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

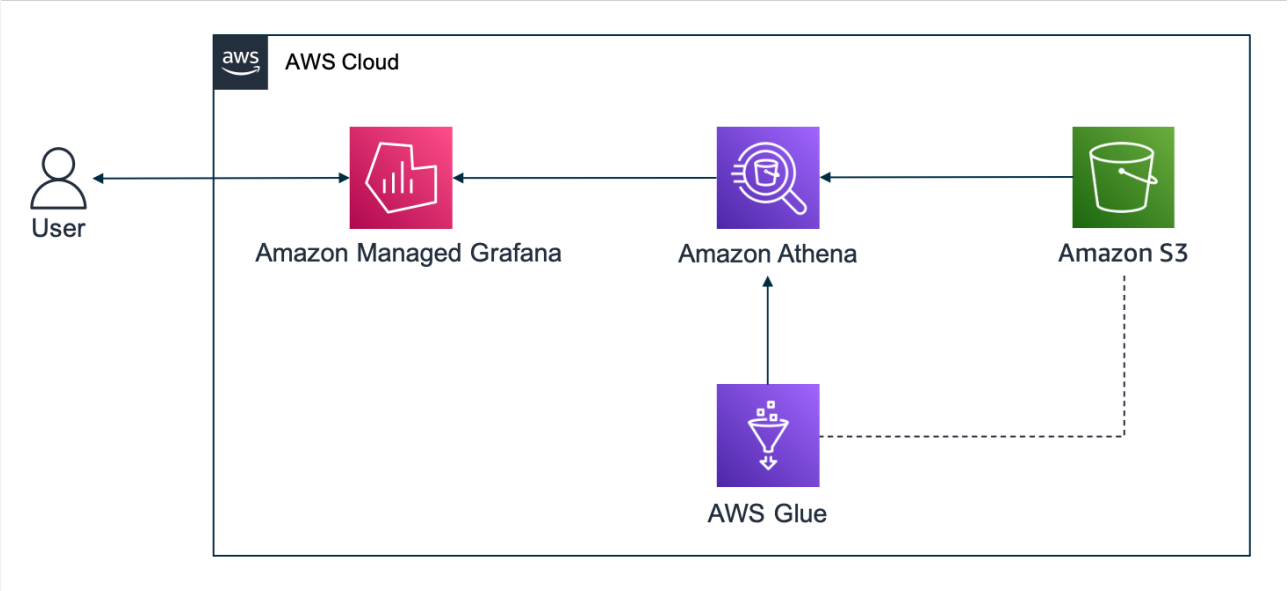
- Need quick insights from S3 data
- Ad hoc queries, log analysis, data exploration
- No ETL needed — query raw data directly

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://...';
```

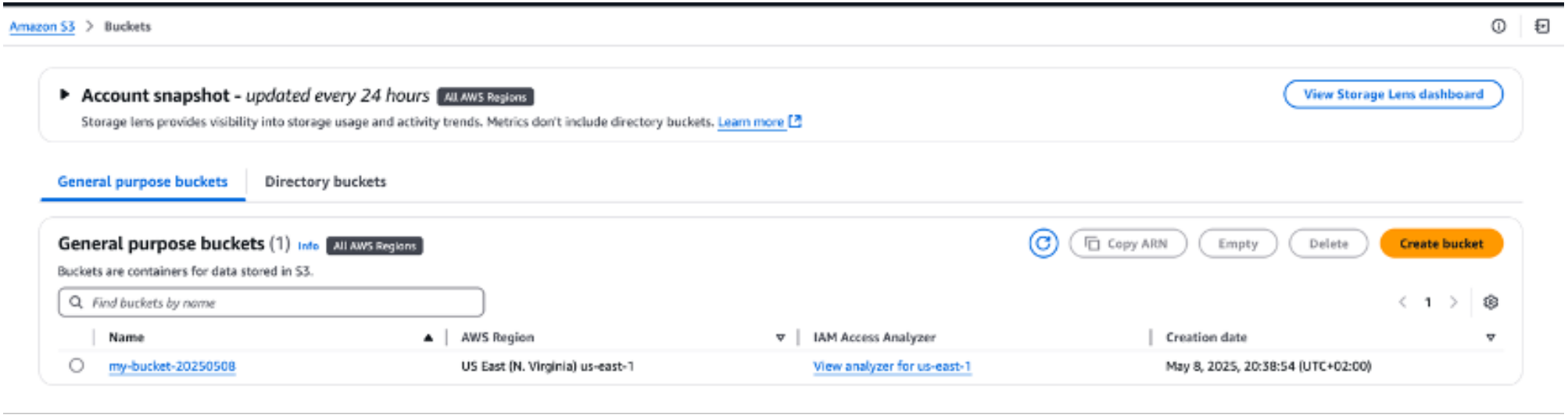
Step 2

Run a Query

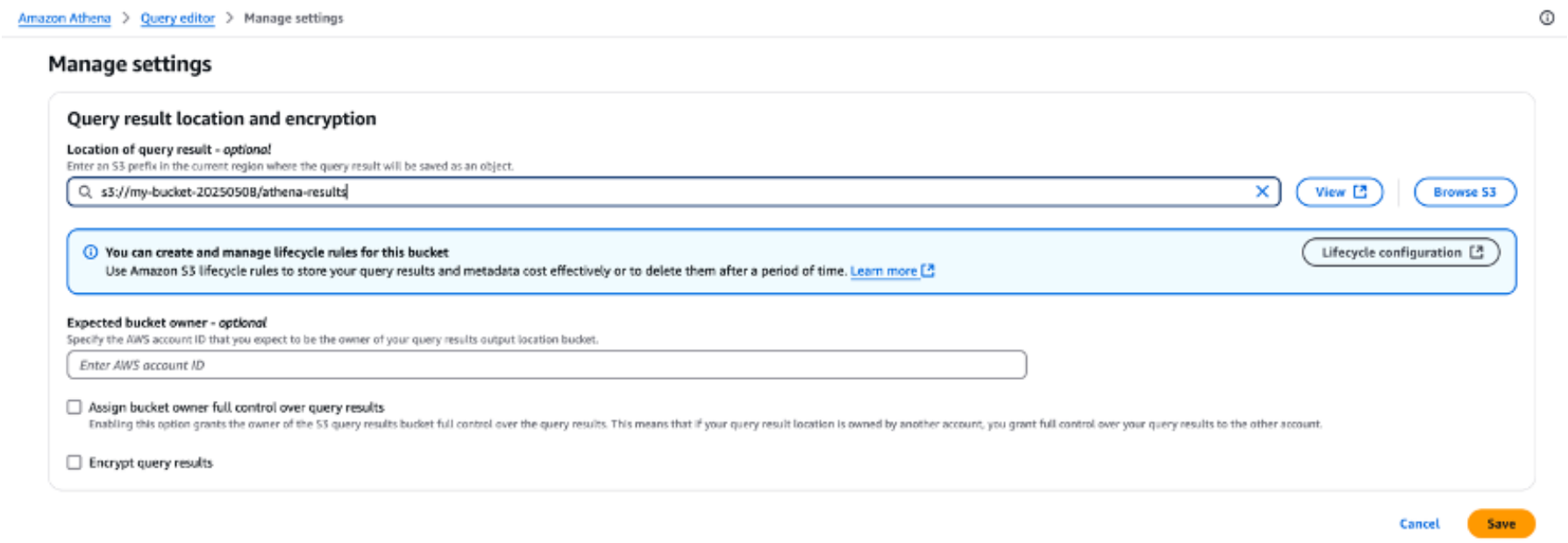
```
SELECT os, COUNT(*) FROM cloudfront_logs
WHERE date BETWEEN '2014-07-05' AND '2014-08-05'
GROUP BY os;
```

Step 3

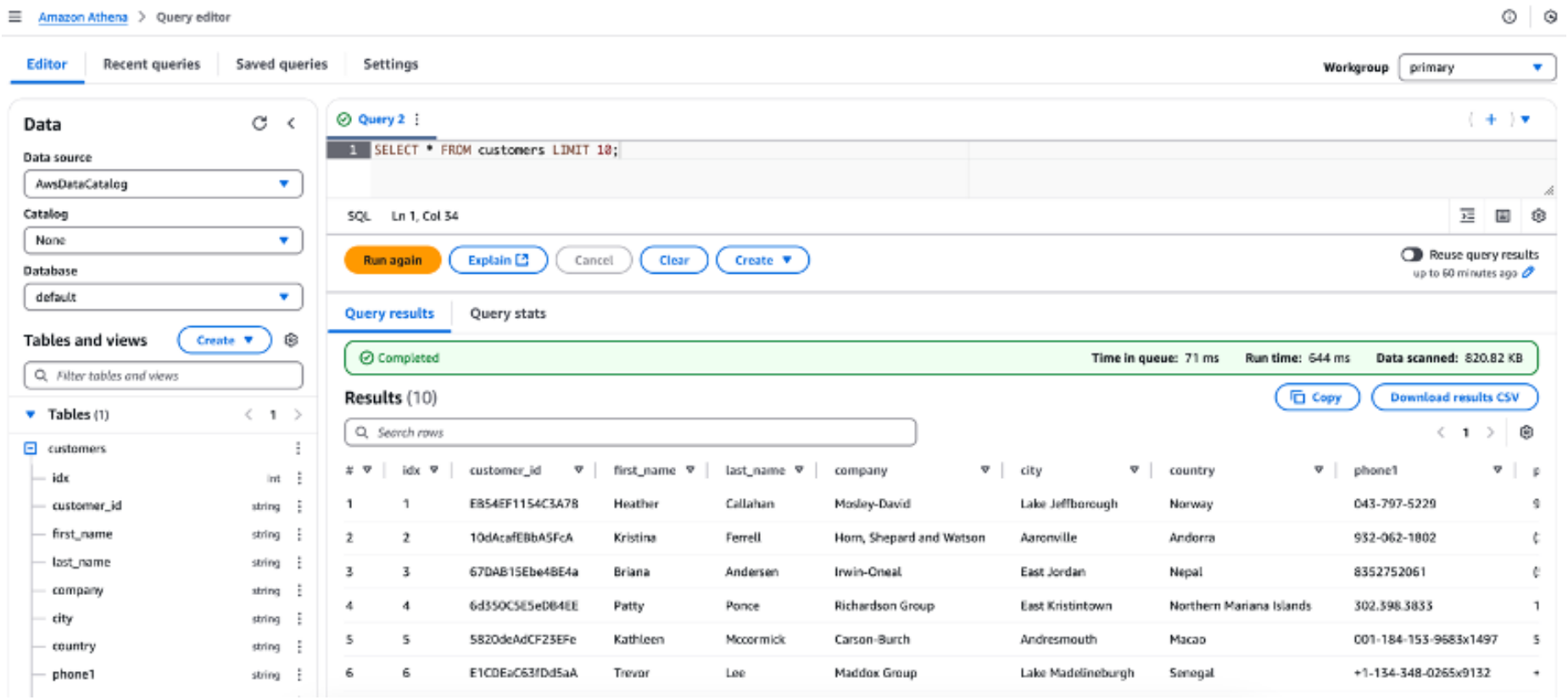
Create a new bucket to store data



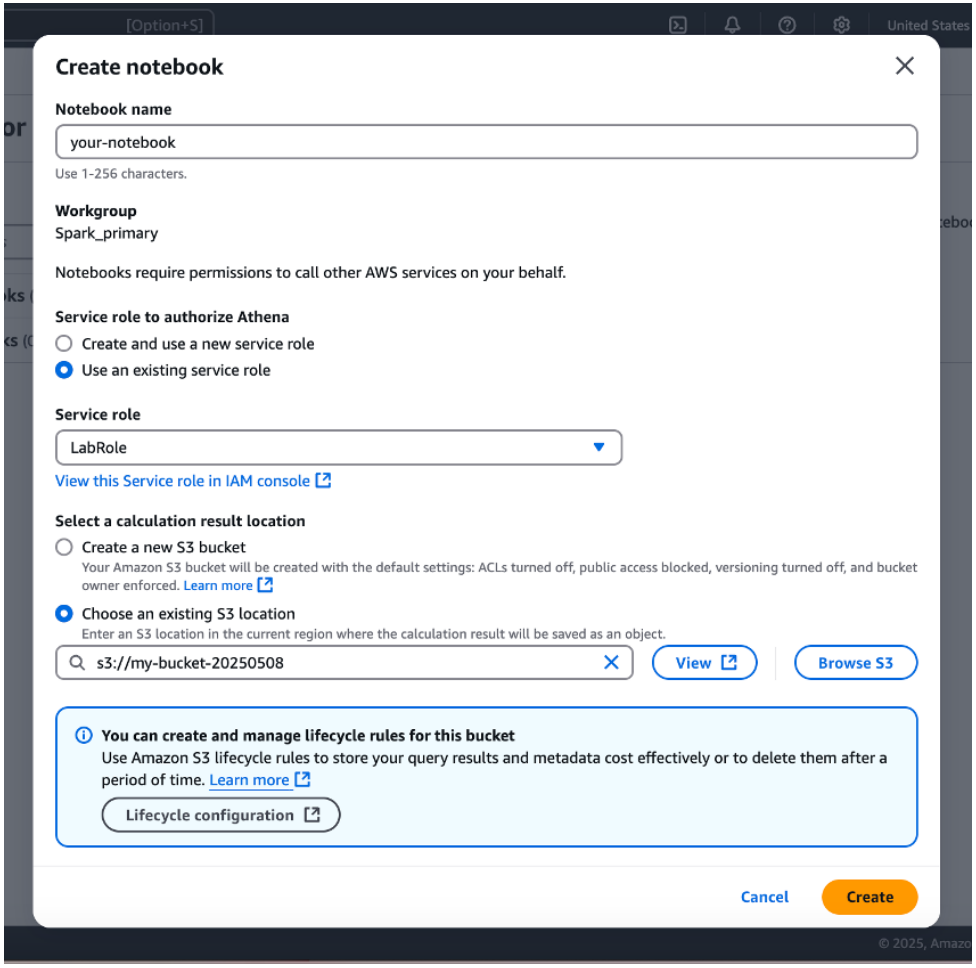
Create a new Table to store query results



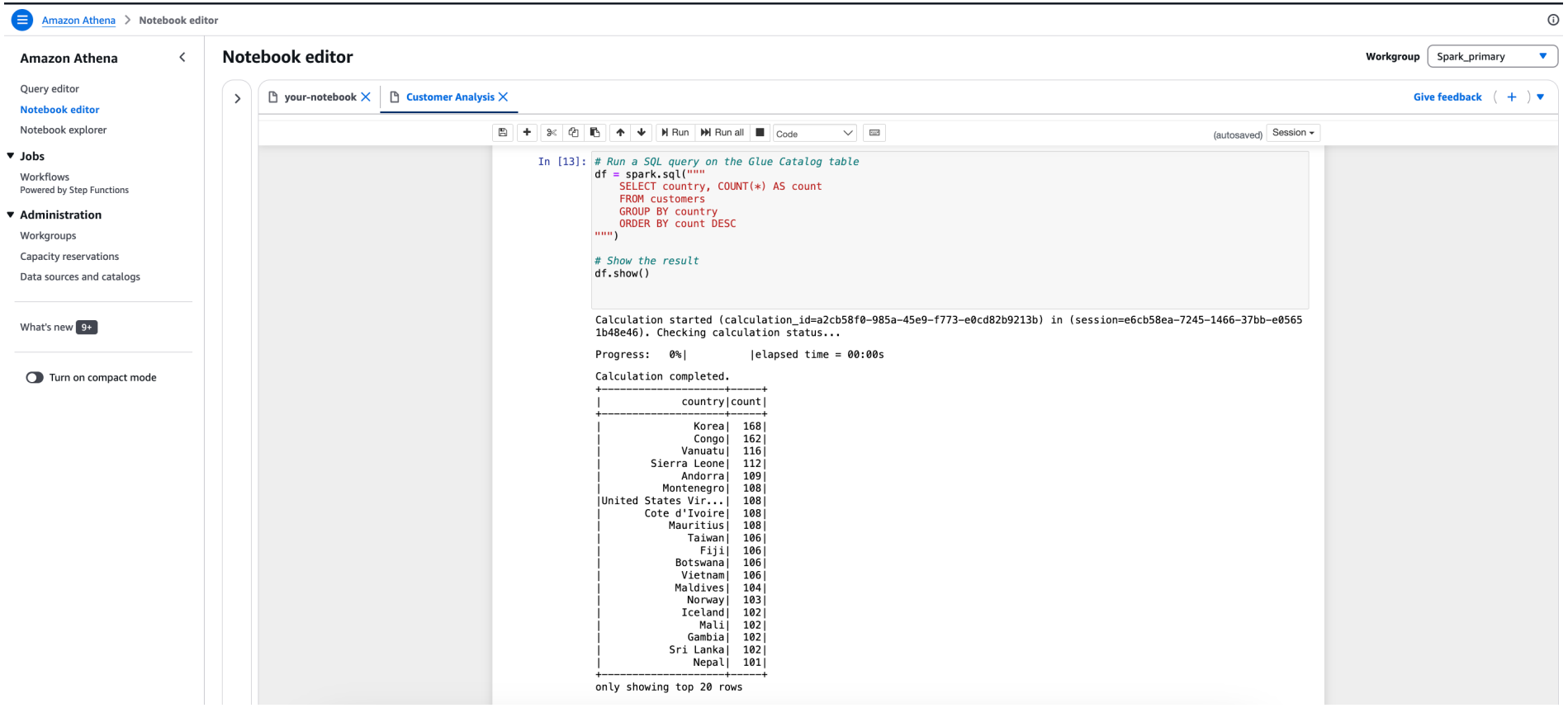
Query for results directly in Query Editor



Query using PySpark, by creating a notebook



PySpark Results



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!