# PROJECT NAME: Batch Time Analysis of Transactional Data

# **DESCRIPTION**

Lenodo is a multinational e-commerce organization that 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.

**OBJECTIVE:**

To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights

# **Steps to perform:**

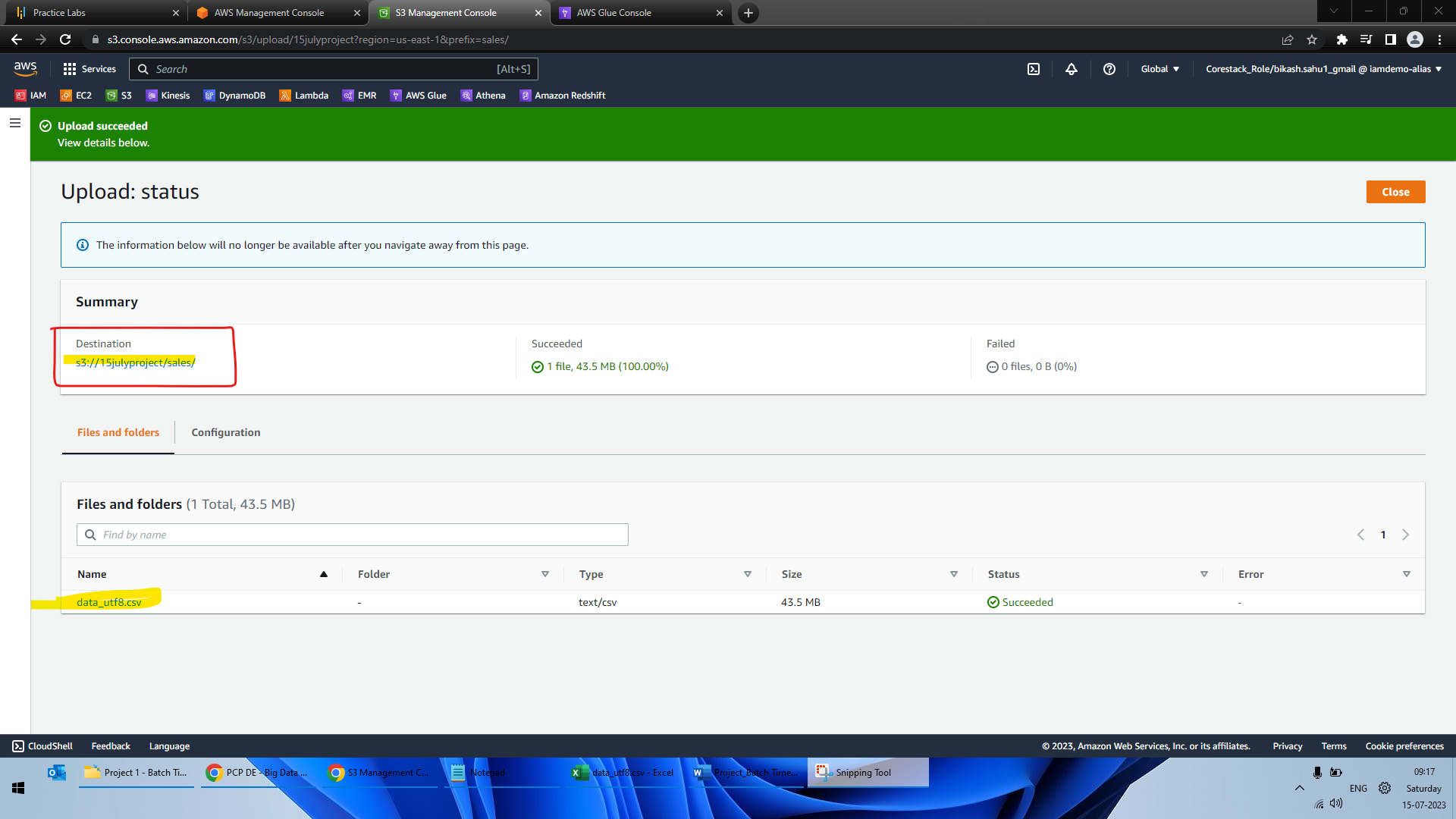
1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)



1. Create a crawler to crawl the CSV data and generate a metadata catalog
2. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries
3. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena
4. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

# **Results:**

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)



1. Create a crawler to crawl the CSV data and generate a metadata catalog

A screenshot of a computer

Description automatically generated

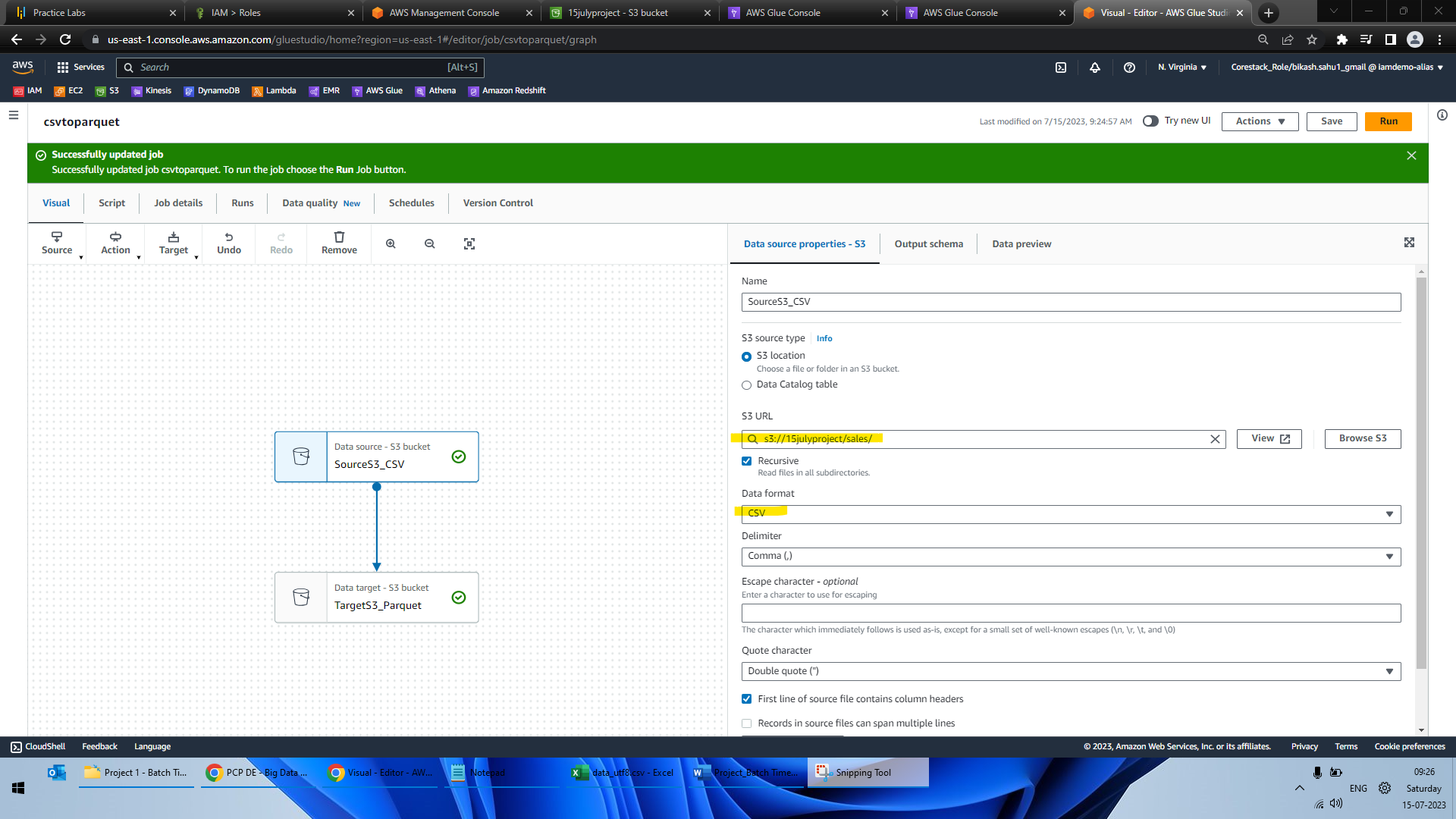
Table is created.

A screenshot of a computer

Description automatically generated

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

* **Source Job**



* **Target Job**

A screenshot of a computer

Description automatically generated

* *Run Job to get parquet file in S3*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena

* **New crawler**

A screenshot of a computer

Description automatically generated

* **New table created from parquet file**

A screenshot of a computer

Description automatically generated

1. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

* **Best selling item**

select count(\*) as quantity, description from sales\_parquet group by description order by quantity desc limit 1;

A screenshot of a computer

Description automatically generated

* **Countries where customers have bought the most-sold item**

select distinct country, description from sales\_parquet where description in (select description from (select count(\*) as quantity, description from sales\_parquet group by description order by quantity desc limit 1));

A screenshot of a computer

Description automatically generated

A computer screen shot of a computer

Description automatically generated