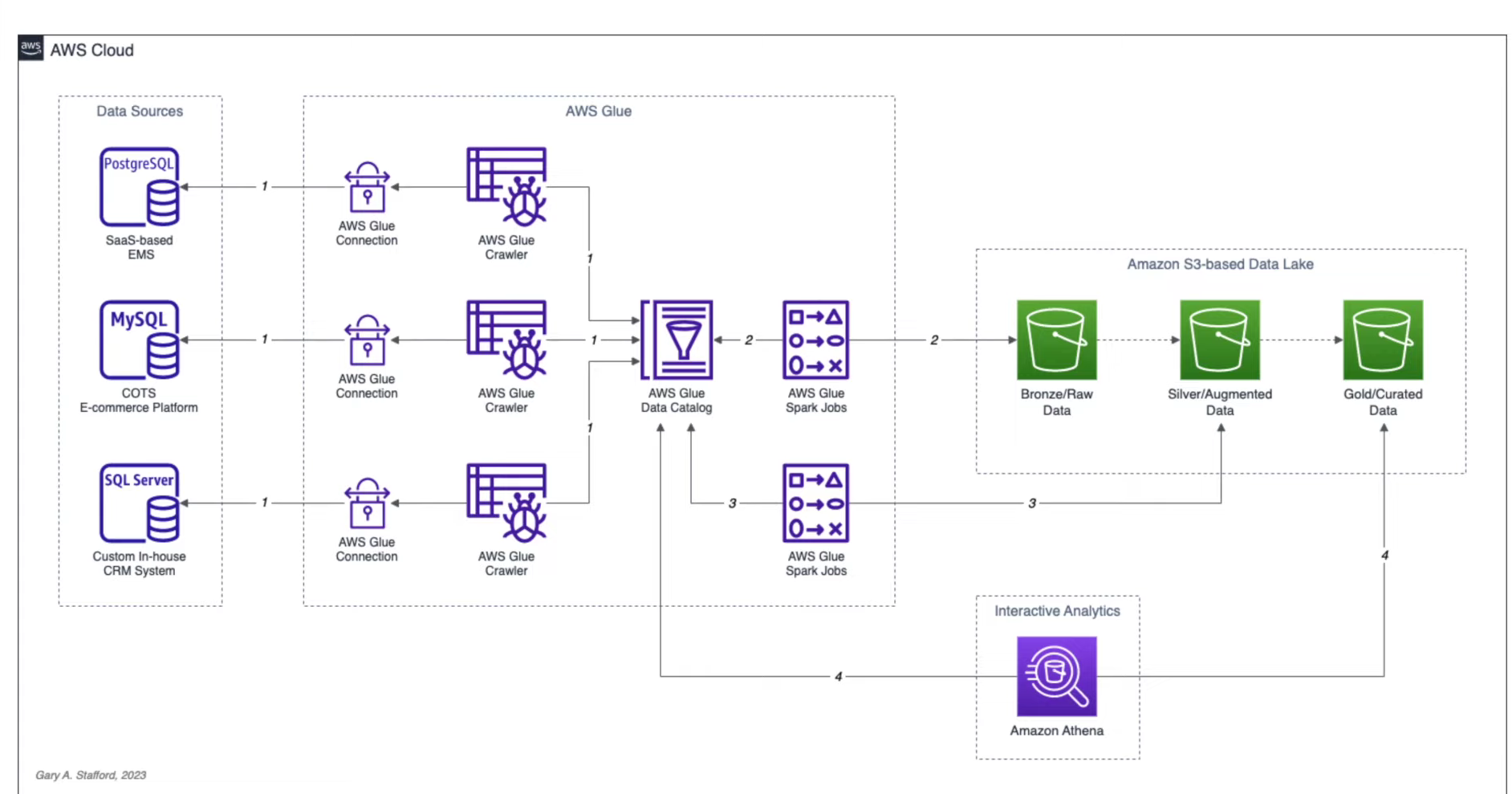
**AWS**

**Step1: Glue crawler:**

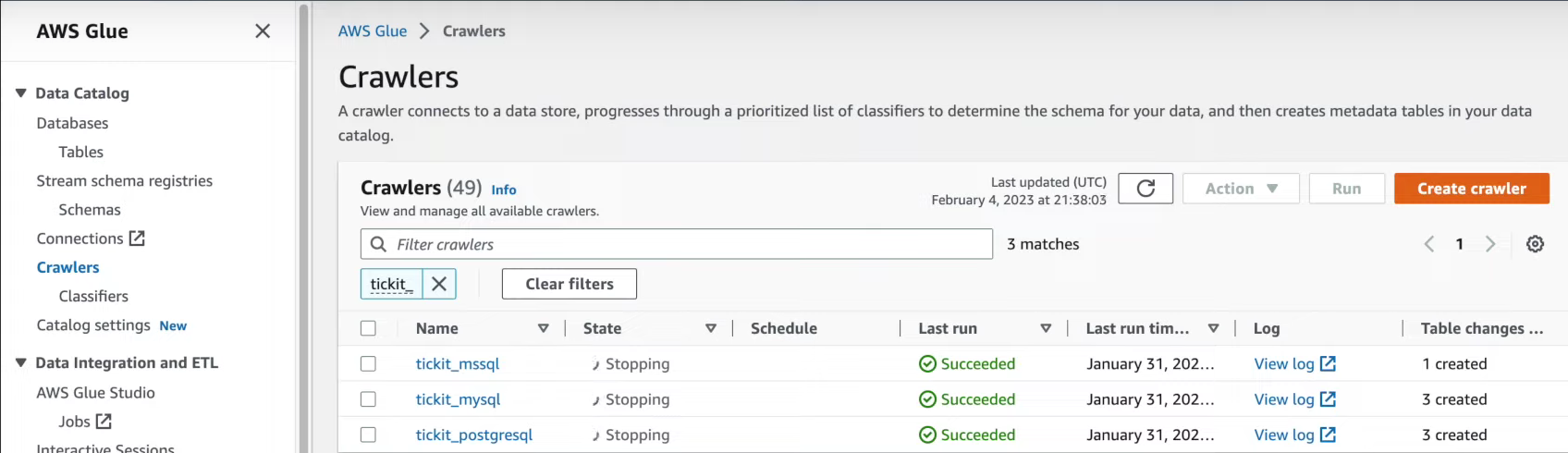
A crawler is a program that connects to a data store (source or target), progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata in the AWS Glue Data Catalog. AWS Glue crawlers can scan, classify, and extract schema information, and then store the metadata automatically in the AWS Glue Data Catalog. The metadata is stored in the form of tables. Crawlers can also detect schema changes and version the tables for previously crawled data stores and detect hive-style partitions on Amazon S3. For previously crawled Amazon S3 partitioned data, the crawler will add new partitions or update the changed partitions by default. Crawlers run on demand or on a schedule to suit your needs.

Create 3 Glue crawler that can connect to 3 different databases from source (MySQL, SQL server, sPostgreSQL)

aws glue start-crawler –name ticket\_postgresql;

aws glue start-crawler –name ticket\_mysql;

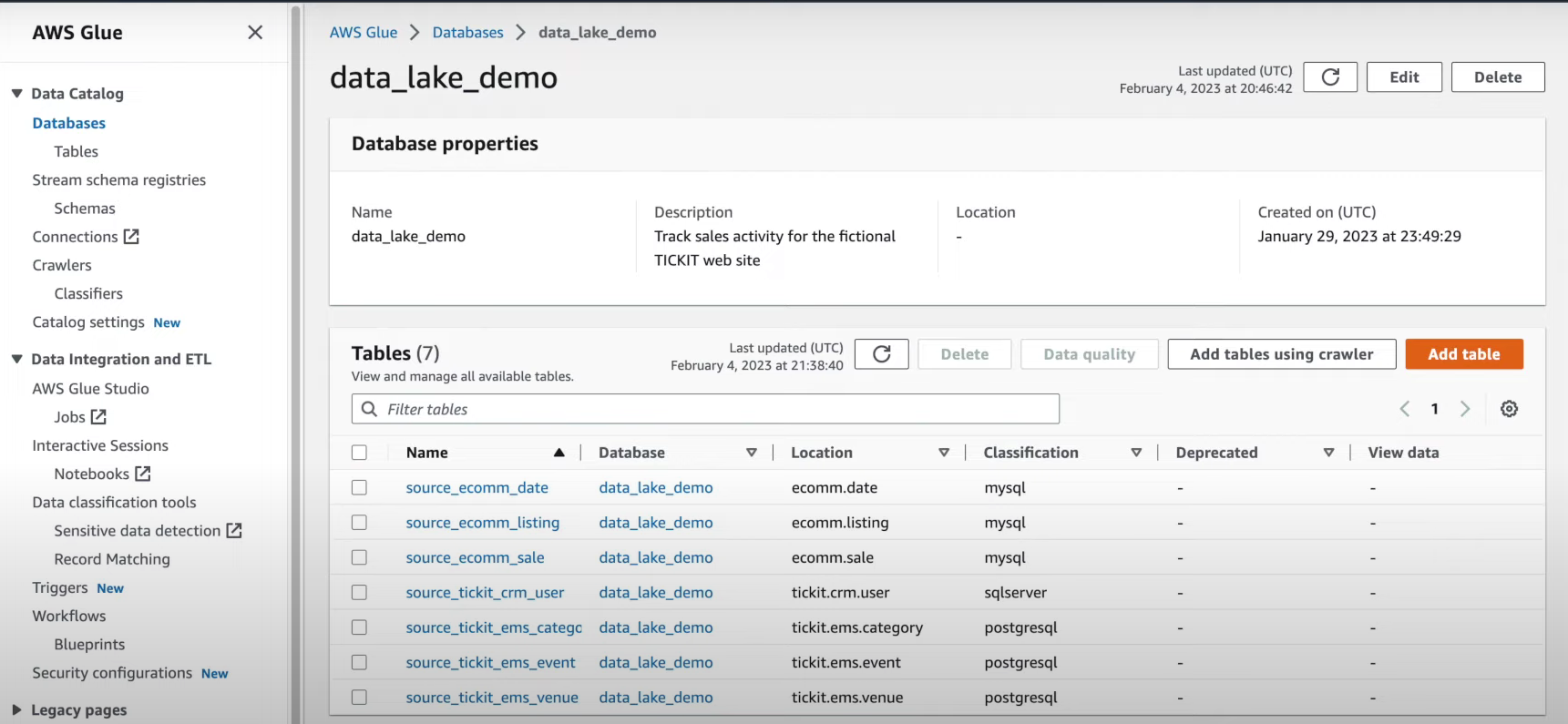
aws glue start-crawler –name ticket\_mssql



**Glue data catalog:**

The AWS Glue Data Catalog is a persistent metadata store in AWS Glue. The Data Catalog stores, annotates, and shares metadata in the AWS cloud similar to Apache Hive metastore. AWS Glue Data Catalog provides a uniform repository where disparate systems can store and find the metadata that keeps track of data in data silos, and use that metadata to query and transform the data. The Data Catalog is integrated with services like Amazon S3, Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. It allows you to search for key data attributes across all the datasets inside the catalog.

The above command will create meta data in Glue data catalog for the 7 tables from different database.



**Step3: Glue Studio**

A job is the business logic that performs the ETL work in AWS Glue. When you start a job, AWS Glue runs a script that extracts data from sources, transforms the data, and loads it into targets.