Project Title:
Performance Comparison of Data Processing: Cloud vs. Local Setup

Overview:
The goal of this project is to compare the performance of a data processing pipeline on both a cloud-based setup (Google Cloud) and a local environment. The project involves uploading data to Google Cloud Storage, processing it, and measuring key performance metrics such as execution time and resource usage. The same process is replicated locally, and a comparative analysis is done.

System Design:

1. **Input Data**
    - Large datasets generated using Python (50MB and 100MB CSV files).
    - Uploaded to Google Cloud Storage for cloud processing.
    - Stored in a local folder for local processing.
2. **Processing**
    - A Python script processes the data on both setups.
    - Performance metrics (execution time, memory usage) are recorded.
3. **Comparison**
    - A bar graph is generated comparing the performance metrics from both setups.
4. **Output**
    - A bar graph representing the performance difference.
    - A report summarizing the experiment and key findings.

Folder Structure:

- **barGraph**
  Contains the generated bar graphs for the performance comparison.
- **code**
  Contains Python code files:
    - `process_data.py`: Script for processing the data.
    - `generate_data.py`: Script for generating synthetic data.
    - `generate_bar_graph.py`: Script for generating the performance comparison bar graph.
- **data**
  Contains the data files (CSV files for processing).
- **runReport**
  Contains the metrics and performance comparison report.

Cloud Architecture:

- Google Cloud Storage: Stores data files.
- Google Cloud Shell: Runs Python scripts to process the data.

Performance Metrics:

- Execution time
- Resource usage (CPU, memory)

Tools Used:

- Google Cloud Platform (Cloud Storage, Cloud Shell)
- Python (for data processing, bar graph generation)
- Timeit and htop (for performance measurements)