Assignment 7: Data Analytics with AWS S3, AWS Glue, and Amazon Athena

Overview

This assignment provides hands-on experience with the core AWS serverless data analytics stack. You will learn to build a modern data lake pipeline using a "bronze-silver-gold" architecture.

You will use **Amazon S3** for scalable data storage, **AWS Glue** to automatically catalog your raw data, and **Amazon Athena** to run interactive SQL queries for transformation and analysis. This process allows you to perform complex data analytics and ETL (Extract, Transform, Load) without provisioning or managing any servers.

Finally, you will analyze real-world weather data, answer business-level questions, and reflect on the significant cost and performance benefits of using serverless tools, partitioning, and columnar formats like Parquet.

Deliverables

Each student (or group) must submit the following:

1. **Presentation Slides (10–14 slides maximum)**
    - Title slide with student name(s), course, and assignment title.
    - Clear explanation of each completed task:
        - Task 1 – Select and Acquire Raw Data
        - Task 2 – Set Up S3 Data Lake Foundation
        - Task 3 – Catalog the Raw Data (Bronze Layer)
        - Task 4 – Transform Data from Bronze to Silver (CTAS)
        - Task 5 – Create Aggregated Gold-Layer Views
        - Task 6 – Answer Analytics Questions
        - Task 7 – Cost, Performance, and Cleanup
        - Task 8 – Answer Reflection Questions
    - Screenshots or videos of key steps/results.
    - Reflection question responses.

- Lessons learned and challenges encountered.

2. **Demonstration (in-class or recorded)**
    - A 8–11 minute walkthrough of the slides and discussion of your results.

Presentation Guidelines

- Use professional formatting (legible fonts, consistent colors, organized layout).
- Ensure clarity: avoid overcrowding slides with text; use bullet points and visuals.
- Provide screenshots or videos as evidence of work completed.
- Keep presentations within the time limit (8–11 minutes).
- Be prepared to answer questions about the process and challenges.

Detailed Grading Rubric (100 points total)

All team members are required to contribute to the presentation. The final score will be proportionally adjusted based on the number of members who actually present. For example, if a group consists of five students but only four deliver the presentation, and the maximum score is 100 points, the adjusted score will be calculated as:

$$Adjusted\ Score\ =\ (\frac{4}{5})\ \times 100\ =\ 80$$

1. Task Completion (60 points total)

Each AWS task is evaluated based on correctness, completeness, and evidence (e.g., screenshots/videos output).

- Task 1 & 2 – S3 Data Lake Setup **(6 pts)**
    - S3 bucket and folder structure (bronze/, silver/, gold/) created correctly. (3 pts)
    - Raw data (.csv) correctly uploaded to the bronze/ folder. (3 pts)
- Task 3 – AWS Glue Catalog **(10 pts)**
    - Glue Crawler created, including a new IAM role. (4 pts)
    - isd_db database created and targeted by the crawler. (3 pts)
    - Crawler runs successfully, and the bronze table is created and verifiable in Athena. (3 pts)
- Task 4 – Transform to Silver (CTAS) **(15 pts)**
    - CTAS query is correct (parses TMP/WND, filters reports, casts types). (8 pts)
    - Query successfully creates the isd_silver table. (3 pts)
    - Evidence that the data is correctly stored in S3 as Parquet and partitioned by year/month. (4 pts)
- Task 5 – Create Gold View **(8 pts)**
    - CREATE VIEW query is correct (uses AVG, MIN, MAX, GROUP BY). (5 pts)
    - View (isd_gold_daily) is created and returns correct-looking aggregated data. (3 pts)
- Task 6 – Analytics SQL Queries **(10 pts)**

    - Provides correct SQL queries for all 4 questions. (5 pts)
    - Provides clear answers (e.g., "Warmest day was...") based on the query results. (5 pts)
- Task 7 – Cost Analysis & Cleanup **(5 pts)**
    - Correctly explains why Parquet scanned less data than CSV. (3 pts)
    - Notes the "Data Scanned" difference and mentions the cleanup process. (2 pts)
- Task 8 – Reflection Questions **(6 pts)**
    - Complete and thoughtful answer for the required Glue/Athena question. (2 pts)
    - Complete and thoughtful answers for the 2 chosen questions. (4 pts – 2 for each)

2. Presentation Quality (30 points total)

- Clear structure (intro, results, discussion, conclusion) **(10 pts)**
- Slides visually organized and professional, for example, adding pictures, transitions, animations **(10 pts)**
- Screenshots or videos included as evidence **(5 pts)**
- Proper timing (within 6–9 min) **(5 pts)**

3. Professionalism & Delivery (10 points total)

- Confident and clear presentation delivery **(4 pts)**
- Ability to answer questions from peers/instructor **(4 pts)**
- Team collaboration and equal participation **(2 pts)**

Submission Instructions

- Submit slides (PowerPoint or PDF) to Moodle by the deadline.
- Each group will present during the next class session.

Task 1: Select and Acquire Raw Data

Learn to identify, understand, and ingest raw source data.

Instructions:

1. Familiarize yourself with the NCEI (NOAA) Global Hourly dataset. This is a common source for public weather data.
   https://www.ncei.noaa.gov/data/global-hourly/

2. Review the CSV_HELP.pdf file (often found with the data) to understand the meaning of key columns like TMP (temperature) and WND (wind).
   https://www.ncei.noaa.gov/data/global-hourly/doc/CSV_HELP.pdf

3. Choose one data file to analyze. For this assignment, we will use a pre-selected file to ensure consistency.

4. Download the following sample file (Washington Dulles, 2019) to your local machine:
   https://www.ncei.noaa.gov/data/global-hourly/access/2019/72403093738.csv

Task 2: Set Up S3 Data Lake Foundation

Establish the "Bronze-Silver-Gold" storage layers in S3. This pattern is an industry best practice for data lakes.

- **Bronze:** Raw, unmodified data, just as it was ingested.
- **Silver:** Cleaned, transformed, and query-optimized data (e.g., in Parquet format).
- **Gold:** Aggregated data, ready for business intelligence and reporting.

Instructions:

1. In the AWS Console, navigate to **Amazon S3**.
2. Create a new, globally unique S3 bucket.
   - **Recommended name:** `isd-<YOUR_GROUP>` (e.g., `isd-group11`)
3. Inside your new bucket, create three folders (which act as prefixes):
   - `bronze/`
   - `silver/`
   - `gold/`
4. Upload the `.csv` file you downloaded in Task 1 into the `bronze/` folder.

Task 3: AWS Glue Catalog

Use AWS Glue to automatically discover the schema of your raw data and make it queryable with SQL.

Instructions:

1. In the AWS Console, navigate to **AWS Glue**.
2. In the left navigation pane, select **Crawlers**.
3. Click **Create crawler**.
4. **Crawler name:** `isd-bronze-crawler`
5. **Custom classifiers → Add new classifier**

→ Add a CSV Classifier using default settings Open CSV SerDe with name `isd-bronze-classifier`.

6. **Data source:** Point the crawler to your S3 `bronze/` folder.
    - **Data source path:** `s3://isd-<YOUR_GROUP>/bronze/`
7. **IAM Role:** Allow Glue to create a new service role for you.
8. **Database:** Create a new database to store your tables.
    - **Database name:** `isd_db`
9. Finish creating the crawler, then **Run** it.
10. Wait for the crawler to complete (this may take 1-2 minutes). It will report that it has created one table.
11. Navigate to **Amazon Athena** in the AWS Console.
12. In the query editor, you should now see `isd_db` in the **Database** dropdown and a new table (likely named `bronze`) in the **Tables** list.
13. Run a test query to confirm it works:

```
SELECT * FROM "isd_db"."bronze" LIMIT 10;
```

Task 4: Transform Data from Bronze to Silver (CTAS)

Convert the raw string data into a clean, typed, and efficient format. You will use a **CREATE TABLE AS SELECT (CTAS)** query in Athena to:

- Parse complex string fields (like `TMP` and `WND`).
- Cast data to correct types (e.g., `INTEGER`, `TIMESTAMP`).
- Filter out summary rows (e.g., `SOD`, `SOM`).
- Save the results as **Parquet**, a columnar format that is much faster and cheaper to query.
- **Partition** the data by year and month for efficient filtering.

Instructions:

1. In the Athena query editor, make sure your `isd_db` is selected.
2. Run the following CTAS query. **You must modify** the `external_location` to match your S3 bucket and the `FROM` line to match your bronze table name.

```
CREATE TABLE "AwsDataCatalog"."isd_db"."silver"
WITH (
  format = 'PARQUET',
  write_compression = 'SNAPPY',
  external_location = 's3://<YOUR_S3_BUCKET>/silver/', -- <-- MODIFY THIS
  partitioned_by = ARRAY['year','month']
) AS
SELECT
  station,
  CAST(from_iso8601_timestamp(date) AS timestamp) AS ts,

  -- Temperature (°C) from TMP "+0078,1"
  CASE WHEN tmp IS NOT NULL AND substr(tmp,1,5) <> '+9999'
       THEN CAST(substr(tmp,1,5) AS integer)/10.0 END AS temp_c,

  -- Wind speed (m/s) from WND "170,1,N,0015,1"
  CASE WHEN wnd IS NOT NULL AND split(wnd, ',')[4] <> '9999'
       THEN CAST(split(wnd, ',')[4] AS integer)/10.0 END AS wind_ms,

  report_type,
  year(from_iso8601_timestamp(date)) AS year,
  month(from_iso8601_timestamp(date)) AS month

FROM "AwsDataCatalog"."isd_db"."bronze" -- <-- MODIFY THIS if your table name is different
WHERE report_type NOT LIKE 'SOD%' AND report_type NOT LIKE 'SOM%';
```

3. After the query succeeds, go to your S3 bucket's `silver/` folder. You will see new folders for `year=2019/` and `month=1/`, containing Parquet files.

Task 5: Create Aggregated Gold-Layer Views

Create a final "gold" view that aggregates the clean data into daily metrics. This view is what a dashboard or business report would use.

Instructions:

1. In the Athena query editor, run the following `CREATE VIEW` query. It builds upon your new `isd_silver` table.

```
CREATE OR REPLACE VIEW "AwsDataCatalog"."isd_db"."gold_daily" AS
SELECT
  date(ts) AS day_utc,
  count_if(temp_c IS NOT NULL) AS obs_count,
  MIN(temp_c) AS t_min_c,
  MAX(temp_c) AS t_max_c,
  AVG(temp_c) AS t_avg_c,
  AVG(wind_ms) AS wind_avg_ms
FROM "isd_db"."silver"
GROUP BY 1;
```

2. Test your new view to see the daily aggregated weather:

```
SELECT * FROM "AwsDataCatalog"."isd_db"."gold_daily" ORDER BY day_utc;
```

Task 6: Answer Analytics Questions

Use your new data pipeline to answer questions. For each question, provide your SQL query and a brief answer.

Instructions:

1. Run queries against your `silver` and `gold_daily` tables to find the answers.
2. **Question 1:** What were the 3 warmest and 3 coldest days (based on max/min temp) in the dataset?
3. **Question 2:** What is the average temperature by the hour of the day (diurnal profile)?
4. **Question 3:** What is the average wind speed by hour? When is it windiest?
5. **Question 4:** Are there any days with incomplete data (e.g., fewer than 24 hourly observations)?

Sample Queries (Adapt these):

```
-- Q1: Warmest / coldest days
SELECT * FROM "isd_db"."gold_daily" ORDER BY t_max_c DESC LIMIT 3;
SELECT * FROM "isd_db"."gold_daily" ORDER BY t_min_c ASC LIMIT 3;

-- Q2: Diurnal profile (avg temp by hour)
SELECT hour(ts) AS hour_utc, AVG(temp_c) AS avg_temp_c
FROM "isd_db"."silver"
GROUP BY 1 ORDER BY 1;

-- Q4: Data completeness
SELECT * FROM "isd_db"."gold_daily" WHERE obs_count < 24 ORDER BY day_utc;
```

Task 7: Cost, Performance, and Cleanup

Reflect on the cost-saving benefits of this architecture and clean up your resources to avoid charges.

Instructions:

1. **Reflection:** In the Athena console, look at the **Recent queries** tab. Find your queries from Task 3 (querying `bronze`) and Task 6 (querying `silver` / `gold_daily`).
    - Note the **"Data scanned"** (e.g., 10 MB) for the query on the raw `bronze` (CSV) table.
    - Note the **"Data scanned"** (e.g., 1.5 MB) for an equivalent query on the `silver` (Parquet) table.
    - **Answer:** Why did the query on Parquet scan significantly less data? How does this (and partitioning) directly relate to query speed and cost in Athena?
2. **Cleanup (Avoid Charges):**

- Navigate to **Amazon S3**.
- **Empty** your S3 bucket (`isd-<YOUR_GROUP>`) by deleting all files and folders inside it.
- **Delete** the S3 bucket itself.
- Navigate to **AWS Glue**.
- Delete the **table** (`bronze`) created by your crawler.
- Delete the **database** (`isd_db`).
- Delete the **crawler** (`isd-bronze-crawler`).
- (Athena views/tables are just metadata in the Glue Catalog, so deleting the database removes them).

Task 8: Reflection Questions

Answer the required question below, and then choose **any 2** of the 5 remaining questions to answer (for a total of 3 answers).

Required Question (Answer this):

1. **AWS Glue & Athena:** Based on your experience in Tasks 3-7, explain the main benefit of using AWS Glue (a serverless catalog) and Athena (a serverless query engine) compared to setting up a traditional, server-based data warehouse. Why was the data scanned in Task 7 so much lower for Parquet than for the raw CSV?

General Analytics Questions (Choose 2):

2. **Amazon Kinesis:** When would you use a service like Kinesis Data Streams (which handles real-time, streaming data) instead of the batch-loading (CSV files) approach we used in this assignment?
3. **Amazon EMR:** Athena is good for interactive SQL, but what kind of large-scale data processing jobs (e.g., using Spark) is Amazon EMR designed for that would be difficult or impossible to do in Athena?
4. **Hadoop & MapReduce:** In simple terms, what core problem did the original Hadoop MapReduce framework solve for data processing?
5. **Amazon QuickSight:** We finished with a "Gold" table (Task 5). How would a service like Amazon QuickSight connect to our data in Athena to build visualizations and dashboards for business users?
6. **EMR vs. Glue:** Both Amazon EMR and AWS Glue (specifically Glue ETL jobs) can be used to run Spark jobs. Why might you choose one over the other for a production ETL pipeline?