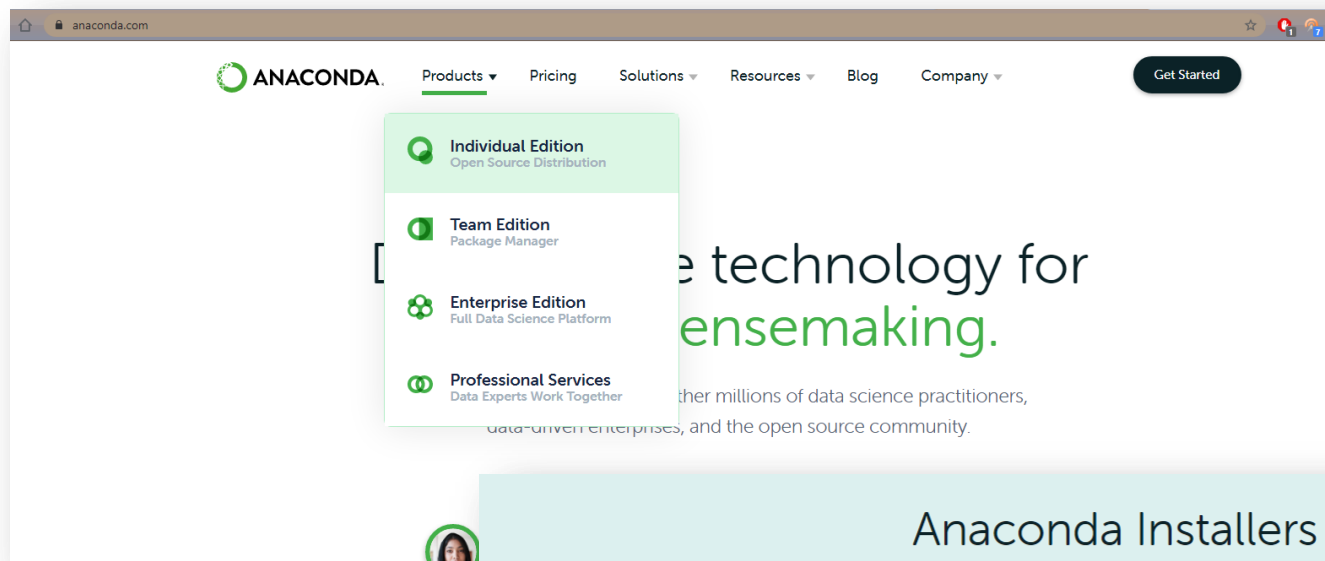


Big Data architektúrák és elemző módszerek Gyakorlat


Gombos Gergő

Spark telepítés

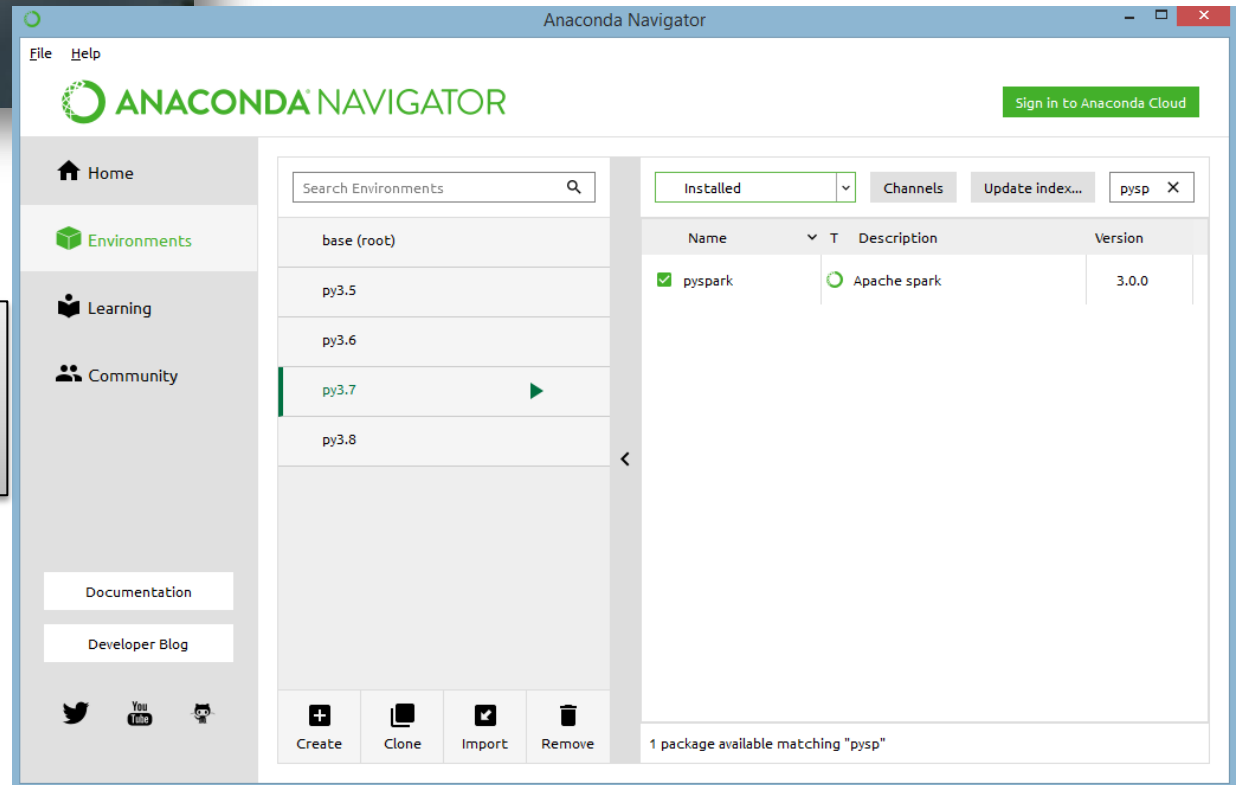
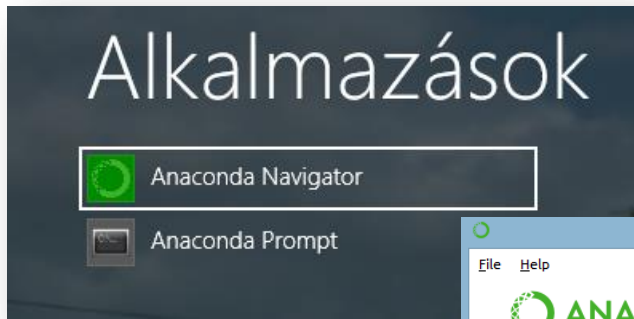
- Anaconda letöltése, telepítése



Anaconda Installers

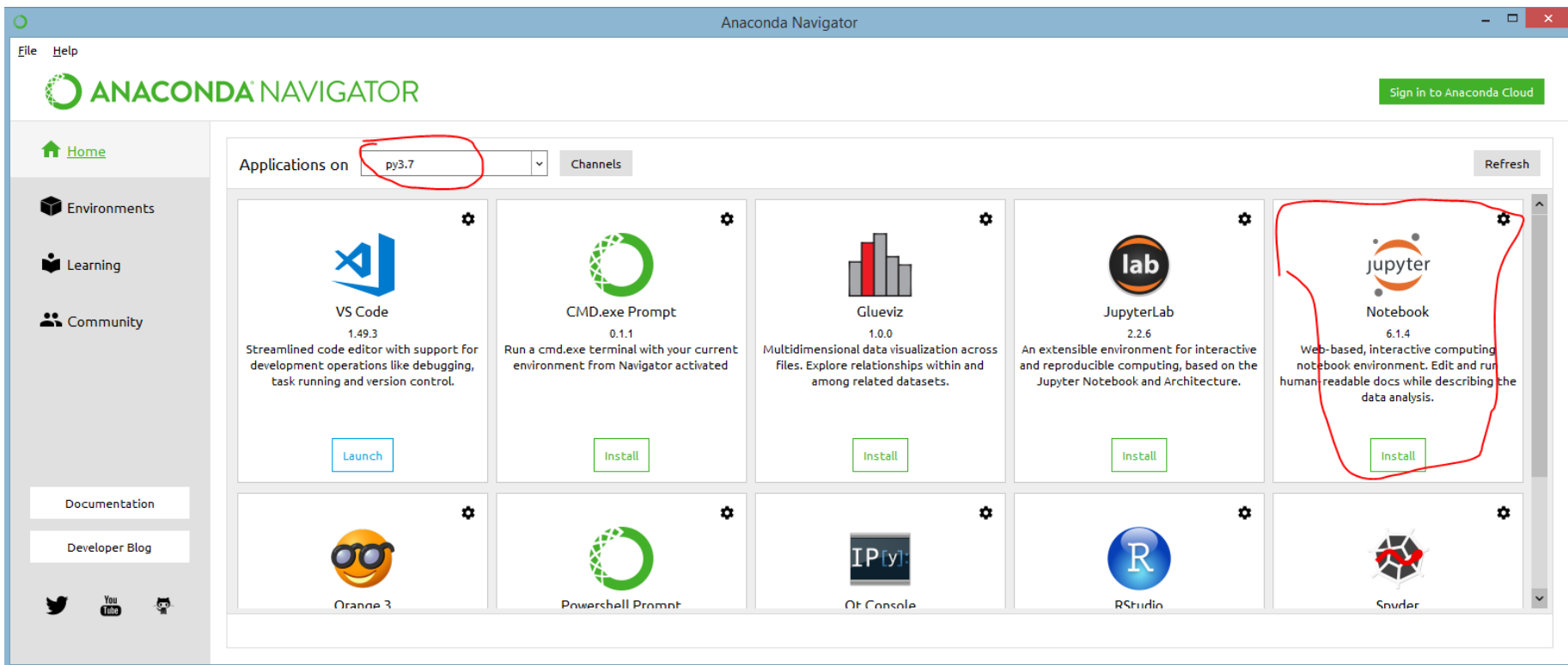
Windows 	MacOS 	Linux 
Python 3.8 64-Bit Graphical Installer (466 MB) 32-Bit Graphical Installer (397 MB)	Python 3.8 64-Bit Graphical Installer (462 MB) 64-Bit Command Line Installer (454 MB)	Python 3.8 64-Bit (x86) Installer (550 MB) 64-Bit (Power8 and Power9) Installer (290 MB)

Spark telepítés



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

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
```

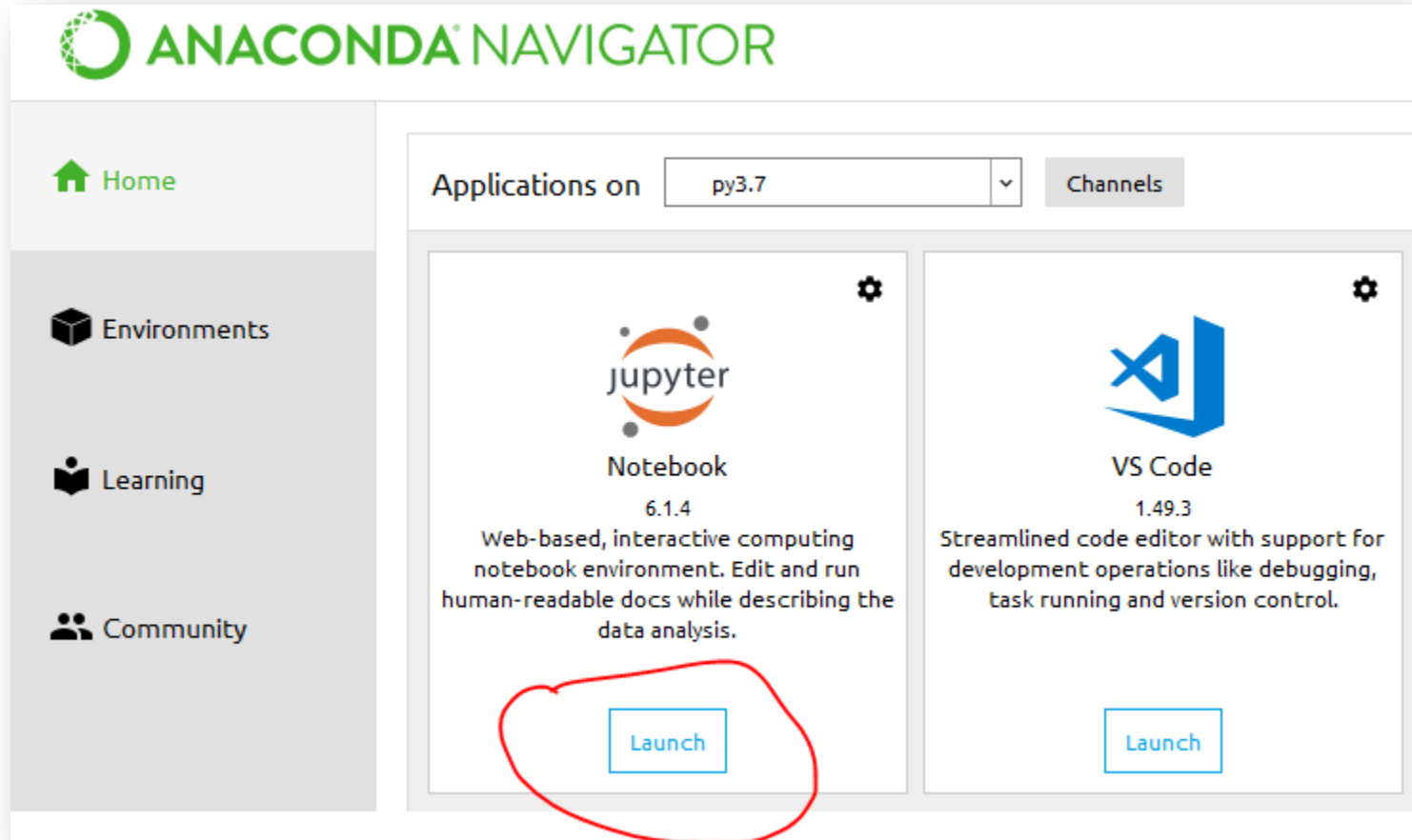
- Környezeti változók közé vegyük fel:

```
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:

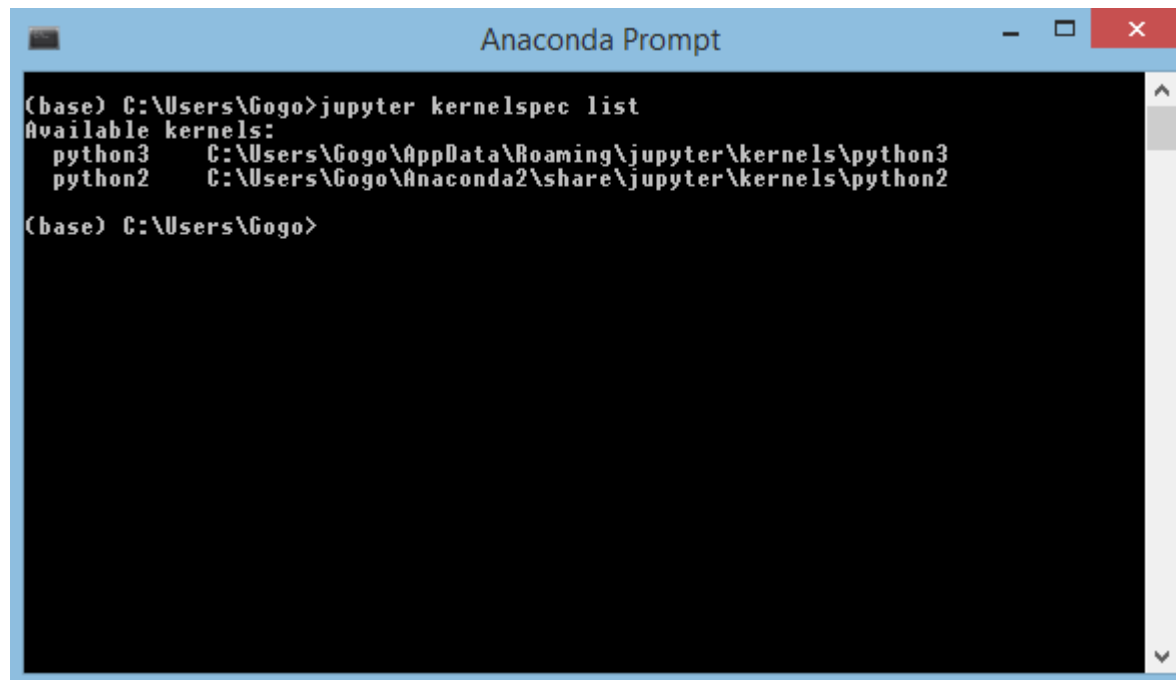
```
jupyter-notebook
```

Jupyter indítása



Jupyter kernel ellenőrzése

- Ha valami hiba van le lehet ellenőrizni a kernelt!



```
Anaconda Prompt
(base) C:\Users\Gogo>jupyter kernelspec list
Available kernels:
  python3    C:\Users\Gogo\AppData\Roaming\jupyter\kernels\python3
  python2    C:\Users\Gogo\Anaconda2\share\jupyter\kernels\python2
(base) C:\Users\Gogo>
```

Jupyter

The screenshot shows the Jupyter web interface in a browser window. The address bar displays 'localhost:8888/tree'. The page title is 'jupyter'. There are 'Quit' and 'Logout' buttons in the top right. Below the title, there are tabs for 'Files', 'Running', and 'Clusters'. A message says 'Select items to perform actions on them.' Below this is a file browser showing a directory structure with folders like 'abevjava', 'Anaconda2', 'ANRW', etc. A 'New' button is highlighted, and its dropdown menu is open, showing options: 'Notebook: Python 3', 'Other:', 'Text File', 'Folder', and 'Terminal'. The 'Notebook: Python 3' option is circled in red. Below the file browser, there is a list of files with their names, sizes, and creation dates.

Name	Size	Created
abevjava		
Anaconda2		
ANRW		
ANRWimages		
ANRWimages-1590332199		
ansel		
autohotkey		
IRTSZ_eroforras.ipynb	4.53 kB	2 éve
Math-PPV.ipynb	26.2 kB	3 hónapja
MonroeYoutube-Copy1.ipynb	2.85 MB	2 éve
MonroeYoutube-dashboard.ipynb	16 kB	2 éve
MonroeYoutube.ipynb	2.84 MB	2 éve
MT-ETVF.ipynb	31.7 kB	2 éve
mtreePlot.ipynb	31.3 kB	8 hónapja
NETPPV.ipynb	21.8 kB	3 hónapja
plotDash-anrw.ipynb	61.9 kB	egy éve
plotDash.ipynb	498 kB	egy éve
ppvDowntime-Delays.ipynb	163 kB	2 éve
ppvDowntime-Sink.ipynb	112 kB	2 éve
ppvDowntime.ipynb	181 kB	2 éve

Jupyter

cella futtatása

cella hozzáadása

cella típusa

futó kernel



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

- 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!