- Week 3: Project Walkthru: Connecting to Redshift
- **Data Ingestion Practices**

)

Introducing to Google BigQuery

```
AWS: Launching a Redshift Cluster
วิธีเข้าใช้งาน AWS Management Console ฝาน AWS Learner Lab
https://youtu.be/T8noPec75po
วิธีสร้าง Amazon Redshift cluster ขึ้นมาใช้งาน
https://youtu.be/BXBb3E95I6Y
วิธีการทำ Data Ingestion จาก S3 เข้า Amazon Redshift
ข้อมูล Titanic
https://github.com/datasciencedojo/datasets/blob/master/titanic.csv
ตัวอย่างโค้ดที่ใช้ในการสร้างตารางใน Redshift
CREATE TABLE IF NOT EXISTS titanic (
 PassengerId text primary key,
 Survivedt text,
 Pclass text.
 Name text,
 Sex text,
 Age text,
 SibSp text,
 Parch text,
 Ticket text,
 Fare text,
 Cabin text,
 Embarked text
)
ตัวอย่างโค้ดที่ใช้ในการโหลดข้อมูลจาก S3 เข้า Redshift
COPY titanic FROM 's3://zkan-swu-labs/titanic.csv'
CREDENTIALS 'aws iam role=arn:aws:iam::377290081649:role/LabRole'
IGNOREHEADER 1 CSV REGION 'us-east-1'
ตัวอย่างโค้ดในการทำ Transformation โดยใช้ SQL
INSERT INTO
 new_table (
       name
```

```
SELECT
Name
FROM
 titanic
WHERE
Name NOT IN (SELECT DISTINCT name FROM new table)
Last modified: Saturday, 10 September 2022, 2:21 PM
GCP: Setting up BigQuery
Google Cloud Platform: <a href="https://cloud.google.com/">https://cloud.google.com/</a>
ข้อมูล Titanic
https://github.com/datasciencedojo/datasets/blob/master/titanic.csv
ตัวอย่างโค้ดที่ใช้ในการสร้างตารางใน BigQuery
CREATE TABLE IF NOT EXISTS 'swu.titanic' (
 PassengerId STRING,
 Survivedt STRING,
 Pclass STRING,
 Name STRING,
 Sex STRING,
 Age STRING,
 SibSp STRING,
 Parch STRING,
 Ticket STRING,
 Fare STRING,
 Cabin STRING,
 Embarked STRING
)
ตัวอย่างโค้ดที่ใช้ในการโหลดเข้า BigQuery จาก Google Cloud Storage (GCS)
LOAD DATA OVERWRITE 'swu.titanic'
FROM FILES (
format = 'CSV',
uris = ['gs://bucket/path/file.csv']
);
ตัวอย่างโค้ดในการทำ Transformation โดยใช้ SQL
MERGE swu staging.titanic as dest
USING swu.titanic as src
ON dest.PassengerId = src.PassengerId
WHEN NOT MATCHED THEN
INSERT (PassengerId, Name)
 VALUES (src.PassengerId, src.Name)
```

```
WHEN MATCHED THEN
 UPDATE SET
  dest.PassengerId = src.PassengerId
  , dest.Name = src.Name
ดูรายละเอียดเพิ่มเติมเกี่ยวกับ Merge statement ใน BigQuery ได้ที่ Data manipulation language (DML)
statements in Google Standard SQL
Last modified: Saturday, 10 September 2022, 2:33 PM
Week 3
https://voutu.be/cvdYqx5UMLo
Week 3 Office Hour
https://youtu.be/DXCVY8obmZQ
git config pull.rebase false
เสร็จแล้วก็ git pull
create table if not exists category_staging (
catname text
);
insert into category staging (catname)
select
 case
  when catname = 'NFL' then 'Hello World'
 end
from category
where catname not in (select distinct catname from category staging);
select * from category staging
delete from category staging where catname = 'NFL'
User: introde413@gmail.com
Pass: 5qGn@0udHZ!7XW9U7fh8Sp
Week 4 (Part 1)
https://youtu.be/iVlkcoe_IGc
Week 4 (Part 2)
https://youtu.be/BKFPZkwSUkc
```

Project: Building an ETL Pipeline for a Cloud Data Warehouse (Redshift/Google BigQuery) สามารถเริ่มต้นจากโค้ดที่ https://github.com/zkan/swu-ds525 ได้

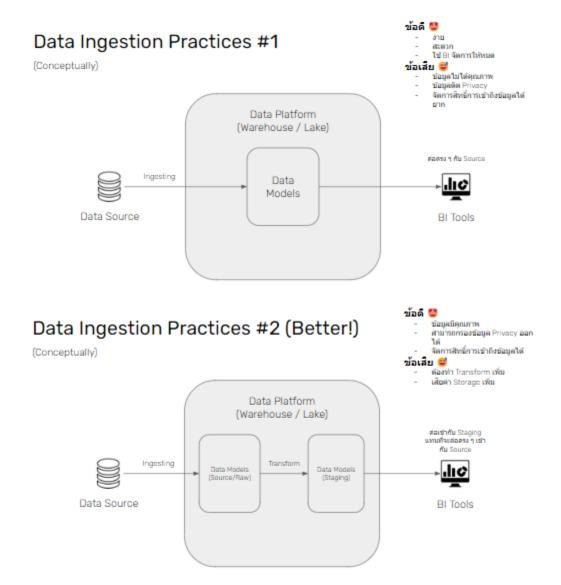
ในโปรเจคนี้เราจะใช้ข้อมูล GitHub event data จาก <u>API</u> หรือสามารถดาวน์โหลดไฟล์ JSON ได้ที่ <u>URL</u> นี้ สิ่งที่คาดหวังในโปรเจคนี้

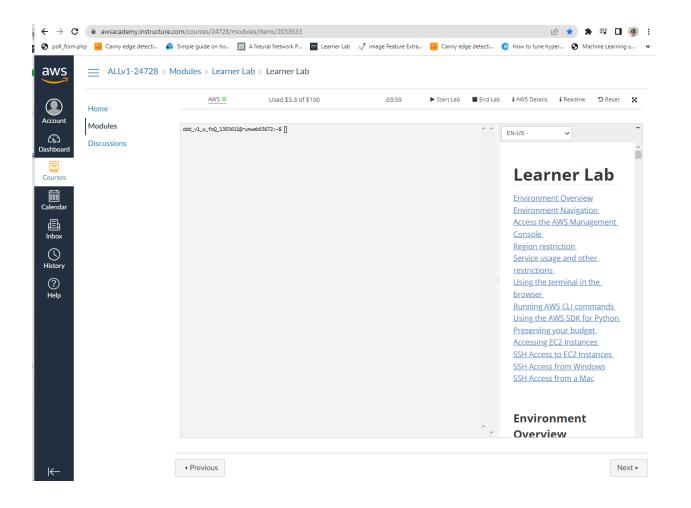
ในโปรเจคนี้เราจะตั้งต้นว่าเรามีไฟล์อยู่บน S3 อยู่แล้ว ดังนั้นทุกคนสามารถที่จะอัพโหลดไฟล์ขึ้นไปไว้บน S3 ได้เลย

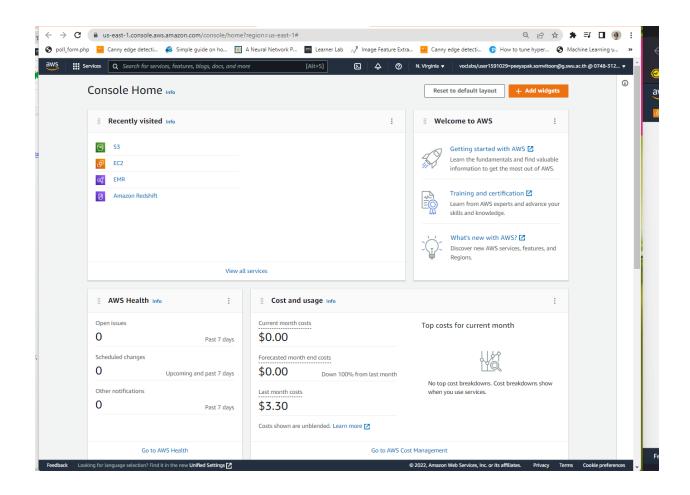
- 1. มีโค้ดที่ทำ ETL จากข้อมูล JSON โหลดเข้าไปใน Redshift ตาม data model ที่เราออกแบบไว้
- 2. มีการเขียน documentation อธิบายสิ่งที่ตัวเองทำลงไป รวมไปถึงการออกแบบ data model
- 3. มี instruction ในการรันโค้ดของตัวเอง

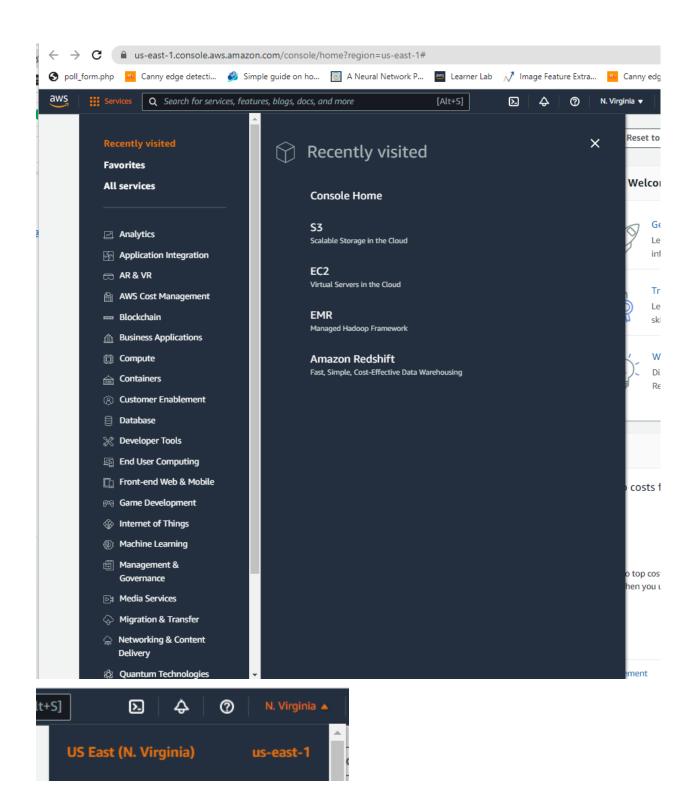
อ่านวิธีโหลด JSON เข้า Redshift ได้ที่ <u>COPY from JSON format</u> Last modified: Saturday, 3 September 2022, 8:05 AM

```
COPY titanic FROM 's3://zkan-swu-labs/titanic.csv'
CREDENTIALS 'aws_iam_role=arn:aws:iam::377290081649:role/LabRole'
IGNOREHEADER 1 CSV REGION 'us-east-1'
```











Recently visited

Favorites

All services

Analytics

- Application Integration
- AR & VR
- AWS Cost Management
- **Blockchain**
- Business Applications
- Compute
- Containers
- Customer Enablement
- Database
- End User Computing
- Front-end Web & Mobile
- Internet of Things
- Machine Learning
- Management & Governance
- A Migration & Transfer
- Networking & Content Delivery

AWS Data Exchange

Easily exchange data in the cloud

Data Pipeline

Orchestration for Data-Driven Workflows

EMR

Managed Hadoop Framework

Amazon FinSpace

Store, catalog, prepare, and analyze financial industry data

AWS Glue

AWS Glue is a fully managed ETL (extract, transform, and load) service

AWS Glue DataBrew

Visual data preparation tool to clean and normalize data for analytics and machine learning

Kinesis

Work with Real-Time Streaming Data

AWS Lake Formation

AWS Lake Formation makes it easy to set up a secure data lake

MSK

Fully managed, highly available, and secure service for Apache Kafka

Amazon OpenSearch Service

Run and Scale OpenSearch and Elasticsearch Clusters (successor to Amazon Elasticsearch Service)

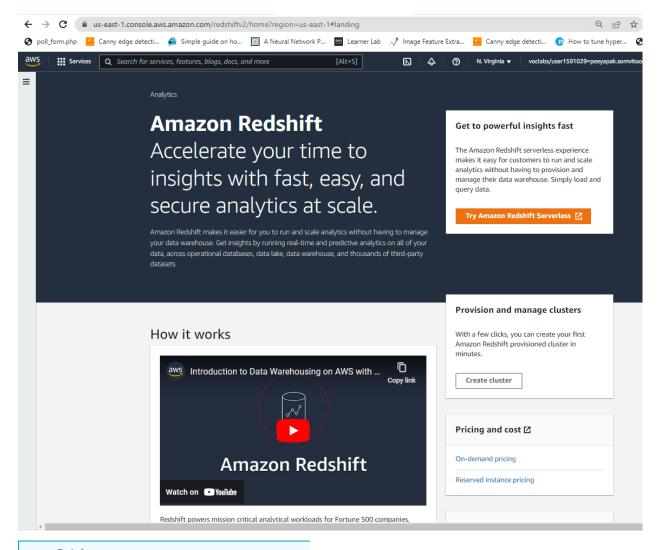
QuickSight 2

Fast, easy to use business analytics

Amazon Redshift

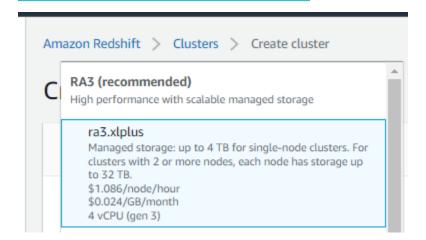
Fast, Simple, Cost-Effective Data Warehousing



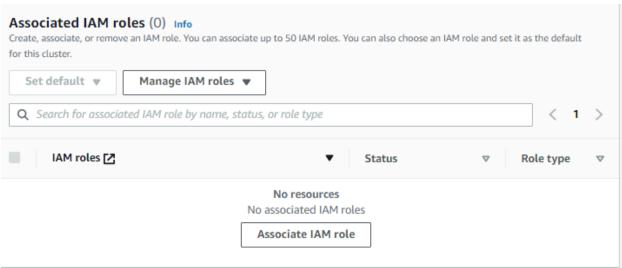


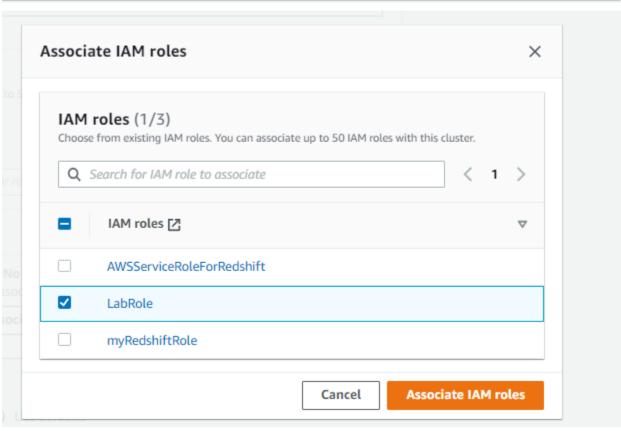
ra3.4xlarge

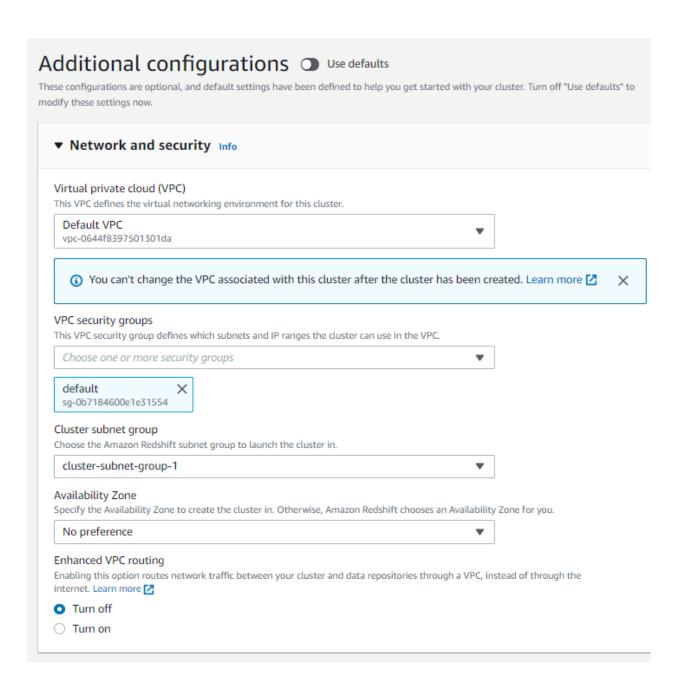
Managed storage: up to 128 TB/node \$3.26/node/hour \$0.024/GB/month 12 vCPU (gen 3)

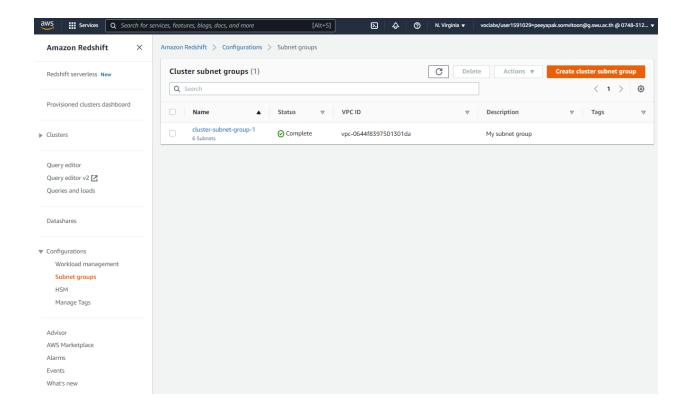


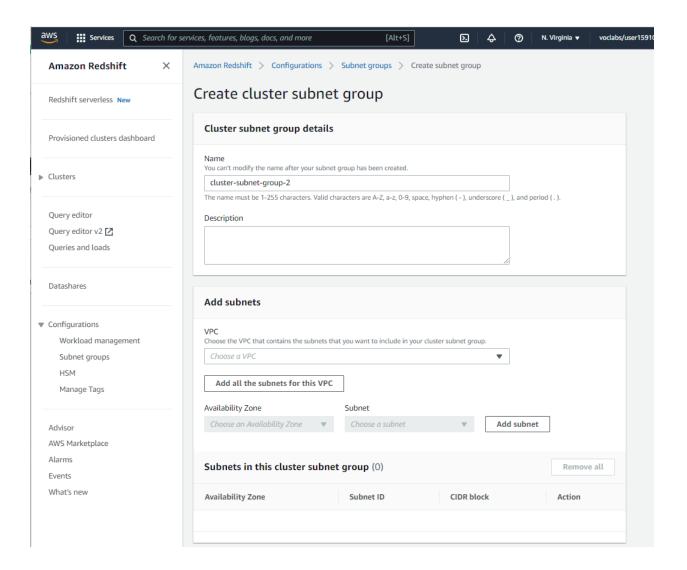
Cluster configuration Cluster identifier This is the unique key that identifies a cluster. redshift-cluster-1 The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen). What are you planning to use this cluster for? Production Free trial Configure for fast and consistent performance at the best Configure for learning about Amazon Redshift. This configuration is free for a limited time if your price. organization has never created an Amazon Redshift Choose the size of the cluster I'll choose Help me choose Node type Info Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements. ra3.xlplus Number of nodes Enter the number of nodes that you need. Range (1-32) Configuration summary Info ra3.xlplus | 1 node Sample data Info Load sample data Load sample data to your Redshift cluster to start using the query editor to query data. Tickit (28 MB) Tickit is the sample data set that uses a sample database called TICKIT. Tickit contains individual sample data files: two fact tables and five dimensions.



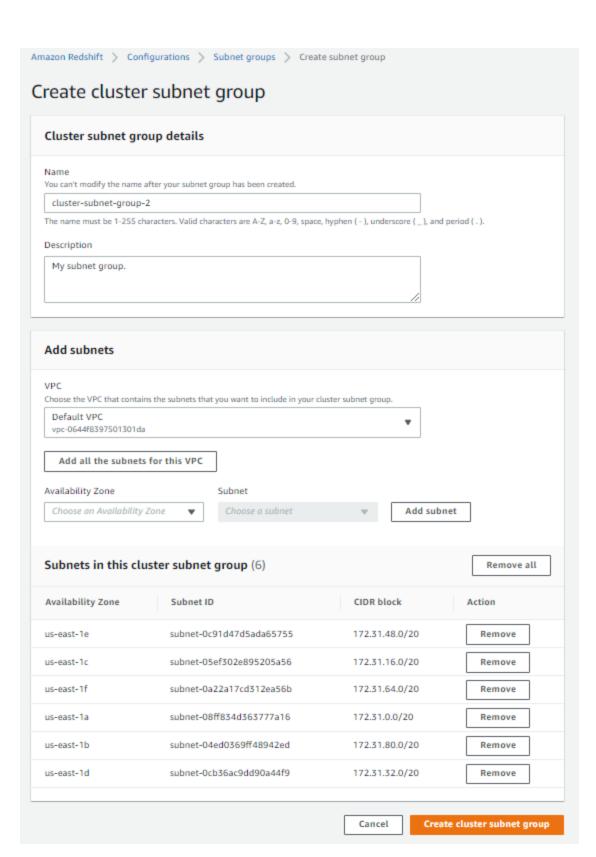






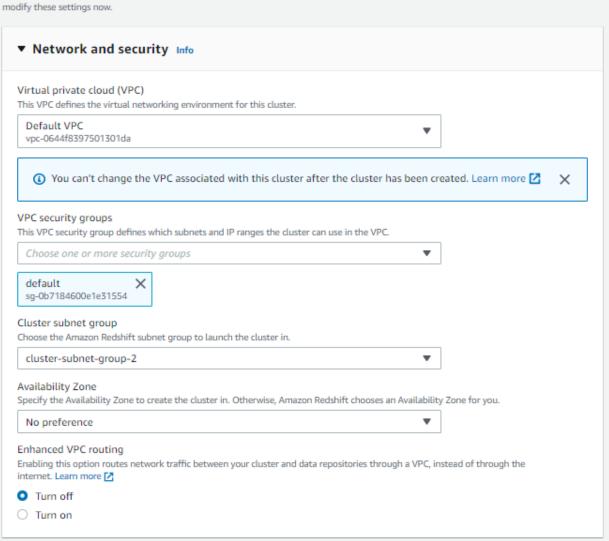


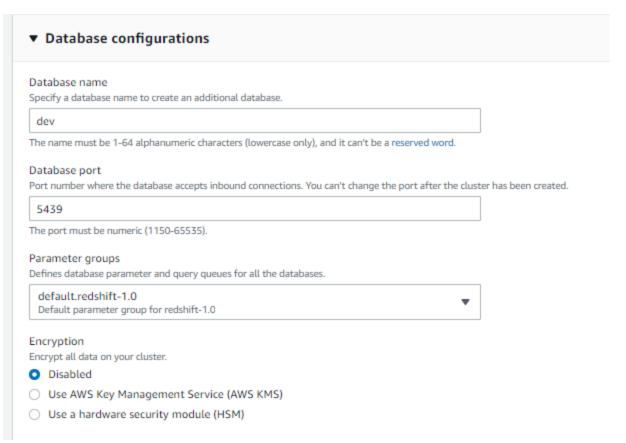
Cluster subnet group de	etails
Name You can't modify the name after your	r subnet group has been created.
cluster-subnet-group-2	
The name must be 1-255 characters.	Valid characters are A-Z, a-z, 0-9, space, hyphen (-), underscore (_), and period (.).
Description	
Add subnets	
Add subnets	
VPC	posts that you want to include in your cluster subnet group
VPC Choose the VPC that contains the sub	onets that you want to include in your cluster subnet group.
Default VPC	onets that you want to include in your cluster subnet group.
VPC Choose the VPC that contains the sub	
VPC Choose the VPC that contains the sub Default VPC	▼
VPC Choose the VPC that contains the sub Default VPC vpc-0644f8397501301da	▼
VPC Choose the VPC that contains the sub Default VPC vpc-0644f8397501301da	▼

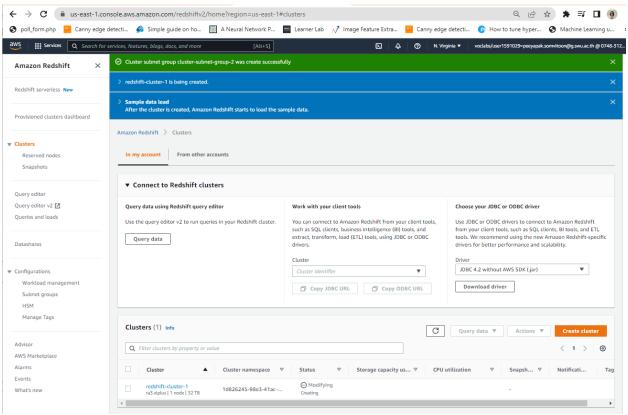


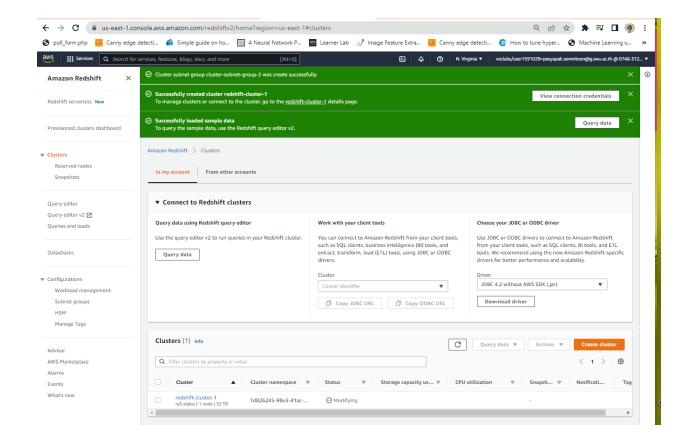
Additional configurations • Use defaults

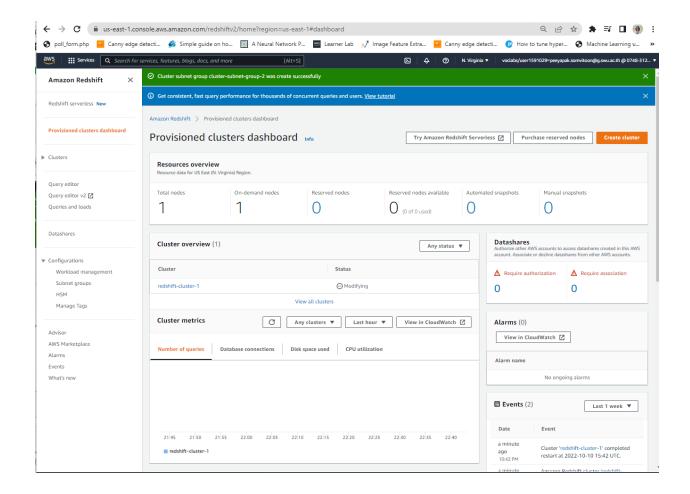
These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

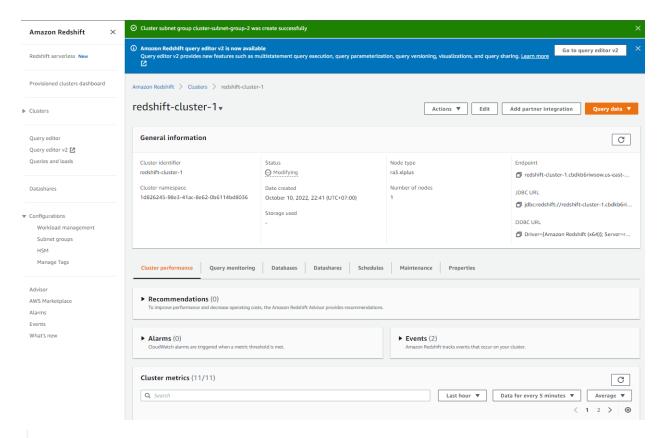






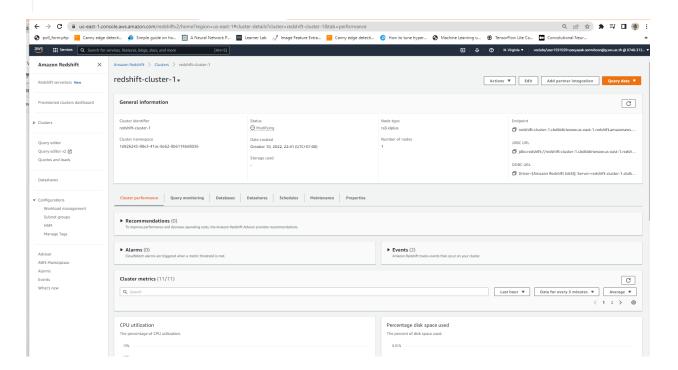


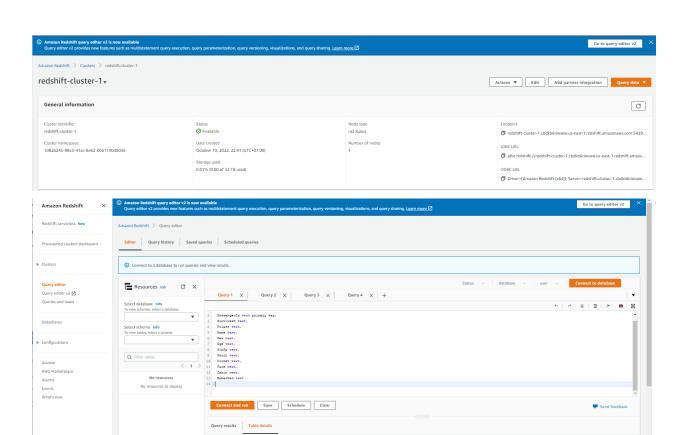




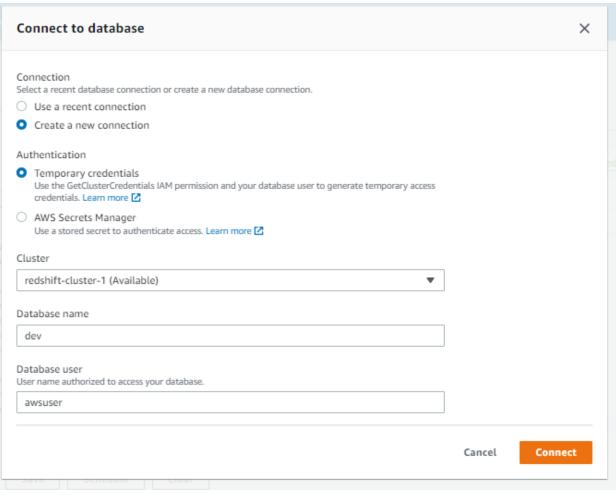
Endpoint

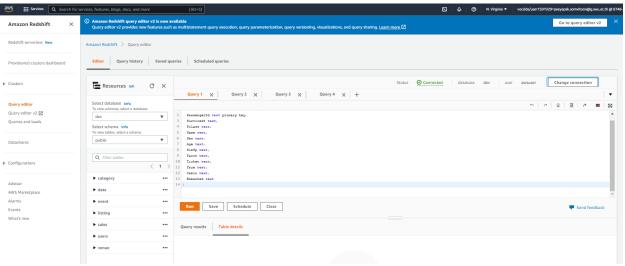
redshift-cluster-1.cbdkb6riwsow.us-east-...

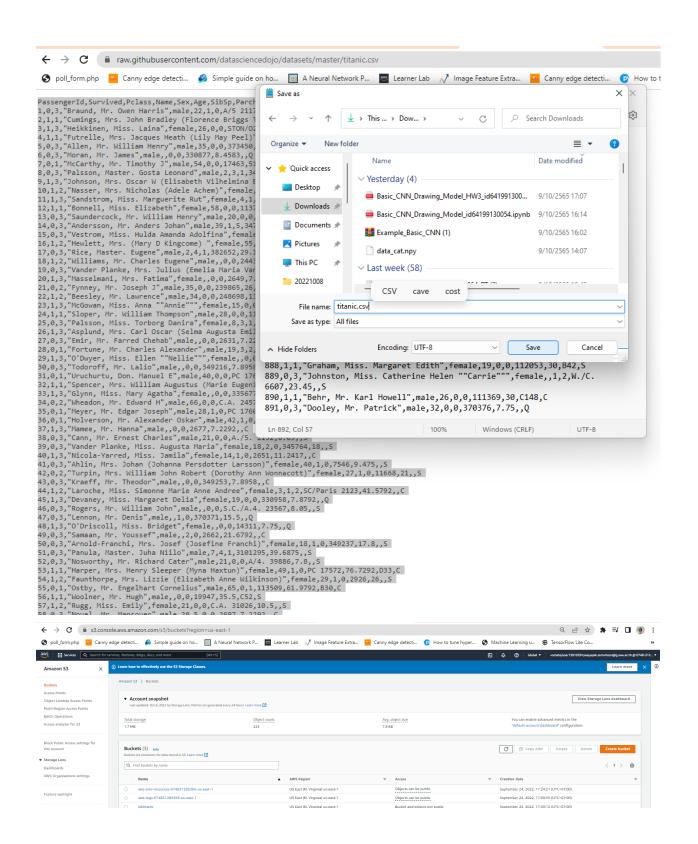




Connect to database







AWS: Launching a Redshift Cluster

วิธีเข้าใช้งาน AWS Management Console ผ่าน AWS Learner Lab

https://youtu.be/T8noPec75po

วิธีสร้าง Amazon Redshift cluster ขึ้นมาใช้งาน

https://youtu.be/BXBb3E95I6Y

วิธีการทำ Data Ingestion จาก S3 เข้า Amazon Redshift

ข้อมูล Titanic

https://github.com/datasciencedojo/datasets/blob/master/titanic.csv

ตัวอย่างโค้ดที่ใช้ในการสร้างตารางใน Redshift

```
CREATE TABLE IF NOT EXISTS titanic (
   PassengerId text primary key,
   Survivedt text,
   Pclass text,
   Name text,
   Sex text,
   Age text,
   SibSp text,
   Parch text,
   Ticket text,
   Fare text,
   Cabin text,
   Embarked text
)
```

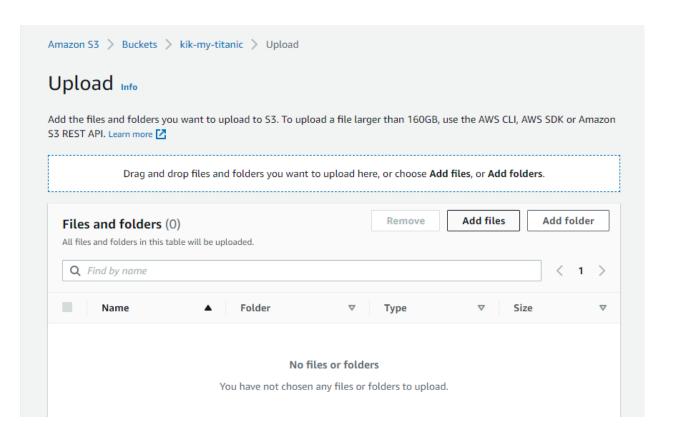
ตัวอย่างโค้ดที่ใช้ในการโหลดข้อมูลจาก S3 เข้า Redshift

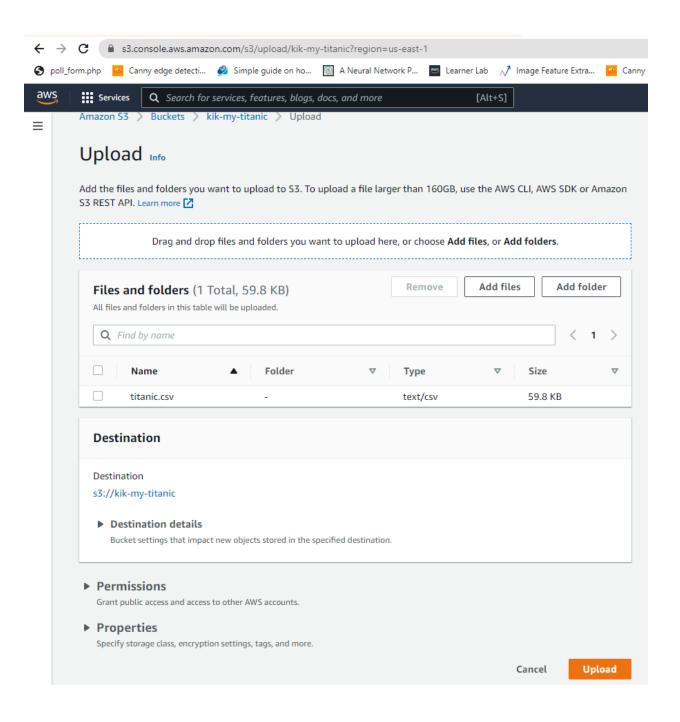
```
COPY titanic FROM 's3://zkan-swu-labs/titanic.csv'
CREDENTIALS 'aws_iam_role=arn:aws:iam::377290081649:role/LabRole'
IGNOREHEADER 1 CSV REGION 'us-east-1'
```

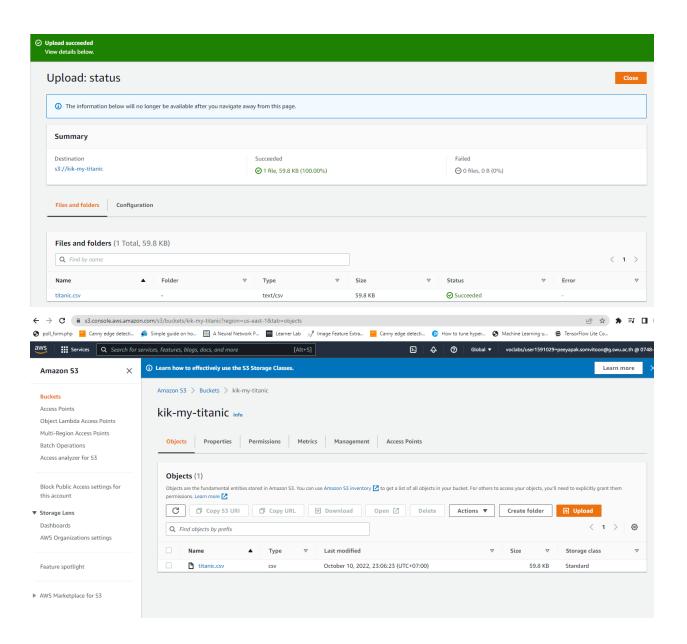
ตัวอย่างโค้ดในการฟา Transformation โดยใช้ SOL

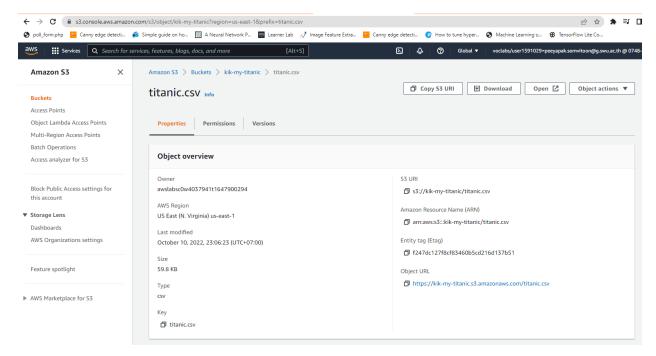
```
INSERT INTO
   new_table (
       name
  )
SELECT
   Name
FROM
   titanic
WHERE
   Name NOT IN (SELECT DISTINCT name FROM new_table)
```

Last modified: Saturday, 10 September 2022, 2:21 PM





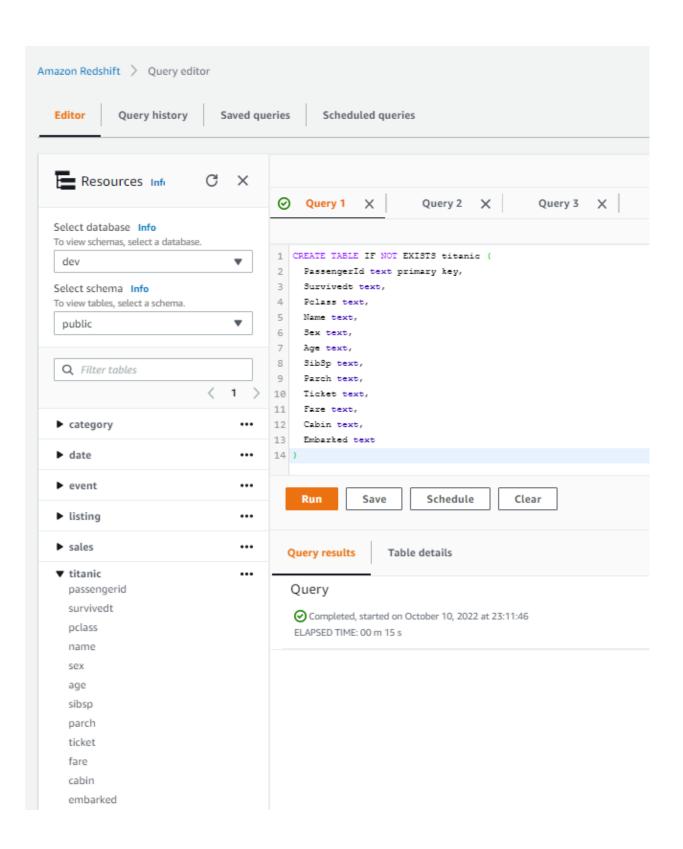


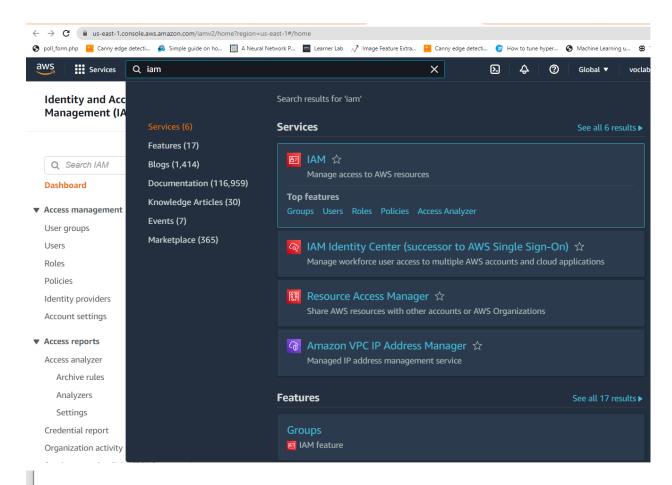


S3 URI

Object URL

https://kik-my-titanic.s3.amazonaws.com/titanic.csv





▼ Access management

User groups

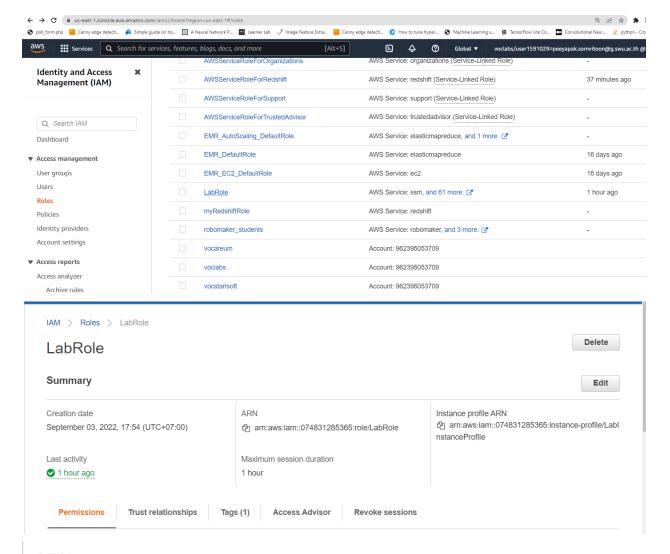
Users

Roles

Policies

Identity providers

Account settings



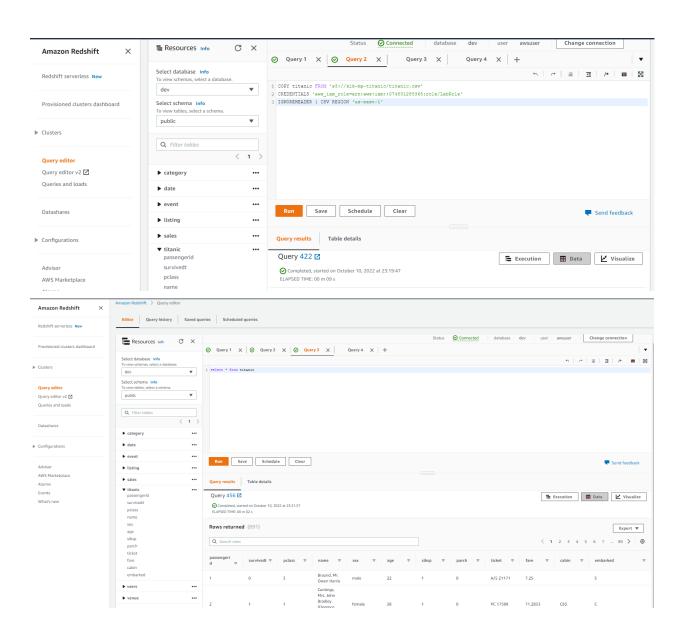
ARN

arn:aws:iam::074831285365:role/LabRole

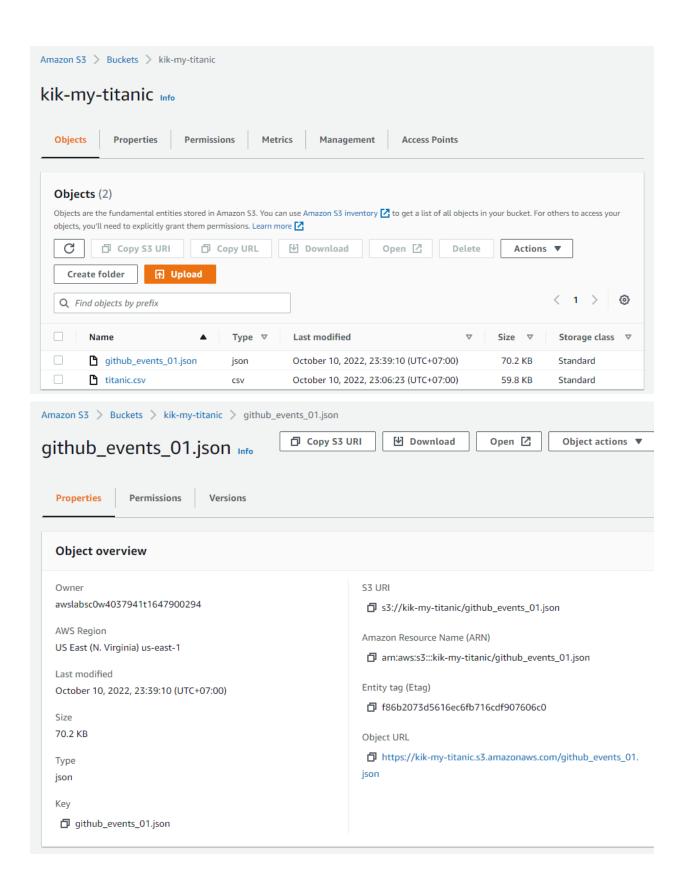
```
COPY titanic FROM 's3://kik-my-titanic/titanic.csv'

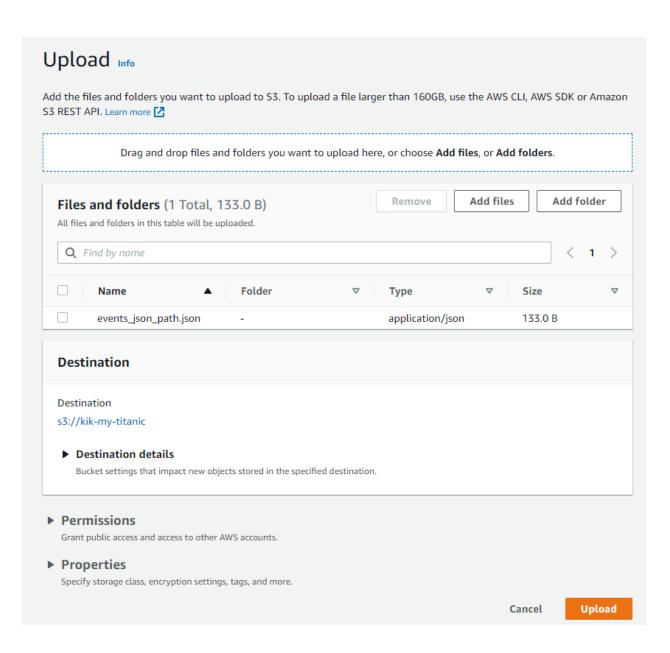
CREDENTIALS 'aws_iam_role=arn:aws:iam::074831285365:role/LabRole'

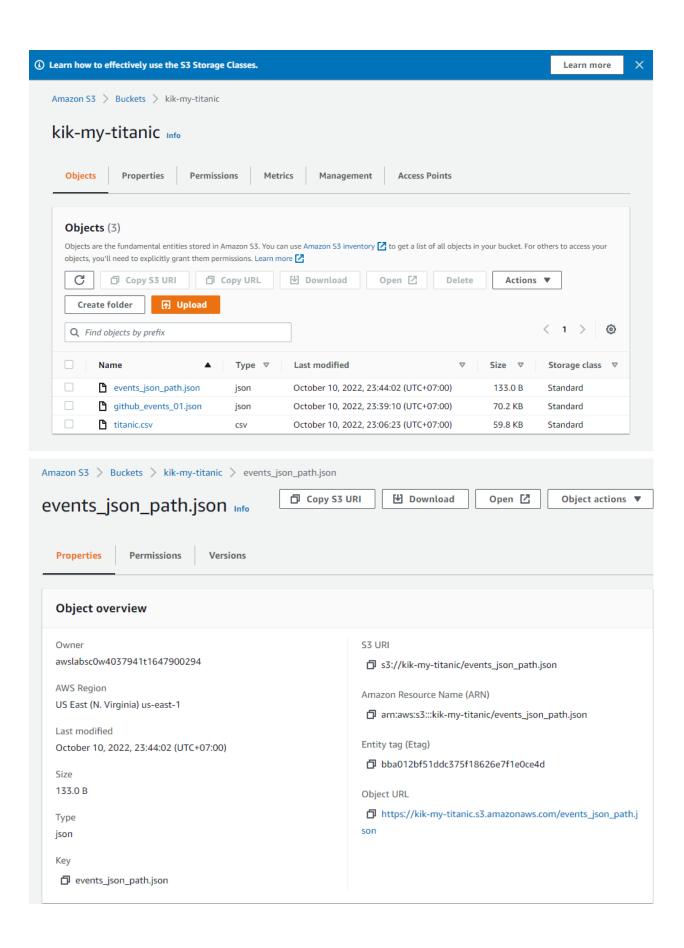
IGNOREHEADER 1 CSV REGION 'us-east-1'
```



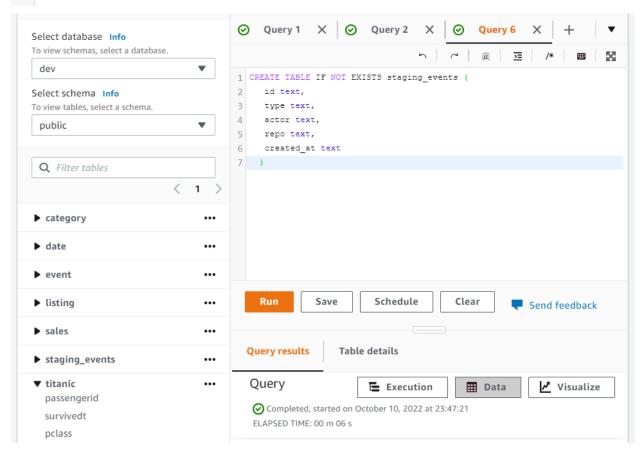
Upload Info Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more 🔀 Drag and drop files and folders you want to upload here, or choose Add files, or Add folders. Remove Add files Add folder Files and folders (1 Total, 70.2 KB) All files and folders in this table will be uploaded. Q Find by name Name Folder Size github_events_01.json application/json 70.2 KB Destination Destination s3://kik-my-titanic **▶** Destination details Bucket settings that impact new objects stored in the specified destination. Permissions Grant public access and access to other AWS accounts. ▶ Properties Specify storage class, encryption settings, tags, and more. Cancel Upload

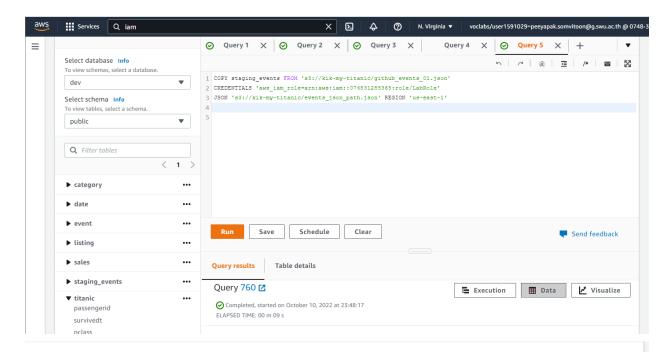






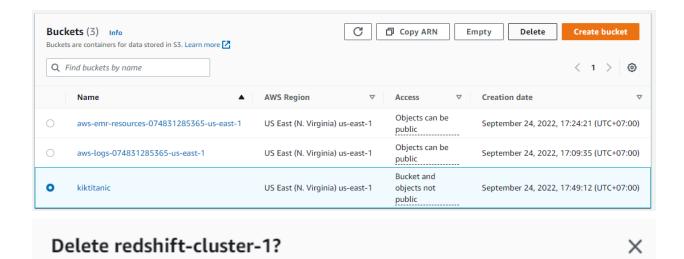
```
1 CREATE TABLE IF NOT EXISTS staging_events (
2 id text,
3 type text,
4 actor text,
5 repo text,
6 created_at text
7
```





To confirm deletion, type permanently delete in the text input field.

permanently delete



Deleting the cluster causes the following results:

- · Deletes all databases (and data) in the cluster.
- · Deletes the automated snapshot.
- Retains all manual snapshots until you manually delete them (none exist).
- · You can't rotate keys for encrypted manual snapshots if you delete this cluster.
- Removes access to the data in datashares for data consumers, including subscribers.

Are you sure that you want to permanently delete redshift-cluster-1?

Final snapshot You can create a final manual snapshot of your cluster before it's deleted so you can later restore it. Restoring it enables you to resume running the cluster and querying data. Create final snapshot	
To confirm deletion, enter <i>delete</i> in the field and choose Delete. delete	

Cancel

Delete cluster