRticeot8OMLWoip69ZSWRlwViXcpGh86Dv3O3MSbCt8Ki3aYyzs

Encryption
Disabled

AWS KMS key ID
-

Audit logging
Disabled

Successfully runed the query on Redshift query



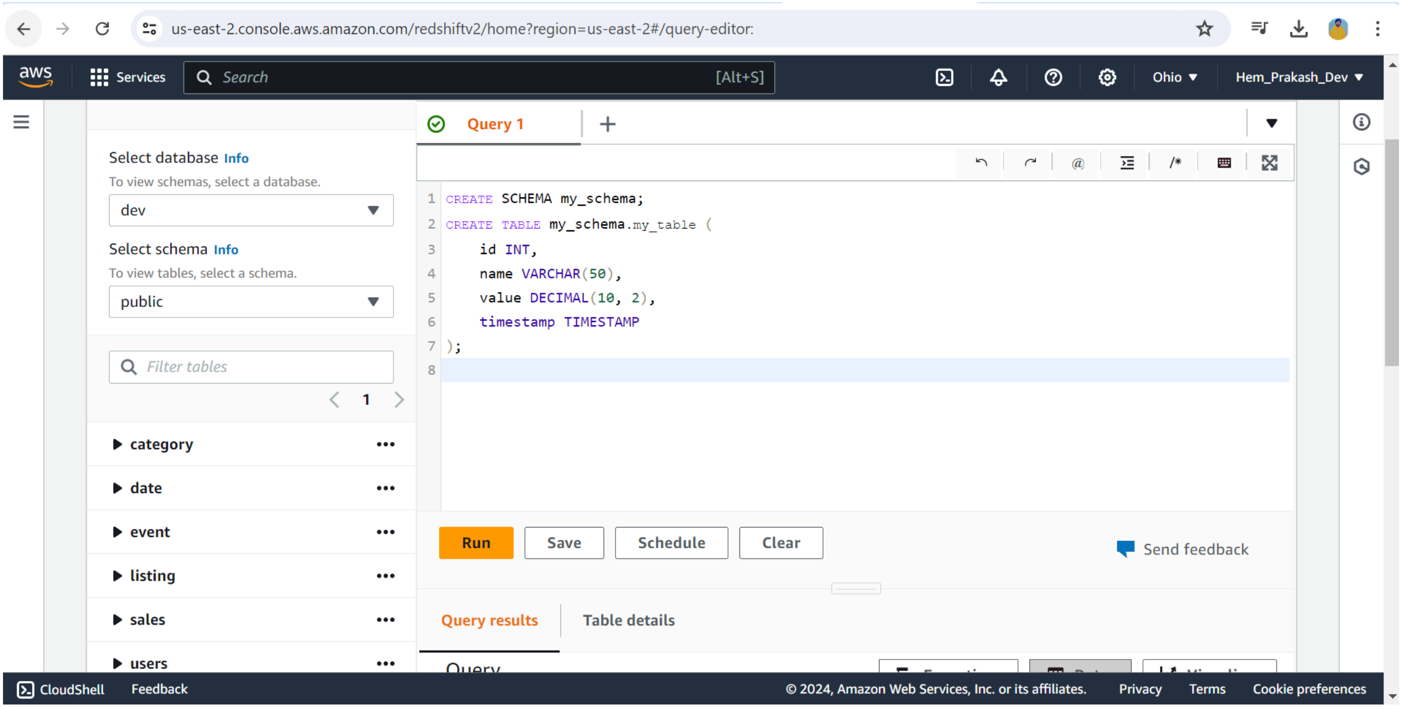
Status Connected database dev user awsuser

Select database Info
To view schemas, select a database.
dev

Select schema Info
To view tables, select a schema.
public

Filter tables

- category
- date
- event
- listing
- sales



Select database Info
To view schemas, select a database.
dev

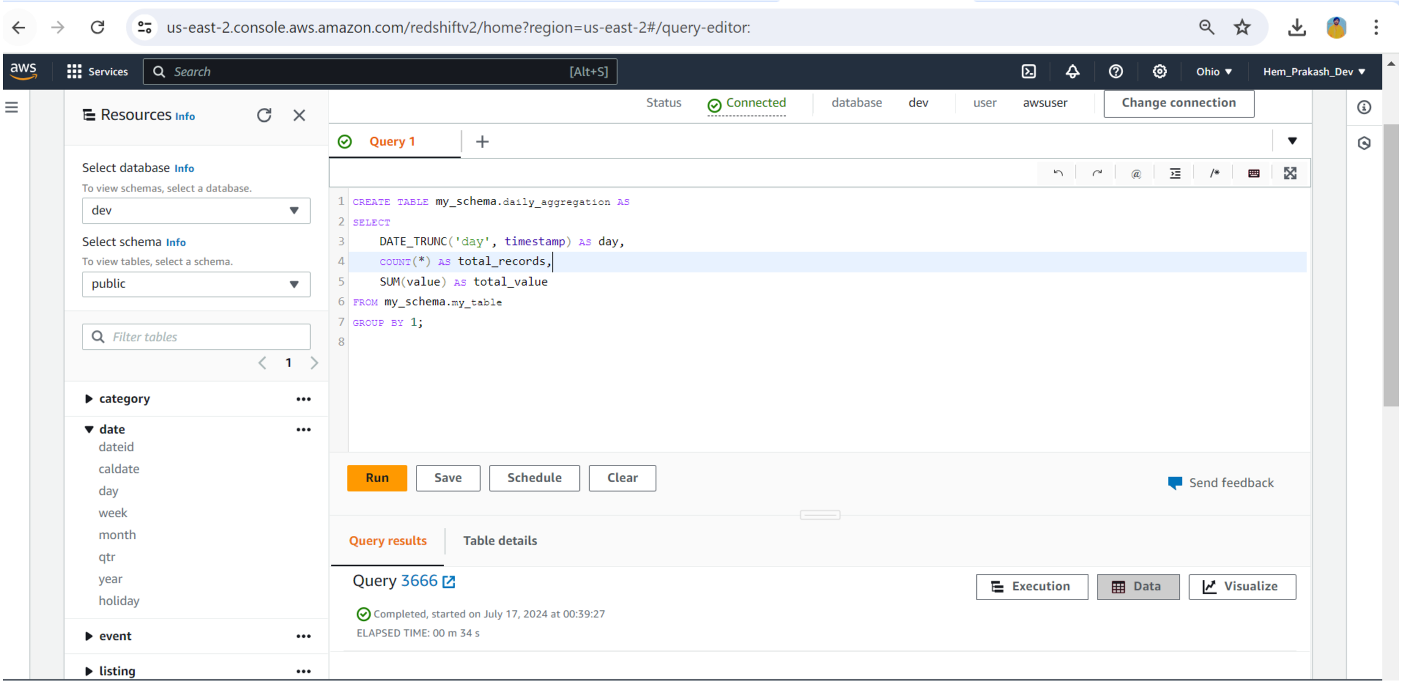
Select schema Info
To view tables, select a schema.
public

```
CREATE SCHEMA my_schema;
CREATE TABLE my_schema.my_table (
    id INT,
    name VARCHAR(50),
    value DECIMAL(10, 2),
    timestamp TIMESTAMP
);
```

2. Design and implement the data aggregation queries and optimize them for performance.

Example Aggregation Queries for Redshift

Daily Aggregation:



Status Connected database dev user awsuser

Select database Info
To view schemas, select a database.
dev

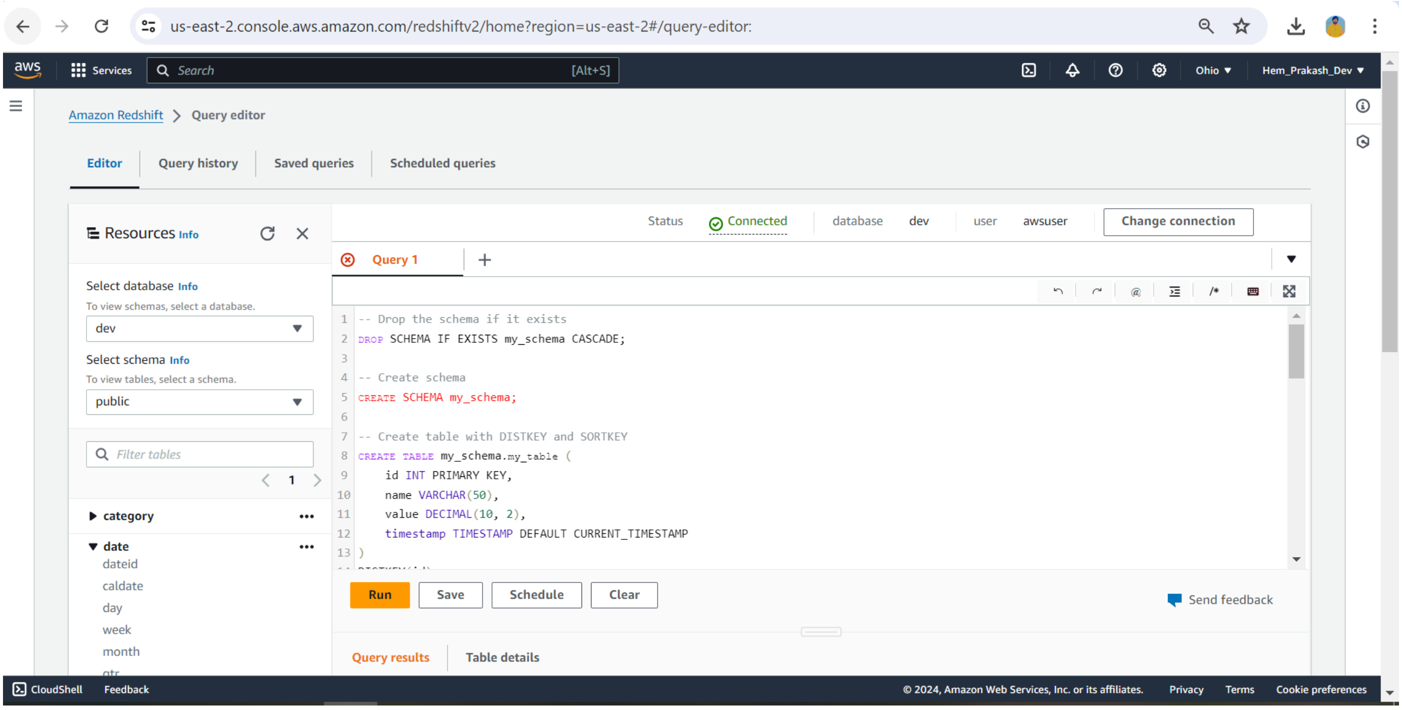
Select schema Info
To view tables, select a schema.
public

```
CREATE TABLE my_schema.daily_aggregation AS
SELECT
    DATE_TRUNC('day', timestamp) AS day,
    COUNT(*) AS total_records,
    SUM(value) AS total_value
FROM my_schema.my_table
GROUP BY 1;
```

Query 3666

Completed, started on July 17, 2024 at 00:39:27
ELAPSED TIME: 00 m 34 s

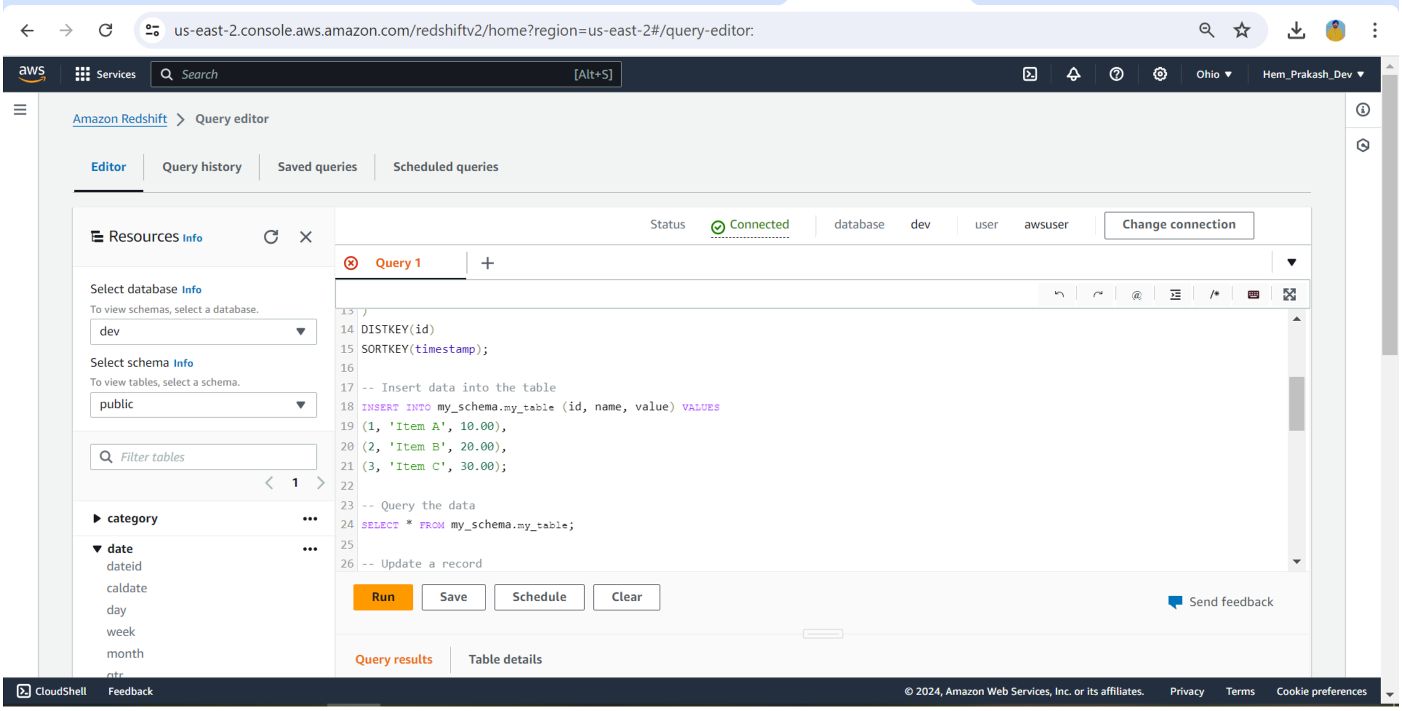
Optimizing queries in a database involves creating appropriate indexes, using **SORTKEY** and **DISTKEY** in databases like Amazon Redshift, and performing regular maintenance like analyzing and vacuuming tables.



```sql
-- Drop the schema if it exists
DROP SCHEMA IF EXISTS my_schema CASCADE;

-- Create schema
CREATE SCHEMA my_schema;

-- Create table with DISTKEY and SORTKEY
CREATE TABLE my_schema.my_table (
    id INT PRIMARY KEY,
    name VARCHAR(50),
    value DECIMAL(10, 2),
    timestamp TIMESTAMP DEFAULT CURRENT_TIMESTAMP
)
```

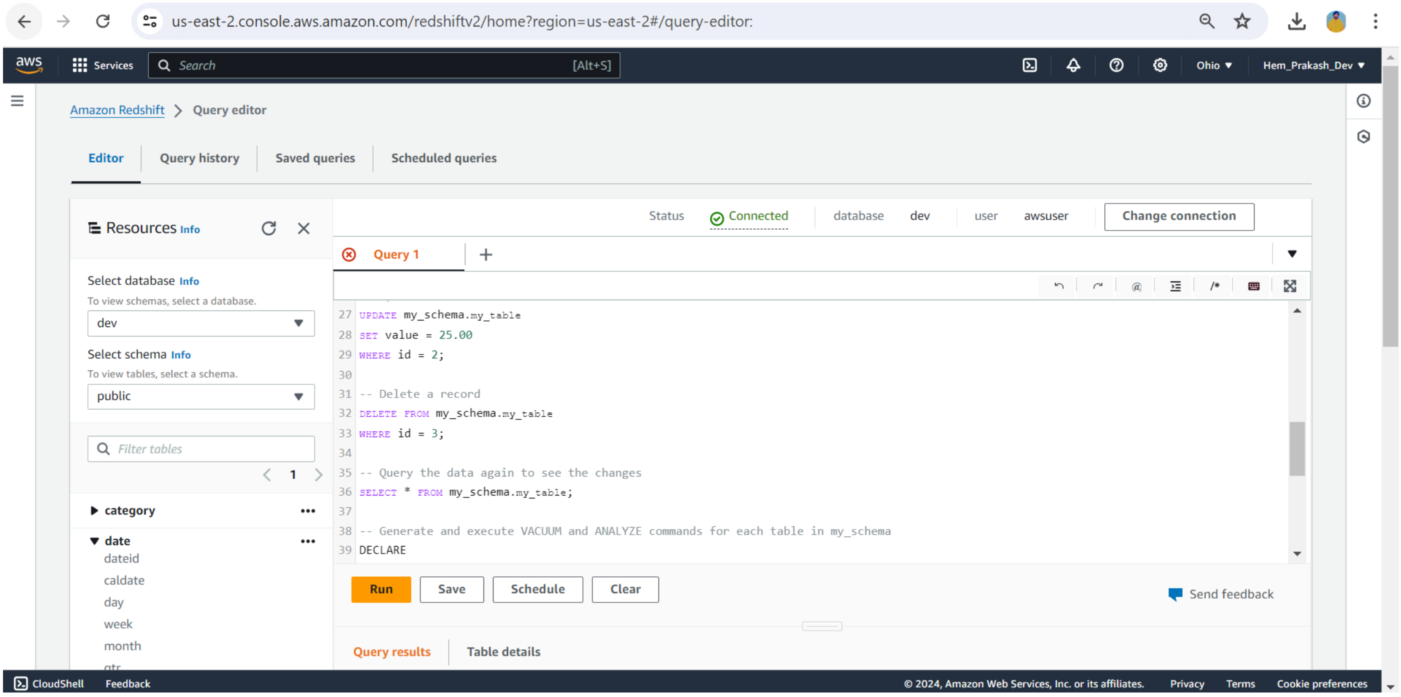


```sql
)
DISTKEY(id)
SORTKEY(timestamp);

-- Insert data into the table
INSERT INTO my_schema.my_table (id, name, value) VALUES
(1, 'Item A', 10.00),
(2, 'Item B', 20.00),
(3, 'Item C', 30.00);

-- Query the data
SELECT * FROM my_schema.my_table;

-- Update a record
```



us-east-2.console.aws.amazon.com/redshiftv2/home?region=us-east-2#/query-editor:

Select database Info: dev

Select schema Info: public

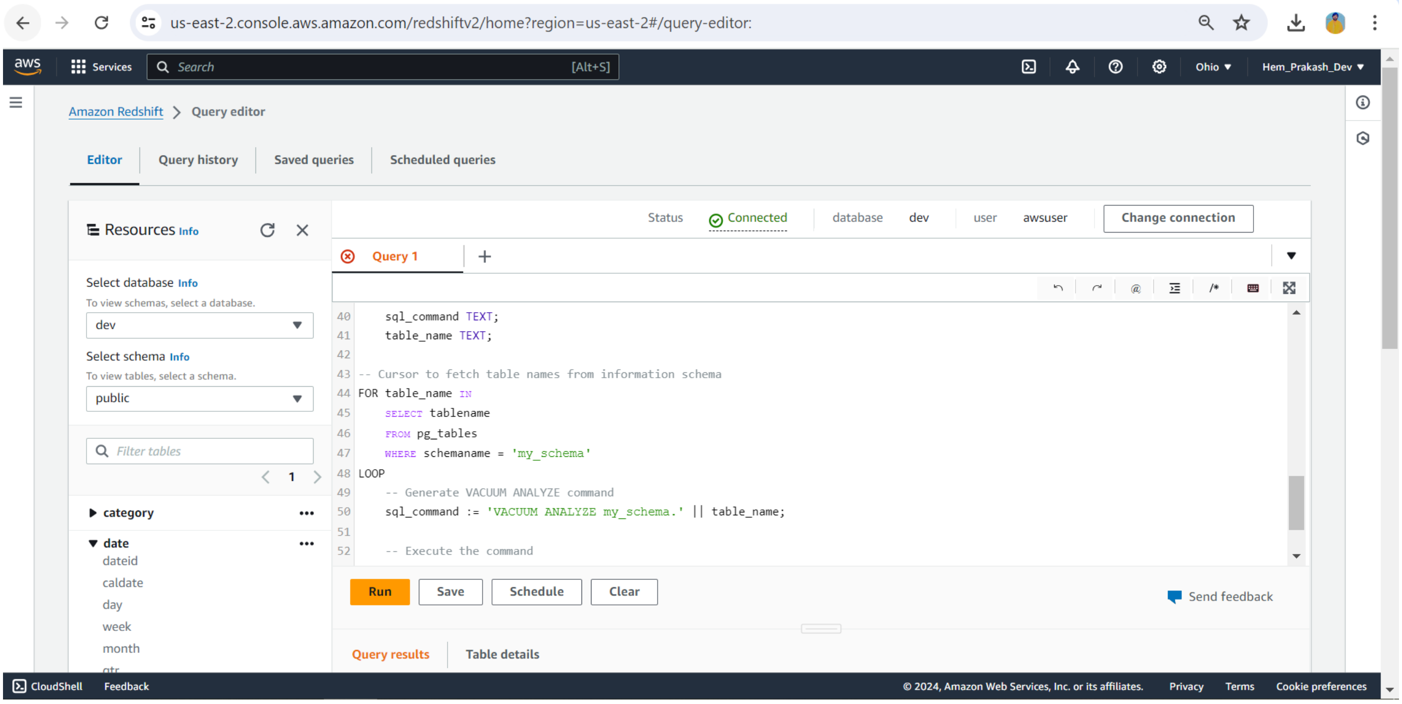
Status: Connected | database: dev | user: awsuser

```sql
UPDATE my_schema.my_table
SET value = 25.00
WHERE id = 2;

-- Delete a record
DELETE FROM my_schema.my_table
WHERE id = 3;

-- Query the data again to see the changes
SELECT * FROM my_schema.my_table;

-- Generate and execute VACUUM and ANALYZE commands for each table in my_schema
DECLARE
```



us-east-2.console.aws.amazon.com/redshiftv2/home?region=us-east-2#/query-editor:

Select database Info: dev

Select schema Info: public

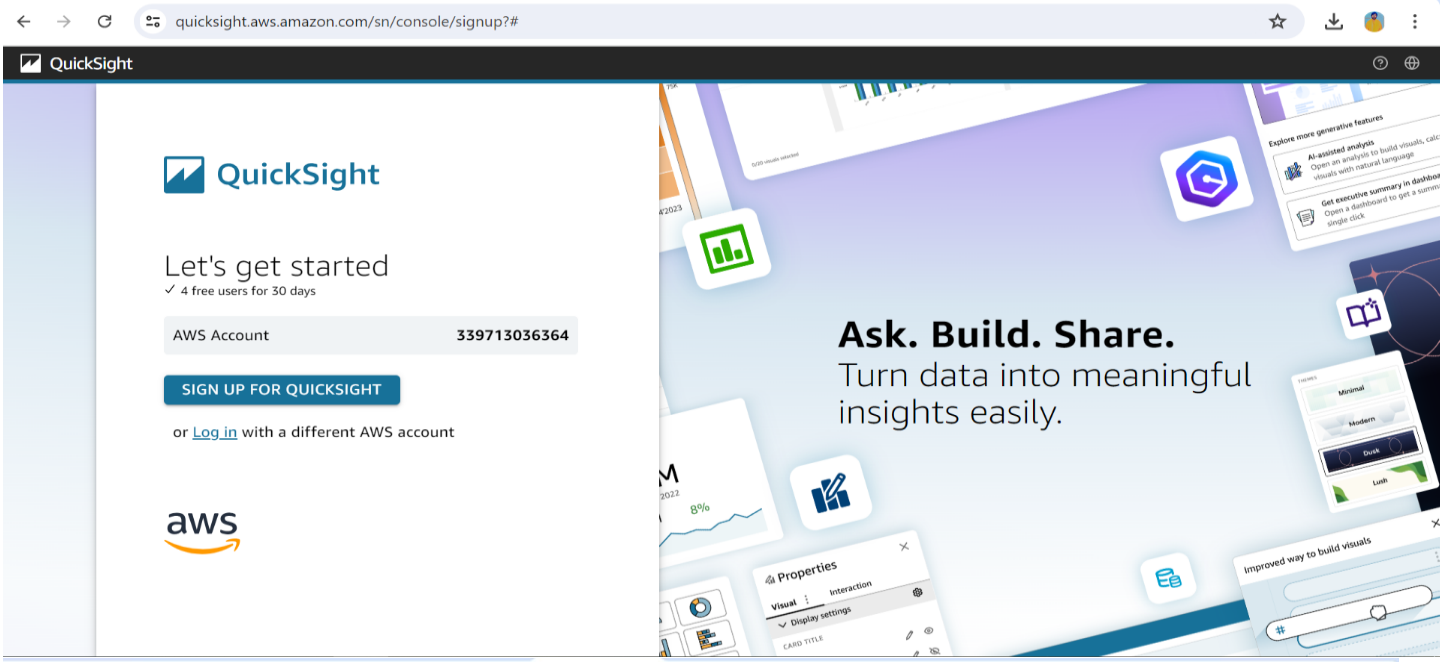
Status: Connected | database: dev | user: awsuser

```sql
    sql_command TEXT;
    table_name TEXT;

-- Cursor to fetch table names from information schema
FOR table_name IN
    SELECT tablename
    FROM pg_tables
    WHERE schemaname = 'my_schema'
LOOP
    -- Generate VACUUM ANALYZE command
    sql_command := 'VACUUM ANALYZE my_schema.' || table_name;

    -- Execute the command
```

3. Integrate Amazon QuickSight or a third-party BI tool for data visualization.



quicksight.aws.amazon.com/sn/console/signup?#

QuickSight

Let's get started

✓ 4 free users for 30 days

AWS Account	339713036364

SIGN UP FOR QUICKSIGHT

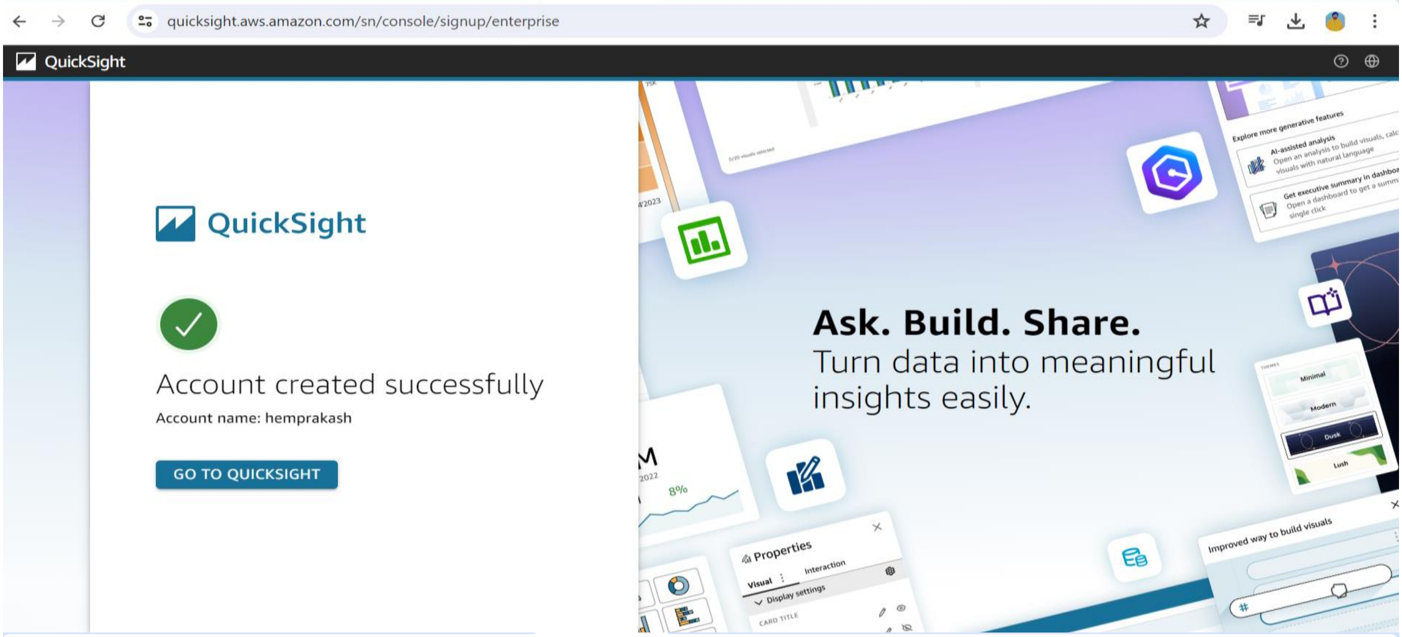
or Log in with a different AWS account

aws

Ask. Build. Share.

Turn data into meaningful insights easily.

Explore more generative features
AI-assisted analysis
Open an analysis to build visuals, calc visuals with natural language
Get executive summary in dashboard
Open a dashboard to get a summ single click
Minimal
Modern
Dusk
Lush
Properties
Visual
Interaction
Display settings
CARD TITLE
Improved way to build visuals
8%



quicksight.aws.amazon.com/sn/console/signup/enterprise

QuickSight

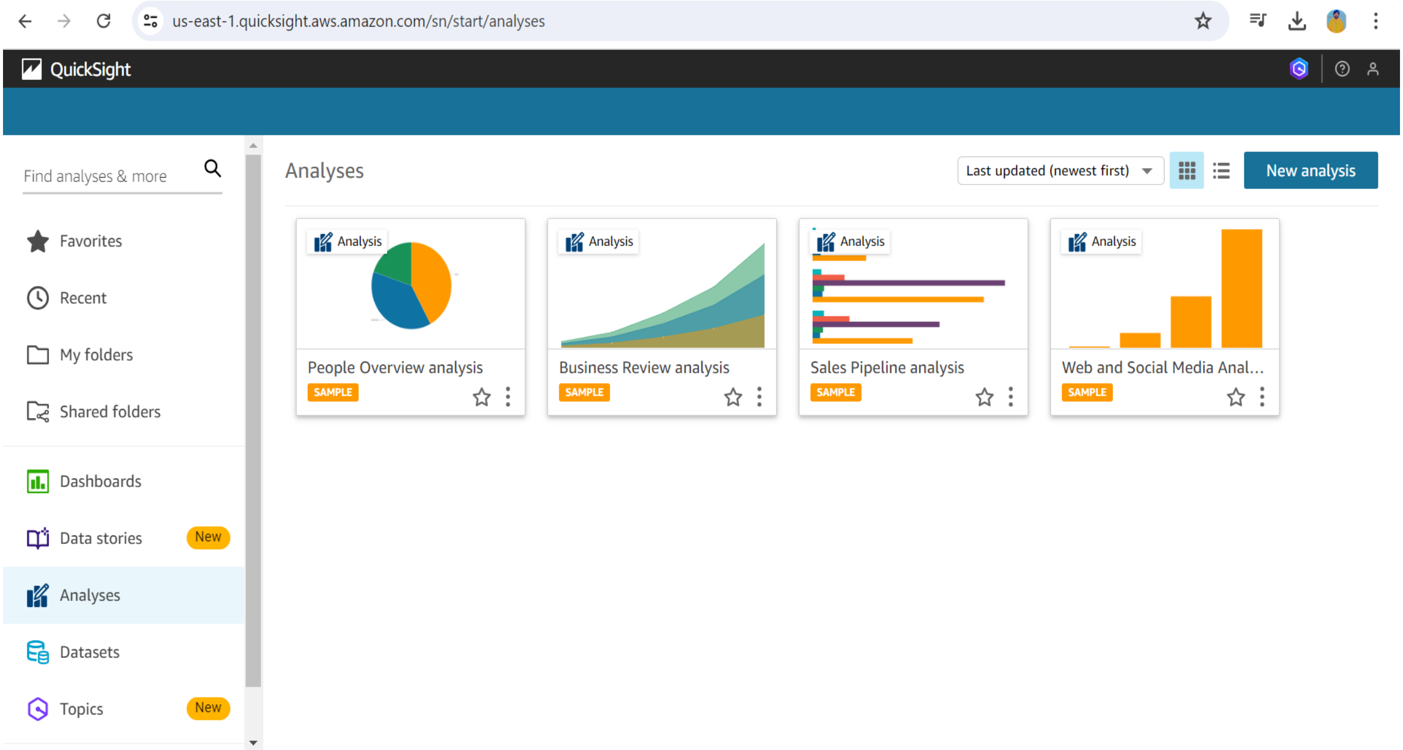
Account created successfully

Account name: hemprakash

GO TO QUICKSIGHT

Ask. Build. Share.

Turn data into meaningful insights easily.



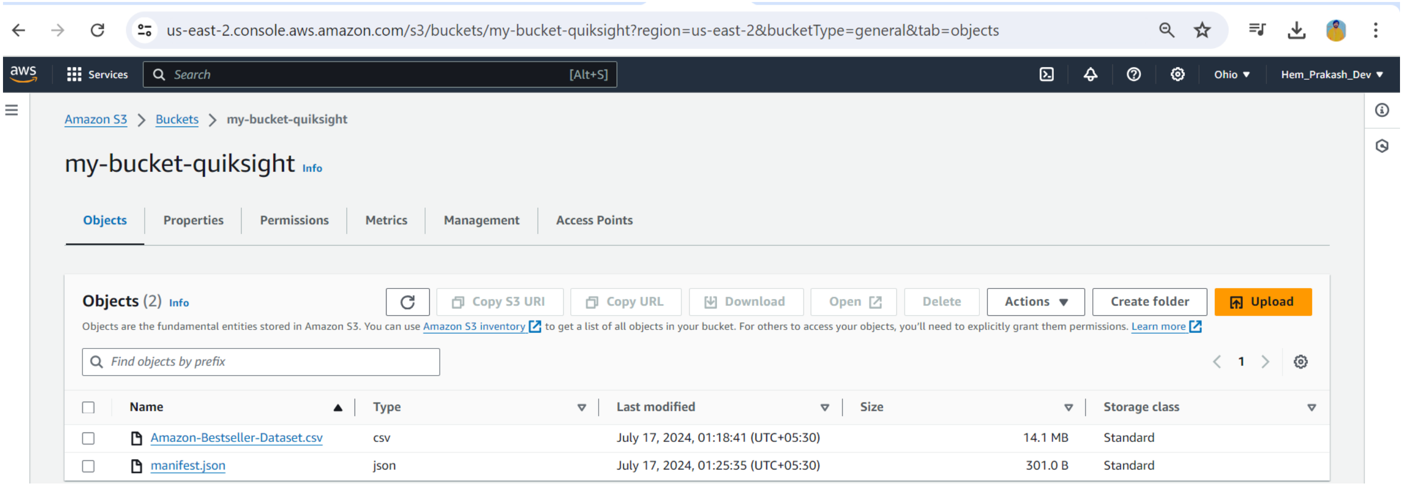
us-east-1.quicksight.aws.amazon.com/sn/start/analyses

QuickSight

Analyses

Analysis	Analysis	Analysis	Analysis
People Overview analysis	Business Review analysis	Sales Pipeline analysis	Web and Social Media Anal...
SAMPLE	SAMPLE	SAMPLE	SAMPLE

Uploaded CSV file (amazon –bestseller-dataset) in amazon s3 bucket



us-east-2.console.aws.amazon.com/s3/buckets/my-bucket-quiksight?region=us-east-2&bucketType=general&tab=objects

Amazon S3 > Buckets > my-bucket-quiksight

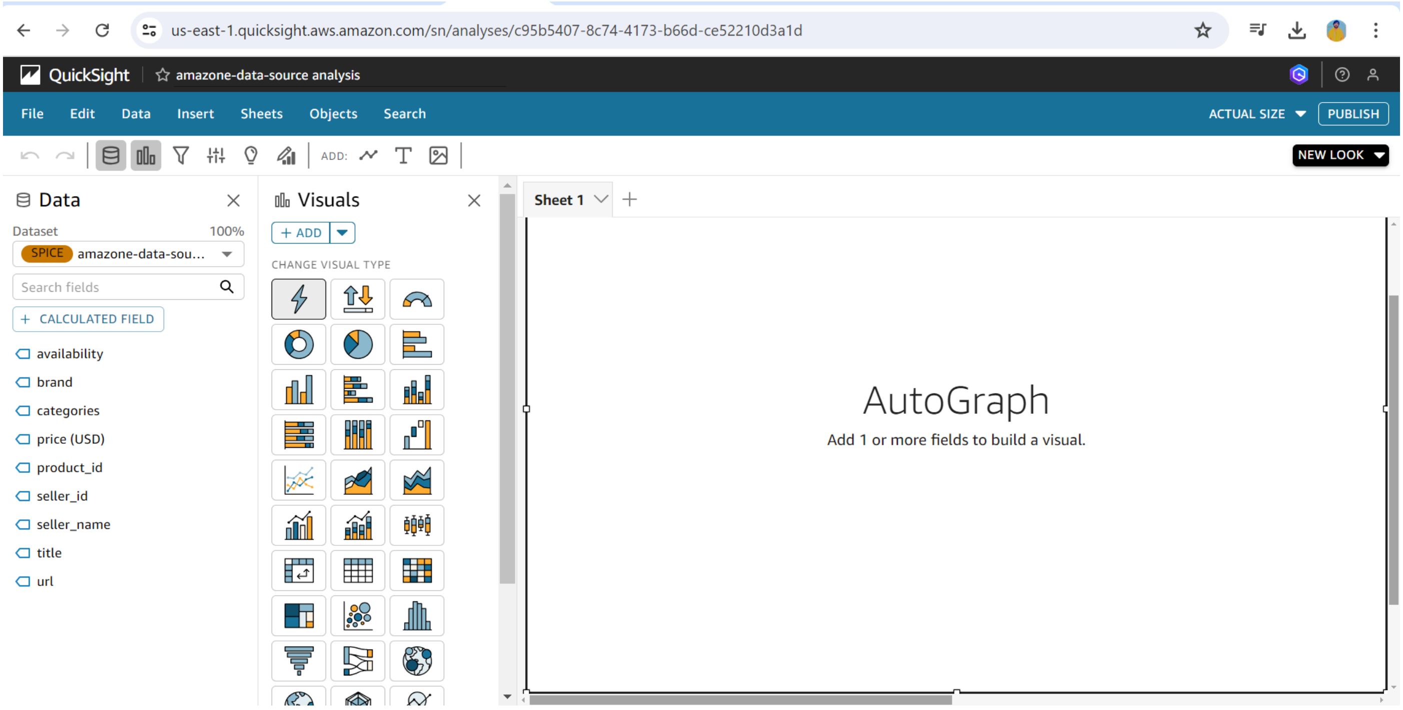
my-bucket-quiksight Info

Objects (2) Info

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
□	Amazon-Bestseller-Dataset.csv	csv	July 17, 2024, 01:18:41 (UTC+05:30)	14.1 MB	Standard
□	manifest.json	json	July 17, 2024, 01:25:35 (UTC+05:30)	301.0 B	Standard

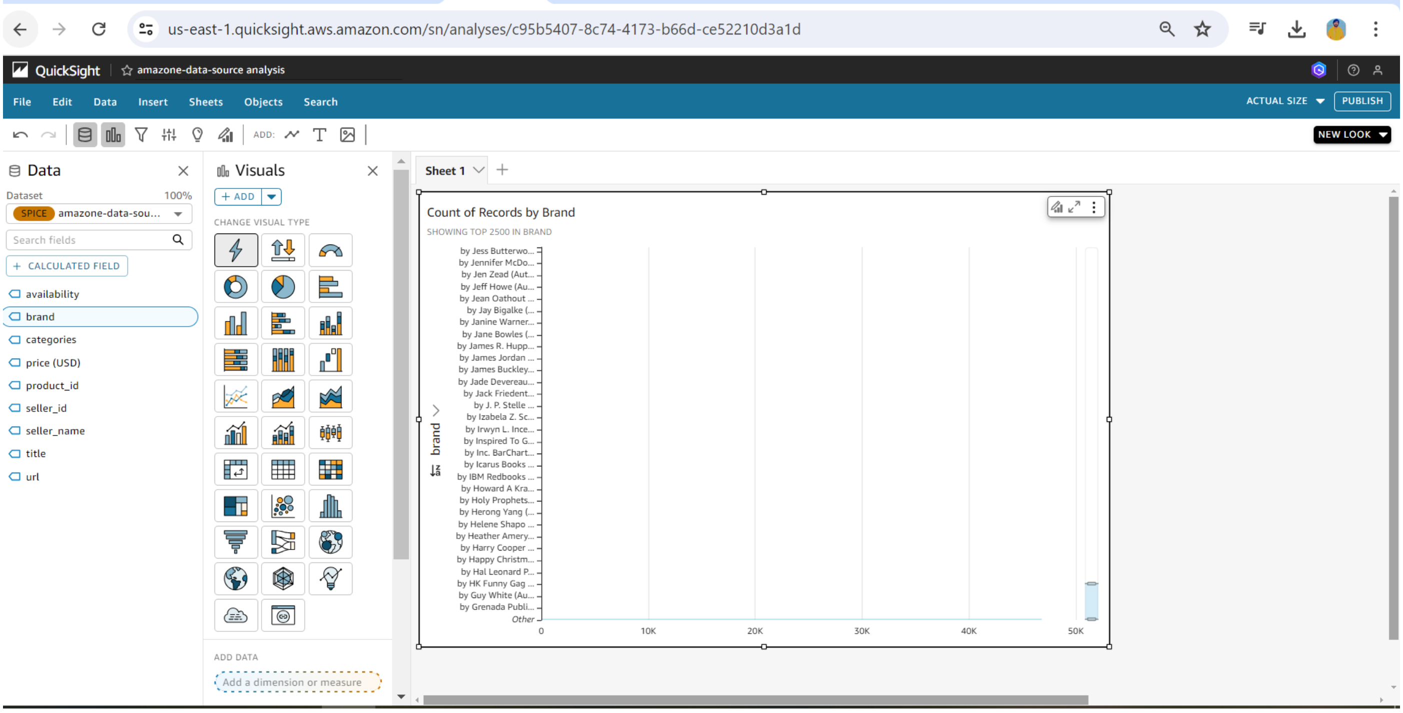


us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

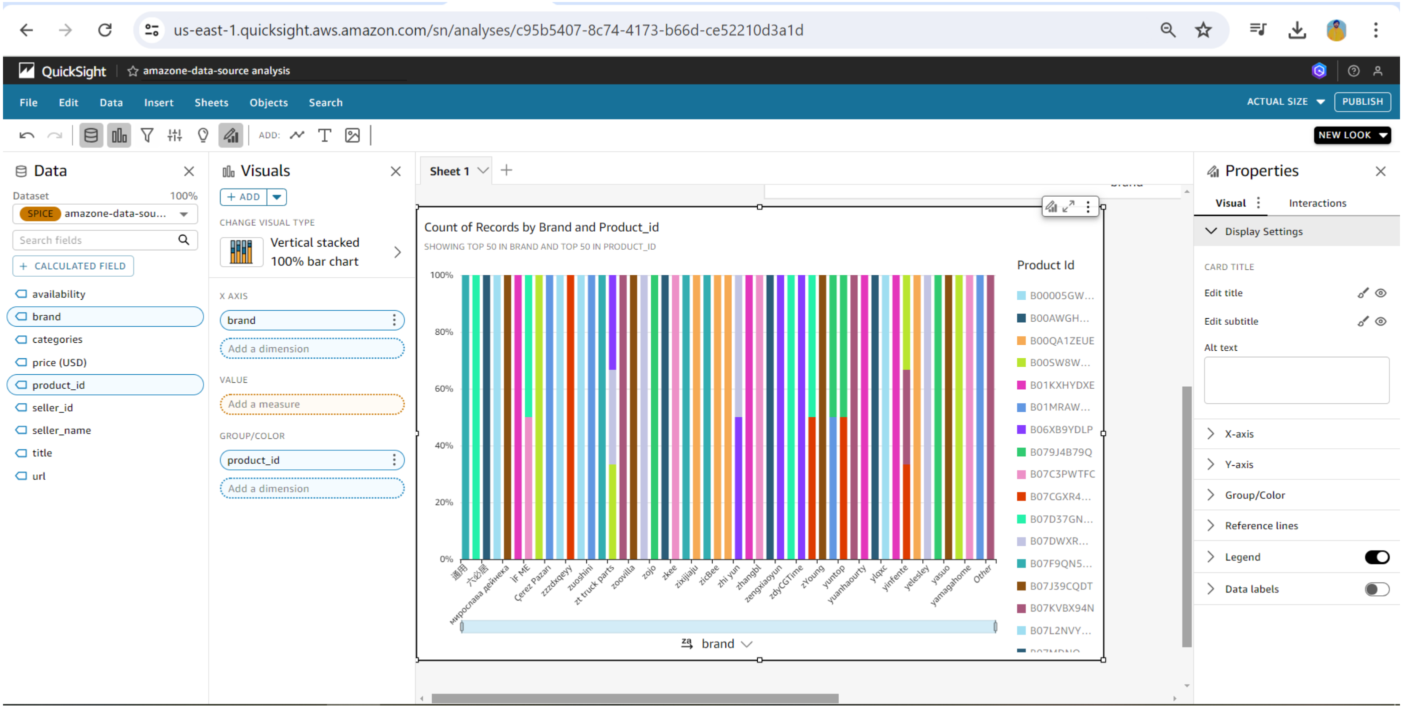
availability
brand
categories
price (USD)
product_id
seller_id
seller_name
title
url

AutoGraph

Add 1 or more fields to build a visual.



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

QuickSight ☆ amazone-data-source analysis

Data — Dataset: SPICE amazone-data-sou… 100%

Fields: availability, brand, categories, price (USD), product_id, seller_id, seller_name, title, url

Visuals — CHANGE VISUAL TYPE: Vertical stacked 100% bar chart

X AXIS: brand

VALUE: Add a measure

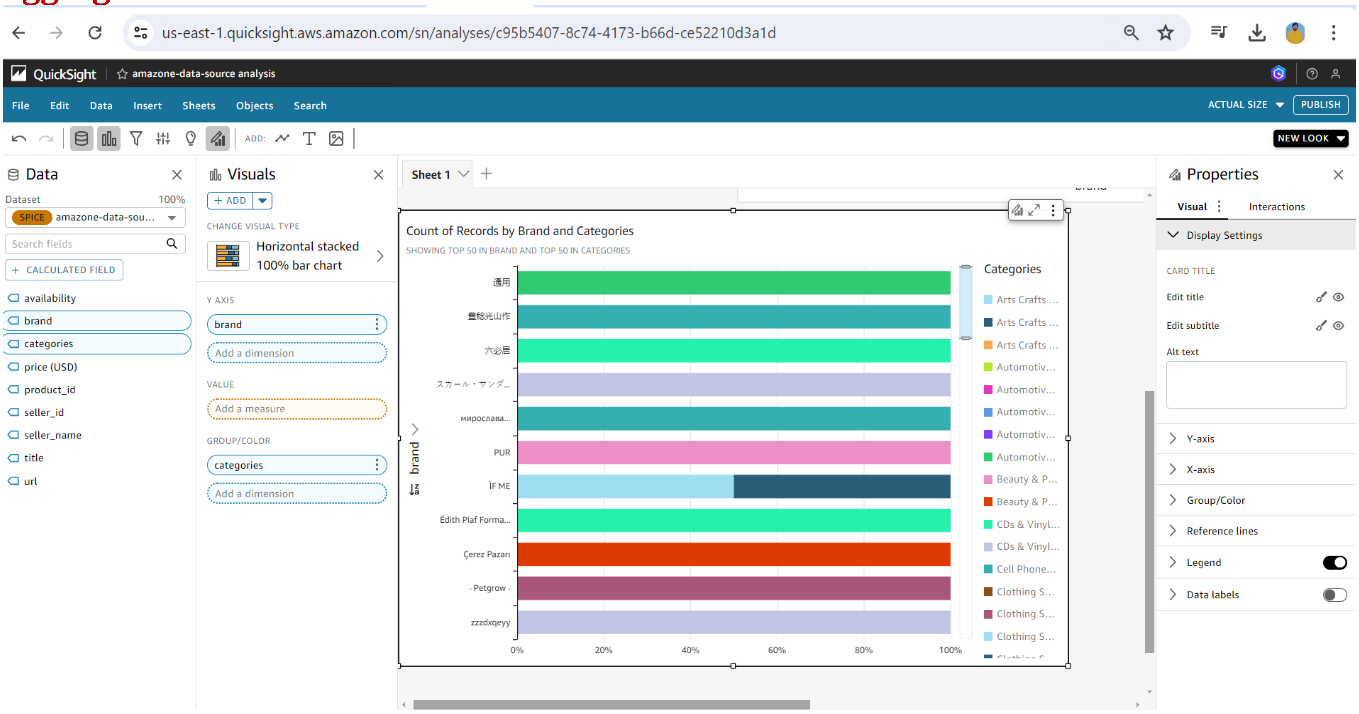
GROUP/COLOR: product_id

Count of Records by Brand and Product_id

SHOWING TOP 50 IN BRAND AND TOP 50 IN PRODUCT_ID

Properties — Visual: Display Settings; Legend on; Data labels off

4. Create sample dashboards and reports to showcase the insights derived from the aggregated data.



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

QuickSight ☆ amazone-data-source analysis

Data — Dataset: SPICE amazone-data-sou… 100%

Fields: availability, brand, categories, price (USD), product_id, seller_id, seller_name, title, url

Visuals — CHANGE VISUAL TYPE: Horizontal stacked 100% bar chart

Y AXIS: brand

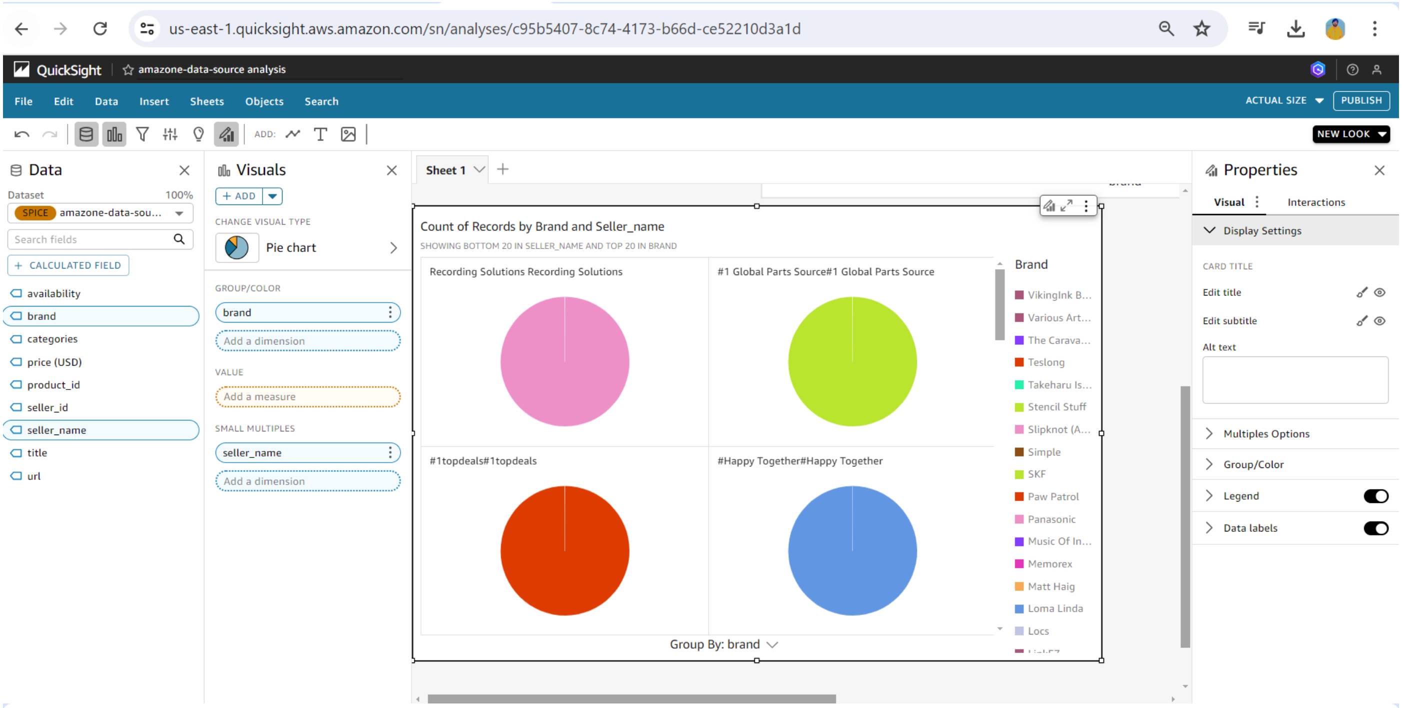
VALUE: Add a measure

GROUP/COLOR: categories

Count of Records by Brand and Categories

SHOWING TOP 50 IN BRAND AND TOP 50 IN CATEGORIES

Properties — Visual: Display Settings; Legend on; Data labels off



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

QuickSight ☆ amazone-data-source analysis

Dataset: SPICE amazone-data-sou... 100%

Fields: availability, brand, categories, price (USD), product_id, seller_id, seller_name, title, url

CHANGE VISUAL TYPE: Pie chart

GROUP/COLOR: brand

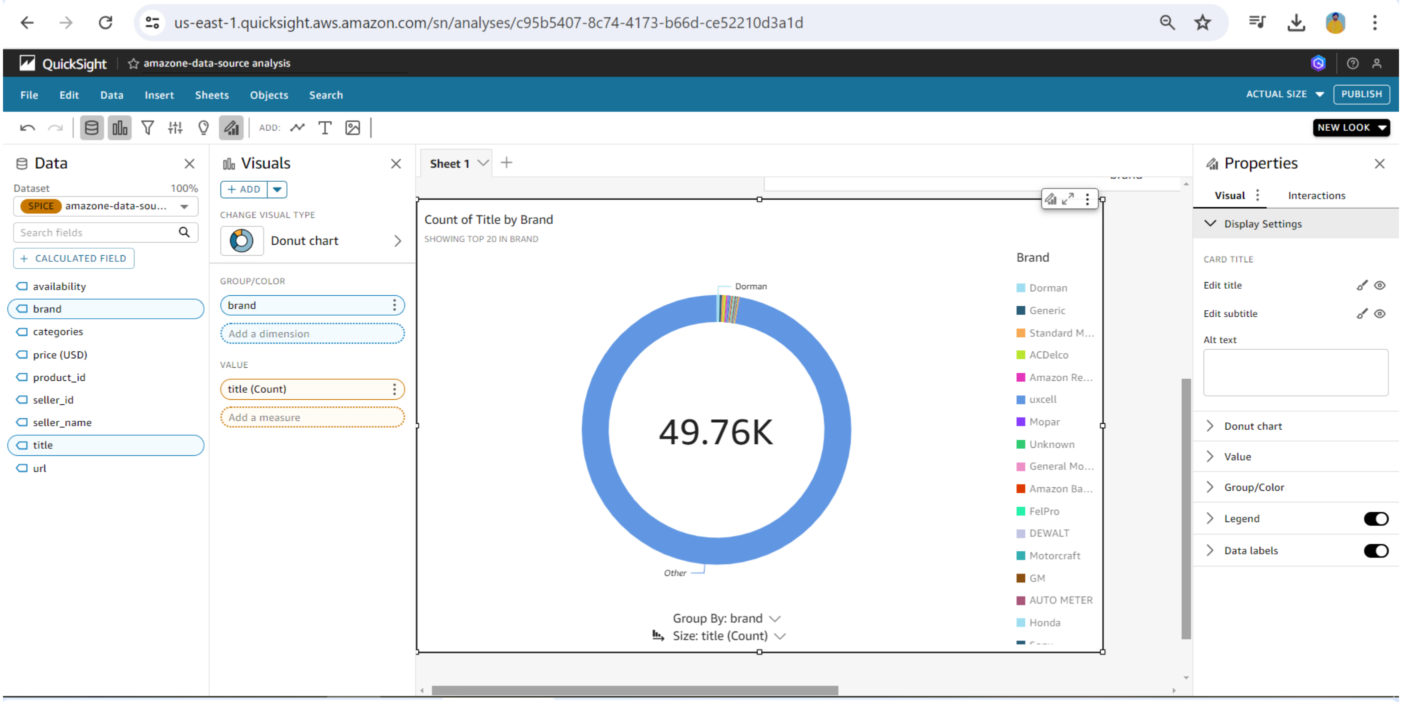
VALUE: Add a measure

SMALL MULTIPLES: seller_name

Count of Records by Brand and Seller_name

SHOWING BOTTOM 20 IN SELLER_NAME AND TOP 20 IN BRAND

Properties — Visual — Display Settings: Legend ●, Data labels ●



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

QuickSight ☆ amazone-data-source analysis

CHANGE VISUAL TYPE: Donut chart

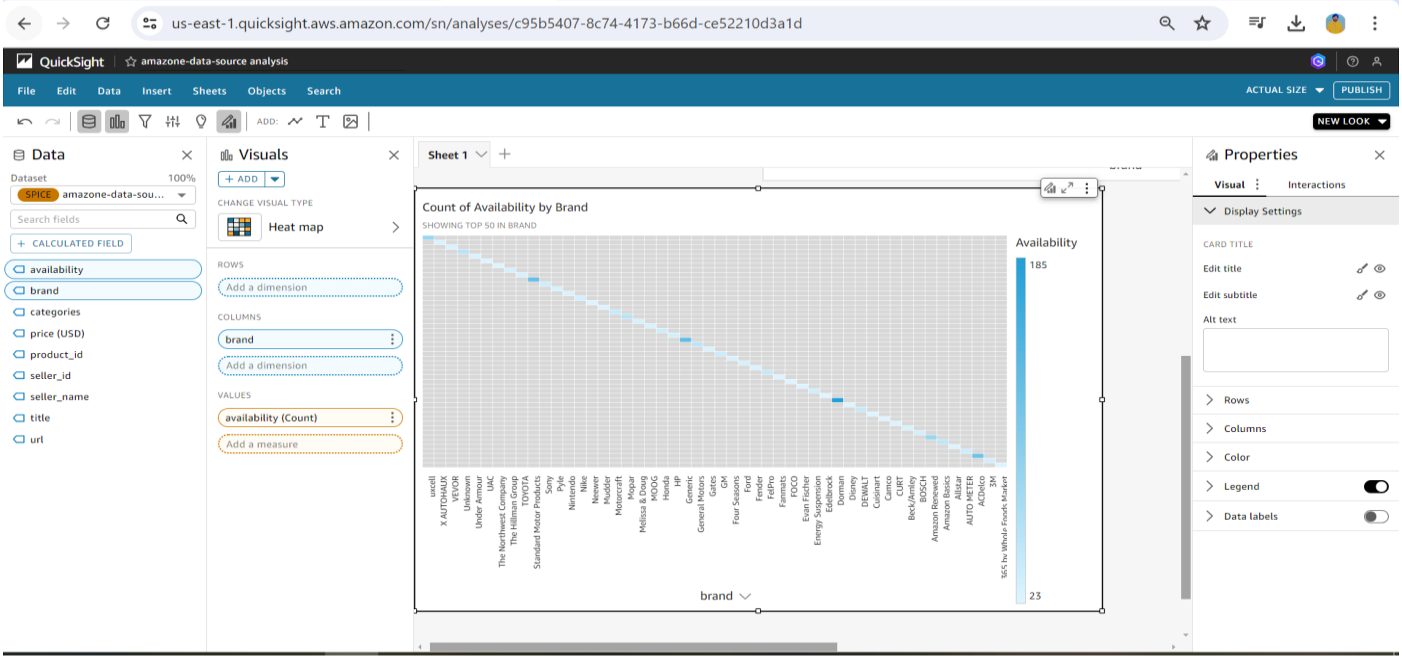
GROUP/COLOR: brand

VALUE: title (Count)

Count of Title by Brand

SHOWING TOP 20 IN BRAND

Properties — Visual — Display Settings: Legend ●, Data labels ●



us-east-1.quicksight.aws.amazon.com/sn/analyses/c95b5407-8c74-4173-b66d-ce52210d3a1d

QuickSight ☆ amazone-data-source analysis

CHANGE VISUAL TYPE: Heat map

ROWS: Add a dimension

COLUMNS: brand

VALUES: availability (Count)

Count of Availability by Brand

SHOWING TOP 50 IN BRAND

Properties — Visual — Display Settings: Legend ●, Data labels ○

5. Update the GitHub repository with the code and configuration files for data aggregation and visualization

```
git init
git remote add origin https://github.com/Hem-Prakash-Dev-Bharadwaj/data-engineering
git add .
git commit -m "Initial commit with Redshift setup, aggregation queries, and QuickSight configuration"
git push -u origin master
```