Big Data Analysis on E-commerce Behavior Using Python, Pandas, Dask, and Parquet Compression

1. Introduction

This project performs comprehensive big data analysis on a real-world e-commerce dataset from Kaggle. The objectives include understanding customer interactions across different event types (view, cart, purchase) and benchmarking various data processing frameworks for large CSV files.

2. Dataset Overview

- Source: Kaggle's "E-commerce Behavior Data from Multi-category Store"
- Files: 2019-Oct.csv and 2019-Nov.csv (~5 GB each, millions of records)
- Fields: event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session

3. Project Objectives

- Efficiently process and analyze very large datasets using Python
- Compare and benchmark Pandas (chunked), Dask, and Parquet compression
- Generate actionable insights into user behavior, event distribution, price trends, and brand conversion rates

4. Implementation Steps

A. Data Acquisition & Preparation

- Installed Kaggle and required Python packages
- Downloaded and extracted large CSV datasets
- Explored sample data to understand schema and content

B. Chunked Data Processing with Pandas

- Read CSV files in 50,000 row chunks for memory-efficient processing
- Counted occurrences of each event type and calculated processing time
- Extracted top 3 brands per event type using groupby and value_counts

C. Benchmarking: Pandas vs. Dask vs. Parquet Compression

- Compared execution time, memory consumption, and MB/s for:
  - Pandas chunked processing
  - Dask dataframe processing
  - Compressed Parquet format
- Summarized results in a comparison table
- Visualized results using matplotlib bar charts

D. Detailed Analytical Tasks with Dask
- Calculated event distribution and average price per event type
- Analyzed user behavior hourly across event types
- Computed purchase conversion rates for products and brands
- Identified top brands per event and calculated brand conversion rates

E. Result Export and Visualization
- Saved analytical results to CSV files
- Generated distribution charts and comparison figures for reporting

5. Key Results
- Efficient handling of multi-million row datasets using chunked and parallel processing
- Dask and Parquet demonstrated significant improvements in speed and memory over Pandas chunking
- Brand-level and hourly user activity insights reveal trends for marketing and optimization

6. Conclusion
This project successfully achieved high-performance data analysis on large-scale e-commerce datasets by integrating efficient frameworks and benchmarking their real-world advantages. The documented workflow offers a repeatable blueprint for similar big data analytics tasks.

7. Technologies Used
- Python (Pandas, Dask, Matplotlib)
- Kaggle API
- Google Colab
- Parquet data format

8. References

- Kaggle Dataset: E-commerce Behavior Data from Multi-category Store