Cloud vs Local: Performance Analysis with Large Dataset

A mini project on software & data engineering

Project Overview

1

Objective

Compare performance of data processing between cloud and local setup

2

Key Components

Data generation, processing, performance measurement, and visualization

Project Design

Cloud Services Used

Google Cloud Storage for storing large datasets

Cloud Run for processing data

Tools & Technologies

Python for data generation and processing

Matplotlib for visualizations

Google Cloud SDK for cloud communication

Local Setup vs Cloud Setup

Local Setup

Process large datasets stored locally

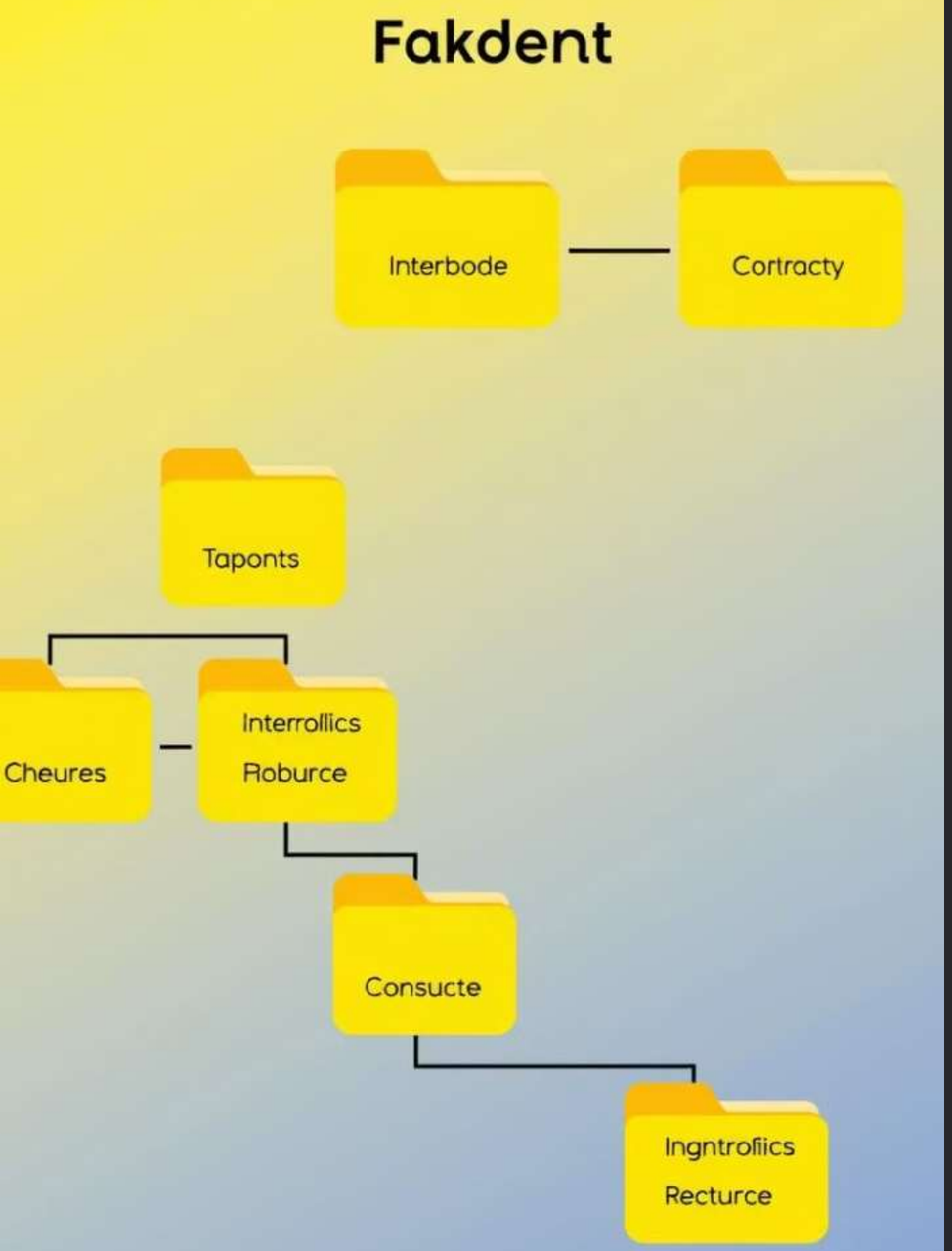
Cloud Setup

Upload datasets to Google Cloud Storage

Process the dataset using Cloud Run

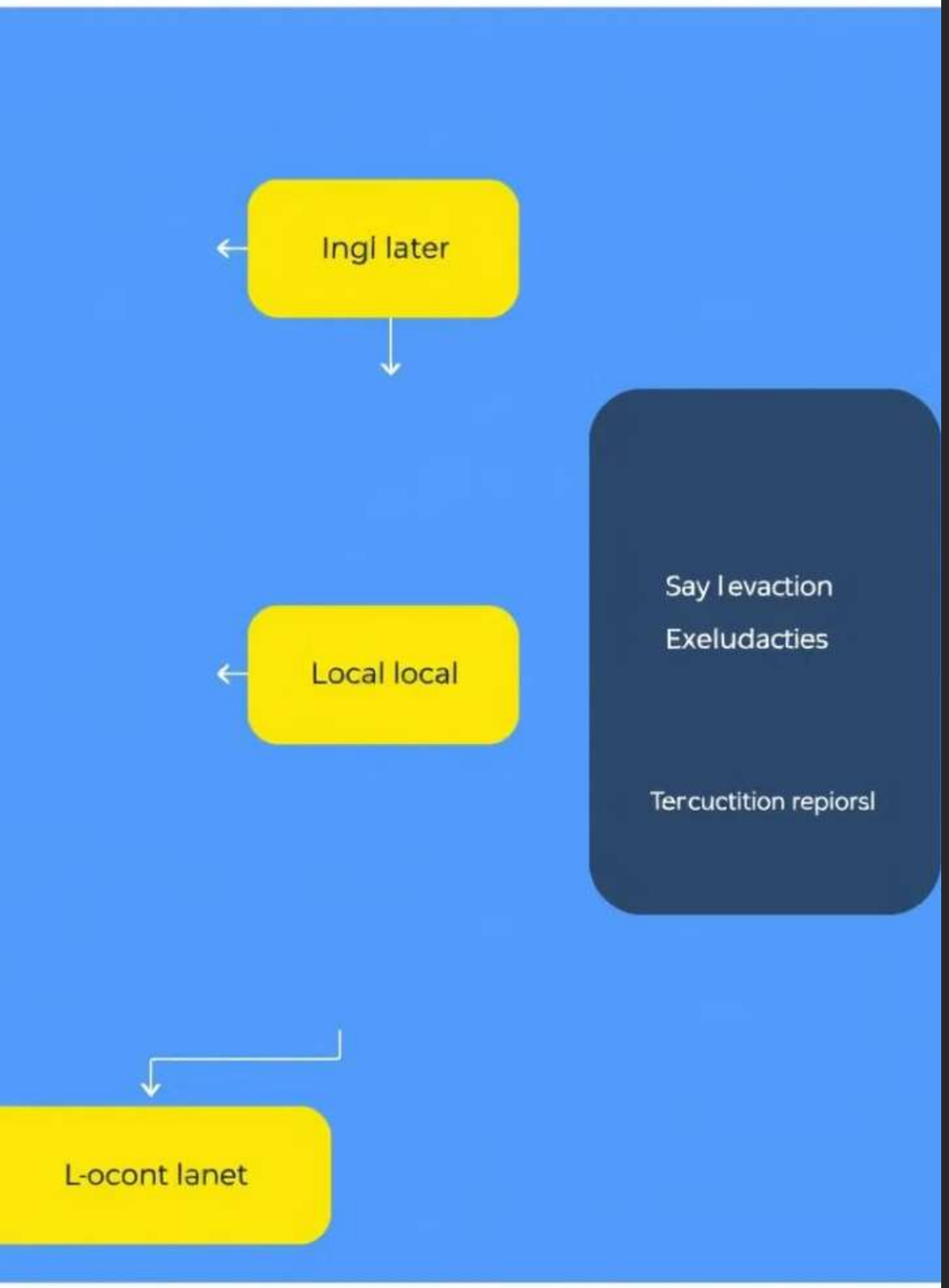
Metric

Measure execution time for both setups



Code Structure

Folders	Description
barGraph/	Contains bar graph images comparing performance
code/	Contains Python scripts for data generation, processing, and visualization
data/	Contains generated datasets
runReport/	Contains execution results for both cloud and local runs



Procedure: Local vs Cloud Processing

1

Local Processing

Execute process_data.py --local

2

Cloud Processing

Upload dataset to Google Cloud Storage

3

Cloud Processing

Execute process_data.py --cloud from Google Cloud Shell

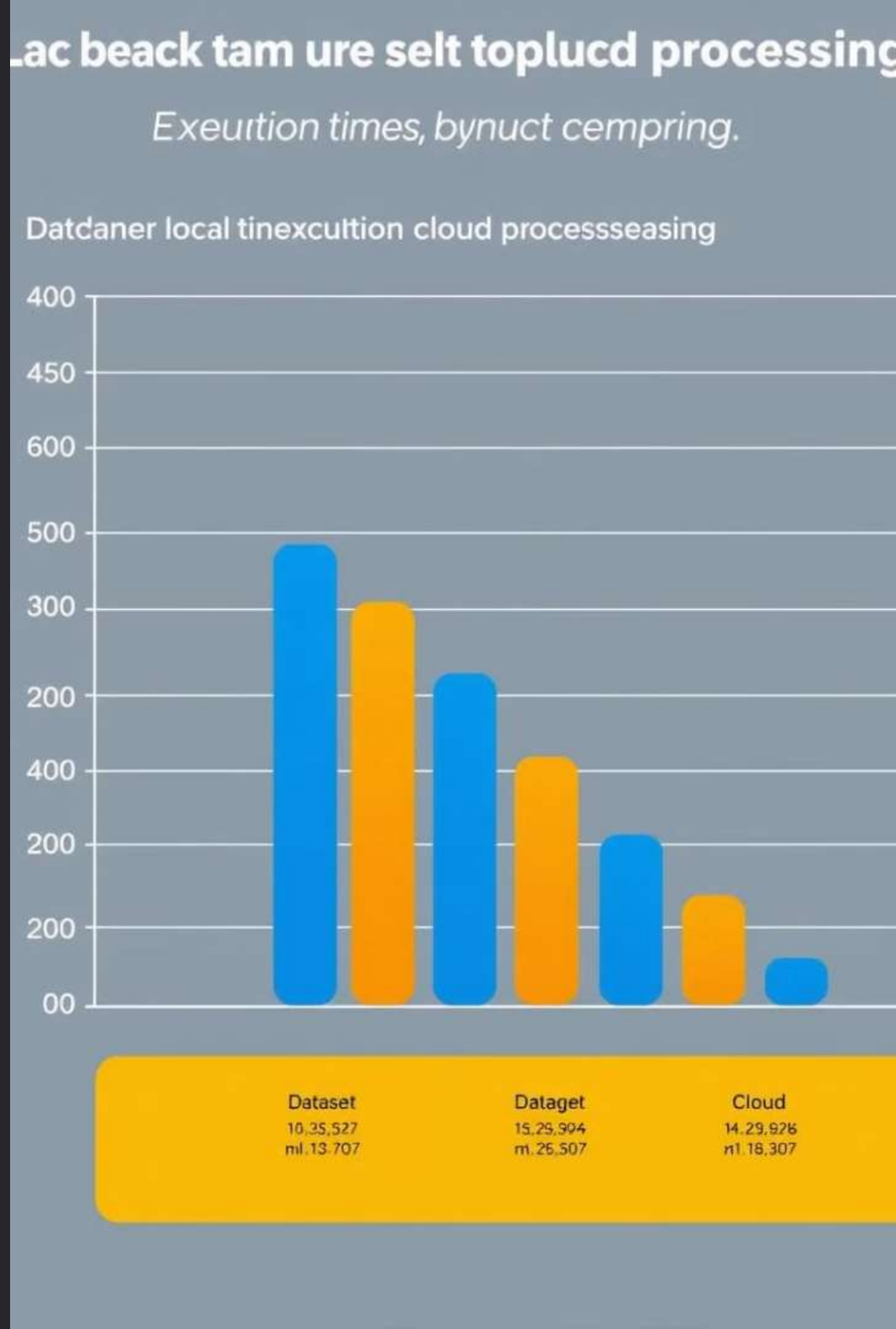
Bar Graph Visualization

Explanation

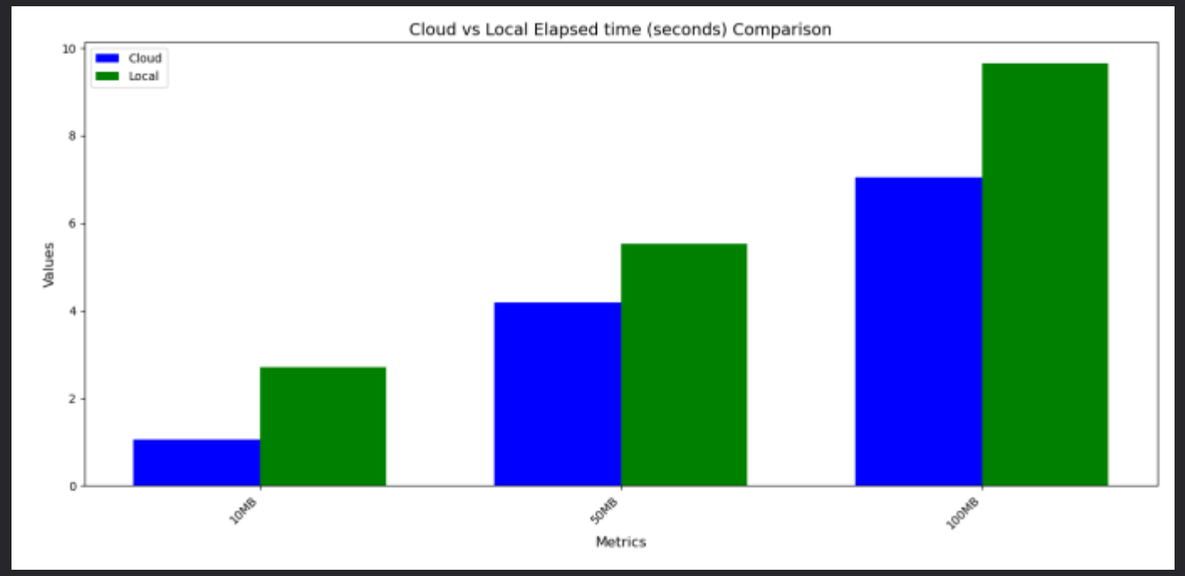
Bar graph comparing execution times for cloud vs local processing

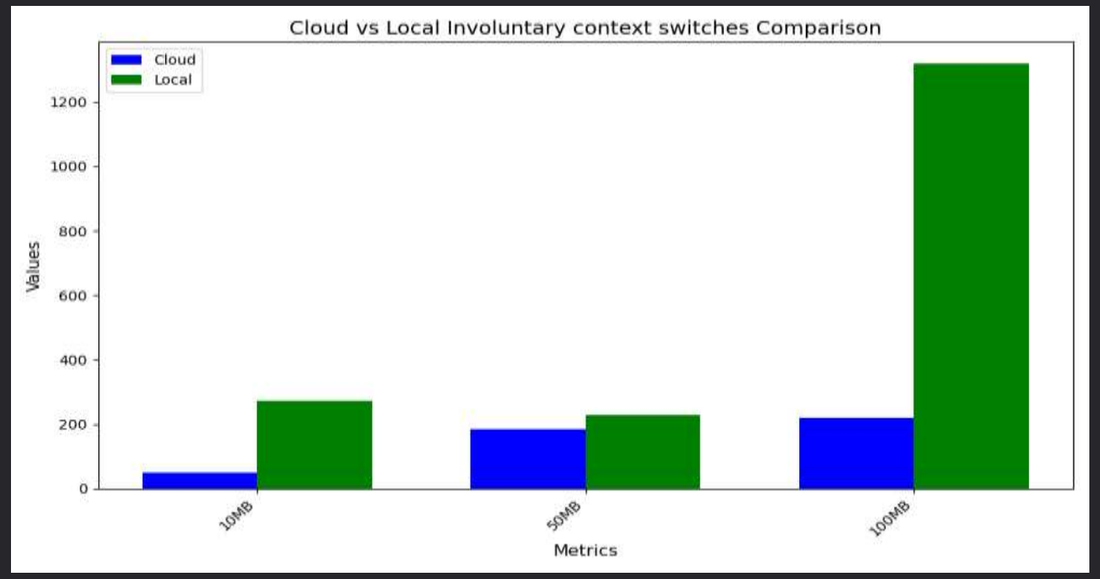
Interpretation

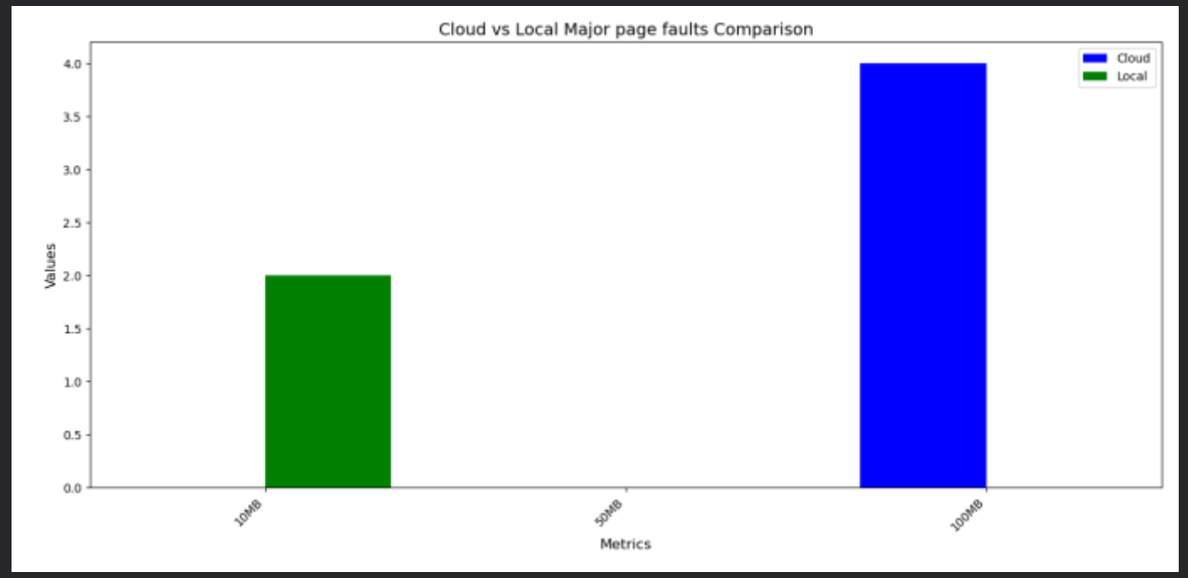
Cloud setup is expected to perform better for large datasets

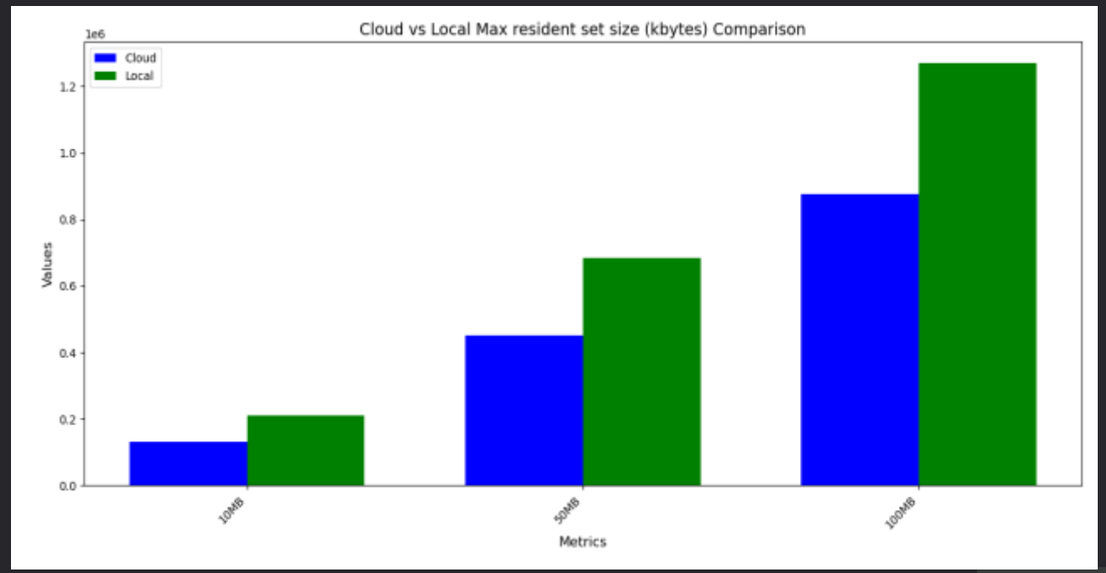


Snapshot Slides for Bar Graphs

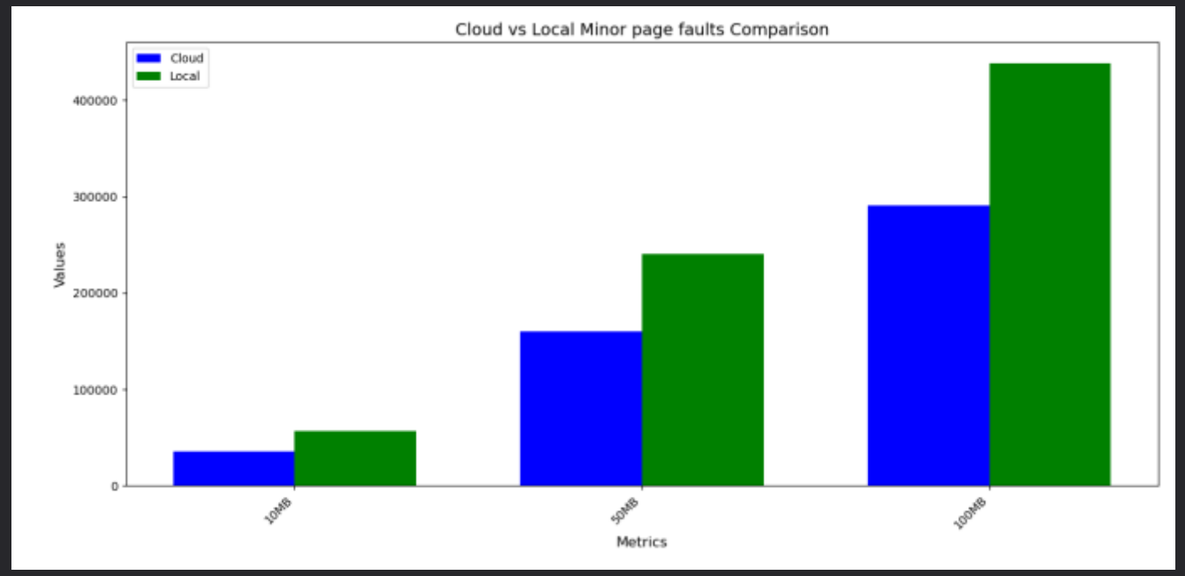


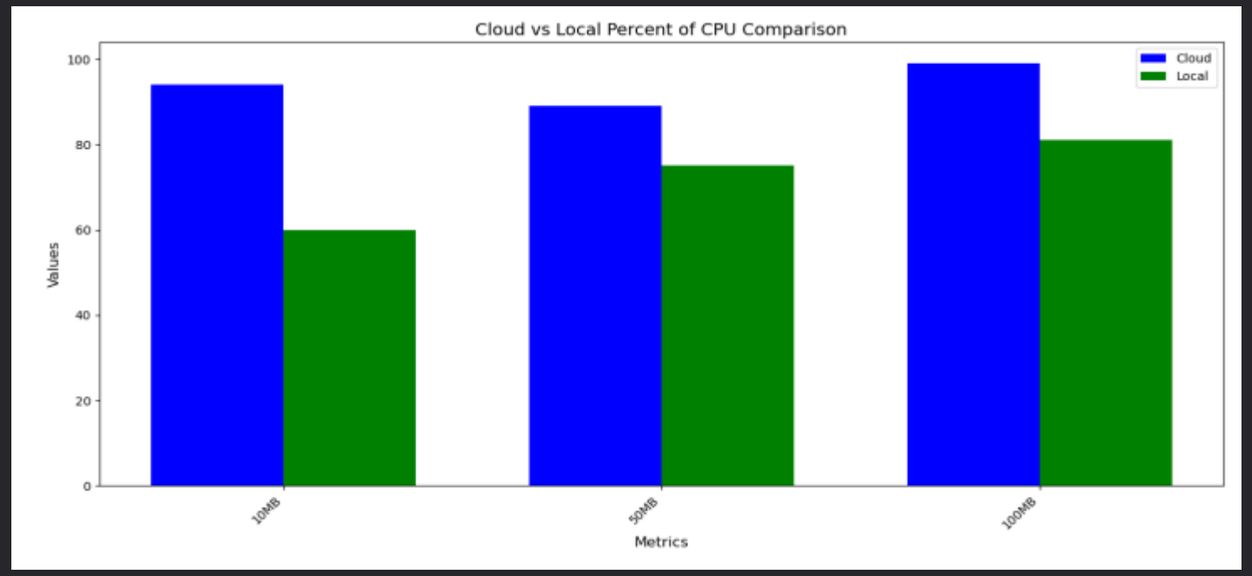


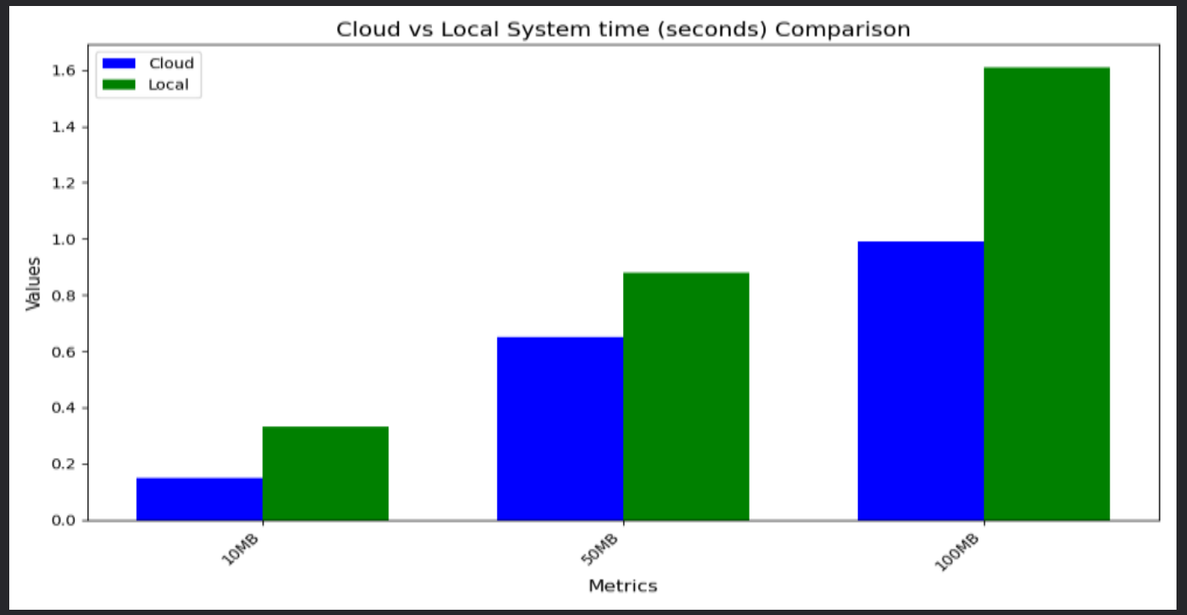


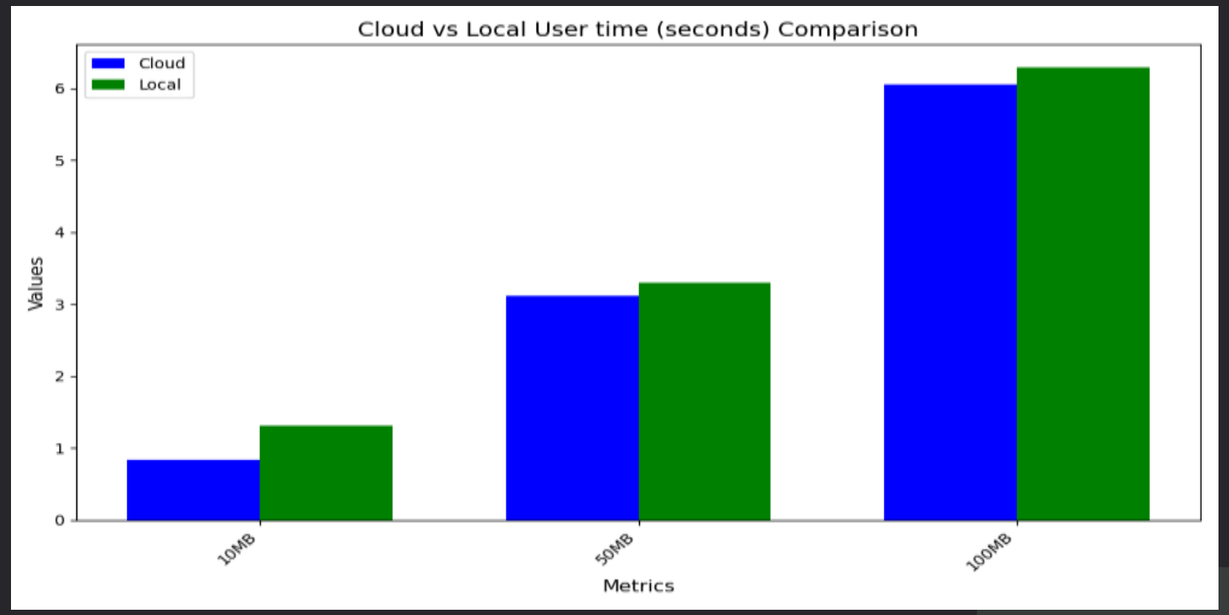


Snapshot Slides for Bar Graphs









Conclusions

1

Results

Cloud setup outperforms local setup for processing large datasets

2

Key Learnings

Cloud services can optimize resource usage and reduce execution time

Performance benefits become significant with larger datasets