Performance Comparison of Data Reading Methods

Table 1: Comparison between different data reading methods

Method	Time (seconds)	Max Memory per Chunk (MB)	Total Memory Used (MB)
Pandas	652.06	227.99	587.36
Dask	542.68	—	863.29
Compressed (Gzip)	760.83	227.99	623.32

Results and Discussion

From the results, we noticed that the reading method using **Dask** is the fastest (542.68 s) of the three methods, due to its reliance on parallel processing, while consuming more memory compared to the other methods (863.29 MB).

Reading using **Pandas** with chunk size was slightly slower than Dask, taking (652.06 s), but it consumed less memory, with total memory consumption reaching (587.36 MB), while the memory consumed for a chunk was (227.99 MB).

For the last method, i first tried to read the entire compressed file, but I couldn't because the system crashed due to the large file size, so I switched to reading the file using chunks. which was the slowest of the three methods, taking (760.83 s) despite the total memory used being (623.32 MB).

Recommendations

- It is preferable to use **Pandas with chunking** in resource-limited environments in order to achieve a balance between time taken and memory consumption.
- It is recommended to use **Dask** in the case of big data due to its reliance on parallel processing.