hansa cequity

ENRICHING CUSTOMER EQUITY

ISO/ IEC 27001:2013 CERTIFIED
ISO/ IEC 27701:2019 CERTIFIED
ISO 9001:2015 CERTIFIED

Data Engineering Analyst (Round 2: Technical)- Data Sol

Objective:

This task simulates a real-world data ingestion and analytics pipeline in the context of the following data-set titled **AI-driven digital transformation**.

Dataset:

A global dataset capturing the **impact of AI-generated content** across industries, regions, and time.

Provided File: Global_AI_Content_Impact_Dataset.csv

Assignment Instructions

AWS Environment Setup

1. Create an account on AWS
2. Launch an EC2 instance (free tier instances are available)
3. Generate key pairs and set it up such that only you can SSH into it
4. Install python3.6, docker on it

S3 & REST API

5. Create a new S3 bucket
6. Write a REST API (using python3.6/flask) to upload file to S3

Dockerization & Deployment

7. Dockerize the API created in Step 6 and deploy it on the EC2 machine
8. Now deploy the REST API created in Step 6 on API Gateway

Lambda Integration

9. Invoke a Lambda function on the s3 event when the file is uploaded on S3
    a. The lambda function should read the file (Data file attached in the email)

b. Perform following transformations (using python pandas) on the data

1. **Clean Up**:
   - Drop rows where `Country`, `Industry`, or `AI Adoption Rate (%)` is null
   - Convert `AI Adoption Rate (%)` column to `whole number` format rounded off to nearest integer from `floating point number`
2. **Rename Columns**:
   - `AI Adoption Rate (%)` → `ImpactScore` in `whole number` format
   - `Industry` → `ConsumerIndustry`
3. **Filter Rows**:
   - Keep rows only where `ImpactScore >= 50`
4. **Transformations**:
   - Calculate average `ImpactScore` **per country**
   - Count of records **per ConsumerIndustry**
   - Max `ImpactScore` **by ConsumerIndustry**
5. **Save & Store Output**:
   - Cleaned file → `filename_cleaned.csv`
   - Summary metrics → `filename_summary.csv`
   - Upload both files back to the S3 bucket

Analytics & Visualization Task

Analytics Task

Perform a **brief exploratory analysis** on the transformed data (can be done in Jupyter or a simple script):

- Top 5 countries with the highest average AI content impact
- Most affected content category and industry
- Trend: How has ImpactScore changed over time (basic plot)

QuickSight Dashboard Creation

Use **Amazon QuickSight** to create a simple dashboard (based on your cleaned or summarized S3 data):

Include the following visuals:

1. **Bar Chart**:
    a. X-axis: `Country`
    b. Y-axis: Avg `ImpactScore`
    c. Sorted descending
2. **Pie or Donut Chart**:
    a. Distribution of `ContentCategory`
3. **Line Chart**:
    a. X-axis: `Date`
    b. Y-axis: Avg `ImpactScore`
    c. Grouped by `Industry`

Optional Bonus Chart Ideas:

- Heatmap of `ImpactScore` by `Industry` vs. `Country`
- KPI card showing:
    - Max ImpactScore
    - Number of entries after cleaning
    - Total affected countries

Submission Checklist

Please email the following:

- Code files (zipped or GitHub repo link)
- Dockerfile & instructions
- Output files written to S3 (or screenshots + sample outputs)
- Mini-analysis results
- Summary of your approach, assumptions, and challenges (1 pager)
- Screenshots of your QuickSight dashboard/ Visualizations
- (Optional) Embed the dashboard using a public/shared link if your AWS account allows
- The underlying cleaned or summary CSV used for QuickSight

- A brief description of each chart: what it shows, and one insight from it

Candidate Notes:

If you've never used QuickSight before, don't worry:

- You can use the **free trial of Amazon QuickSight**
- If you run into permission issues (e.g., accessing S3 via QuickSight), feel free to mock the dashboard using Excel/PowerBI/Tableau and describe what you would have done in QuickSight

Time Estimate:

4–6 hours, depending on experience. You're not expected to polish everything, but show your **problem-solving and thought process. Data Transformations, Analysis & Visualization task holds higher weightage.**