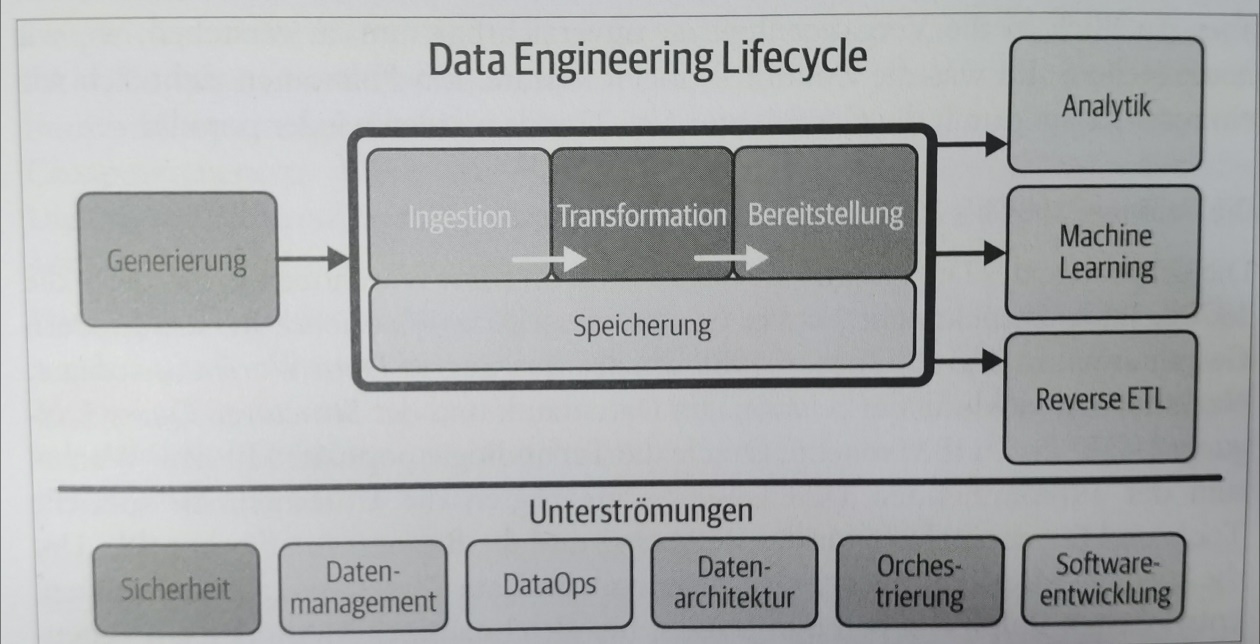
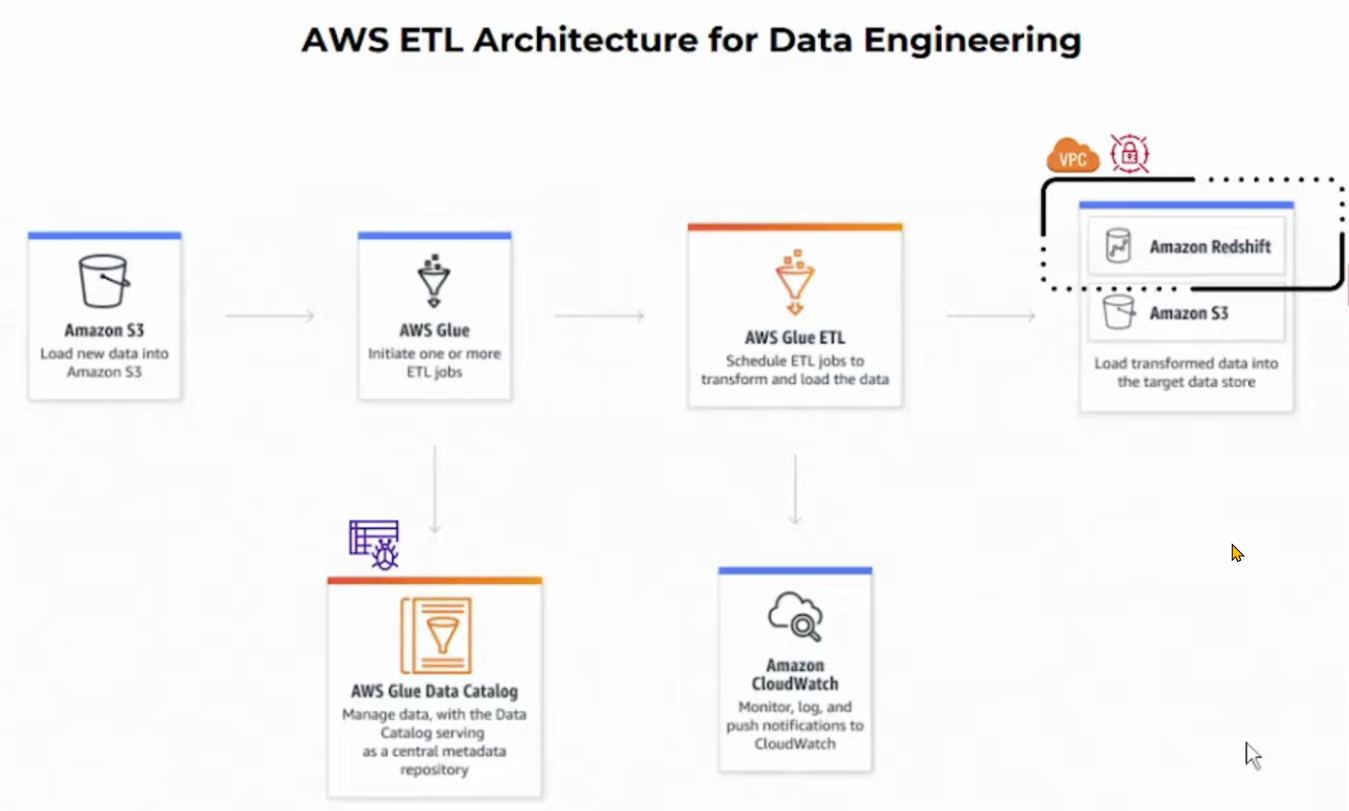
Project by Gero Krikawa   
AWS Cloud, Datapipelines, Apache Spark, Pyspark, Jupyter Notebook, Redshift **End to End Data Engineering Project in AWS using Spark (Pyspark)**

**Overview: Data Engineering Lifecycle**



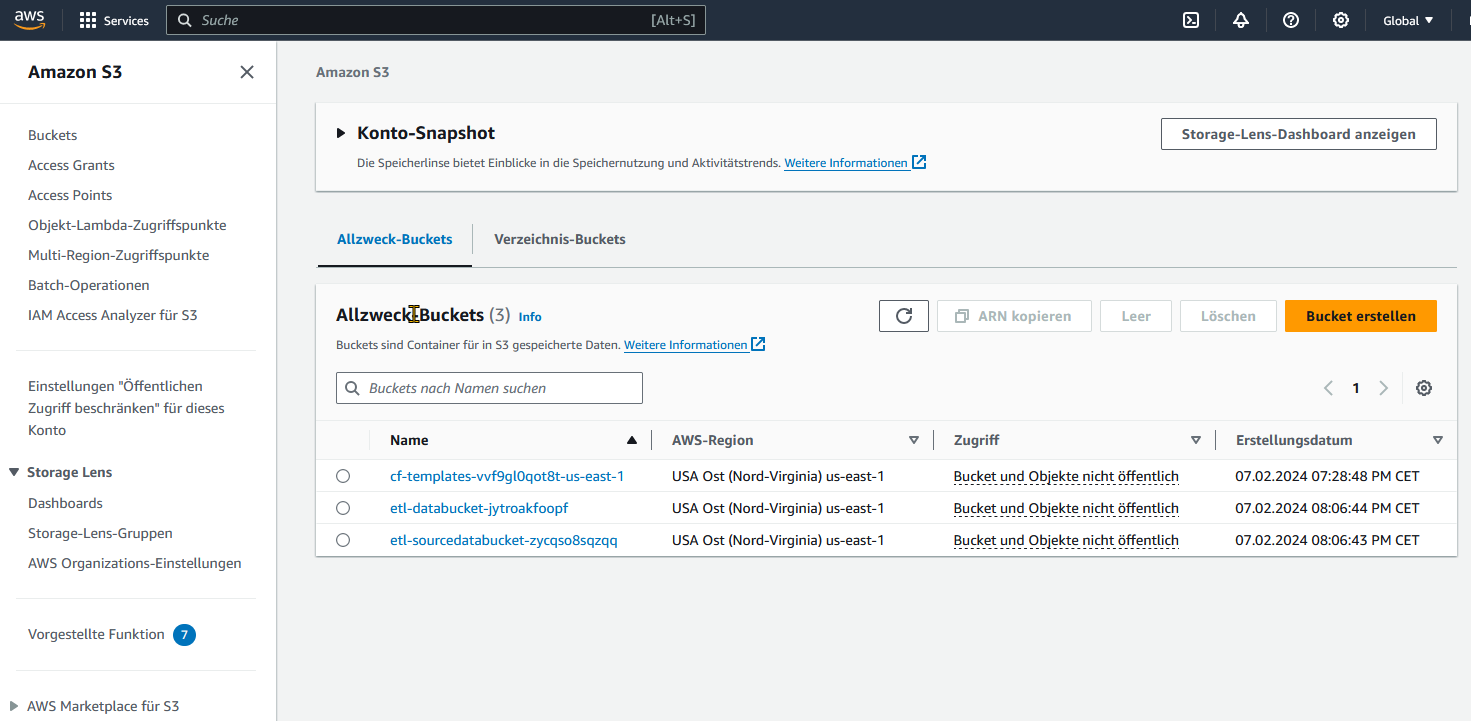
**AWS ETL Architektur for Data Engineering**

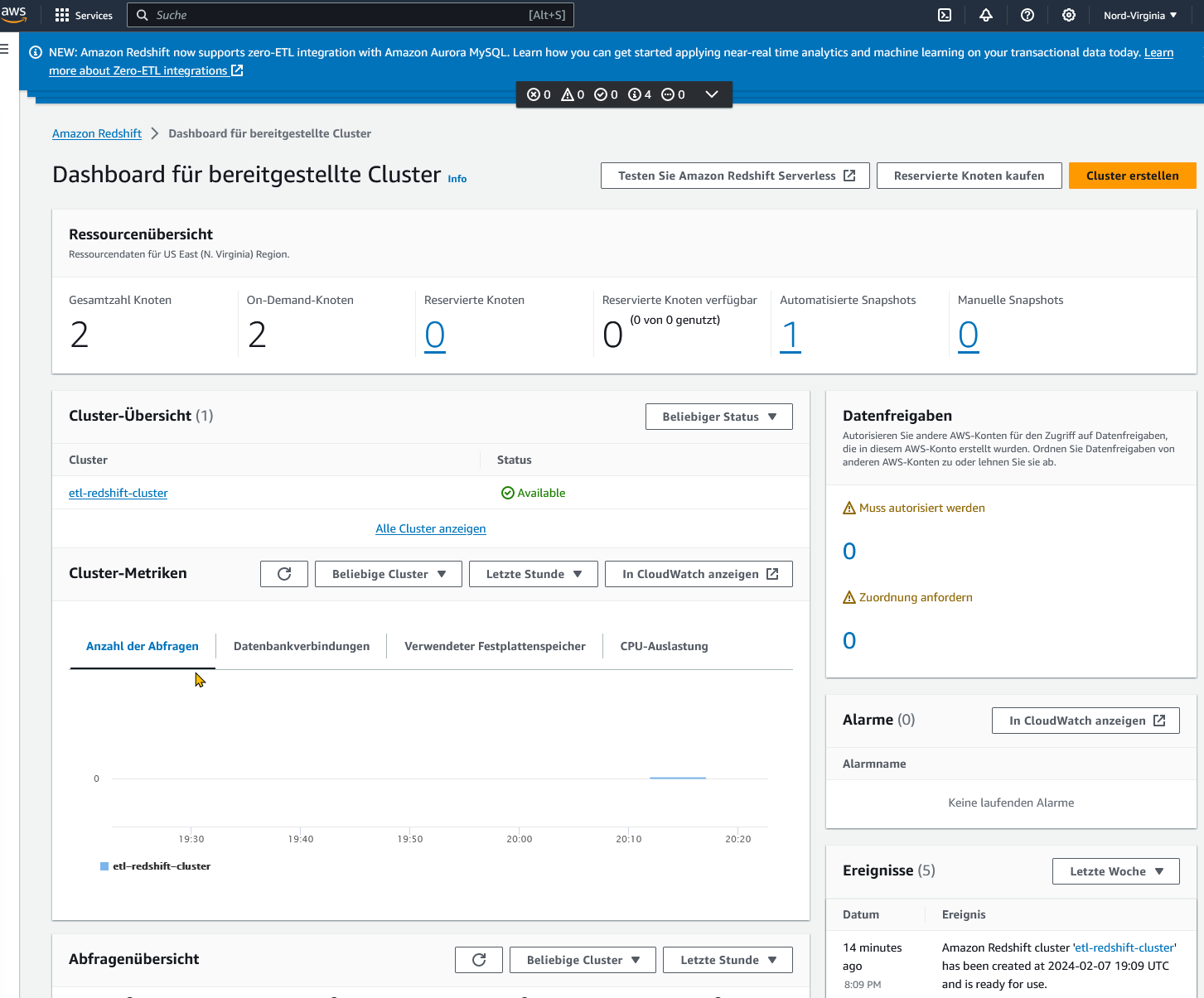


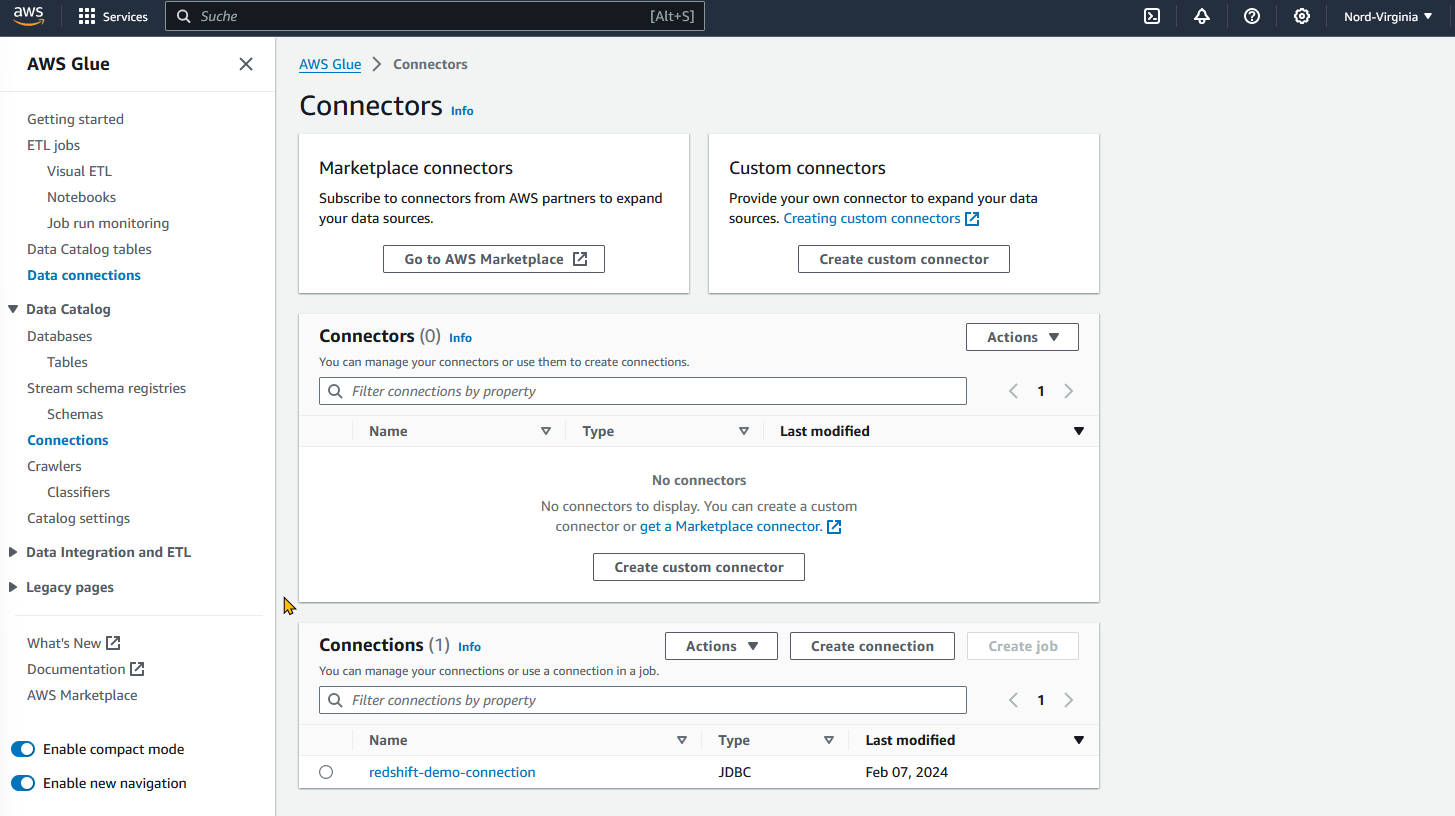
**Project Workflow**

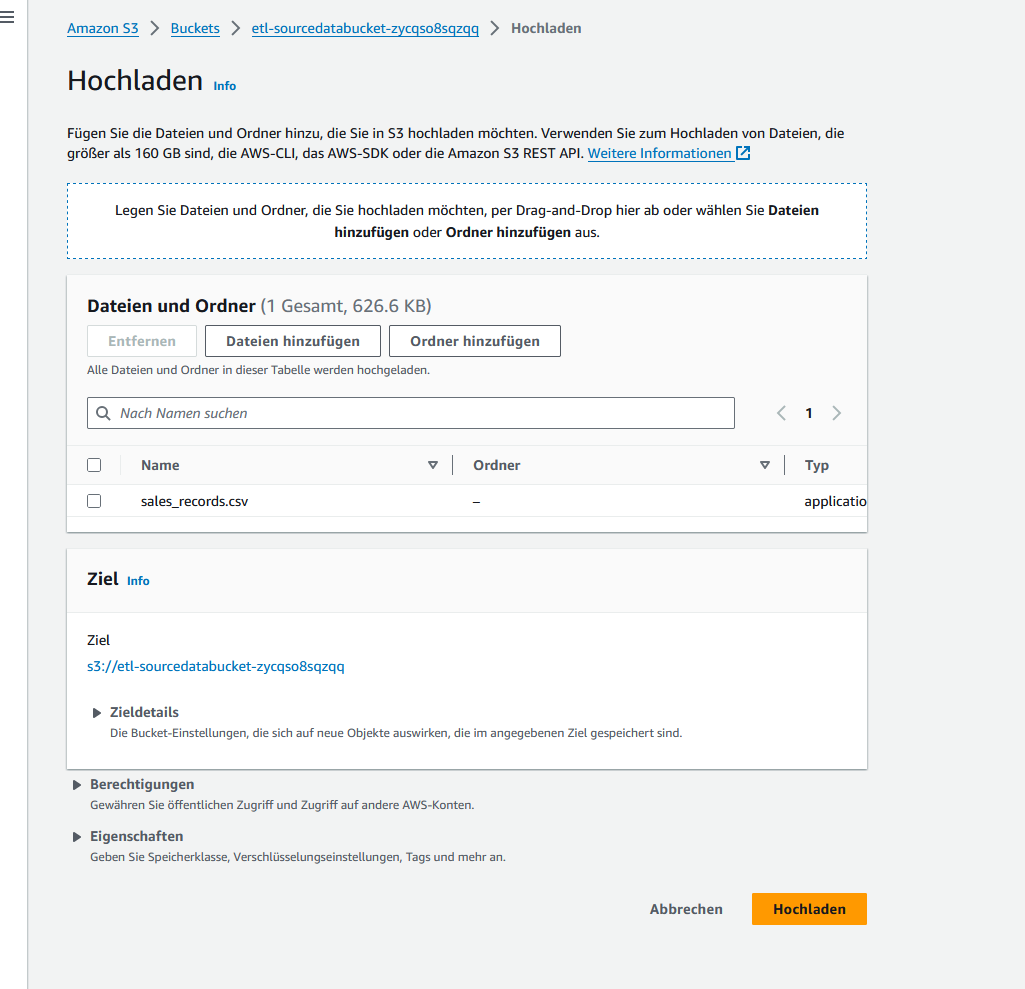
* Create Data Engineering System in AWS using „Infrastructure as Code“
* End to End Data Pipeline
  + S3 Storage & Source of Data Pipeline
  + AWS Clue – using crawler to catalog data
  + Processing data using Pyspark within interactive Jupyter Notebook in Glue
    - Build Data Pipeline in Pyspark using Glue Jupyter interactive Notebook
    - Reading data from s3 storage, processing it in Spark and then loading it into Redshift (using dynamic frames and spark data frames

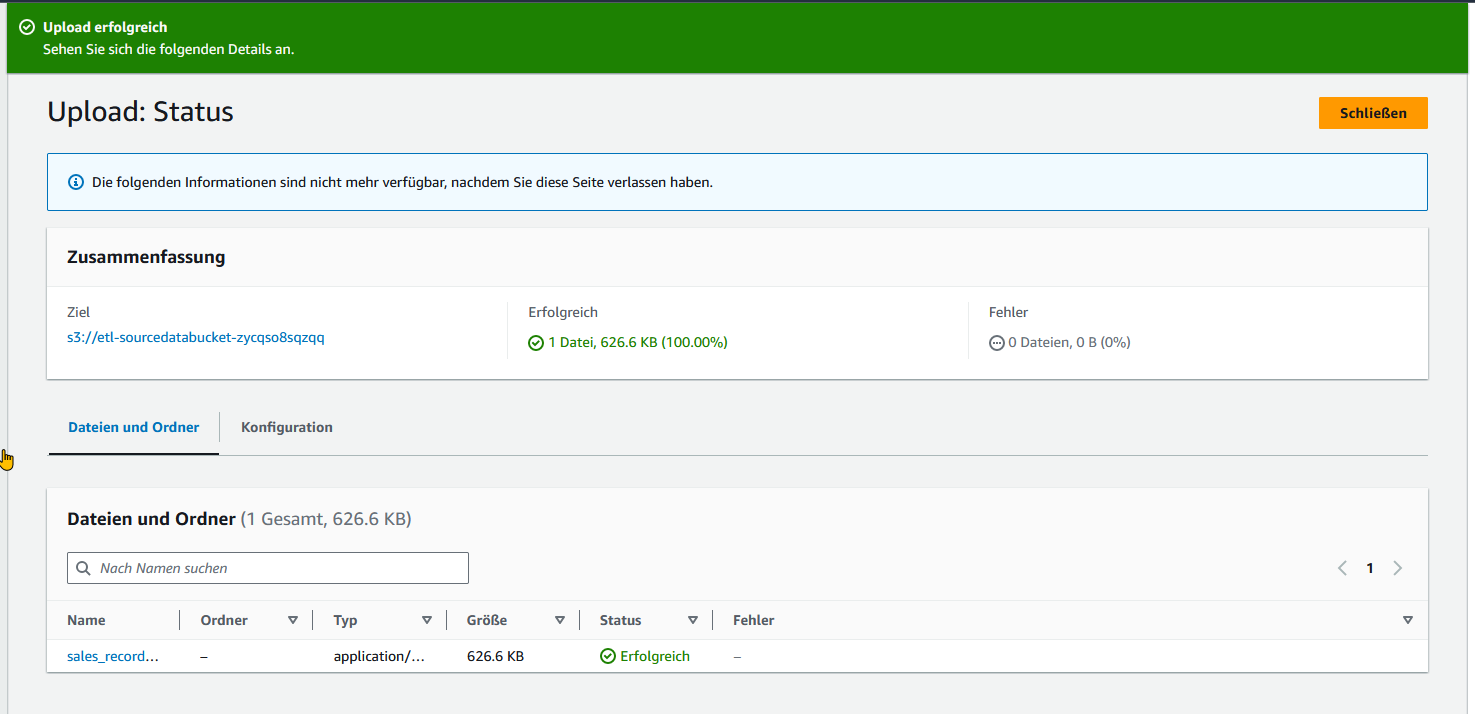
**Screenshots and explanation during development**

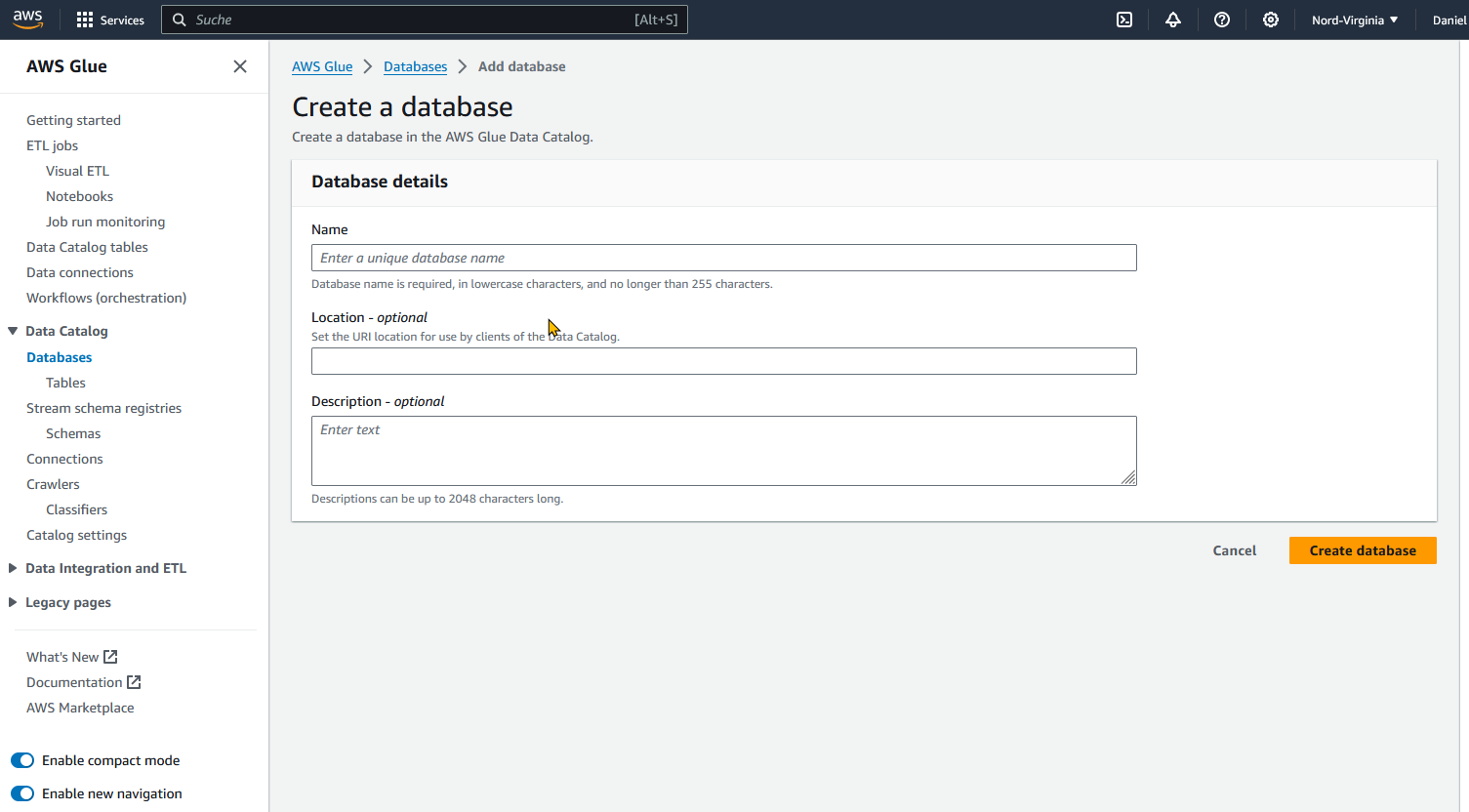


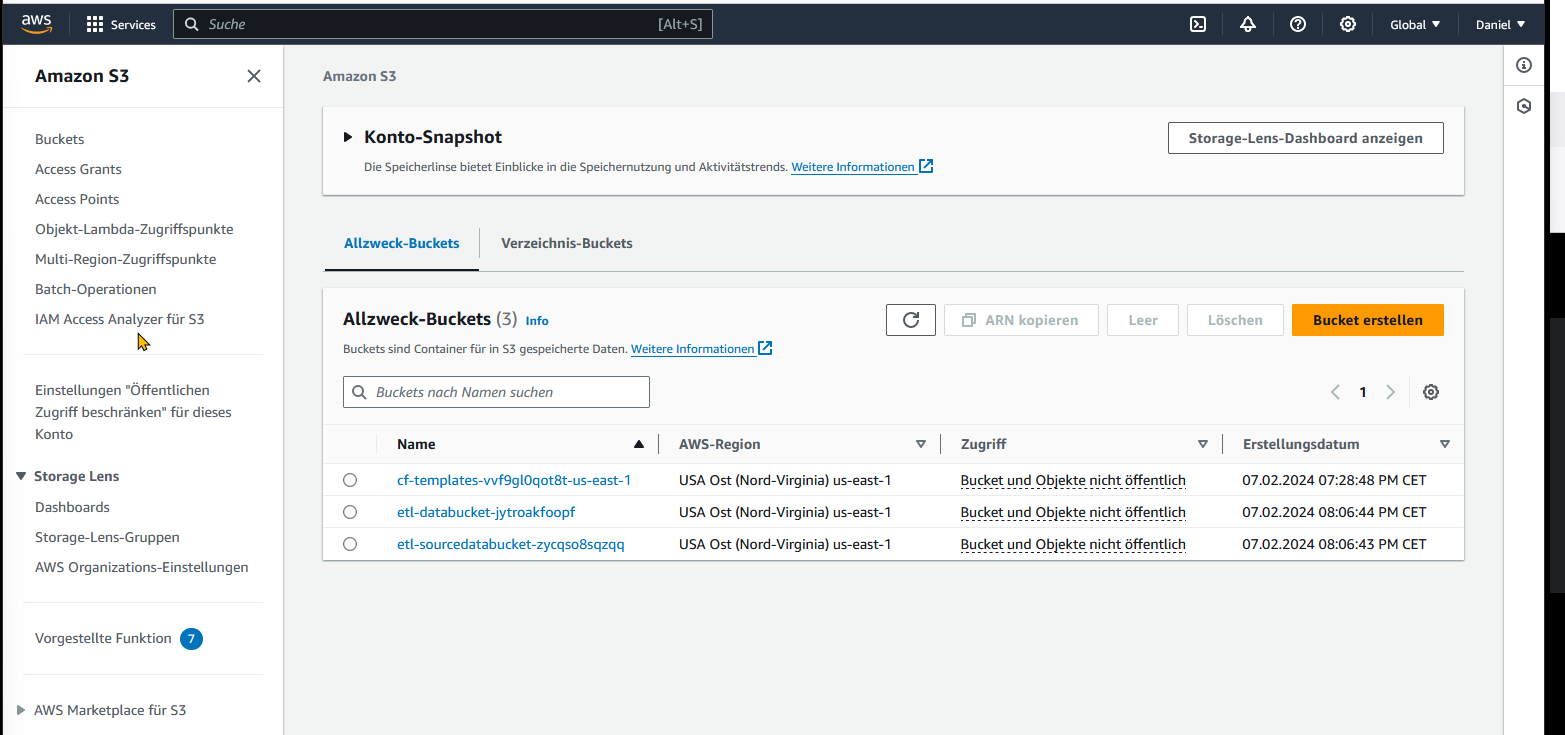


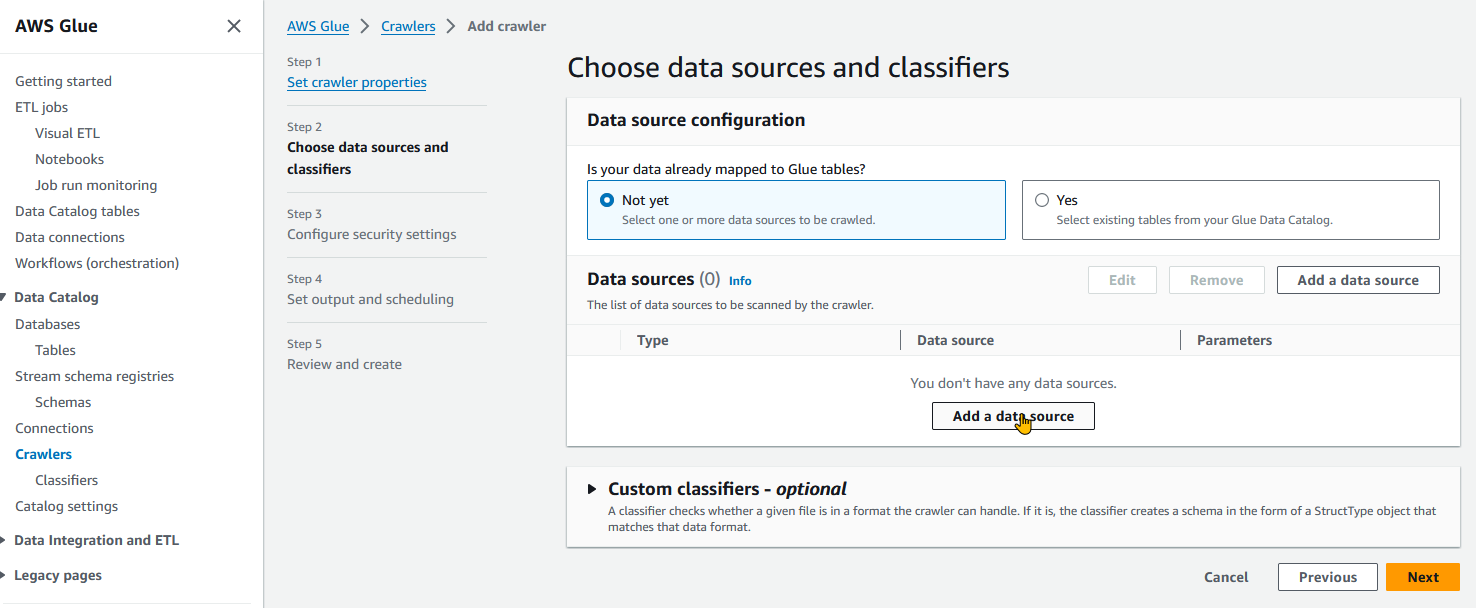


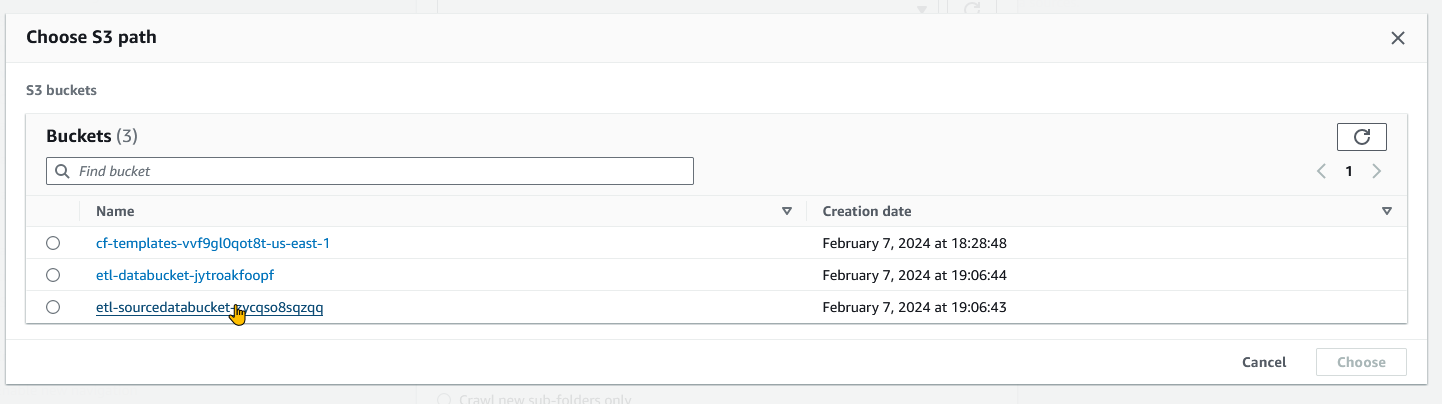


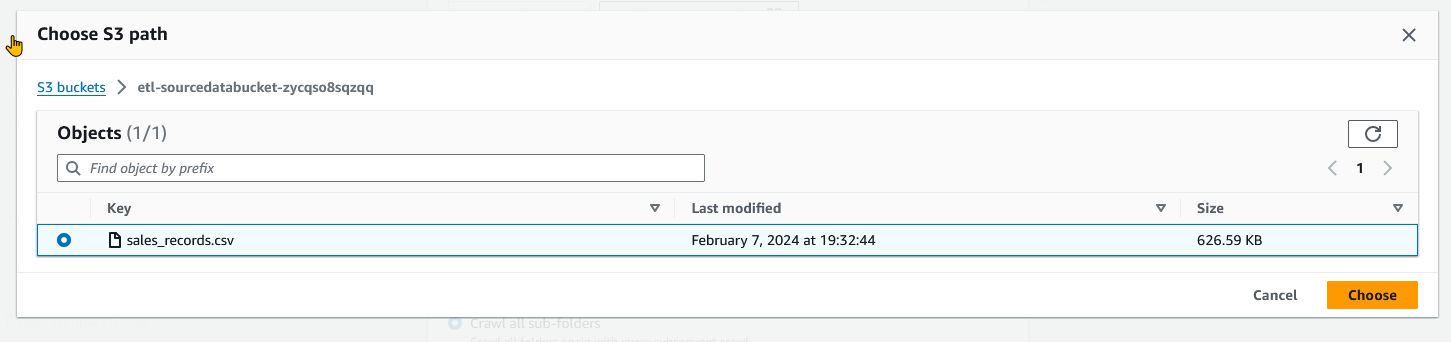


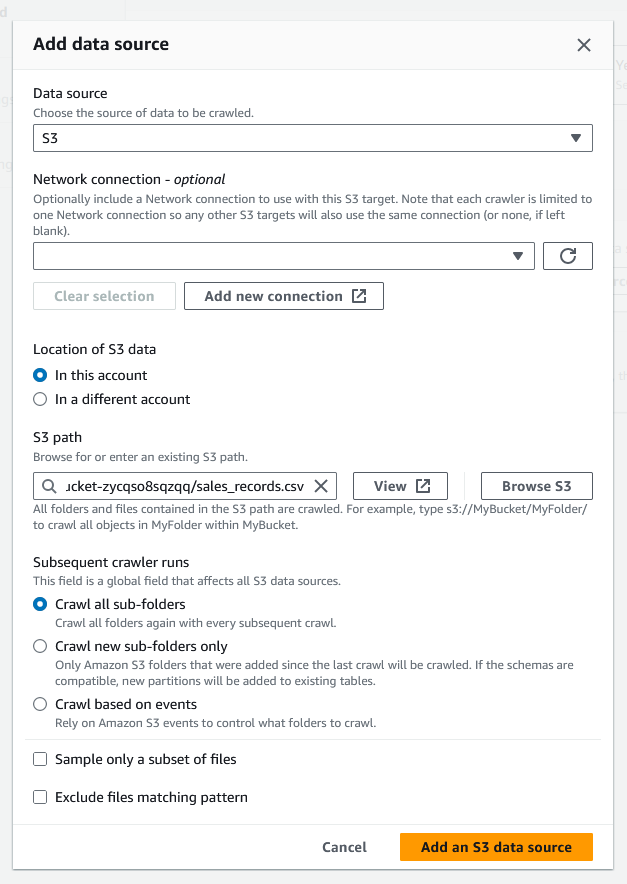


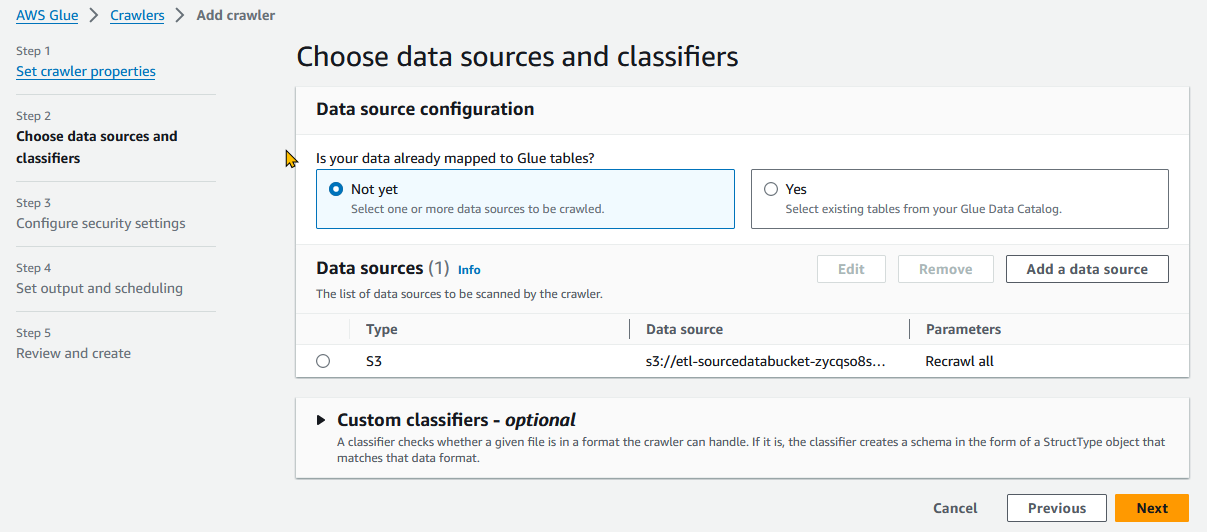


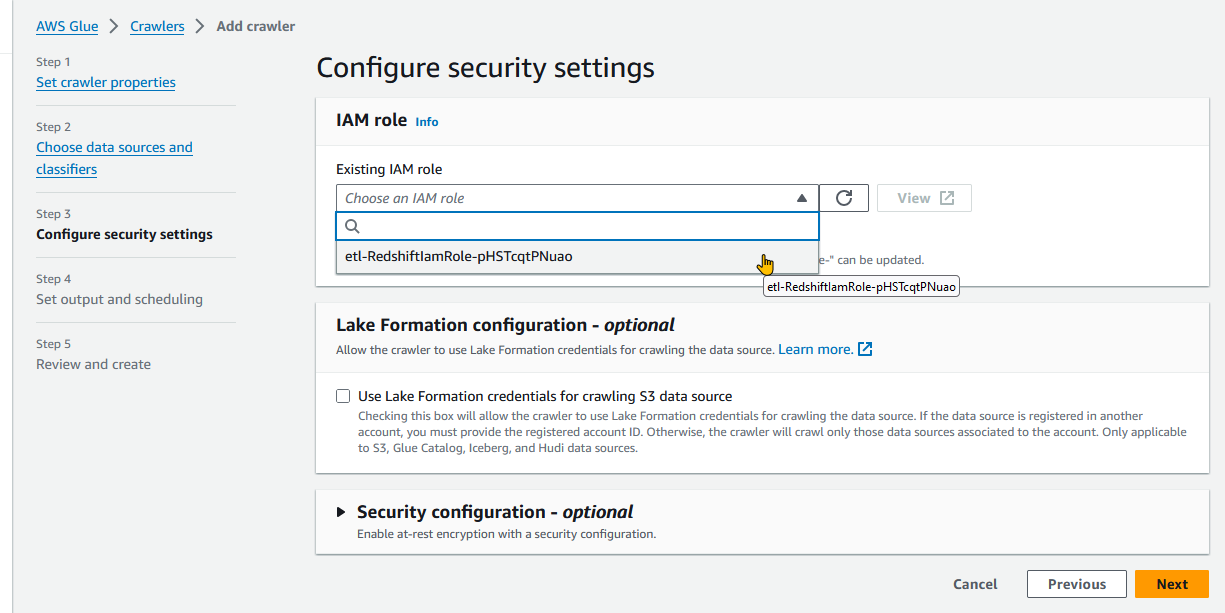


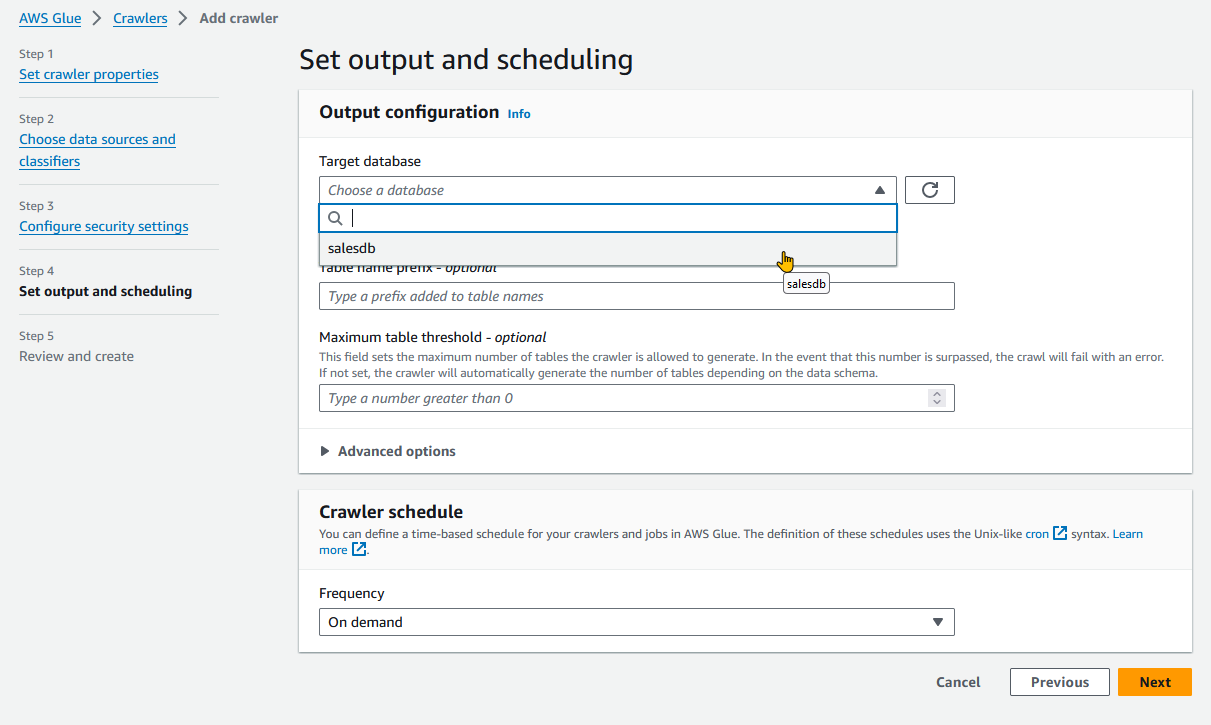


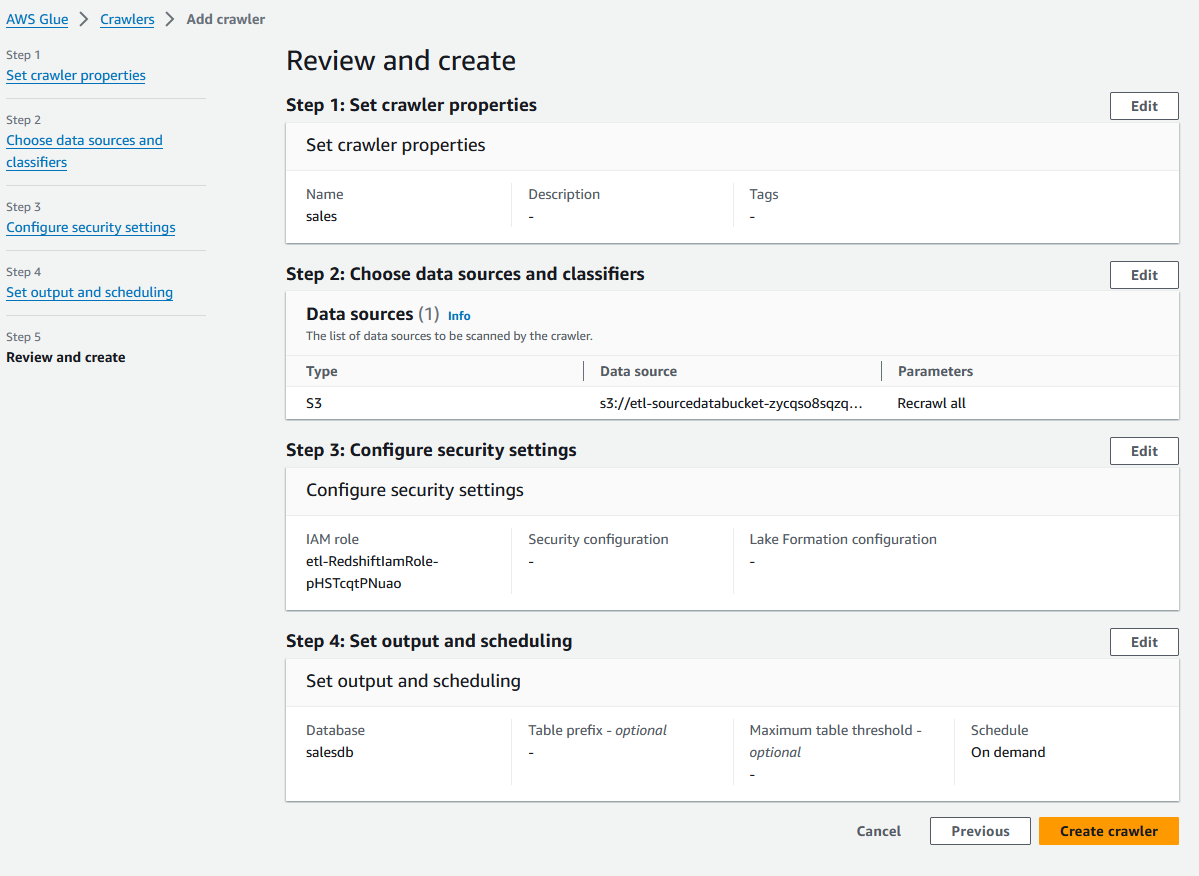


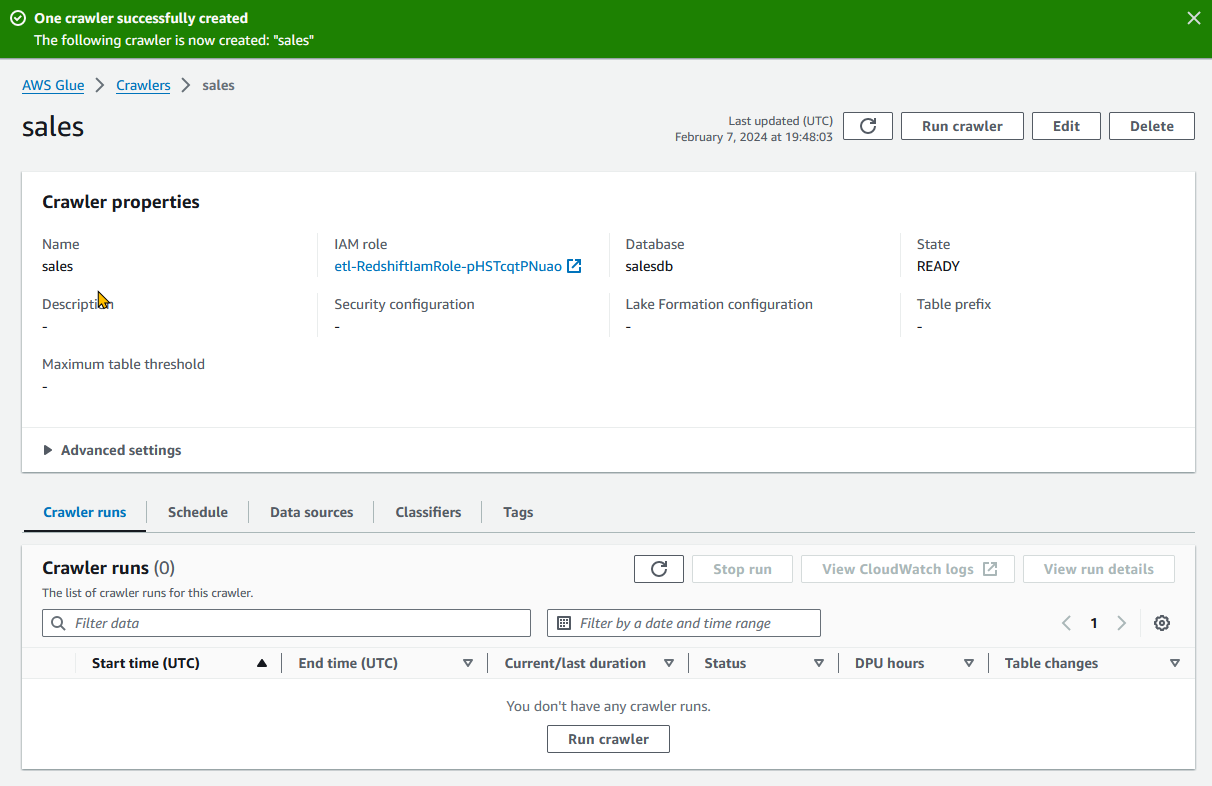


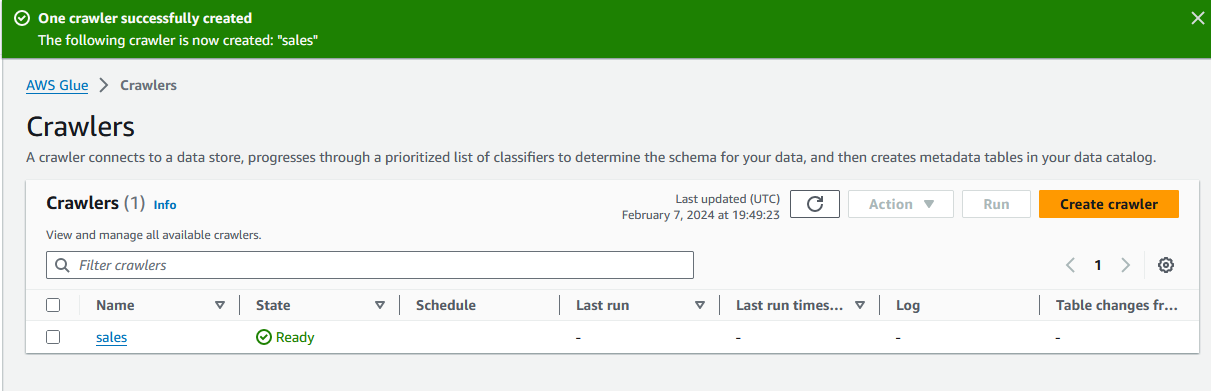


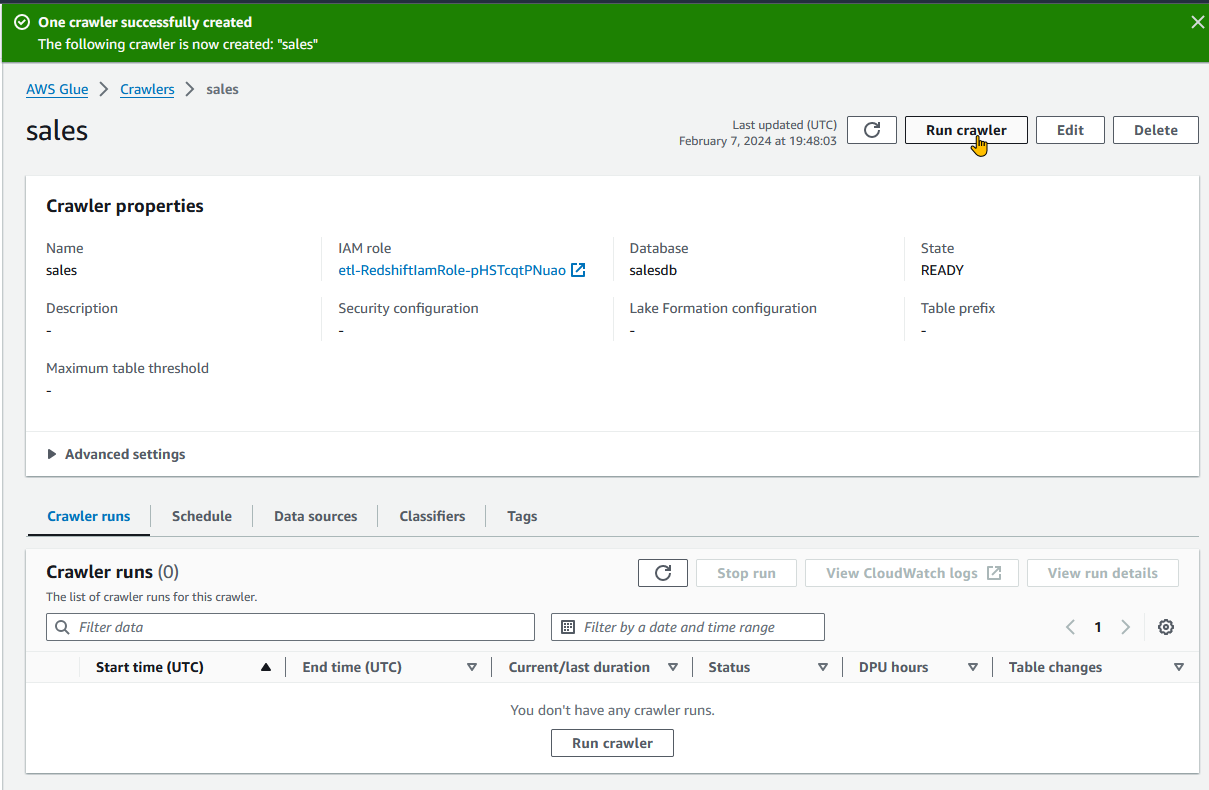


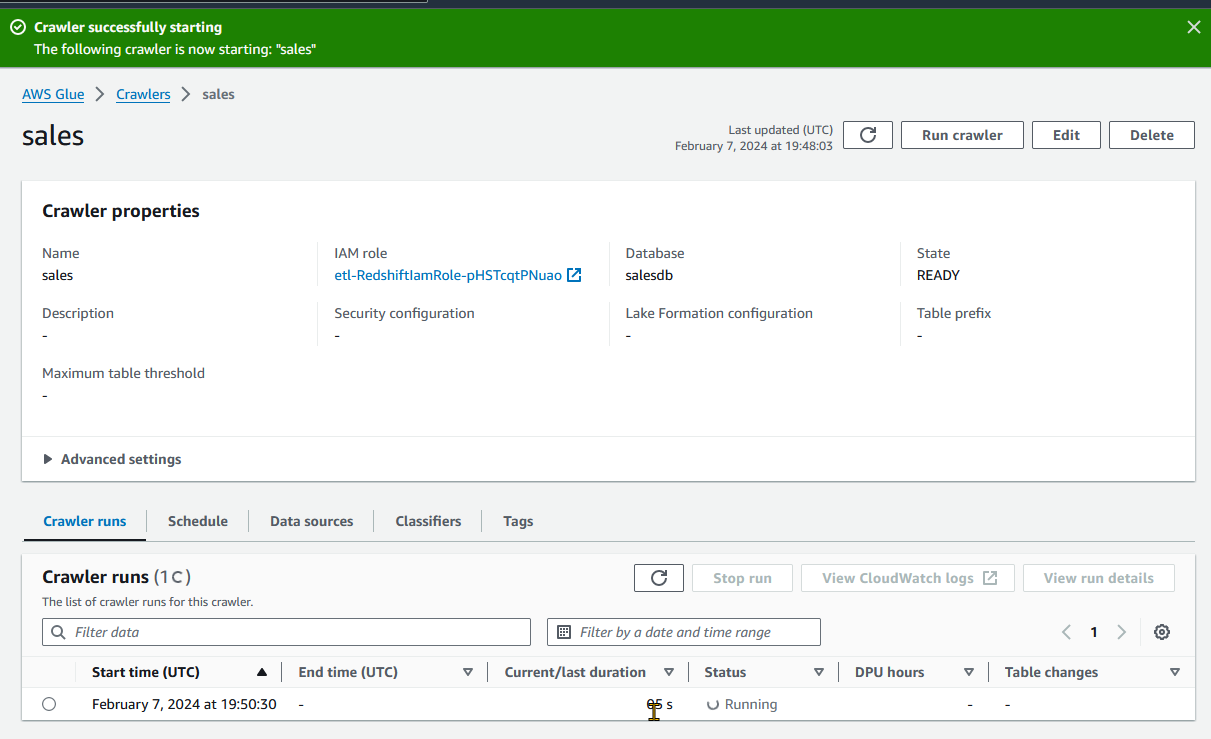


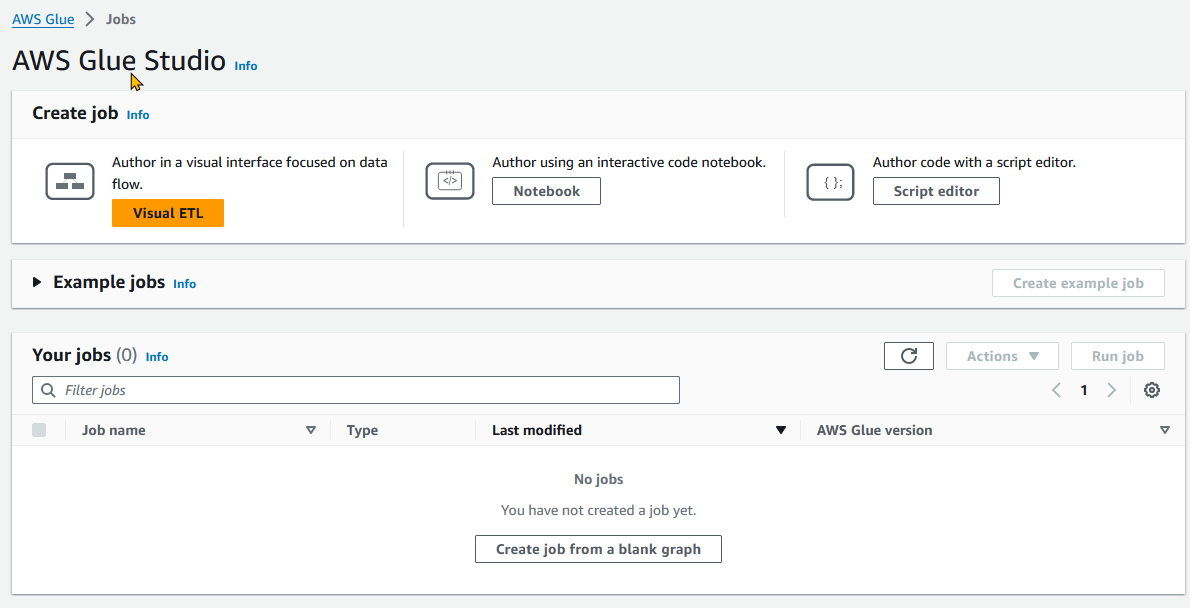


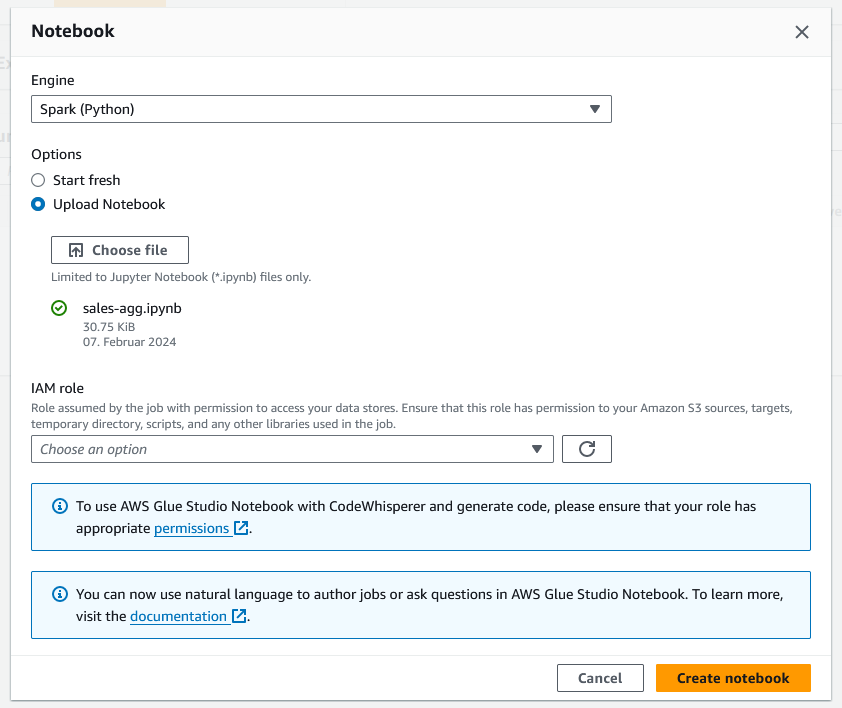


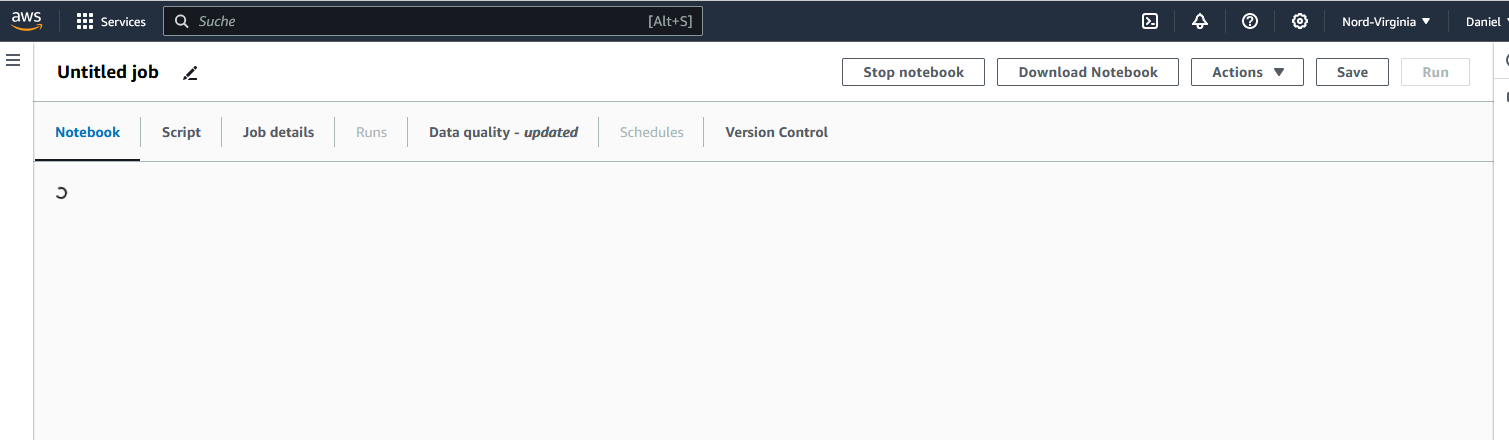


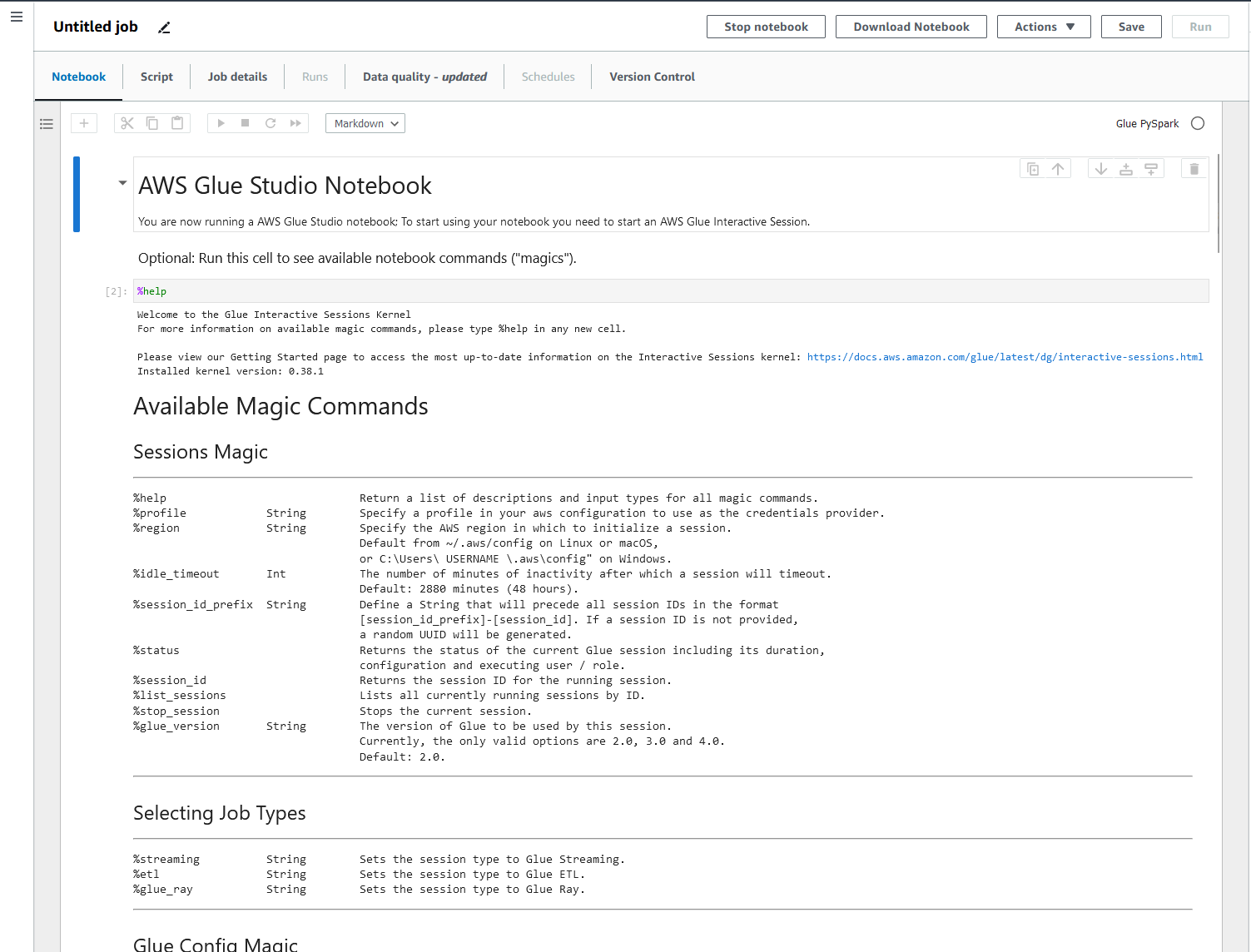


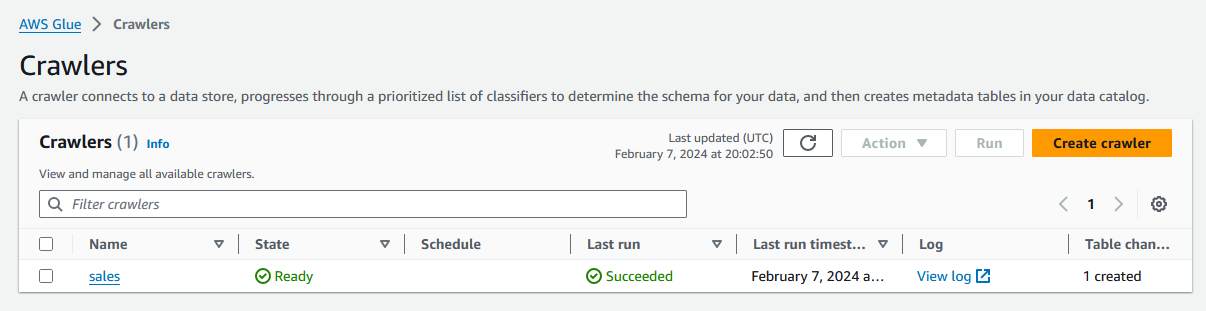






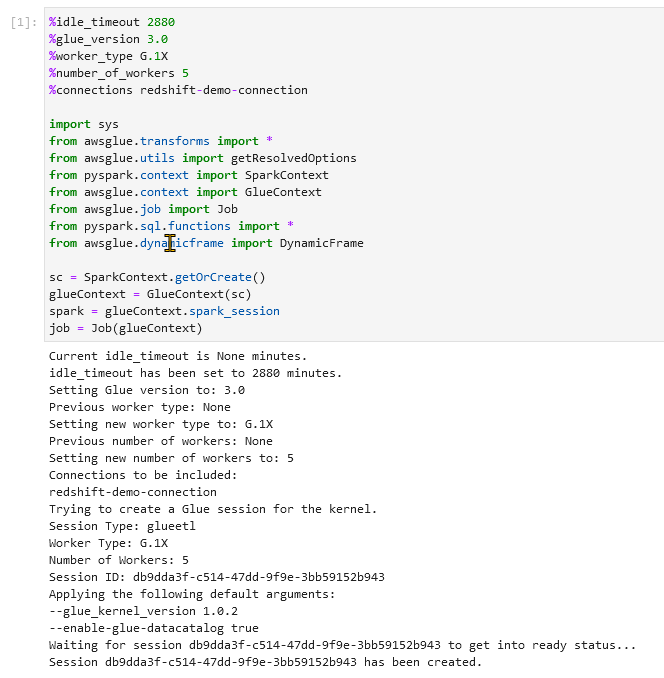


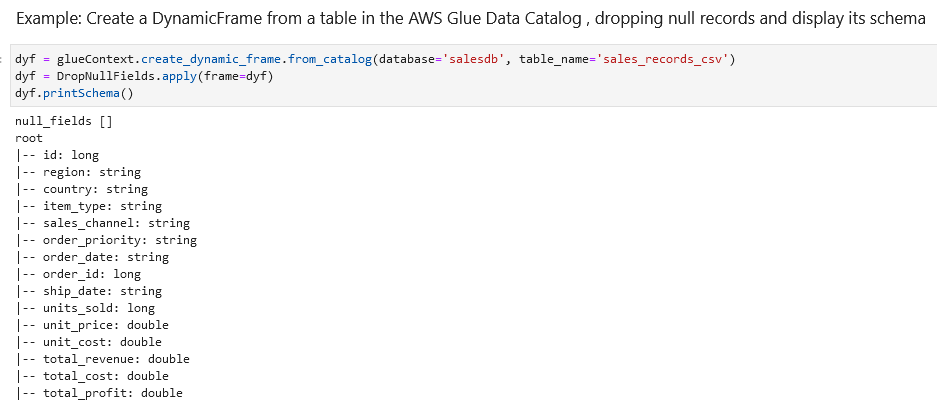




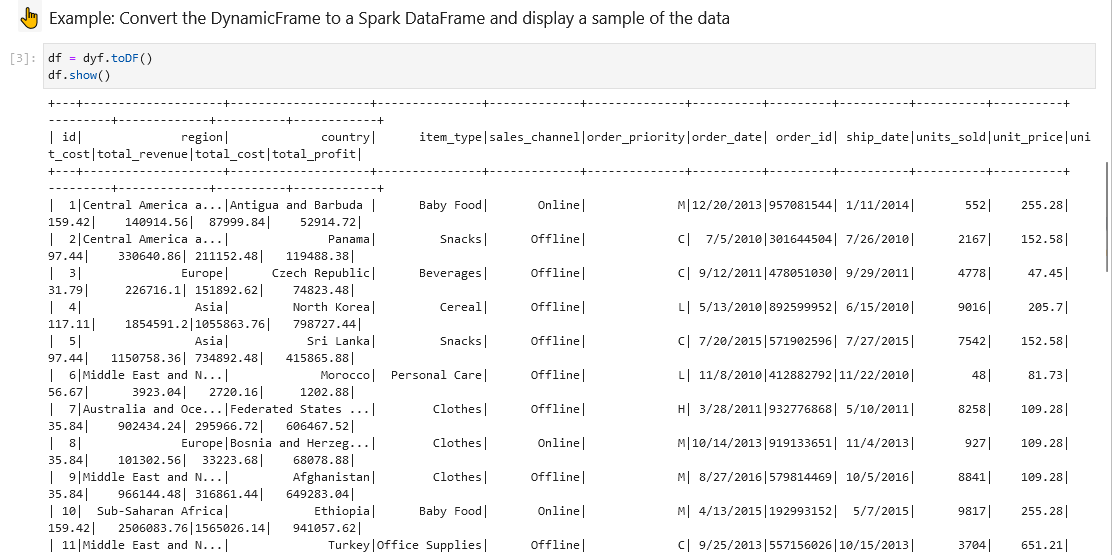
Jupyter Notebook for ETL Process using Spark / Pyspark

Import Libaries & initializing Spark and glue context (main entry point for AWS glue ETL)

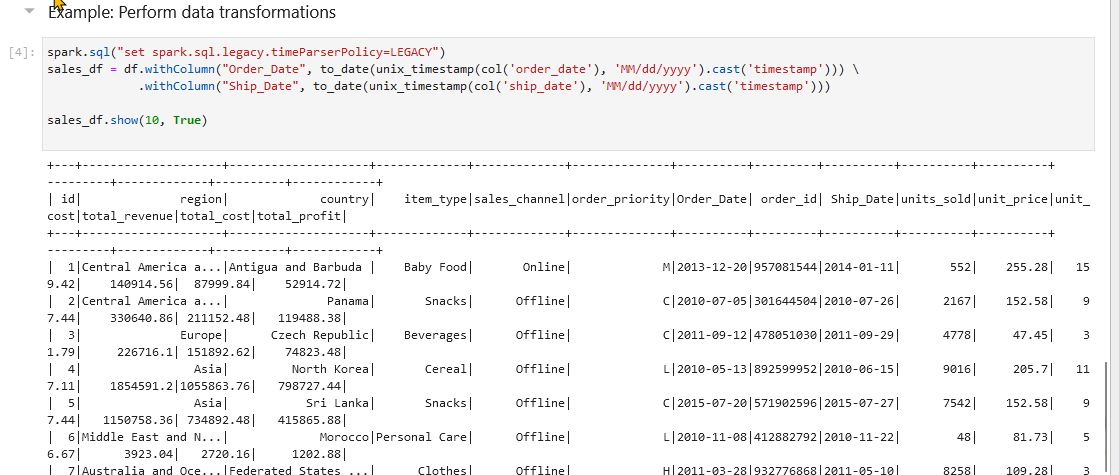




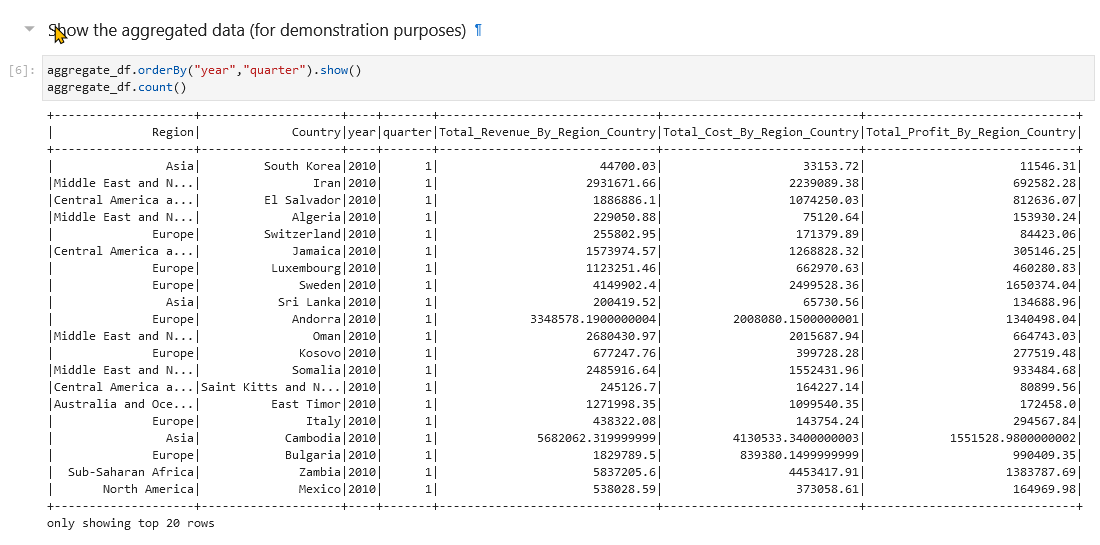
Dynamic DataFrame similar to spark dataframe but different syntax to perfom action

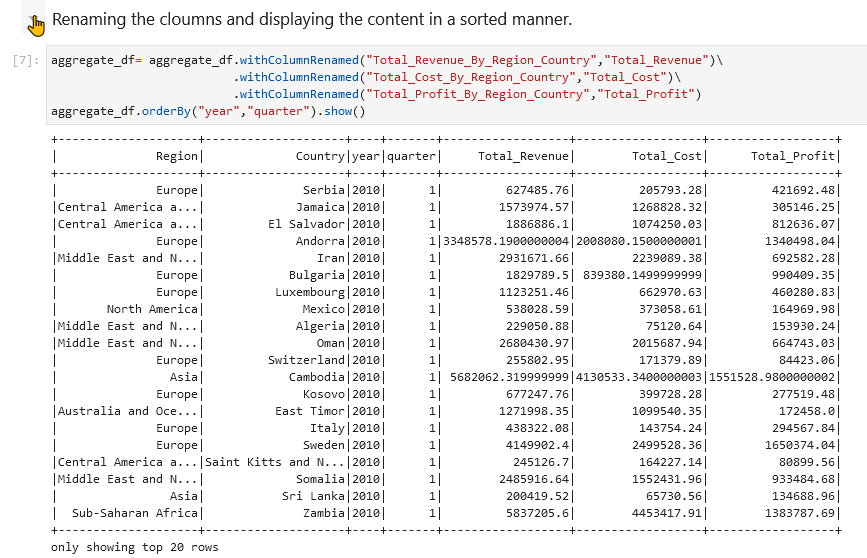


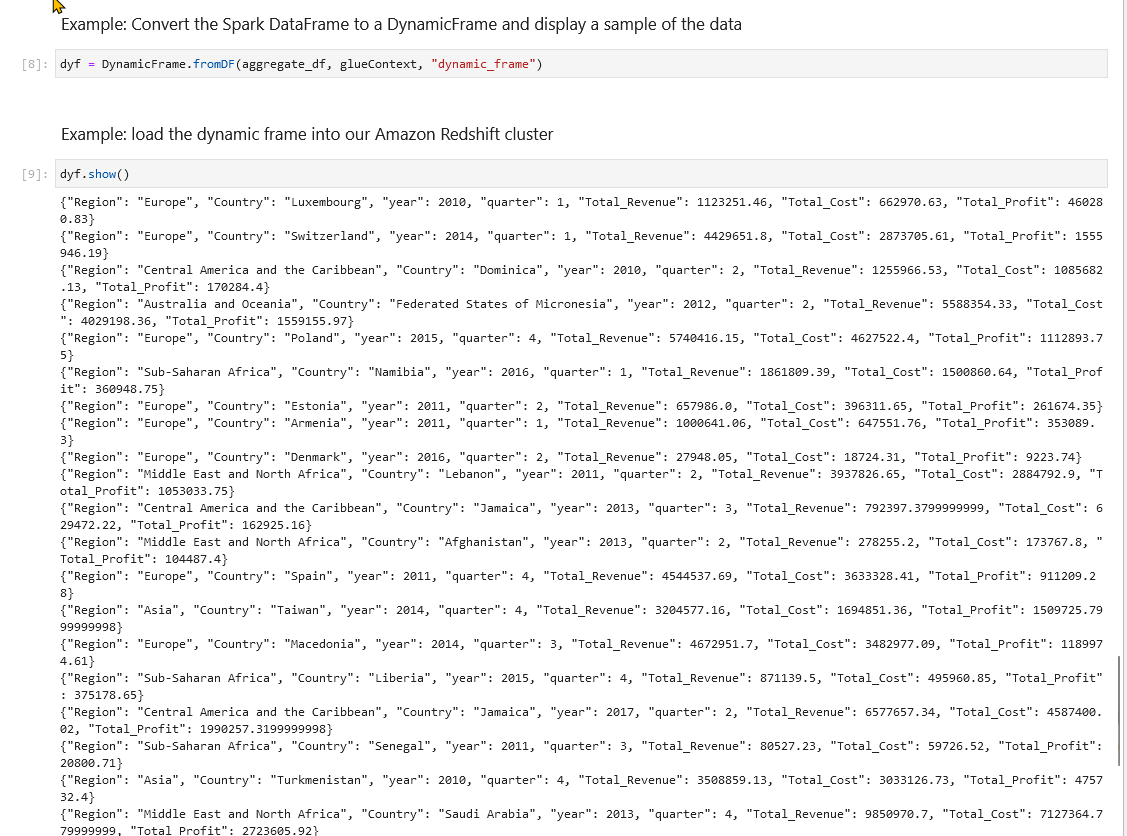
Date Transformation











Writing dynamicFrame to Redshift   
Authorization access problem to solve



