Big Data architektúrák és elemző módszerek Gyakorlat

Gombos Gergő

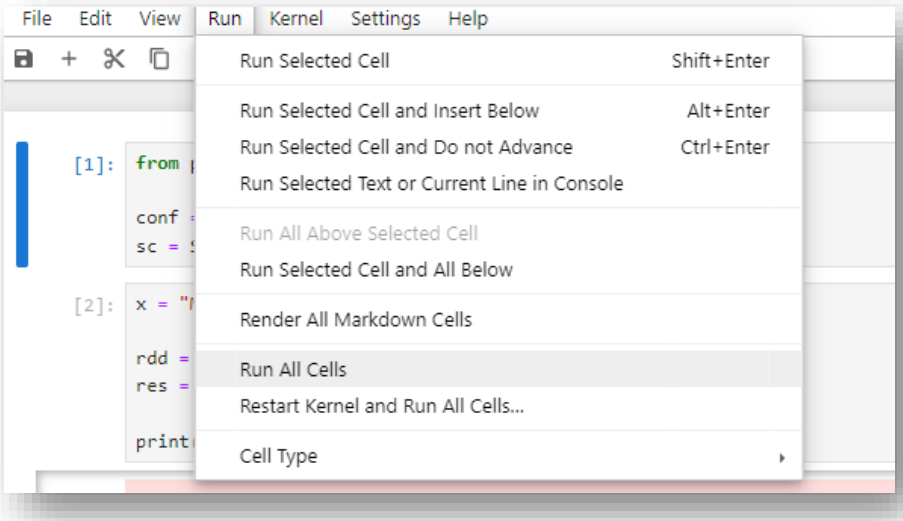
Spark telepítés emlékeztető

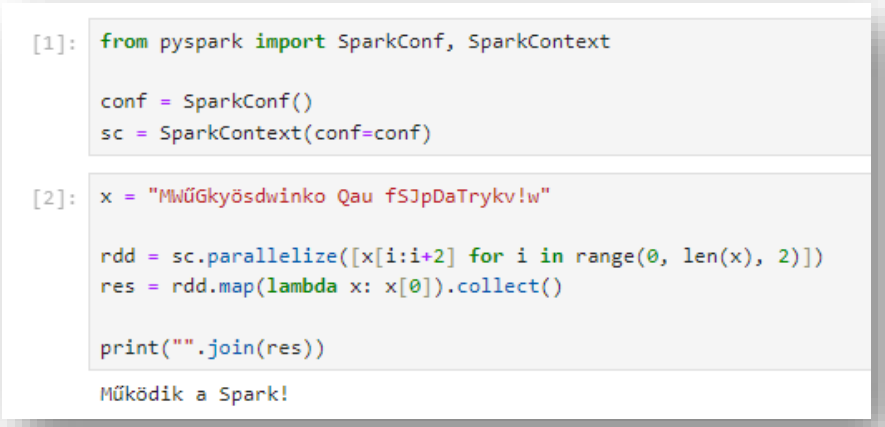
1. Nyiss egy parancsori ablakot (cmd)
2. Add ki a következő utasításokat
    1. pip install pyspark
    2. pip install jupyter
    3. pip install notebook

Környezet elindítása, kipróbálása (Spark)

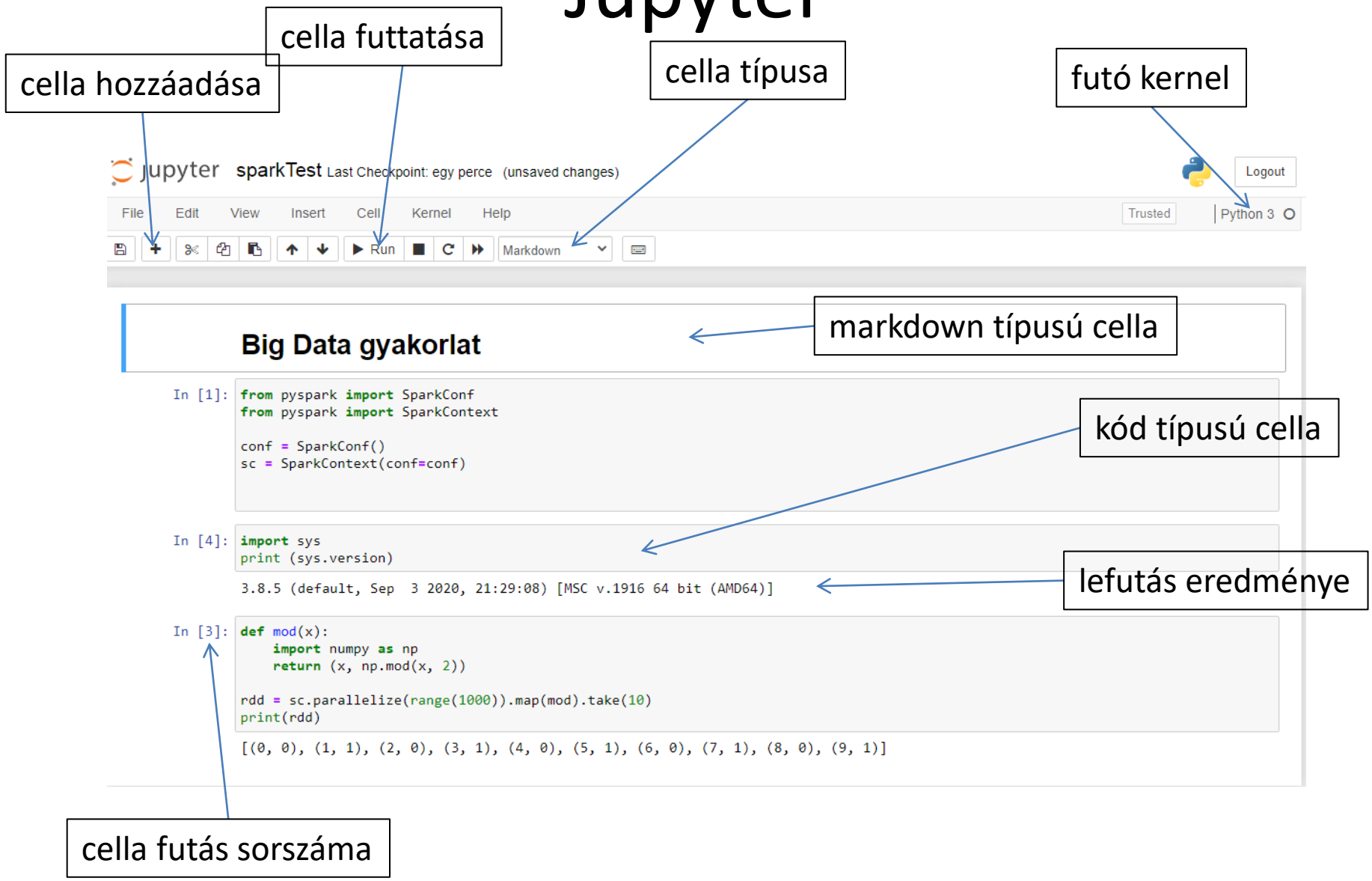
1. Indítsuk el a startJupyterSpark.bat
2. Indítsuk el a SparkTest.ipynb-t
3. Futtassuk le az összes mezőt

```
SET PYSPARK_PYTHON=python
SET JAVA_HOME=
python -m notebook
```





Jupyter


cella hozzáadása
cella futtatása
cella típusa
futó kernel
jupyter sparkTest Last Checkpoint: egy perce (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Markdown
Big Data gyakorlat
markdown típusú cella
In [1]:
from pyspark import SparkConf
from pyspark import SparkContext
conf = SparkConf()
sc = SparkContext(conf=conf)
kód típusú cella
In [4]:
import sys
print (sys.version)
3.8.5 (default, Sep 3 2020, 21:29:08) [MSC v.1916 64 bit (AMD64)]
lefutás eredménye
In [3]:
def mod(x):
import numpy as np
return (x, np.mod(x, 2))
rdd = sc.parallelize(range(1000)).map(mod).take(10)
print(rdd)
[(0, 0), (1, 1), (2, 0), (3, 1), (4, 0), (5, 1), (6, 0), (7, 1), (8, 0), (9, 1)]
cella futás sorszáma

Spark környezet létrehozása

- **!!! 1 sparkContext lehet 1 kernelben.**
- **Csak 1x-szer futtassuk!!!!!**

```
from pyspark import SparkConf
from pyspark import SparkContext

conf = SparkConf()
sc = SparkContext(conf=conf)
```

Feladatok

- RDD, partíciók, SPARK felület, DAG
- Action, Transformation műveletek

- Feladat 1 (parallelize)
  - írassuk ki a számok 2-es modulóját
  - szűrjük le azokra, amik 3-mal is oszthatóak

- Feladat 2 (WordCount)
  - map vs flatMap
  - szűrni azokra amik nem üresek
  - lecserélni a sorvégi '.', ',', '@', '#', stb.-t
  - reduceByKey
  - groupByKey + reduce
  - rendezni (sortBy, SortByKey)

- Feladat 3 (leghosszabb szó)
  - max()
  - reduce()

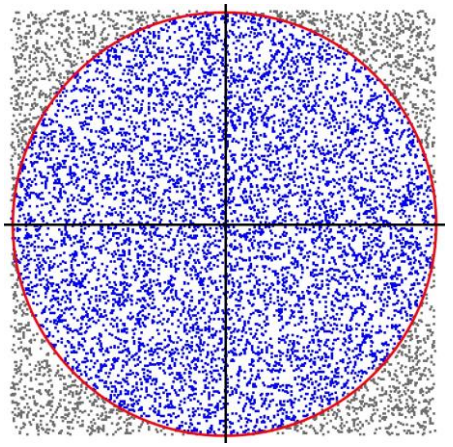
Feladatok

- Feladatok
  - Accumulator
  - Broadcast
  - PI számítás
    - Generáljunk random pontokat [0,1] intervallumon és nézzük meg milyen messze esik az origótól. Ha közelebb van, mint 1, akkor a körbe esik.

$$\frac{\text{kör területe}}{\text{négyzet területe}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$

$$\frac{\pi}{4} = \frac{\text{körbe eső pontok száma}}{\text{négyzetbe eső pontok száma}}$$

$$\pi = 4 \times \frac{\text{körbe eső pontok száma}}{\text{négyzetbe eső pontok száma}}$$



Feladatok

- Feladatok
  - Timing, lazy evaluation, cache
  - toDebugString()
  - RDD join
  - (K-means)

Köszönöm a Figyelmet!