Project Documentation

Project Title:

Querying AWS Glue Catalog Tables using Apache Hive on EMR

Real-World Scenario:

An analytics team at a **retail corporation** is looking to process vast amounts of **transactional data stored in Amazon S3** using complex analytical queries. Their goal is to utilize **Amazon EMR with Apache Hive** for batch processing while maintaining a centralized and queryable metadata catalog.

Best Architecture Decision:

Configure the EMR cluster to use the AWS Glue Data Catalog as a drop-in replacement for the Hive Metastore.

Objective

- Launch an EMR cluster with Hive configured to use AWS Glue Data Catalog as the Hive Metastore
- Connect to the master node via SSH
- Run SQL queries on a Glue-registered table using Hive CLI on EMR

Tools & Services Used

Service	Purpose
AWS EMR	Hadoop framework with Hive and Spark for distributed compute
AWS Glue	Centralized metadata store used as Hive Metastore
IAM	Secure permission handling for Glue and S3 access
EC2	Hosts the EMR master node and Hive CLI environment
PuTTY / SSH	Used to connect securely to the EMR master node

⚠ Issues Encountered & How They Were Resolved

1 Insufficient IAM Permissions

- **Error:** `Not authorized to perform: glue:GetDatabase`
- **Cause:** EMR EC2 instance profile lacked the correct Glue permissions.
- **Solution:** Attached `AWSGlueConsoleFullAccess` policy to EMR's EC2 role (AmazonEMR-InstanceProfile-).

2 SSH Access Timeout

- **Error:** `ssh: connect to host ... port 22: Connection timed out`
- **Cause:** EC2 Security Group didn't allow port 22 access.
- **Solution:** Added an inbound rule to allow SSH access from the current IP.

3 Misuse of Shell Context

- **Error:** `-bash: SHOW: command not found`
- **Cause:** SQL was mistakenly executed in the Linux shell.
- **Solution:** Accessed Hive CLI first using the command:

```
hive
```

☑ Final Success

Once the setup was corrected:

📜 Commands Run via Hive CLI:

```
SHOW DATABASES;
SHOW TABLES IN nyc_taxi_db;
SELECT * FROM nyc_taxi_db.nyc_taxi_dataset_bucket LIMIT 10;
```

☑ Output:

- Glue table was successfully queried via Hive on EMR
- Queries returned expected results

🧠 Key Learnings

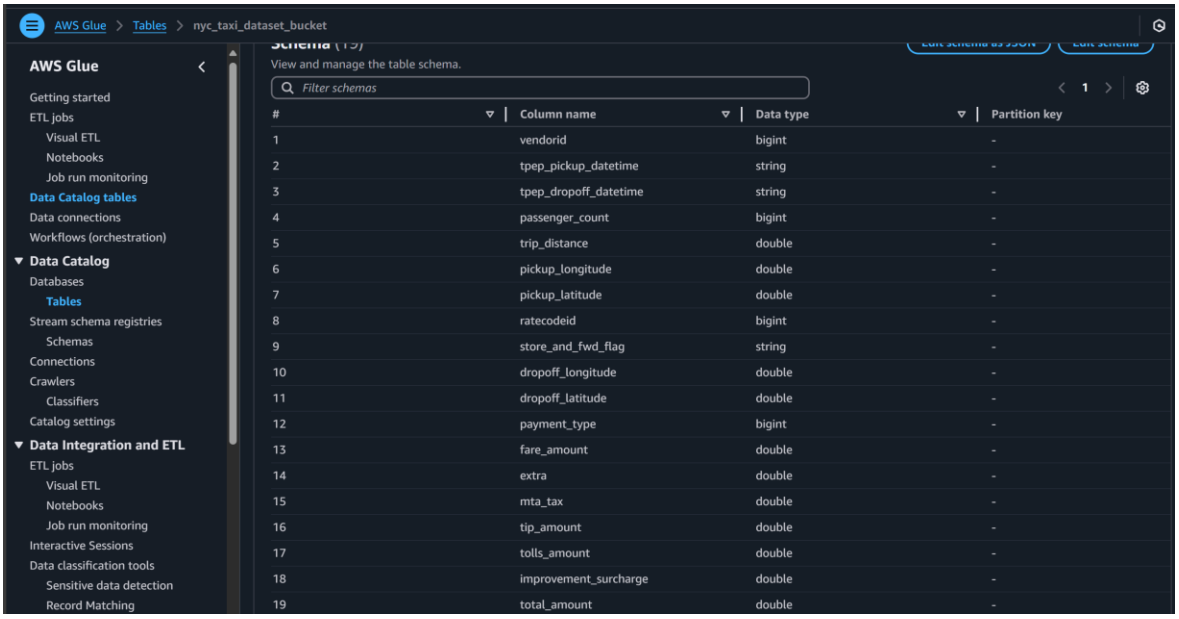
- AWS Glue can **seamlessly act as a Hive Metastore** for EMR without deploying RDS or DynamoDB.
- Permissions for Glue and S3 access are crucial — **configure IAM roles properly**.
- Always ensure Hive CLI is launched before running SQL commands.
- EMR clusters in **WAITING** state are ready for use.

- Security groups should be updated for each SSH session — **especially if IP address changes**.

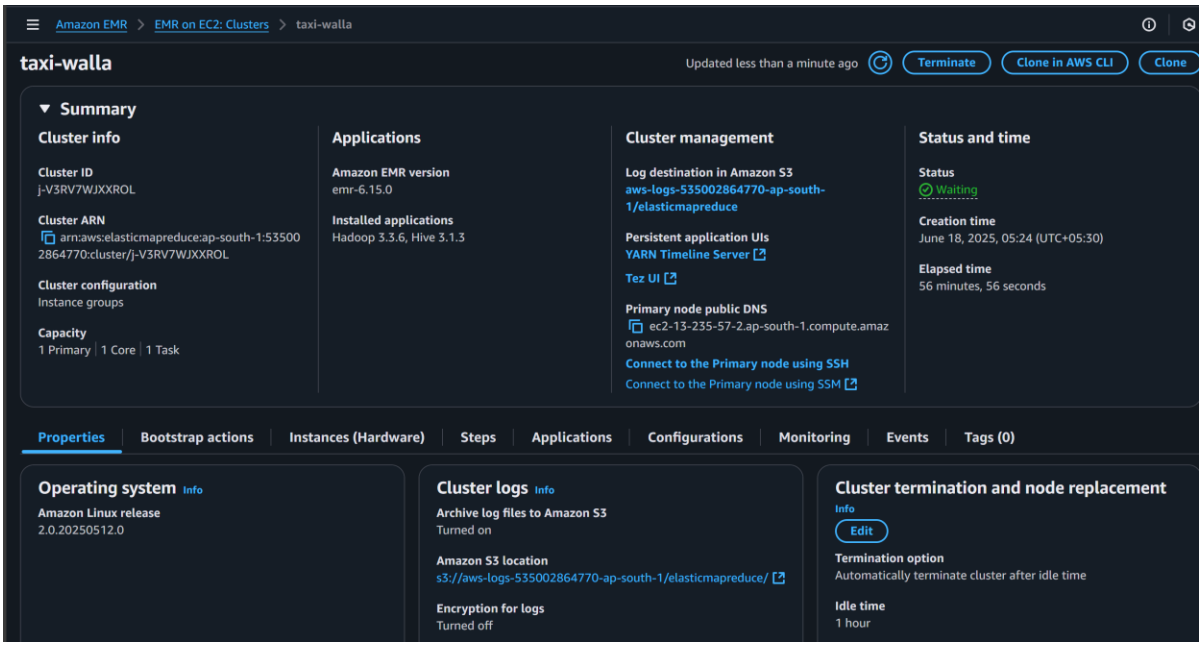
Workspace Output (Paste Your Results Below)

📝 Paste screenshots, query outputs, Hive logs, and S3 listings here:

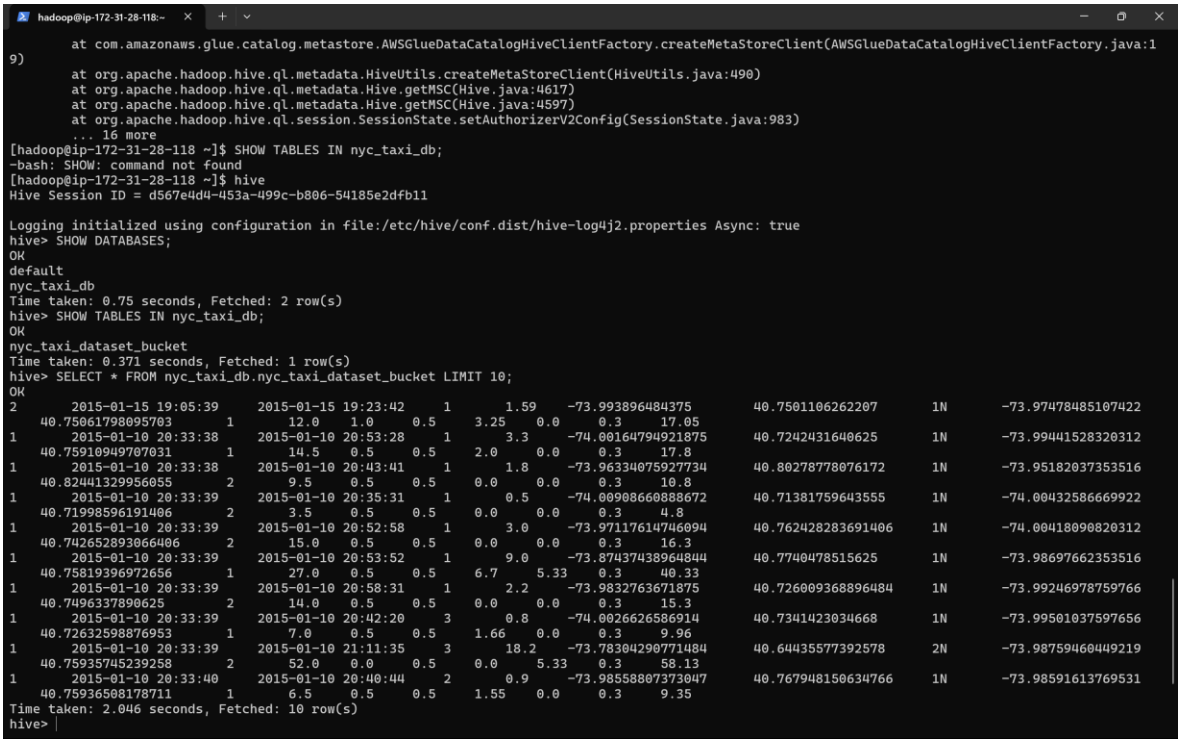
[Screenshot: Glue table definition]



[Screenshot: EMR cluster configuration]



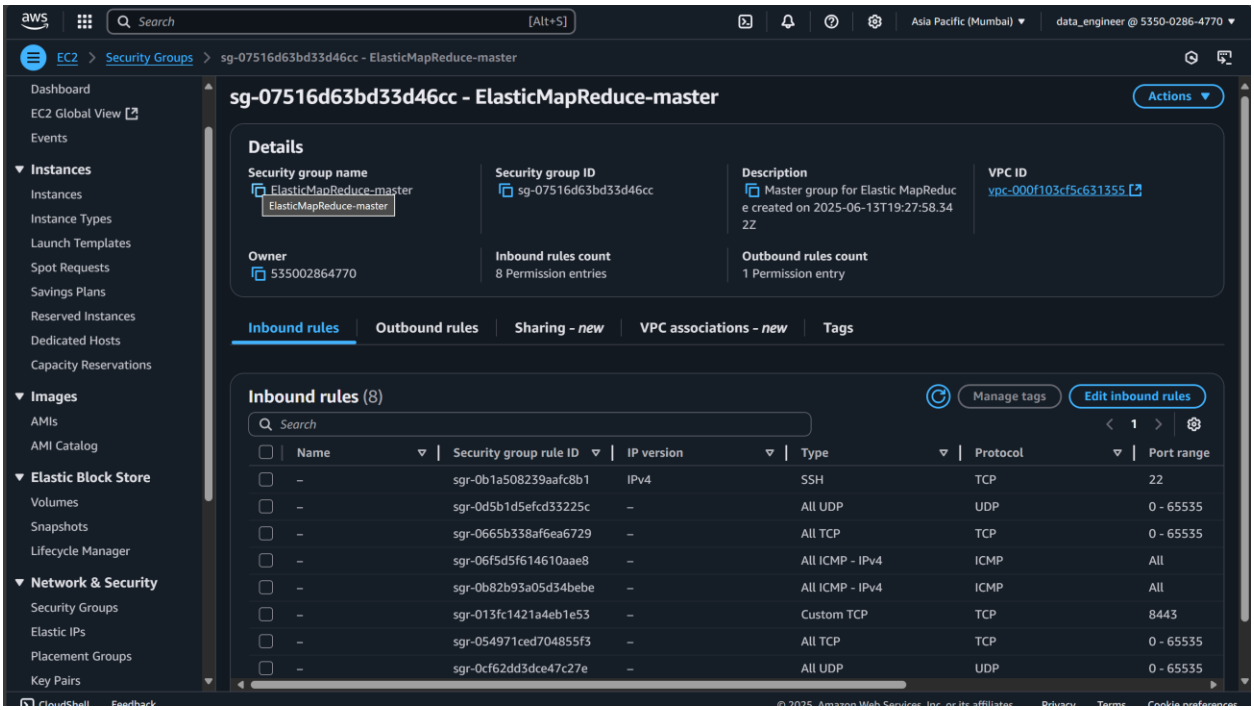
[Hive query result logs]



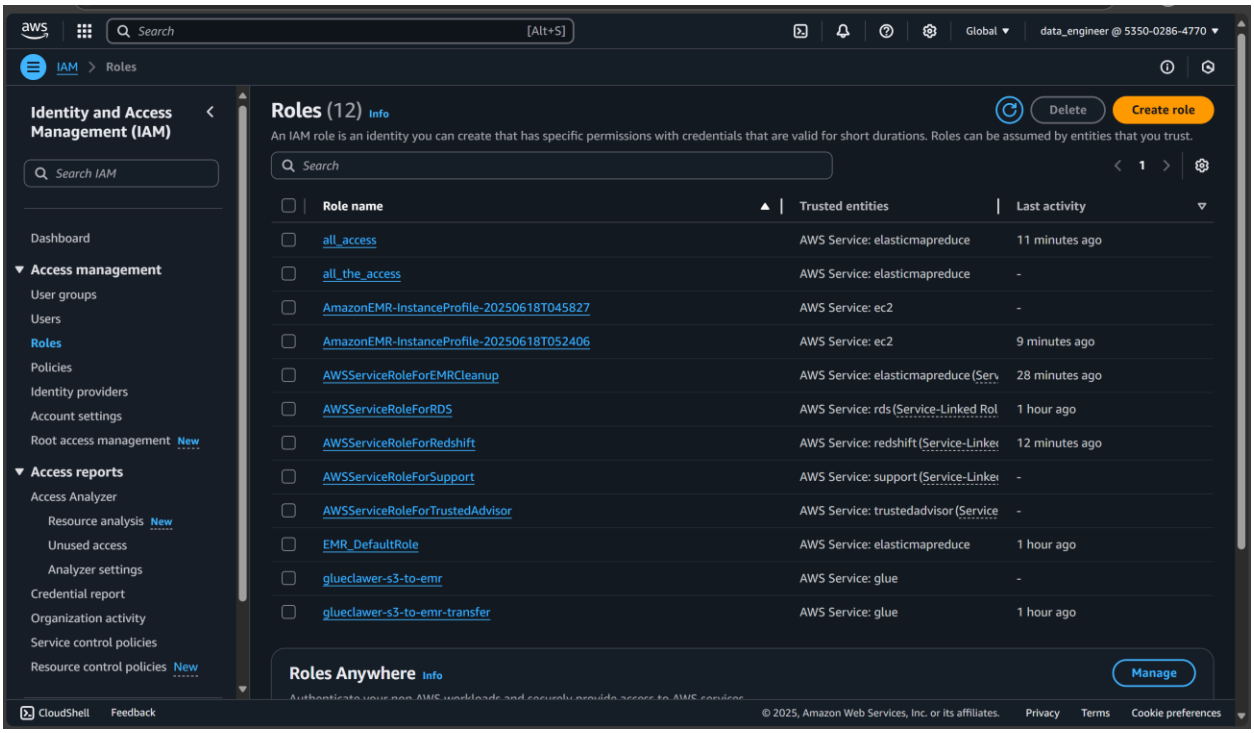
```
hadoop@ip-172-31-28-118:~
        at com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory.createMetaStoreClient(AWSGlueDataCatalogHiveClientFactory.java:1
9)
        at org.apache.hadoop.hive.ql.metadata.HiveUtils.createMetaStoreClient(HiveUtils.java:490)
        at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:4617)
        at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:4597)
        at org.apache.hadoop.hive.ql.session.SessionState.setAuthorizerV2Config(SessionState.java:983)
        ... 16 more
[hadoop@ip-172-31-28-118 ~]$ SHOW TABLES IN nyc_taxi_db;
-bash: SHOW: command not found
[hadoop@ip-172-31-28-118 ~]$ hive
Hive Session ID = d567e4d4-453a-499c-b806-54185e2dfb11

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> SHOW DATABASES;
OK
default
nyc_taxi_db
Time taken: 0.75 seconds, Fetched: 2 row(s)
hive> SHOW TABLES IN nyc_taxi_db;
OK
nyc_taxi_dataset_bucket
Time taken: 0.371 seconds, Fetched: 1 row(s)
hive> SELECT * FROM nyc_taxi_db.nyc_taxi_dataset_bucket LIMIT 10;
OK
2       2015-01-15 19:05:39     2015-01-15 19:23:42     1       1.59    -73.9938964843375       40.7501106262207        1N      -73.97478485107422
    40.75061798095703       1       12.0    1.0     0.5     3.25    0.0     0.3     17.05
1       2015-01-10 20:33:38     2015-01-10 20:53:28     1       3.3     -74.00164794921875      40.7242431640625        1N      -73.99441528320312
    40.75910949707031       1       14.5    0.5     0.5     2.0     0.0     0.3     17.8
1       2015-01-10 20:33:38     2015-01-10 20:43:41     1       1.8     -73.96334075927734      40.80278778076172       1N      -73.95182037353516
    40.82441329956055       2       9.5     0.5     0.5     0.0     0.0     0.3     10.8
1       2015-01-10 20:33:39     2015-01-10 20:35:31     1       0.5     -74.00908660888672      40.71381759643555       1N      -74.00432586669922
    40.71998596191406       2       3.5     0.5     0.5     0.0     0.0     0.3     4.8
1       2015-01-10 20:33:39     2015-01-10 20:52:58     1       3.0     -73.97117614746094      40.762428283691406      1N      -74.00418090820312
    40.742652893066406      2       15.0    0.5     0.5     0.0     0.0     0.3     16.3
1       2015-01-10 20:33:39     2015-01-10 20:53:52     1       9.0     -73.87437438964844      40.7740478515625        1N      -73.98697662353516
    40.75819396972656       1       27.0    0.5     0.5     6.7     5.33    0.3     40.33
1       2015-01-10 20:33:39     2015-01-10 20:58:31     1       2.2     -73.9832763671875       40.7260093688964844     1N      -73.99246978759766
    40.7496337890625        2       14.0    0.5     0.5     0.0     0.0     0.3     15.3
1       2015-01-10 20:33:39     2015-01-10 20:42:20     3       0.8     -74.0026626586914       40.7341423034668        1N      -73.99501037597656
    40.72632598876953       1       7.0     0.5     0.5     1.66    0.0     0.3     9.96
1       2015-01-10 20:33:39     2015-01-10 21:11:35     3       18.2    -73.78304290771484      40.644355773925781      2N      -73.98759460449219
    40.75935745239258       2       52.0    0.0     0.5     0.0     5.33    0.3     58.13
1       2015-01-10 20:33:40     2015-01-10 20:40:44     2       0.9     -73.98558807373047      40.7679481506347766     1N      -73.98591613769531
    40.75936508178711       1       6.5     0.5     0.5     1.55    0.0     0.3     9.35
Time taken: 2.046 seconds, Fetched: 10 row(s)
hive>
```

[Security group setup confirmation]



[Role attachment confirmation in IAM console]



📌 Summary

You replicated an enterprise-level EMR + Hive setup with Glue integration:

- ✅ Solved multiple setup errors
- ✅ Queried real Glue metadata from Hive
- ✅ Improved metadata management and system reliability

This solution scales and can be reused across data lake architectures.

📎 Tags

#AWS #EMR #GlueMetastore #HiveCLI #BigData #RetailAnalytics