

Cloud computing

Lecture 2

Ramesh Shankar

UConn

Taxi dataset

- Column names (yellow)
- Raw data (white, below)
- Comma-separated

```
1 medallion,hack_license,vendor_id,rate_code,store_and_fwd_flag,pickup_datetime,dropoff_datetime,passenger_count,trip_time_in_secs,trip_distance,pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
2 89D227B655E5C82AECF13C3F540D4CF4,BA96DE419E711691B9445D6A6307C170,CMT,1,N,2013-01-01 15:11:48,2013-01-01 15:18:10,4,382,1.00,-73.978165,40.757977,-73.989838,40.751171
3 0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-06 00:18:35,2013-01-06 00:22:54,1,259,1.50,-74.006683,40.731781,-73.994499,40.75066
4 0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-05 18:49:41,2013-01-05 18:54:23,1,282,1.10,-74.004707,40.73777,-74.009834,40.726002
5 DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:54:15,2013-01-07 23:58:20,2,244,.70,-73.974602,40.759945,-73.984734,40.759388
6 DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:54:15,2013-01-07 23:58:20,2,244,.70,-73.974602,40.759945,-73.984734,40.759388
```

AWS Glue – Create (schema) database and table

- <https://www.youtube.com/watch?v=qNojanBn1NY>
- Data catalog – read your raw file (in S3) and attach a schema (field names, data types) – preferably automatically

Your problem...

- So you now have your massive datasets loaded on S3
- How do you query it?
- Consider **AWS Glue**

What is AWS GLUE:

- Glue is a serverless, fully managed, cloud-optimized, Extract-Transform-Load (ETL) service
- Glue automatically infers data format, schema and partitions for semi-structured and structured data
 - It means you may not have to create tables or type column names
- You run Glue, point to a dataset on S3, and Glue will create a **table** out of that data, in the **Glue Data Catalog**, which you can query in **AWS Athena**.
 - So, the data is in S3
 - Glue runs a **crawler** that runs through your data and infers schemas and partitions
 - The crawler saves the schema in the Glue Data Catalog
 - You query the schema using Athena, and view/export the data summary extract

AWS Glue Console

https://console.aws.amazon.com/glue/home?region=us-east-1#

aws Services Resource Groups

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Settings

ETL

Crawlers A crawler connects to a data store, p

Add crawler

Run crawler

Action

| Name | Schedule |
|------|----------|
|------|----------|

AWS Glue Console

https://console.aws.amazon.com/addCrawler:

aws Services Re:

Add crawler

Add information about

- Crawler info
- Data store
- IAM Role
- Schedule
- Output
- Review all steps

Crawler name

taxi-crawler

Description, security configuration, and classifiers

Grouping behavior for S3 data (optional)

Next

Add a data store

Choose a data store

1

S3

Crawl data in

- ☒ Specified path in my account
☐ Specified path in another account

Include path

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

All folders and files contained in the include path are crawled. For example, type `s3://MyBucket/MyFolder/` within MyBucket.

- Exclude patterns (optional)

Back

Next

4

Choose S3 path

2

S3

nyctaxi-jan-2013-trip1

Select

Add another data store

- ☐ Yes
☒ No

Back

Next


Glue recognizes many of the standard data formats, including CSV and JSON

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

- ☐ Update a policy in an IAM role
- ☒ Choose an existing IAM role
- ☐ Create an IAM role

IAM role ⓘ



AWSGlueServiceRole-DefaultRole

ramesh-glue

- s3://nyctaxi-jan-2013-trip1

You can also create an IAM role on the [IAM console](#).

Back

Next

Create a schedule for this crawler

Frequency



Back

Next

Configure the crawler's output

Database ⓘ

default

Add database

Prefix added to tables (optional) ⓘ

Type a prefix added to table names

► Configuration options (optional)

Back

Next

Add database

Database name

taxitrips

► Description and location (optional)

Create

Crawler info

Name taxi-crawler
Create a single schema for each S3 path false

Data stores

Data store S3
Include path s3://nyctaxi-jan-2013-trip1
Exclude patterns

IAM role

IAM role arn:aws:iam::371521246937:role/service-role/AWSGlueServiceRole-DefaultRole

Schedule

Schedule Run on demand

Output

Database taxitrips
Prefix added to tables (optional)
▸ Configuration options

Back

Finish



AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Settings

ETL

Jobs

Triggers

Dev endpoints

Crawlers A crawler connects to a data store, progresses through a priori

Crawler **taxi-crawler** was created to run on demand. [Run it now?](#)

Add crawler

Run crawler

Action

Filter by attributes



Name



taxi-crawler

Crawlers A crawler connects

atalog.

[User preferences](#)

Add crawler

Run crawler

Showing: 1 - 1



Name



taxi-crawler

Tables added

1

Running it will take 1 or 2 mins

aws Services ▾ Resource Groups ▾

AWS Glue

Data catalog

- Databases** 1
- Tables
- Connections
- Crawlers
- Classifiers
- Settings

Databases A database is a set of associated tables

Add database View tables 3 Action ▾

| <input type="checkbox"/> | Name |
|-------------------------------------|-------------|
| <input type="checkbox"/> | default |
| <input checked="" type="checkbox"/> | taxitrips 2 |

aws Services ▾ Resource Groups ▾

AWS Glue

Data catalog

- Databases
- Tables**
- Connections
- Crawlers
- Classifiers
- Settings

Tables A table is the metadata definition that represents your data, including its schema. A table

Add tables ▾ Action ▾ Database : taxitrips × Filter or search for tables...

| <input type="checkbox"/> | Name | Database | Location |
|--------------------------|------------------------|-----------|------------------------------|
| <input type="checkbox"/> | nyctaxi_jan_2013_trip1 | taxitrips | s3://nyctaxi-jan-2013-trip1/ |

Tables > nyctaxi_jan_2013_trip1

Edit table

Delete table

Name nyctaxi_jan_2013_trip1

Description

Schema

| | Column name | Data type |
|---|--------------|-----------|
| 1 | medallion | string |
| 2 | hack_license | string |
| 3 | vendor_id | string |
| 4 | rate_code | bigint |

Last updated 31 Jan 2019

Table Version (Current version) ▼

View properties

Compare versions

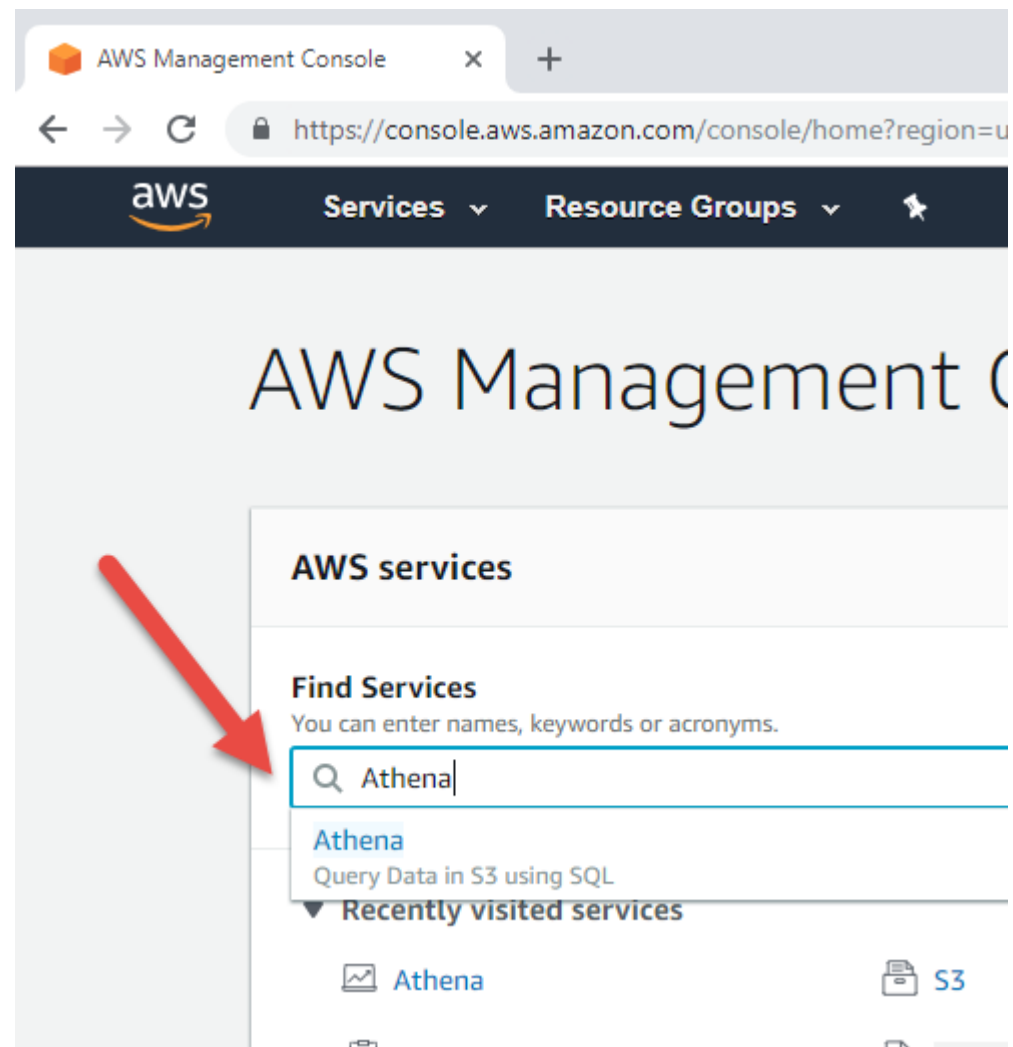
Edit schema



Showing: 1 - 14 of 14 < >

AWS Athena

- Now that you have the data schema in the Data Catalog, you can query it using Athena



Athena

Services Resource Groups

Athena Query Editor Saved Queries History AWS Glue Data Catalog

Database

taxitrips

Filter tables and views...

Tables (1) Create table

nyctaxi_jan_2013_trip1

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

Preview table
Show properties
Delete table
Generate Create Table DDL

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

✓ New query 1



```
1 SHOW CREATE TABLE nyctaxi_jan_2013_trip1;
```

Run query

Save as

Create ▾

(Run time: 1.22 seconds, Data scanned: 0 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

```
CREATE EXTERNAL TABLE `nyctaxi_jan_2013_trip1`(  
  `medallion` string,  
  `hack_license` string,  
  `vendor_id` string,  
  `rate_code` bigint,  
  `store_and_fwd_flag` string,  
  `pickup_datetime` string,  
  `dropoff_datetime` string,  
  `passenger_count` bigint,  
  `trip_time_in_secs` bigint,  
  `trip_distance` double,  
  `pickup_longitude` double,  
  ...
```



✓ New query 1



```
1 select * from nyctaxi_jan_2013_trip1 limit 5;  
2  
3
```

1

2

Run query

Save as

Create ▾

(Run time: 1.61 seconds, Data scanned: 3.66 MB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

| | medallion | hack_license | vendor_id | rate |
|---|----------------------------------|----------------------------------|-----------|------|
| 1 | 6C8C5507F1928059FBBCB7E4C7D3627A | E68694776C382CBB6160D15E94844105 | CMT | 1 |
| 2 | A3281E8510FED7EE0371C2690A243880 | 40BBE16556A9F0F8CF3A4C01F9C5F29F | CMT | 1 |
| 3 | 927C59F57F43537DA492555F5B557326 | 79AB34B5FBC943E315068D01C6A9D8A9 | CMT | 1 |
| 4 | FEBFB5478D15AE3E06E1D0CA674A4C38 | 87C723C9E83E19D8AF0424DFC2865846 | CMT | 1 |
| 5 | 275AF4D0E47451563A4DD853CE352DAC | 334AB3D18DB63C600527E49DE883A624 | CMT | 1 |