Learn Amazon Athena

The Basics

Learn Amazon Athena basics. Amazon Athena is meant for querying copious amounts of data on the cheap. It’s a straightforward service based on Presto with some interesting integrations. The Presto Foundation calls Amazon Athena and Amazon EMR Presto Cloud on their website.

It’s not dirt cheap to use Amazon Athena, but it is convenient. It costs $5.00 per TB of data scanned. It’s possible to pay a third of that with compression. Additional savings occur when querying a single column. The big win with Athena is the cluster of servers that AWS has at the ready. It wouldn’t be straightforward to set up your own Presto analytics engine.

Primary Use Cases

- Querying vast amounts of
  - Log Data
  - Behavioral Data
  - Noncritical data

Importing Data

I’m facing a scenario where I need to query a 58G CSV file. I was able to load the ~1 billion rows into PostgreSQL, but any queries on the data caused my work machine to crash. I wrote a little python script to query the data which is working but is taking too long to process. So, I brought out the big guns.

I decided to try Athena out as a solution to my problem. Here’s how it went down.

I created this table to store my CSV file. This table is for storing the output of an AWS command-line query.

```
DROP TABLE learnamazonathena.s3_objects;
CREATE EXTERNAL TABLE IF NOT EXISTS learnamazonathena.s3_objects (
         time string,
         bytes bigint,
         object string
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE
LOCATION 's3://learn-amazon-athena/csv';
```

Amazon S3 is now the datastore for our CSV file. We generate this CSV file with these commands.

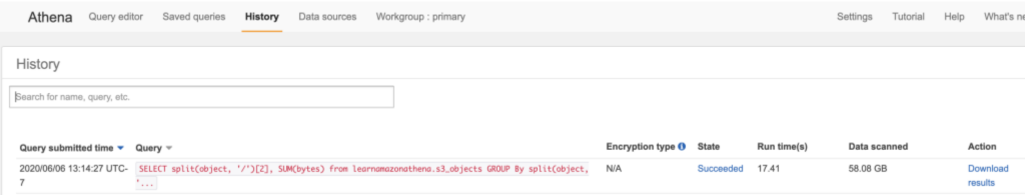
```
# Query all S3 objects recursively
aws s3 ls --recursive learn-amazon-athena > s3_objects.csv

# Format the output as a CSV file
perl -p -i -e
's/(^.{0,19})(\s+)([0-9]+)(\s)(.+)$/"\1","\3","\5"/g'
applications.csv
```

I then ran this query against the newly created table.

```
SELECT split(object,
         '/')[2], SUM(bytes)
FROM learnamazonathena.s3_objects
GROUP BY  split(object, '/')[2]
```

This query took seventeen seconds to complete.



Query submitted time	Query	Encryption type	State	Run time(s)	Data scanned	Action
2020/06/06 13:14:27 UTC-7	SELECT split(object, '/')[2], SUM(bytes) from learnamazonathena.s3_objects GROUP By split(object, '...	N/A	Succeeded	17.41	58.08 GB	Download results

I love how fast and straightforward Amazon Athena is. The CSV file I created is small in the grand scheme of things. Facebook created Presto to query obscene amounts of data then later open-sourced it.

Learn Amazon Athena – Beyond the Basics

- **Print**
  - Getting Started – User Guide (AWS)
  - Amazon Athena – User Guide (AWS)
  - Athena FAQ (AWS)
- **Video**
  - Amazon Web Services: Data Analytics (LinkedIn)