<https://www.youtube.com/watch?v=dQnRP6X8QAU>

**AWS Glue Tutorial for Beginners [FULL COURSE in 45 mins]**

Johnny Chivers

AWS Glue –

* Is Fully Managed ETL Service
* Consists of Central Meta Data Repository – Glue data catalog
* A spark ETL engine (serverless)
* Flexible Scheduler

AWS Glue is Needed when there is requirement for ETL service in AWS

Examples of Meta data🡪 Data Location/Schema/Data Types/Data classification

Set-Up🡪 Create S3 Bucket/ Create folders/ create sub folders..

athenaResults – data -Scripts -temp-dir (folders)

Data/customer\_database/customers\_csv/dataload=yyyymmdd

Create folder as Customers table and sub folder as customer\_database

Upload file Upload samaple .csv data file under dataload = yyyymmdd

Create IAMROLE – select glue – select adminrole – create

Console – AWS glue – create database – (name must have \_ as spark uses \_ customer\_database)

Location – (get copy s3 url from dataload file in S3)- create.

Console – AWS Glue- Select database – select tables in customer\_database-

We can add tables – manually or add using crawler

Crawler name – sorce s3 – etc… finish (creates table for the s3 data we uploaded cuatomer data)

Select table and click run crawler… to see data displayed

Tables – View partitions (we have provided dataload as our partition)

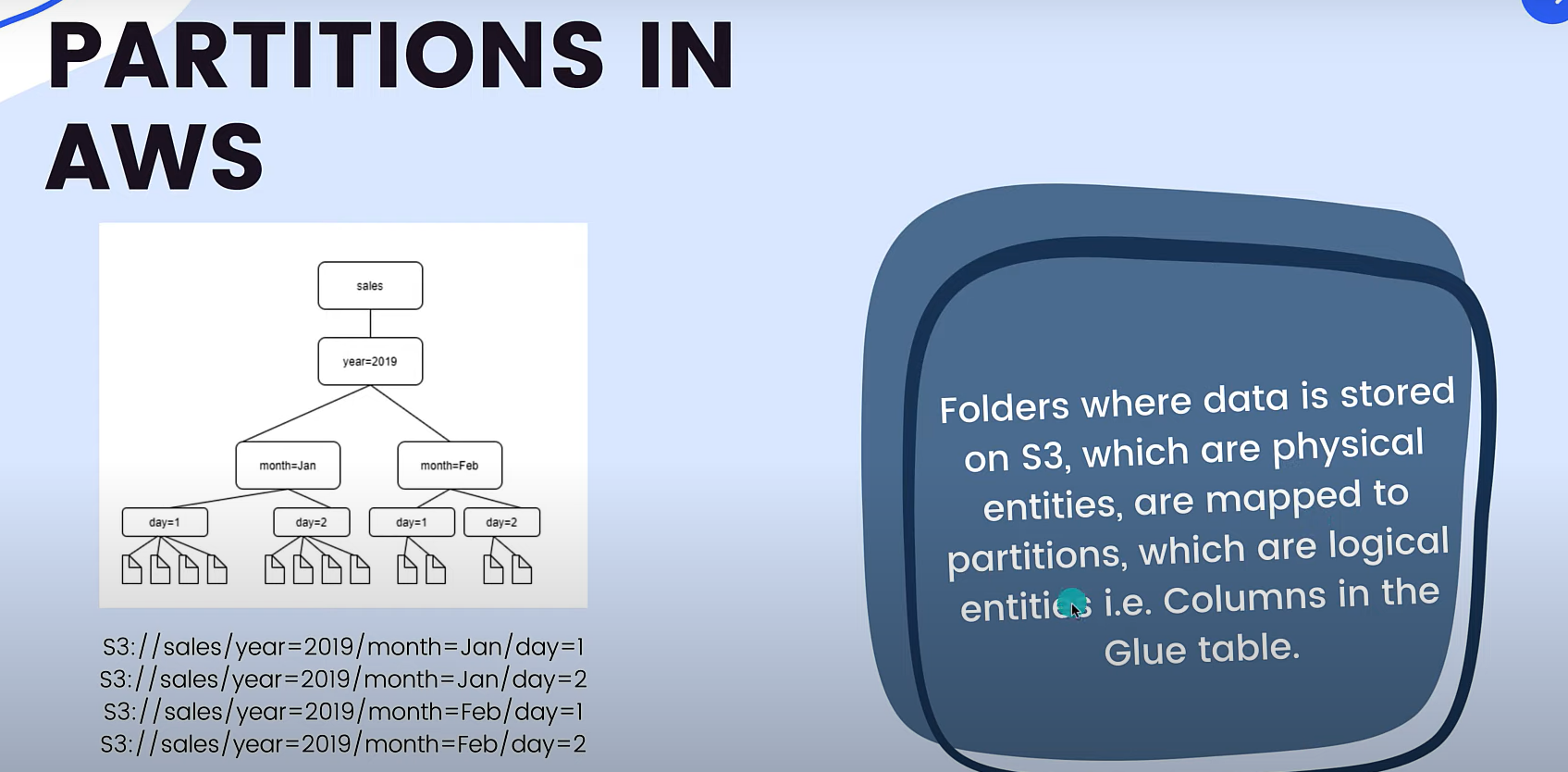
S3 – create folder as – athena\_results

AWS Athena – Query our data – settings – manage -browse s3 – glue bucket -athena results chose.

Select data source as awsdatacatalog and database as customer\_database- select your table – preview-

Will display all customer table data..

S3 – create folder as – athena\_results



Partitions help to query the data more efficiently..

AWS Glue Crawler 🡪 A program that connects to data source (source or target), progresses thru a prioritized list of classifiers to determine schema for your data, then creates metadata tables in AWS glue data catalog.

AWS Glue Jobs🡪 AWS Glue – jobs – create job- follow steps – save – run job..

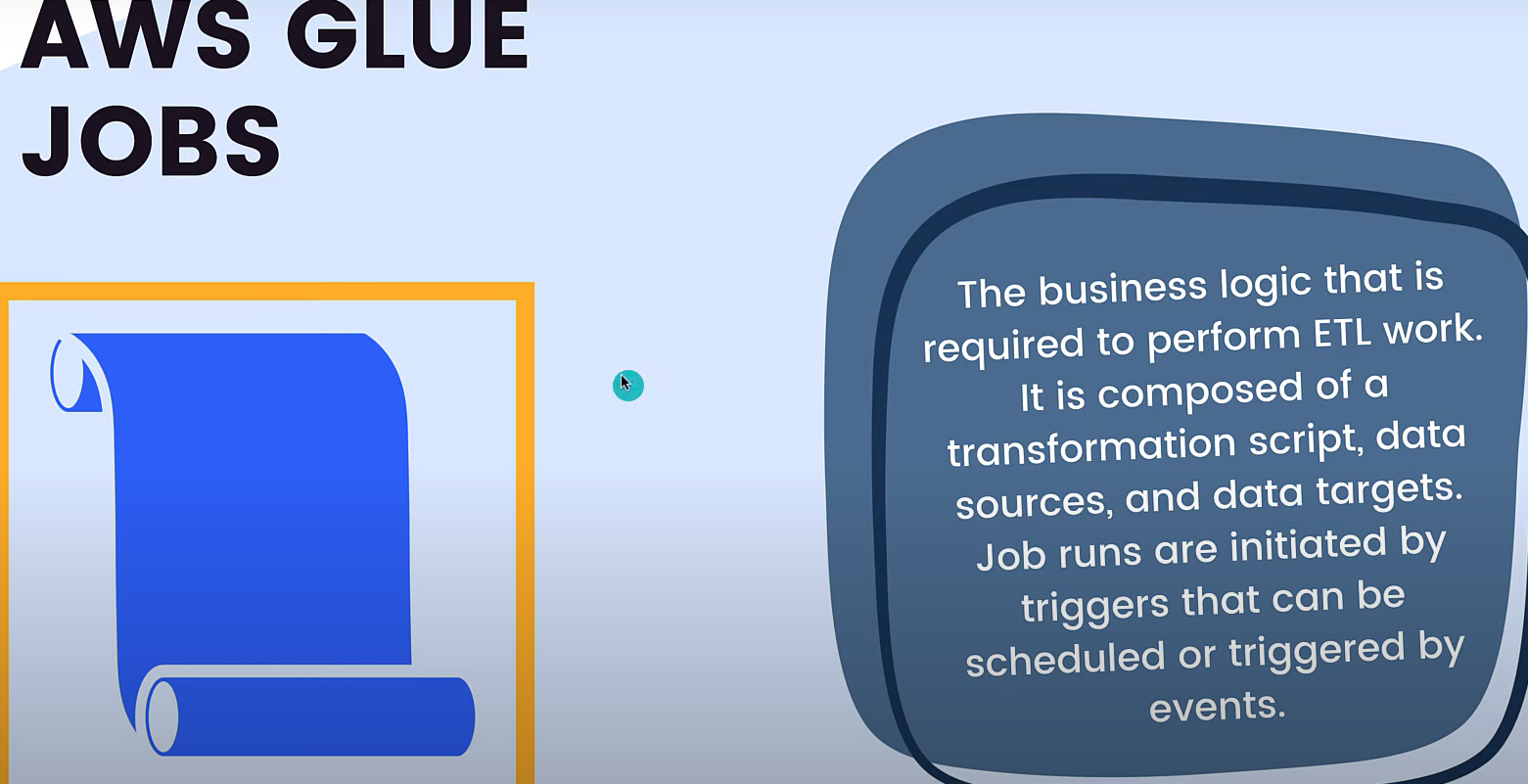
We are moving data csv data to parque fromat.. by following steos aws generates script code for same..

AWS Glue trigger – Initiates an ETL job..

Trigger types are – schedule – job events – on-demand-event bridge

Frequency daily/weekly/hourly/monthly/custom

Job events – When 1st job succeeds run second job….



AWS Glue – jobs – create job-

Check video its clear and good to follow steps..

Look for video on create a database connection on AWS…