AWS Athena Intro

About me

- I’ve been around data for 20+ years
- Owner of Inventive Data Solutions
- bobhaffner on LinkedIN, Medium and Twitter
- AWS Certified Solution Architect - Associate
- AWS Certified Big Data Speciality

Polls

- S3?
- SQL?
- Athena or another distributed SQL engine?

Athena Intro

- Interactive, Ad-Hoc tool for querying your S3 files with SQL
- Serverless
- Based on Presto
- In the distributed SQL engine class of tools
- Redshift Spectrum and S3 Select
- Structured, Semi-Structured and Unstructured(?) data

Athena Intro (cont'd)

- Much like Presto and Hive you have to create schemas/tables
- Supports CSV, JSON and columnar storage formats like Parquet and ORC
- Supports data partitioning
- ODBC and JDBC support
- Pay-Per-Query pricing model ($5/TB scanned)

- AWS Console
- DataGrip

SQL Editor

BI Tool

- AWS QuickSight
- Tableau
- Cognos

JDBC or ODBC

Amazon Athena

- Tabular files
- Log files
- JSON files
- Parquet and ORC

Amazon Simple Storage Service (S3)

Demo Data

- NYC Taxi Data
- 6 Months
- Roughly 1.5 million rows
- Pickup and dropoff datetimes
- Pickup and dropoff coordinates
- Basic info about each trip

DEMO

What we covered

- Quick and easy to start querying your data
- Create Tables, Selects and Create Table as Select (CTAS)
- Columnar storage format like Parquet
- Partitioning to reduce scans/save money
- Spatial queries

What we didn’t

- Athena
  - Complex joins and Window functions
  - Complex data types like Arrays and Structs
  - Bucketing
- Glue
  - Classifiers
  - Jobs

In Closing

- Git Repo
  - Slides
  - SQL
  - The Glue job I used to create that partitioned dataset
  - Other Resources

Thank You!