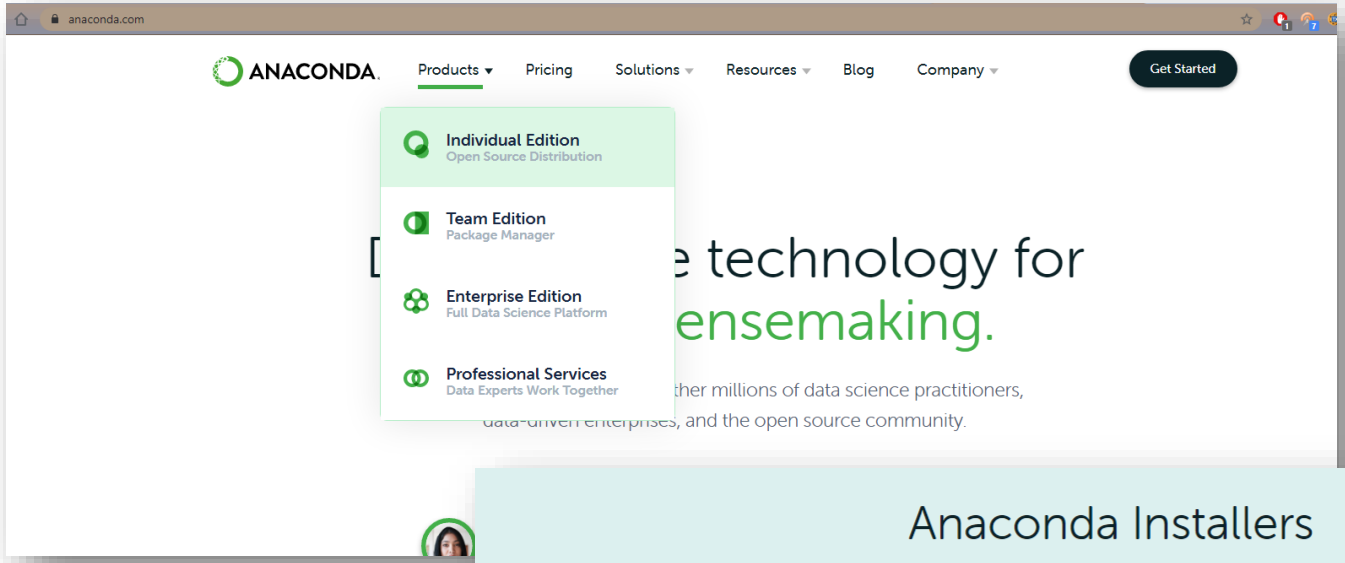
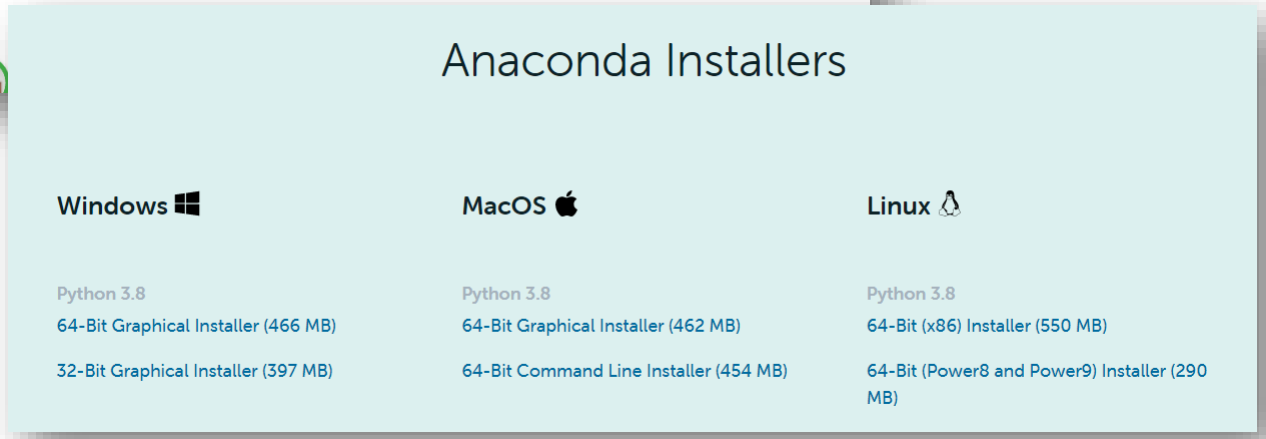
Big Data architektúrák és elemző módszerek Gyakorlat

Gombos Gergő

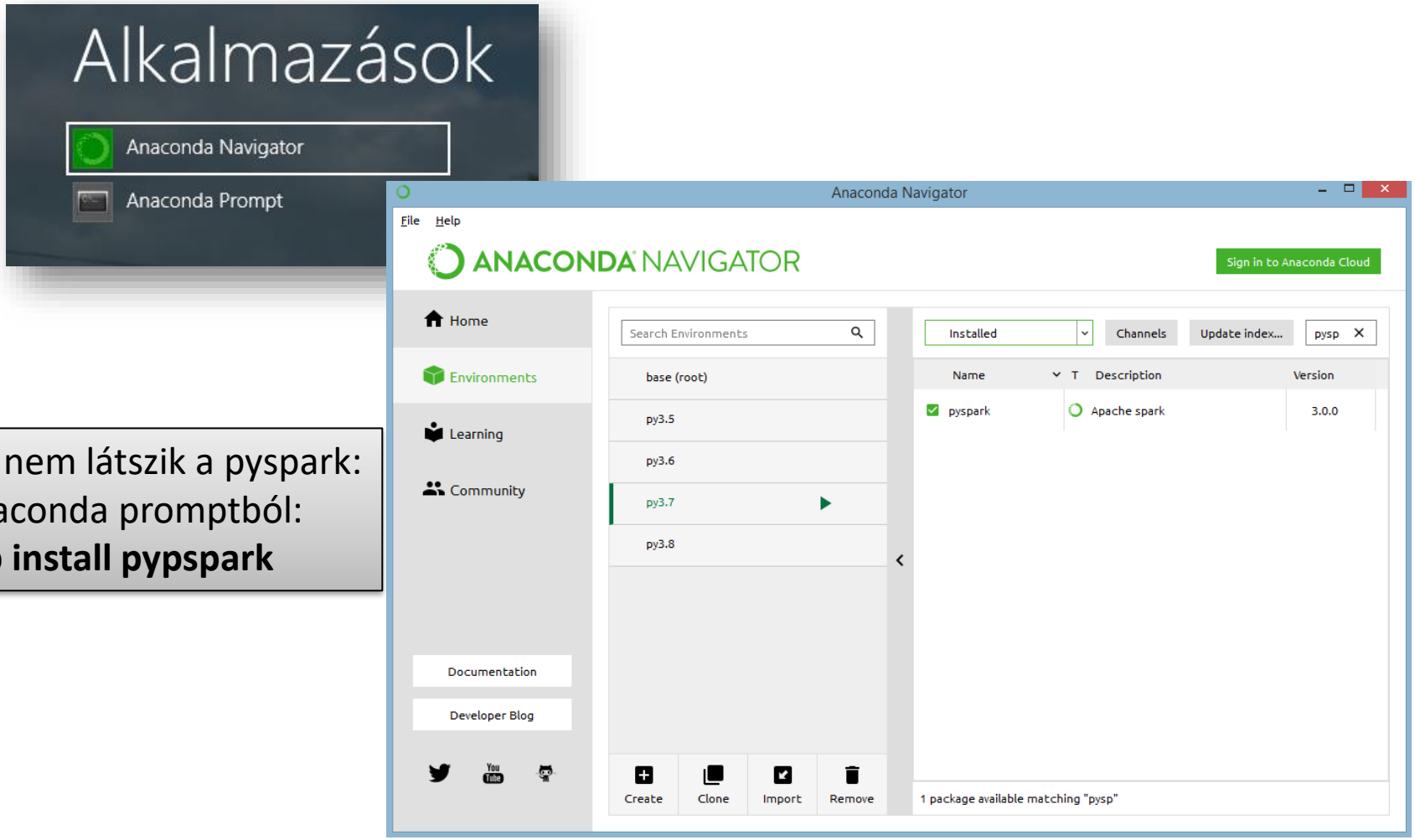
Spark telepítés

- Anaconda letöltése, telepítése



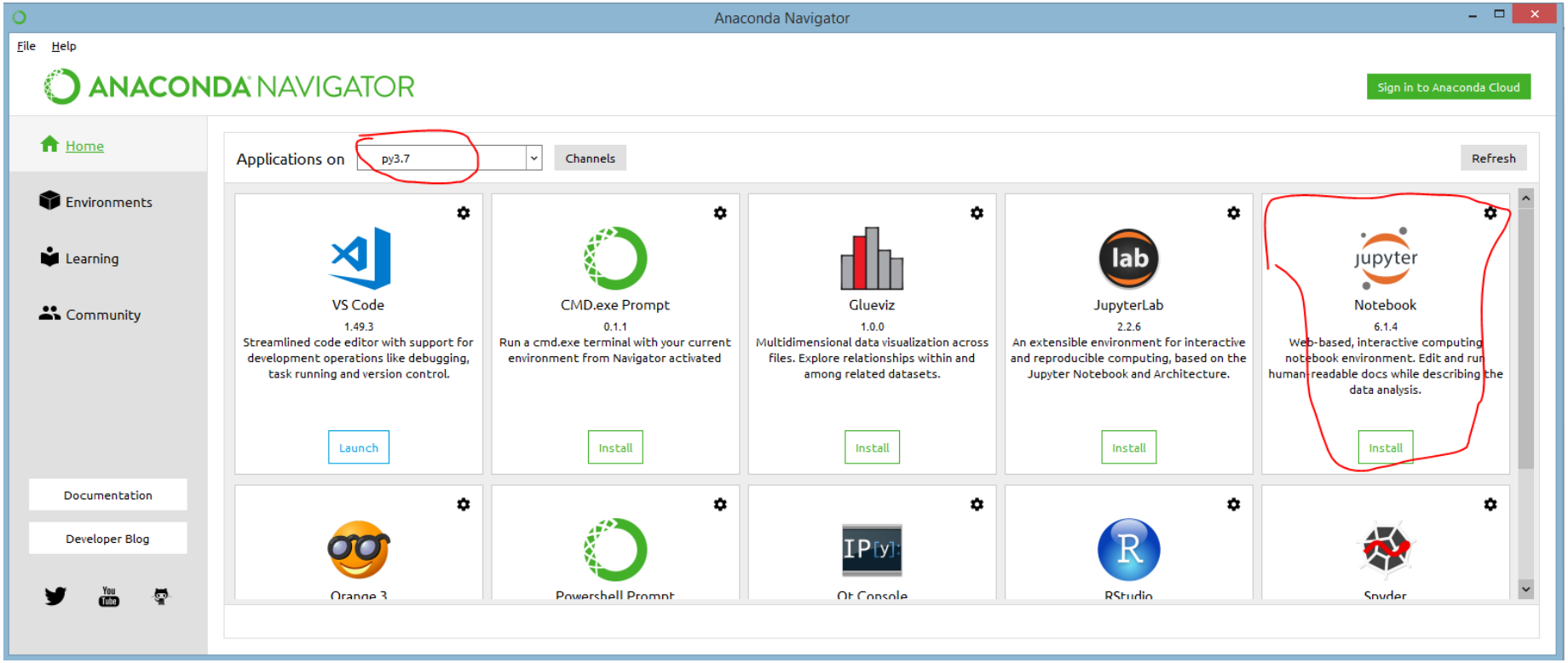


Spark telepítés



Ha nem látszik a pyspark:
anaconda promptból:
pip install pypspark

Spark telepítés



Spark telepítése másik megoldás

- Cmd-ből:

```
py -m pip install pyspark
py -m pip install notebook
```

- PATH-be adjuk meg:

```
c:\Users\userneved\AppData\Roaming\Python\Python39\Scripts
```

- Környezeti változók közé vegyük fel:

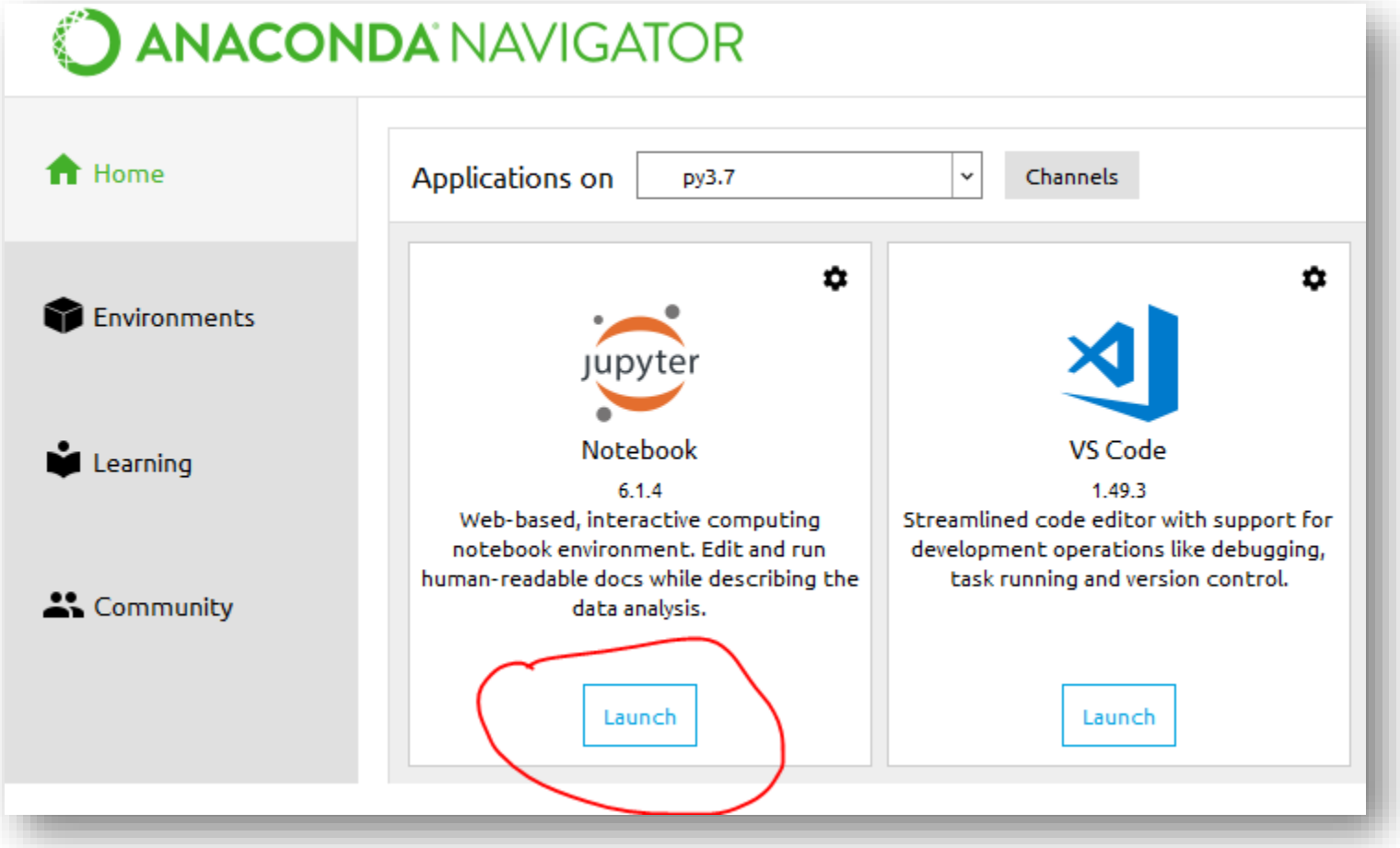
```
PYSPARK_DRIVER_PYTHON=jupyter
PYSPARK_DRIVER_PYTHON_OPTS=notebook
PYSPARK_PYTHON=py     (a python3 parancs amit használni tudsz)

JAVA_HOME=c:\BigData\java    (valamilyen jre)
```

- Indítás:

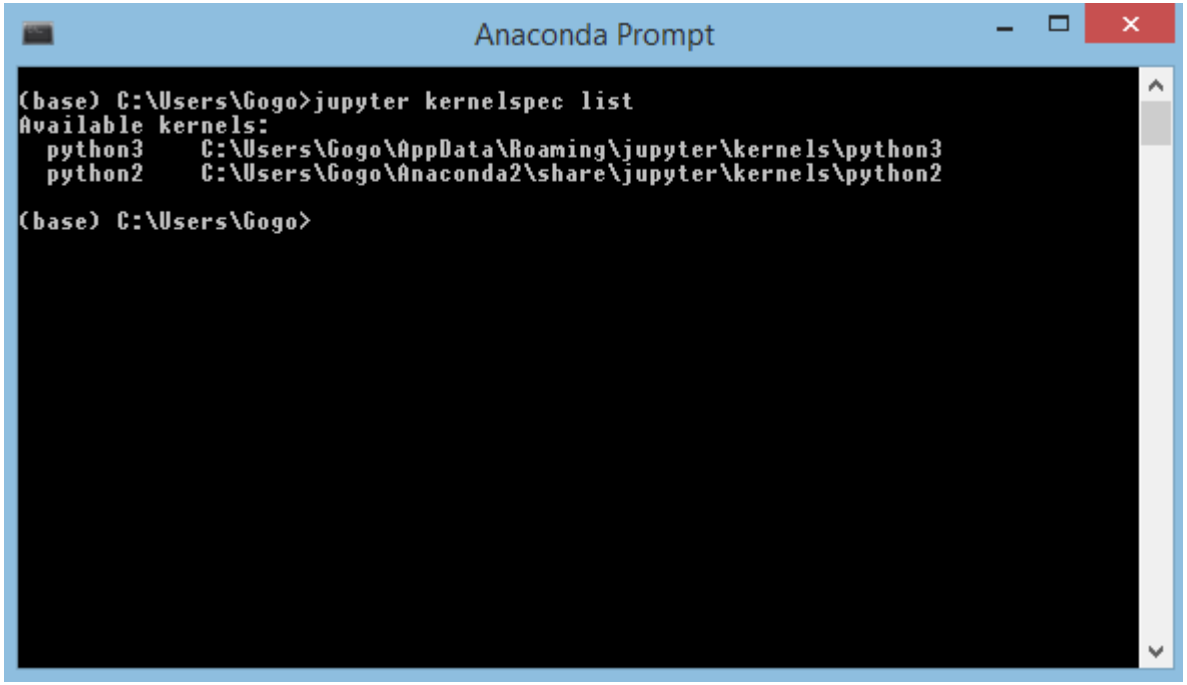
```
jupyter-notebook
```

Jupyter indítása

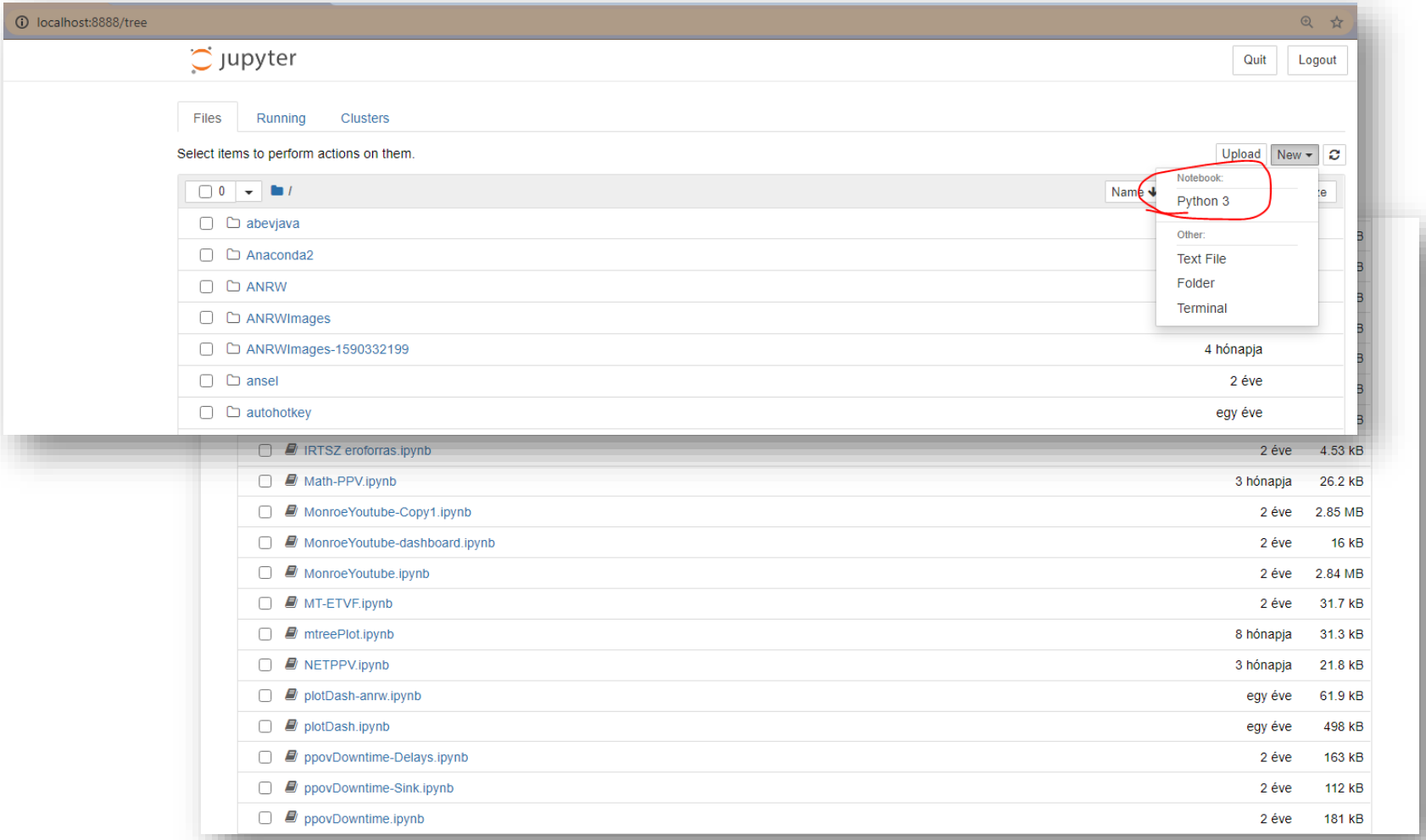


Jupyter kernel ellenőrzése

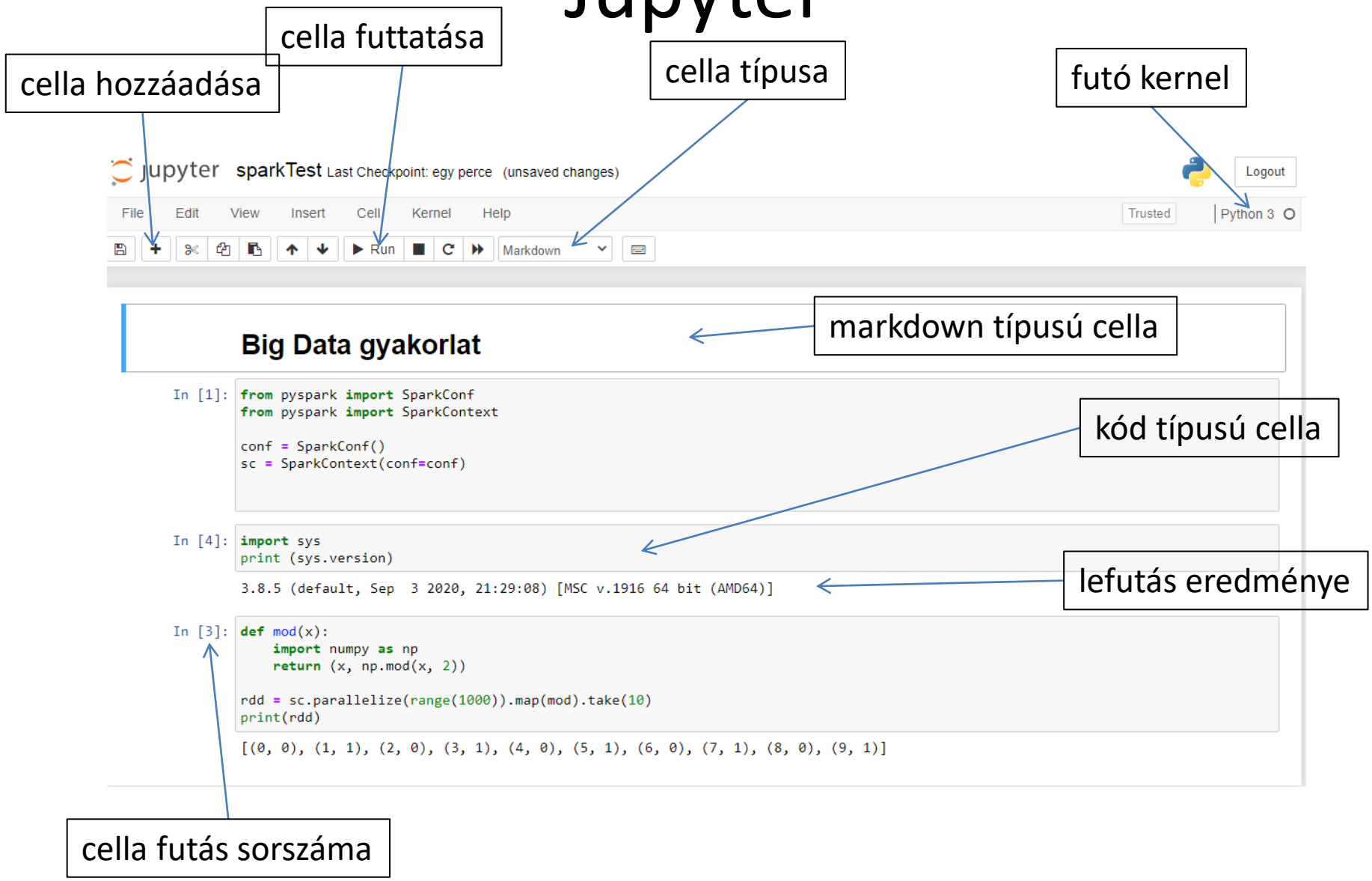
- Ha valami hiba van le lehet ellenőrizni a kernelt!



Jupyter


localhost:8888/tree
jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Notebook:
Python 3
Other:
Text File
Folder
Terminal
Name
abevjava
Anaconda2
ANRW
ANRWImages
ANRWImages-1590332199 4 hónapja
ansel 2 éve
autohotkey egy éve
IRTSZ eroforras.ipynb 2 éve 4.53 kB
Math-PPV.ipynb 3 hónapja 26.2 kB
MonroeYoutube-Copy1.ipynb 2 éve 2.85 MB
MonroeYoutube-dashboard.ipynb 2 éve 16 kB
MonroeYoutube.ipynb 2 éve 2.84 MB
MT-ETVF.ipynb 2 éve 31.7 kB
mtreePlot.ipynb 8 hónapja 31.3 kB
NETPPV.ipynb 3 hónapja 21.8 kB
plotDash-anrw.ipynb egy éve 61.9 kB
plotDash.ipynb egy éve 498 kB
ppovDowntime-Delays.ipynb 2 éve 163 kB
ppovDowntime-Sink.ipynb 2 éve 112 kB
ppovDowntime.ipynb 2 éve 181 kB

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

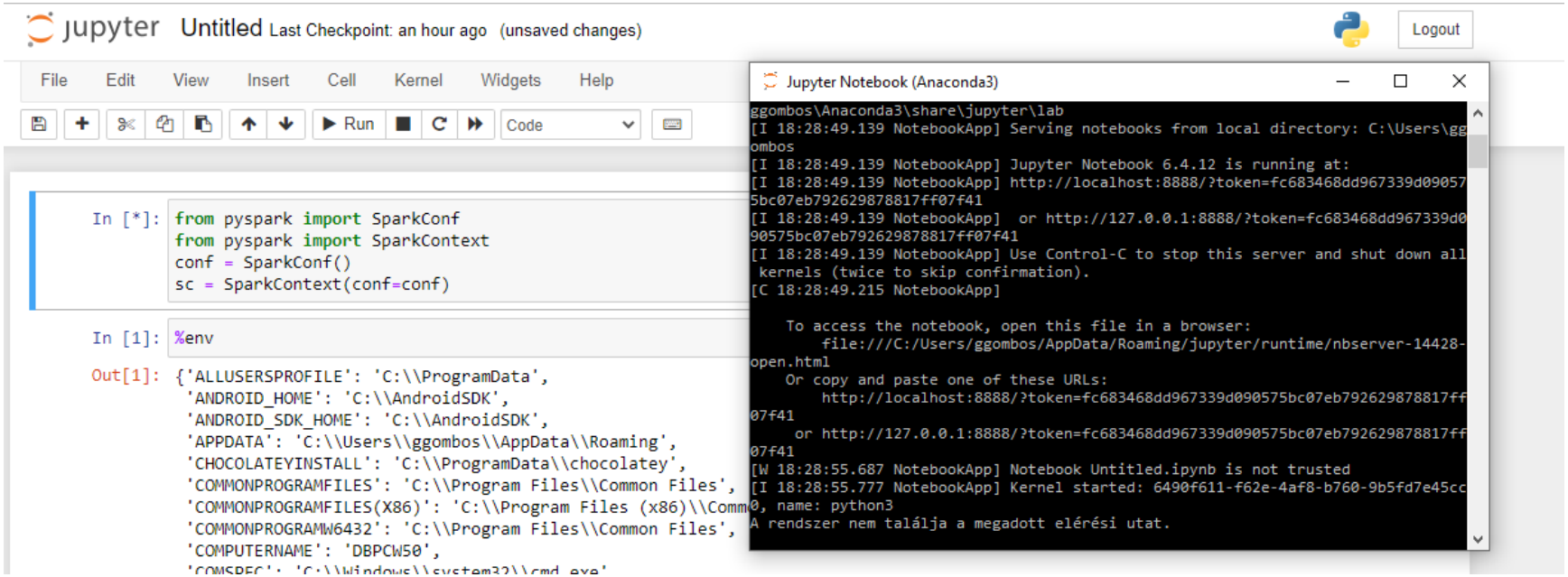
Spark környezet létrehozása

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

```
from pyspark import SparkConf
from pyspark import SparkContext

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

Jupyter + Spark hibák!



Jupyter + Spark hibák!

- Hibaüzenet:
  - „Java gateway process exited before sending its port number”
- Megoldás 1:

```
from pyspark import SparkConf
from pyspark import SparkContext
import findspark
findspark.init()
findspark.find()

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

Jupyter + Spark hibák!

- Hibaüzenet:
  - „Java gateway process exited before sending its port number”
- Megoldás 2:
  - Parancssorba / powershellbe:

```
py -m venv c:\spark-env
```

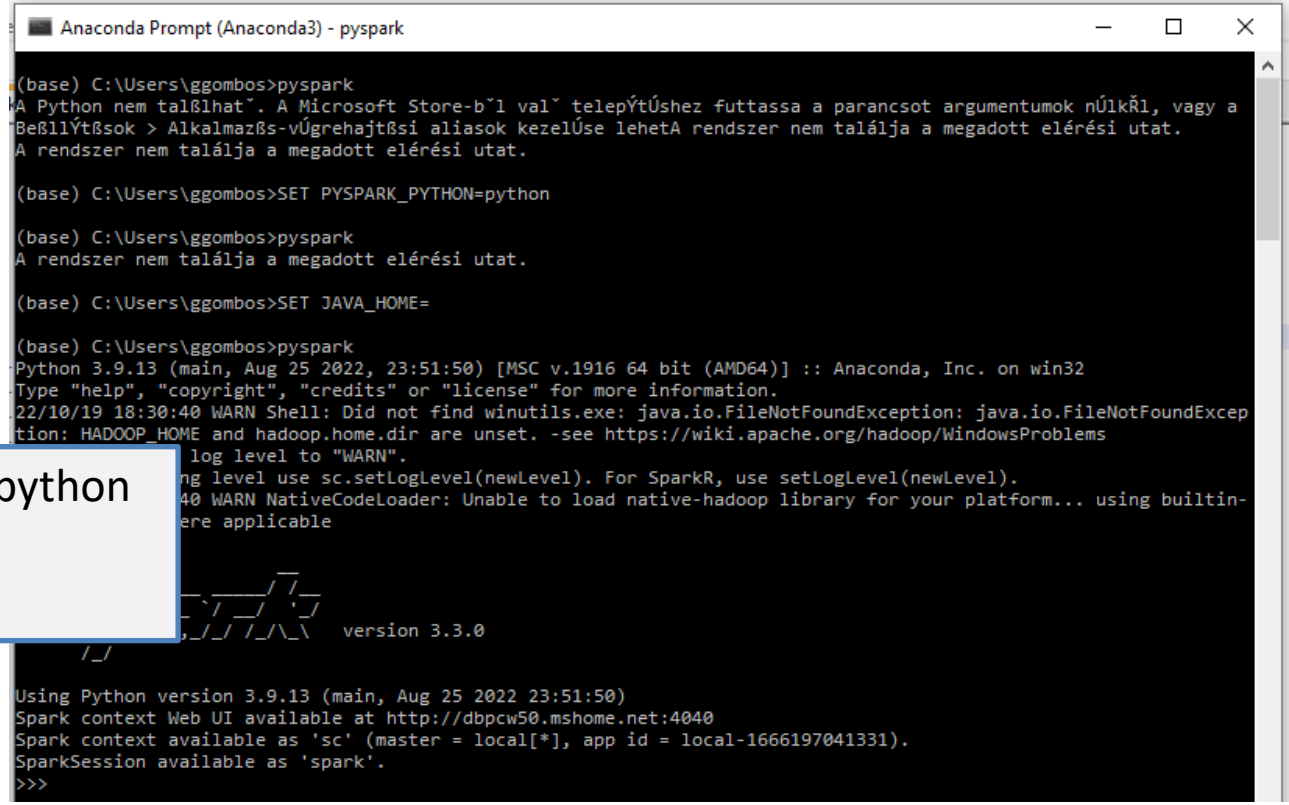
```
.\spark-env\Scripts\activate.bat
```

  - Pip-pel telepíteni a notebookot és a pysparkot

Jupyter + Spark hibák!

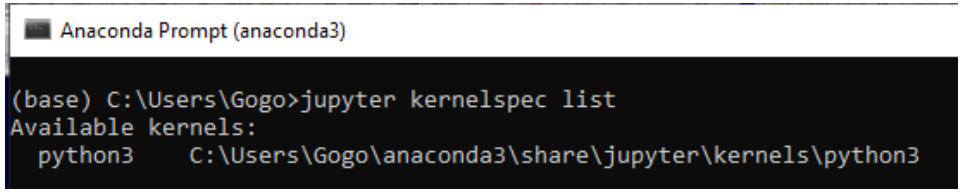
- Hibaüzenet:
  - „Java gateway process exited before sending its port number”
- Megoldás 3:
  - anaconda prompt

SET PYSPARK_PYTHON=python
SET JAVA_HOME=
jupyter-notebook



Jupyter + Spark hibák!

- Hibaüzenet:
  - „Java gateway process exited before sending its port number”
- Megoldás 4:
  - Anaconda promptban:
    - Jupyter kernelspec list



    - A kaporr elérési úton a kernel.json-t szerkeszteni
    - A pirossal írt sort hozzávenni

```
{
 "argv": [
  "python",
  "-m",
  "ipykernel_launcher",
  "-f",
  "{connection_file}"
 ],
 "env": {"PYSPARK_PYTHON":"python","JAVA_HOME":""},
 "display_name": "Python 3 (ipykernel)",
 "language": "python",
 "metadata": {
  "debugger": true
 }
}
```

Feladatok

- 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()

Köszönöm a Figyelmet!