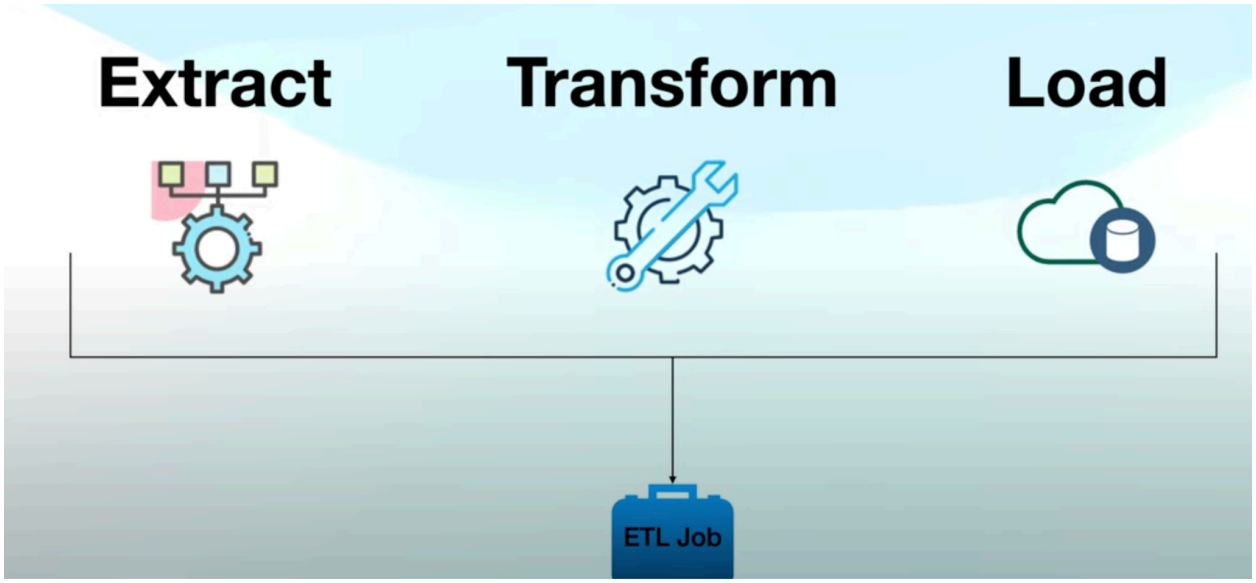
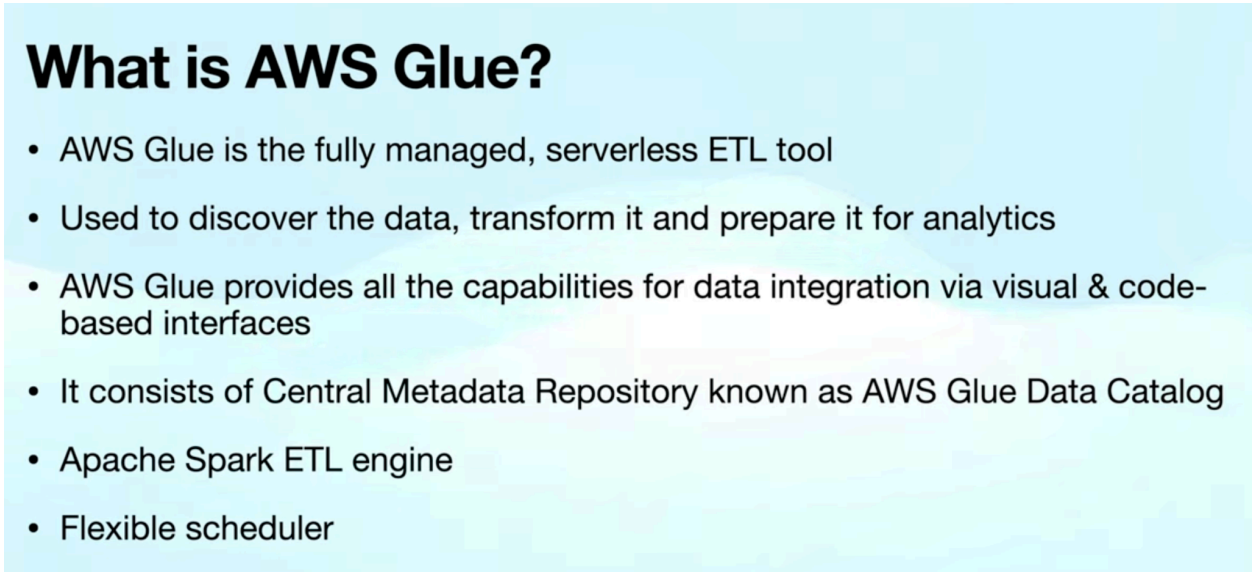
ETL JOB



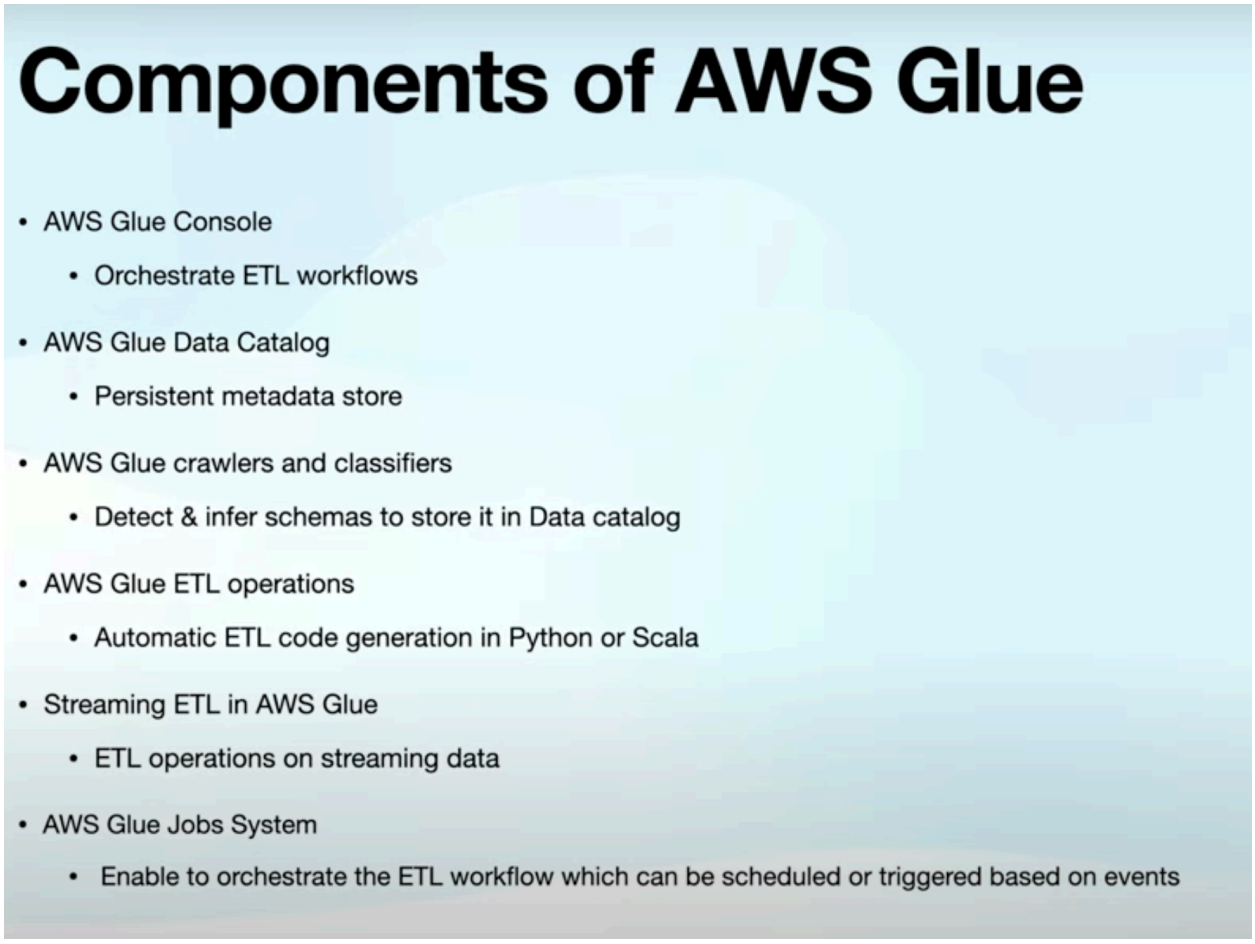
Running ETL options

On-Prem ETL	Serverless ETL
Team has to take care of the infrastructure like servers, its maintenance, software updates, apply patches, etc	The infrastructure part is maintained by the service provider
Over allocation of resources for an ETL task results in idle resources + Before hand investment for servers/infra setup	Only pay for what you use. No need to own a setup/server
Aware of what is provisioned	Little or no knowledge of what is provisioned
Focus on infra, choosing the ETL tools, software for ETL job + building core data pipelines	Allows developers more time to focus on data preparation, building data pipelines
Code for ETL jobs is written from scratch	Generate the serverless code for the ETL job

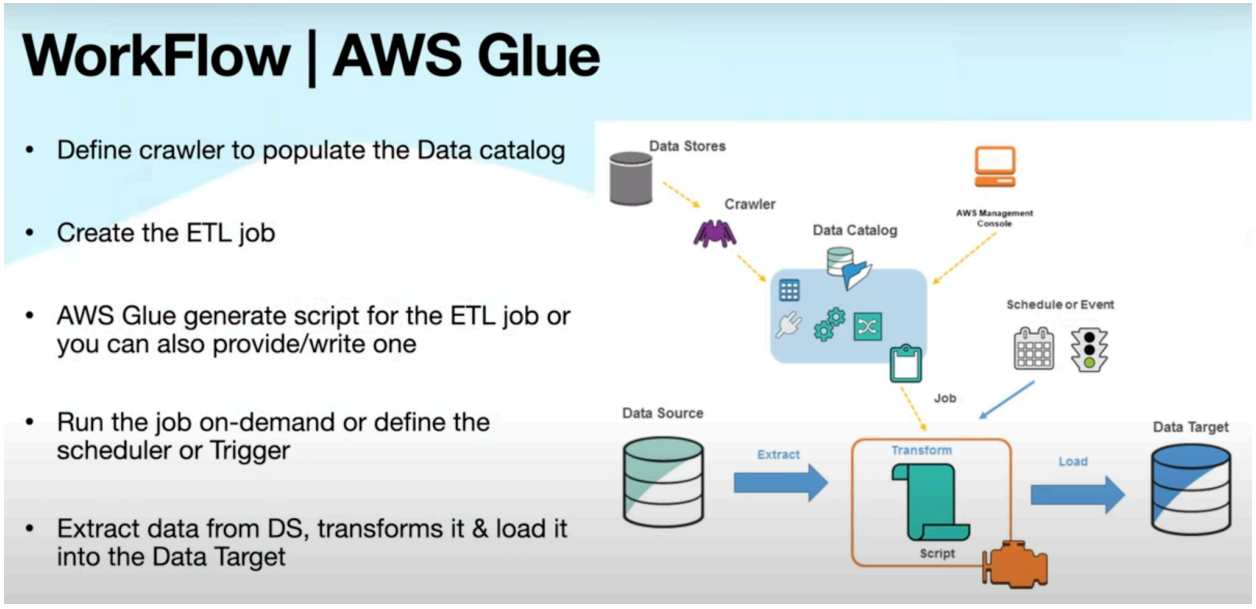
Introduction to aws GLUE



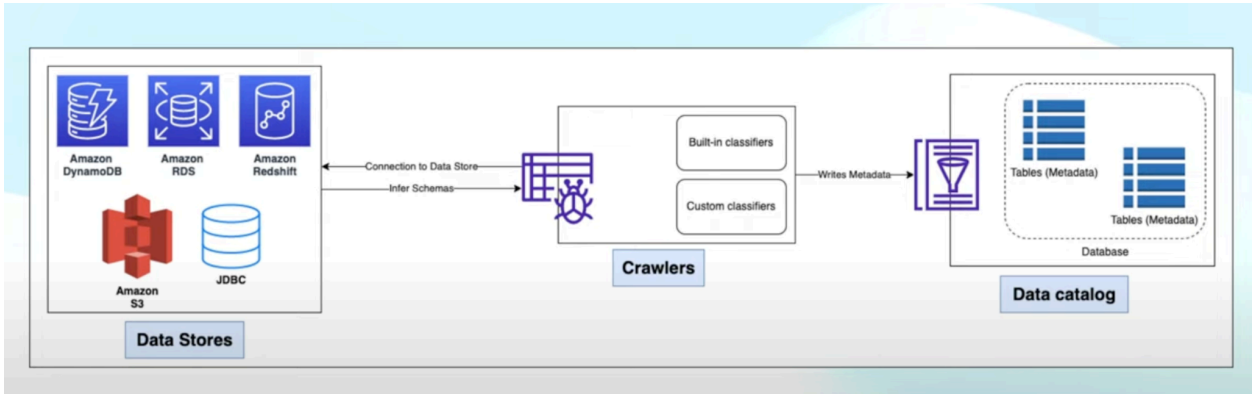
Components of aws GLUE



AWS GLUE WORK flow



More info

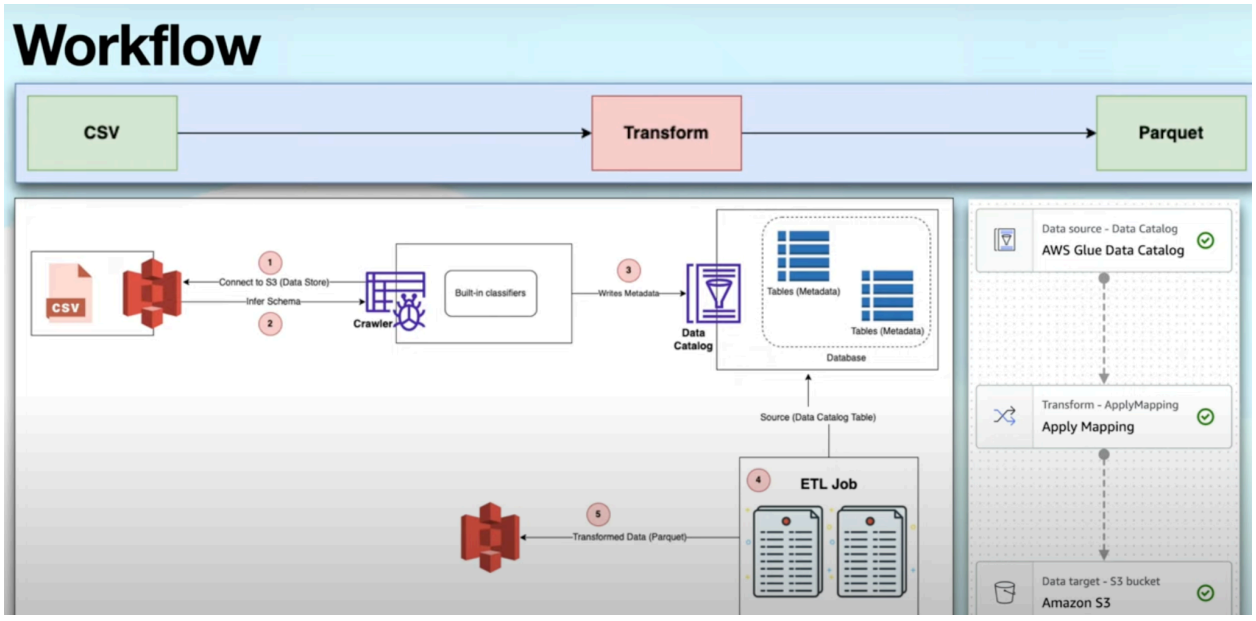


Limitations

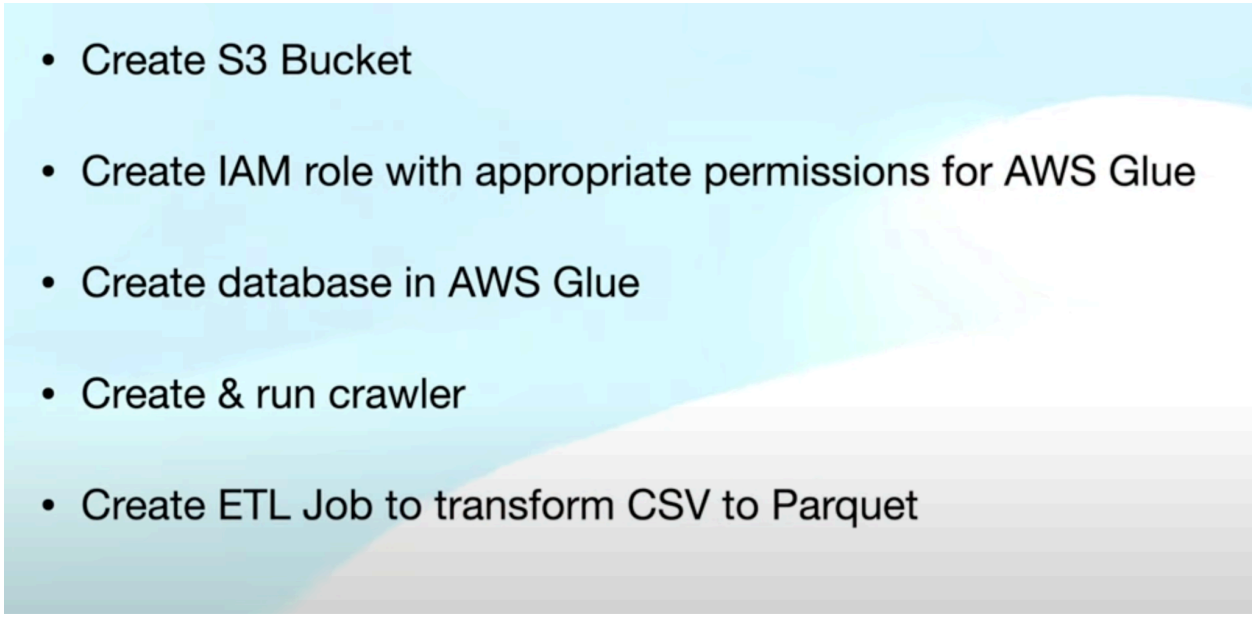
Limitations

- Limited to Python & Scala programming language
- Tightly coupled with AWS eco-system
- Dependency on Spark
- Little or no information about what's going on under the hood

CSV to Parquet :



Steps :


• Create S3 Bucket
• Create IAM role with appropriate permissions for AWS Glue
• Create database in AWS Glue
• Create & run crawler
• Create ETL Job to transform CSV to Parquet