**Lab Exercise- AWS S3, AWS Glue and AWS Athena**

**Objective**

* Upload a CSV file to Amazon S3.
* Create a metadata catalog using AWS Glue.
* Query the data interactively using Amazon Athena.

**Pre-requisites**

* An AWS account with permissions for S3, Glue, and Athena.
* AWS CLI or AWS Management Console access.
* A sample CSV file.

**Step 1 — Prepare a Sample CSV File**

Create a file called ***employees.csv*** with the following content:

***employees.csv***

employee\_id,first\_name,last\_name,department,salary

1,John,Doe,Engineering,75000

2,Jane,Smith,Finance,82000

3,Bob,Johnson,Engineering,65000

4,Alice,Williams,Marketing,72000

5,Tom,Brown,Finance,90000

**Step 2 — Create an S3 Bucket and Upload CSV**

**a) Create an S3 bucket (use a unique name):**

aws s3 mb s3://my-glue-athena-lab-bucket

**b) Upload the CSV file:**

aws s3 cp employees.csv s3://my-glue-athena-lab-bucket/

**Step 3 — Create a Glue Database**

1. Go to the **AWS Glue Console**.
2. In the left pane, choose **Databases** → **Add database**.
3. Enter name: company\_db.
4. Save.

**Step 4 — Create a Glue Crawler**

A crawler scans your S3 data and creates table metadata in Glue Data Catalog.

1. In Glue Console → **Crawlers** → **Add crawler**.
2. Name: employee-crawler.
3. Data source: **S3**, point to your bucket path:
4. s3://my-glue-athena-lab-bucket/
5. IAM Role: Create a new role (Glue needs permissions for S3).
6. Choose database: company\_db.
7. Run the crawler.

After it completes, you should see a new **table** (e.g., employees) in your Glue database with schema columns (employee\_id, first\_name, etc.).

**Step 5 — Query Data with Athena**

1. Open the **Athena Console**.
2. Configure query results location (only the first time):
   * Go to **Settings** → enter a query results bucket, e.g.:
   * s3://my-glue-athena-lab-bucket/query-results/
3. Select the database company\_db.
4. Run SQL queries against the employees table.

**Sample Queries**

**Query 1 — See all data:**

SELECT \* FROM employees;

**Query 2 — Average salary by department:**

SELECT department, AVG(salary) AS avg\_salary

FROM employees

GROUP BY department;

**Query 3 — Highest salary:**

SELECT first\_name, last\_name, salary

FROM employees

ORDER BY salary DESC

LIMIT 1;

**Step 6 — Verify Output**

* Athena will scan the S3 data using the Glue schema.
* You can view results directly in the Athena console or download them as CSV.

**Summary of What You Learned**

* Store raw data in **S3**.
* Use **Glue Crawlers** to automatically create table schema in Glue Data Catalog.
* Use **Athena** to query structured data using SQL — serverless, no infrastructure required.