AWS GLUE

Slides from AWS Popup Lift

Simple, flexible, cost-effective ETL

- AWS Glue is a fully managed ETL (extract, transform, and load) service
- Categorize your data, clean it, enrich it and move it reliably between various data stores
- Once catalogued, your data is immediately searchable and queryable across your data silos
- Simple and cost-effective
- Serverless; runs on a fully managed, auto-scaling Spark environment

Why would AWS get into the ETL space?

We have lots of ETL partners

Amazon Redshift Partner Page for Data Integration

ATTUNITY CloudBeam

ironSource .atom

infoteria

bryte Systems

alooma

informatica Cloud

talend integration at any scale

MATILLION

Segment

Fivetran

etleap

xplenty

FlyData

TREASURE DATA

Qubole

RJMETRICS

snapLogic

but...Customers are still hand-coding ETL…

70% of ETL jobs are hand-coded

With no use of ETL tools

Actually…

It’s over 90% in the cloud

Why do we see so much hand-coding?

Code is flexible | Code is powerful

You can unit test | You can deploy with other code | You know your dev tools

Hand-coding involves a lot of undifferentiated heavy lifting…

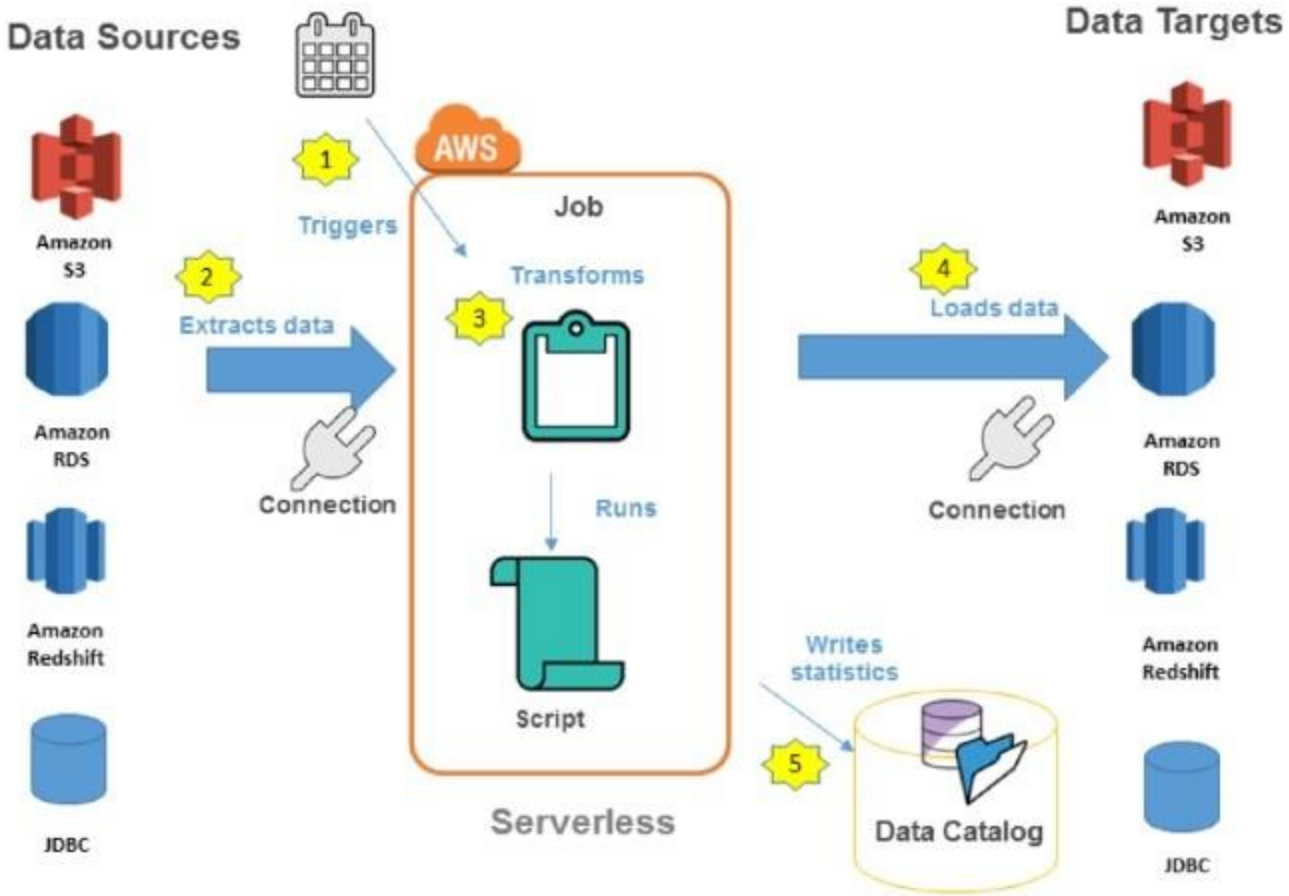
Brittle | Error-prone | Laborious

- As data formats change
- As target schemas change
- As you add sources
- As data volume grows

Glue automates the undifferentiated heavy-lifting of ETL

- ✓ Discover and organize data, regardless of where it lives
- ✓ Focus on writing transformations, not handling undifferentiating heavy lifting
- ✓ ETL jobs run under a Serverless execution model

AWS Glue: big picture



AWS Glue: components

Data Catalog

- Discover and Organize your data in various databases, data warehouses and data lakes

Job Authoring

- Focus on the writing transformations
  - Generate code through a wizard
  - Write your own code

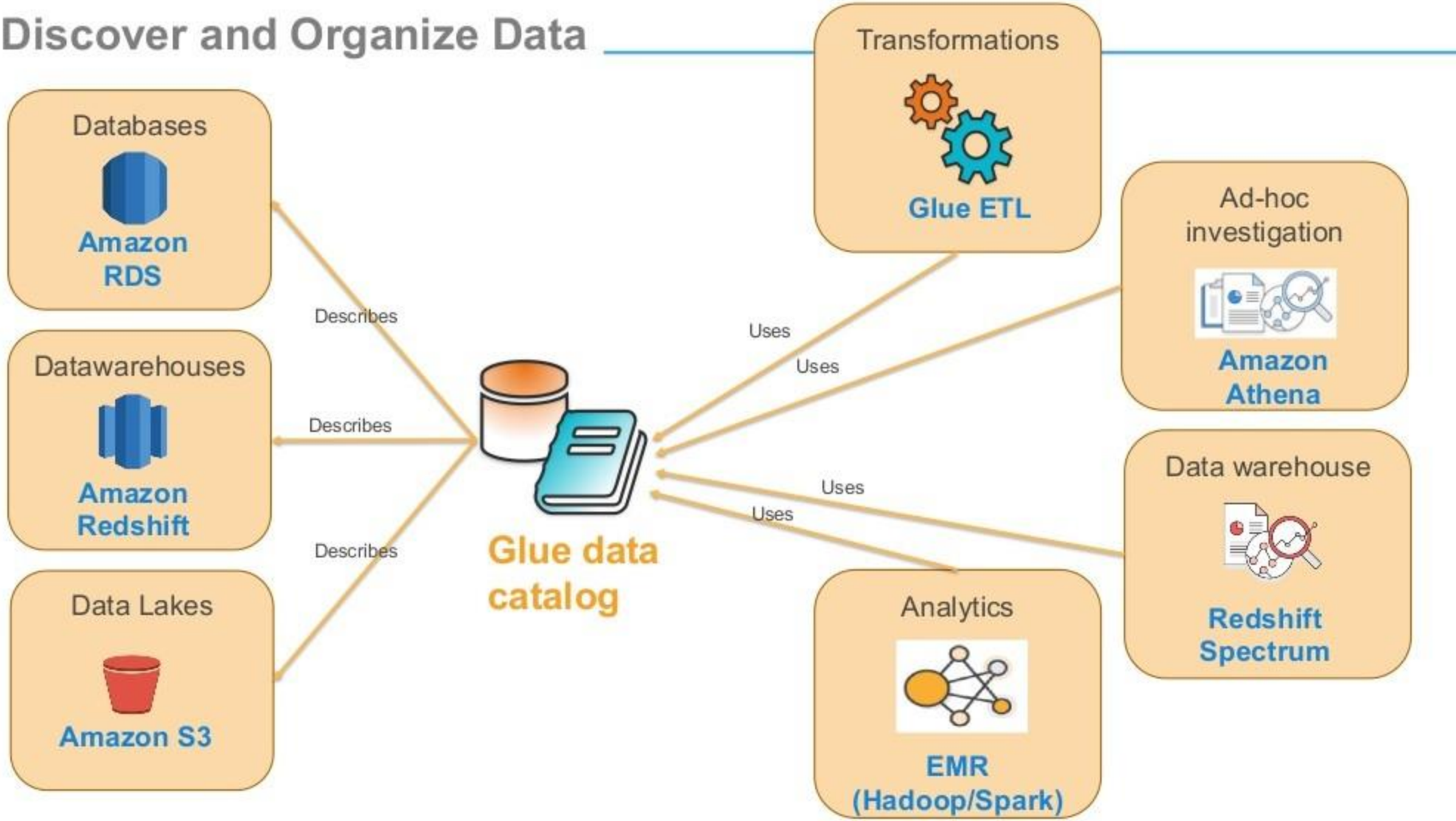
Job Execution

- Runs jobs in Spark containers – automatic scaling based on SLA
- Glue is serverless – only pay for the resources you consume

Glue data catalog

Discover and organize your data sets

Discover and Organize Data

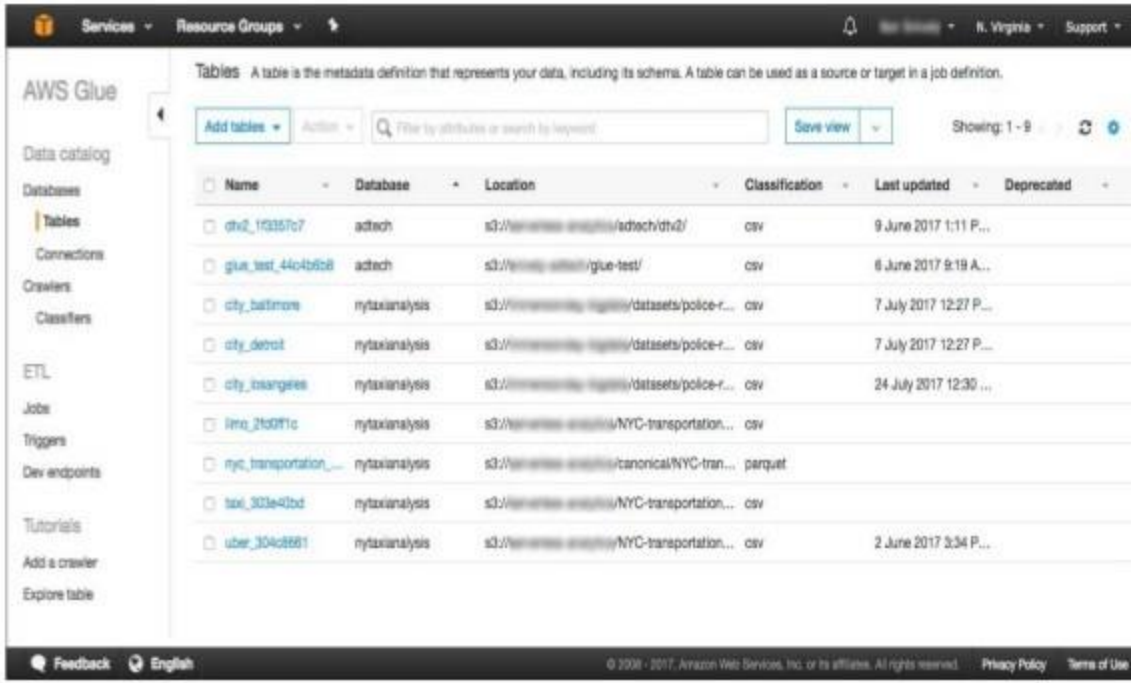


Glue data catalog

Manage table metadata through a Web Interface, Hive metastore API, Hive SQL, or automated through crawlers

Listening to our customers, we've added:

- **Search** metadata for data discovery
- **Connections** to RDS, Redshift and JDBC
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas



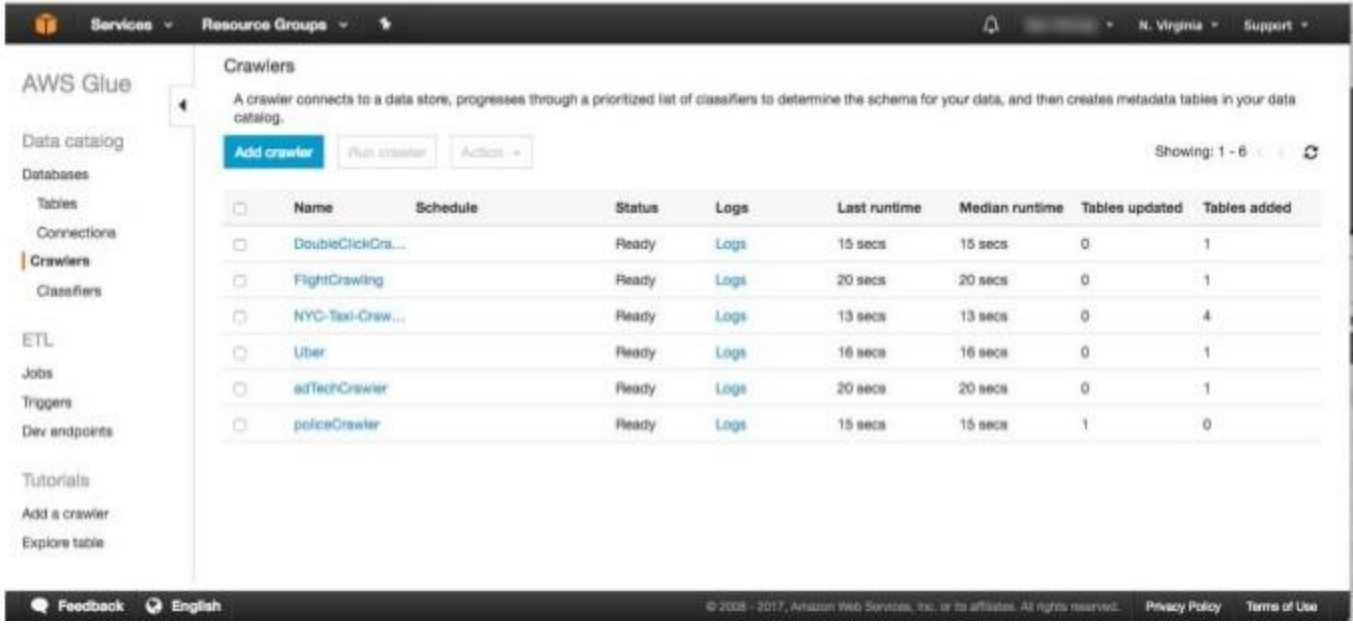
Crawlers: auto-populate data catalog

Run crawlers on-demand and on a schedule to **discover** new data and **schema changes.**

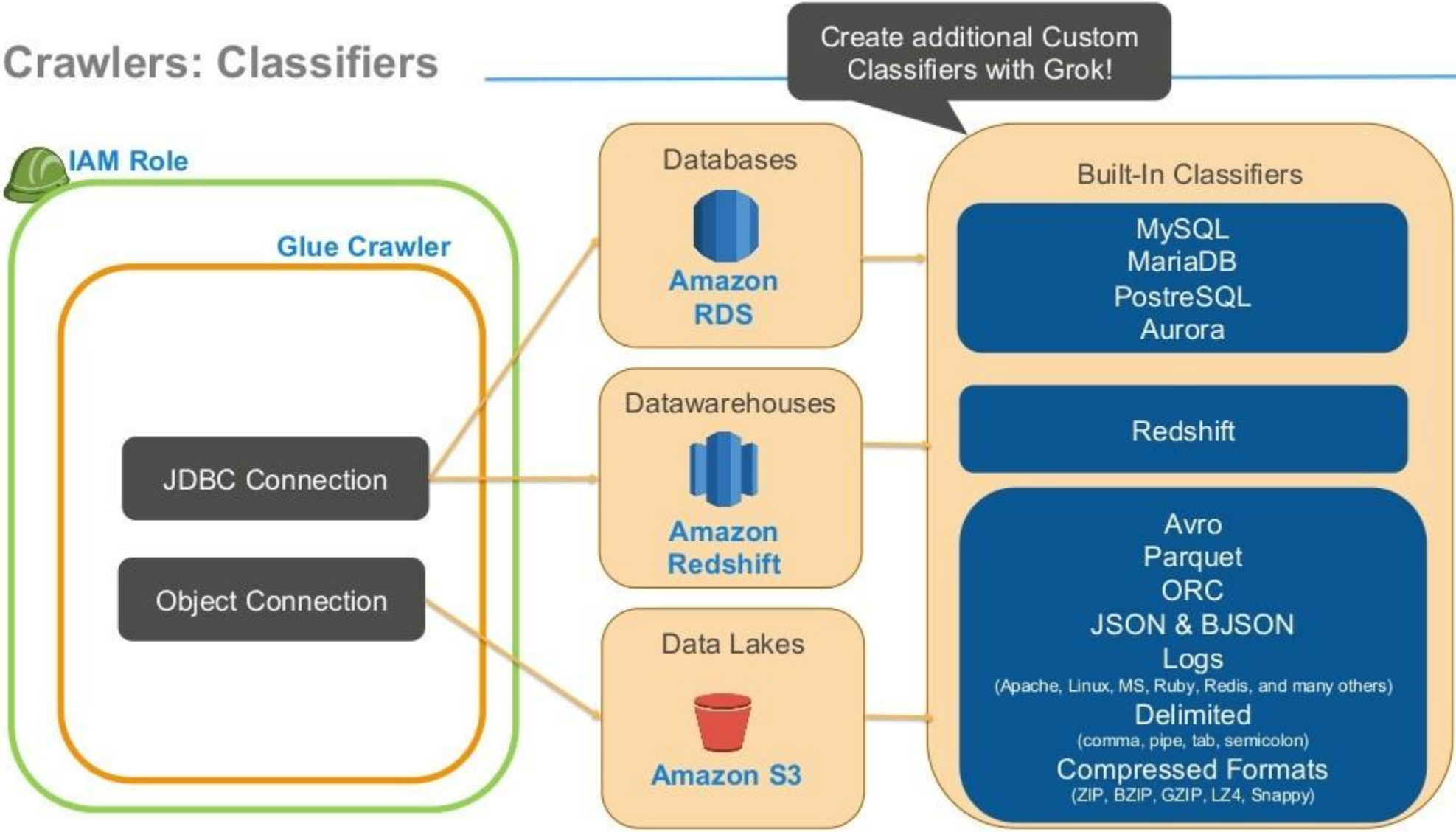
Serverless – only pay when crawls run.

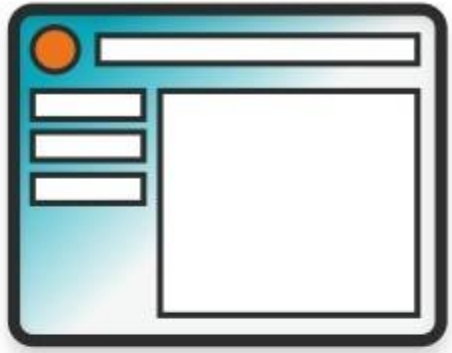
Automatic schema inference:

- Built-in classifiers
  - **Detect file type**
  - **Extract schema**
  - **Identify partitions**
- Add your own classifiers
  - Grok for each of use



Crawlers: Classifiers



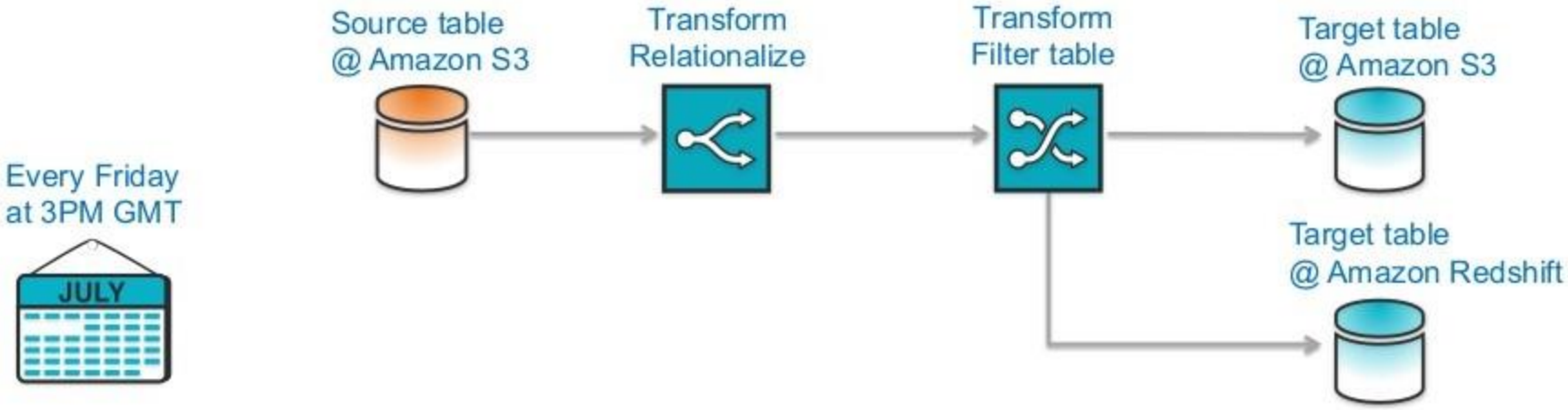


Job authoring in Glue

You have choices on how to get started…

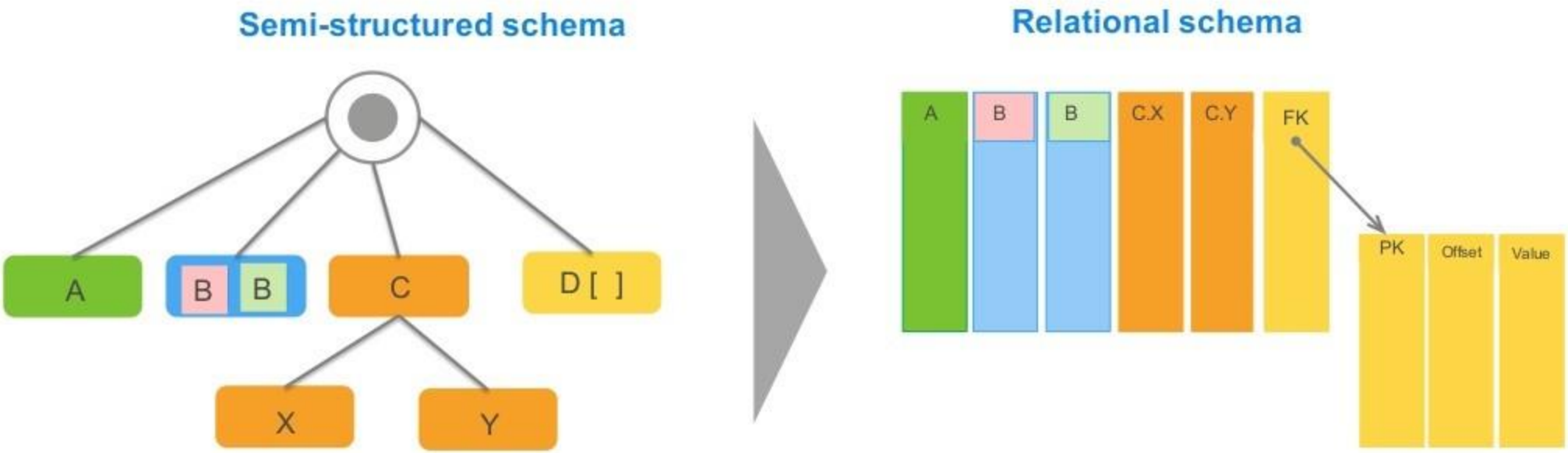
- Script generated by AWS Glue
- Existing script brought into AWS Glue
- Blank script authored by you

Automatic Code Generation



1. Pick sources and targets from the data catalog
2. Glue generates transformation graph and *Python code*
3. Specify trigger condition and execute job

Glue transformations are flexible and adaptive



- Flatten semi-structured objects with arbitrary complexity into relational tables, on-the-fly.
- Pivot arrays and other collection types into a separate table, generating key-foreign key values.
- Modify mapping as the source schema changes, and modify the target schemas as needed.

Glue ETL scripts are forgiving and flexible

- **Human-readable** code run on a scalable platform, PySpark
- **Forgiving** in the face of failures – handles bad data and crashes
- **Flexible**: handles complex semi-structured data, and adapts to source schema changes

Add Custom Modules and Files

- Add external Python modules
- Java JARs required by the script
- Additional files such as configuration, etc.

Parameters (optional)

▸ Advanced properties

▾ Script libraries and job parameters (optional)

☐ Server-side encryption

Python library path

s3://bucket-name/folder-name/file-name

Dependent jars path

s3://bucket-name/folder-name/file-name

Referenced files path

s3://bucket-name/folder-name/file-name

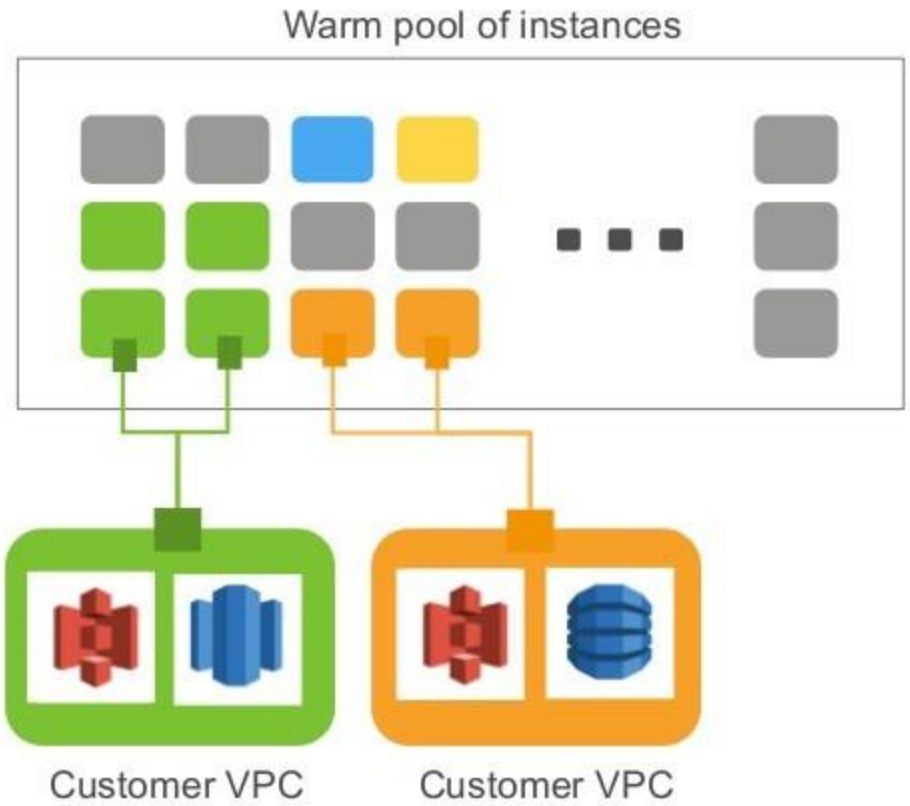
Orchestration & resource management

Fully managed, serverless job execution

Serverless job execution

There is no need to provision, configure, or manage servers

- Warm pools: pre-configured fleets of instances to reduce job startup time
- Auto-configure VPC and role-based access
- Automatically scale resources to meet SLA and cost objectives
- You pay only for the resources you consume while consuming them.



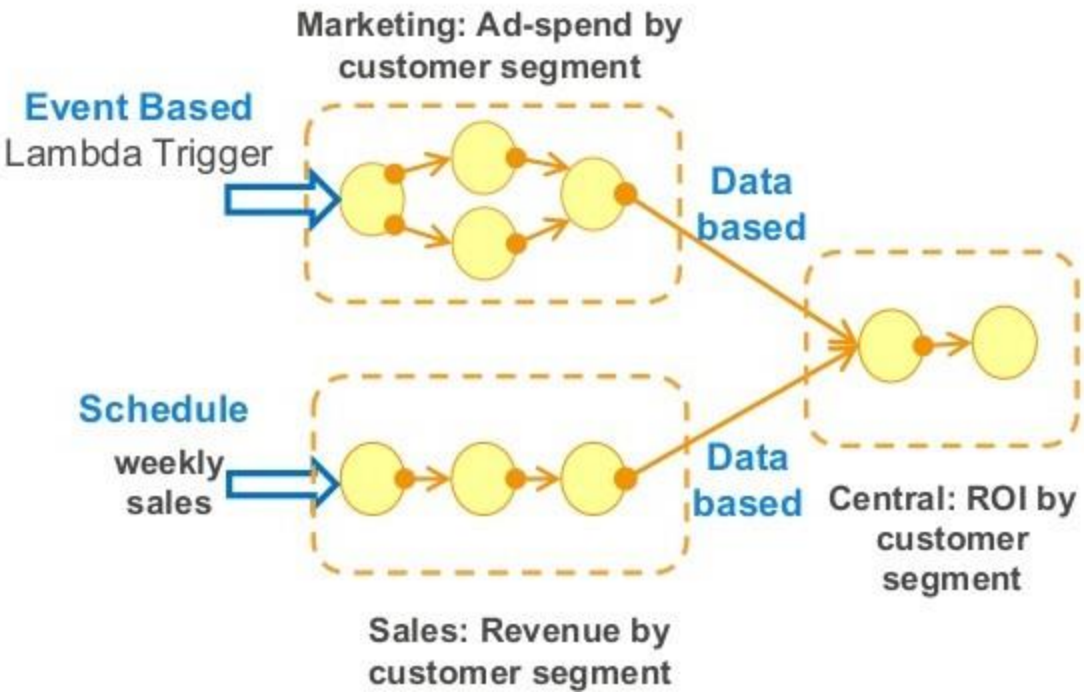
Job composition and triggers

Compose jobs globally with event-based dependencies

- Easy to reuse and leverage work across organization boundaries

Multiple triggering mechanisms

- **Schedule-based:** e.g., time of day
- **Event-based:** e.g., data availability, job completion
- **External sources:** e.g., AWS Lambda

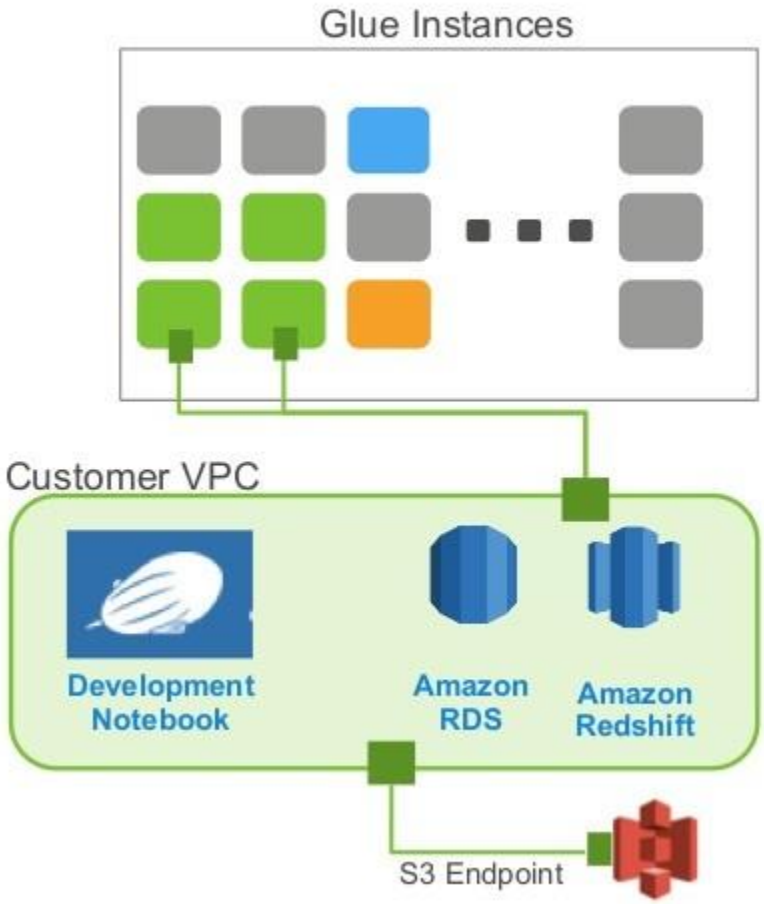


Developer Endpoints

Environment to iteratively develop and test ETL scripts.

Develop your script in a notebook and point to an AWS Glue endpoint to test it.

When you are satisfied with the results of your development process you can create an ETL job that runs your script.



Glue ETL Security Model

- ETL Jobs that do not require special handling can be launched without additional configuration
- For jobs requiring restricted access to data Glue utilizes a VPC endpoint launching dedicated ENIs into the customer's VPC which are assigned private IP address from the customer's subnet.
- For such jobs, Glue also enforces the use of an S3 VPC endpoint

Usage and Pricing

Glue ETL Pricing

- With Glue, you only pay for the time your ETL job takes to run.
  - There are no resources to manage and no upfront costs, and you are not charged for startup or shutdown time.
- You are charged an hourly rate based on the number of Data Processing Units (or DPUs) used to run your ETL job.
  - A single Data Processing Unit (DPU) provides 4 vCPU and 16 GB of memory and corresponding networking capabilities.
- You are billed per hour in increments of 1 minute, rounded up to the nearest minute, with a 10-minute minimum duration for each job.

Glue Data Catalog and Crawler Pricing

Data catalog:

- With the AWS Glue data catalog, you can store up to a million objects per month for free. If you store more than a million objects, you will be charged per 100,000 objects over a million.
  - An object in the AWS Glue data catalog is a table, a partition, or a database.

- The first million access requests per month to the AWS Glue data catalog are free. If you exceed a million requests in a month, you will be charged per million requests over the first million.
  - Some common requests are CreateTable, CreatePartition, GetTable and GetPartitions.

Crawlers:

- You will pay an hourly rate for AWS Glue crawler runtime to populate the Glue data catalog, based on the number of Data Processing Units (or DPUs) used to run your crawler.
  - A single Data Processing Unit (DPU) provides 4 vCPU and 16 GB of memory and corresponding networking capabilities.
  - You are billed in increments of 1 minute, rounded up to the nearest minute.
- Use of AWS Glue crawlers is optional, and you can populate the Glue data catalog directly through the API.