COP 290

ASSIGNMENT 1

SUBTASK 1

SATWIK

2022CS51150

Comparison of File Formats for Storing Stock Data

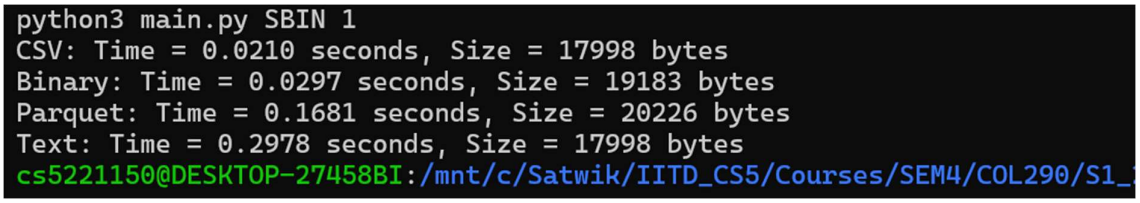
The aim of this benchmarking study is to assess the performance of different file formats for storing stock data. The considered file formats include CSV, Text, Binary, and Parquet. The assessment criteria encompass both the time taken for writing the data to files and the resulting file sizes.

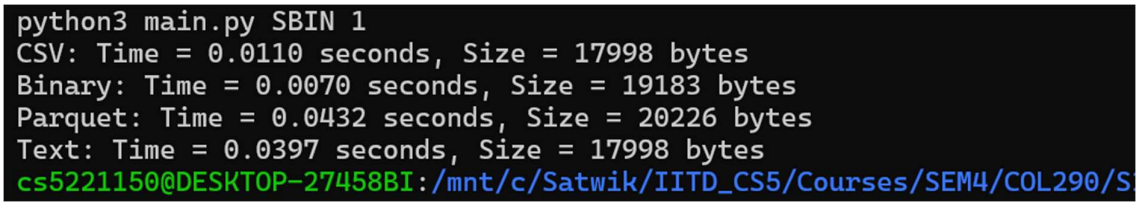
File Formats

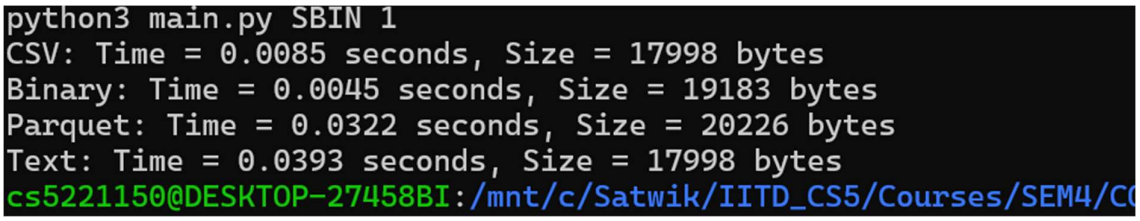
1. **CSV (Comma-Separated Values):** The data was saved in a CSV file format using the pandas `to_csv` method.
2. **Text File:** A text file was created with tab-separated values, providing a straightforward alternative to CSV.
3. **Binary File:** The data was stored in binary format using the `to_pickle` method from pandas.
4. **Parquet File:** The data was saved in the Parquet file format, leveraging the `pyarrow` library.

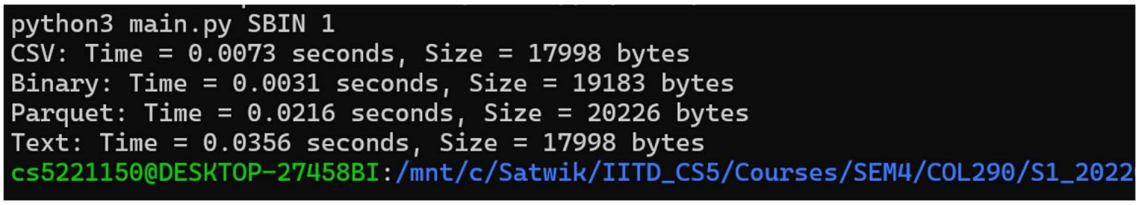
Benchmarking

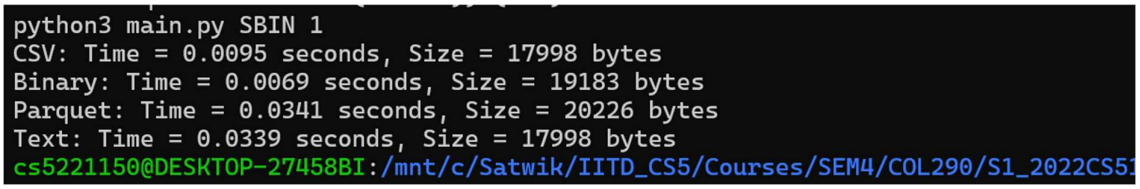
The benchmarking process involved measuring the time taken to write each file format and recording the resulting file sizes. This was achieved using the `time` library for time measurements and the `os` library for file size calculations. For each file type we have recorded the time 5 times. The final time is the average of these values.S











From the above screenshots, we find that the average time required(in seconds) for each of the file formats are:

CSV: 0.0114

Binary: 0.0102

Text: 0.0892

Parquet: 0.0598

The size of the different file formats(bytes) are:

CSV: 17998

Binary: 19183

Text: 17998

Parquet: 20226

From the data above, we can conclude that binary and CSV files require lesser time on average as compared to text and Parquet files. The space requirement of text and CSV files are the same and are also lesser than that of binary and Parquet files.

Although text files take longer time, they offer improved readability.

Overall, CSV files provide the best functionality because of their low time and space requirements.

ll these values.