DATA AND
ARTIFICIAL INTELLIGENCE

# Big Data on AWS

DATA AND
ARTIFICIAL INTELLIGENCE

# Batch Time Analysis of Transactional Data

# Objectives

- To analyze transactions, uncover patterns, and share actionable insights using the AWS Big Data stack

- To provide valuable insights and statistics across products, brands, categories, and segments

# Prerequisites

- Amazon S3
- AWS Glue
- AWS Glue Studio
- Apache Spark
- Athena

# Industry Relevance

- Amazon S3: It provides industry-leading data availability, data security, and performance through object storage.
- AWS Glue: It is a serverless data integration service for discovering, preparing, and combining data to support analytics, machine learning, and application development.
- AWS Glue Studio: It is a graphical interface that simplifies the creation, execution, and monitoring of extract, transform, and load (ETL) jobs in AWS Glue.
- Apache Spark: It is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
- Amazon Athena: It is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

# Problem Statement

Lenodo is a multinational e-commerce organization and sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

# Dataset Description

Download the data_utf8.csv and Demofile.docx files from the course resources section

# Tasks to Perform

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)
   - Create a bucket by providing the Bucket name
   - Upload and add files
   - Create another bucket to store the parquet file that will be generated after executing the script
2. Create a crawler to crawl the CSV data and generate a metadata catalog
   - Search for AWS Glue
   - Add a database in the glue console to hold the metadata tables
   - Now, create a crawler for this

# Tasks to Perform

- Specify the crawler source type as Data Stores
- Choose the data store as S3 and crawl data in a specified path
- Choose an existing IAM role that can execute the glue crawler with the required permissions
- Keep the frequency as Run on Demand
- Select the database that you created earlier
- Now, select the crawler and click on Run crawler
- Click on the table. It will show you the details about the metadata of the CSV file along with all the columns

# Tasks to Perform

3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

   - Select AWS Glue Studio
   - Select the Spark script editor
   - Copy the code given in Demofile.docx and paste it into the query window
   - Provide the name of the script as “Demo Job,” and choose the same IAM role in the job details option you created while creating the crawler to save the script
   - After 5 to 6 minutes, you will see the run status as "Succeeded" in the runs options, which are shown after the Job details option. If the job fails, then you need to check the permissions for the IAM role

# Tasks to Perform

- The IAM role that you created while creating a crawler must have “S3 Bucket full access” and “Administrator Full Access” permissions
- If you are facing any IAM errors, try to give full permissions on your bucket to the glue role
- Search for IAM to provide permissions
- Click on the “Roles” on the left side of the pane to attach policies
- Select the IAM role that you created and click on “Attach policies”
- Search for the S3 full access, select the option and attach a policy
- After attaching the policies, you can execute the code again
- Now, go to the S3 bucket to check if a parquet file has been generated successfully

# Tasks to Perform

4. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file to query it with Athena
   - Click on “Add crawler” in the same database and provide a name to the crawler
   - In the add a data store page, choose the parquet file path in the Include path
   - Choose the same IAM role that you have been using from the start with the attached policies
   - Choose the same database
   - Click on “Finish" and run the crawler
   - Go to Tables. You will be able to see another table for the metadata table generated by the parquet crawler.

# Tasks to Perform

5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena
   - Search for “Athena”
   - Choose the database as the one created earlier
   - Execute the query “which item was sold mostly” by typing the code mentioned below:
     - Select stockcode, quantity, and description from "capstone"."projectparquet1901" group by stockcode, quantity, description order by quantity DESC
   - Execute another query “which country bought the most sold item”
     - Select country, a quantity from "capstone"."projectparquet1901" where stockcode = '23843' group by country, quantity order by quantity DESC

# Project Outcome

- The aim of the project is to analyze transactions, uncover patterns, and share actionable insights using the AWS Big Data stack.

DATA AND
ARTIFICIAL INTELLIGENCE

# Submission Process

1. Complete the project in the Simplilearn lab
2. Complete each task listed in the problem statement
3. Take screenshots of the results for each question and the corresponding code
4. Save it as a document and submit using the assessment tab
5. Tap the "Submit" button (this will present you with three choices)
6. Attach three files and then click "Submit“

Note: Be sure to include screenshots of the output

DATA AND
ARTIFICIAL INTELLIGENCE

Thank You



simplilearn