# Snowflake End-to-End ETL Assignment

## Business Case

Your company’s data engineering team wants to build a data pipeline in Snowflake that ingests, cleans, and analyzes real-world data from a public S3 bucket. You are asked to demonstrate the full ETL flow: Extract → Transform → Load, including error handling and reporting.  
  
The dataset is available in the public S3 bucket:  
s3://noaa-ghcn-pds/csv/2024.csv.gz  
  
This file contains daily global weather observations such as temperature, precipitation, and snow depth.

## Assignment Tasks

### 1. Data Ingestion (Bronze Layer)

- Create a Snowflake database and schemas (BRONZE, SILVER, GOLD).  
- Define a file format for the dataset (CSV,PARQUET).  
- Create an external stage that points to the S3 URL.  
- Ingest the raw data into a Bronze table while capturing metadata (filename, load timestamp).  
- Validate the row count after ingestion.

### 2. Data Validation & Rejected Records

- Some rows may contain missing values, invalid quality flags, or corrupted entries.  
- Design a process to segregate rejected records into a separate table (BRONZE.REJECTED\_DATA).  
- Document which validation rules you applied (e.g., invalid dates, NULL station IDs, failed quality flags).

### 3. Data Transformation (Silver Layer)

- Convert raw dates (YYYYMMDD) into proper Snowflake DATE.  
- Apply unit conversions:  
 • TMAX, TMIN (tenths of °C → °C).  
 • PRCP (tenths of mm → mm).  
- Filter out rejected/invalid rows.  
- Restructure data so each row contains station, date, TMAX, TMIN, PRCP.  
- Load the cleaned and structured data into the Silver table.

### 4. Reference Data Integration

- Ingest reference files:  
 • ghcnd-stations.txt → Stations metadata (station ID, lat, lon, elevation, state, name).  
 • ghcnd-countries.txt → Country codes and names.  
- Join weather data with station and country details.

### 5. Reporting Layer (Gold Layer)

- Build reporting tables such as:  
 1. Extreme Heat Days Report – Identify days where max temperature > 35°C.  
 2. Monthly Precipitation by Country – Aggregate precipitation totals per country per month.  
 3. Top 10 Wettest Stations – Find stations with highest rainfall in the last 30 days.

### 6. Automation with Tasks & Streams

- Create a Stream on the Bronze table to track new data.  
- Implement a Task pipeline:  
 • Task 1: Ingest new data from stage → Bronze.  
 • Task 2: Transform Bronze → Silver.  
 • Task 3: Refresh Gold reporting tables.  
- Schedule tasks to run daily (e.g., 01:30 AM IST).

### 7. Deliverables

- Snowflake DDL/DML scripts for ingestion, transformation, reporting.  
- Evidence of segregated rejected data.  
- Task pipeline diagram (showing flow Bronze → Silver → Gold).  
- Two final reporting queries (Extreme Heat, Monthly Precipitation).  
- Screenshots of task history proving automation ran successfully.