Data Processing and Analytics with AWS EMR and Athena

Objective: Explore big data processing and serverless query services.

Task 1: Create an AWS Elastic MapReduce (EMR) Cluster
1. In the AWS Console, navigate to the EMR service.
2. Click “Create cluster” then go to “Advanced options”.
3. Select EMR release and software (e.g., Hadoop, Spark).
4. Configure hardware (e.g., instance types, number of instances).
5. Set up networking options, choose the VPC, and assign roles.
6. Launch the cluster.

Task 2: Run a Spark Job on EMR
1. SSH into the master node of your EMR cluster:

   ```bash
   ssh -i /path/to/myKeyPair.pem hadoop@<MasterNode-Public-DNS>
   ```

2. Submit a Spark job. For example, a word count job:

   ```python
   spark-submit --master yarn --deploy-mode client /path/to/wordcount.py /path/to/input.txt /path/to/output
   ```

3. Monitor the job in the EMR console and review the output.

Task 3: Set Up AWS Athena for Serverless SQL Queries
1. Go to the Athena service in the AWS Console.
2. Setup a query editor and connect to an S3 bucket.
3. Create a database and table reflecting the structure of data in S3:

   ```sql
   CREATE EXTERNAL TABLE IF NOT EXISTS mydatabase.mytable (
       column1 string,
       column2 string
   )
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   STORED AS TEXTFILE
   LOCATION 's3://mybucket/mydata/';
   ```

Task 4: Data Analysis with Athena

1. Execute SQL queries in Athena to analyze data in S3:

   ```sql
   SELECT * FROM mytable WHERE condition;
   ```

2. Explore results and learn about Athena’s capabilities in handling large datasets.