E-commerce - Automate ETL to sync data for batch processing

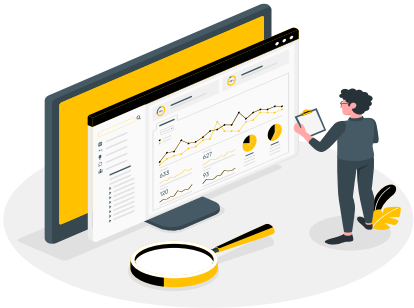


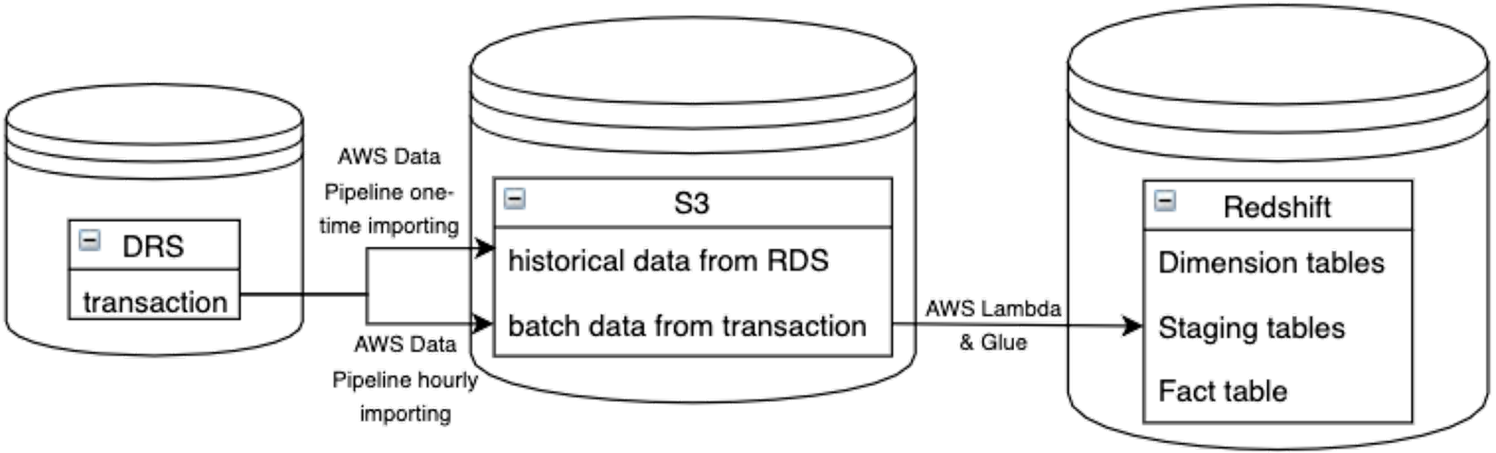
Table of Contents

01 Project Background & Subjective

- Data comes from Kaggle
- E-commerce company X wants to migrate their historical data from traditional database RDS to Data Warehouse
- And then build automate data pipeline to do batch processing hourly to sync data for last 30 days (assuming orders' status will not change after 30 days)

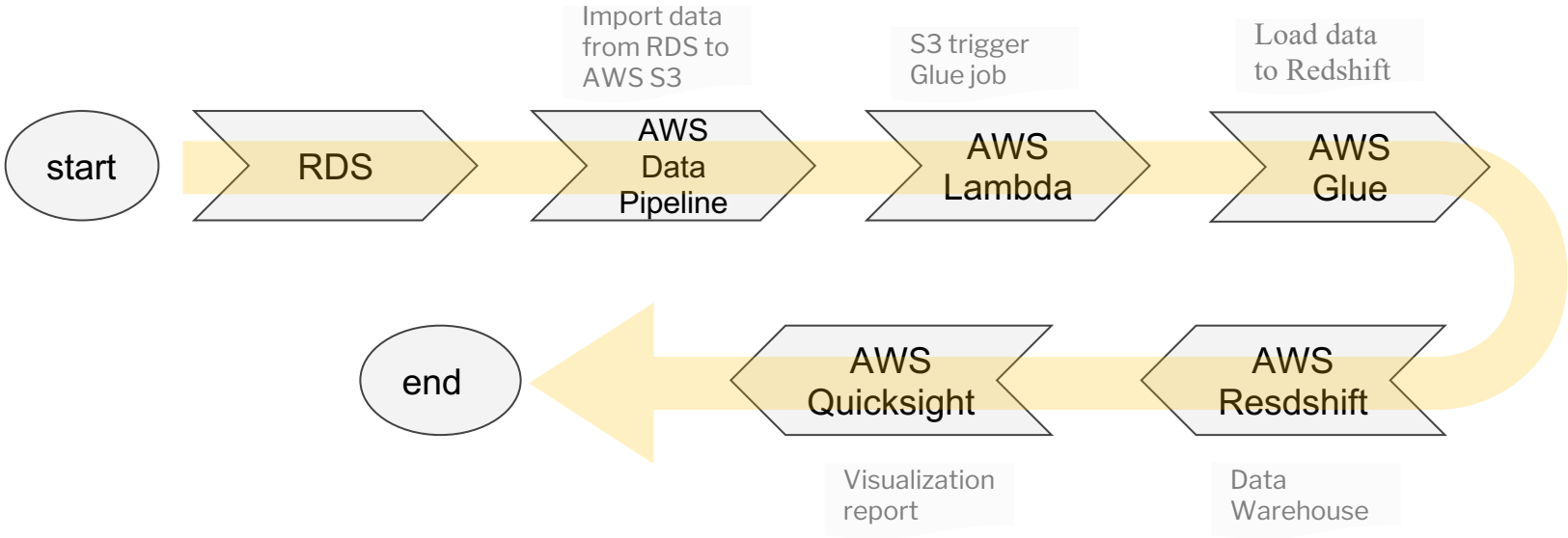
E-commerce company X

02 Data Warehouse Design

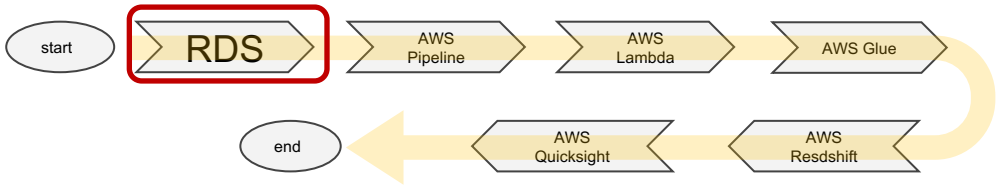


03 Auto ETL Process

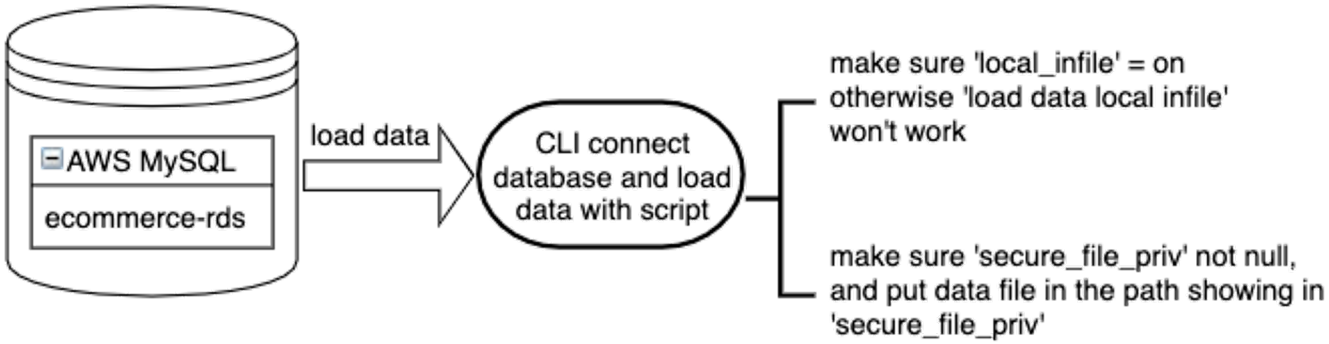
Project Process Flow Chart



03 Auto ETL Process

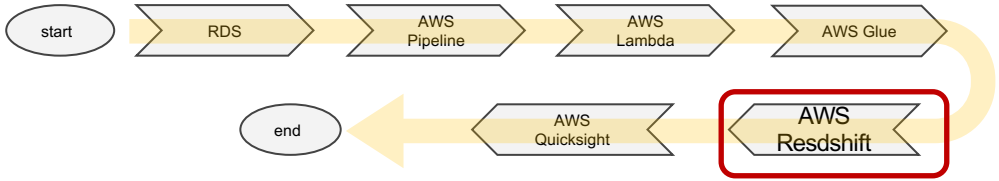


Since we are using Kaggle data to mimic real world job, we need to prepare data in Relational Database first.

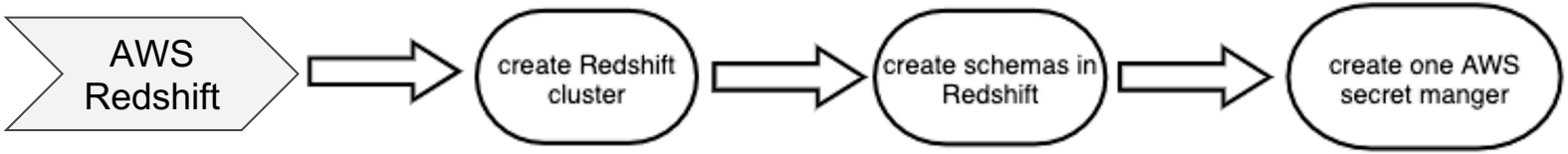


script

03 Auto ETL Process

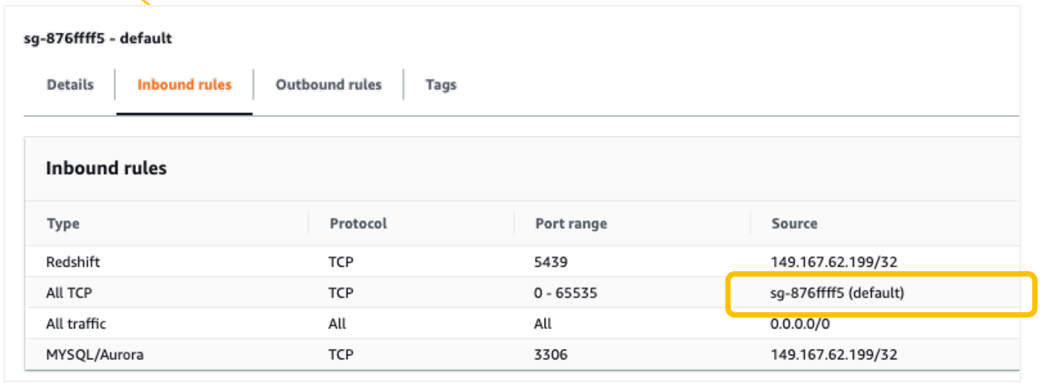


Firstly, schemas were prepared in Data Warehouse Redshift



create Redshift cluster

- DB name, username, password
- Modify 'network &security' to 'public access'
- Role for accessing Redshift, S3, Glue
- Choose default VPC but change its Security Group inbound and outbound rule



sg-876ffff5 - default

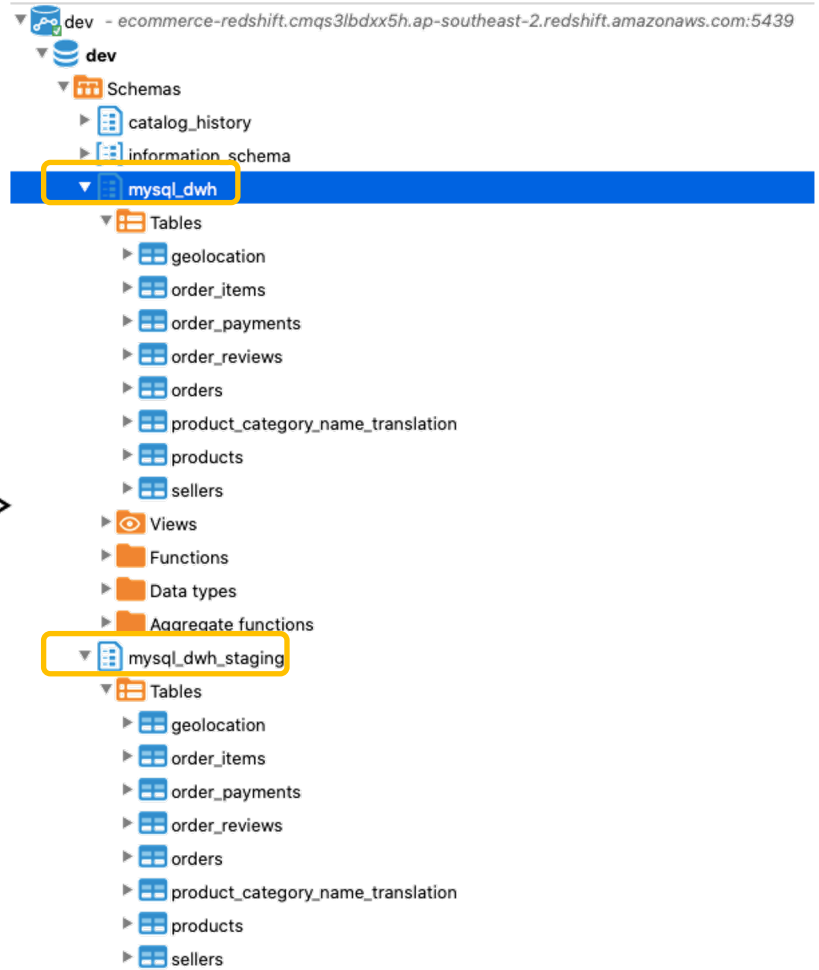
Details | Inbound rules | Outbound rules | Tags

Inbound rules

Type	Protocol	Port range	Source
Redshift	TCP	5439	149.167.62.199/32
All TCP	TCP	0 - 65535	sg-876ffff5 (default)
All traffic	All	All	0.0.0.0/0
MYSQL/Aurora	TCP	3306	149.167.62.199/32

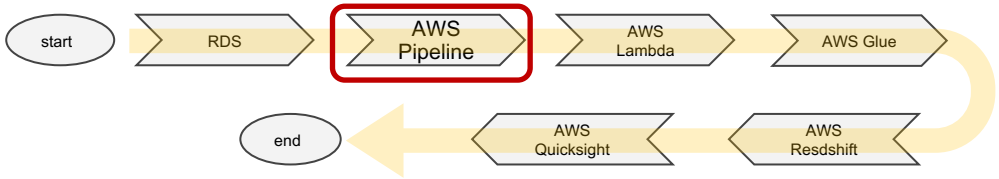
create schemas in Redshift

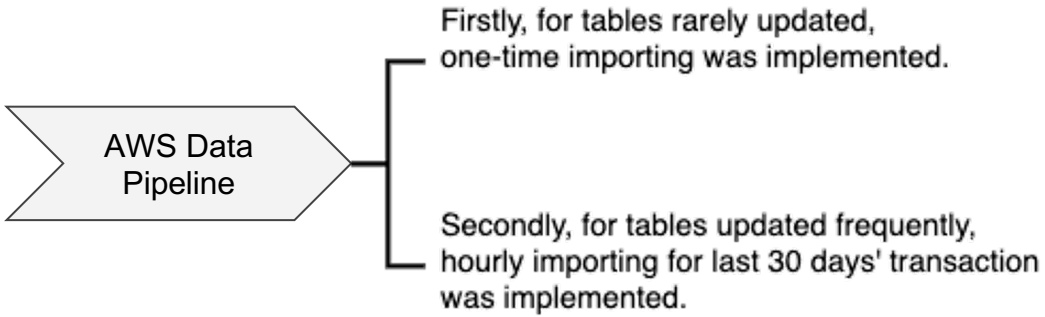
- Connect to cluster using command line
- source script to create schema



script

03 Auto ETL Process





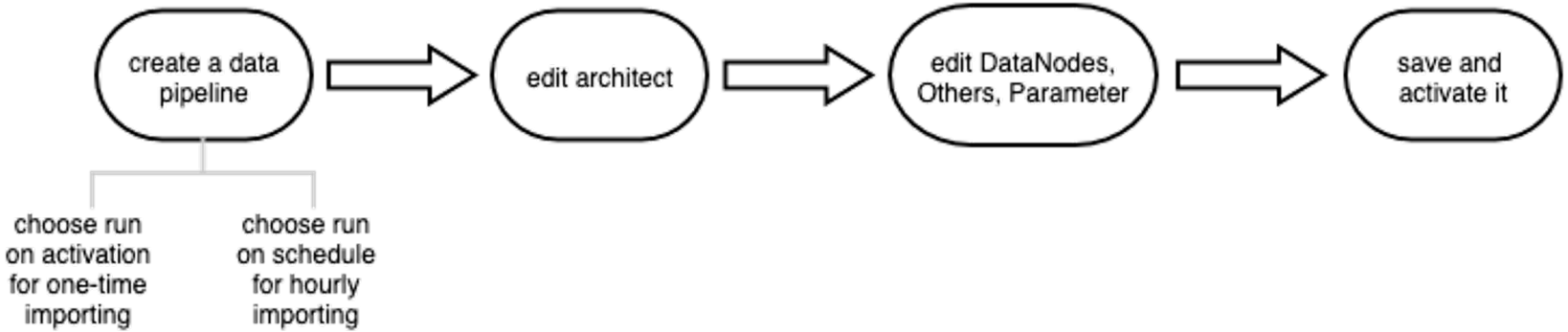
Note:

- One-time importing: click ‘activate’ to run data-pipeline whenever needed
- Hourly importing: auto-run data-pipeline every hour for last 30 days transaction

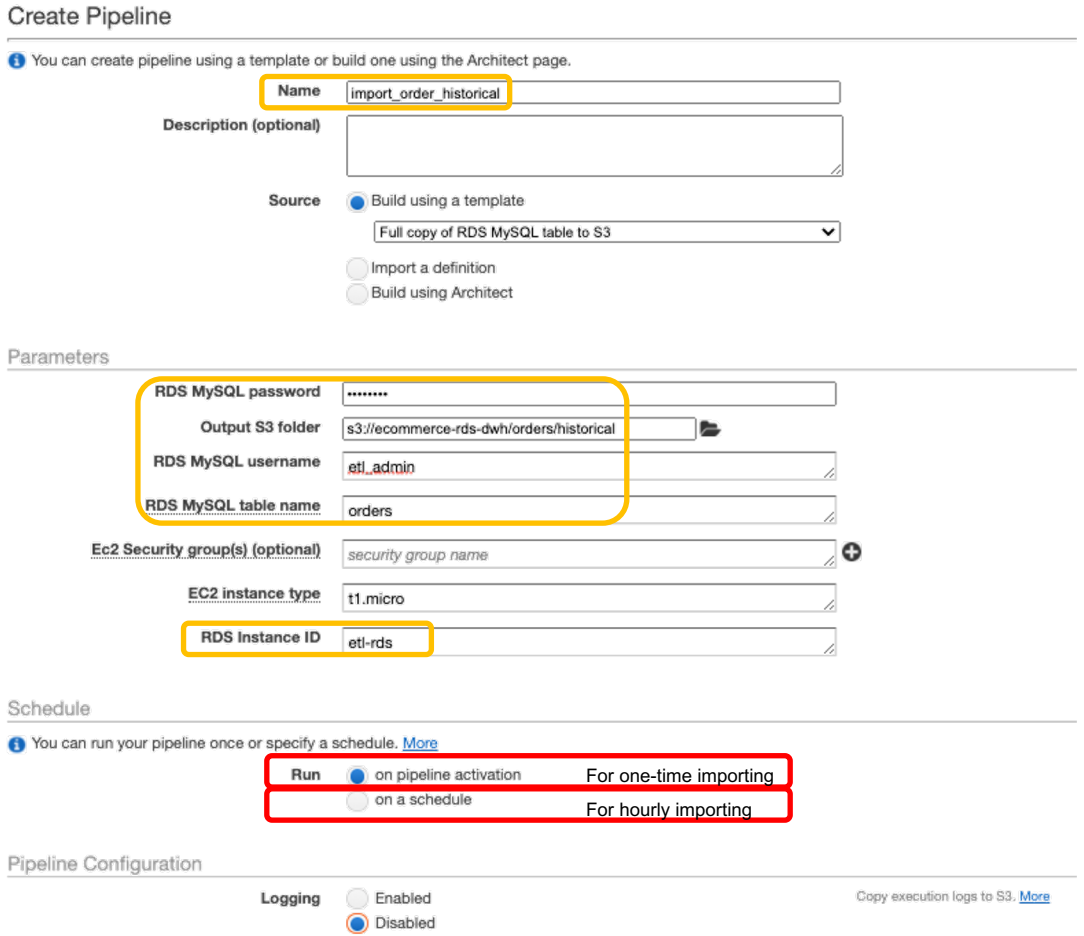
Script for DataNodes

03 Auto ETL Process

AWS Data Pipeline

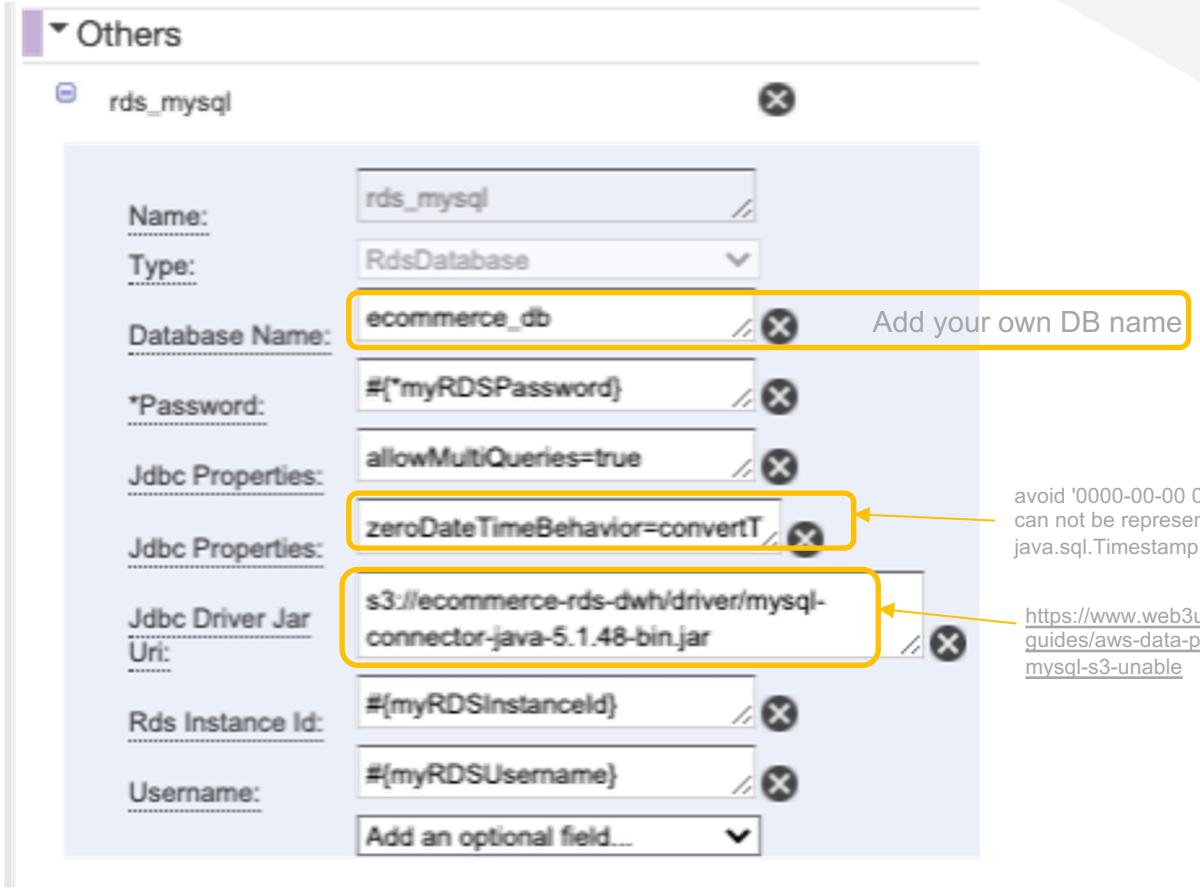


create a data pipeline

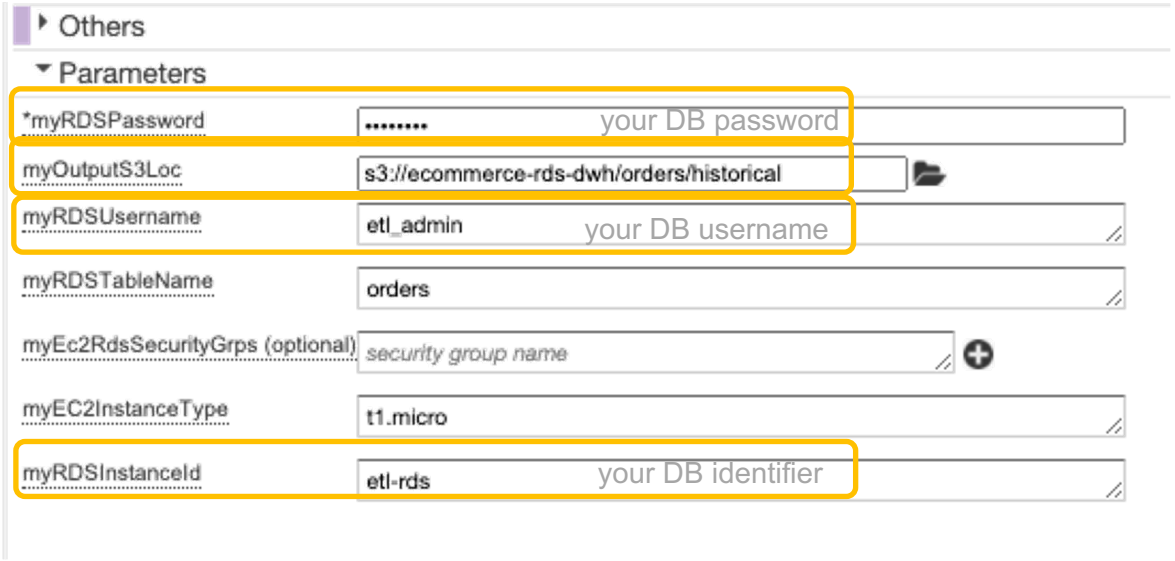


edit DataNodes,
Others, Parameter

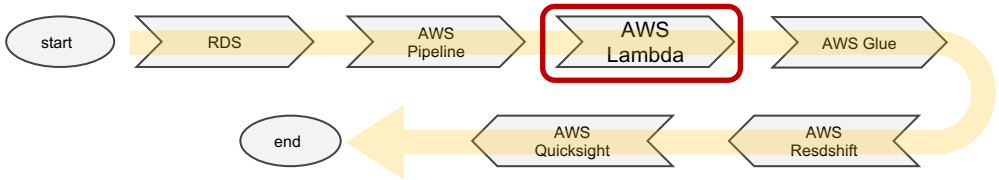
edit DataNodes, Others, Parameter

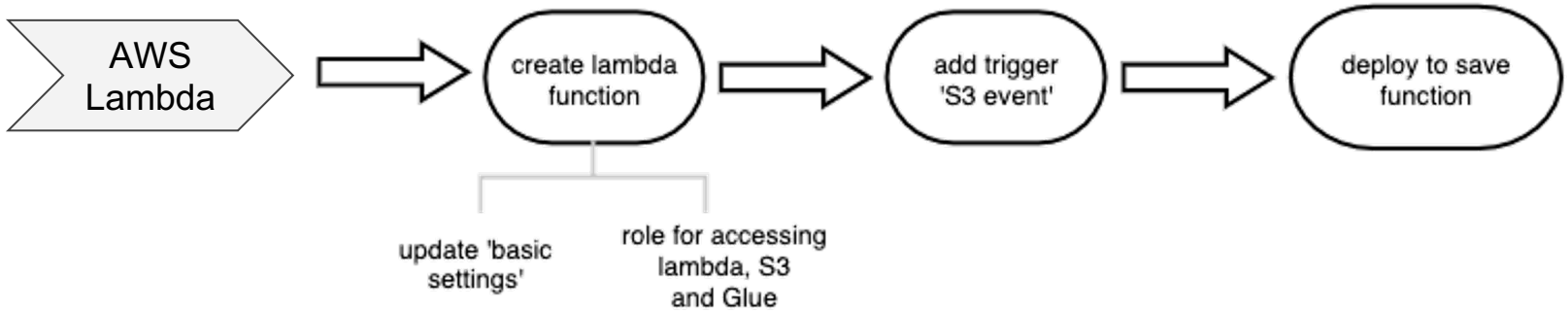


edit DataNodes.
Others. Parameter



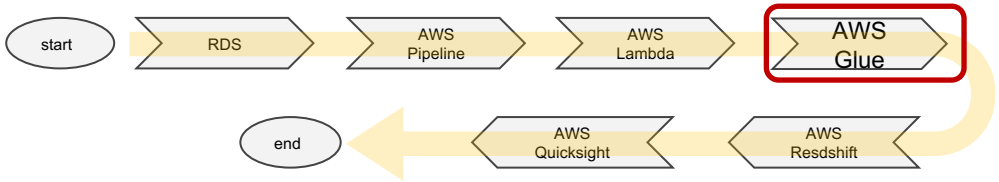
03 Auto ETL Process

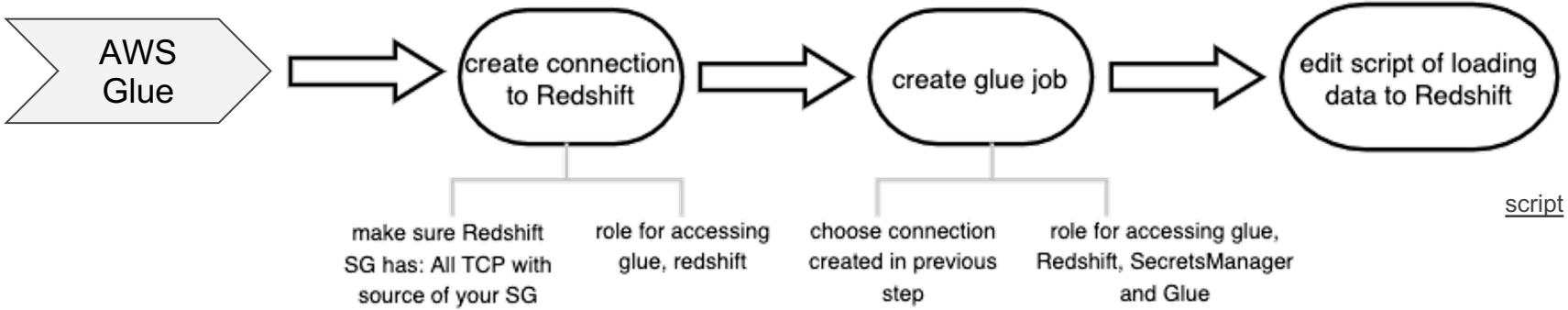




trigger script

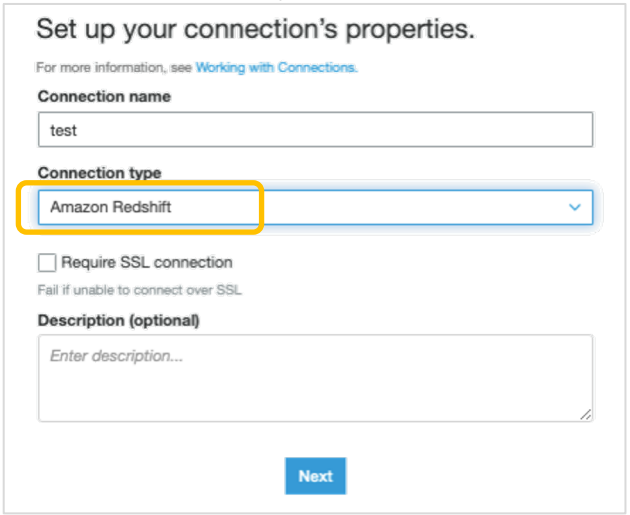
03 Auto ETL Process

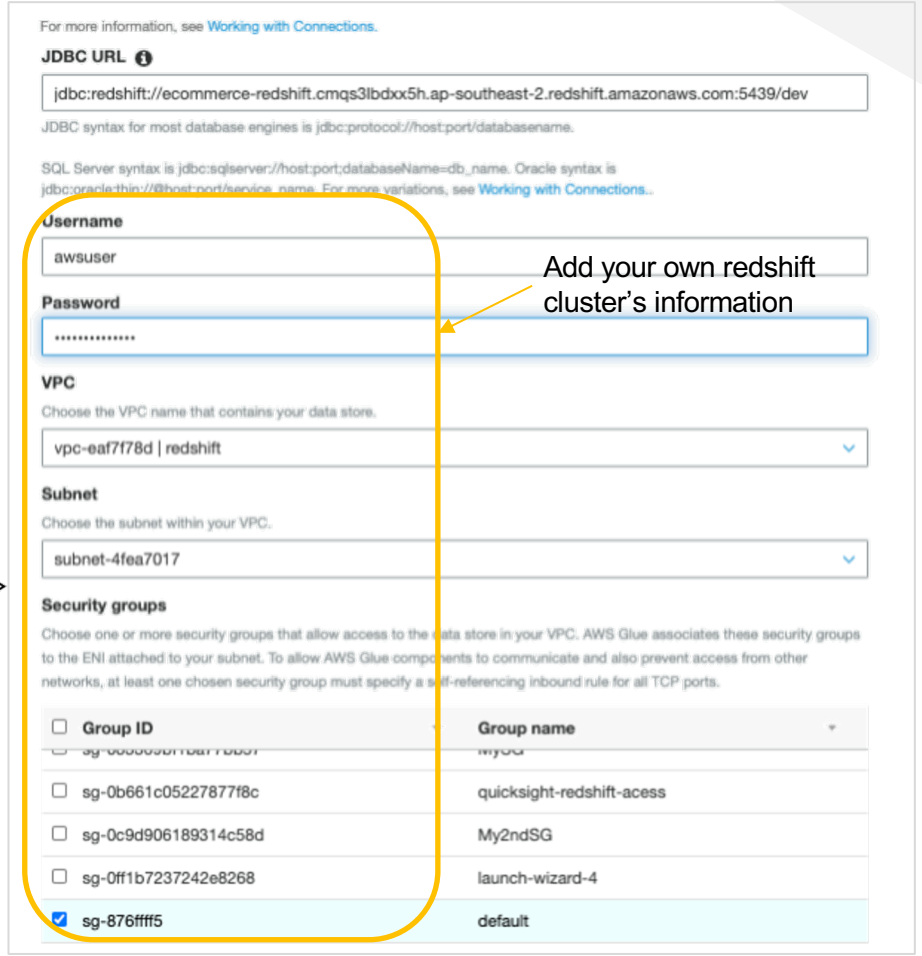




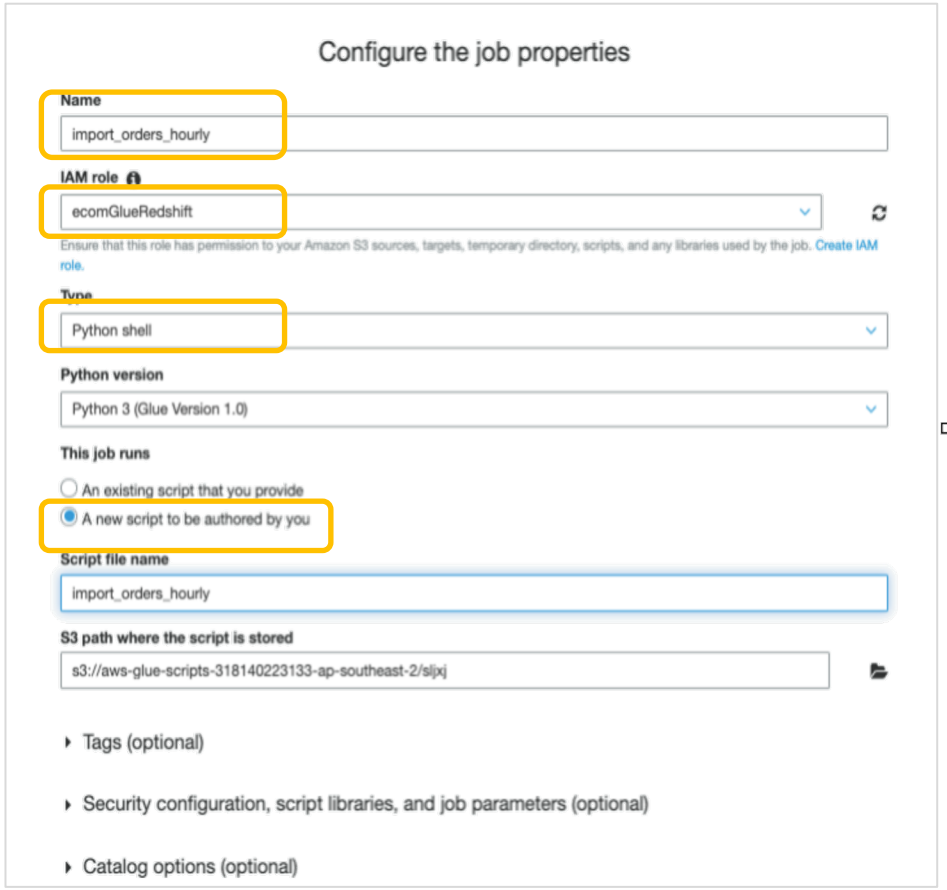
script

create connection to Redshift





create glue job

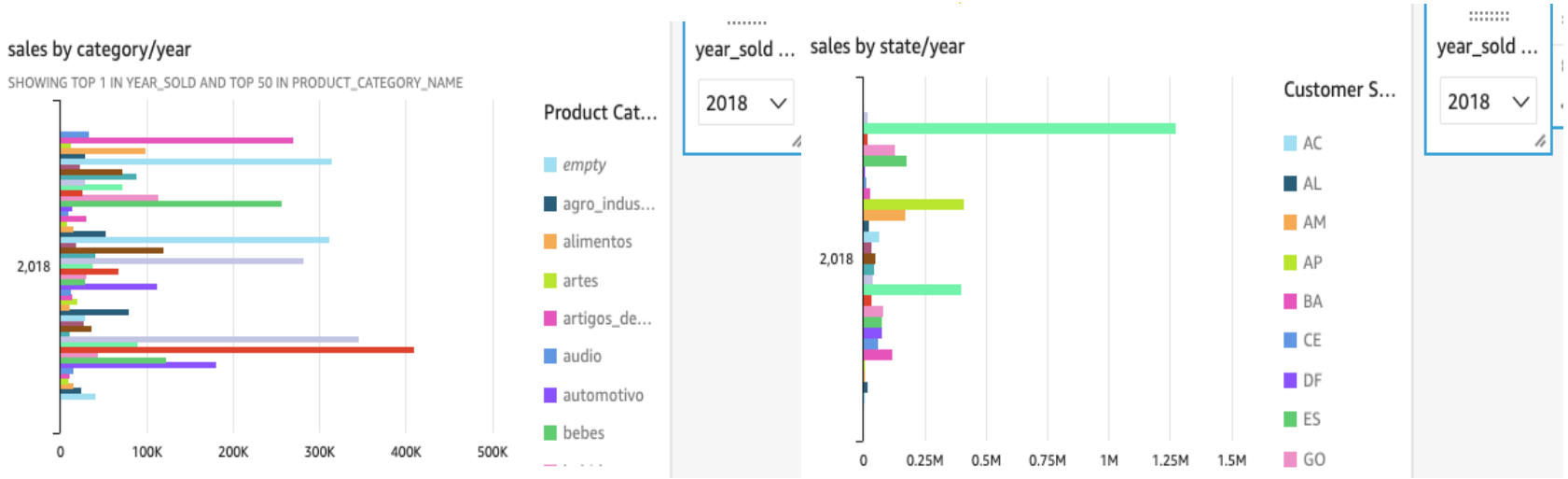


connect to created connection

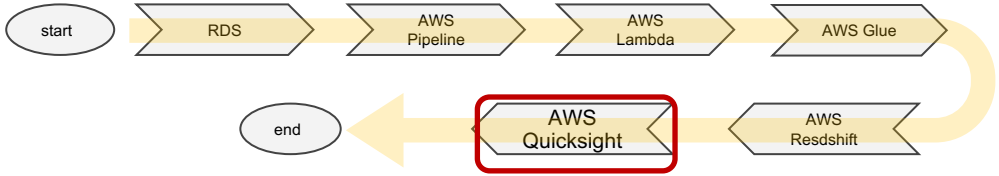
edit script and save

04 Data Visualization

- Use QuickSight, users can easily create reports,
  eg. report of year sales by category or state below

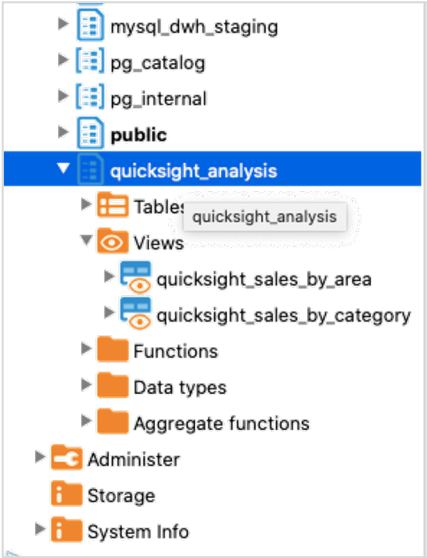


04 Data Visualization



create 'quick-sight' schema and views

- To manage everything easily, a separate schema for QuickSight was created to store all related views
- QuickSight can also do join too




```
create schema quicksight_analysis;

create or replace view quicksight_analysis.quicksight_sales_by_category as
select
p.product_category_name,
EXTRACT (year from date(op.order_purchase_timestamp)) as year_sold,
sum(op.payment_value) as value
FROM mysql_dwh.order_payments op
join mysql_dwh.order_items oi
on op.order_id = oi.order_id
join mysql_dwh.products p
on p.product_id = oi.product_id
group by 1,2;

create or replace view quicksight_analysis.quicksight_sales_by_area as
select
c.customer_city,
EXTRACT (year from date(op.order_purchase_timestamp)) as year_sold,
sum(op.payment_value) as value
FROM mysql_dwh.orders o
join mysql_dwh.customers c
on o.customer_id = c.customer_id
join mysql_dwh.order_payments op
on op.order_id = o.order_id
group by 1,2;
```

To Add On...

- Scenario: combine user_behavior data from the third party to do some analysis
- Solution: a) upload file to S3
  b) Glue crawler to glue data to data catalog
  c) check Athena to make sure table's importing
  d) use Redshift spectrum get external data from catalog, then do join with other tables if needed