ELTE | IK

INFORMATIKAI KAR

BigData architektúrák és elemző módszerek

PySpark telepítés egyetemi gépeken

- 1. lépés: pyspark és jupyterlab csomagok telepítése (parancssorból)

```
python -m pip install jupyterlab
python -m pip install pyspark
```

- 2. lépés: Szükséges fájlok letöltése:
  https://vargadaniel.web.elte.hu/bigdata24/SparkFiles.zip

- 3. lépés: Indítás – futtassuk a start_spark.bat fájlt parancssorból

```
.\start_spark.bat
```

- 4. lépés: Teszteljük le, hogy működik-e a környezet. Futtassuk a *spark_test_code.ipynb* fájlban lévő blokkokat. Ha megjelenik a „Működik a Spark!” üzenet, akkor minden rendben.

PySpark telepítés GNU/Linux rendszeren

- 1. lépés: Java telepítése

```
sudo apt install openjdk-8-jdk-headless
```

- 2. lépés: pyspark és jupyterlab csomagok telepítése

```
python -m pip install jupyterlab
python -m pip install pyspark
```

- 3. lépés: Teszteljük le, hogy működik-e a környezet. Futtassuk a *spark_test_code.ipynb* fájlban lévő blokkokat. Ha megjelenik a „Működik a Spark!” üzenet, akkor minden rendben. A fájl itt megtalálható: https://vargadaniel.web.elte.hu/bigdata24/SparkFiles.zip

- Megjegyzés: ha a „python” parancs helyett a „py” vagy a „python3” van használatban, akkor mindnehol hazsnáljuk azt.

Hibaelhárítás

- (1) Saját gépen lehetséges, hogy nincs fent Python és Java. Ez esetben ezeket kell legelőször telepíteni.
  - https://www.oracle.com/java/technologies/downloads
  - https://www.python.org/downloads/

- (2) Bizonyos hibákat megoldhat a findspark csomag feltelepítése…

```
python –m pip install findspark
```

- …és a notebook első blokkjában a következő sorok futtatása:

```
import findspark
findspark.init()
findspark.find()
```

Hibaelhárítás

- Bizonyos hibák esetén (pl. Java gateway process exited before sending its port nubmer, szóköz a felhasználónévben) érdemes lehet új Python virtual environmentet létrehozni
- Parancssorba/PowerShellbe (tetszőleges útvonal megadható):

```
python -m venv C:\spark-venv
```

- Létrehozás után aktiváljuk a venv-et a létrehozott mappában található activate.bat futtatásával:

```
.\spark-venv\Scripts\activate.bat
```

- A venv létrehozása után a venv-ben ugyanúgy fel kell telepíteni a pysaprkot és a jupyterlab-et és utána ugyanazzal a paranccsal indítható (lásd 2. dia)

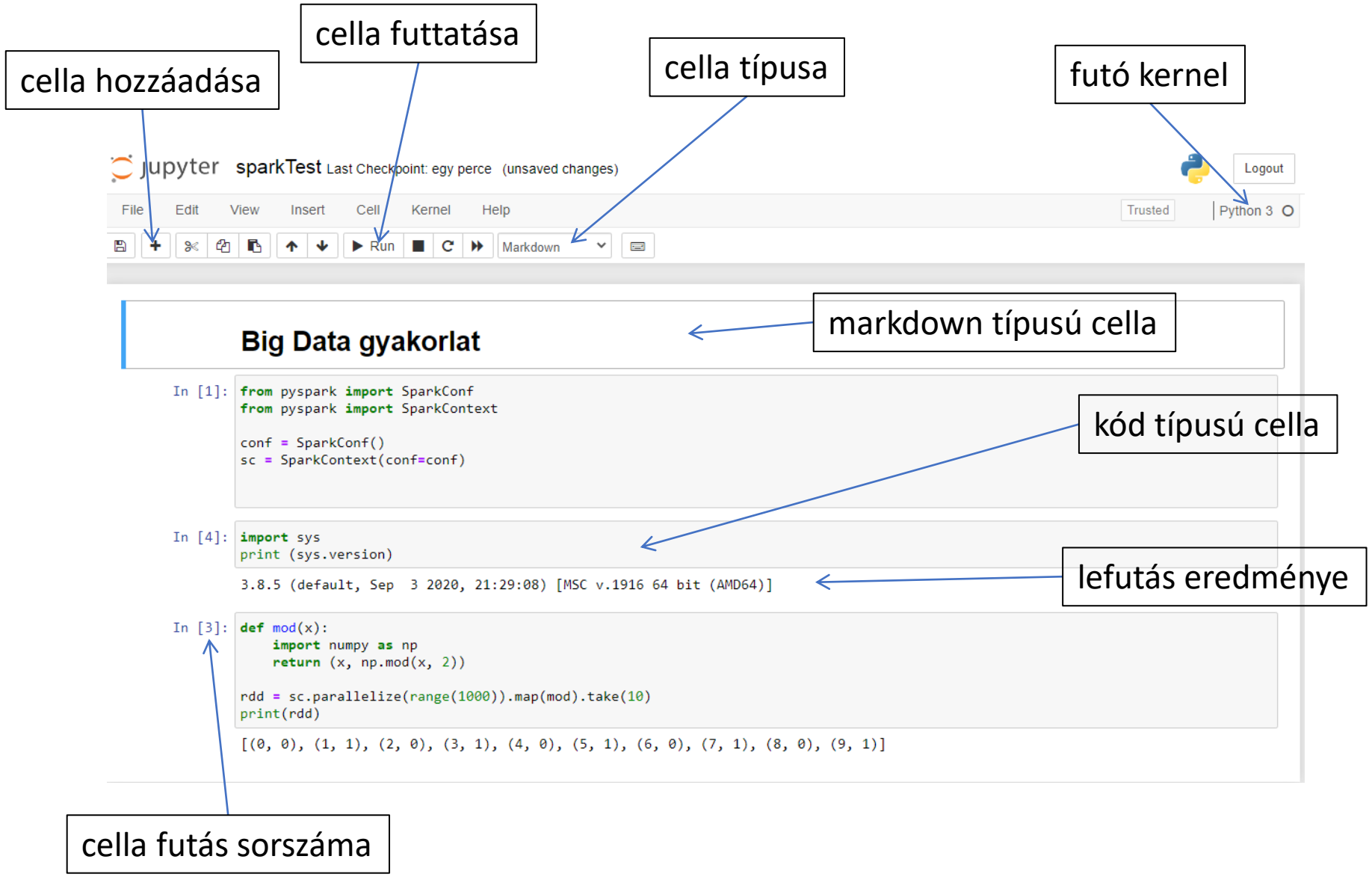
Alternatíva: Google Colab

- Ha nem sikerül megfelelelően beállítani a környezetet, de mindenképp fontos az otthoni munka, akkor a Google Colab használatával ingyen és könnyeddén lehet Spark kódokat írni és futtatni.
- https://colab.research.google.com/
- Első lépésben fel kell telepítenünk a pysparkot a következő utasítással (futtassuk az első blokkban):

```
!pip install pyspark
```

- Ezután már el is kezdhetünk PySparkban dolgozni (lásd mintakód):
- https://vargadaniel.web.elte.hu/bigdata24/SparkFiles.zip

Jupyter Notebook felhasználói felülete



Spark alkalmazás készítése

- Egy SparkContext lehet egy kernelben!
- **Mindenképp csak egyszer futtassuk!**

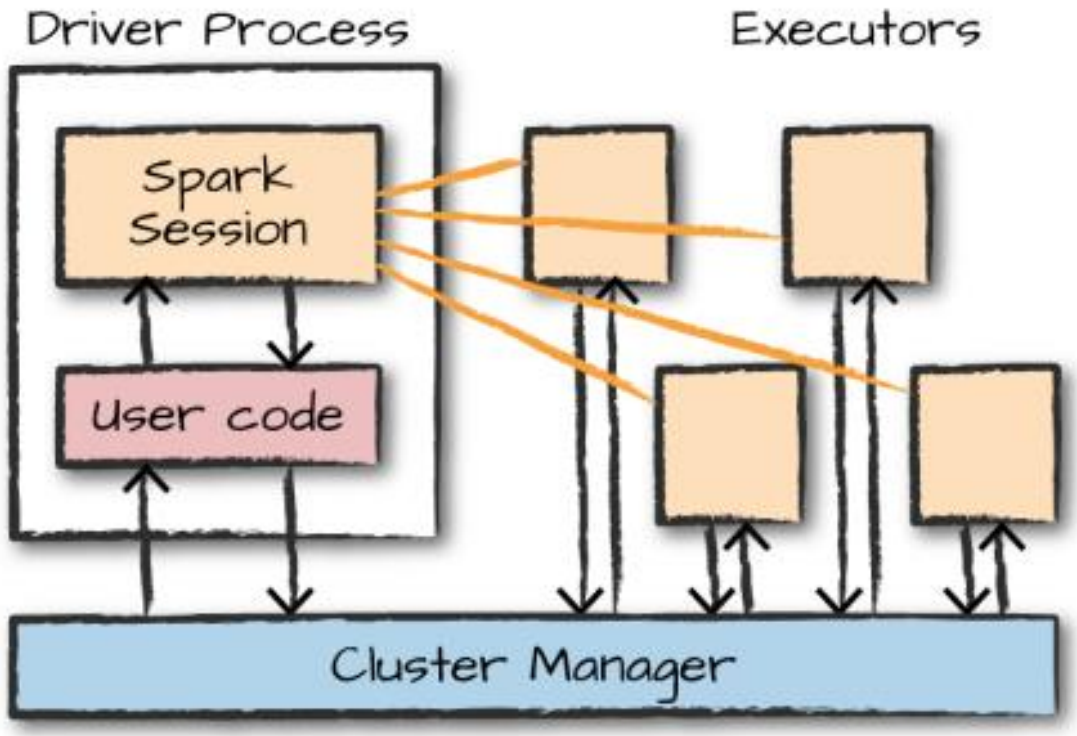
```
from pyspark import SparkConf
from pyspark import SparkContext

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

- A további kódokat új blokkokban futtassuk.
- A következő kóddal letesztelhető, hogy jól működik-e a PySpark:

```
In [2]: sc.parallelize(range(100)).count()
Out[2]: 100
```

Spark architektúra



Spark felépítése