**Project 1: 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:**

* 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):
  + S3 > Create Bucket > Name: [batch-time-analysis-of-transaction-data](https://s3.console.aws.amazon.com/s3/buckets/batch-time-analysis-of-transaction-data?region=us-east-1):
    - Create folder: input
    - Create folder: output
    - Upload CSV data into: s3://batch-time-analysis-of-transaction-data/input/
* Create a crawler to crawl the CSV data and generate a metadata catalog:
  + AWS Glue > Crawlers > Add Crawler:
    - Name: input-crawler
    - Data Stores Path: s3://batch-time-analysis-of-transaction-data/Products\_dataset.csv
    - IAM Role: AWSGlueServiceRole-Admin
    - Schedule: Run on Demand
    - Add Database: input-database
  + Run Crawler: input-crawler
  + AWS Glue> Databases:
    - Create Database: output-database
* Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries:
  + Create Job > Source S3 to Target S3:
    - Job Name: csv-to-parquet
    - Data Source:
      * S3 Location URL: s3://batch-time-analysis-of-transaction-data/input/Products\_dataset.csv
      * Data Format: CSV
      * Delimiter: Tab
      * Optimize CSV: True
    - ApplyMapping:
      * Adjust data types
      * Drop unwanted columns
    - Data Target
      * S3 Location URL: s3://batch-time-analysis-of-transaction-data/output/
      * Data Format: Parquet
      * Compression Type: Snappy
    - Save Job > Run Job > Run Details
* 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:
  + AWS Glue > Crawlers > Add Crawler
    - Name: parquet-crawler
    - Data Stores Path: s3://batch-time-analysis-of-transaction-data/output/Products\_dataset.parquet
    - IAM Role: AWSGlueServiceRole-Admin
    - Schedule: Run on Demand
    - Database: output-database
    - Run Crawler: parquet-crawler
* Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena.

SELECT \* FROM "input-database"."products\_dataset\_csv" limit 10;

SELECT country AS Country, description AS Description, quantity AS Quantity

FROM "input-database"."products\_dataset\_parquet"

GROUP BY quantity DESC

limit 2;

ANSWER:  
United Kingdom – PAPEL CRAFT, LITTLE BIRDIE – 80995

United Kingdom - MEDIUM CERAMIC TOP STORAGE JAR – 74215