**22/12/2023**

**Introduction of PySpark**

PySpark is an Apache Spark library written in Python to run Python applications using Apache Spark capabilities. Using PySpark we can run applications parallelly on the distributed cluster (multiple nodes).

In other words, PySpark is a Python API which is an analytical processing engine for large-scale powerful distributed data processing and machine learning applications

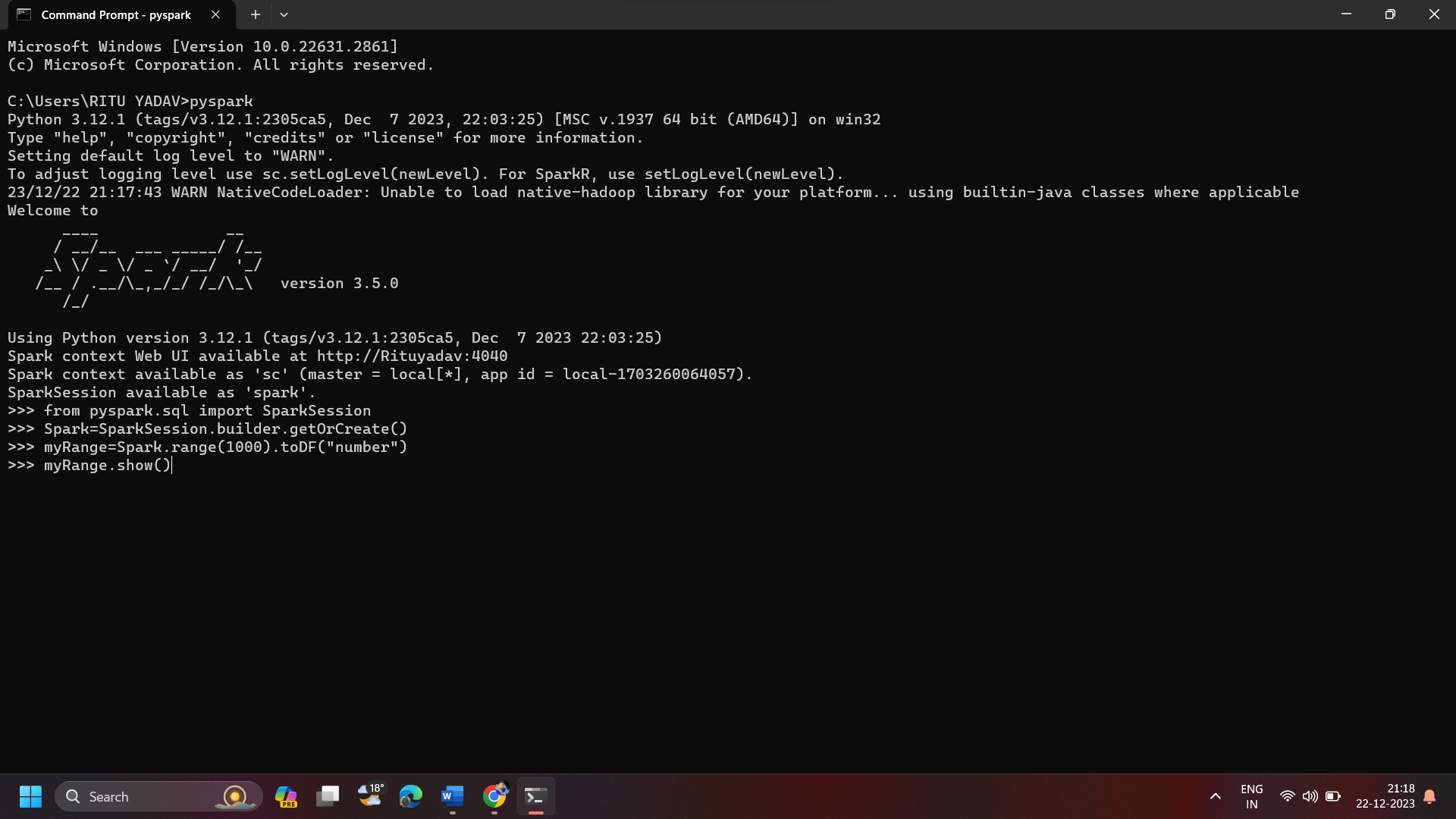
PySpark combines Python’s learnability and ease of use with the power of Apache Spark to enable processing and analysis of data at any size for everyone familiar with Python.

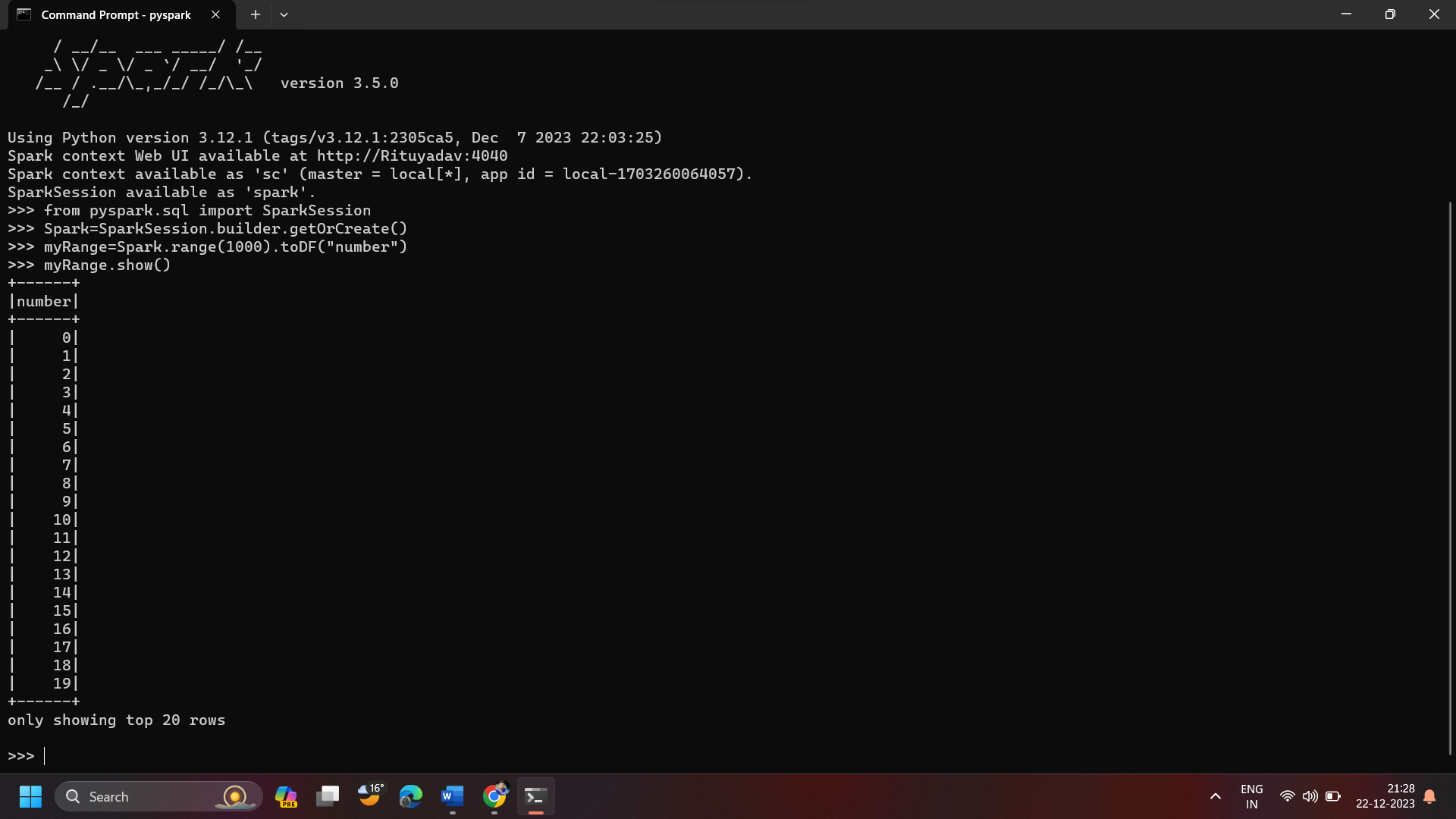
**Apache Spark**

Apache Spark is an open-source unified analytics engine used for large-scale data processing, hereafter referred it as Spark. Spark is designed to be fast, flexible, and easy to use, making it a popular choice for processing large-scale data sets. Spark runs operations on billions and trillions of data on distributed clusters 100 times faster than traditional applications.

Spark can run on single-node machines or multi-node machines(Cluster). It was created to address the limitations of MapReduce, by doing in-memory processing. Spark reuses data by using an in-memory cache to speed up machine learning algorithms that repeatedly call a function on the same dataset.  This lowers the latency making Spark multiple times faster than MapReduce, especially when doing machine learning, and interactive analytics.  Apache Spark can also process real-time streaming.

PySpark statements running

****

****