TP02: How to work with big data files (5gb+) in Python Pandas!

Data plays a key role in building machine learning and the AI model. In today's world where data is being generated at an astronomical rate by every computing device and sensor, it is important to handle huge volumes of data correctly. One of the most common ways of storing data is in the form of Comma-Separated Values (CSV). Directly importing a large amount of data leads to out-of-memory error and reading the entire file at once leads to system crashes due to insufficient RAM.

Working with large CSV files in Python

The following are a few ways to effectively handle large data files in .csv format and read large CSV files in Python. The dataset we are going to use is gender_voice_dataset:

- ✓ **Using pandas.read_csv(chunk size)**
- ✓ **Using Dask**
- ✓ **Use Compression**

Questions

1. Choose which large CSV files and apply these methods to it?
2. Compare them in terms of **time** and **storage**?