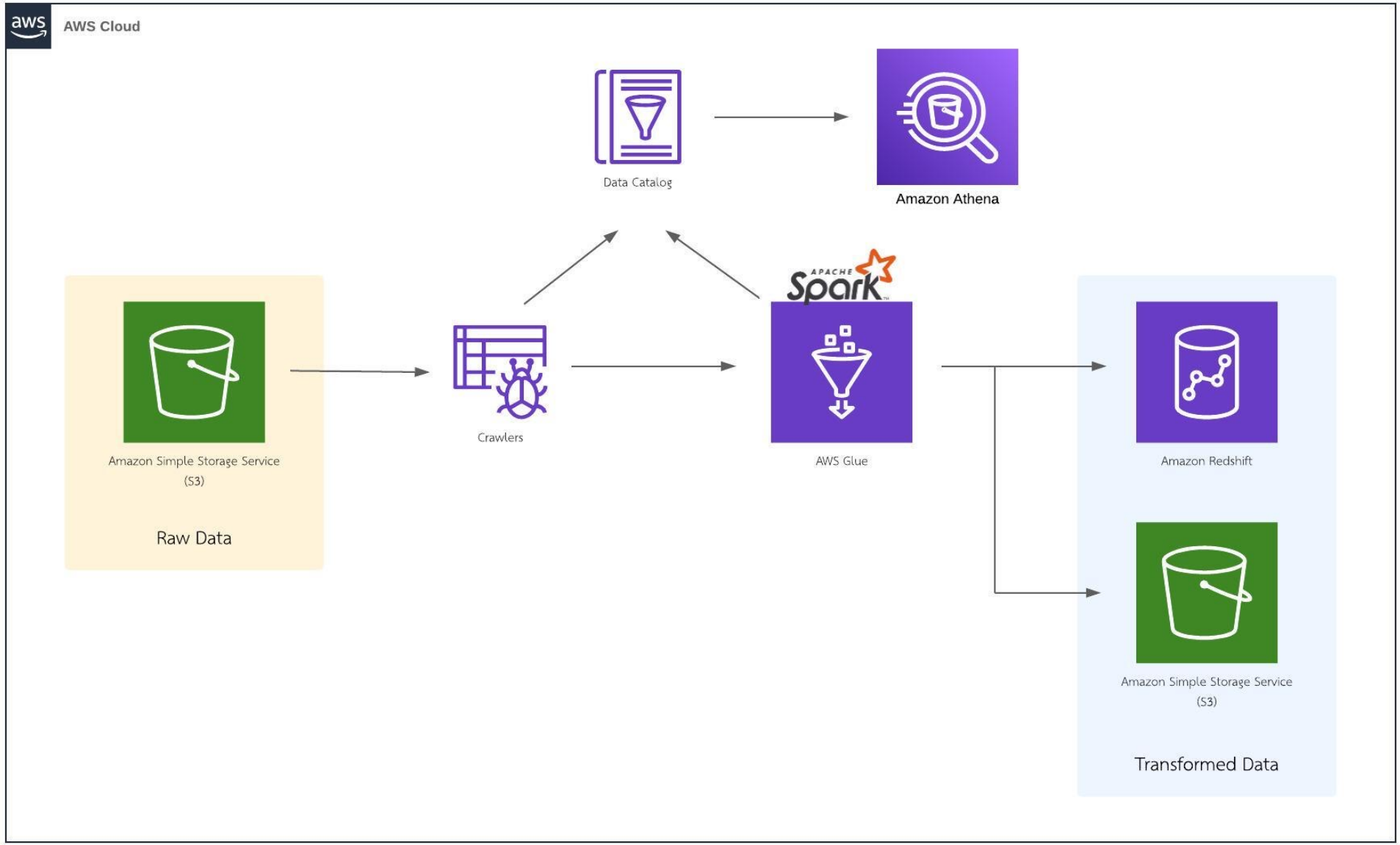
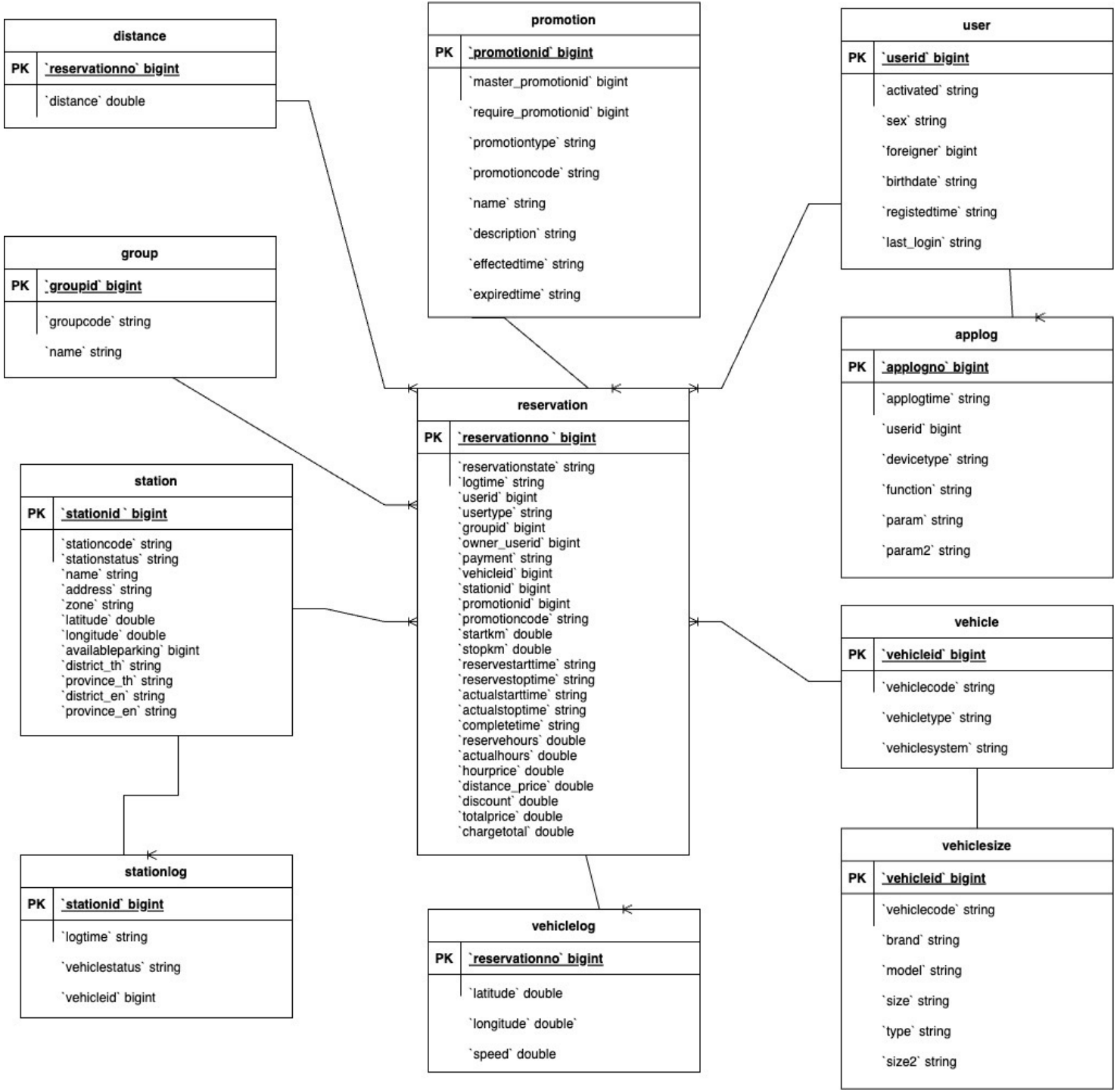
ETL Data Pipeline Development for Car-Sharing Data

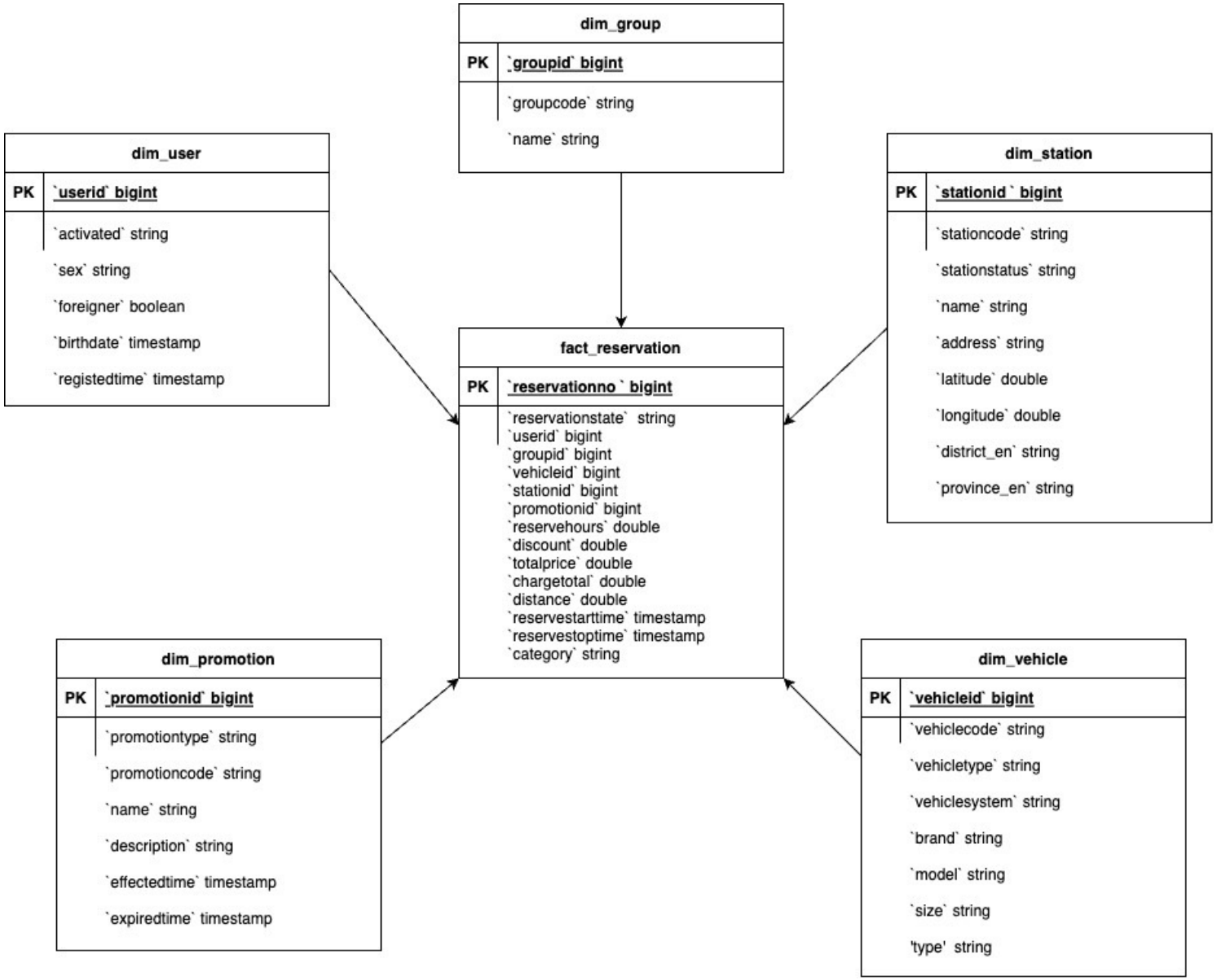
Designed and built a robust ETL data pipeline to transfer raw car-sharing data from the data lake (Amazon S3) to the data warehouse (Amazon Redshift) in the AWS cloud platform

https://github.com/bmbenz/ETL-Data-Pipeline-Development-for-Car-Sharing-Data.git


AWS Cloud
Data Catalog
Amazon Athena
APACHE Spark
Amazon Simple Storage Service (S3)
Raw Data
Crawlers
AWS Glue
Amazon Redshift
Amazon Simple Storage Service (S3)
Transformed Data

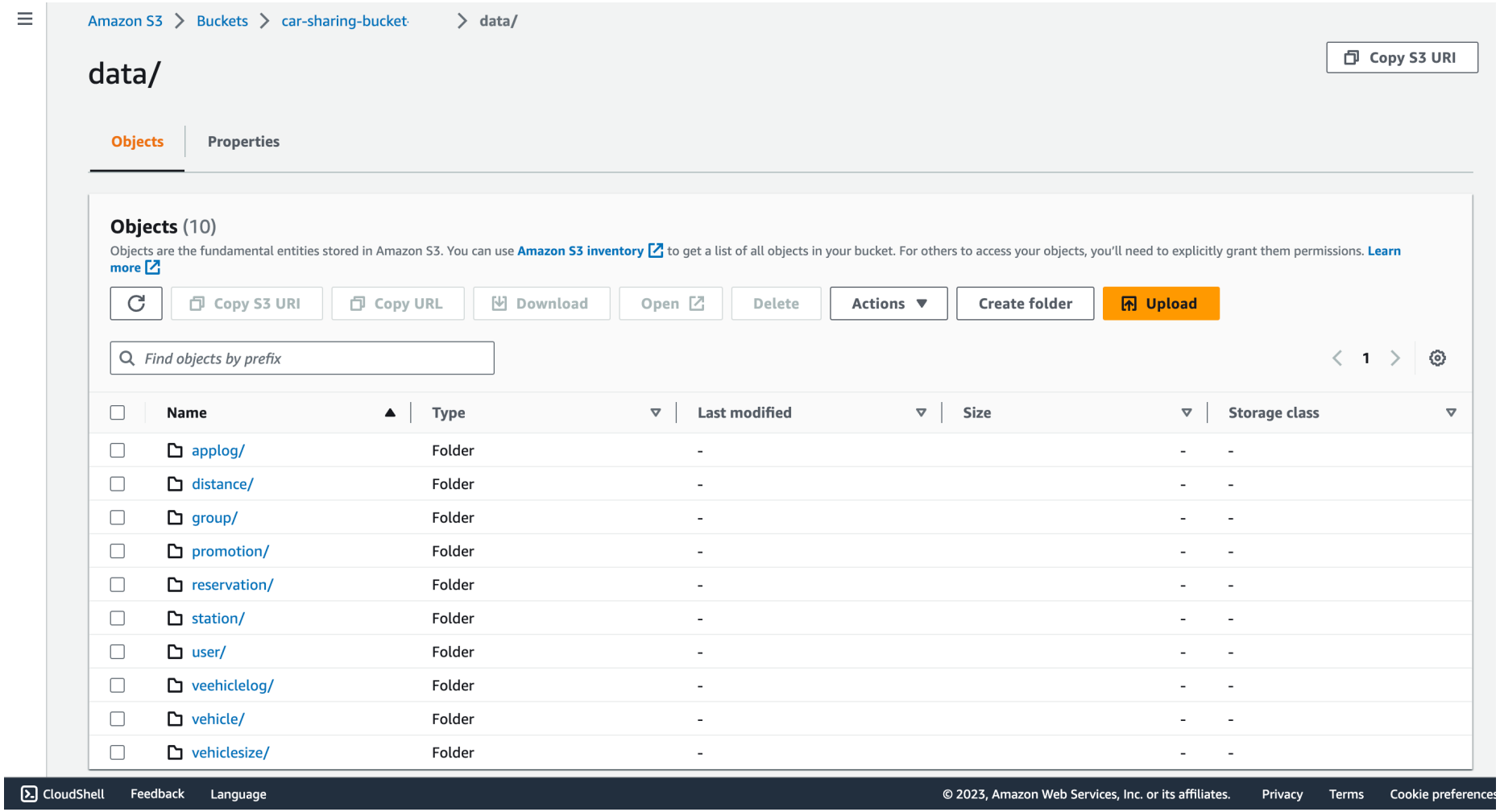
ER Diagram


distance
PK `reservationno` bigint
`distance` double
promotion
PK `promotionid` bigint
`master_promotionid` bigint
`require_promotionid` bigint
`promotiontype` string
`promotioncode` string
`name` string
`description` string
`effectedtime` string
`expiredtime` string
user
PK `userid` bigint
`activated` string
`sex` string
`foreigner` bigint
`birthdate` string
`registedtime` string
`last_login` string
group
PK `groupid` bigint
`groupcode` string
`name` string
applog
PK `applogno` bigint
`applogtime` string
`userid` bigint
`devicetype` string
`function` string
`param` string
`param2` string
reservation
PK `reservationno ` bigint
`reservationstate` string
`logtime` string
`userid` bigint
`usertype` string
`groupid` bigint
`owner_userid` bigint
`payment` string
`vehicleid` bigint
`stationid` bigint
`promotionid` bigint
`promotioncode` string
`startkm` double
`stopkm` double
`reservestarttime` string
`reservestoptime` string
`actualstarttime` string
`actualstoptime` string
`completetime` string
`reservehours` double
`actualhours` double
`hourprice` double
`distance_price` double
`discount` double
`totalprice` double
`chargetotal` double
station
PK `stationid ` bigint
`stationcode` string
`stationstatus` string
`name` string
`address` string
`zone` string
`latitude` double
`longitude` double
`availableparking` bigint
`district_th` string
`province_th` string
`district_en` string
`province_en` string
vehicle
PK `vehicleid` bigint
`vehiclecode` string
`vehicletype` string
`vehiclesystem` string
stationlog
PK `stationid` bigint
`logtime` string
`vehiclestatus` string
`vehicleid` bigint
vehiclelog
PK `reservationno` bigint
`latitude` double
`longitude` double`
`speed` double
vehiclesize
PK `vehicleid` bigint
`vehiclecode` string
`brand` string
`model` string
`size` string
`type` string
`size2` string

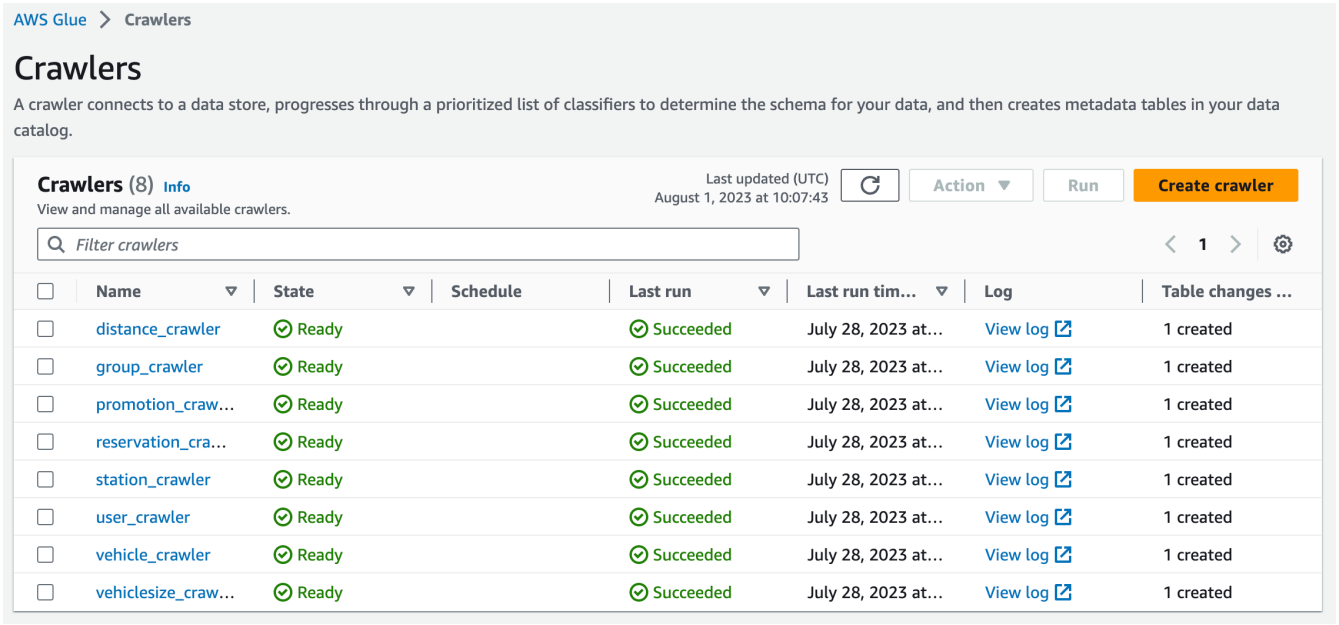

dim_group
PK
`groupid` bigint
`groupcode` string
`name` string
dim_user
PK
`userid` bigint
`activated` string
`sex` string
`foreigner` boolean
`birthdate` timestamp
`registedtime` timestamp
dim_station
PK
`stationid ` bigint
`stationcode` string
`stationstatus` string
`name` string
`address` string
`latitude` double
`longitude` double
`district_en` string
`province_en` string
fact_reservation
PK
`reservationno ` bigint
`reservationstate` string
`userid` bigint
`groupid` bigint
`vehicleid` bigint
`stationid` bigint
`promotionid` bigint
`reservehours` double
`discount` double
`totalprice` double
`chargetotal` double
`distance` double
`reservestarttime` timestamp
`reservestoptime` timestamp
`category` string
dim_promotion
PK
`promotionid` bigint
`promotiontype` string
`promotioncode` string
`name` string
`description` string
`effectedtime` timestamp
`expiredtime` timestamp
dim_vehicle
PK
`vehicleid` bigint
`vehiclecode` string
`vehicletype` string
`vehiclesystem` string
`brand` string
`model` string
`size` string
'type' string

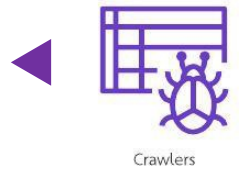
Star Schema

1. Create a bucket and upload data

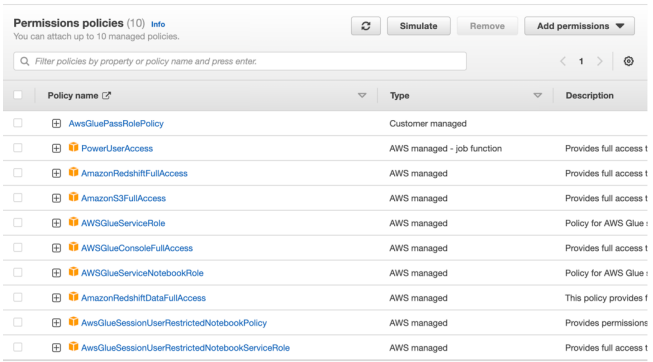


2. Create data catalog using crawler



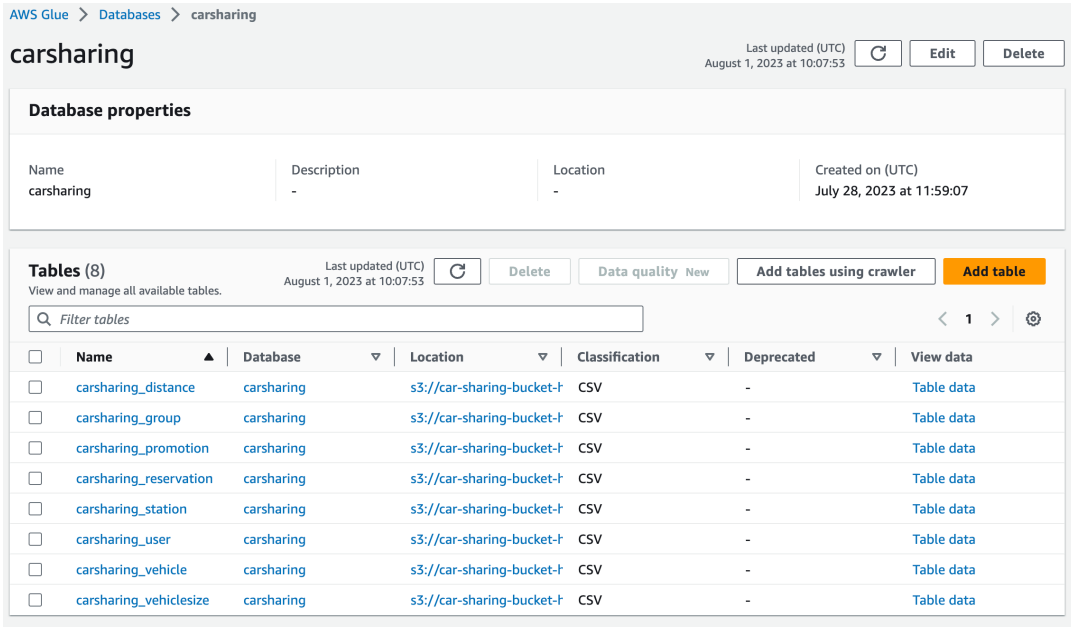


Crawler

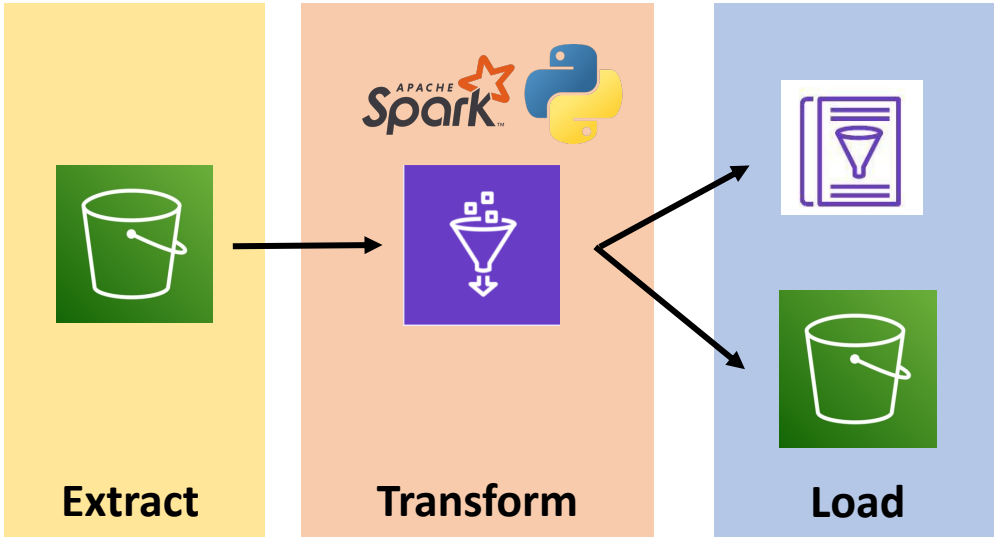


IAM Roles

Data Catalog

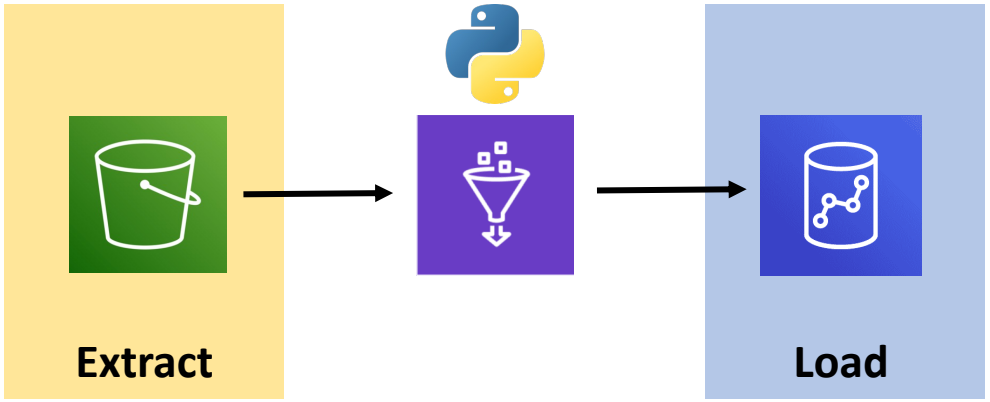


3. ETL raw data using AWS Glue



ETL_S3_Athena: Python Script for Extracting data from Amazon S3, Transforming data on Glue Notebook Studio using PySpark, Loading data to Amazon S3, and creating tables on the AWS Glue Catalog.

EL_Redshift: Python Script for Extracting data from Glue Catalog, and Loading data to Amazon Redshift with Redshift connector.

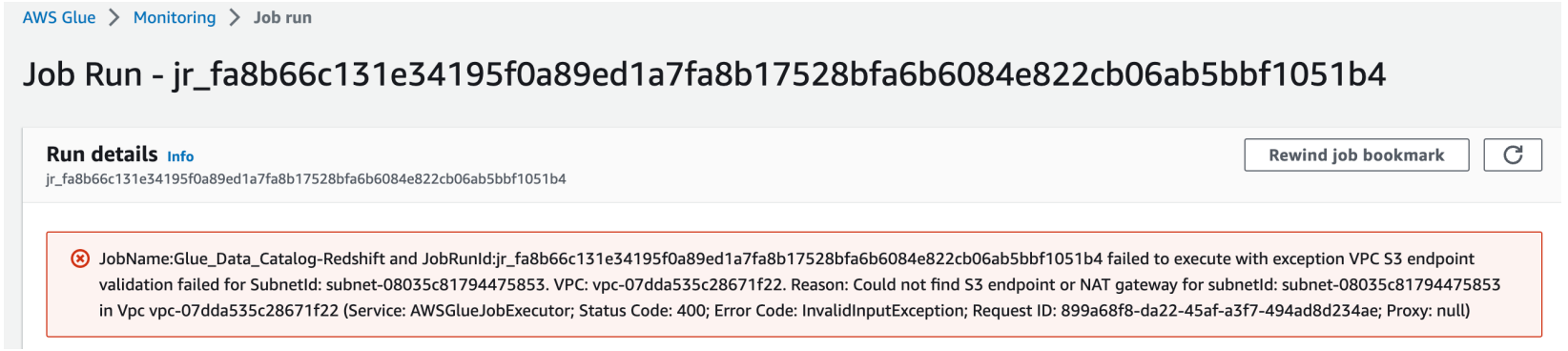


Problem

Can not load data to Redshift

Code with many script but it’s not work, and finally…

1. Create Glue Connector (Connector type: Redshift) but…



2. Solved by create a new VPC Endpoint

https://repost.aws/knowledge-center/glue-s3-endpoint-validation-failed