Beginner From Zero To Hero

AWS GLUE FULL COURSE

Johnny Chivers

Tables, Crawlers, Jobs and More

TABLE OF CONTENTS

Presentation Outline

Course Overview
What is AWS Glue?
Why Do We Use AWS Glue?
Set Up Work
AWS Glue Data Catalog
AWS Glue Databases
AWS Glue Tables
Partitions in AWS
AWS Glue Crawlers
AWS Glue Connections
AWS Glue Jobs
AWS Glue Triggers
AWS Glue Dev Endpoints

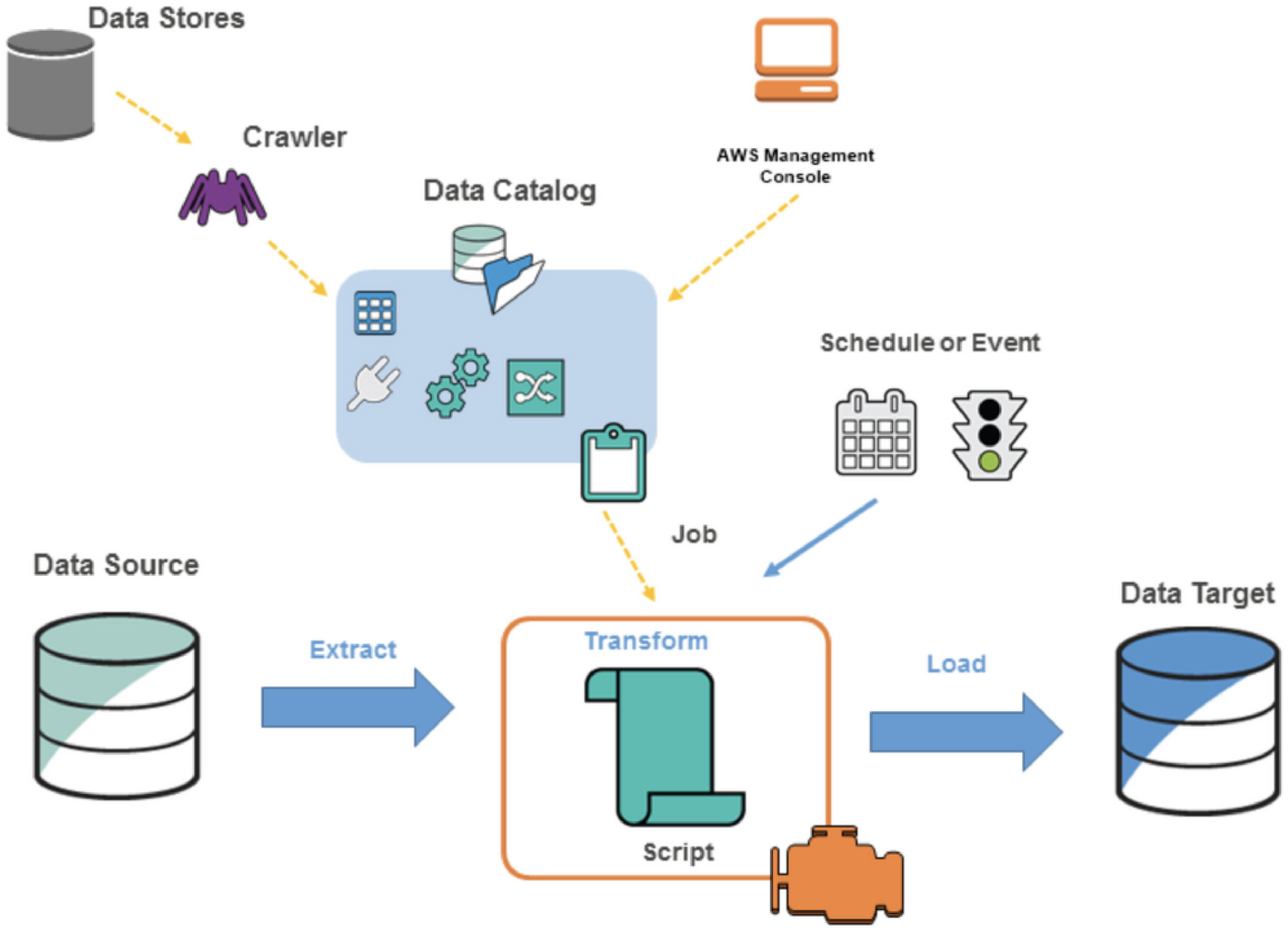
JOHNNYCHIVERS.CO.UK

Course Overview

- What is AWS Glue?
- Why Do We Use AWS Glue?
- Set Up Work
- AWS Glue Data Catalog
- AWS Glue Databases
- AWS Glue Tables
- Partitions in AWS
- AWS Glue Crawlers
- AWS Glue Connections
- AWS Glue Jobs
- AWS Glue Triggers
- AWS Glue Dev Endpoints

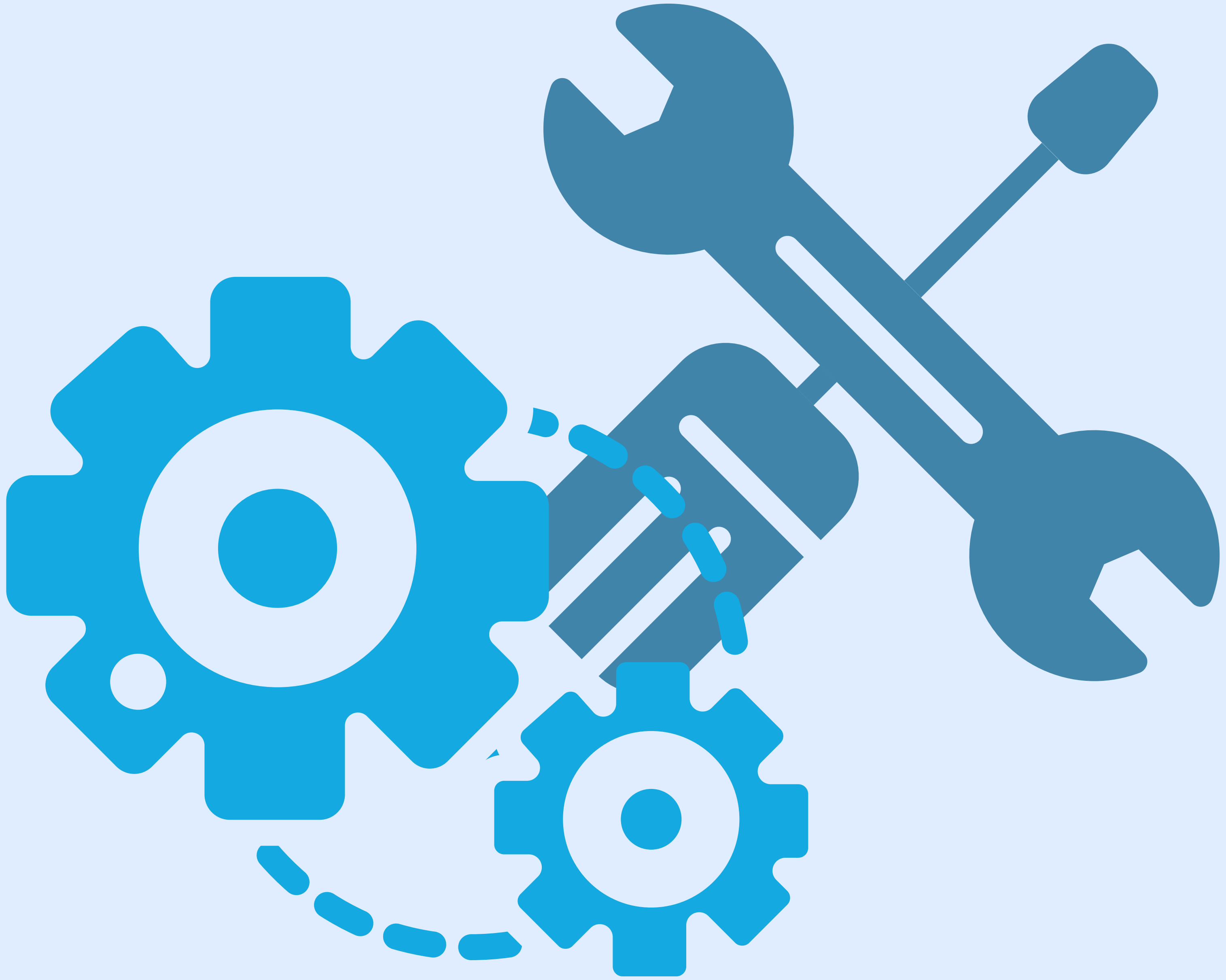
What is AWS Glue?

- Fully Managed ETL Service
- Consists of a Central Metadata Repository - Glue Data Catalog
- A Spark ETL Engine
- Flexible Scheduler



Why use AWS Glue?

AWS Glue offers a fully managed serverless ETL Tool. This removes the overhead, and barriers to entry, when there is a requirement for a ETL service in AWS.



SET UP WORK

Just A Couple Of Things

- S3 bucket
    - Data folder and sub folders
        - Upload Data
    - Script location
    - Temp Dir
- IAM role

AWS GLUE DATA CATALOG

Persistent Metadata Store

- It is a managed service that lets you store, annotate, and share metadata which can be used to query and transform data
- One AWS Glue Data Catalog per AWS region
- Identity and Access Management (IAM) policies control access
- Can be used for data governance

Data Location

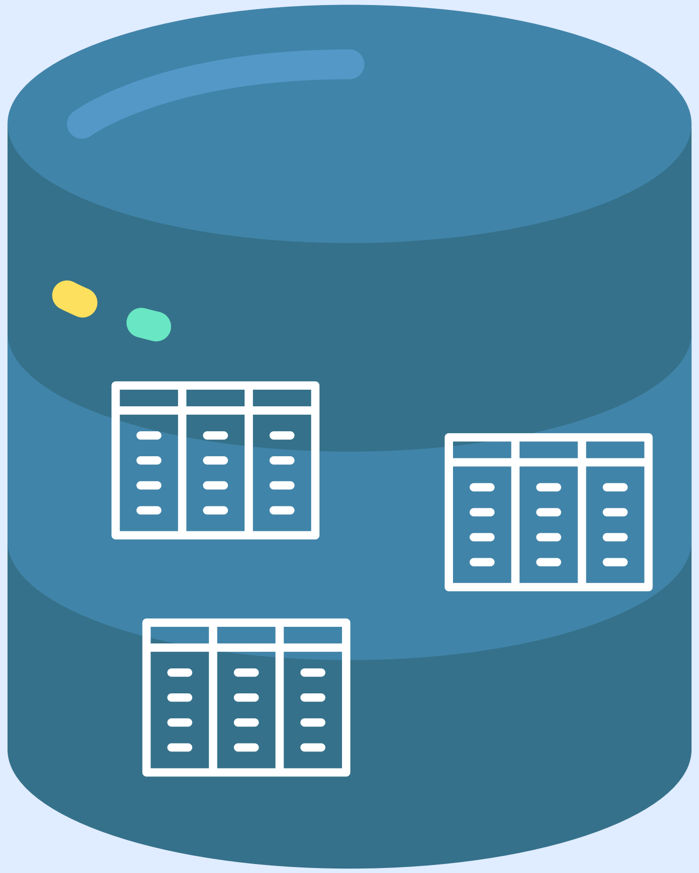
Schema

Data Types

Data Classification

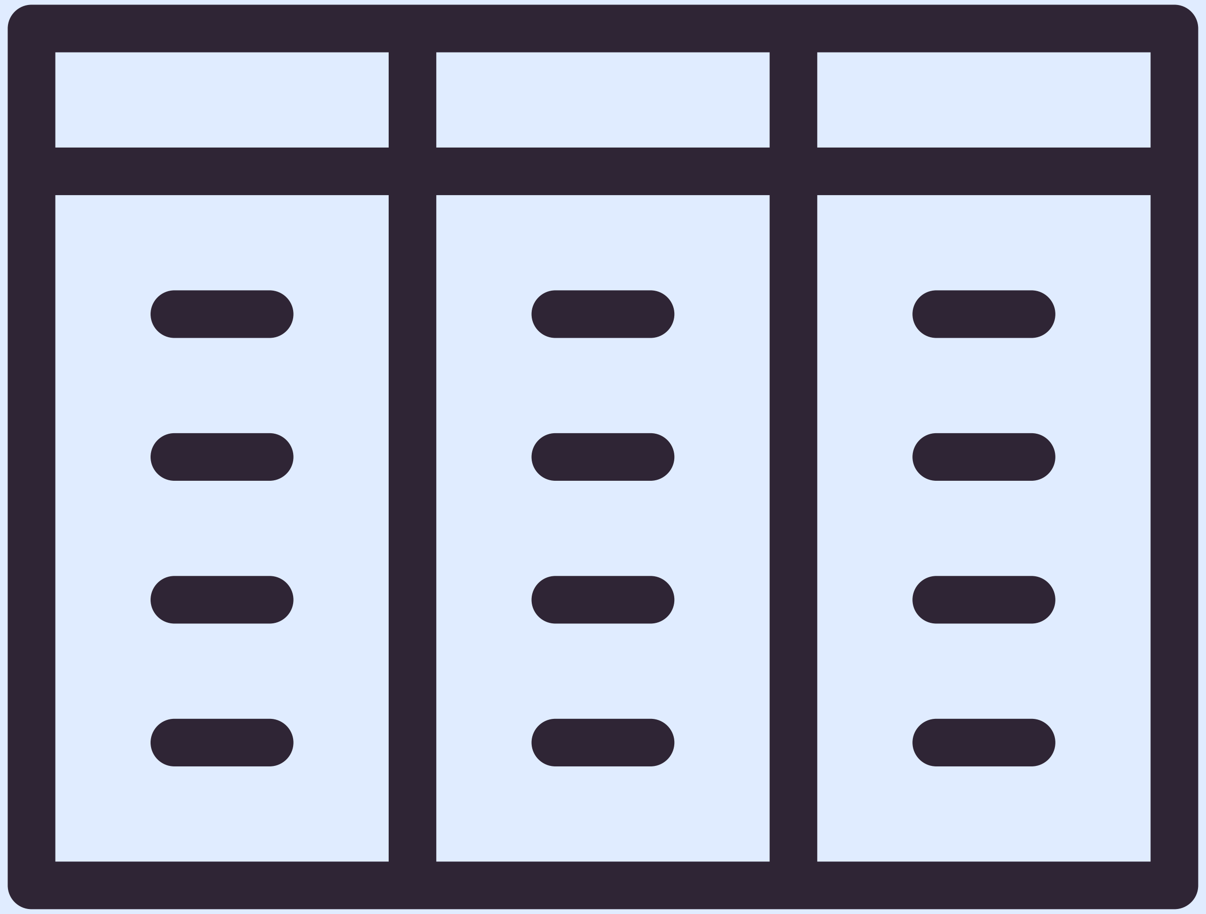
Examples of Meta Data

AWS GLUE DATABASES



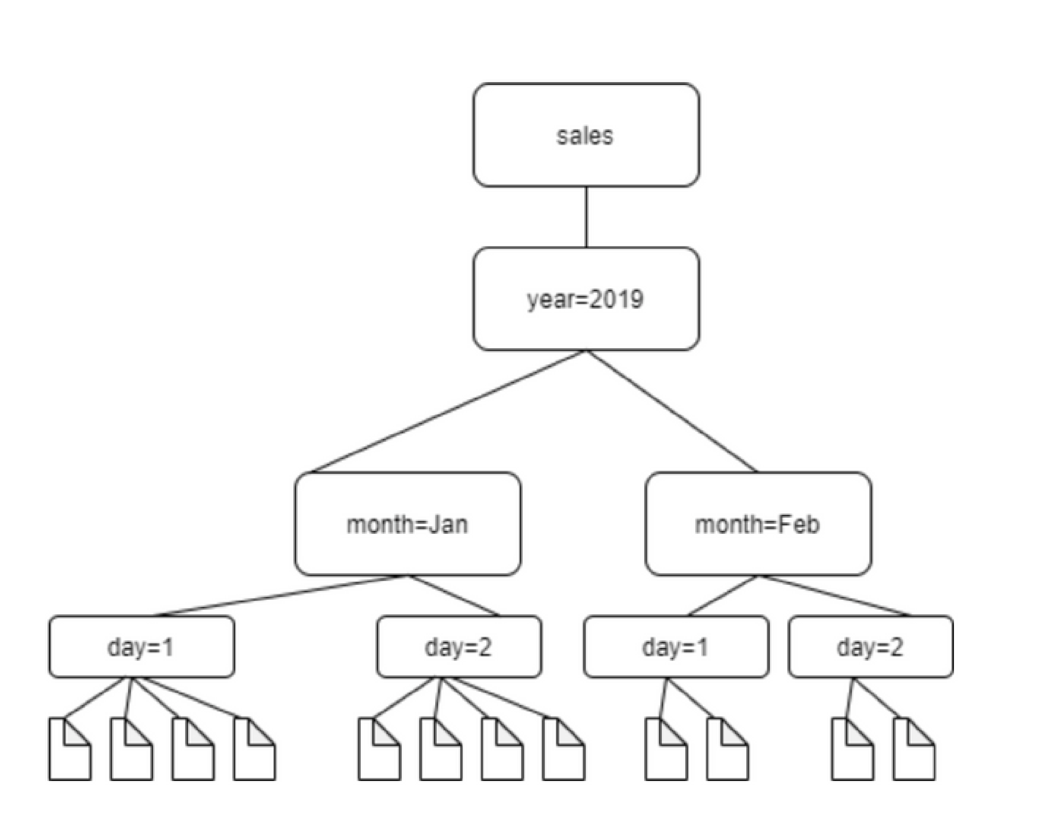
A set of associated Data Catalog table definitions organized into a logical group.

AWS GLUE TABLES



The metadata definition that represents your data. The data resides in its original store. This is just a representation of the schema.

PARTITIONS IN AWS



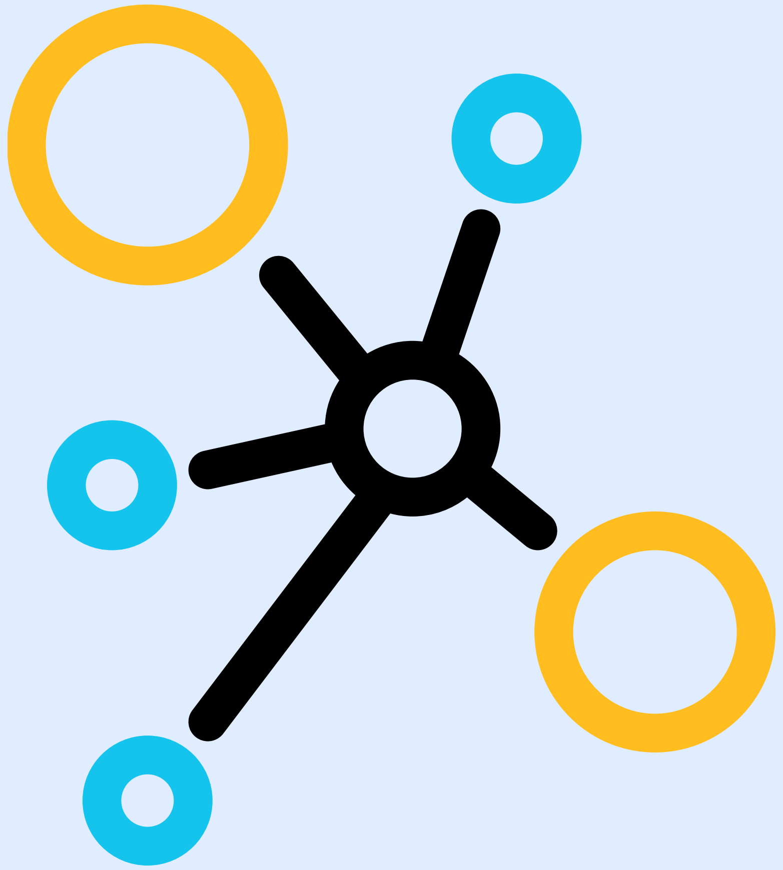
S3://sales/year=2019/month=Jan/day=1
S3://sales/year=2019/month=Jan/day=2
S3://sales/year=2019/month=Feb/day=1
S3://sales/year=2019/month=Feb/day=2

Folders where data is stored on S3, which are physical entities, are mapped to partitions, which are logical entities i.e. Columns in the Glue table.

AWS GLUE CRAWLER

A program that connects to a data store (source or target), progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in the AWS Glue Data Catalog.

AWS GLUE CONECTIONS

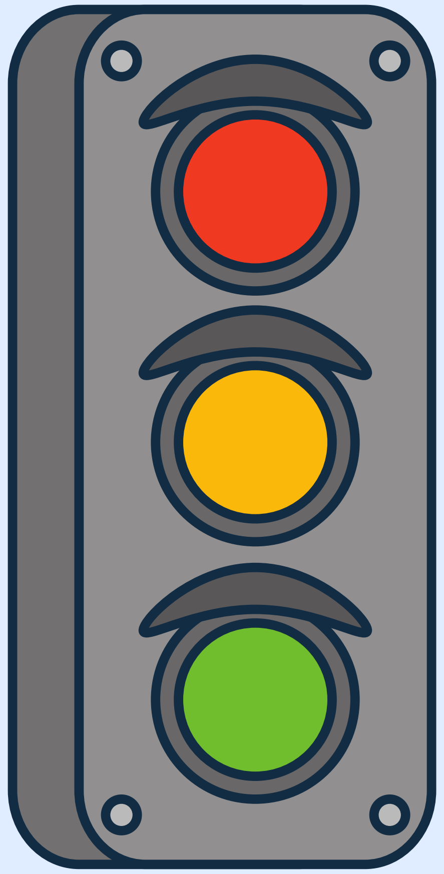


A Data Catalog object that contains the properties that are required to connect to a particular data store.

AWS GLUE JOBS

The business logic that is required to perform ETL work. It is composed of a transformation script, data sources, and data targets. Job runs are initiated by triggers that can be scheduled or triggered by events.

AWS GLUE TRIGGERS



Initiates an ETL job. Triggers can be defined based on a scheduled time or an event.

AWS GLUE DEV ENDPOINTS

A development endpoint is an environment that you can use to develop and test your AWS Glue scripts. Its essentially an abstracted cluster. NB The cost can add up.