A CLOUD GURU

AWS Data Preparation Helper Tools

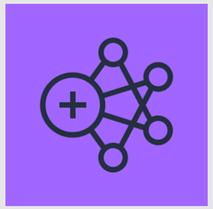
Brock Tubre
INSTRUCTOR

Data Preparation Helper Tools

Data Preparation

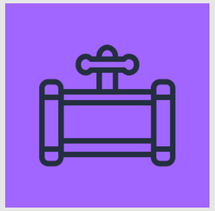
AWS Glue

SageMaker



EMR

Athena



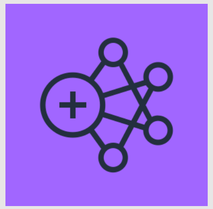
Data Pipeline

AWS Glue

Data Preparation

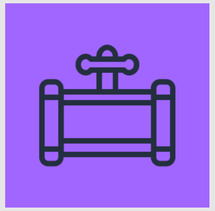
AWS Glue

SageMaker



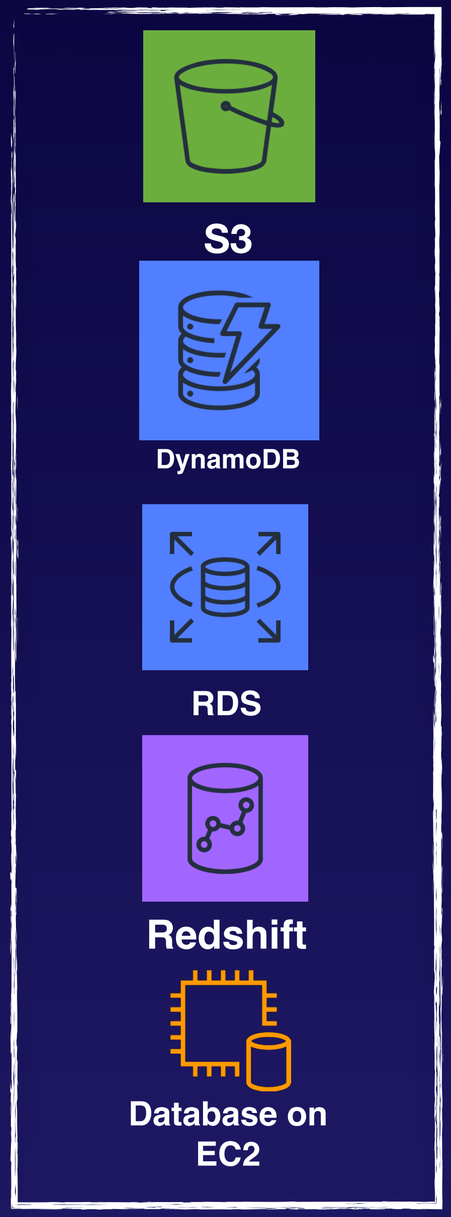
EMR

Athena

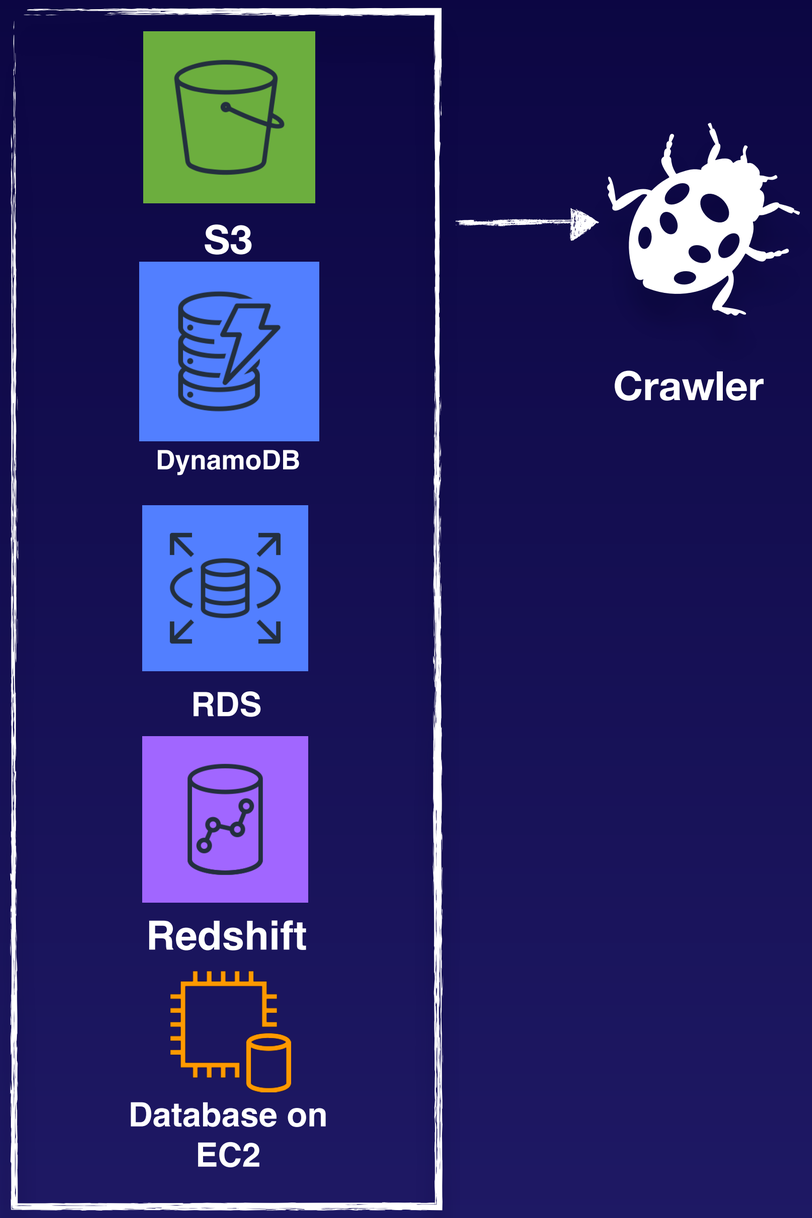


Data Pipeline

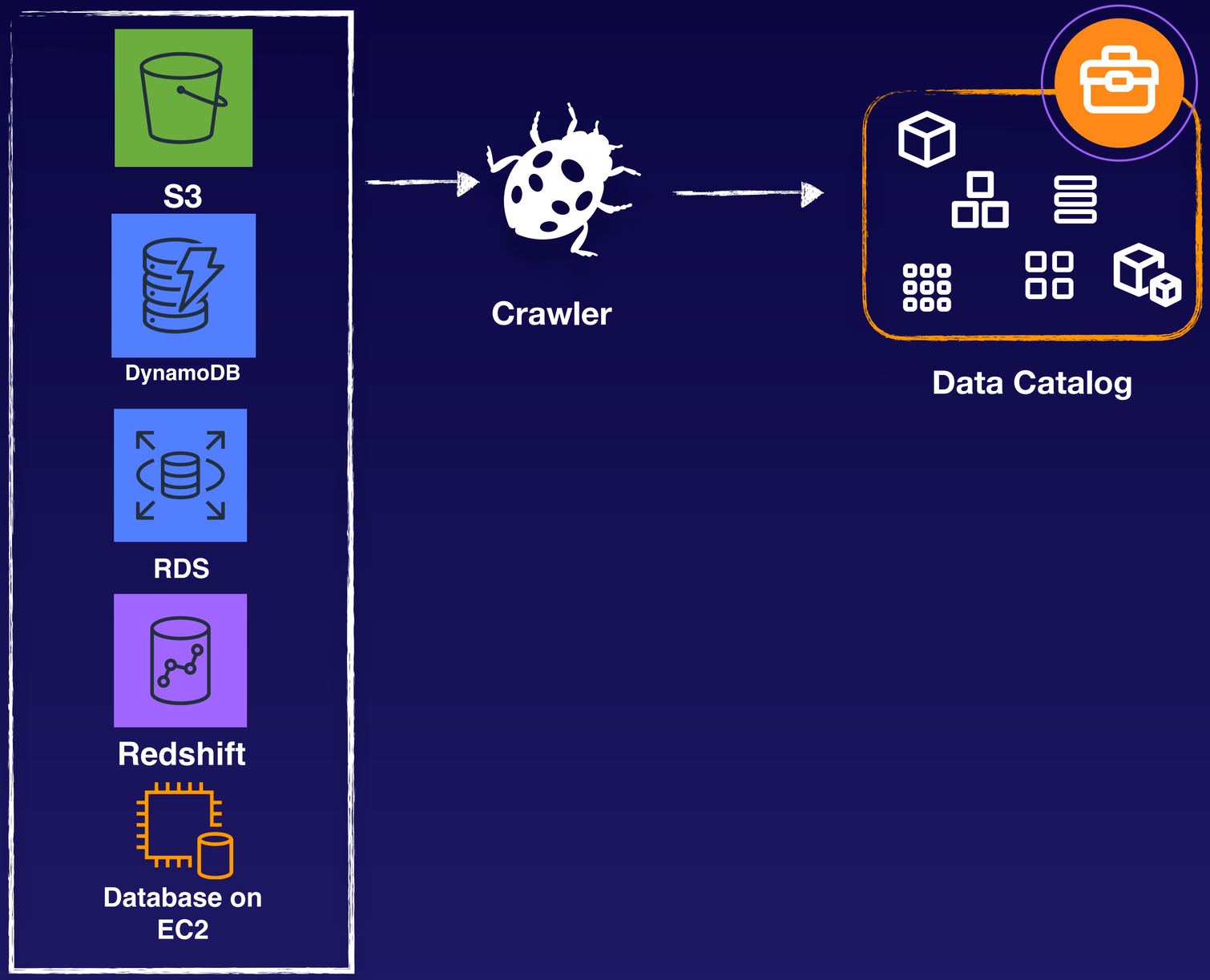
AWS Glue



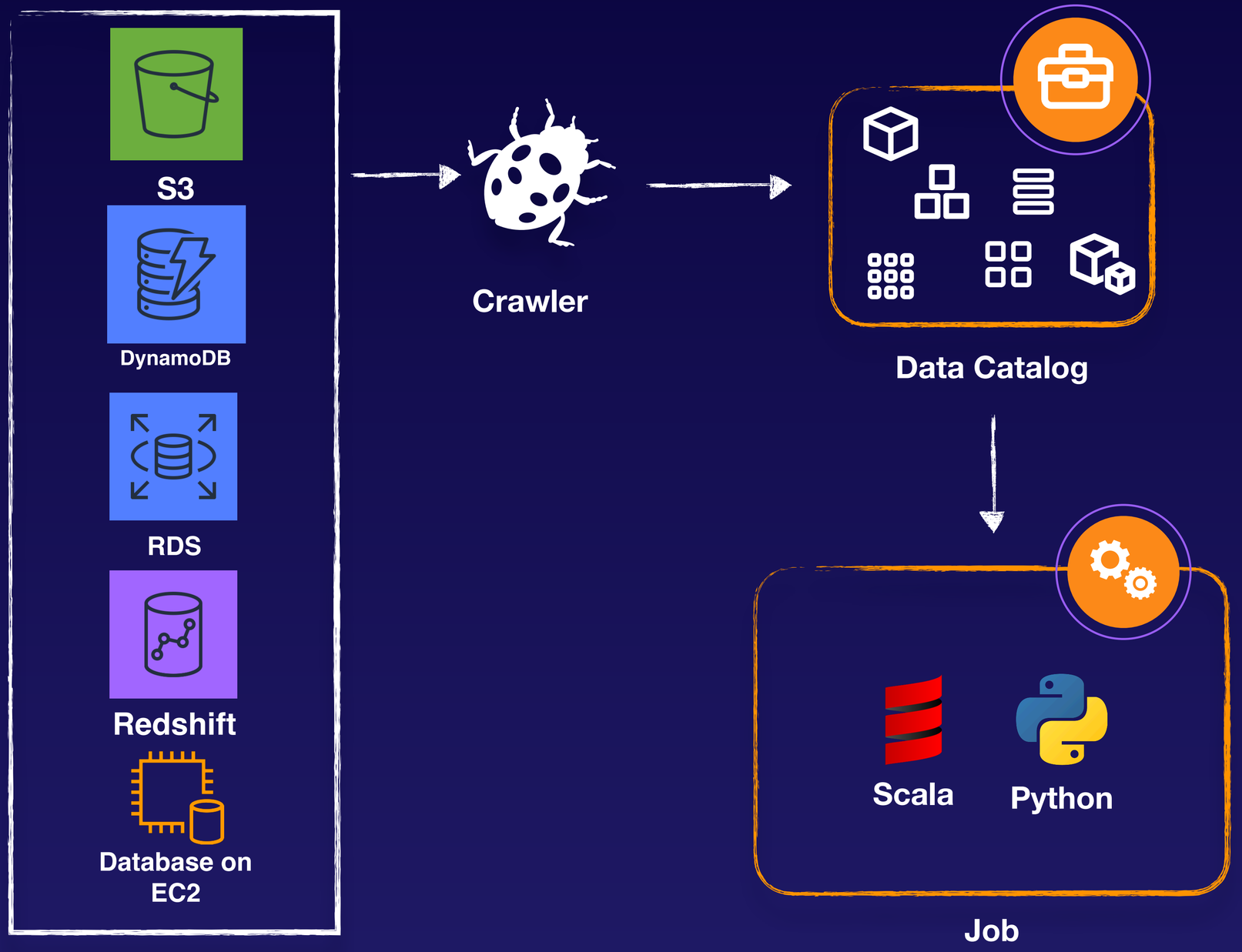
AWS Glue



AWS Glue


S3
DynamoDB
RDS
Redshift
Database on EC2
Crawler
Data Catalog

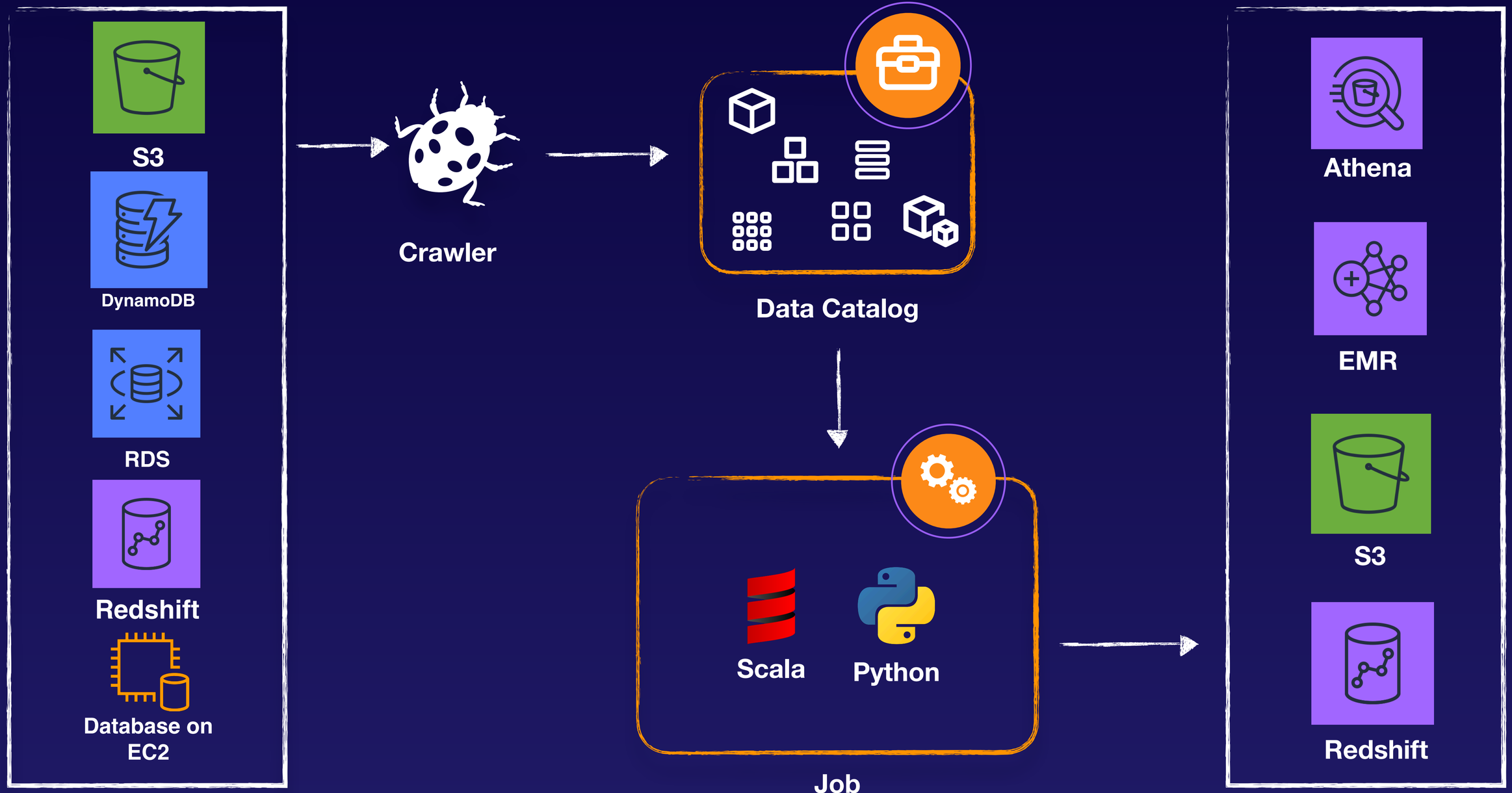
AWS Glue


S3
DynamoDB
RDS
Redshift
Database on EC2
Crawler
Data Catalog
Scala
Python
Job

AWS DATA PREPARATION HELPER TOOLS

AWS Glue

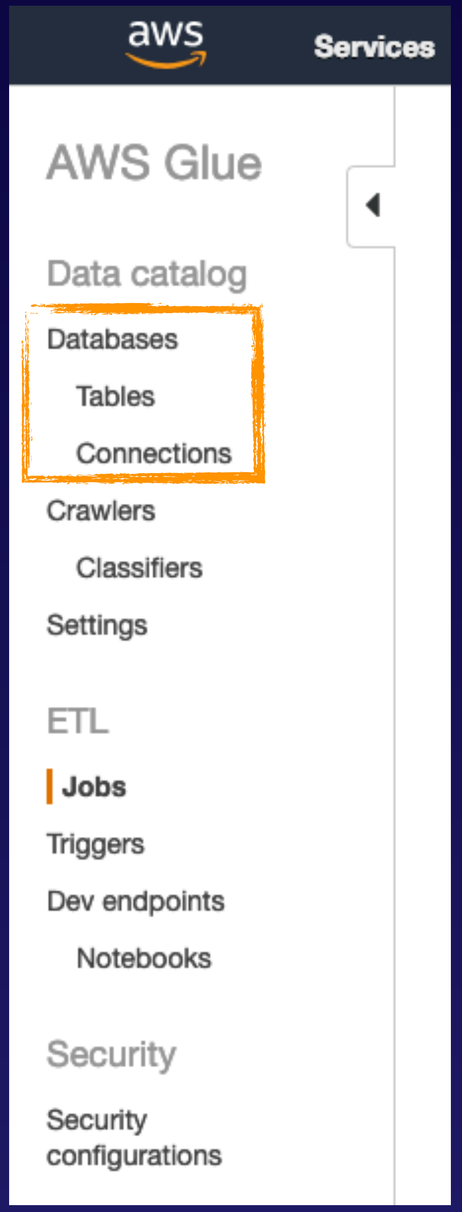
A CLOUD GURU



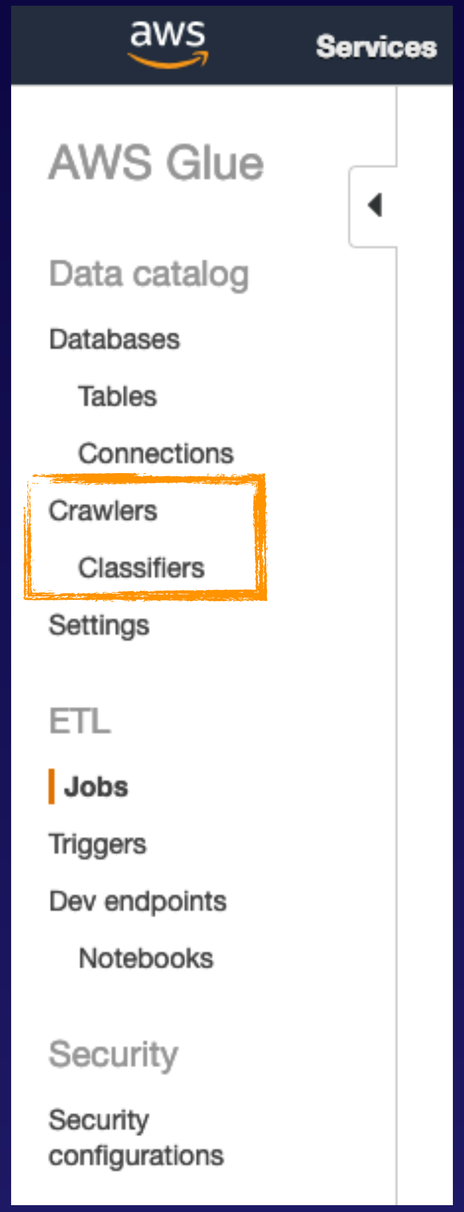
AWS Glue


aws
Services
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Settings
ETL
Jobs
Triggers
Dev endpoints
Notebooks
Security
Security configurations

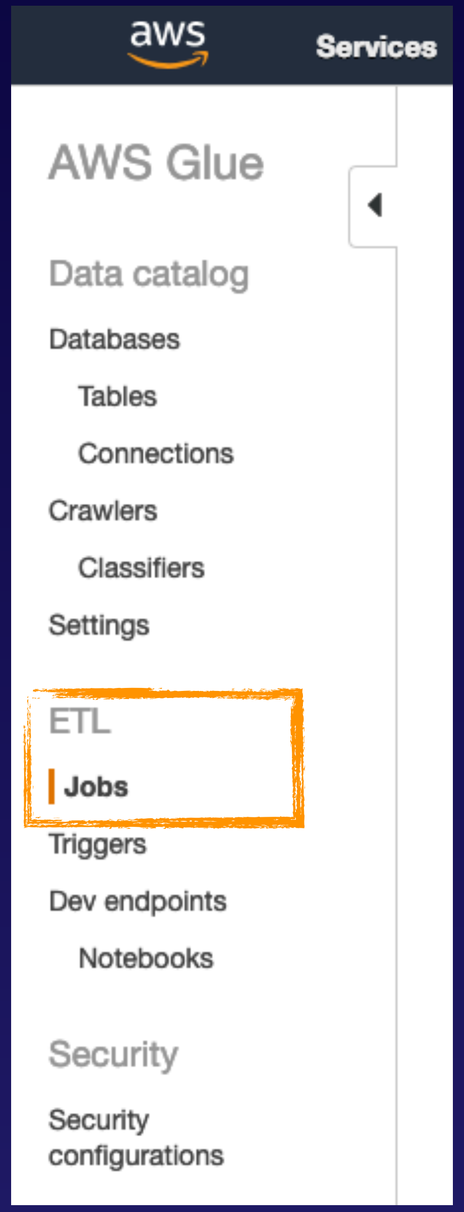
AWS Glue


aws
Services
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Settings
ETL
Jobs
Triggers
Dev endpoints
Notebooks
Security
Security configurations

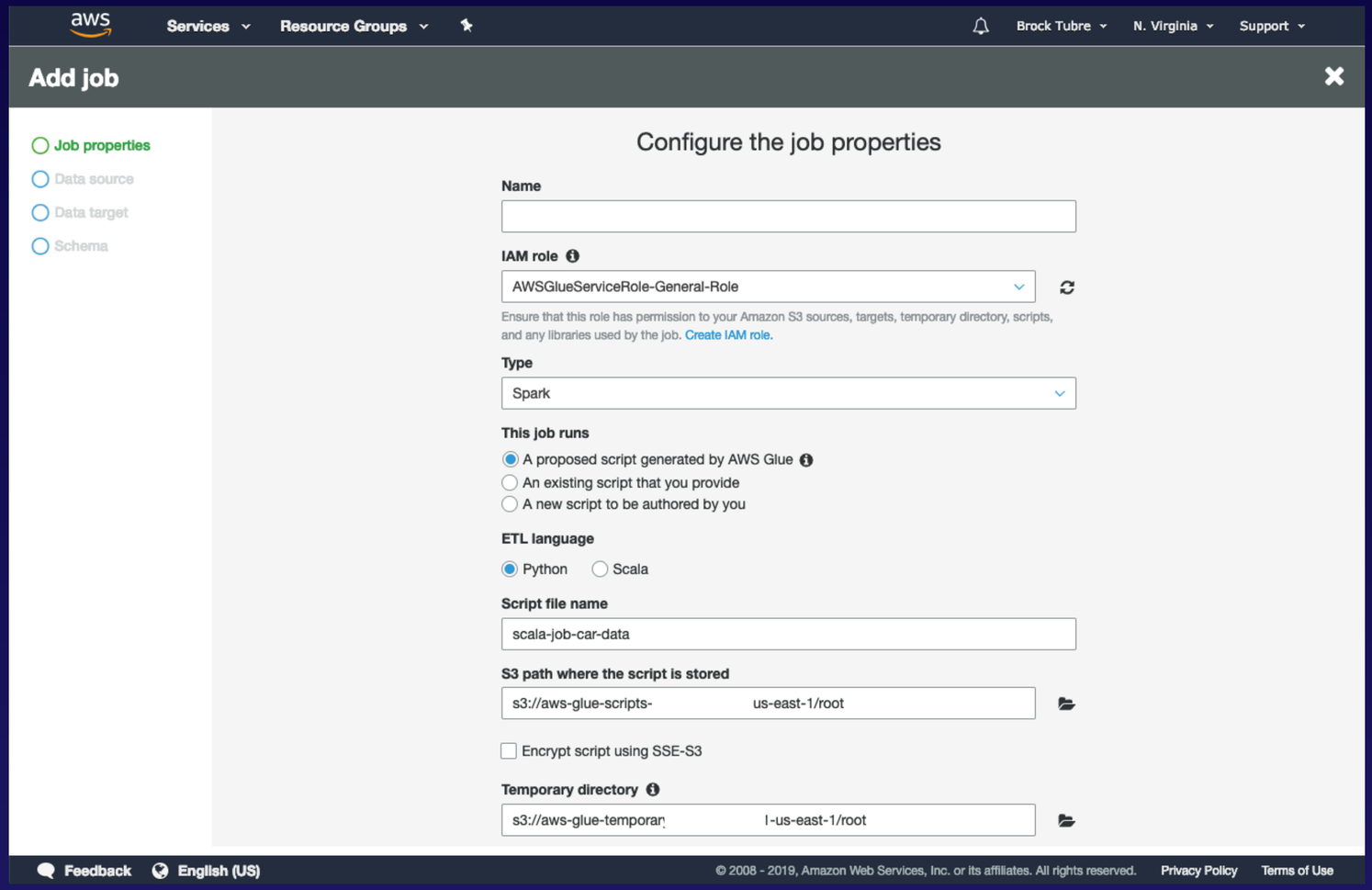
AWS Glue


aws
Services
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Settings
ETL
Jobs
Triggers
Dev endpoints
Notebooks
Security
Security configurations

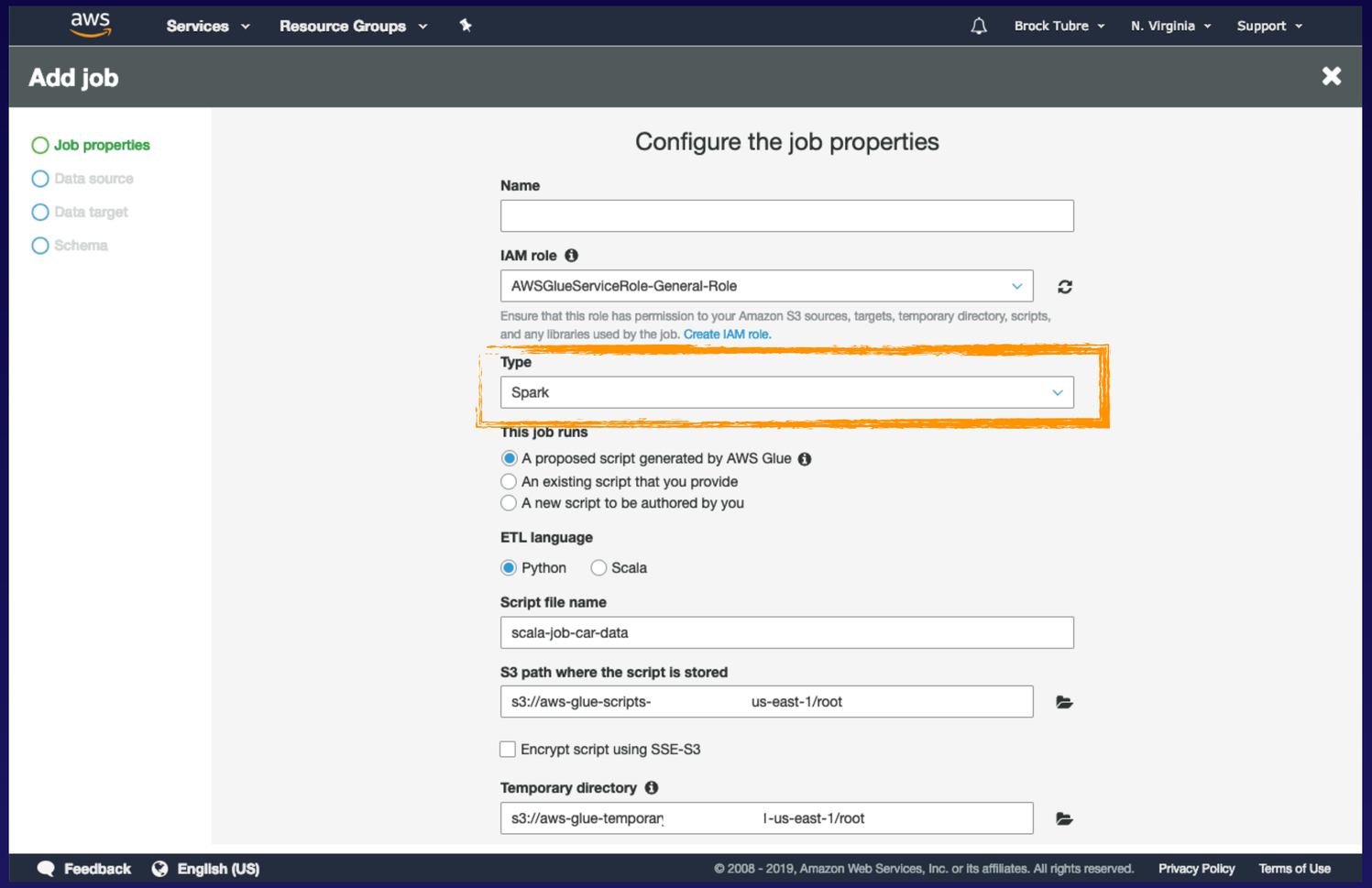
AWS Glue


aws
Services
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Settings
ETL
Jobs
Triggers
Dev endpoints
Notebooks
Security
Security configurations

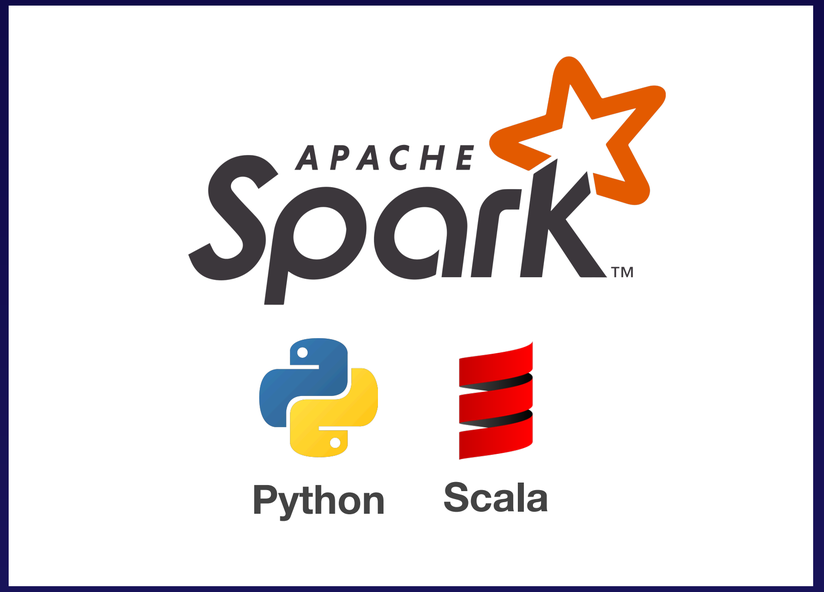
AWS Glue Jobs

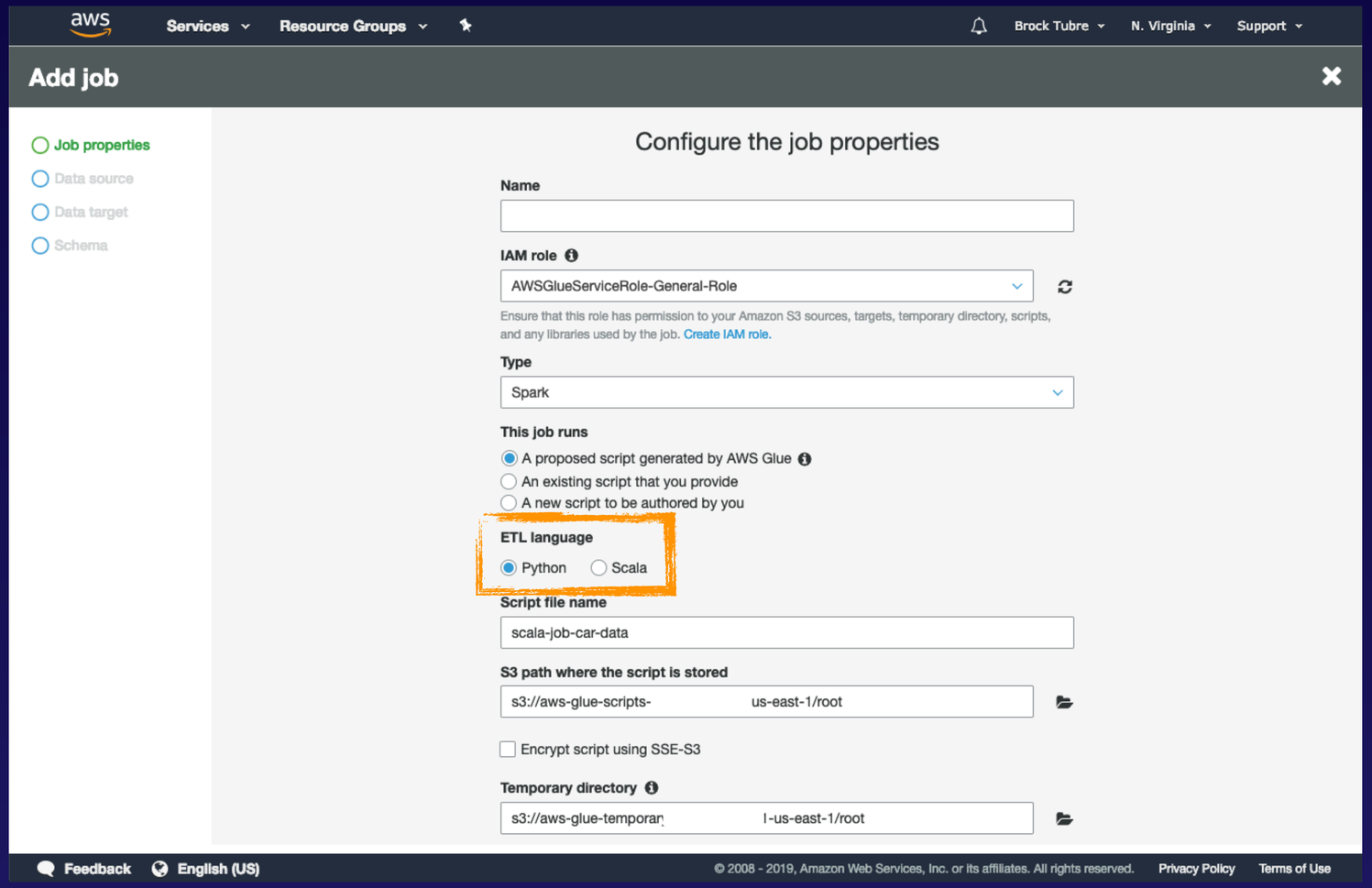


AWS Glue Jobs

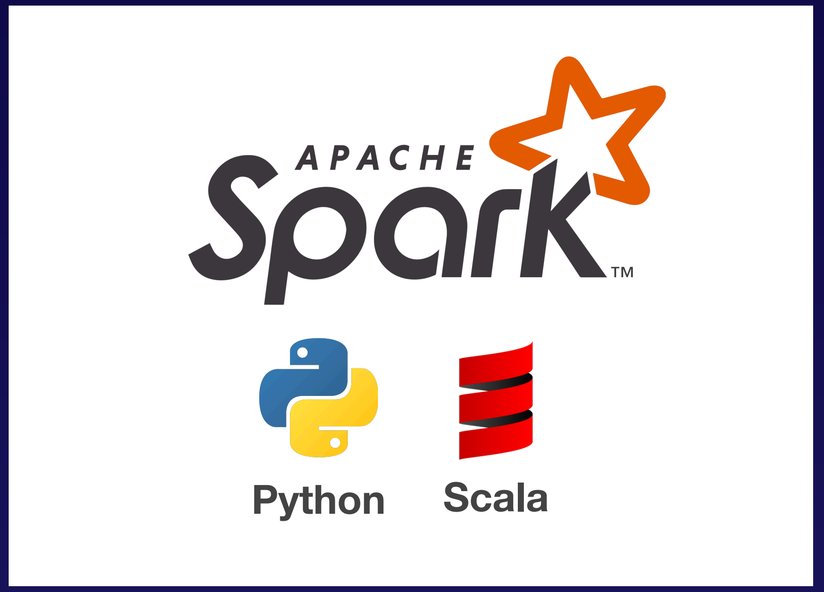


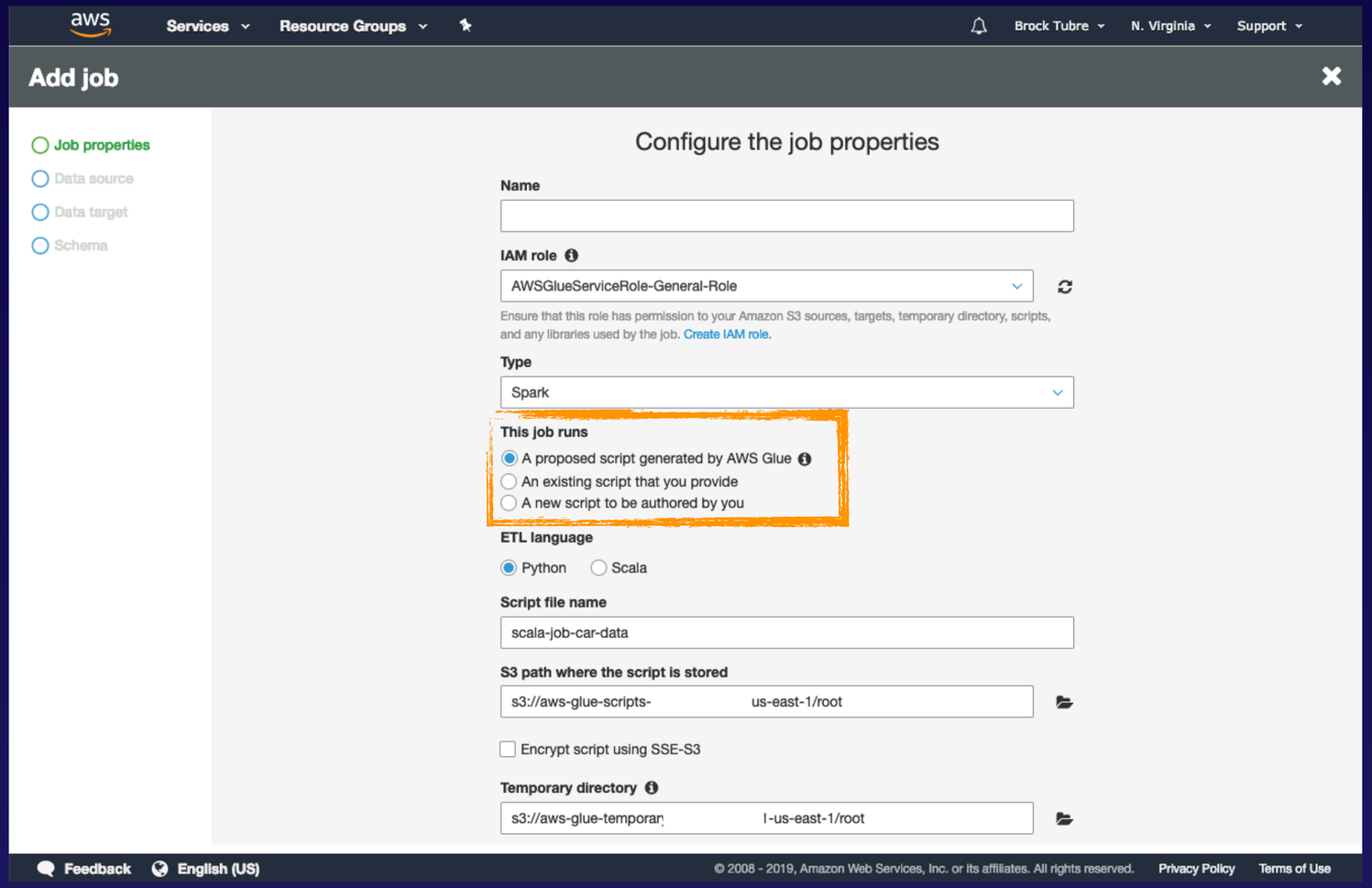
AWS Glue Jobs



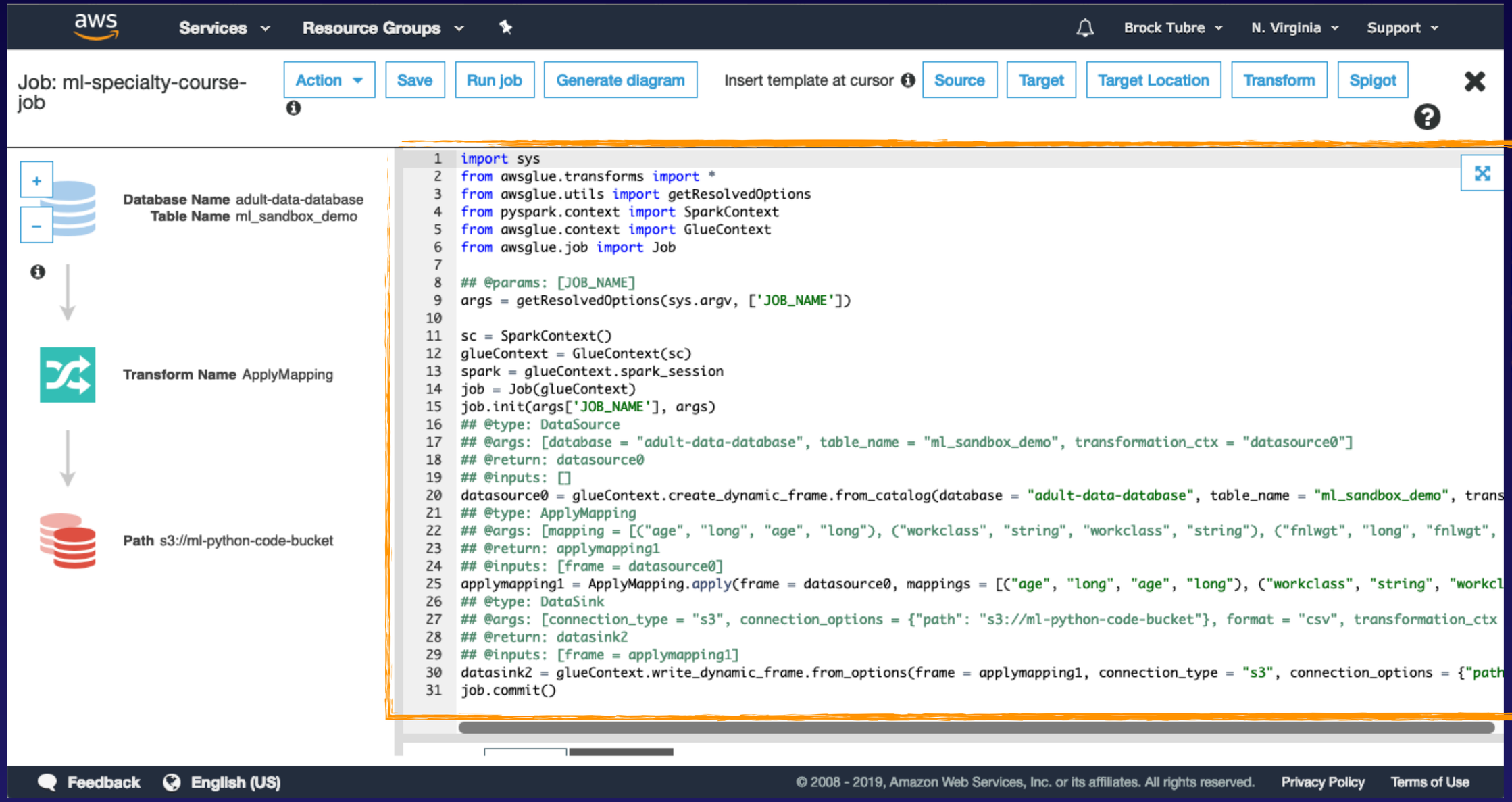


AWS Glue Jobs

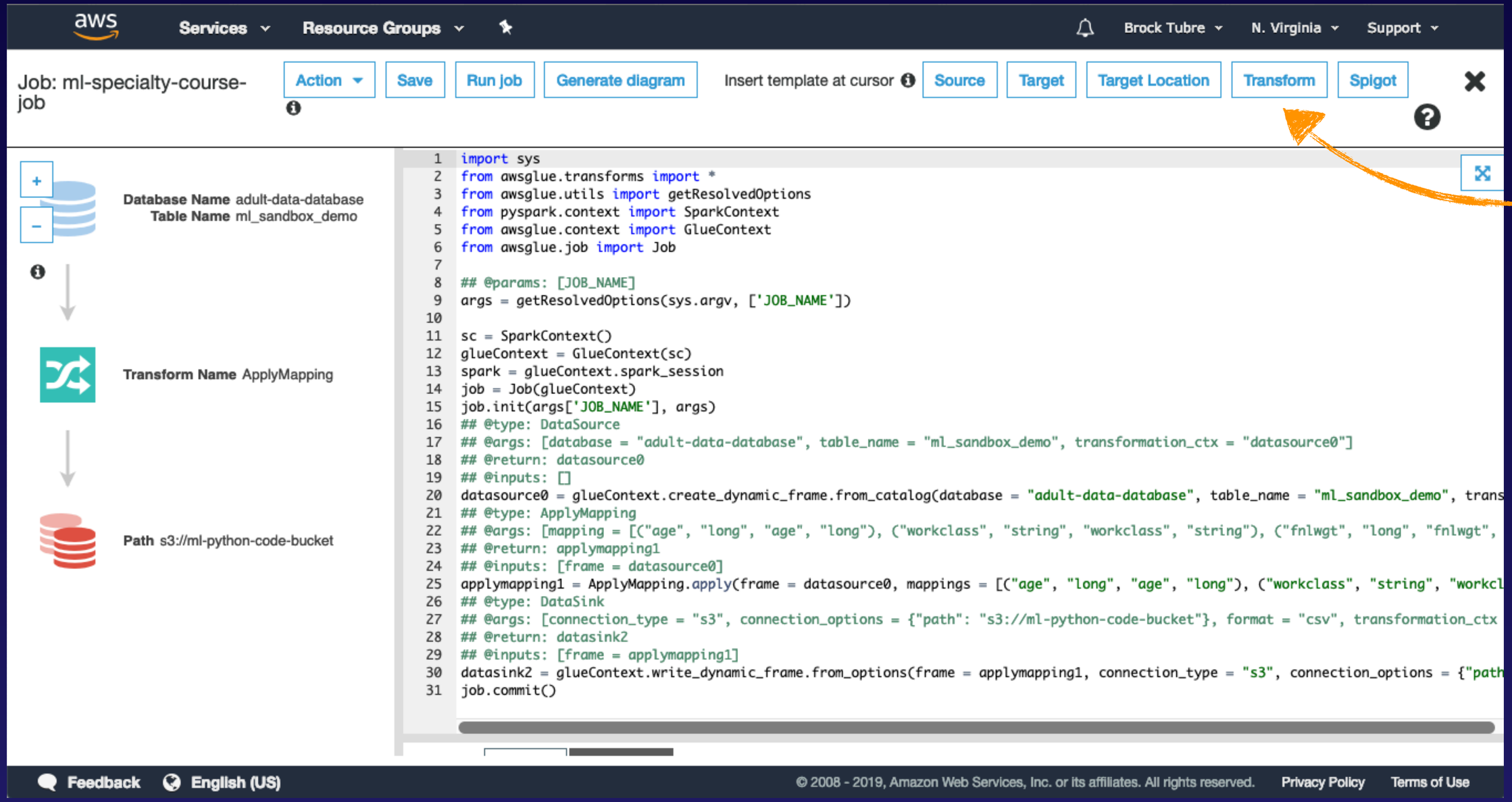




AWS Glue Jobs



AWS Glue Jobs



AWS Glue Jobs PySpark

AWS Glue PySpark Transforms Reference

AWS Glue has created the following transform Classes to use in PySpark ETL operations.

- GlueTransform Base Class
- ApplyMapping Class
- DropFields Class
- DropNullFields Class
- ErrorsAsDynamicFrame Class
- Filter Class
- Join Class
- Map Class
- MapToCollection Class
- Relationalize Class
- RenameField Class
- ResolveChoice Class
- SelectFields Class
- SelectFromCollection Class
- Spigot Class
- SplitFields Class
- SplitRows Class
- Unbox Class
- UnnestFrame Class

Output Formats

avro

csv

ion

grokLog

json

orc

parquet

xml

AWS Glue Jobs Scala

APIs in the AWS Glue Scala Library

com.amazonaws.services.glue

The **com.amazonaws.services.glue** package in the AWS Glue Scala library contains the following APIs:

- ChoiceOption
- DataSink
- DataSource trait
- DynamicFrame
- DynamicRecord
- GlueContext
- MappingSpec
- ResolveSpec

APIs in the AWS Glue Scala Library

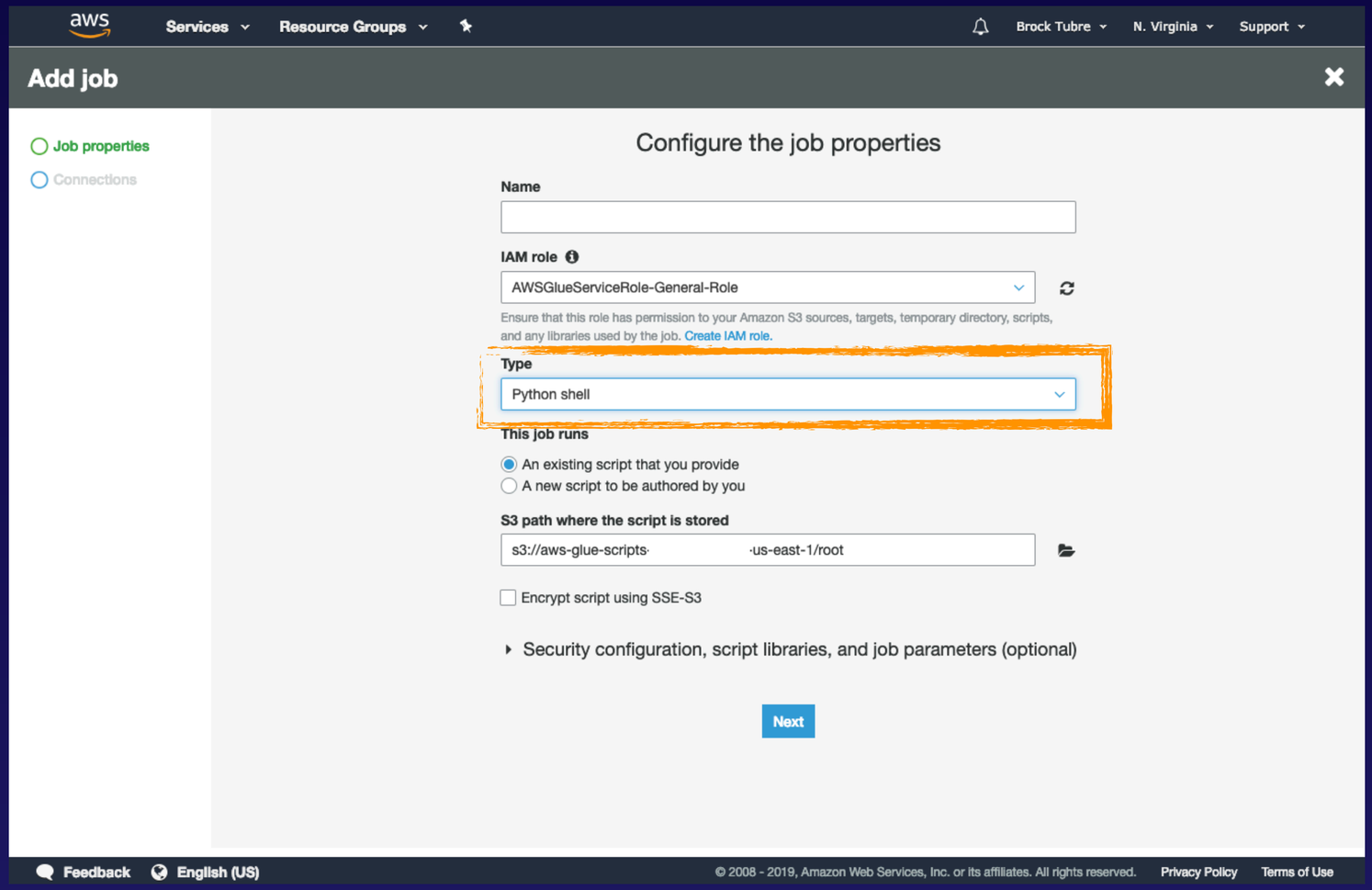
com.amazonaws.services.glue.types

The **com.amazonaws.services.glue.types** package in the AWS Glue Scala library contains the following APIs:

- ArrayNode
- BinaryNode
- BooleanNode
- ByteNode
- DateNode
- DecimalNode
- DoubleNode
- DynamicNode
- FloatNode
- IntegerNode
- LongNode
- MapLikeNode
- MapNode
- NullNode
- ObjectNode
- ScalarNode
- ShortNode
- StringNode
- TimestampNode

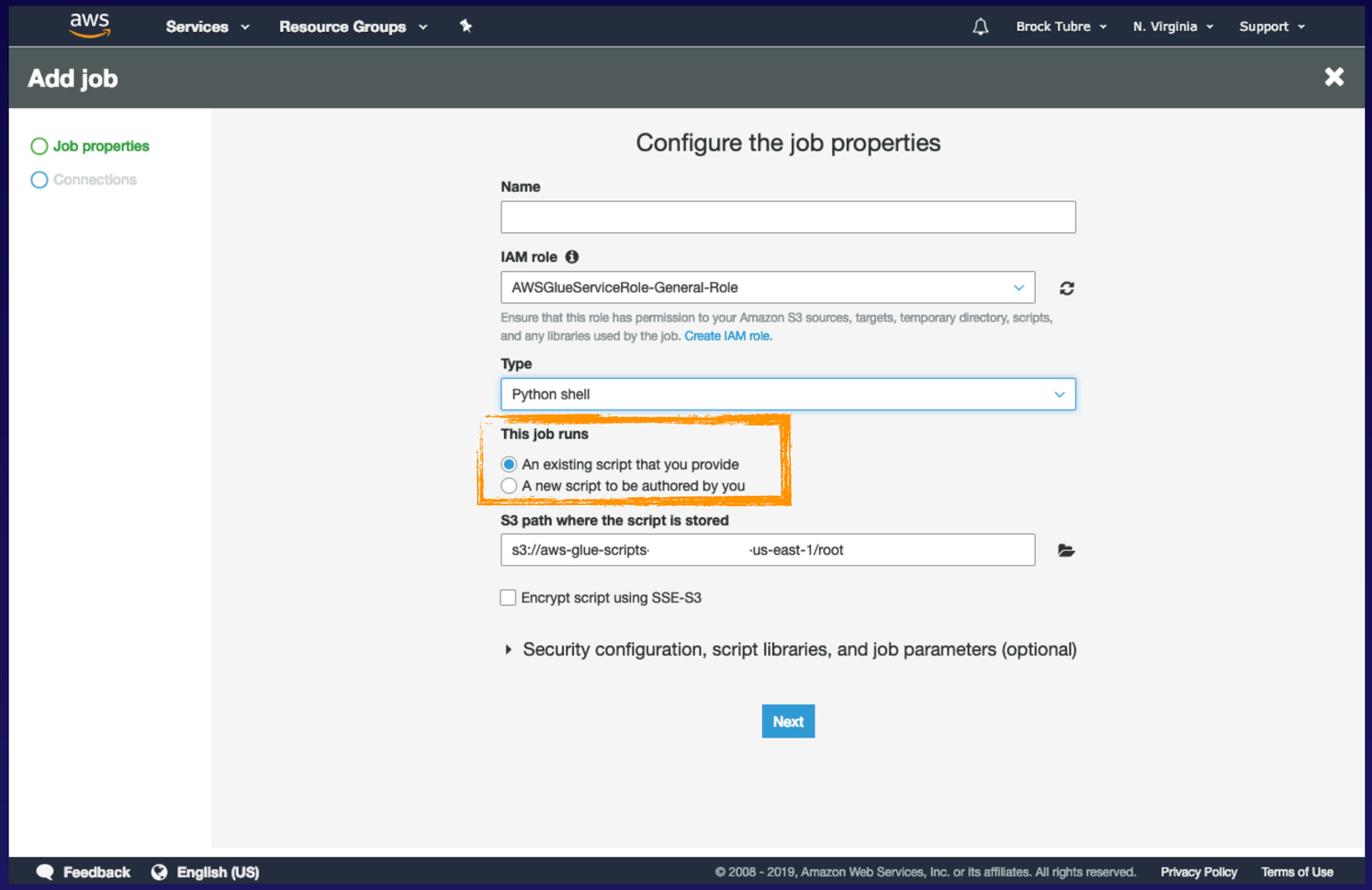
AWS Glue Jobs

Python Shell

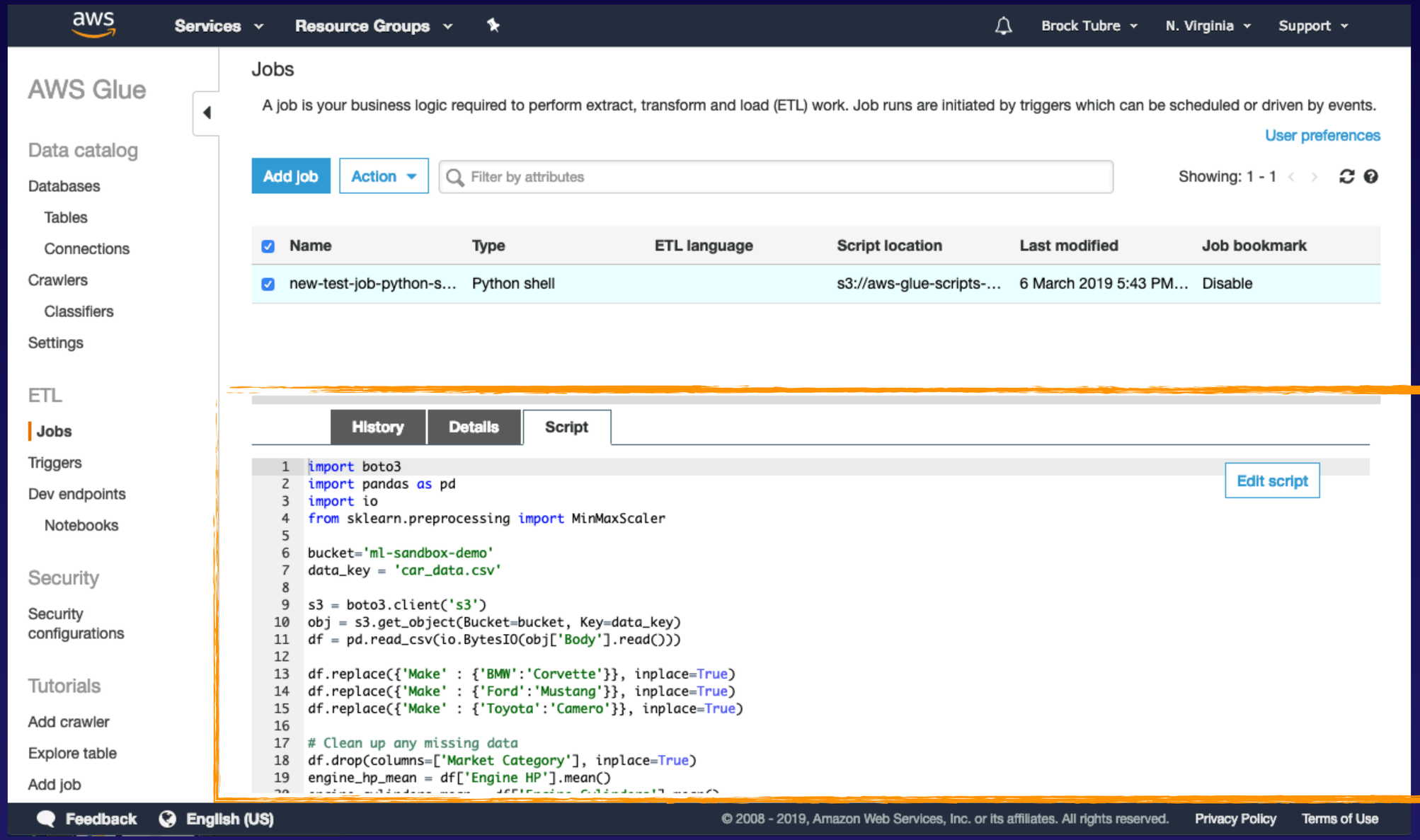


AWS Glue Jobs

Python Shell



AWS Glue Jobs



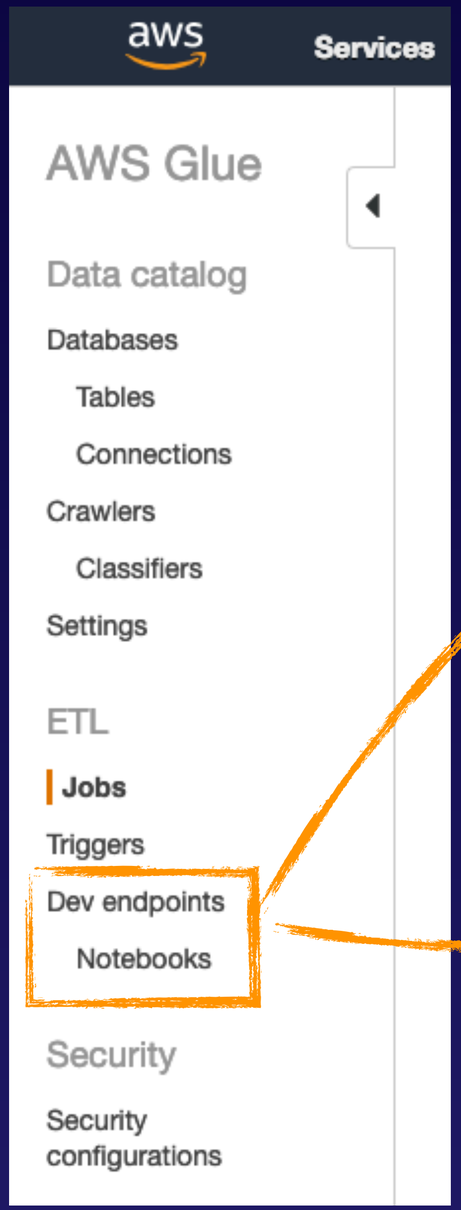
AWS Glue Jobs

Supported Libraries for Python Shell Jobs

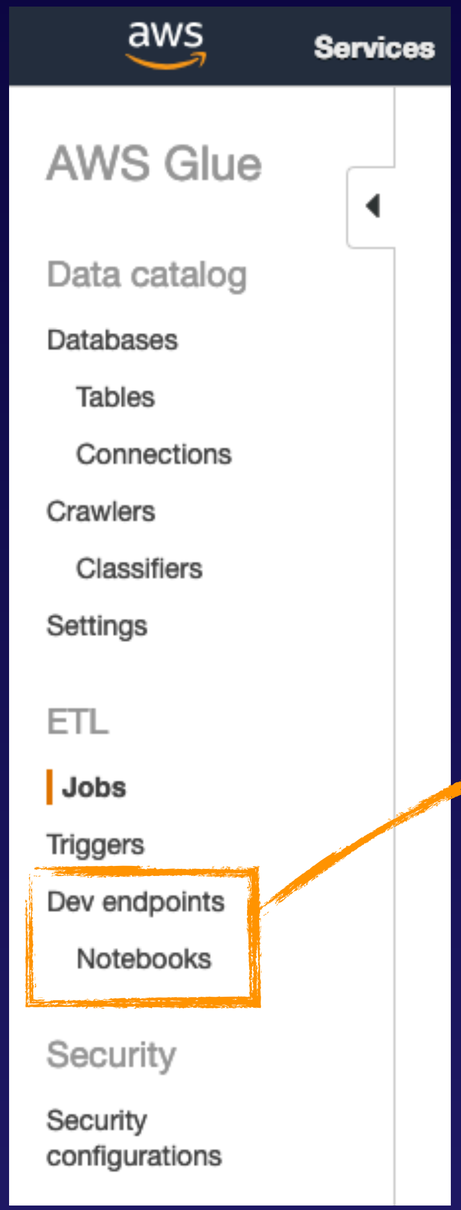
The environment for running a Python shell job supports the following libraries:

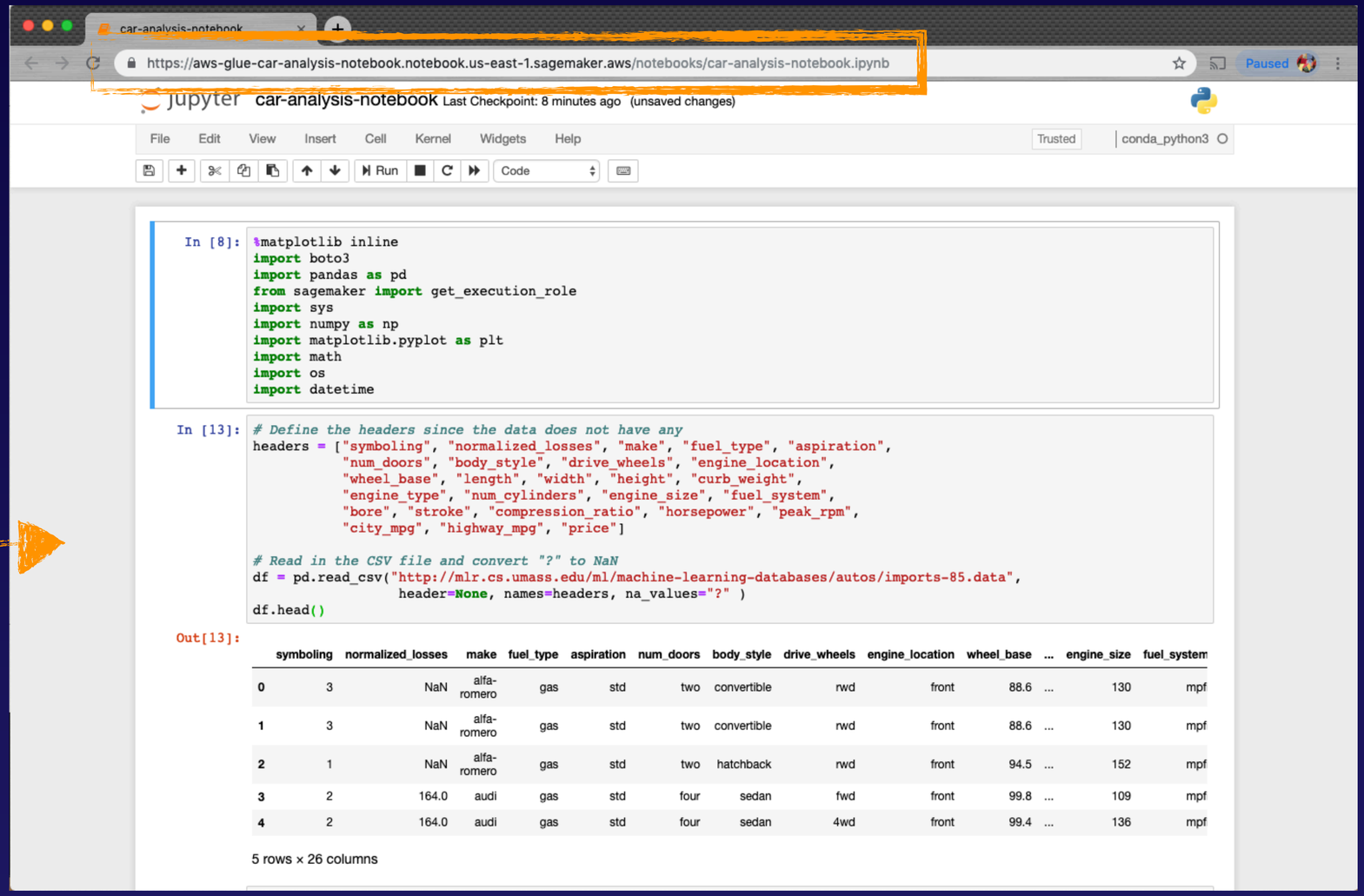
- Boto3
- collections
- CSV
- gzip
- multiprocessing
- NumPy
- pandas
- pickle
- re
- SciPy
- sklearn
- sklearn.feature_extraction
- sklearn.preprocessing
- xml.etree.ElementTree
- zipfile

AWS Glue Notebooks



AWS Glue Notebooks



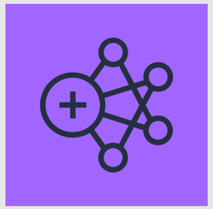


SageMaker

Data Preparation

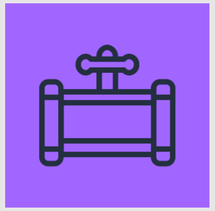
AWS Glue

SageMaker



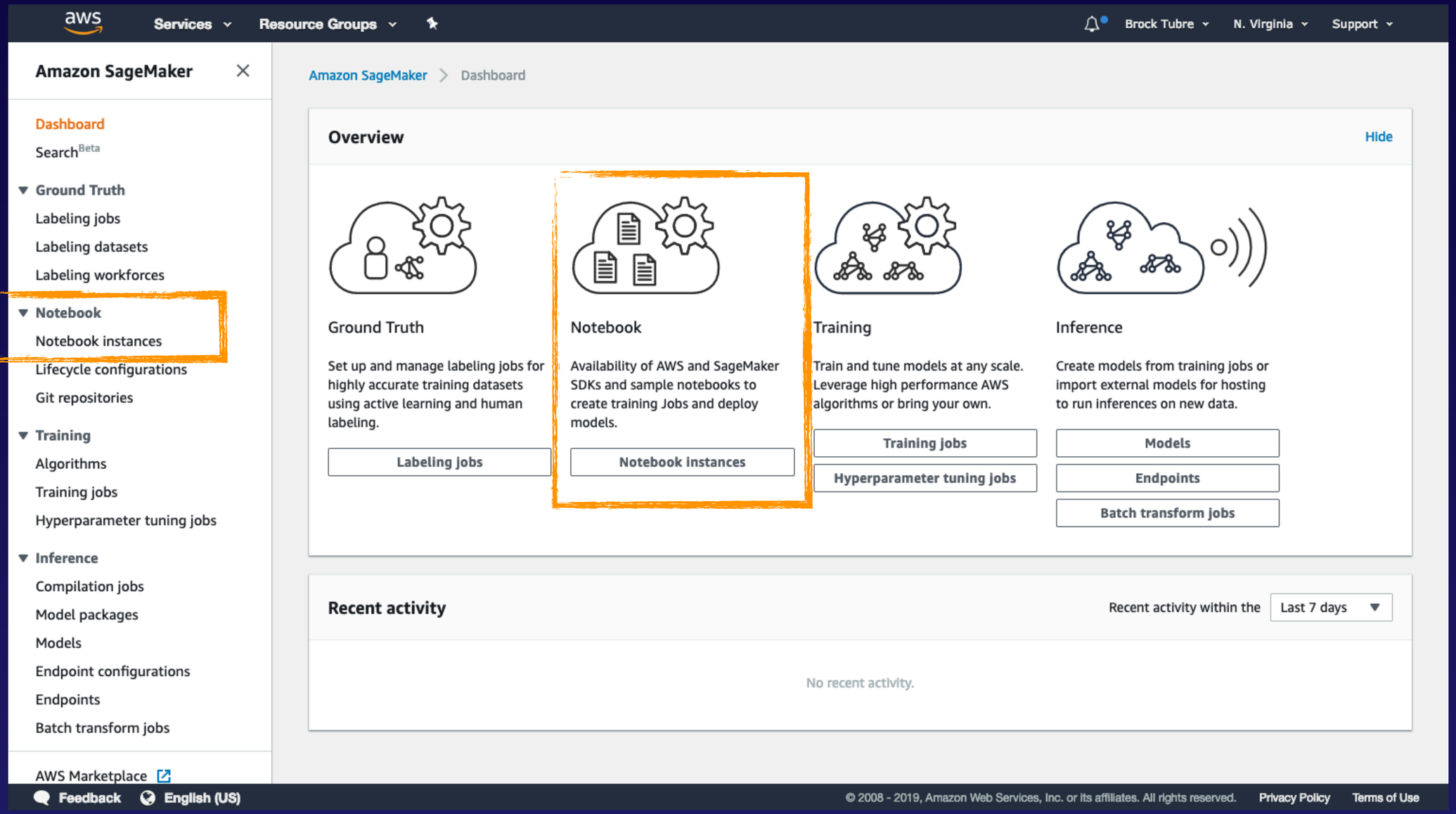
EMR

Athena

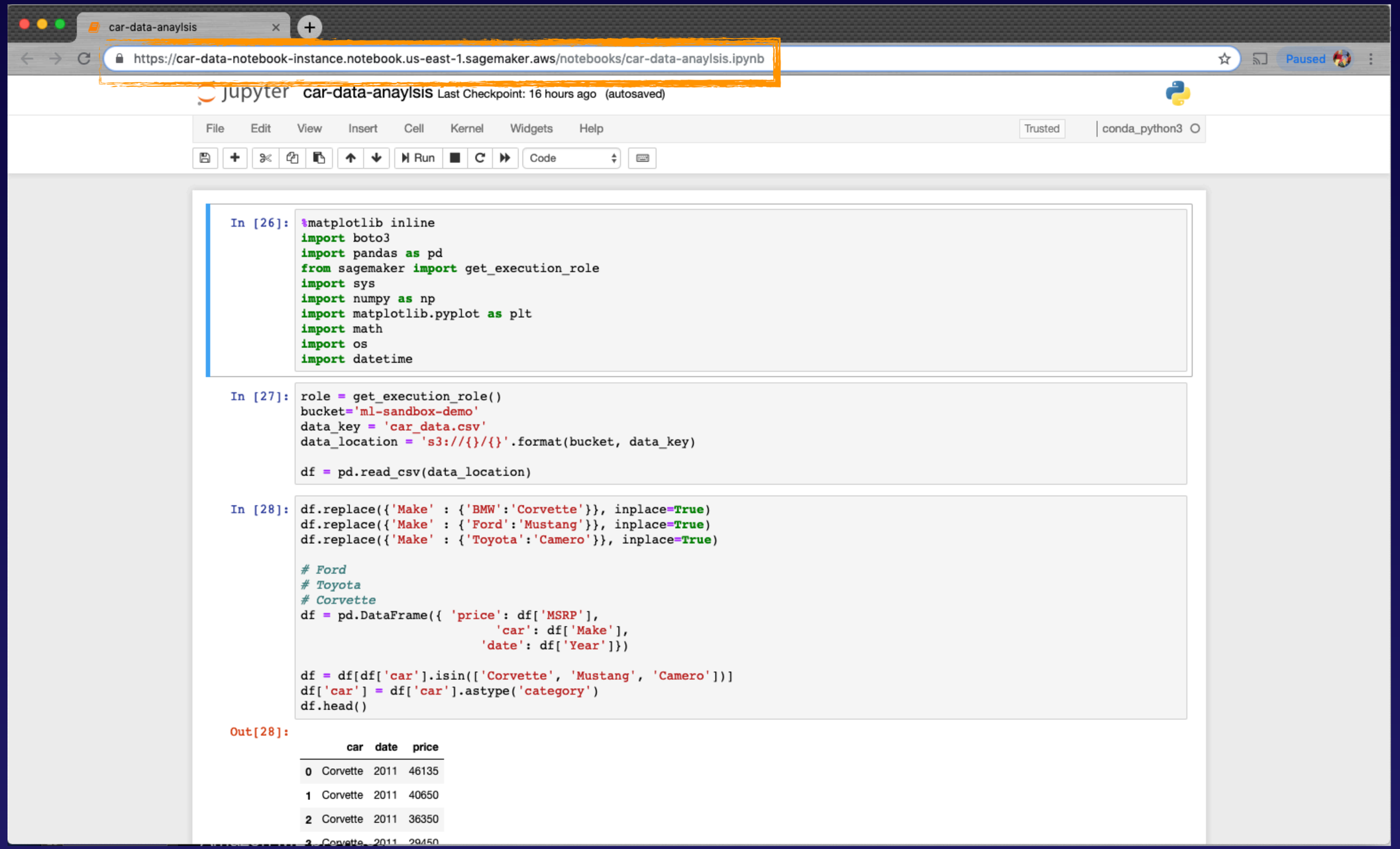


Data Pipeline

SageMaker Jupyter Notebooks

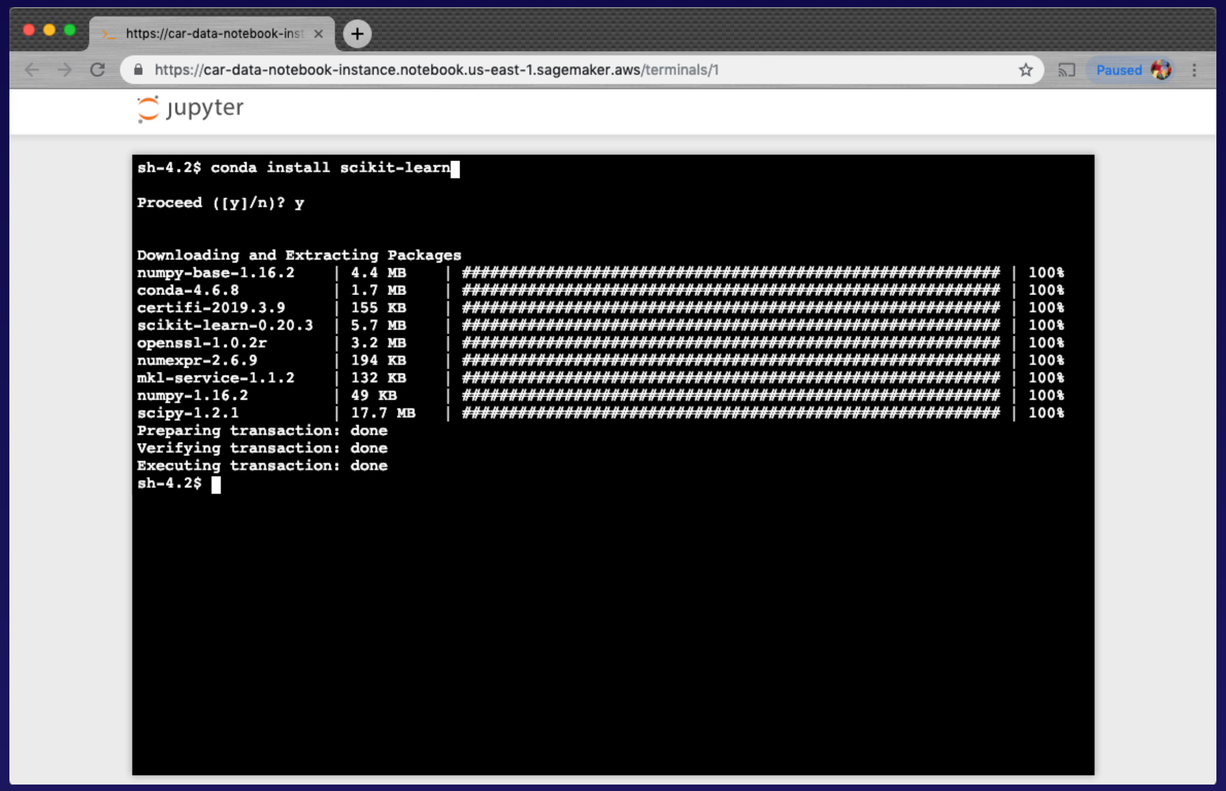


SageMaker Jupyter Notebooks

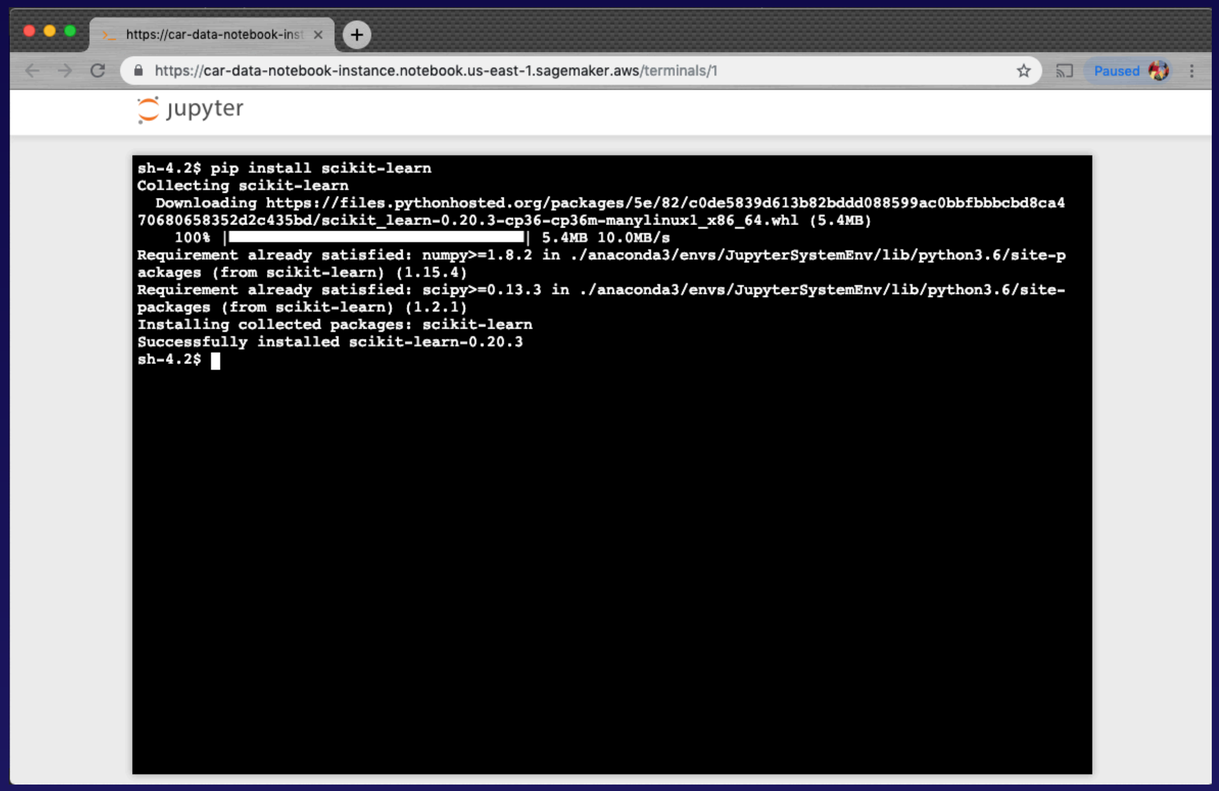


SageMaker Jupyter Notebooks

conda



pip



Using AWS Glue or SageMaker

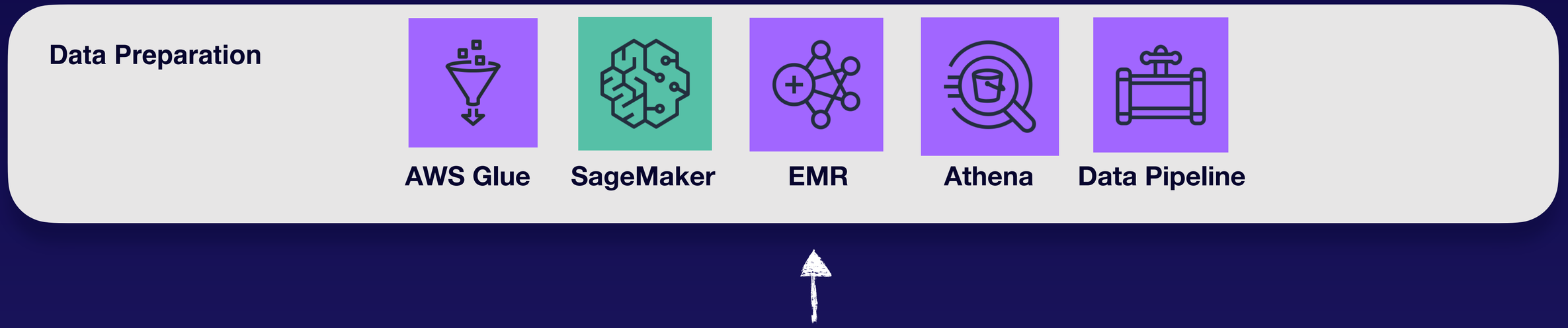
SageMaker & Juypter Notebooks

ETL jobs in AWS Glue

Adhoc

Reusable

Elastic Map Reduce (EMR)



Elastic Map Reduce (EMR)

APACHE PIG

presto

MAHOUT

mxnet

oozie

HIVE

hadoop HDFS

APACHE HBASE

FLUME

APACHE Spark™

jupyter

APACHE Zookeeper

Elastic Map Reduce (EMR)

Spark ML and MLib
ETL and Machine Learning Library

Presto
SQL Query Engine

Mahout
Machine Learning Framework

Hive
ETL Service

Jupyter Notebooks
Code Sharing

TensorFlow
Machine Learning Framework

Hadoop Distributed File System
Persistant Datastore

MXNet
Machine Learning Framework

Elastic Map Reduce (EMR)

Spark ML and MLib
ETL and Machine Learning Library

Presto
SQL Query Engine

Mahout
Machine Learning Framework

Hive
ETL Service

Jupyter Notebooks
Code Sharing

TensorFlow
Machine Learning Framework

Hadoop Distributed File System
Persistant Datastore

MXNet
Machine Learning Framework

Elastic Map Reduce (EMR)

Spark ML and MLib
ETL and Machine
Learning Library

Elastic Map Reduce (EMR) Cluster

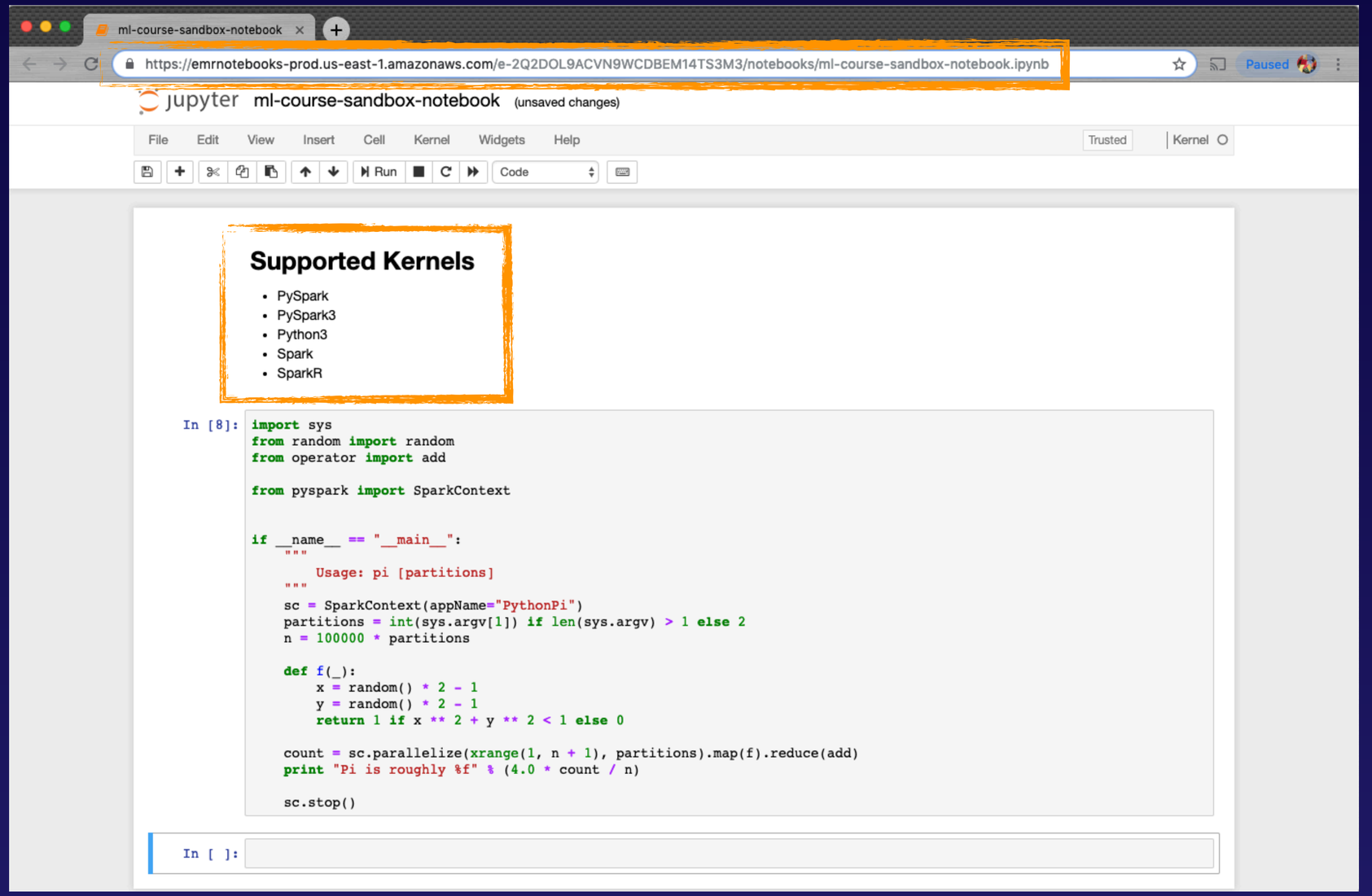
Create Cluster - Advanced Options

Software Configuration

Release emr-5.20.0

☑ Hadoop 2.8.5	☐ Zeppelin 0.8.0	☐ Livy 0.5.0
☑ JupyterHub 0.9.4	☐ Tez 0.9.1	☐ Flink 1.6.2
☐ Ganglia 3.7.2	☐ HBase 1.4.8	☐ Pig 0.17.0
☑ Hive 2.3.4	☑ Presto 0.214	☐ ZooKeeper 3.4.13
☑ MXNet 1.3.1	☐ Sqoop 1.4.7	☑ Mahout 0.13.0
☐ Hue 4.3.0	☐ Phoenix 4.14.0	☐ Oozie 5.0.0
☑ Spark 2.4.0	☐ HCatalog 2.3.4	☑ TensorFlow 1.12.0

Elastic Map Reduce (EMR) Notebooks

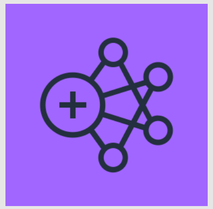


Athena

Data Preparation

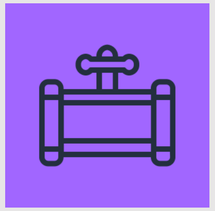
AWS Glue

SageMaker



EMR

Athena

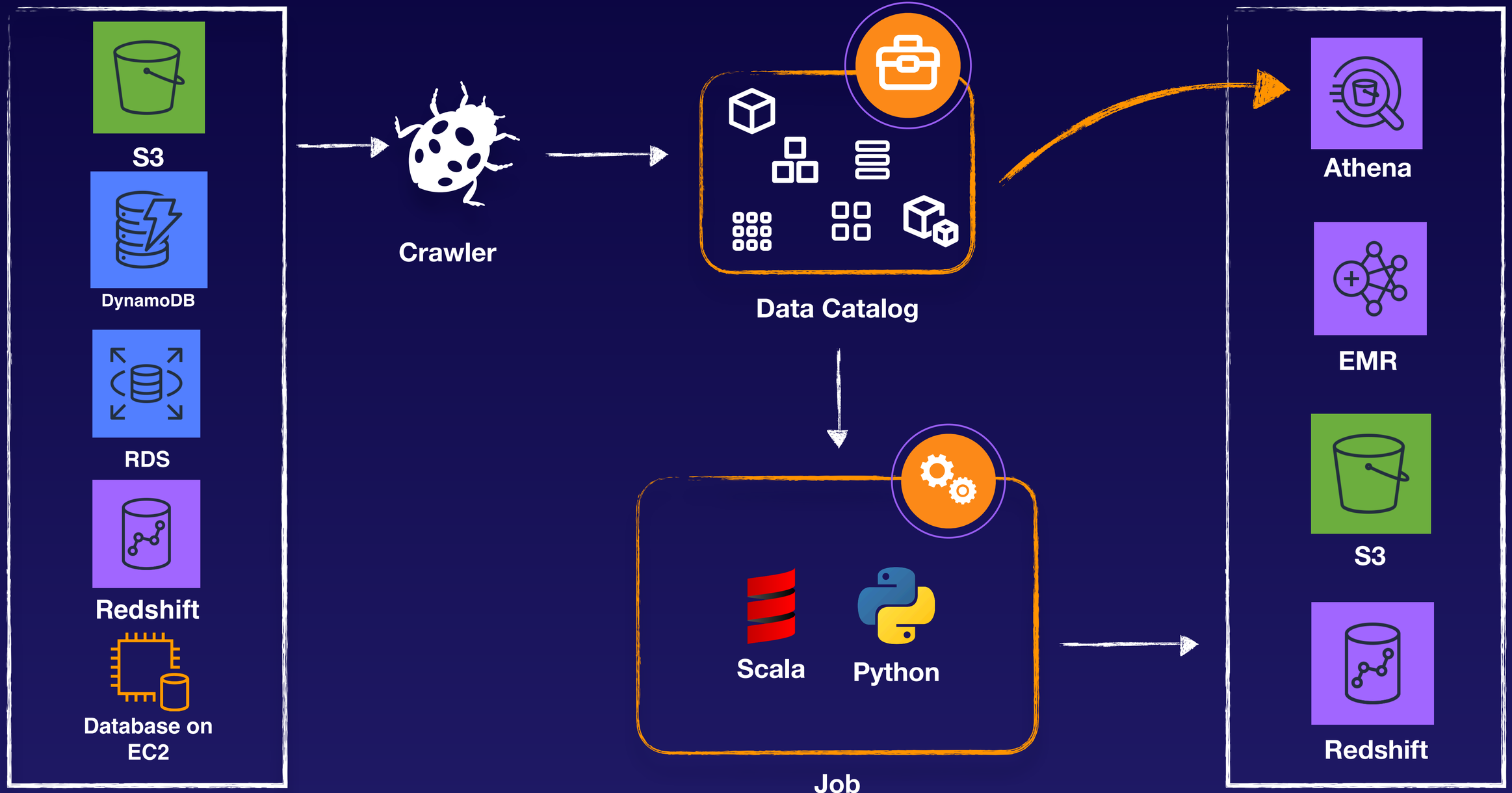


Data Pipeline

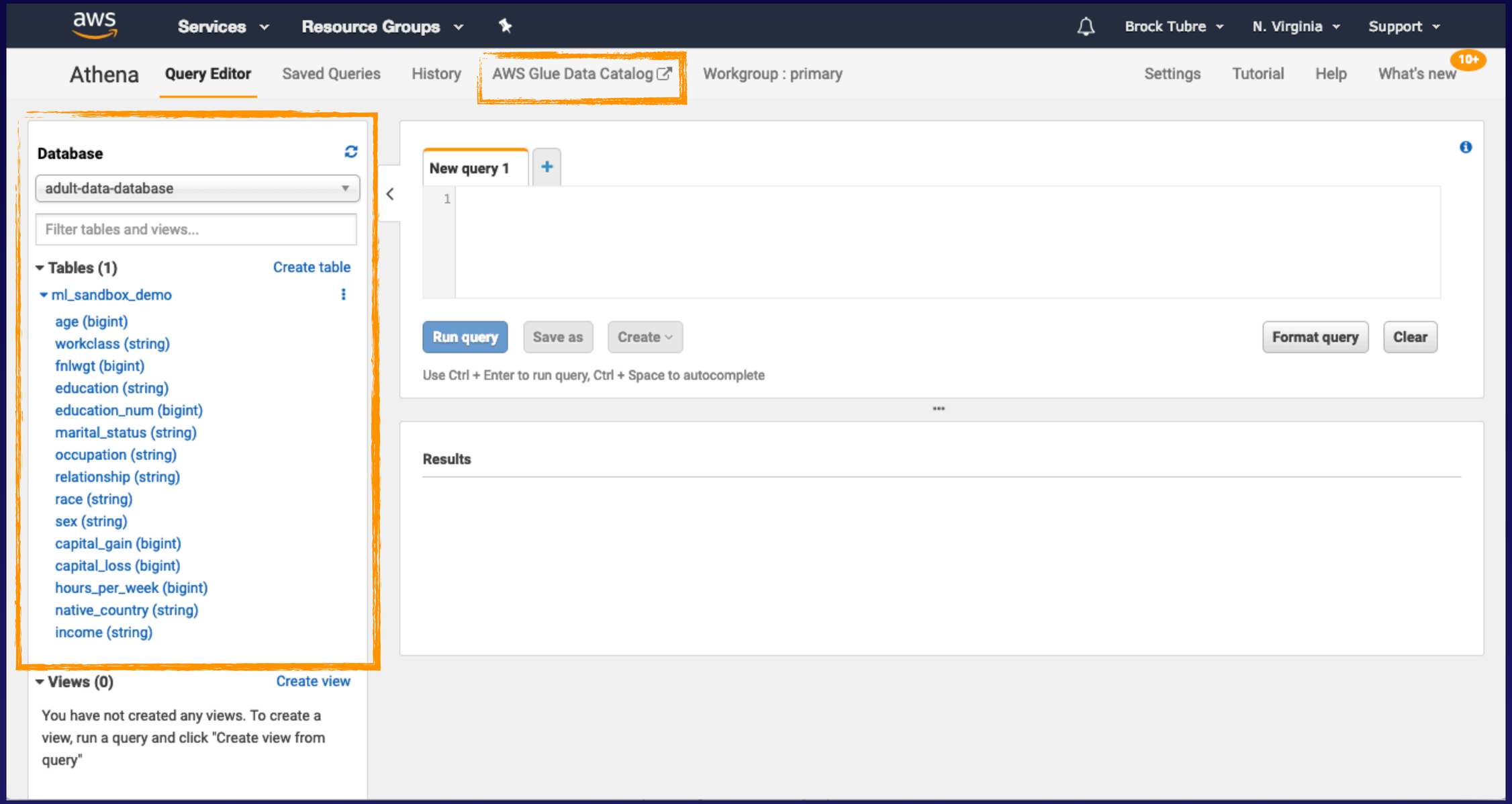
AWS DATA PREPARATION HELPER TOOLS

AWS Glue

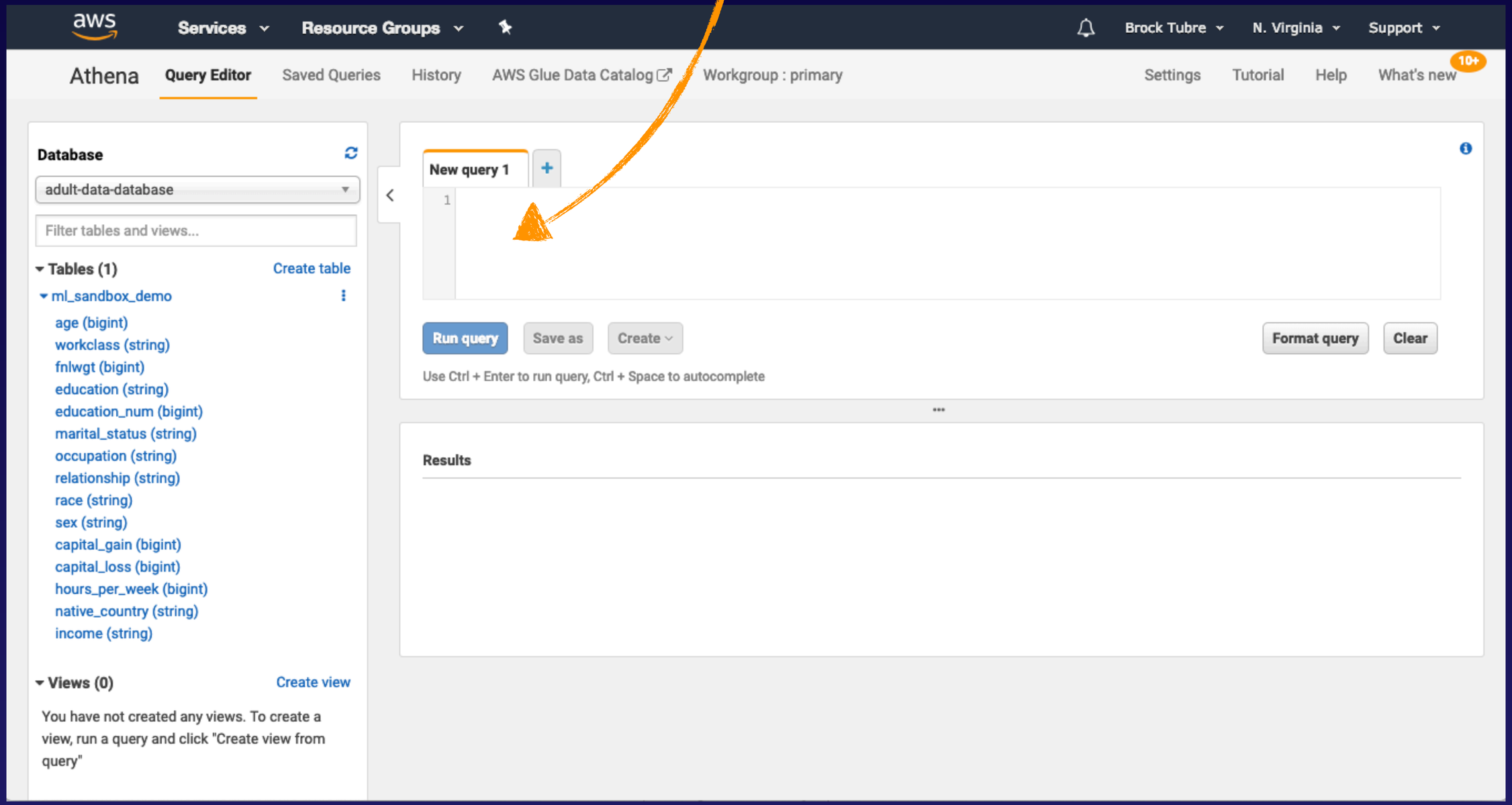
A CLOUD GURU



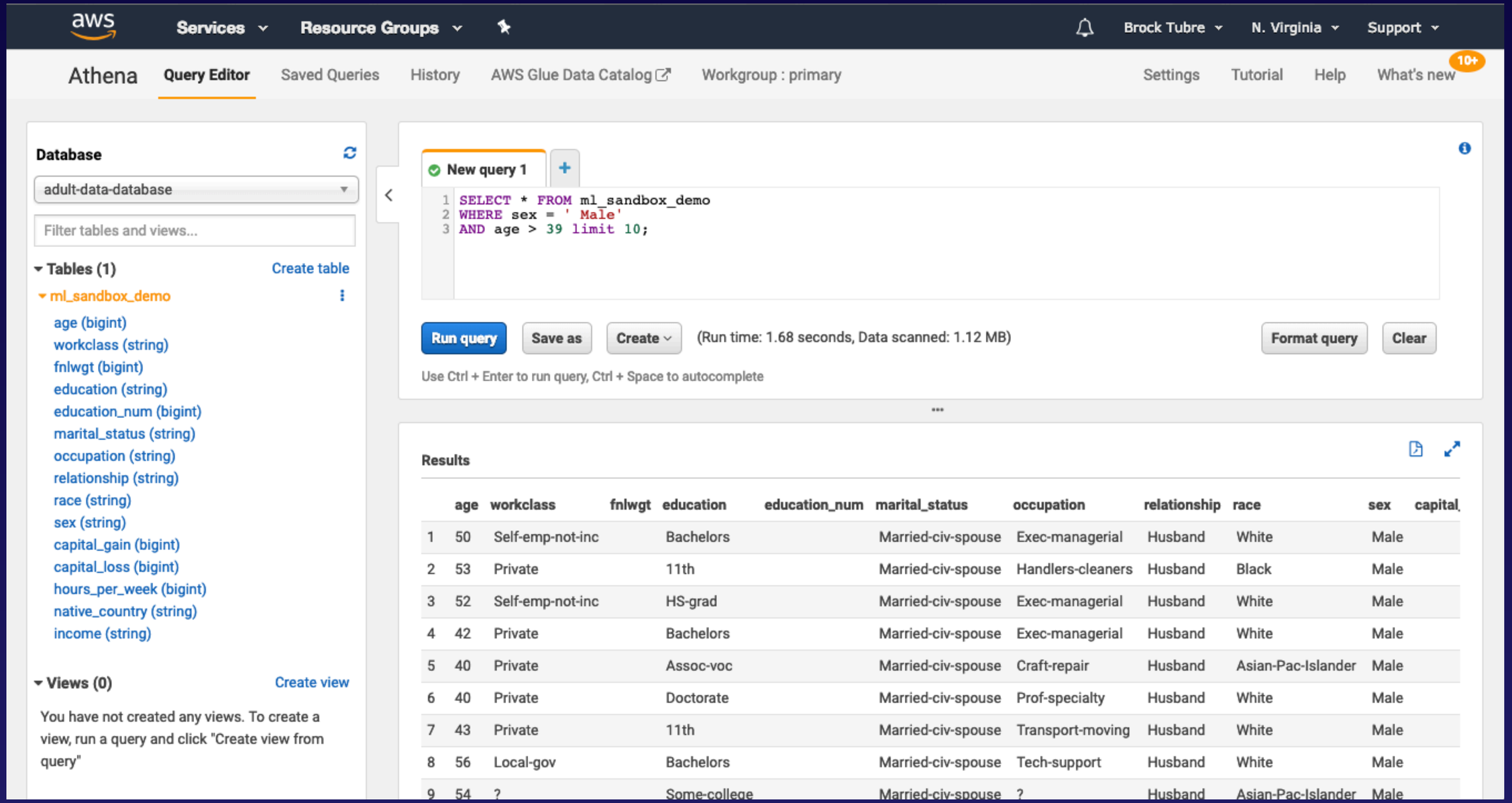
Athena


aws
Services
Resource Groups
Brock Tubre
N. Virginia
Support
Athena
Query Editor
Saved Queries
History
AWS Glue Data Catalog
Workgroup : primary
Settings
Tutorial
Help
What's new
10+
Database
adult-data-database
Filter tables and views...
Tables (1)
Create table
ml_sandbox_demo
age (bigint)
workclass (string)
fnlwgt (bigint)
education (string)
education_num (bigint)
marital_status (string)
occupation (string)
relationship (string)
race (string)
sex (string)
capital_gain (bigint)
capital_loss (bigint)
hours_per_week (bigint)
native_country (string)
income (string)
Views (0)
Create view
You have not created any views. To create a view, run a query and click "Create view from query"
New query 1
1
Run query
Save as
Create
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results

Athena


aws
Services
Resource Groups
Brock Tubre
N. Virginia
Support
Athena
Query Editor
Saved Queries
History
AWS Glue Data Catalog
Workgroup : primary
Settings
Tutorial
Help
What's new
10+
Database
adult-data-database
Filter tables and views...
Tables (1)
Create table
ml_sandbox_demo
age (bigint)
workclass (string)
fnlwgt (bigint)
education (string)
education_num (bigint)
marital_status (string)
occupation (string)
relationship (string)
race (string)
sex (string)
capital_gain (bigint)
capital_loss (bigint)
hours_per_week (bigint)
native_country (string)
income (string)
Views (0)
Create view
You have not created any views. To create a view, run a query and click "Create view from query"
New query 1
1
Run query
Save as
Create
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results

Athena



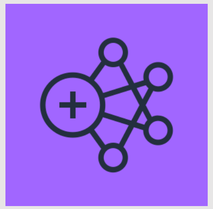
	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital
1	50	Self-emp-not-inc		Bachelors		Married-civ-spouse	Exec-managerial	Husband	White	Male	
2	53	Private		11th		Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
3	52	Self-emp-not-inc		HS-grad		Married-civ-spouse	Exec-managerial	Husband	White	Male	
4	42	Private		Bachelors		Married-civ-spouse	Exec-managerial	Husband	White	Male	
5	40	Private		Assoc-voc		Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	
6	40	Private		Doctorate		Married-civ-spouse	Prof-specialty	Husband	White	Male	
7	43	Private		11th		Married-civ-spouse	Transport-moving	Husband	White	Male	
8	56	Local-gov		Bachelors		Married-civ-spouse	Tech-support	Husband	White	Male	
9	54	?		Some-college		Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	

Data Pipeline

Data Preparation

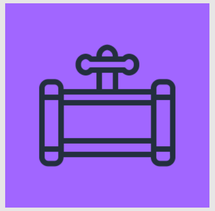
AWS Glue

SageMaker



EMR

Athena

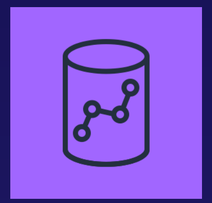


Data Pipeline

Data Pipeline

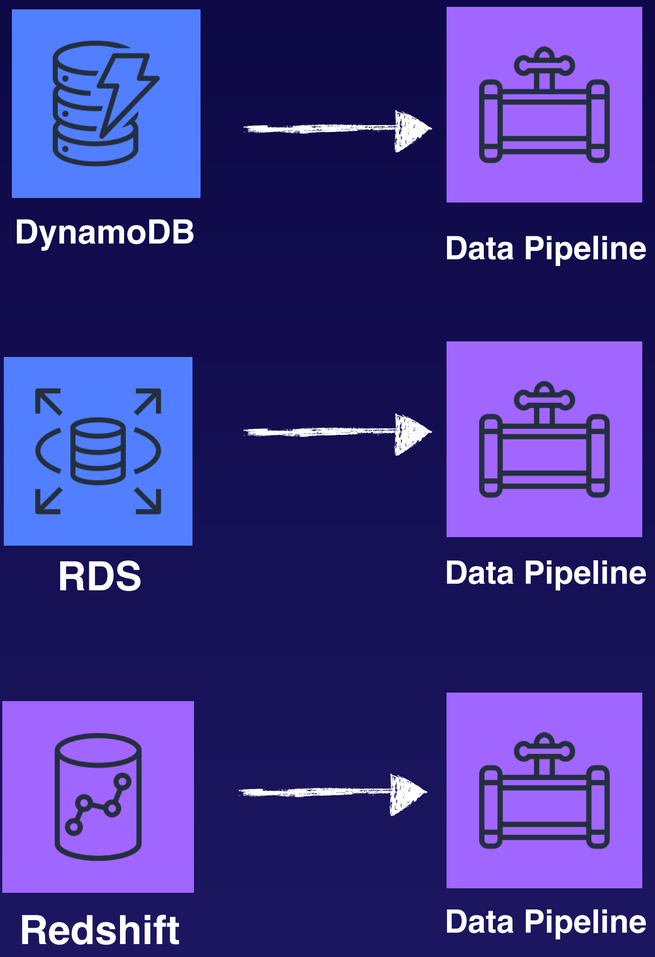
DynamoDB

RDS

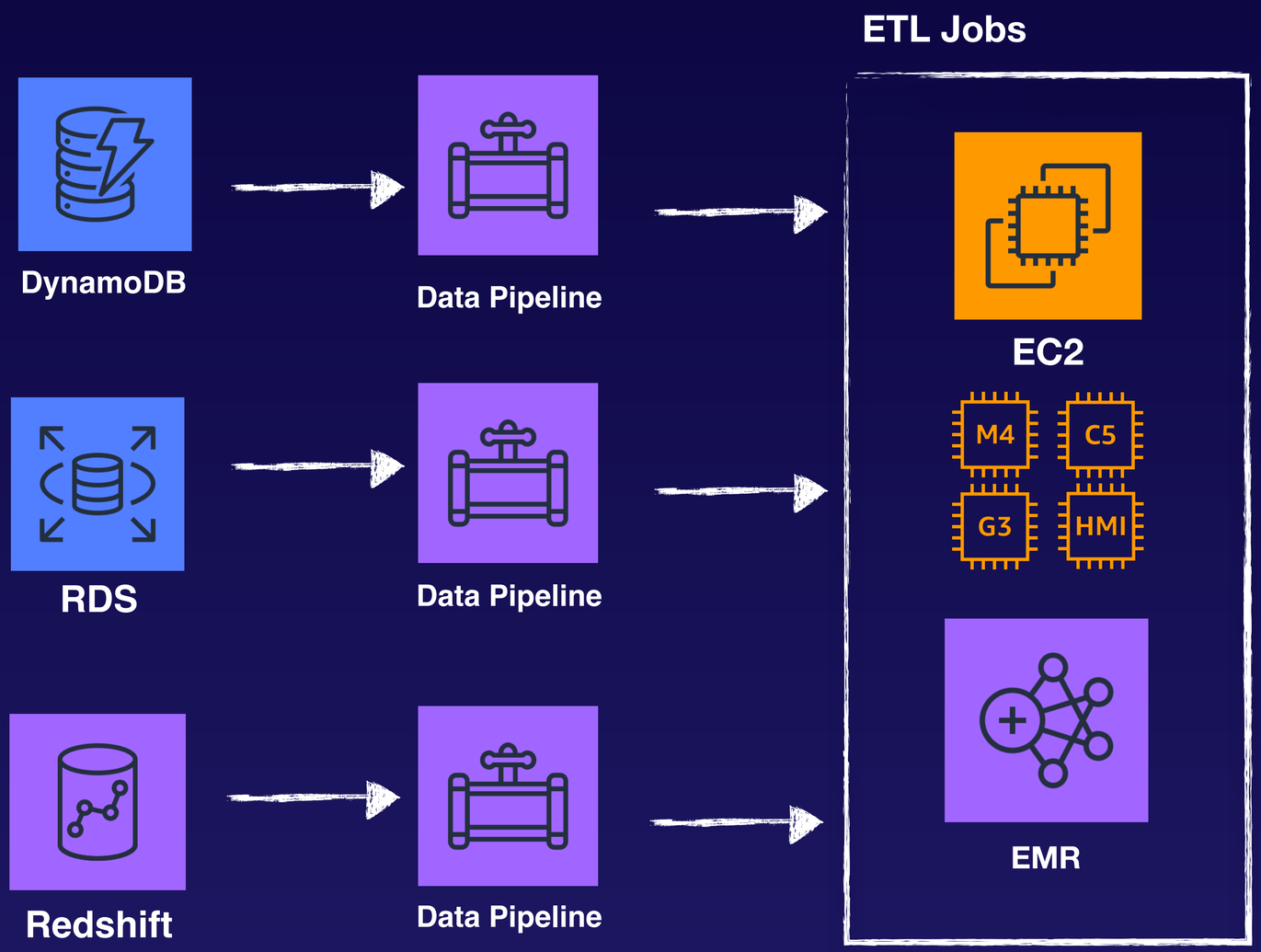


Redshift

Data Pipeline



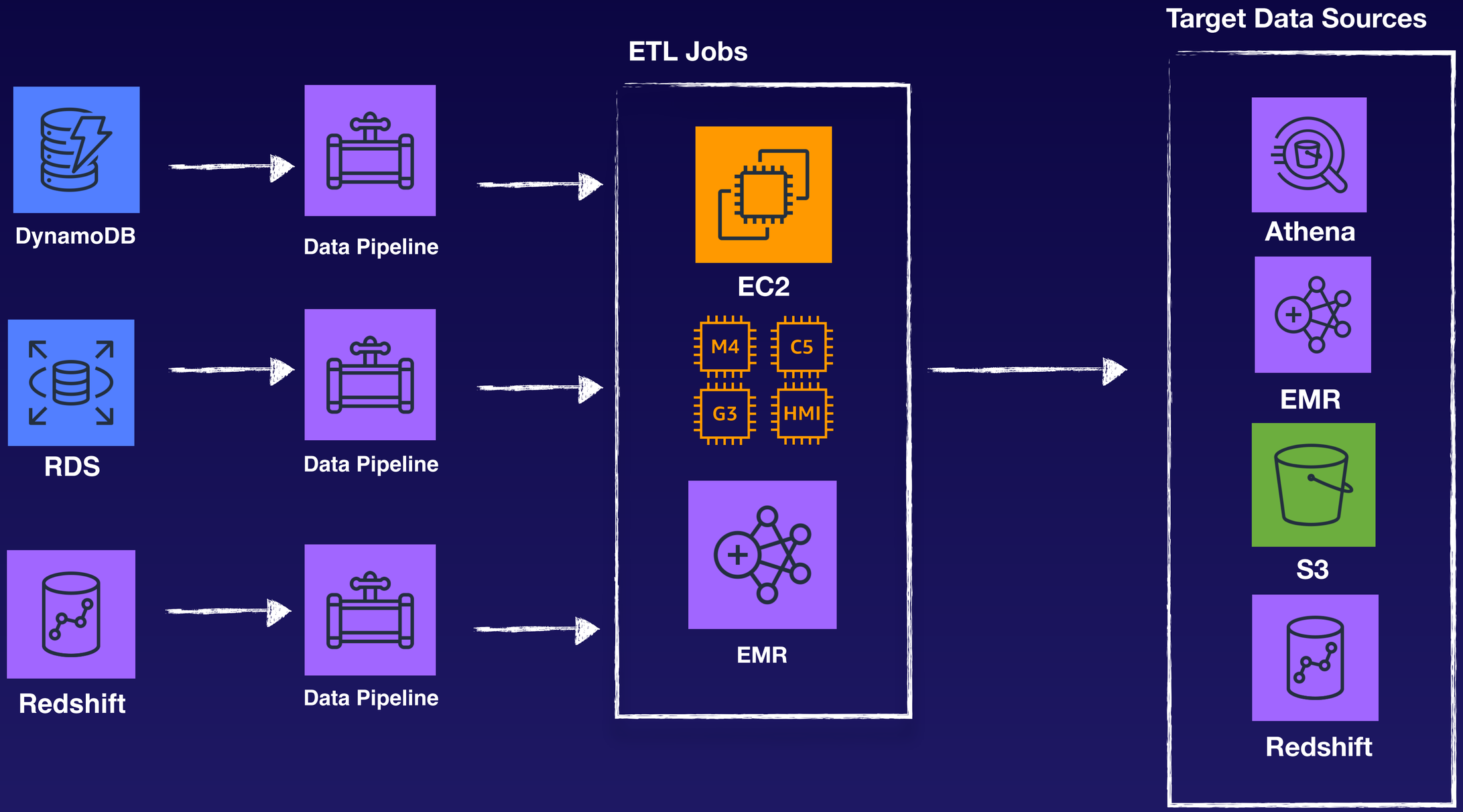
Data Pipeline



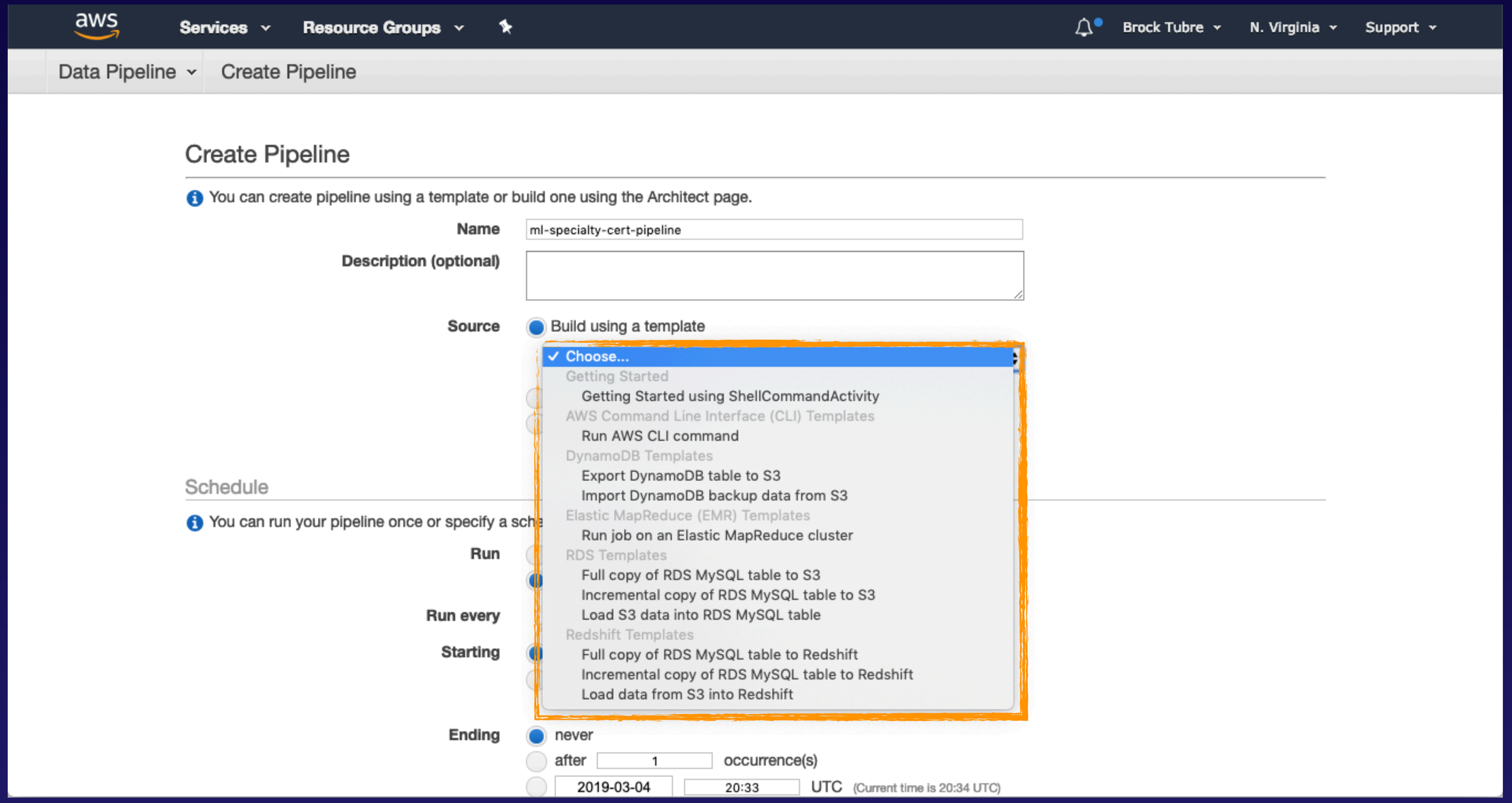
AWS DATA PREPARATION HELPER TOOLS

Data Pipeline

A CLOUD GURU



Data Pipeline



Which service to use?

Datasource	Data Preparation Tool	Why
S3, Redshift, RDS, DynamoDB, On Premise DB	AWS Glue	Use Python or Scala to transform data and output data into S3
S3	Athena	Query data and output results into S3
EMR	PySpark/Hive in EMR	Transform petabytes of distributed data and output data into S3
RDS, EMR, DynamoDB, Redshift	Data Pipeline	Setup EC2 instances to transform data and output data into S3