Day - 10 (03rd June)

03 June 2024 10:00

1. AWS Glue.
2. AWS Redshift.
3. AWS Athena.
4. SQS & SNS
5. API gateway.
6. AWS EMR
7. cloudwatch & Monitoring.
8. Apache Hudi
9. Lambda function.

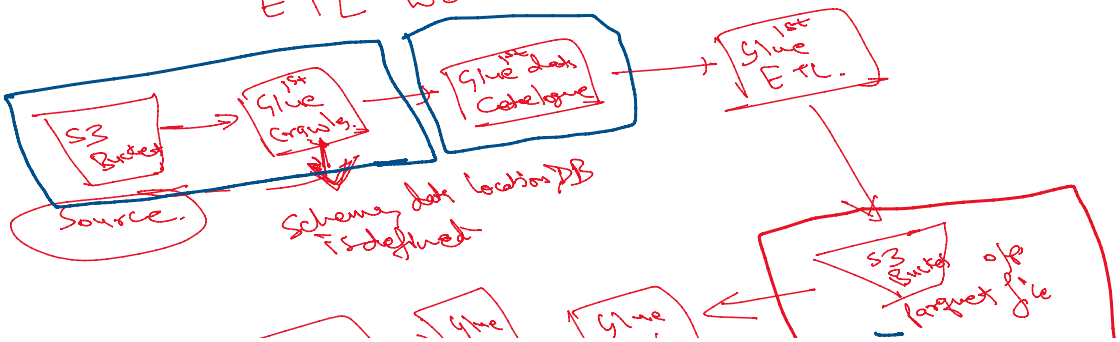
Quiz :-

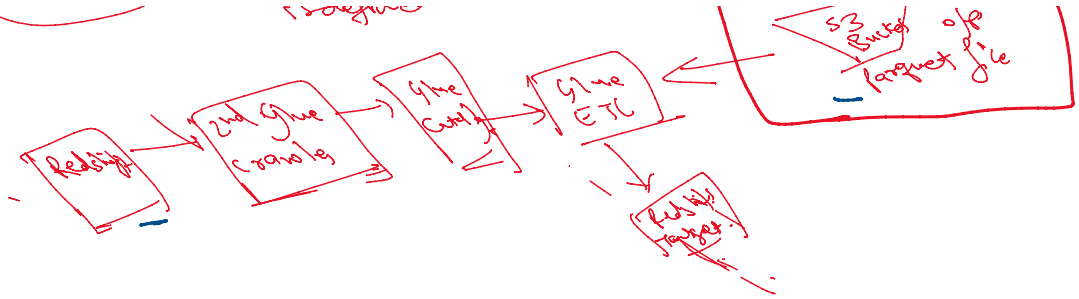
1. 7th june
2. 14th june (Post training Quiz)

1) AWS Glue

1. AWS Crawler
2. Glue ETL Script
3. Aws Glue workflow.
4. Glue concept
5. Read CSV file from S3 bucket using Glue catelogue.
6. Insert data into Redshift w/o Glue Catelogue.
7. S3 + Glue + Athena.
8. S3 + Glue + Redshift.
9. Pyspark concept.
10. Frame Concept.

AWS Glue provide console of API operation to setup & manage your ETL workload.





① Crawler:- Connect with source or target - to store the schema or the metadata information of the data.

② Glue for ETL:- Processing the data - Python or Pyspark -

Python:- Small dataset. why? not distributed processing, Use Pandas, Numpy to perform.

Pyspark:- Large dataset. distributed processing → RAM → speed very high as compared to Python

Glue
↓
large dataset
Processing → upto 4 hrs
Serverless service

Lambda
→ small dataset
Processing → 15 min
Serverless service.

Store code → No cost
↓
Runs of execution

Steps to create the crawler:-

① Save data in S3 Bucket.
② Read data from S3 using Crawler
③ Create the crawler to reference S3 Bucket.
④ Create IAM Role → S3 fullaccess, AWSGlueService & cloud watch
⑤ Column as partition column. Data will be visible in Athena.
⑥ [illegible] update. Append or modify files

5. Column as partition column.

6. Crawler → incremental update. Append or modify files to the existing Catalogue tables. It will not insert the old data file again