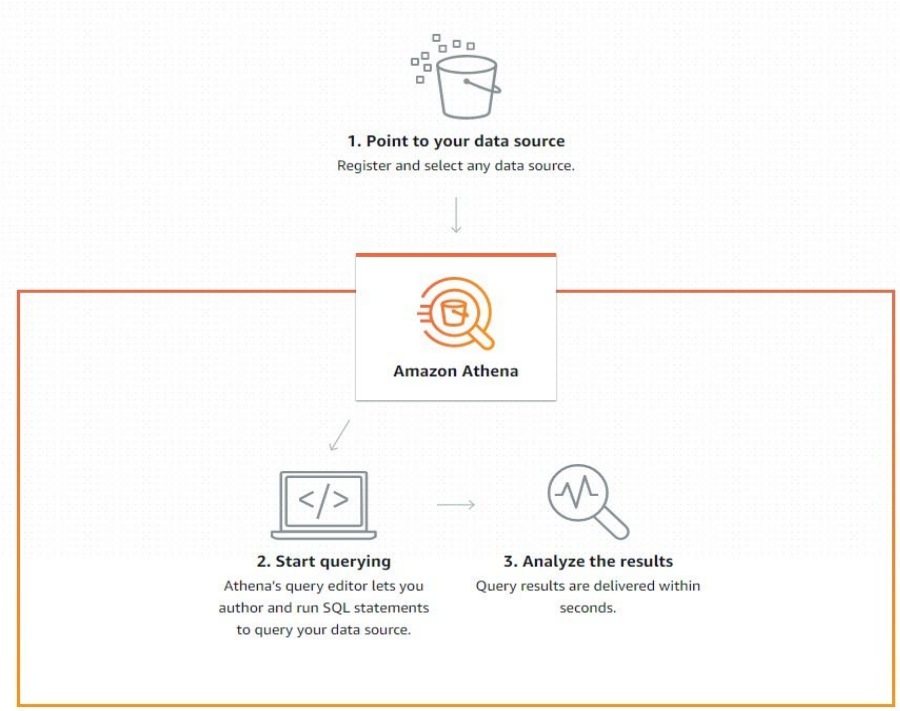
AWS Athena

Group -10
Team: Ayushi Ameta, Ananya Rao and Ashwin Meena

What is Athena

- Amazon Athena is an interactive query service used to analyze data directly from Amazon S3 using standard SQL easily.
- Athena is serverless, so no infrastructure to set up or manage and you can start analyzing your data immediately.
- Athena is not a database - It is a query engine that uses SQL as the query definition language.



Athena competitors/alternatives

- Amazon Athena service was launched in Nov 2016.
- Similar AWS products:
  - Amazon redshift
  - Amazon EMR
- Market competitors:
  - Google BigQuery
  - Azure Data lake analytics

Why athena

- Simple query engine to query data on S3 buckets.
- Serverless solution.
- Quick ad-hoc querying with no complex ETL operations
- Analyze large datasets using SQL.
- Can also handle complex analysis, including large joins, window functions, and arrays for the data stored in S3.

  Window functions: A window function combines the values of a field (or fields) from a group of rows and return a value for each row in a generated column in the result set.

  Arrays: array is an ordered list consisting of zero or more values of the same data type.

- No administration required.

Features evolved to date in athena

- Amazon QuickSight integration
- AWS Glue integration
- Federated queries for analyzing multiple datasource formats.
- Now Athena also supports user defined functions.
- Supported by AWS step functions and AWS Lambda.

Athena's Architecture - 1

- **Presto** with ANSI SQL support.
- Presto is an open source, distributed SQL query engine optimized for low latency.
- Supports both relational and non relational datasources.
- Can query data where it is stored directly.
- Query execution in parallel
- Supports separation of compute and storage.

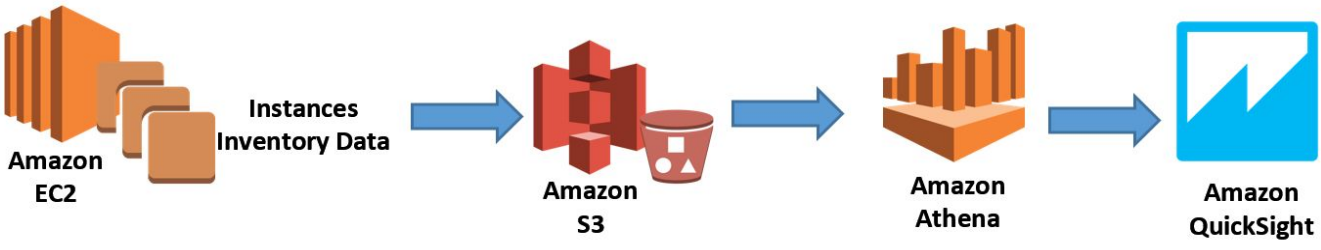
Athena's ARCHITECTURE - 2

- **Apache Hive** to create, drop, and alter tables and partitions.
- Schema-on-read (Hive supports schema-on-read) allows you to project your schema on to your data at the time you execute a query. This eliminates the overhead of loading data.

Athena use cases

Amazon Athena can be used for

- Archival log analysis
- Quickly check new datasets for validity
- Athena is a great tool for data scientists to train ML models.



Athena Customers

AWS customers running Presto on Amazon EMR, got an opportunity to eliminate Presto and adopt Amazon Athena to query on the fly to explore data.

Early adopters of Athena:

- OLX
- Movable Ink:
  - Able to query 7 years' worth of data (hundreds of TB).
  - got results at least 50 percent faster.
  - saved nearly $15,000 per month.
- Atlassian

Advantages

- Start querying instantly and interactively
- Built to scale automatically due to serverless design
- Open, powerful, highly available
- Achieves high performance through massively parallel queries
- Read optimization through columnar data storage
- Versatile - supports multiple data formats
- Pay per query

DISADVANTAGES

- Unlike other databases, it doesn't have its own storage layer.
- Speed depends on how data is organized in S3.
- Fluctuating query performance due to queuing.
- Stored procedures are not supported.
- Following SQL statements are not supported.
  - CREATE TABLE LIKE
  - DESCRIBE INPUT and DESCRIBE OUTPUT
  - EXECUTE … USING
  - MERGE
  - UPDATE
- Not suitable for latency sensitive applications

BEST PRACTICES

Best way to structure your data to get optimal performance:

- Partition your data based on columns
- Partition your data into buckets
- Writing queries based on how Presto works.

Code

```
SELECT dest, origin FROM flights WHERE year = 1991
```

Query	Non- Partitioned Table		Cost	Partitioned table		Cost	Savings
	Run time	Data scanned		Run time	Data scanned		
SELECT count(*) FROM lineitem WHERE l_shipdate = '1996-09-01'	9.71 seconds	74.1 GB	$0.36	2.16 seconds	29.06 MB	$0.0001	99% cheaper 77% faster
SELECT count(*) FROM lineitem WHERE l_shipdate >= '1996-09-01' AND l_shipdate < '1996-10-01'	10.41 seconds	74.1 GB	$0.36	2.73 seconds	871.39 MB	$0.004	98% cheaper 73% faster

Usability

- Stream data directly from Amazon S3 by defining your schema using DDL statements and querying directly.
- Query results for each query that runs can be stored in a query result location that you can specify in Amazon S3.
- 34 Actions that enable Athena operations are defined.
  - GetDataBase
  - ListDataBase
  - GetQueryResults
  - StartQueryExecution
- Query history is retained for 45 days.

Cost discussion

- Free: Data Definition Language (DDL) queries, failed queries, cancelled queries.
- Price per query : $5/TB of data scanned or $0.000004768 per MB scanned
- Standard S3 data transfer and usage costs apply.
- Extra rates for using AWS Glue and other AWS services.

COST DISCUSSION - When to use ATHENA?

S3 select

1. Query single object at once.
2. Basic filter out options can be performed before loading data from S3.
3. Supports CSV, JSON and Paraquet.
4. Works only with S3-API (Python boto3 SDK)
5. Just used to filter out data.

Athena

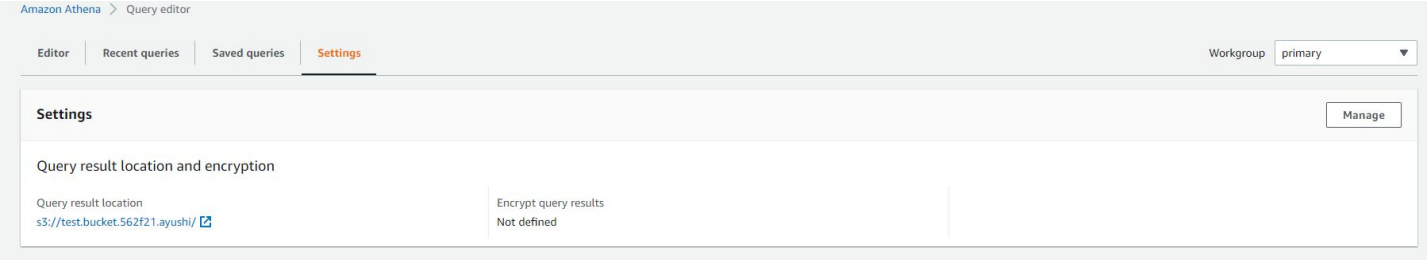
1. Query multiple S3 objects at once.
2. Can encapsulate complex business logic using ANSI compliant queries.
3. Supports CSV, TSV, JSON, Textfiles, Apache ORC, Apache Parquet, Snappy, Zlib, LZO and Gzip.
4. Can query directly from management console or SQL clients using JDBC
5. Allows optimization techniques

Conclusion- Athena in a nutshell

- It is a **query engine** that uses SQL as Run ad-hoc SQL queries on **S3 data** in minutes, supports multiple data formats.
- Results on running a simple query on the sales dataset with 1.5 million rows, 178.5 MB:
  - Query: select * from "test_athena"."my_athena_input_bucket";
  - Runtime: 13.781 seconds
  - Data scanned: 178.51 MB
- Even though slower than Redshift on 8XL nodes, can act like a mini Redshift cluster.

Athena Demo

- AWS Management Console > Athena.
- Query data residing in S3 buckets
  - **Input** s3://athena-examples-us-east-2/cloudfront/plaintext/
  - **Result** s3://test.bucket.562f21.ayushi/
- Specify the Query result location.



Database creation

- To create database, run the query in the query editor

```
CREATE DATABASE mydatabase
```

- Choose the database from list of databases.

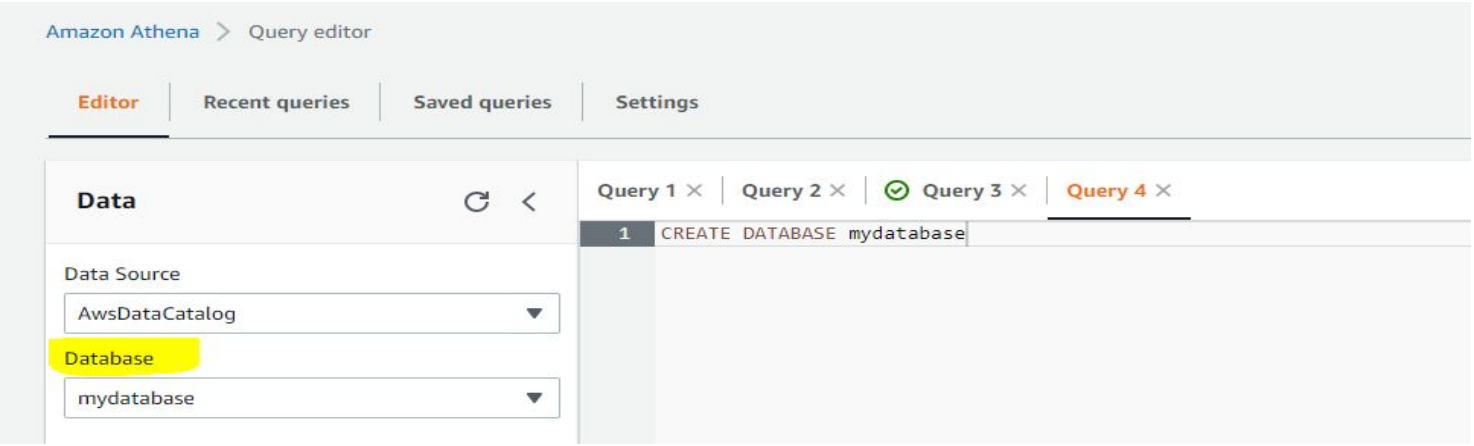
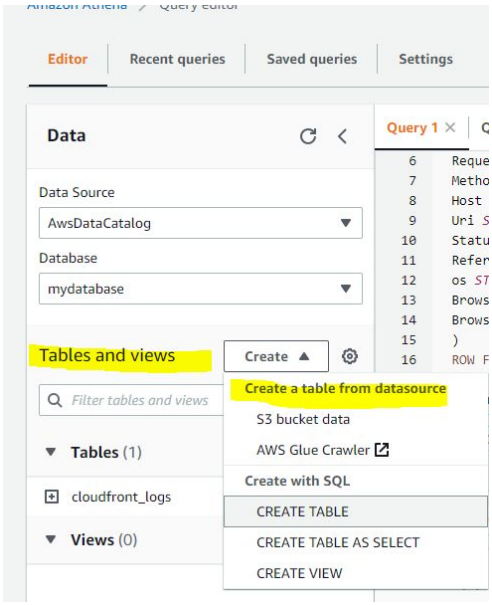


Table creation - 1

- In Athena, there are mainly 3 ways to create tables.
  - AWS Glue Crawler
  - Manually defining schema
  - SQL DDL queries

Table creation - 2

- From the list of create Table dropdown, choose S3 bucket data to define schema manually. (Or use AWS Glue to generate schema automatically)



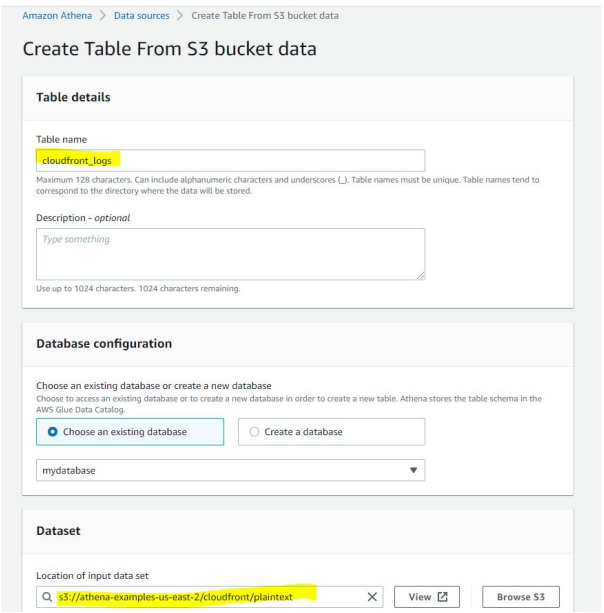


Table creation - 3

Table creation using SQL query.

Query 1 | Query 2 | Query 3

```
CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (
  `Date` DATE,
  Time STRING,
  Location STRING,
  Bytes INT,
  RequestIP STRING,
  Method STRING,
  Host STRING,
  Uri STRING,
  Status INT,
  Referrer STRING,
  os STRING,
  Browser STRING,
  BrowserVersion STRING
  )
  ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
  WITH SERDEPROPERTIES (
  "input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^\(]+[\(]([^\;]+).*\%20([^\/]+)[\/]
      (.*)$"
  ) LOCATION 's3://athena-examples-us-east-2/cloudfront/plaintext/';
```


Running Queries and Checking results

After table creation, run SQL queries.

Data

Data Source

AwsDataCatalog

Database

Query 1 | Query 2 | Query 3

```
SELECT os, COUNT(*) count
FROM cloudfront_logs
WHERE date BETWEEN date '2014-07-05' AND date '2014-08-05'
GROUP BY os
```

Recent queries (1/4)

Search recent queries

Cancel | Download results

Execution ID	Query	Start time	Status	Run time	Data scanned	Query engine version use
6b37446b-bfa6-4bf0-8178-148acdcccf20	SELECT os, COUNT(*) count FROM cloudfront_logs WHERE dat...	2021-11-28T18:08:28.809-08:...	Completed	3.465 sec	992.88 KB	Athena engine version 2
037d5709-dd63-4d74-aee2-ccf754bb5703	SELECT os, COUNT(*) count FROM cloudfront_logs WHERE dat...	2021-11-27T19:30:27.574-08:...	Completed	3.56 sec	992.88 KB	Athena engine version 2
04dd9db2-05bf-4d9b-b59e-cf19578cdfeb	CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (`D...	2021-11-27T19:29:57.324-08:...	Completed	0.346 sec	0 MB	Athena engine version 2
9472df5d-2c16-445e-8d06-476113797c5c	CREATE DATABASE mydatabase	2021-11-27T18:25:46.425-08:...	Completed	0.315 sec	0 MB	Athena engine version 2

Thank you

Q & A