

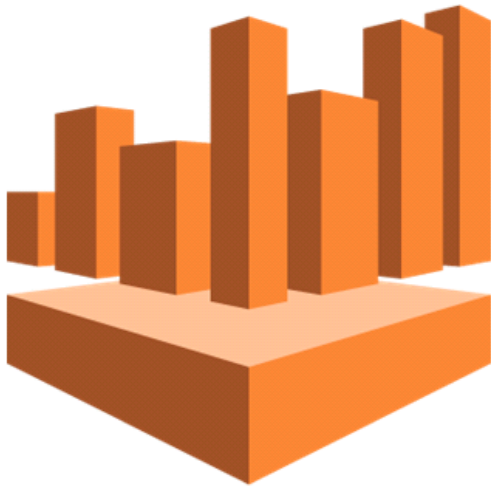
Understanding **AWS Athena**



Serverless Interactive Query service on AWS S3

BRUNA BARRAQUER
NAYARA COSTA
QIUCHI CHEN
ZHENGYONG JI

What is Amazon Athena?



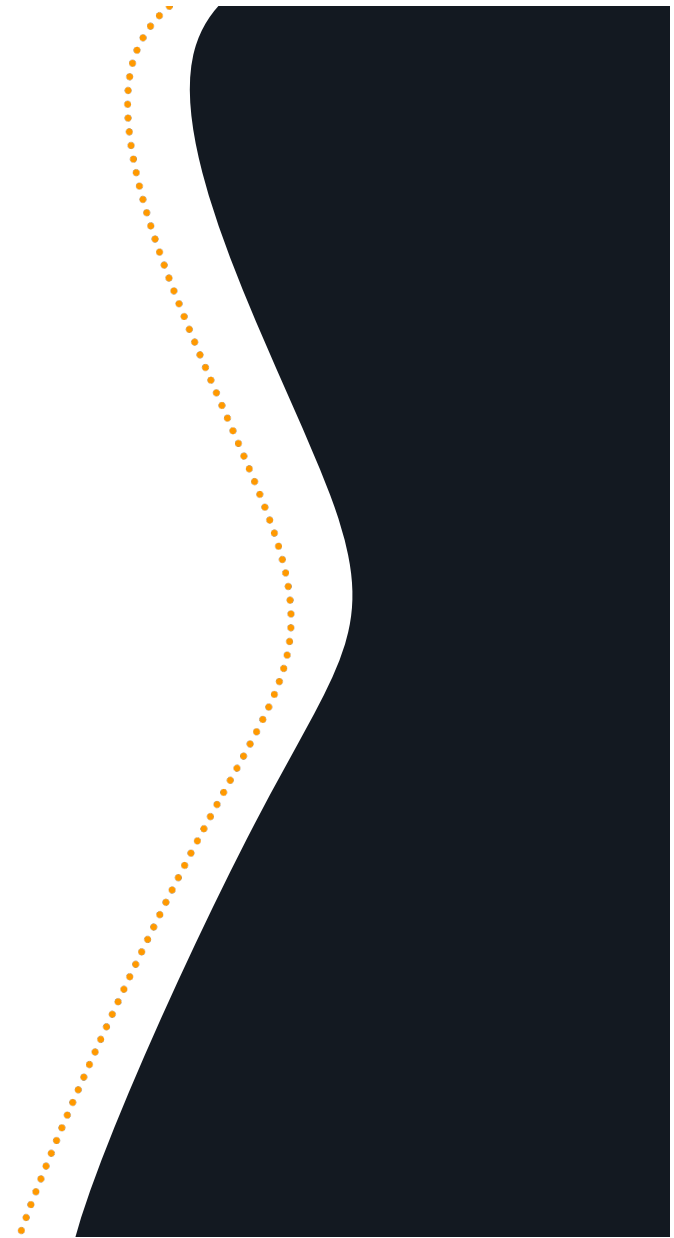
Amazon Athena

- Athena is a serverless interactive query service.
- Query data in Amazon S3 using standard SQL.
- No infrastructure to manage; pay per query.
- Supports CSV, JSON, Parquet, ORC formats.



When to Use Athena

- Need quick insights from S3 data
- Ad hoc queries, log analysis, data exploration
- No ETL needed — query raw data directly



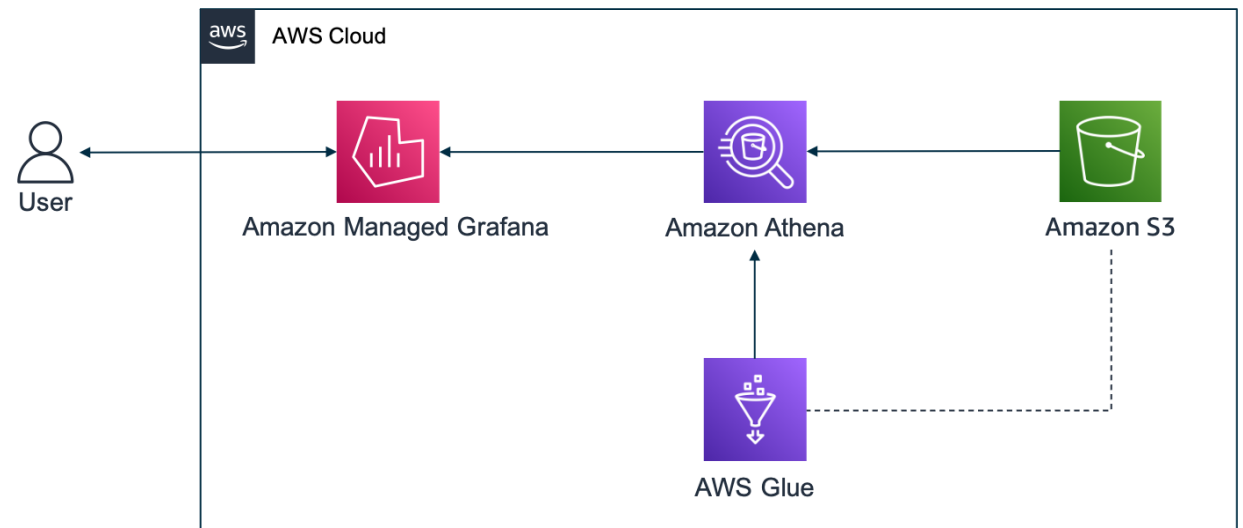
Athena vs. Other Tools

- **Athena:** Serverless, good for quick S3 queries
- **EMR:** Custom big data processing (Spark, Hadoop)
- **Redshift:** Structured, long-term analytical storage



Architecture Overview

- Raw data in S3
- Glue Data Catalog holds metadata (schema)
- Athena queries this using SQL
- Query results saved to S3 output bucket



Getting Started

1. Create S3 bucket for results
2. Open Athena Console
3. Set query result location
4. Use Query Editor for SQL queries

Create S3 Bucket

```
aws s3 mb s3://your-  
bucket-name --region your-  
region
```

Step 1

Create a Database and Table

```
CREATE EXTERNAL TABLE  
cloudfront_logs (...)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY '\t'  
LOCATION 's3://...';
```

Step 2

Run a Query

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

Step 3

Create a new bucket to store data

[Amazon S3](#) > Buckets

► **Account snapshot - updated every 24 hours** All AWS Regions

Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets. [Learn more](#)

[View Storage Lens dashboard](#)

[General purpose buckets](#) | Directory buckets

General purpose buckets (1) Info All AWS Regions

Buckets are containers for data stored in S3.

[Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

Name	AWS Region	IAM Access Analyzer	Creation date
my-bucket-20250508	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 8, 2025, 20:38:54 (UTC+02:00)

Create a new Table to store query results

[Amazon Athena](#) > [Query editor](#) > Manage settings

①

Manage settings

Query result location and encryption

Location of query result - optional

Enter an S3 prefix in the current region where the query result will be saved as an object.

X

[View](#)

[Browse S3](#)

You can create and manage lifecycle rules for this bucket

Use Amazon S3 lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time. [Learn more](#)

[Lifecycle configuration](#)

Expected bucket owner - optional

Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

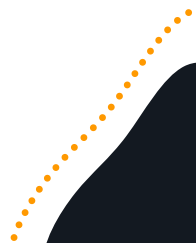
☐ Assign bucket owner full control over query results

Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

☐ Encrypt query results

[Cancel](#)

[Save](#)



Query for results directly in Query Editor

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup primary

Data

Data source: AwsDataCatalog

Catalog: None

Database: default

Tables and views: Create

Filter tables and views

Tables (1)

- customers
 - idx: int
 - customer_id: string
 - first_name: string
 - last_name: string
 - company: string
 - city: string
 - country: string
 - phone1: string

Query 2

```
1 SELECT * FROM customers LIMIT 10;
```

SQL Ln 1, Col 34

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

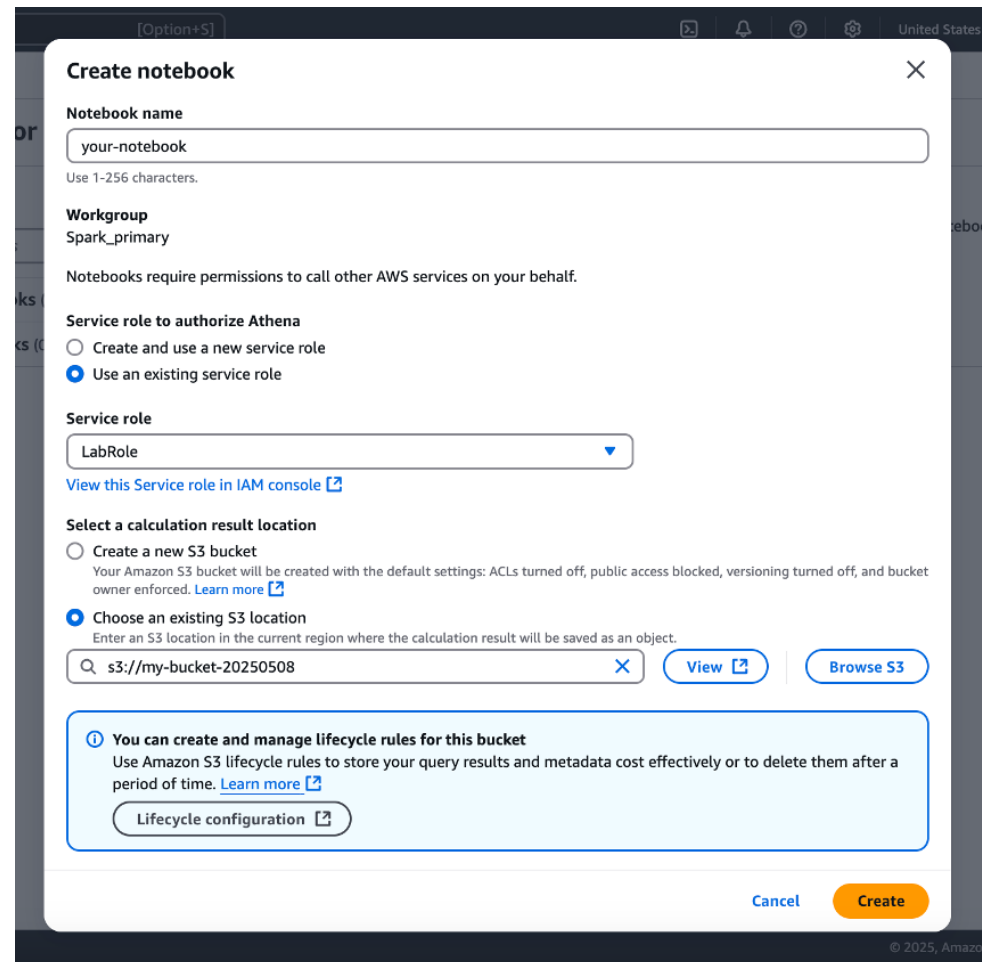
Completed Time in queue: 71 ms Run time: 644 ms Data scanned: 820.82 KB

Results (10) Copy Download results CSV

Search rows

#	idx	customer_id	first_name	last_name	company	city	country	phone1	p
1	1	E854EF1154C3A78	Heather	Callahan	Mosley-David	Lake Jeffborough	Norway	043-797-5229	9
2	2	10dAcafEBbA5FcA	Kristina	Ferrell	Horn, Shepard and Watson	Aaronville	Andorra	932-062-1802	6
3	3	67DAB15Ebe48E4a	Briana	Andersen	Irwin-Oneal	East Jordan	Nepal	8352752061	0
4	4	6d350C5E5eD84EE	Patty	Ponce	Richardson Group	East Kristintown	Northern Mariana Islands	302.398.3833	1
5	5	5820deAdCF23FEe	Kathleen	Mccormick	Carson-Burch	Andresmouth	Macao	001-184-153-9683x1497	5
6	6	E1CDEaC63fDd5aA	Trevor	Lee	Maddox Group	Lake Madelineburgh	Senegal	+1-134-348-0265x9132	+

Query using PySpark, by creating a notebook



The screenshot shows the 'Create notebook' dialog in the AWS IAM console. The dialog is titled 'Create notebook' and has a close button (X) in the top right corner. It contains the following sections:

- Notebook name:** A text input field containing 'your-notebook'. Below it, a note says 'Use 1-256 characters.'
- Workgroup:** A dropdown menu showing 'Spark_primary'.
- Permissions:** A note stating 'Notebooks require permissions to call other AWS services on your behalf.'
- Service role to authorize Athena:** Two radio buttons: 'Create and use a new service role' (unselected) and 'Use an existing service role' (selected).
- Service role:** A dropdown menu showing 'LabRole'. Below it, a link says 'View this Service role in IAM console'.
- Select a calculation result location:** Two radio buttons: 'Create a new S3 bucket' (unselected) and 'Choose an existing S3 location' (selected). Below the selected option, a note says 'Enter an S3 location in the current region where the calculation result will be saved as an object.'
- S3 location:** A text input field containing 's3://my-bucket-20250508'. To the right of the field are 'View' and 'Browse S3' buttons.
- Information box:** A light blue box with an information icon (i) and the text: 'You can create and manage lifecycle rules for this bucket. Use Amazon S3 lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time. Learn more'. Below the text is a 'Lifecycle configuration' button.
- Buttons:** 'Cancel' and 'Create' buttons at the bottom right.

The background of the screenshot shows the AWS IAM console interface with a dark theme.



PySpark Results



Amazon Athena

Query editor

Notebook editor

Notebook explorer

Jobs

Workflows

Powered by Step Functions

Administration

Workgroups

Capacity reservations

Data sources and catalogs

What's new 9+

Turn on compact mode

Notebook editor

your-notebook Customer Analysis

Workgroup Spark_primary

Give feedback (+)

(autosaved) Session

In [13]:
Run a SQL query on the Glue Catalog table
df = spark.sql("""
 SELECT country, COUNT(*) AS count
 FROM customers
 GROUP BY country
 ORDER BY count DESC
 """)

Show the result
df.show()

Calculation started (calculation_id=a2cb58f0-985a-45e9-f773-e0cd82b9213b) in (session=e6cb58ea-7245-1466-37bb-e05651b48e46). Checking calculation status...
Progress: 0% | elapsed time = 00:00s
Calculation completed.
+-----+-----+
| country | count |
+-----+-----+
Korea	168
Congo	162
Vanuatu	116
Sierra Leone	112
Andorra	109
Montenegro	108
United States Vir...	108
Cote d'Ivoire	108
Mauritius	108
Taiwan	106
Fiji	106
Botswana	106
Vietnam	106
Maldives	104
Norway	103
Iceland	102
Mali	102
Gambia	102
Sri Lanka	102
Nepal	101
+-----+-----+
only showing top 20 rows

Integrations

- AWS Glue (metadata management)
- Amazon QuickSight (visualization)
- CloudTrail, VPC Logs, Step Functions

Pros and Cons

Pros:

- Fast setup, SQL-based, scalable

Limitations:

- Only queries data in S3
- Costs based on data scanned





thanks!

