📘Project Title: Centralized Metadata Management Using Amazon EMR and AWS Glue Data Catalog

🧰Problem Scenario:

A multinational company leverages multiple AWS big data services, such as Amazon EMR for batch processing and Amazon Athena for ad hoc querying. However, the company struggles to maintain a consistent and centralized metadata repository, especially as their existing metadata is fragmented across an old Apache Hive metastore and multiple services.

This leads to:

- Data duplication
- Inconsistent schema definitions
- Complex governance and management overhead

🕖Objective:

To implement a **centralized metadata repository** using **AWS Glue Data Catalog** that integrates seamlessly with **Amazon EMR and Amazon Athena**, ensuring scalability and reducing development effort.

🔗Proposed Solution:

Use **AWS Glue Data Catalog** as the unified Hive-compatible metastore. Configure Amazon EMR to point to Glue instead of a traditional Hive metastore. This enables interoperability across:

- EMR (Hive/Spark)
- Athena
- Redshift Spectrum

🛠️Solution Implementation Steps

✂️IAM Role Setup

Service Role for EMR (``)

Attach these AWS-managed policies:

- `AmazonElasticMapReduceRole`

- `AmazonEC2FullAccess`
- `AmazonS3FullAccess`
- `AmazonSSMManagedInstanceCore` *(optional for SSM access)*

Trust Relationship:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "elasticmapreduce.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

Instance Profile Role (``)

Attach these policies:

- `AmazonElasticMapReduceforEC2Role`
- `AmazonEC2FullAccess`
- `AmazonS3FullAccess`
- `AWSGlueConsoleFullAccess`

Trust Relationship:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "ec2.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

✒️S3 Bucket Setup

Create or reuse a bucket for:

- EMR logs
- Workspace and job files
- Input/output datasets

Examples:

- `s3://company-emr-logs/`
- `s3://company-emr-data/input/`
- `s3://company-emr-data/output/`

🛏️Create EMR Cluster with Glue Integration

In the **Amazon EMR console**:

- Click **Create cluster**
- Software Configuration:
- Applications: **Hadoop, Hive, Spark, Livy**
- EMR Release: **emr-6.x or higher**
- IAM Roles:
- Service Role: `emr-iam-role`
- EC2 Instance Profile: `AmazonEMR-InstanceProfile-*`

💡Enable Glue Catalog Integration:

Under **Edit Software Settings**, add this `hive-site` configuration:

```
[
  {
    "Classification": "hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
"com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  }
]
```

🌂Launch and Validate

- Launch the cluster
- Monitor **CloudTrail logs** and **cluster events**

🔗**Check:**

- `RunInstances` permissions
- `AddInstanceProfile` permissions
- IAM trust policies correctness

Once up, connect to Master Node via SSH or EMR Studio to validate Glue table visibility:

```
SHOW DATABASES;
SHOW TABLES;
```

🌡️ (Optional) EMR Studio for Notebooks

- Navigate to **EMR Studio > Create Studio**
- Choose `emr-iam-role` as service role
- Provide S3 workspace location and name
- Launch Studio and use **Jupyter Notebook** or **PySpark** with access to Glue Data Catalog

---

🧾Outcome

✅Achievements:

- Centralized, unified schema management via AWS Glue
- Seamless compatibility across EMR, Athena, and Redshift Spectrum
- Easier governance, auditing, and reuse of metadata

Benefits:

- No need to maintain separate Hive metastores
- EMR clusters automatically get schema from Glue
- Athena can query the same datasets without duplication

📈Next Steps

- ✡️Add Glue Crawlers to auto-register new data in S3
- 📵Use AWS Lake Formation for access control
- ♀️ Automate EMR cluster provisioning using CloudFormation or Terraform

---

📌Conclusion

By integrating EMR with the AWS Glue Data Catalog, you establish a **scalable**, **consistent**, and **cost-effective** metadata layer. This foundation enables cross-team collaboration, faster time-to-insight, and stronger governance in enterprise-scale analytics.

📂Workspace Output (Paste Your Results Below)

> ♌Paste screenshots, logs, EMR configuration, and output screenshots here:

```
[Screenshot of EMR configuration]
[S3 folder structure for logs/data]
[EMR Studio or SSH validation output]
```

> 🆖**Name:** Atmakuru Siva Sandeep\ 🆘**Date:** 20-06-2025