Extract, Transform, and Load Certification Training

Course-End Project Problem Statement

Course-End Project - 2
Social Media Data Integration and Analysis using AWS Glue

Overview

This project aims to enhance data management capabilities through AWS Glue, an ETL service, to streamline data processing workflows and derive actionable insights from social media data. Specifically, the project focuses on integrating and analyzing data from Twitter and Blogs to understand customer sentiments and trends. The insights gained from this analysis will help optimize marketing strategies and overall business performance.

Instructions

- Review the learning materials in the ETL course
- Carefully read the situation, tasks, actions, and results sections to grasp the assignment
- Complete and submit your assignment via the Learning Management System (LMS)
- Follow the provided guidelines closely, ensuring your report includes all required analyses and interpretations

Situation

You are a data analyst tasked with improving data processing and analysis workflows. Your company aims to leverage AWS Glue to streamline its data processing workflows and derive insights from social media data (Twitter and blogs). Your role is crucial in unlocking insights from this data to drive business growth and enhance operational efficiency.

Task

Your task is to use AWS Glue to integrate Twitter and blog data, ensuring data integrity. Develop a script to cleanse and transform the data and summarize the average sentiments by user ID.

Action

1. **Login to the AWS Console**
    - Open your web browser and navigate to the AWS Management Console
    - Log into the AWS Management Console with your account credentials
    - Navigate to **S3**
    - Click on **Create bucket** and add the bucket name as **etl-twitter-blog**
    - Scroll down the screen and click on the **Create bucket** button
2. **Create two folders inside the etl-twitter-blog**
    - Click on **etl-twitter-blog** and **Create folder**. Add the folder name as **etl-social-media** for the first folder. Scroll down the screen and click on the **Create folder** button.
    - Inside the folder, click **Upload** and click **Add files**. Select the **sample_tweets** file from your local system and click on **Open**.
    - Scroll down and click **Upload.**
    - Repeat the same step for **blog-data** and upload the **sample_blogs** dataset.
3. **Create a new bucket**
    - Create a new bucket named **etl-cep-output,** as you did above.

4. **Navigate to AWS Glue and create a new database**
    - In the AWS management console, search for **AWS Glue** and select it.
    - In the AWS Glue console, navigate to the **Databases** section
        - Click on **Add database** and provide the name as **social_media_data**. Scroll down and click on **Create database.**
5. **Set up two classifiers to read transaction data and product data**
    - Navigate to the Data Catalog, click on **Classifiers**, and click **Add Classifier.**
    - Fill the first classifier details as given below, and click on **Create**
        - **Classifier name** as **twitter_data**
        - **Classifier type and properties** as **CSV**
        - **CSV Serde – optional** as **None**
        - **Column delimiter** as **comma(,)**
        - **Quote symbol** as **Double-quote(“)**
        - **Column headings** as **Has headings**
    - Fill in the second classifier details as given below, and click on **Create**
        - **Classifier name** as **blog_data**
        - **Classifier type and properties** as **CSV**
        - **CSV Serde – optional** as **None**
        - **Column delimiter** as **comma(,)**
        - **Quote symbol** as **Double-quote(“)**
        - **Column headings** as **Has headings**
6. **Create an IAM role**
    - In the AWS management console, search for **IAM** service and select it
    - Navigate to **Roles** and click on **Create role**
    - Select **AWS service** as the **Trusted entity type** and **Glue** as the **Use case** and click **Next**
    - Select the **AdministratorAccess** policy. Scroll down and click on **Next**

- Enter the name as **glue-role**. Scroll down and click on **Create role**.

7. **Set up a Crawler**
   - Navigate to **AWS Glue**, click on **Databases** from the Data Catalog, and select **social_media_data** database
   - Click on **Add tables using crawler**
   - Enter the name as **tweet-crawl** and click on **Next**
   - Click on **Add a data source**
   - Click on **Browse S3**, click on **etl-twitter-blog**, select **etl-social-media,** and click on **Choose**
   - Click on **Add an S3 data source**
   - Choose classifier as **twitter_data** from the drop-down of **custom classifiers – optional** and click **Next**
   - Choose **glue-role** in the **IAM role** section and click **Next**
   - Choose **Target database** as **social_media_data** and click **Next**
   - Click on **Create crawler**
   - Click on **Run crawler**

> **Note:** Create a Crawler for Blog Data
>
> Repeat the above steps for the **blog data**, naming the crawler **blog-crawler** and using the **blog-data** folder and the classifier for **blog data**.

8. **Create an ETL job**
   - Navigate to **AWS Glue**, click on **ETL jobs,** and click on **Visual ETL**
   - In the **Add nodes**, double-click on **AWS Glue Data Catalog**
   - Select **Join** from the add nodes and link **Join** to both the **AWS Glue Data Catalog**
   - Click on the **Join** box and select **Drop Fields** from the **Add nodes**
   - Click on the **Drop Fields** box and select **Regex Extractor** from the **Add nodes**
   - Click on the **Regex Extractor** box and select **Aggregate** from the **Add nodes**
   - Click on the **Aggregate box** and select **Amazon S3** from the Targets in **Add nodes**

- Click on the first **AWS Glue Data Catalog** box, select **social_media_data** in the **Database** dropdown, and select **sample_tweets_csv** under the **Table** dropdown
- Click on the second **AWS Glue Data Catalog** box, select **social_media_data** in the **Database** dropdown, and select **sample_blogs_csv** under the **Table** dropdown
- Select the **Join** box and add the **AWS Glue Data Catalog** in the node parents. Select **Inner join** in the **Join type**. In the **Join conditions** box, select the **user id** in both the **AWS Glue Data Catalog** boxes
- Click on the **Drop Fields** box, and in the **Drop Fields** section, select the **user id** as it appears twice
- We need to extract tweet text values as they have a # symbol. Click on **Regex Extractor** and fill in the following fields:
    - **Column to extract from** as **tweet text**
    - **Regular expression** as **#(\w+)**
    - **Extracted column** as **hashtags**
- Now, let us create our summary report using **Aggregate** block. Fill in the following fields:
    - **Fields to group by** as **right_user id**
    - **Field to aggregate** as **tweet id** and **Aggregation function** as **count**
    - Click on **Aggregate a column,** then in **Field to aggregate** as **timestamp** and **Aggregation function** as **min**
- Click on the **Amazon S3** block, click on **Browse S3,** and select **etl-cep-output**
- On the top left corner, click **Untitled job** and give the **etl-cep-job**. Now, click **Save** and click **Run**
- Check progress in **Runs**

9. **Check the output and run the sql query**
    - Navigate to **S3** bucket, click on **etl-cep-output** bucket to check the output
    - Click on **run-1716301995938-part-r-00000**

- Navigate to **Object actions** and select **Query with S3. Select**
- In the **Output** settings, select **CSV** format and click **Run SQL query**

Result

Create a Word document detailing the steps you have performed, including screenshots. Upload the solution document to the Learning Management System (LMS).