ETL Project

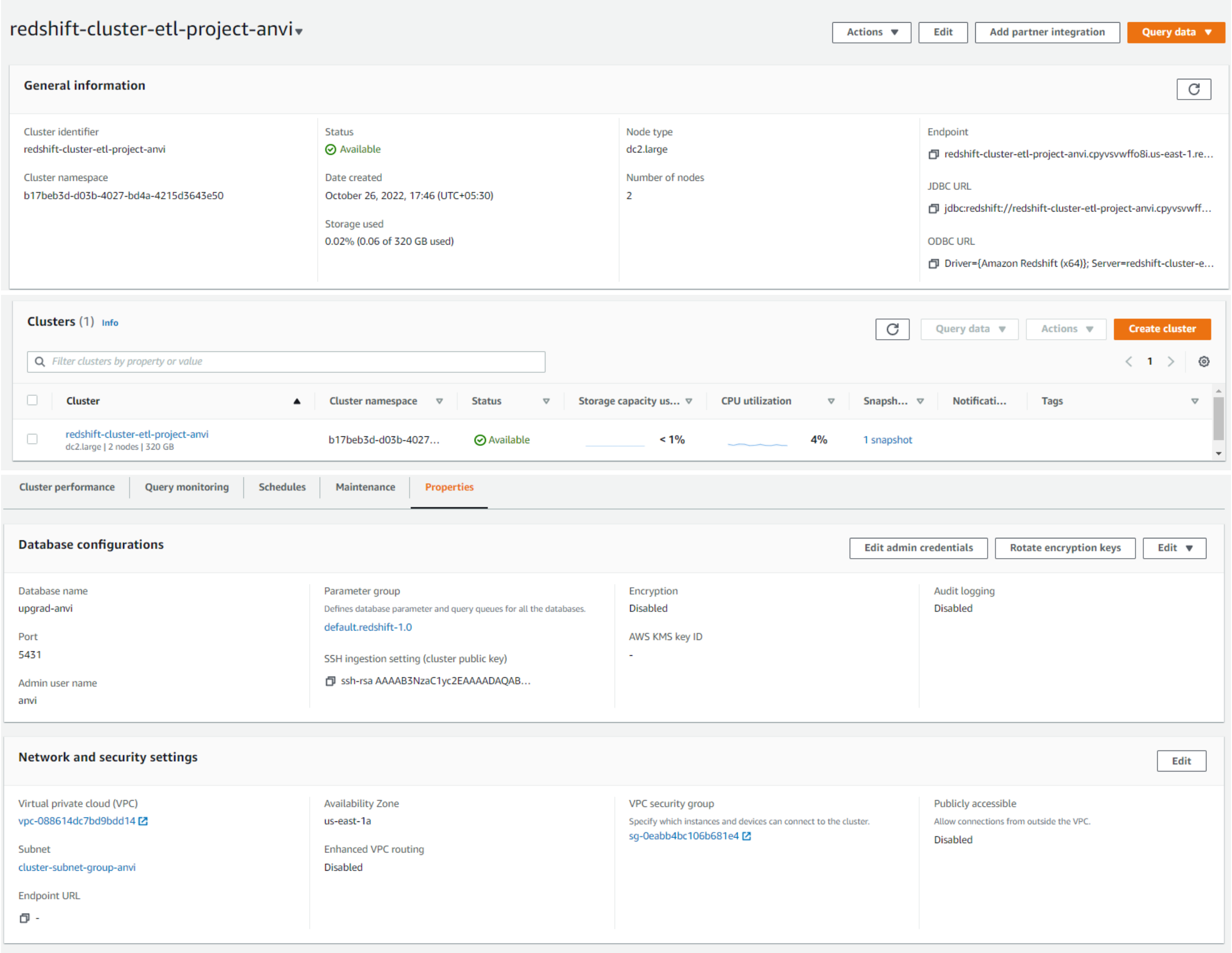
By Anviksha Singh

anviksha.singh0110s@gmail.com

Step 3: Redshift Set Up and Data Loading

Creation of a Redshift Cluster

Screenshots of the configuration of the Redshift cluster that you have created:



redshift-cluster-etl-project-anvi▾

General information

Cluster identifier
redshift-cluster-etl-project-anvi

Cluster namespace
b17beb3d-d03b-4027-bd4a-4215d3643e50

Status
✓ Available

Date created
October 26, 2022, 17:46 (UTC+05:30)

Storage used
0.02% (0.06 of 320 GB used)

Node type
dc2.large

Number of nodes
2

Endpoint
redshift-cluster-etl-project-anvi.cpyvsvwffo8i.us-east-1.re...

JDBC URL
jdbc:redshift://redshift-cluster-etl-project-anvi.cpyvsvwff...

ODBC URL
Driver={Amazon Redshift (x64)}; Server=redshift-cluster-e...

Clusters (1) Info

Filter clusters by property or value

□	Cluster	Cluster namespace	Status	Storage capacity us...	CPU utilization	Snapsh...	Notificati...	Tags
□	redshift-cluster-etl-project-anvi dc2.large | 2 nodes | 320 GB	b17beb3d-d03b-4027...	✓ Available	< 1%	4%	1 snapshot		

Cluster performance | Query monitoring | Schedules | Maintenance | Properties

Database configurations

Database name
upgrad-anvi

Port
5431

Admin user name
anvi

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

SSH ingestion setting (cluster public key)
ssh-rsa AAAAB3NzaC1yc2EAAAADAQAB...

Encryption
Disabled

AWS KMS key ID
-

Audit logging
Disabled

Network and security settings

Virtual private cloud (VPC)
vpc-088614dc7bd9bdd14

Subnet
cluster-subnet-group-anvi

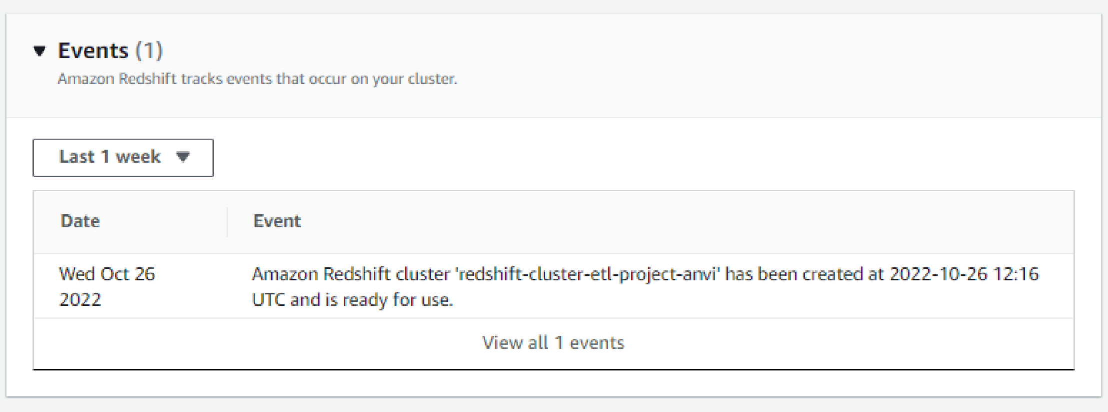
Endpoint URL
-

Availability Zone
us-east-1a

Enhanced VPC routing
Disabled

VPC security group
Specify which instances and devices can connect to the cluster.
sg-0eabb4bc106b681e4

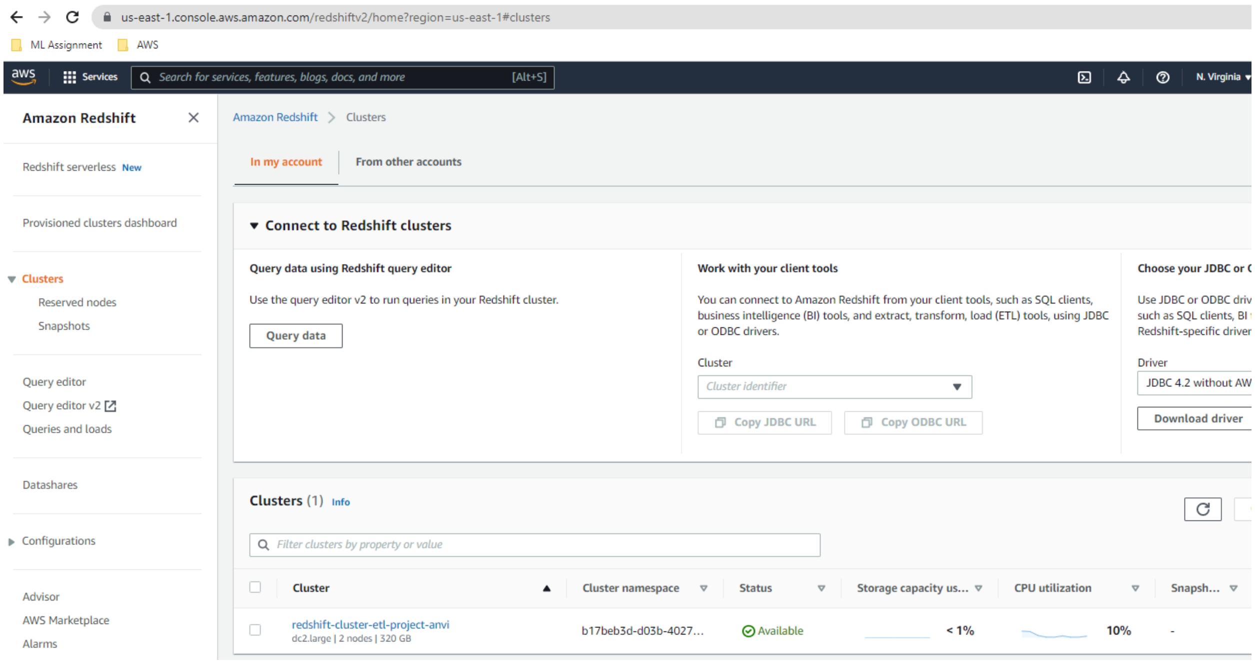
Publicly accessible
Allow connections from outside the VPC.
Disabled



▼ Events (1)
Amazon Redshift tracks events that occur on your cluster.

Date	Event
Wed Oct 26 2022	Amazon Redshift cluster 'redshift-cluster-etl-project-anvi' has been created at 2022-10-26 12:16 UTC and is ready for use.

View all 1 events



Connect to Redshift clusters

Query data using Redshift query editor

Use the query editor v2 to run queries in your Redshift cluster.

Work with your client tools

You can connect to Amazon Redshift from your client tools, such as SQL clients, business intelligence (BI) tools, and extract, transform, load (ETL) tools, using JDBC or ODBC drivers.

Cluster

Choose your JDBC or (

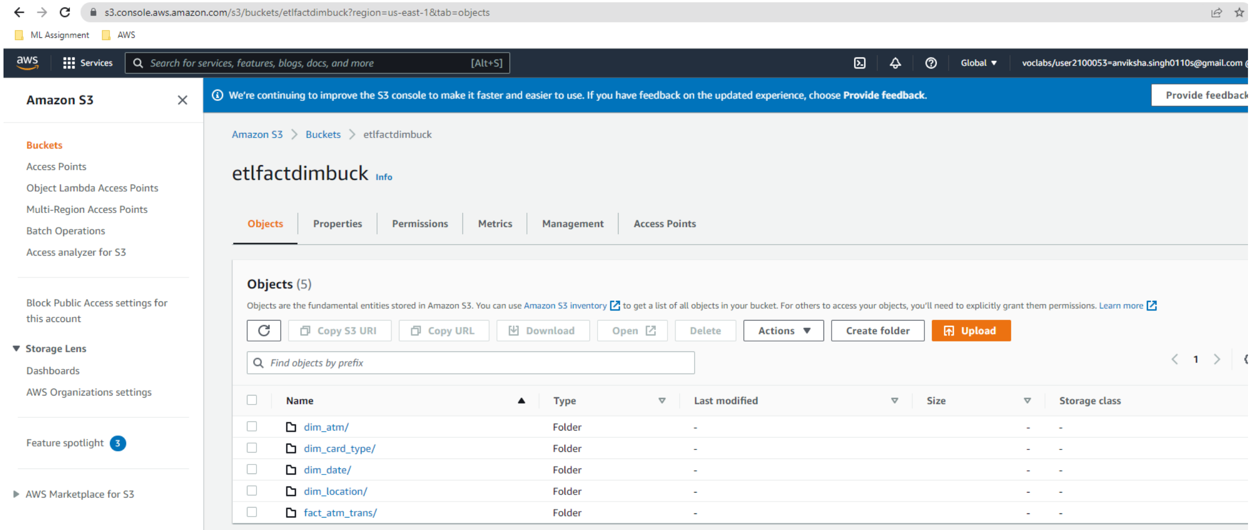
Use JDBC or ODBC driv
such as SQL clients, BI
Redshift-specific driver

Driver

Clusters (1) Info

□	Cluster	Cluster namespace	Status	Storage capacity us...	CPU utilization	Snapsh...
□	redshift-cluster-etl-project-anvi dc2.large | 2 nodes | 320 GB	b17beb3d-d03b-4027...	✓ Available	< 1%	10%	-

Screenshot of data loaded in S3


s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&tab=objects
ML Assignment
AWS
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
voclabs/user2100053=anviksha.singh0110s@gmail.com
Provide feedback
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Amazon S3 > Buckets > etlfactdimbuck
Objects
Properties
Permissions
Metrics
Management
Access Points
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

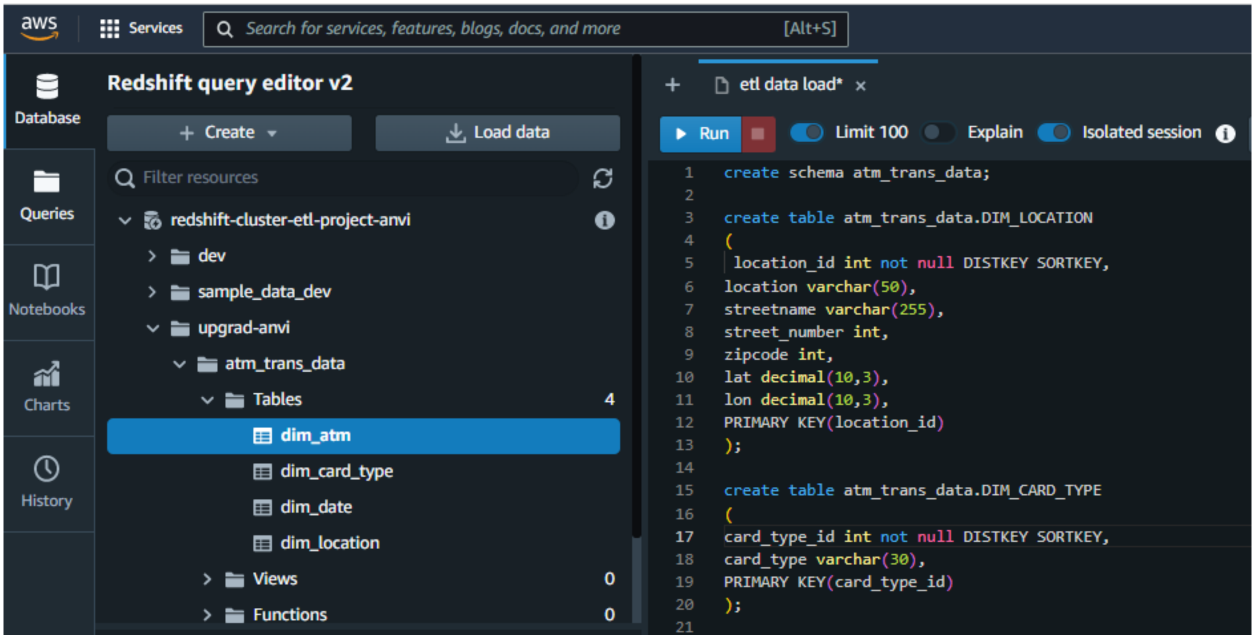
etlfactdimbuck Info

Objects (5)

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

□	Name	Type	Last modified	Size	Storage class
□	dim_atm/	Folder	-	-	-
□	dim_card_type/	Folder	-	-	-
□	dim_date/	Folder	-	-	-
□	dim_location/	Folder	-	-	-
□	fact_atm_trans/	Folder	-	-	-

Setting up a database in the Redshift cluster and running queries to create the dimension and fact tables



```sql
create schema atm_trans_data;

create table atm_trans_data.DIM_LOCATION
(
 location_id int not null DISTKEY SORTKEY,
location varchar(50),
streetname varchar(255),
street_number int,
zipcode int,
lat decimal(10,3),
lon decimal(10,3),
PRIMARY KEY(location_id)
);

create table atm_trans_data.DIM_CARD_TYPE
(
card_type_id int not null DISTKEY SORTKEY,
card_type varchar(30),
PRIMARY KEY(card_type_id)
);
```

Queries to create the various dimension and fact tables with appropriate primary and foreign keys:

Step 1 : Create Schema

```sql
create schema atm_trans_data;
```

Step 2 : Create dimension tables

DIM_LOCATION

```sql
create table atm_trans_data.DIM_LOCATION
(
 location_id int not null DISTKEY SORTKEY,
location varchar(50),
streetname varchar(255),
street_number int,
zipcode int,
lat decimal(10,3),
lon decimal(10,3),
PRIMARY KEY(location_id)
);
```

DIM_CARD TYPE

```sql
create table atm_trans_data.DIM_CARD_TYPE
(
card_type_id int not null DISTKEY SORTKEY,
card_type varchar(30),
PRIMARY KEY(card_type_id)
);
```

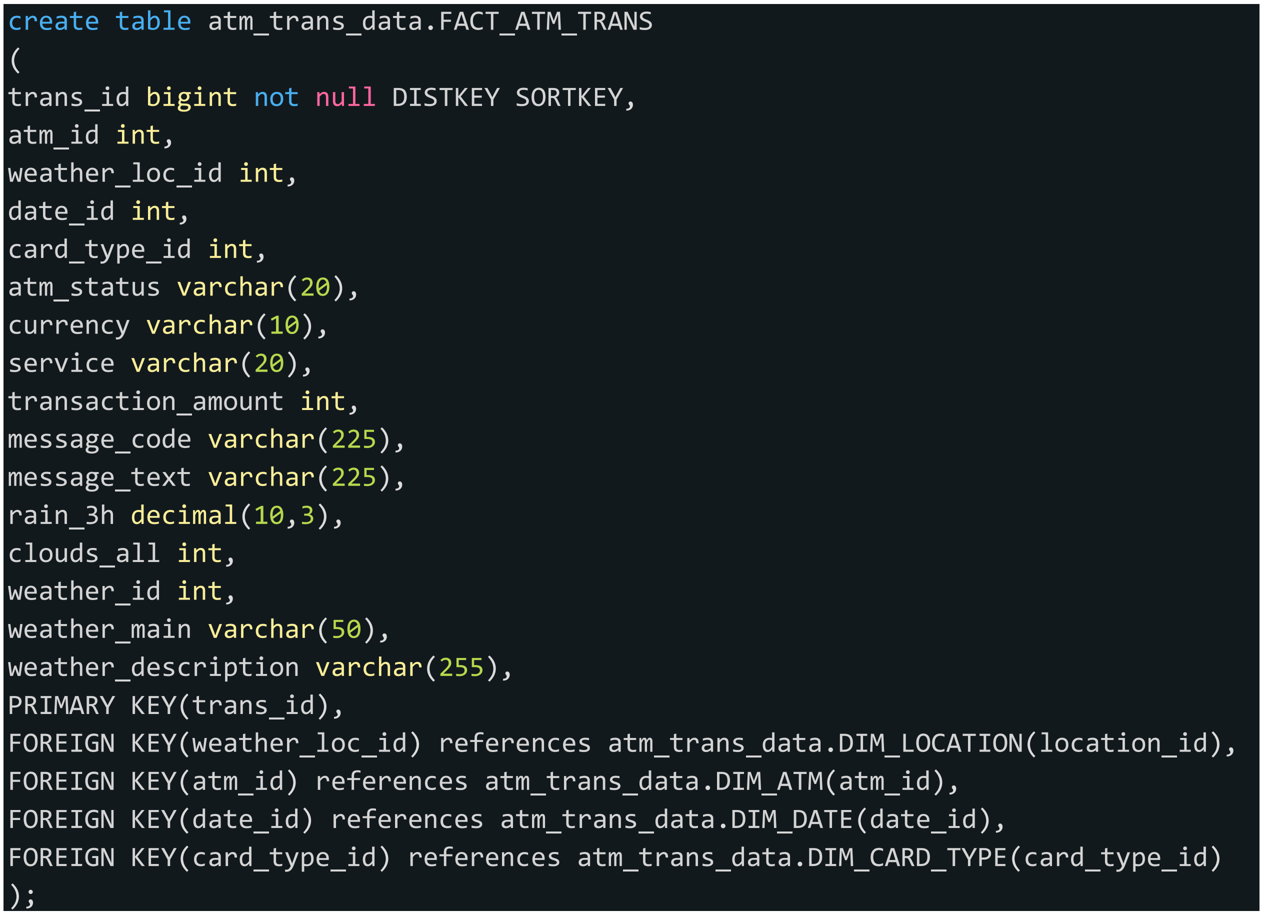
DIM_ATM

```sql
create table atm_trans_data.DIM_ATM
(
atm_id int not null DISTKEY SORTKEY,
atm_number varchar(20),
atm_manufacturer varchar(50),
atm_location_id int,
PRIMARY KEY(atm_id),
FOREIGN KEY(atm_location_id) references atm_trans_data.DIM_LOCATION(location_id)
);
```

DIM_DATE

```sql
create table atm_trans_data.DIM_DATE
(
date_id int not null DISTKEY SORTKEY,
full_date_time timestamp,
year int,
month varchar(20),
day int,
hour int,
weekday varchar(20),
PRIMARY KEY(date_id)
);
```

Step 3 : Create Fact table

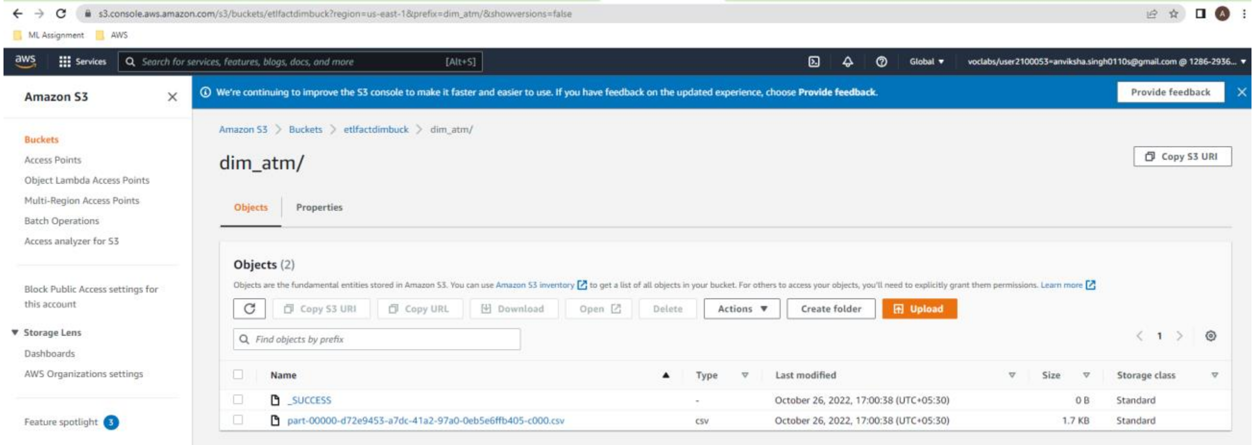


```
create table atm_trans_data.FACT_ATM_TRANS
(
trans_id bigint not null DISTKEY SORTKEY,
atm_id int,
weather_loc_id int,
date_id int,
card_type_id int,
atm_status varchar(20),
currency varchar(10),
service varchar(20),
transaction_amount int,
message_code varchar(225),
message_text varchar(225),
rain_3h decimal(10,3),
clouds_all int,
weather_id int,
weather_main varchar(50),
weather_description varchar(255),
PRIMARY KEY(trans_id),
FOREIGN KEY(weather_loc_id) references atm_trans_data.DIM_LOCATION(location_id),
FOREIGN KEY(atm_id) references atm_trans_data.DIM_ATM(atm_id),
FOREIGN KEY(date_id) references atm_trans_data.DIM_DATE(date_id),
FOREIGN KEY(card_type_id) references atm_trans_data.DIM_CARD_TYPE(card_type_id)
);
```

Loading data into a Redshift cluster from Amazon S3 bucket

Queries to copy the dimension dim_atm from S3 buckets to the Redshift cluster

A. S3 Bucket – Dim_atm



s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&prefix=dim_atm/&showversions=false

ML Assignment
AWS
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global ▾
voclabs/user2100053=anviksha.singh0110s@gmail.com @ 1286-2936... ▾
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
▾ Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Provide feedback
Amazon S3 > Buckets > etlfactdimbuck > dim_atm/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions ▾
Create folder
Upload
Find objects by prefix
< 1 >

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

dim_atm/

Objects (2)

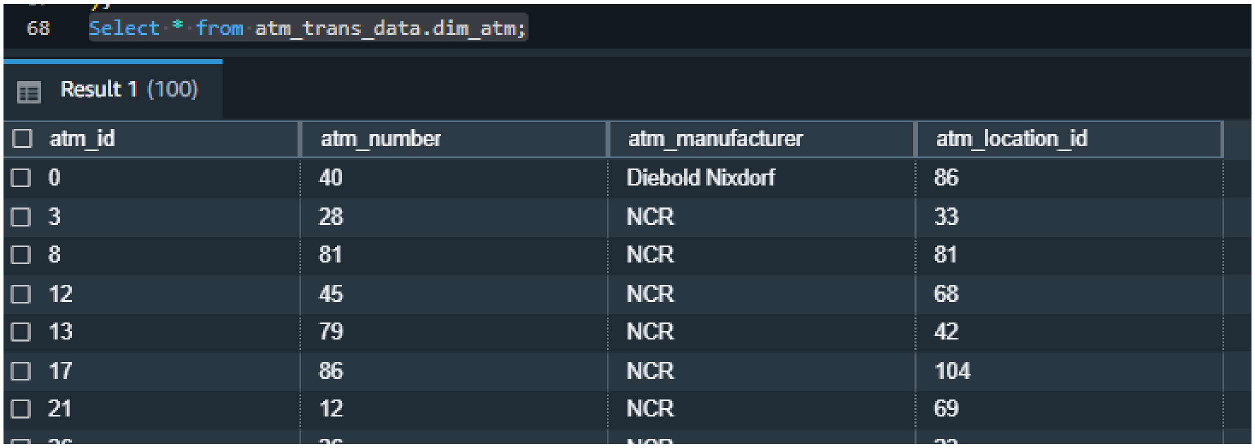
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	_SUCCESS	-	October 26, 2022, 17:00:38 (UTC+05:30)	0 B	Standard
□	part-00000-d72e9453-a7dc-41a2-97a0-0eb5e6ffb405-c000.csv	csv	October 26, 2022, 17:00:38 (UTC+05:30)	1.7 KB	Standard

B. Query to copy the dim_atm from S3 buckets to the Redshift cluster

```
COPY "upgrad-anvi".atm_trans_data.dim_atm FROM
's3://etlfactdimbuck/dim_atm/part-00000-d72e9453-a7dc-41a2-97a0-
0eb5e6ffb405-c000.csv' IAM_ROLE
'arn:aws:iam::128629367978:role/myRedshiftRole' FORMAT AS CSV DELIMITER
',' QUOTE '"' REGION AS 'us-east-1'
```

C. Loaded Data



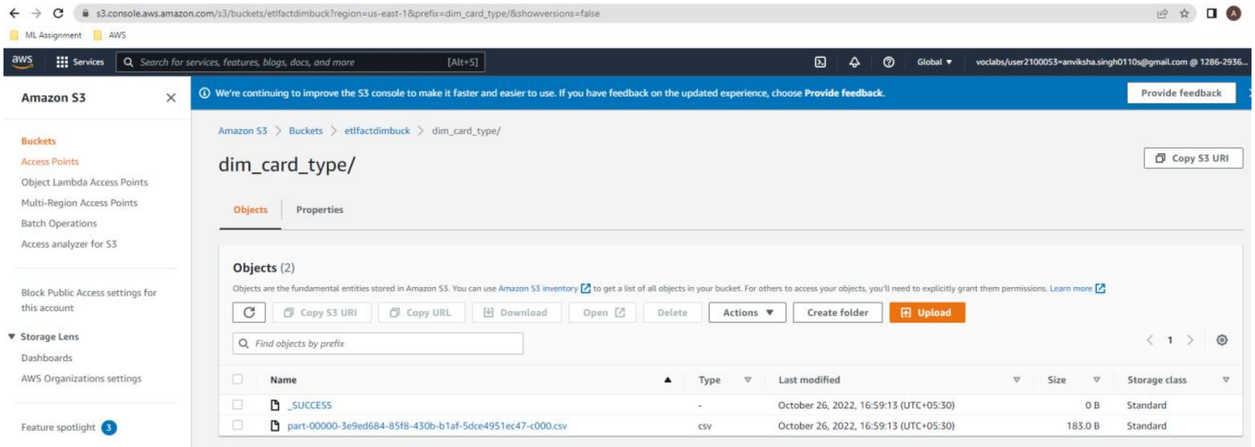
```
68   Select * from atm_trans_data.dim_atm;
```

Result 1 (100)

□ atm_id	atm_number	atm_manufacturer	atm_location_id
□ 0	40	Diebold Nixdorf	86
□ 3	28	NCR	33
□ 8	81	NCR	81
□ 12	45	NCR	68
□ 13	79	NCR	42
□ 17	86	NCR	104
□ 21	12	NCR	69

Queries to copy the dimension dim_card_type from S3 buckets to the Redshift cluster

A. S3 Bucket – Dim_card_type


s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&prefix=dim_card_type/&showversions=false
ML Assignment
AWS
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
voclabs/user2100053=anviksha.singh0110s@gmail.com @ 1286-2936...
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Provide feedback
Amazon S3 > Buckets > etlfactdimbuck > dim_card_type/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

dim_card_type/

Objects (2)

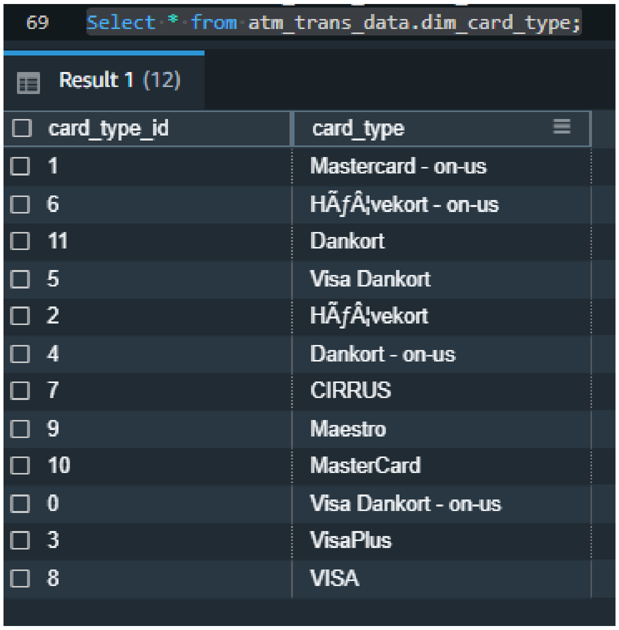
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	_SUCCESS	-	October 26, 2022, 16:59:13 (UTC+05:30)	0 B	Standard
□	part-00000-3e9ed684-85f8-430b-b1af-5dce4951ec47-c000.csv	csv	October 26, 2022, 16:59:13 (UTC+05:30)	183.0 B	Standard

B. Query to copy the dim_card_type from S3 buckets to the Redshift cluster

```
COPY "upgrad-anvi".atm_trans_data.dim_card_type FROM
's3://etlfactdimbuck/dim_card_type/part-00000-3e9ed684-85f8-430b-b1af-
5dce4951ec47-c000.csv' IAM_ROLE
'arn:aws:iam::128629367978:role/myRedshiftRole' FORMAT AS CSV DELIMITER
',' QUOTE '"' REGION AS 'us-east-1'
```

C. Loaded Data

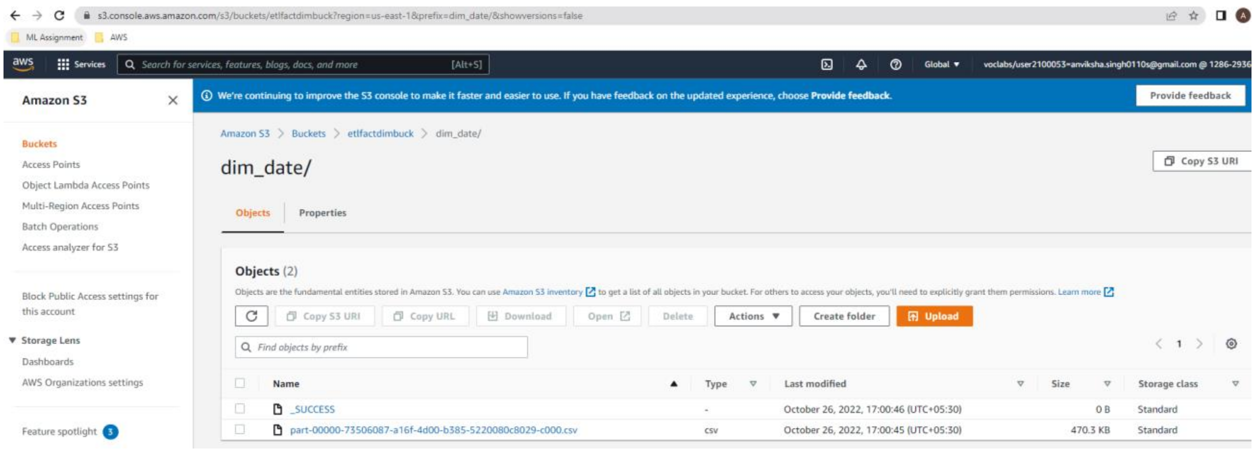


```
69    Select * from atm_trans_data.dim_card_type;
```

card_type_id	card_type
1	Mastercard - on-us
6	HÃƒÂ¦vekort - on-us
11	Dankort
5	Visa Dankort
2	HÃƒÂ¦vekort
4	Dankort - on-us
7	CIRRUS
9	Maestro
10	MasterCard
0	Visa Dankort - on-us
3	VisaPlus
8	VISA

Queries to copy the dimension dim_date from S3 buckets to the Redshift cluster

A. S3 Bucket – Dim_date


s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&prefix=dim_date/&showversions=false
ML Assignment
AWS
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global ▾
voclabs/user2100053=anviksha.singh0110s@gmail.com @ 1286-2936
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
▼ Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Provide feedback
Amazon S3 > Buckets > etlfactdimbuck > dim_date/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions ▾
Create folder
Upload
Find objects by prefix
< 1 >

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

dim_date/

Objects (2)

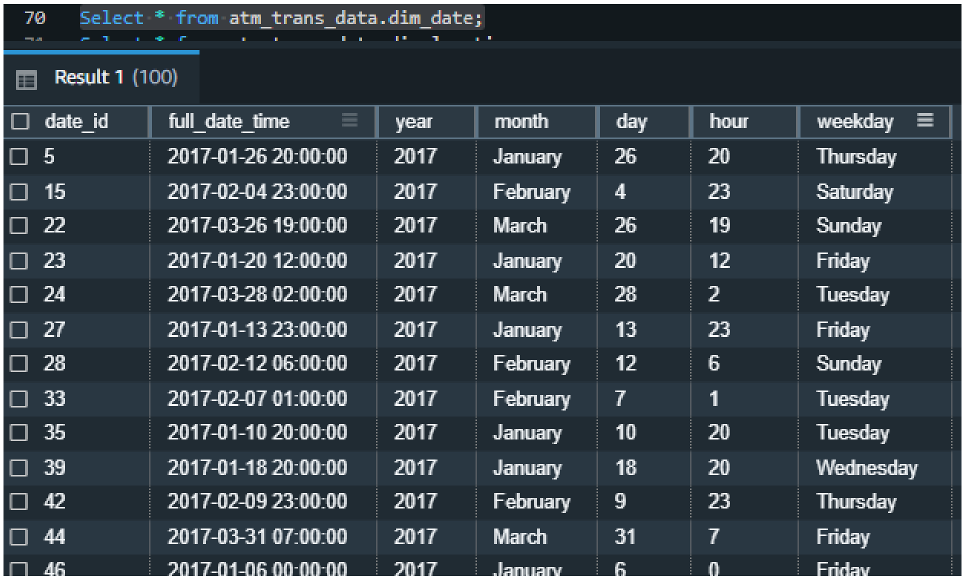
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	_SUCCESS	-	October 26, 2022, 17:00:46 (UTC+05:30)	0 B	Standard
□	part-00000-73506087-a16f-4d00-b385-5220080c8029-c000.csv	csv	October 26, 2022, 17:00:45 (UTC+05:30)	470.3 KB	Standard

B. Query to copy the dim_date from S3 buckets to the Redshift cluster

```
COPY "upgrad-anvi".atm_trans_data.dim_date FROM
's3://etlfactdimbuck/dim_date/part-00000-73506087-a16f-4d00-b385-
5220080c8029-c000.csv' IAM_ROLE
'arn:aws:iam::128629367978:role/myRedshiftRole' FORMAT AS CSV DELIMITER
',' QUOTE '"' TIMEFORMAT 'auto' REGION AS 'us-east-1'
```

C. Loaded Data

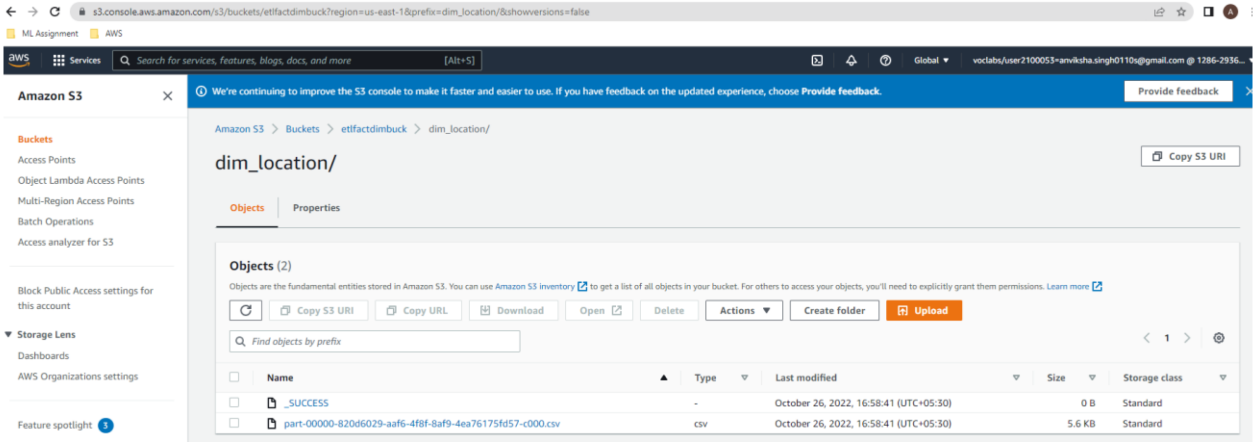


```
Select * from atm_trans_data.dim_date;
```

□	date_id	full_date_time	year	month	day	hour	weekday
□	5	2017-01-26 20:00:00	2017	January	26	20	Thursday
□	15	2017-02-04 23:00:00	2017	February	4	23	Saturday
□	22	2017-03-26 19:00:00	2017	March	26	19	Sunday
□	23	2017-01-20 12:00:00	2017	January	20	12	Friday
□	24	2017-03-28 02:00:00	2017	March	28	2	Tuesday
□	27	2017-01-13 23:00:00	2017	January	13	23	Friday
□	28	2017-02-12 06:00:00	2017	February	12	6	Sunday
□	33	2017-02-07 01:00:00	2017	February	7	1	Tuesday
□	35	2017-01-10 20:00:00	2017	January	10	20	Tuesday
□	39	2017-01-18 20:00:00	2017	January	18	20	Wednesday
□	42	2017-02-09 23:00:00	2017	February	9	23	Thursday
□	44	2017-03-31 07:00:00	2017	March	31	7	Friday
□	46	2017-01-06 00:00:00	2017	January	6	0	Friday

Queries to copy the dimension dim_location from S3 buckets to the Redshift cluster

A. S3 Bucket – Dim_location



s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&prefix=dim_location/&showversions=false

ⓘ We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Amazon S3 > Buckets > etlfactdimbuck > dim_location/

dim_location/

Objects (2)

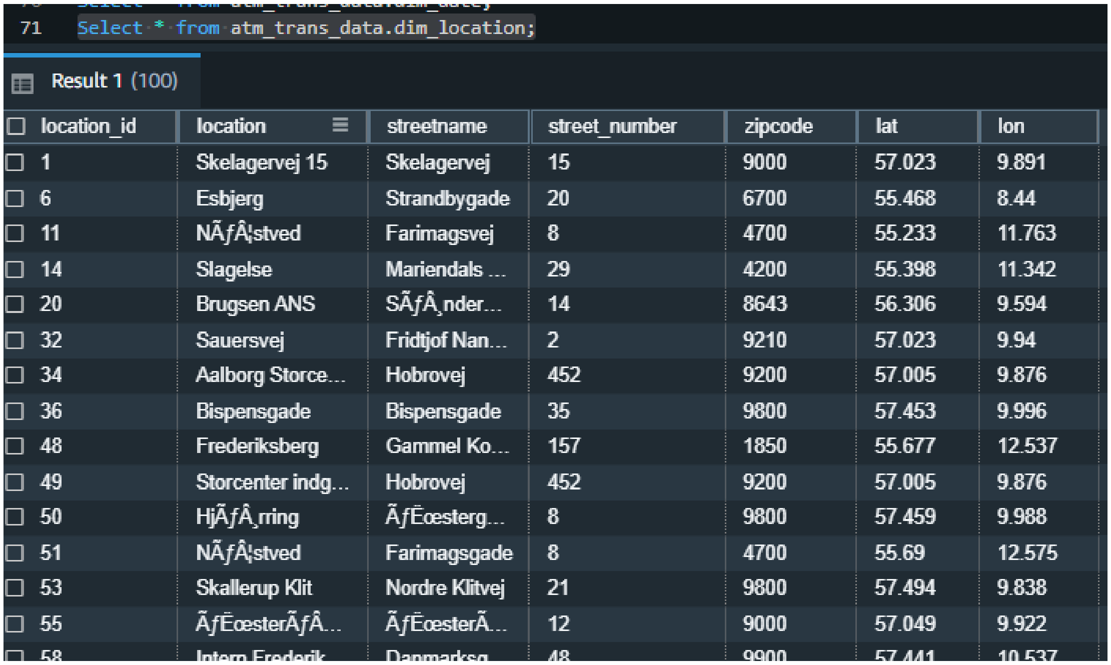
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

☐	Name	Type	Last modified	Size	Storage class
☐	_SUCCESS	-	October 26, 2022, 16:58:41 (UTC+05:30)	0 B	Standard
☐	part-00000-820d6029-aaf6-4f8f-8af9-4ea76175fd57-c000.csv	csv	October 26, 2022, 16:58:41 (UTC+05:30)	5.6 KB	Standard

B. Query to copy the dim_location from S3 buckets to the Redshift cluster

```
COPY "upgrad-anvi".atm_trans_data.dim_location FROM
's3://etlfactdimbuck/dim_location/part-00000-820d6029-aaf6-4f8f-8af9-
4ea76175fd57-c000.csv' IAM_ROLE
'arn:aws:iam::128629367978:role/myRedshiftRole' FORMAT AS CSV DELIMITER
',' QUOTE '"' REGION AS 'us-east-1'
```

C. Loaded Data



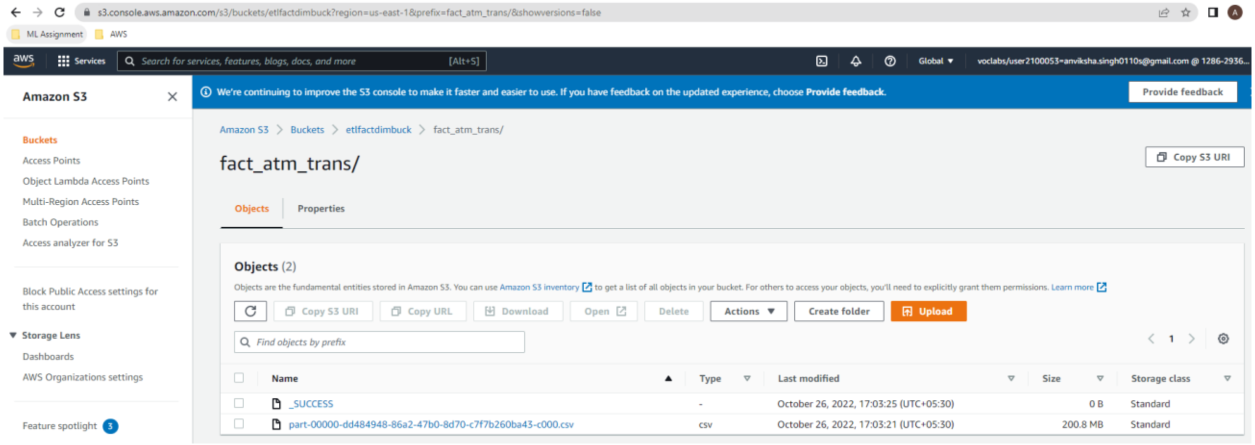
```
Select * from atm_trans_data.dim_date;
Select * from atm_trans_data.dim_location;
```

Result 1 (100)

☐ location_id	location	streetname	street_number	zipcode	lat	lon
☐ 1	Skelagervej 15	Skelagervej	15	9000	57.023	9.891
☐ 6	Esbjerg	Strandbygade	20	6700	55.468	8.44
☐ 11	NÃƒÂ¦stved	Farimagsvej	8	4700	55.233	11.763
☐ 14	Slagelse	Mariendals ...	29	4200	55.398	11.342
☐ 20	Brugsen ANS	SÃƒÂ¸nder...	14	8643	56.306	9.594
☐ 32	Sauersvej	Fridtjof Nan...	2	9210	57.023	9.94
☐ 34	Aalborg Storce...	Hobrovej	452	9200	57.005	9.876
☐ 36	Bispensgade	Bispensgade	35	9800	57.453	9.996
☐ 48	Frederiksberg	Gammel Ko...	157	1850	55.677	12.537
☐ 49	Storcenter indg...	Hobrovej	452	9200	57.005	9.876
☐ 50	HjÃƒÂ¸rring	ÃƒËœsterg...	8	9800	57.459	9.988
☐ 51	NÃƒÂ¦stved	Farimagsgade	8	4700	55.69	12.575
☐ 53	Skallerup Klit	Nordre Klitvej	21	9800	57.494	9.838
☐ 55	ÃƒËœsterÃƒÂ...	ÃƒËœsterÃ...	12	9000	57.049	9.922
☐ 58	Intern Frederik	Danmarksg...	48	9900	57.441	10.537

Queries to copy the fact from S3 buckets to the Redshift cluster

A. S3 Bucket – fact



s3.console.aws.amazon.com/s3/buckets/etlfactdimbuck?region=us-east-1&prefix=fact_atm_trans/&showversions=false

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Amazon S3 > Buckets > etlfactdimbuck > fact_atm_trans/

fact_atm_trans/

Objects (2)

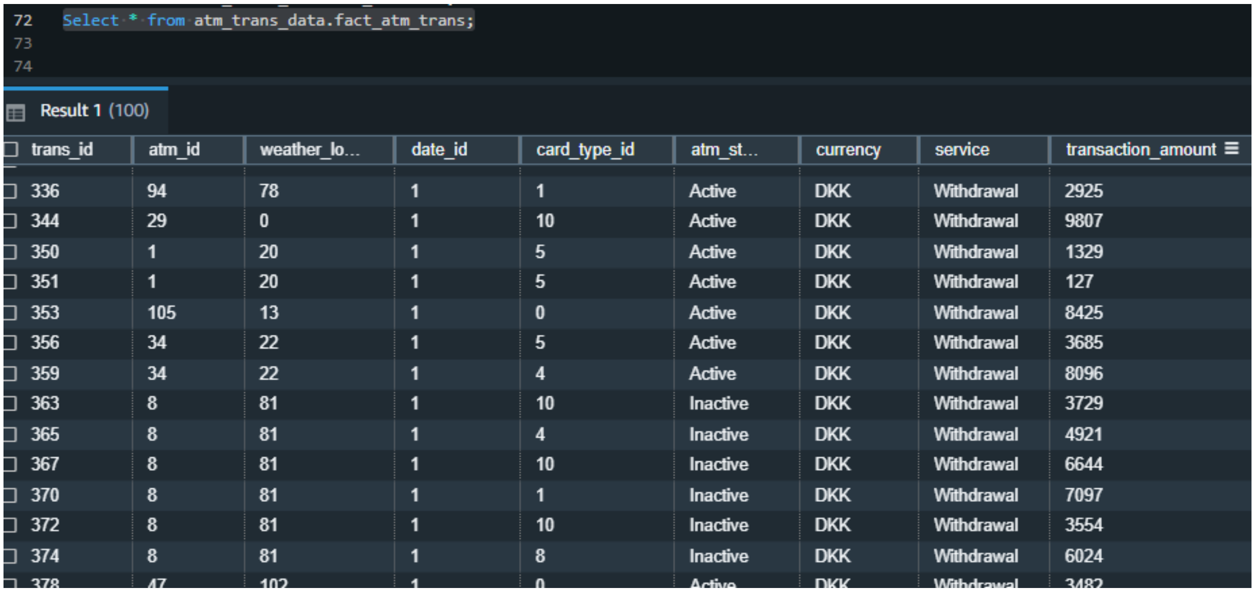
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	_SUCCESS	-	October 26, 2022, 17:03:25 (UTC+05:30)	0 B	Standard
□	part-00000-dd484948-86a2-47b0-8d70-c7f7b260ba43-c000.csv	csv	October 26, 2022, 17:03:21 (UTC+05:30)	200.8 MB	Standard

B. Query to copy the fact from S3 buckets to the Redshift cluster

```
COPY "upgrad-anvi".atm_trans_data.fact_atm_trans FROM
's3://etlfactdimbuck/fact_atm_trans/part-00000-dd484948-86a2-47b0-8d70-
c7f7b260ba43-c000.csv' IAM_ROLE
'arn:aws:iam::128629367978:role/myRedshiftRole' FORMAT AS CSV DELIMITER
',' QUOTE '"' REGION AS 'us-east-1'
```

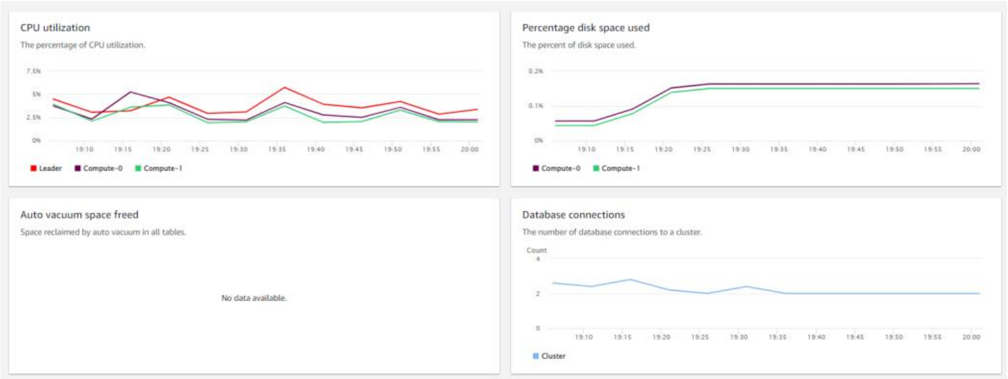
C. Loaded Data

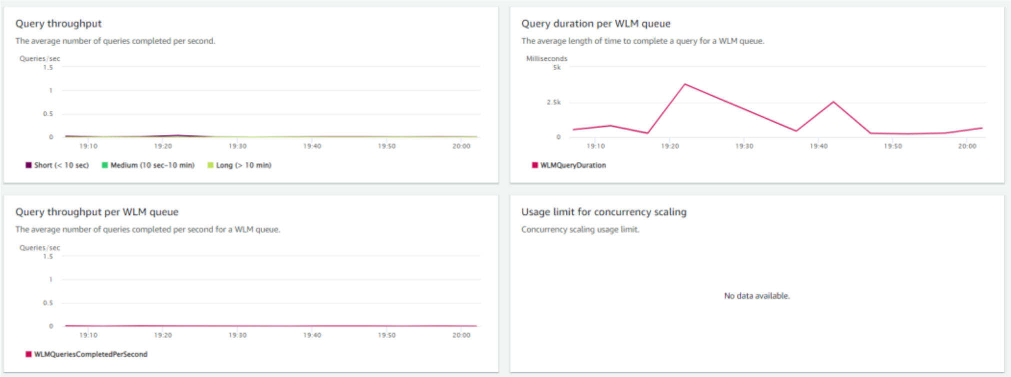


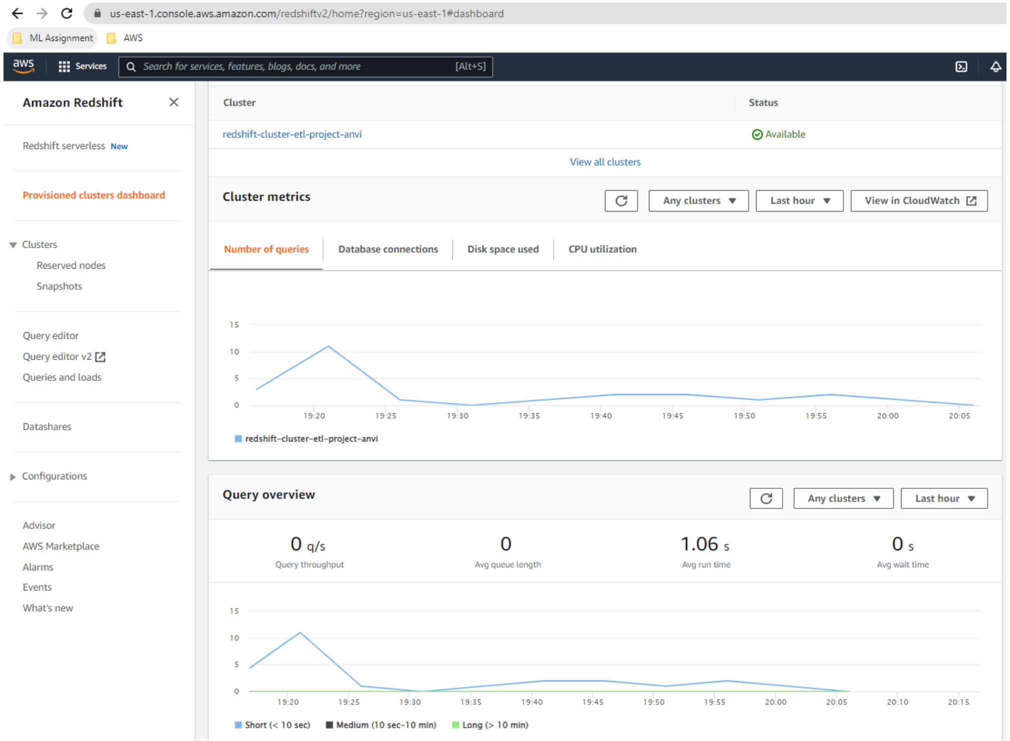
```
Select * from atm_trans_data.fact_atm_trans;
```

trans_id	atm_id	weather_lo...	date_id	card_type_id	atm_st...	currency	service	transaction_amount
336	94	78	1	1	Active	DKK	Withdrawal	2925
344	29	0	1	10	Active	DKK	Withdrawal	9807
350	1	20	1	5	Active	DKK	Withdrawal	1329
351	1	20	1	5	Active	DKK	Withdrawal	127
353	105	13	1	0	Active	DKK	Withdrawal	8425
356	34	22	1	5	Active	DKK	Withdrawal	3685
359	34	22	1	4	Active	DKK	Withdrawal	8096
363	8	81	1	10	Inactive	DKK	Withdrawal	3729
365	8	81	1	4	Inactive	DKK	Withdrawal	4921
367	8	81	1	10	Inactive	DKK	Withdrawal	6644
370	8	81	1	1	Inactive	DKK	Withdrawal	7097
372	8	81	1	10	Inactive	DKK	Withdrawal	3554
374	8	81	1	8	Inactive	DKK	Withdrawal	6024
378	47	102	1	0	Active	DKK	Withdrawal	3482

Performance after all the query execution:




Query throughput
The average number of queries completed per second.
Queries/sec
1.5
1
0.5
0
19:10
19:20
19:30
19:40
19:50
20:00
Short (< 10 sec)
Medium (10 sec-10 min)
Long (> 10 min)
Query duration per WLM queue
The average length of time to complete a query for a WLM queue.
Milliseconds
5k
2.5k
0
WLMQueryDuration
Query throughput per WLM queue
The average number of queries completed per second for a WLM queue.
WLMQueriesCompletedPerSecond
Usage limit for concurrency scaling
Concurrency scaling usage limit.
No data available.


us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#dashboard

Cluster	Status
redshift-cluster-etl-project-anvi	Available

0 q/s	0	1.06 s	0 s
Query throughput	Avg queue length	Avg run time	Avg wait time