Cloud Project

S3 data integrated with Snowflake.

① → create s3 bucket → (etl-project-s3-by-adi)

↳ disable Public

↳ Acknowledge ✓

↳ Go to created bucket

↳ create subfolders

input/

output/

script/

temp/

↳ Go to input folder.

↳ create subfolders as.

input/Product/

↳ Go to input Product folder

↳ create subfolder.

input/Product/year=2021/

→ (upload data into Partition folder after crawler creation).

② create Database

→ Go to Glue service.

→ Go to Databases.

↳ Add database.

↳ name → (ETL-Project-db-by-adi)

↳ create database.

③ Create IAM role ①

→ Go to IAM service.

→ Go to Access management

↳ roles

↳ create role

→ PTO →

↳ AWS service under Trusted entity Type.

↳ use case → Glue →

↳ Next

↳ Permission Policies.

↳ AWSS3FullAccess.

↳ CloudWatchFullAccess.

↳ awsglueservicerole.

↳ Next.

↳ Role details

↳ Role name - (etl-Project-iam-role)-by-adi)

↳ create role.

) create Crawlers

→ Go to AWS Glue service.

→ Go to Data Catalog

↳ select crawlers.

↳ create crawler

↳ name

(etl-Project-crawler-fetch-from-S3-input-by-adi)

→ Next

↳ choose Data source & classifier

↳ Not yet

↳ Add a data source.

↳ Data source - S3

↳ Location of S3 data - In this account

↳ S3 Path : bucket/input/(name of)

→ Add an S3 data source.

→ Next

↳ Configure security settings.

↳ IAM role

↳ select (etl-Project-iam-role-by-adi)

→ Next

↳ Set o/p & scheduling.

↳ target database.

etl-Project-bb-by-adi

↳ crawler schedule

↳ on-demand

↳ review & create.

↳ create crawler.

→ upload 2021 data file to S3 bucket under Year=2021 folder.

→ Run crawler now

-name (etl-Project-crawler-fetch-from-S3input-by-adi)

→ Go to tables under data catalog. and check the table(product) and all

(4i) → Go to Athena service.

↳ check database (eH-Project-blog-by-adi)

↳ check table (Product

↳ observe the result by previe results.

(5) Create ETL job.

⇒ Go to Glue Aws service

↳ ETL jobs.

↳ visual ETL.

↳ under ~~job~~ job details tab

↳ name → eH-project-read-from-S3-job-by-adi

↳ IAM Role

↳ eH-Project-iam-role1-by-adi

↳ Type → Spark

↳ Glue version → 3.0

↳ Language - Python

↳ Worker type - 16GB RAM (Select lower RAM)

↳ job timeout - (5) mins

↳ under Advanced Properties.

↳ script path

↳ bucket/script/

↳ Temporary path

↳ bucket/temp/

↳ Under visual tab

↳ select source

↳ S3

↳ Source type

↳ Data catalog table.

↳ Database (etl-project-db-by-adi)

↳ Table (Product)

↳ Select Target

new columns in Parquet
newYear
Count
Quantity

↳ S3

↳ Format (Parquet)

↳ S3 Target location

s3://bucket/output/

↳ Under script tab

↳ Edit script

↳ Upload data from local device.

→ Run job.

→ After success.

→ Go to S3 bucket

bucket/output/newproduct/

→ Good.

(5i) → Go to Glue Service

↳ under to Workflows (orchestration)

↳ Add workflow

name (etl-Project-Workflow-by-adi)

↳ create workflow

→ open workflow.

↳ Go to workflow details.

↳ Graph

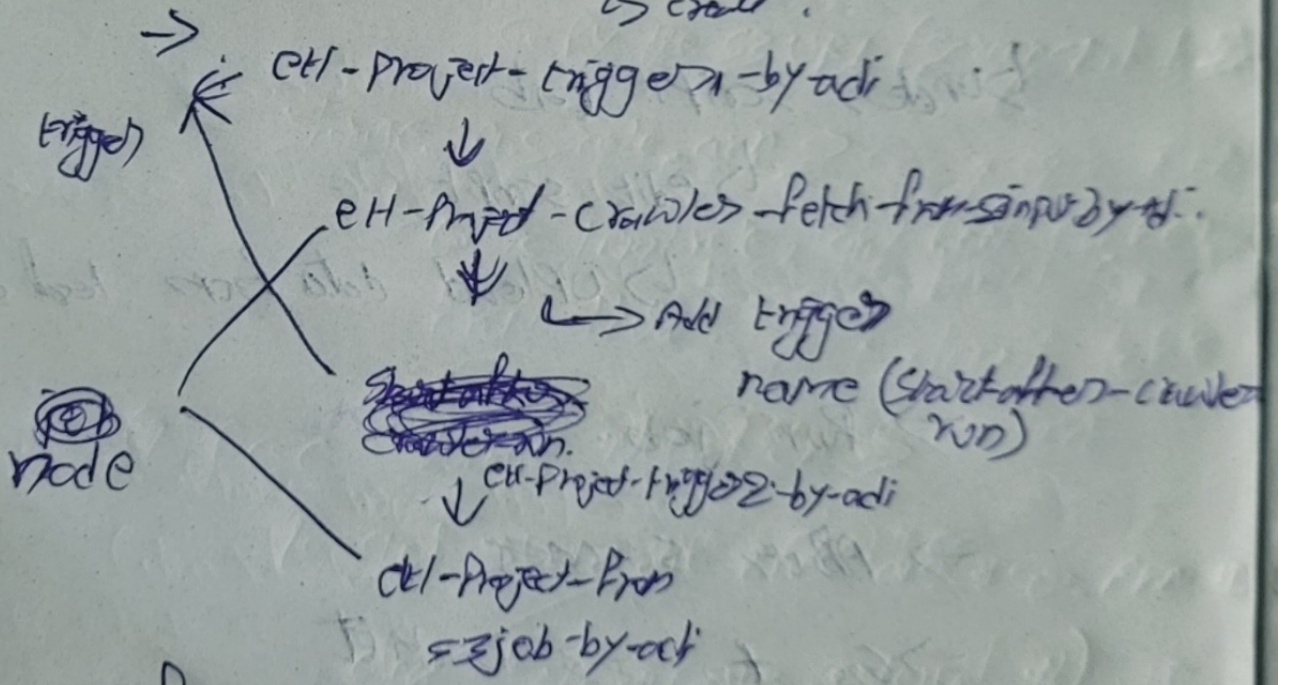
↳ select Add trigger.

↳ Add new

↳ name

(etl-Project-trigger-by-adi)

↳ create.



All Good So far

⑥ Snowflake Work:

① create warehouse
↳ ETL_Project_snowwh_by_adi;

② create Database.
↳ ETL_Project_snowdb_by_adi;

③ Create schema
↳ etl.Project_snow~~schema~~db_by_adi.Prod;

④ Create File Format
↳ etl_project_Parquet_Format_by_adi

⑤ Create IAM role ②
→ Go to IAM service
→ Go to Access management
↳ roles.
↳ Create role.
↳ Trusted entity Type
↳ Aws account
↳ An Aws ~~Acc~~ Account
↳ This account
↳ options
external ID ☑
↳ for now give 0000
↳ Next
↳ Add Permissions.
→ Aws S3 Full Access
→ Aws Secrets Full Access
↳ Next

↳ name, review, and create

↳ Role Name

↳ CH-Project-iam-role2-by-adi

↳ create role

⑥ Create storage integration

↳ name (CH-Project-Sto.intg-by-adi)

⑦ Show

name →

type → External_stage

enabled → true

storage_Provider → S3

storage_allowed-location →

↳ Go to S3 output folder and copy the S3 URL (screenshot taken)

Storage_Aws_role_arn →

↳ Go to IAm role, copy ARN number from there.

⑦ Desc storage integration →

to get → ① storage_AWS_iam_user_ARN ID and ② storage_aws_external_id) ID.

Keep this ID under IDM Trust edit Policy script (screenshot there)

Keep this ID under IDM Trust edit Policy script (screenshot there).

⑧ Create stage

name → etl-Project-stage-by-adi

URL →

Go to S3 output folder and copy the S3 URL (Screen taken).

storage_integration →

name given to storage-integration

(etl-Project-storing-by-adi).

file_format = name given to file format

(etl-Project-target format-by-adi)

⑨ check file loaded or not

list @stagename.

1st way

⑩ create target table.

⑪ copy stage data to target tbl.

⑫ Query results.

2nd way ⑬ Create external table

for that do below.

① create table first (Query available)

name → ext-customer-Product.

② show external tables.

↳ copy notification-channel value.

③ Go to S3 output object bucket

bucket/output/newProduct

bucket

↳ Properties

↳ event notifications

↳ create-event-notification

↳ name -> dll-Project-datarchives-by-adi

↳ object creation

↳ All object create events

↳ object removal

↳ All object removal events

↳ Destination

↳ SQS queue

↳ specify SQS queue

↳ enter SQS queue ARN

↳ Past the ARN copied from external table show command

→ Save changes.

(14) Go to S3 & upload data to S3/output manually or through glue job

(15) Wait for some to query select.