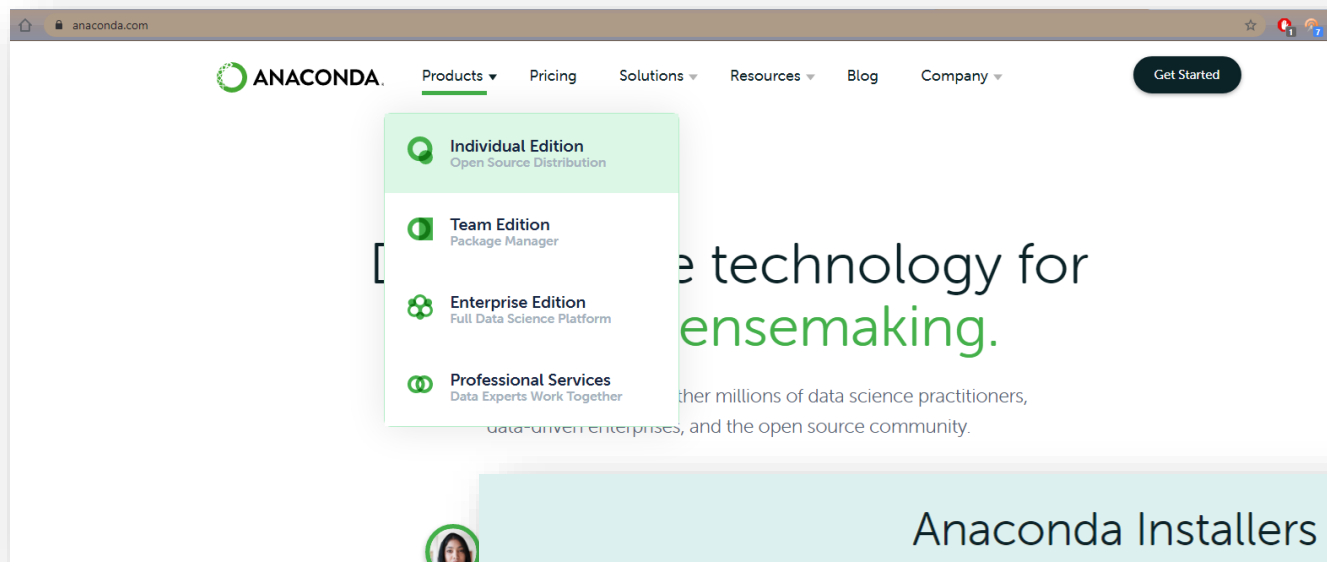


# Big Data architektúrák és elemző módszerek Gyakorlat

Gombos Gergő

# Spark telepítés

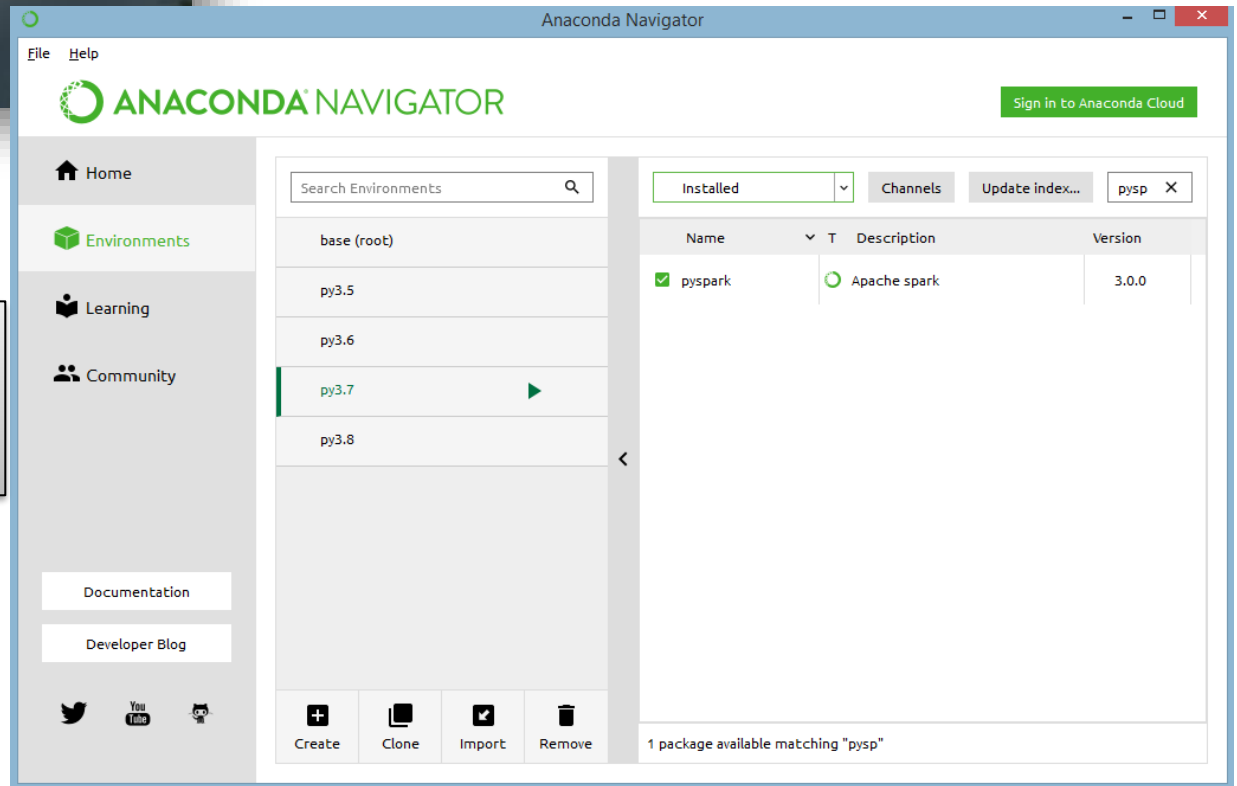
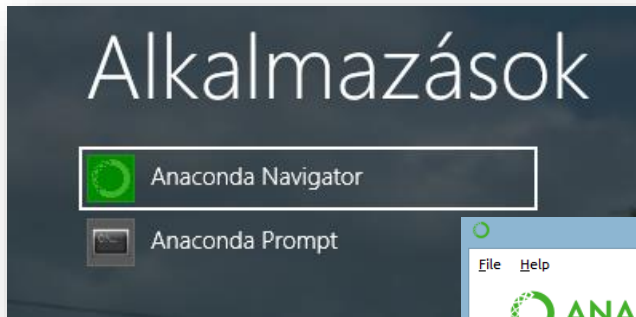
- Anaconda letöltése, telepítése



## 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 pypspark**

# Spark telepítés

The screenshot displays the Anaconda Navigator application window. The title bar reads "Anaconda Navigator". The main interface is divided into a left sidebar and a central content area. The sidebar contains navigation options: Home, Environments, Learning, and Community. The central area shows a grid of applications available for installation on the selected environment, "py3.7".

Applications on **py3.7** Channels Refresh

Application	Version	Description	Action
VS Code	1.49.3	Streamlined code editor with support for development operations like debugging, task running and version control.	Launch
CMD.exe Prompt	0.1.1	Run a cmd.exe terminal with your current environment from Navigator activated	Install
Glueviz	1.0.0	Multidimensional data visualization across files. Explore relationships within and among related datasets.	Install
JupyterLab	2.2.6	An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.	Install
Jupyter Notebook	6.1.4	Web-based, interactive computing notebook environment. Edit and run human readable docs while describing the data analysis.	Install
Orange 3			
Powershell Prompt			
Qt Console			
RStudio			
Snuder			

The Jupyter Notebook application card is highlighted with a red border. The "py3.7" environment name in the top left of the application grid is also circled in red.

# 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
```

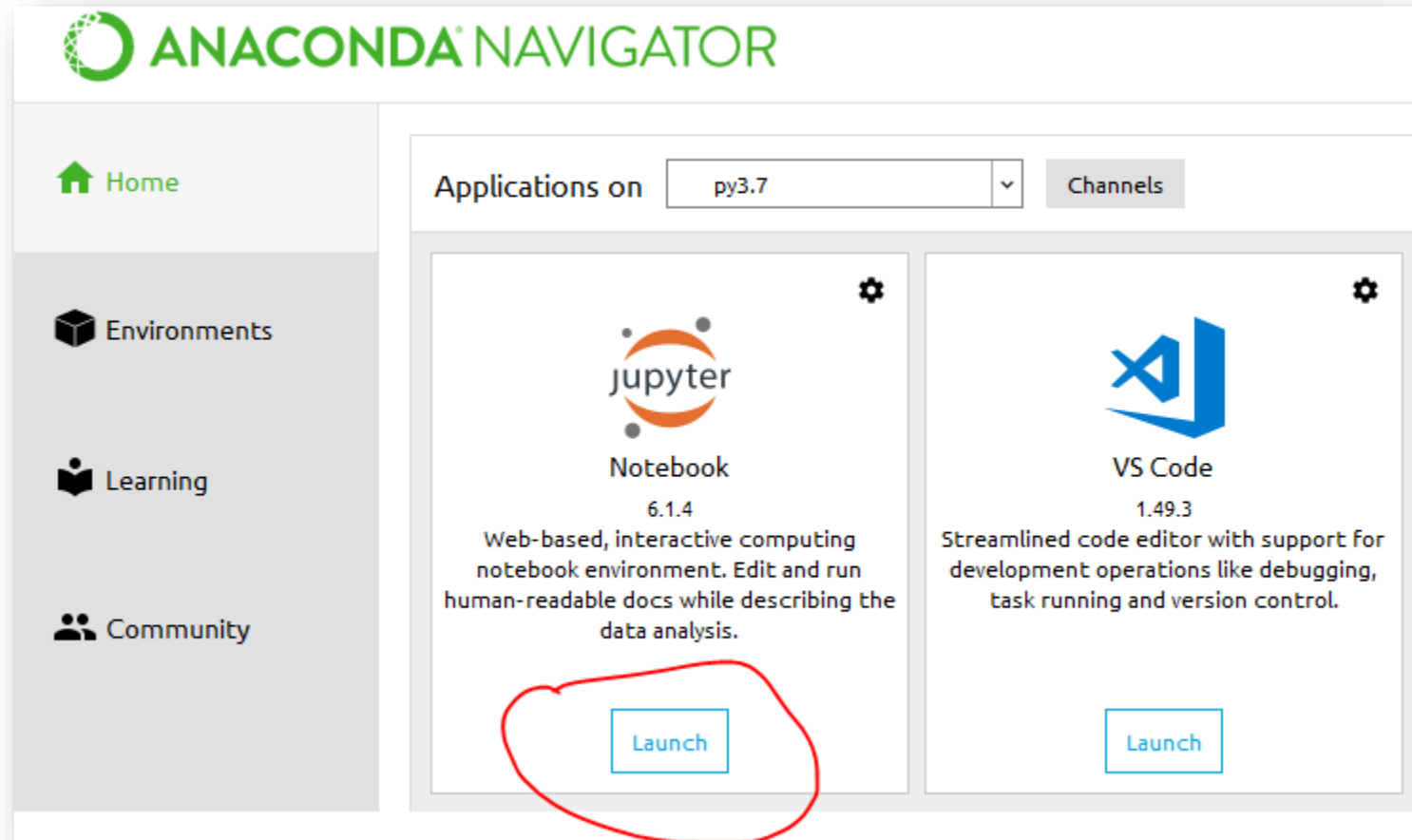
- Környezeti változók közé vegyük fel:

```
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:

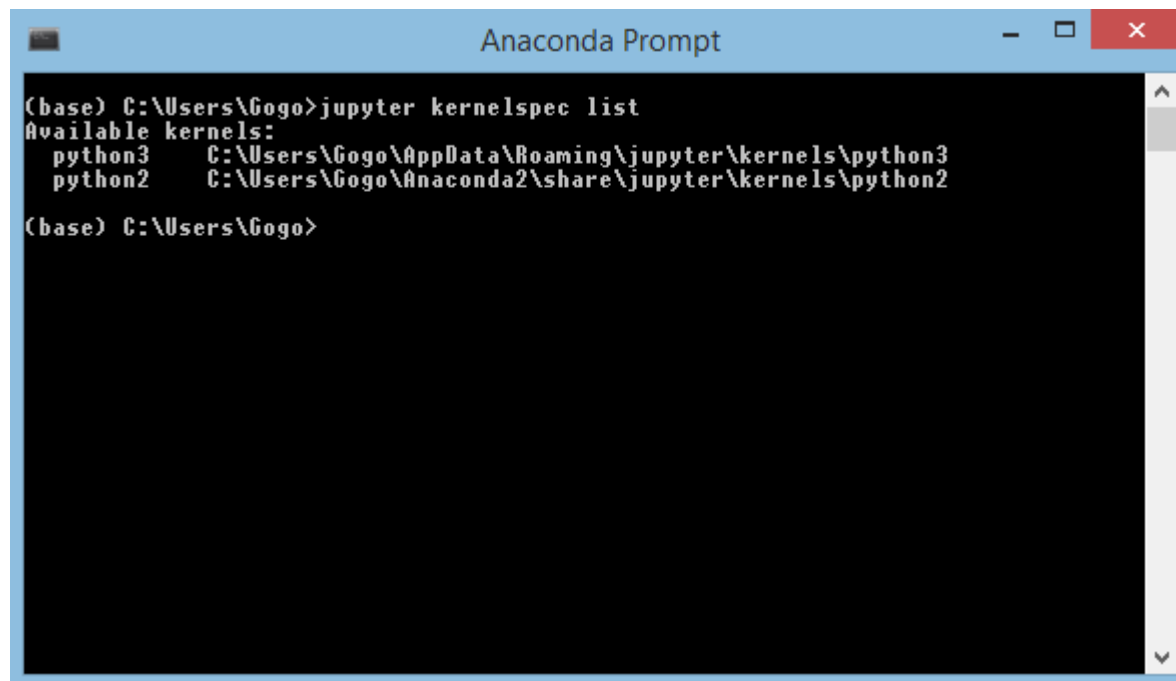
```
jupyter-notebook
```

# 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 at localhost:8888/tree. The interface includes a top navigation bar with 'Quit' and 'Logout' buttons, and a main content area with tabs for 'Files', 'Running', and 'Clusters'. Below the tabs, there is a prompt 'Select items to perform actions on them.' and a file browser view. The file browser shows a list of folders and files. A 'New' dropdown menu is open, showing options for 'Notebook: Python 3', 'Other: Text File', 'Folder', and 'Terminal'. The 'Python 3' option is highlighted with a red circle. 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

Jupyter sparkTest Last Checkpoint: egy perce (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted Python 3

Logout

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)
```

lefutás eredménye

```
3.8.5 (default, Sep 3 2020, 21:29:08) [MSC v.1916 64 bit (AMD64)]
```

cella futás sorszáma

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)]
```

# 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!

The image shows a Jupyter Notebook interface with a code cell and a terminal window. The code cell contains the following Python code:

```
In [*]: from pyspark import SparkConf
        from pyspark import SparkContext
        conf = SparkConf()
        sc = SparkContext(conf=conf)
```

The terminal window shows the following output:

```
ggombos\Anaconda3\share\jupyter\lab
[I 18:28:49.139 NotebookApp] Serving notebooks from local directory: C:\Users\ggombos
[I 18:28:49.139 NotebookApp] Jupyter Notebook 6.4.12 is running at:
[I 18:28:49.139 NotebookApp] http://localhost:8888/?token=fc683468dd967339d090575bc07eb792629878817ff07f41
[I 18:28:49.139 NotebookApp] or http://127.0.0.1:8888/?token=fc683468dd967339d090575bc07eb792629878817ff07f41
[I 18:28:49.139 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 18:28:49.215 NotebookApp]

To access the notebook, open this file in a browser:
file:///C:/Users/ggombos/AppData/Roaming/jupyter/runtime/nbserver-14428-open.html
Or copy and paste one of these URLs:
http://localhost:8888/?token=fc683468dd967339d090575bc07eb792629878817ff07f41
or http://127.0.0.1:8888/?token=fc683468dd967339d090575bc07eb792629878817ff07f41
[W 18:28:55.687 NotebookApp] Notebook Untitled.ipynb is not trusted
[I 18:28:55.777 NotebookApp] Kernel started: 6490f611-f62e-4af8-b760-9b5fd7e45cc0, name: python3
A rendszer nem találja a megadott elérési utat.
```

# 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

```
Anaconda Prompt (Anaconda3) - pyspark
(base) C:\Users\ggombos>pyspark
A Python nem található. A Microsoft Store-ből való telepítéshez futtassa a parancsot argumentumok nélkül, vagy a
Beállítások > Alkalmazás-végrehajtási aliasok kezelése lehetA rendszer nem találja a megadott elérési utat.
A rendszer nem találja a megadott elérési utat.

(base) C:\Users\ggombos>SET PYSPARK_PYTHON=python

(base) C:\Users\ggombos>pyspark
A rendszer nem találja a megadott elérési utat.

(base) C:\Users\ggombos>SET JAVA_HOME=

(base) C:\Users\ggombos>pyspark
Python 3.9.13 (main, Aug 25 2022, 23:51:50) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
22/10/19 18:30:40 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems
log level to "WARN".
log level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
ere applicable

version 3.3.0

Using Python version 3.9.13 (main, Aug 25 2022 23:51:50)
Spark context Web UI available at http://dbpcw50.mshome.net:4040
Spark context available as 'sc' (master = local[*], app id = local-1666197041331).
SparkSession available as 'spark'.
>>>
```

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

```
Anaconda Prompt (anaconda3)
(base) C:\Users\Gogo>jupyter kernelspec list
Available kernels:
python3      C:\Users\Gogo\anaconda3\share\jupyter\kernels\python3
```

```
{
  "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!