1. Налаштувати **WSL**
2. **Hadoop** налаштувати 1 в 1 за цим файлом <https://medium.com/@madihaiqbal606/apache-hadoop-3-3-6-installation-on-ubuntu-22-04-2-lts-wsl-for-windows-bb57ed599bc6>
3. **Apache Spark** (якщо два юзери в wsl –налаштувати потрібно для обох)

* Встановити **scala:** sudo apt install scala –y
* wget <https://dlcdn.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz>
* tar -xvzf spark-3.5.0-bin-hadoop3.tgz
* sudo mv spark-3.5.0-bin-hadoop3 /usr/local/spark
* nano ~/.bashrc
* export SPARK\_HOME=/usr/local/spark
* export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin
* export PYSPARK\_PYTHON=python3
* **Ctrl+S, Ctrl+X**
* source ~/.bashrc

**Можна протестувати:** spark-submit my\_file.py

from pyspark.sql import SparkSession

from pyspark.sql.types import StructType, StructField, StringType, IntegerType

spark = SparkSession.builder \

.appName("MyFirstSparkApp") \

.getOrCreate()

schema = StructType([

StructField("name", StringType(), True),

StructField("age", IntegerType(), True)

])

data = [("Alice", 30), ("Bob", 25)]

df = spark.createDataFrame(data, schema)

df.show()

Іще важливо, щоб pyspark використовував пайтон саме з wsl export PYSPARK\_PYTHON=/usr/bin/python3 (це теж в ~/.bashrc)

Далі можна виконувати код у звичайному .py файлі, а можна створити віртуальне середовище для .ipynb

* python -m venv spark\_env
* source spark\_env/bin/activate
* pip install pyspark findspark ipykernel
* python -m ipykernel install --user --name spark\_env --display-name "Python (Spark)"

Також потрібно налаштувати спарк сесію для цього

import findspark

findspark.init("/шлях до spark")

from pyspark.sql import SparkSession

spark = SparkSession.builder \

.appName("JupyterSpark") \

.master("local[\*]") \

.config("spark.hadoop.fs.defaultFS", "hdfs://localhost:9000") \

.getOrCreate()

**spark-submit --master local[\*] code.py**