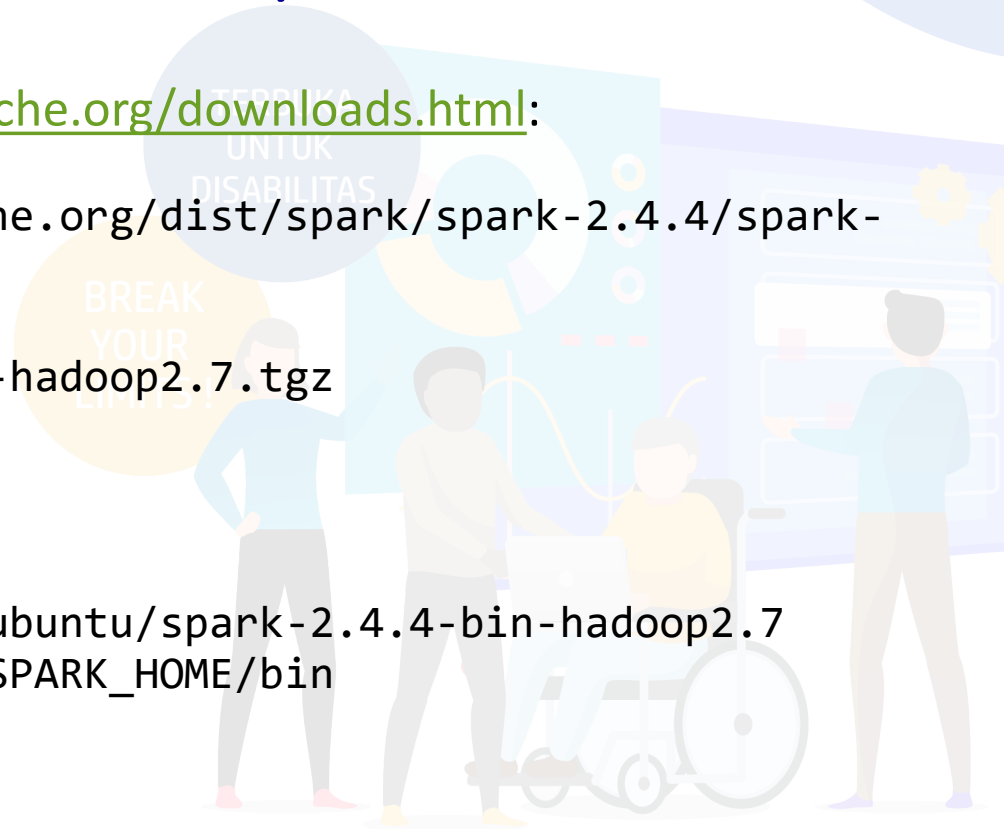


Instalasi Spark+Jupyter di EC2



Instalasi Spark (Minimum)

- *Download* dari <https://spark.apache.org/downloads.html>:
 - Contoh:
 - `wget https://www-eu.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz`
- *Extract*:
 - `tar -xvf spark-2.4.4-bin-hadoop2.7.tgz`
- Ubah `.bashrc`:
 - `nano ~/.bashrc`
- Tambahkan baris:
 - `export SPARK_HOME=/home/ubuntu/spark-2.4.4-bin-hadoop2.7`
 - `PATH=$PATH:$SPARK_HOME:$SPARK_HOME/bin`
- *Reload* `.bashrc`:
 - `source ~/.bashrc`
- *Test* Spark shell atau pyspark:
 - `spark-shell`
- atau
 - `pyspark`



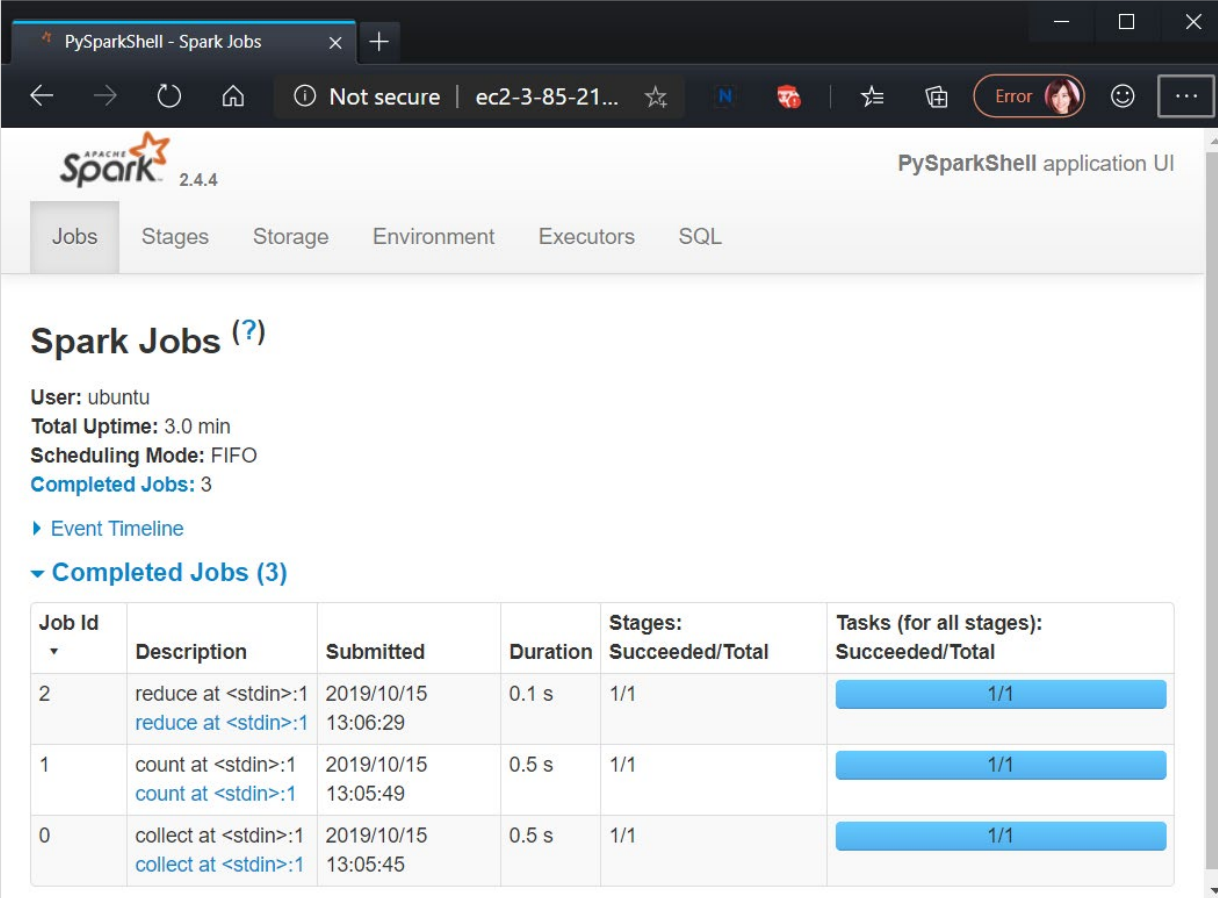
PySpark Shell

- Contoh PySpark shell →
- Untuk membuat RDD salah satunya bisa menggunakan fungsi `parallelize()` dari `SparkContext`

```
ubuntu@ip-172-31-46-241: ~  
se setLogLevel(newLevel).  
Welcome to  
  
      /_/_/ _/_/ _/_/ _/_/  
     / \ V _\ V _\ / \ V _\  
    / _ / . _ \ / \ , / _ / \ / \ \  
   / _ /      / _ / \ / \ \  
  / _ /  
 version 2.4.4  
  
Using Python version 3.7.3 (default, Mar 27 2019 22:11:17)  
SparkSession available as 'spark'.  
>>> x = [1,2,3]  
>>> rdd = sc.parallelize(x)  
>>> rdd.collect()  
[1, 2, 3]  
>>> rdd.count()  
3  
>>> rdd2 = rdd.map(lambda x: 2*x)  
>>> rdd3 = rdd2.reduce(lambda x,y: x+y)  
>>> rdd3  
12
```

Spark Web UI

- Apabila Spark Shell atau PySpark berhasil dijalankan, Spark web UI dapat diakses melalui:
 - `http://hostname-ec2:4040`
 - Misalnya
 - `http://ec2-3-85-216-134.compute-1.amazonaws.com:4040`
- Catatan:
 - Port 4040 adalah default
 - Port 4040 di EC2 (inbound) harus dibuka supaya bisa diakses dari luar



The screenshot shows a web browser window with the title "PySparkShell - Spark Jobs". The address bar shows "Not secure | ec2-3-85-21...". The page header includes the "Spark 2.4.4" logo and the text "PySparkShell application UI". Below the header is a navigation bar with tabs: "Jobs", "Stages", "Storage", "Environment", "Executors", and "SQL". The "Jobs" tab is selected.

Spark Jobs (?)

User: ubuntu
Total Uptime: 3.0 min
Scheduling Mode: FIFO
Completed Jobs: 3

[Event Timeline](#)

▼ Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	reduce at <stdin>:1 reduce at <stdin>:1	2019/10/15 13:06:29	0.1 s	1/1	1/1
1	count at <stdin>:1 count at <stdin>:1	2019/10/15 13:05:49	0.5 s	1/1	1/1
0	collect at <stdin>:1 collect at <stdin>:1	2019/10/15 13:05:45	0.5 s	1/1	1/1

Jupyter+PySpark di AWS EC2

- Kebutuhan:
 - Anaconda (dengan Jupyter)
 - Spark
 - Instalasi module findspark
 - `pip install findspark`

TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR
LIMITS!



SSH Tunneling di Putty ke EC2

- *SSH Tunneling* atau *SSH Port Forwarding* digunakan untuk *mem-forward port* dari *remote server* ke *local computer* melalui koneksi SSH atau sebaliknya
- Dalam kasus ini *port* Jupyter yang digunakan di EC2 akan di-*forward* melalui *port* di komputer lokal
- Untuk mengetahui port mana yang akan digunakan Jupyter, jalankan sekali lalu cek URL dari Jupyter
<http://localhost:8888/?token=xxx>

```

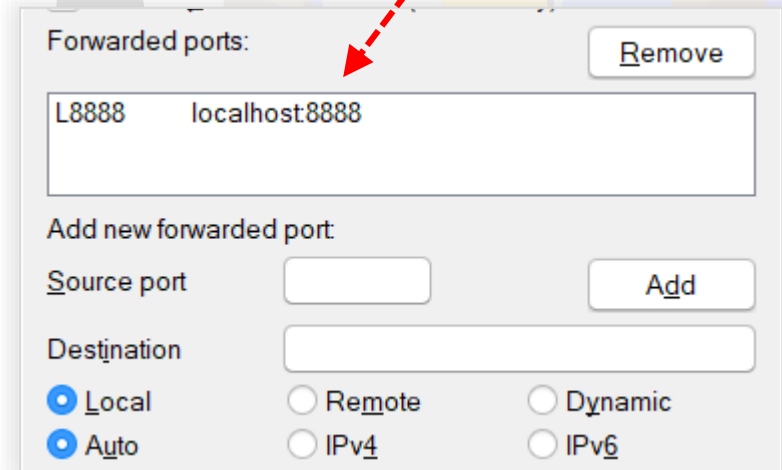
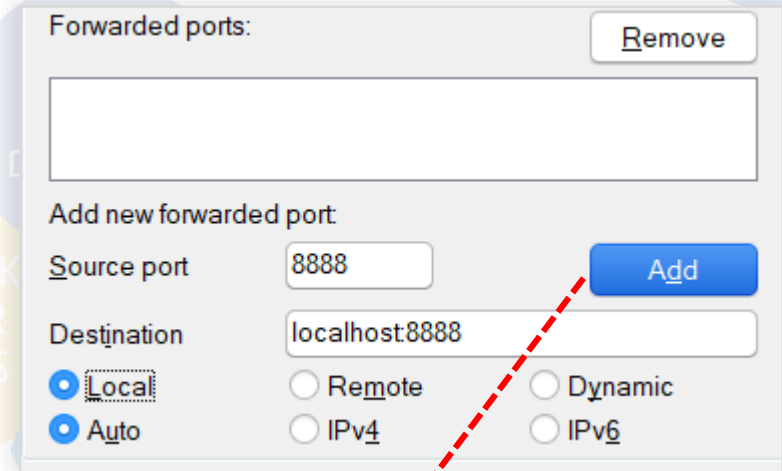
ubuntu@ip-172-31-46-241: ~
conda3/lib/python3.7/site-packages/jupyterlab
[I 13:26:42.950 NotebookApp] JupyterLab application directory is /home/ubuntu/anaconda3/share/jupyter/lab
[I 13:26:42.952 NotebookApp] Serving notebooks from local directory: /home/ubuntu
[I 13:26:42.953 NotebookApp] The Jupyter Notebook is running at:
[I 13:26:42.953 NotebookApp] http://localhost:8888/?token=3bb28cbbd336c50306b9d89f081ad3339d6e6e1311b47794
[I 13:26:42.953 NotebookApp] or http://127.0.0.1:8888/?token=3bb28cbbd336c50306b9d89f081ad3339d6e6e1311b47794
[I 13:26:42.953 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 13:26:42.957 NotebookApp] No web browser found: could not locate runnable browser.
[C 13:26:42.957 NotebookApp]

To access the notebook, open this file in a browser:
    file:///home/ubuntu/.local/share/jupyter/runtime/nbserver-4483-open.html
Or copy and paste one of these URLs:
    http://localhost:8888/?token=3bb28cbbd336c50306b9d89f081ad3339d6e6e1311b47794
    or http://127.0.0.1:8888/?token=3bb28cbbd336c50306b9d89f081ad3339d6e6e1311b47794
  
```

Misalnya dalam contoh ini yang digunakan adalah port 8888

SSH Tunneling di Putty ke EC2

- Jalankan Putty
- Isikan di:
- Session
 - Host Name: *ec2.... <nama hostname ec2-...>*
 - Port: **22**
- Connection
 - Auth
 - Gunakan fail .pem hasil unduh untuk EC2 ybs.
 - Tunnels
 - Source: **8888** (port di komputer local, usahakan sama dengan destination)
 - Destination: **localhost:8888** (Jupyter di EC2)
 - klik tombol **Add**
- Login dengan user: ubuntu
- Jalankan:
 - jupyter notebook
- atau
 - jupyter lab



```

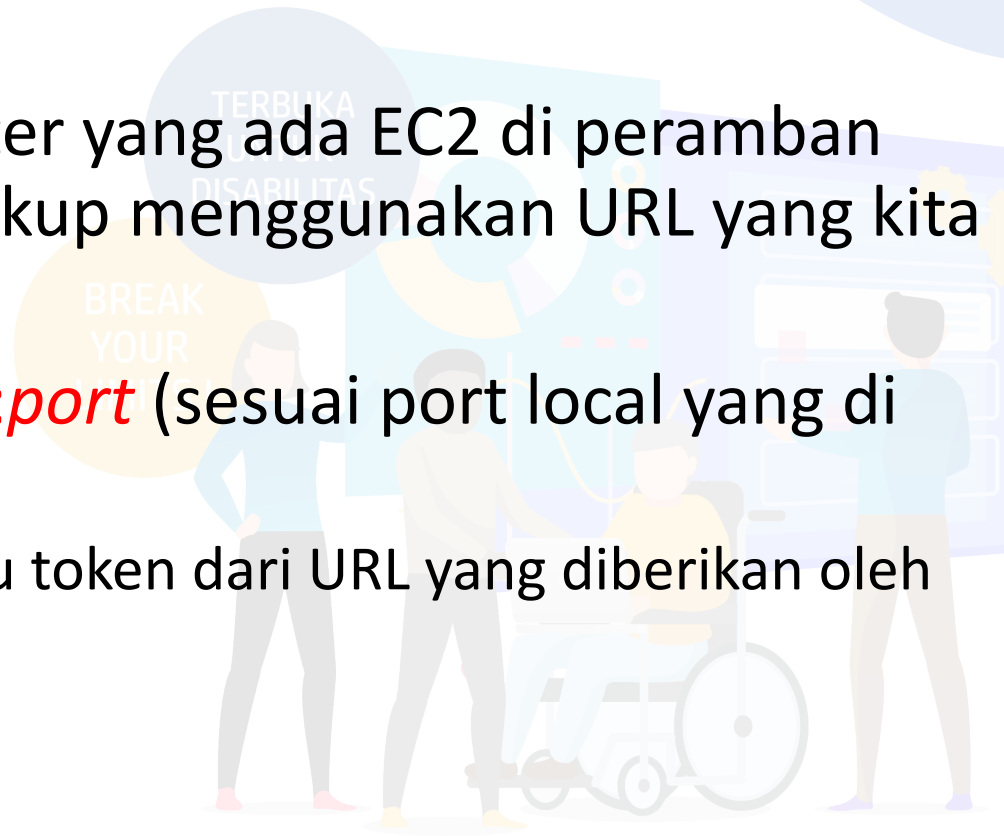
ubuntu@ip-172-31-46-241: ~
conda3/lib/python3.7/site-packages/jupyterlab
[I 13:36:20.763 NotebookApp] JupyterLab application directory is /home/ubuntu/anaconda3/share/jupyter/lab
[I 13:36:20.765 NotebookApp] Serving notebooks from local directory: /home/ubuntu
[I 13:36:20.766 NotebookApp] The Jupyter Notebook is running at:
[I 13:36:20.766 NotebookApp] http://localhost:8888/?token=65d2b35c63f6f69651ddfd5f423b8329966d8a8845fd614f
[I 13:36:20.766 NotebookApp] or http://127.0.0.1:8888/?token=65d2b35c63f6f69651ddfd5f423b8329966d8a8845fd614f
[I 13:36:20.766 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 13:36:20.770 NotebookApp] No web browser found: could not locate runnable browser.
[C 13:36:20.770 NotebookApp]

To access the notebook, open this file in a browser:
    file:///home/ubuntu/.local/share/jupyter/runtime/nbserver-4487-open.html
Or copy and paste one of these URLs:
    http://localhost:8888/?token=65d2b35c63f6f69651ddfd5f423b8329966d8a8845fd614f
    or http://127.0.0.1:8888/?token=65d2b35c63f6f69651ddfd5f423b8329966d8a8845fd614f
  
```

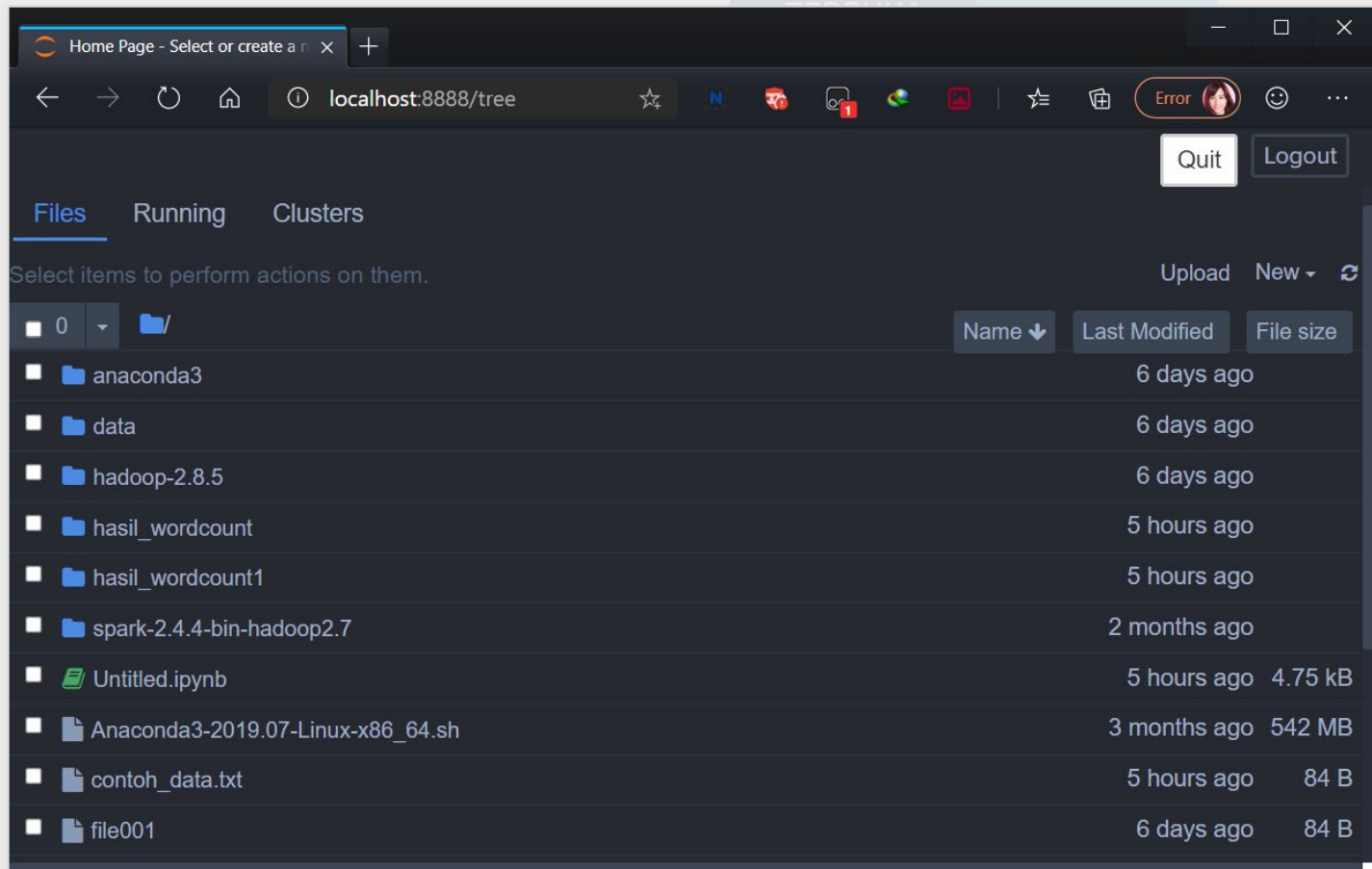
Blok baris ini dengan klik kiri *mouse* (dan otomatis ter-copy)
JANGAN dengan CTRL+C karena akan menghentikan jupyter server

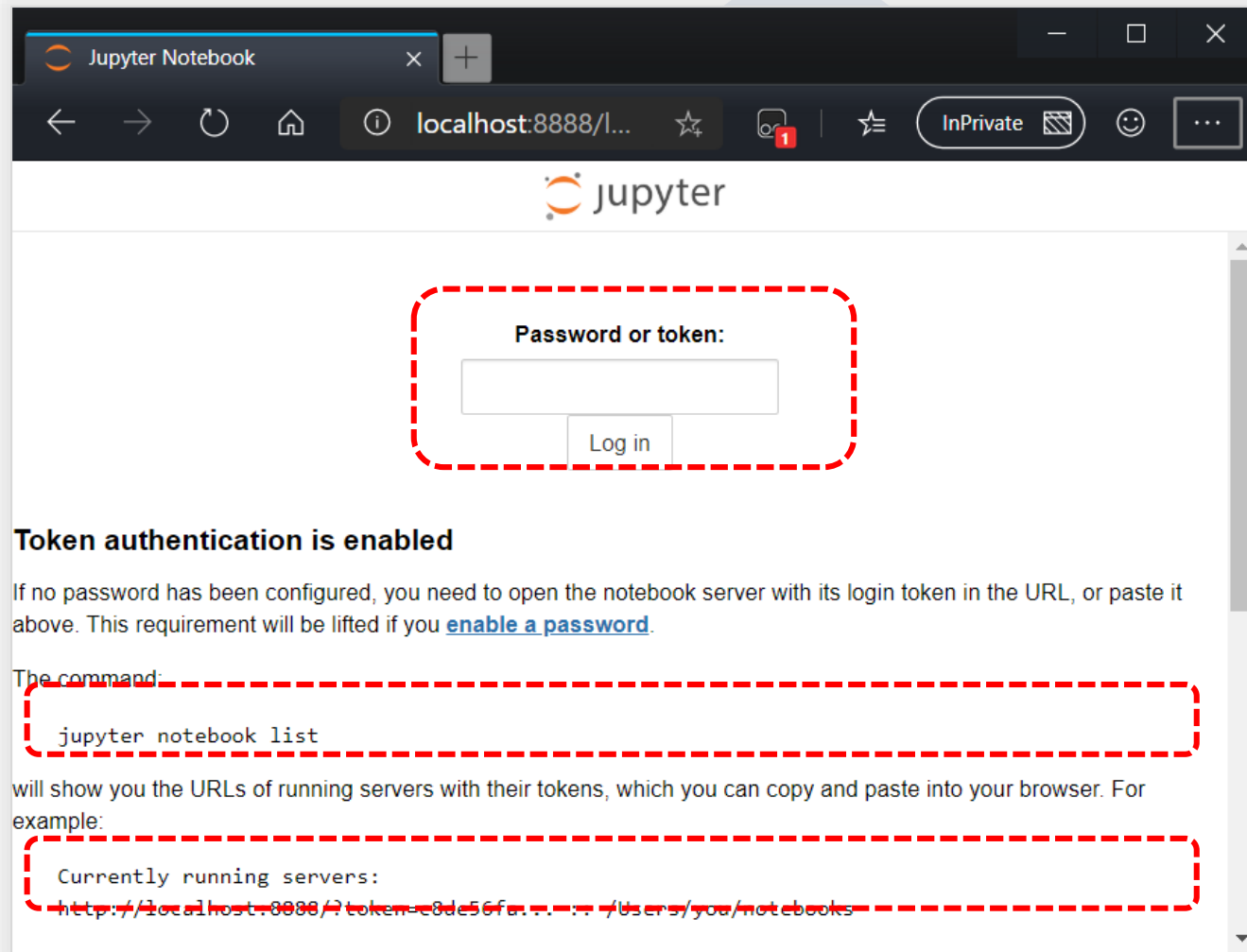
Akses Jupyter

- Untuk mengakses Jupyter yang ada EC2 di peramban komputer kita, maka cukup menggunakan URL yang kita *copy* sebelumnya.
- Atau ketikkan **localhost:port** (sesuai port local yang di SSH tunnelling)
 - Untuk yang ini kita perlu token dari URL yang diberikan oleh Jupyter untuk bisa login



Jupyter Melalui SSH Tunneling ke EC2





Memanggil Spark di Jupyter

- Setelah instalasi modul findspark, ketikkan baris berikut di *cell* pertama Jupyter

```
import findspark
findspark.init()
import pyspark
sc = pyspark.SparkContext(appName="AppTerserah")
```

- Selanjutnya untuk akses Spark melalui SparkContext **sc** layaknya di PySpark Shell



```
In [3]: import findspark
        findspark.init()
        import pyspark
        sc = pyspark.SparkContext(appName="myAppName")
```

```
In [4]: x = [1,2,3,4,5,6,7,8,9,10]
```

```
In [5]: rdd = sc.parallelize(x)
```

```
In [6]: rdd
```

```
ParallelCollectionRDD[0] at parallelize at PythonRDD.scala:195
```

```
In [7]: rdd.collect()
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

```
In [10]: rdd_map = rdd.filter(lambda x: x%2!=0)
```

```
In [11]: rdd_map.collect()
```

```
[1, 3, 5, 7, 9]
```

```
In [12]: hasil_reduce = rdd_map.reduce(lambda a,b: a+b)
```

```
In [13]: hasil_reduce
```