TP Apache Spark Dataframes

v1.1

Table of Contents

Intro

Pour ce TP vous allez utiliser **spark-notebook** qui a ete précédemment installé dans votre VM.

Preparation de l'environnement

> Dans nos exemples l'adresse IP de la VM est ***192.168.56.101***. ***N'oubliez pas de remplacer partout dans les exemples cette adresse par l'adresse affichée dans la console VirtualBox***:
>
> Pour acceder a l'interface spark-notebook il faudra rediriger egalement le port 9001 (en plus des ports habituels).

BASH

```
[andrei@desktop ~]$ ssh  -L 9080:127.0.0.1:8080 \
                         -L 8081:127.0.0.1:8081 \
                         -L 8082:127.0.0.1:8082 \
                         -L 4040:127.0.0.1:4040 \
                         -L 9001:127.0.0.1:9001 \
                   bigdata@192.168.56.101
bigdata@192.168.56.101's password:
Last login: Sun Jan  4 14:53:32 2015 from pc12.home
```

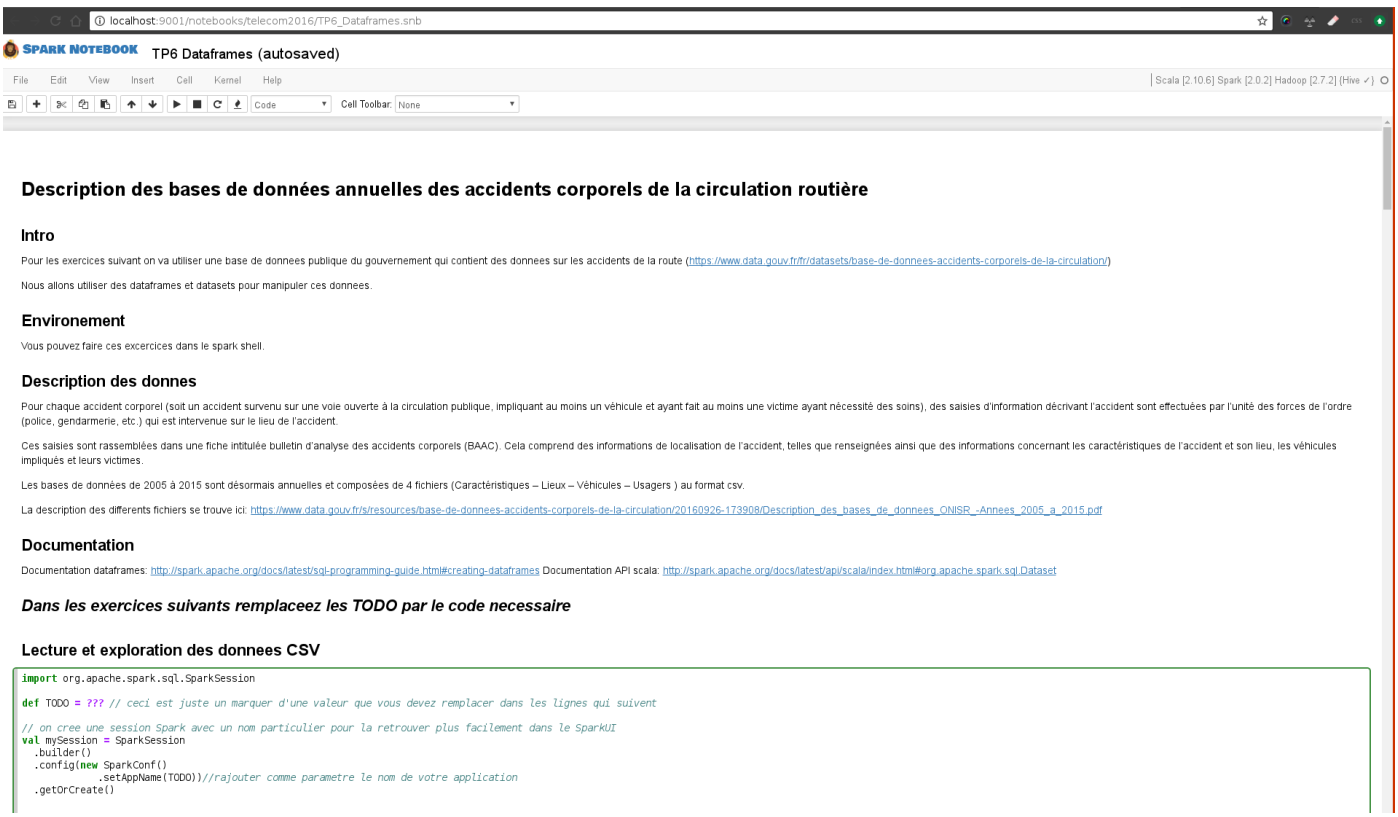
Spark notebook

- (1) lancer le spark-notebook (depuis le bon repertoire !)

BASH

```
[bigdata@bigdata ~]$ cd /home/bigdata/spark-notebook-0.7.0/  1
[bigdata@bigdata spark-notebook-0.7.0]$ bin/spark-notebook  2
Play server process ID is 4582
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/bigdata/spark-notebook-
0.7.0/lib/ch.qos.logback.logback-classic-
1.1.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/bigdata/spark-notebook-
0.7.0/lib/org.slf4j.slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type
[ch.qos.logback.classic.util.ContextSelectorStaticBinder]
[info] play - Application started (Prod)
[info] play - Listening for HTTP on /0:0:0:0:0:0:0:0:9001  3
```

- (2) suivez les exercices sur l'interface du spark-notebook (http://localhost:9001/notebooks/telecom2016/TP6_Dataframes.snb) Vous pouvez editer une cellules en cliquant. Vous pouvez executer le code d'une cellule via le menu/Cell/Run ou via `Shift` + `Enter`. Pour avoir la completion automatique du code vous pouvez utiliser `TAB`. Si le notebook ne reponds pas vous pouvez redemarer le kernel spark via Kernel/Restart ou redemarer le notebook (`Ctrl` + `C` dans le terminal puis relancer bin/spark-notebook).



Identifiants de connexion:

- utilisateur: ***bigdata***
- password: ***bigdatafuret***

Ressources

Spark Programming Guide (https://spark.apache.org/docs/latest/sql-programming-guide.html)

SQL Dataframes Tutorial (http://spark.apache.org/docs/latest/sql-programming-guide.html)

Scala API (http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.package)

Scala Cheat-Sheet
(http://homepage.cs.uiowa.edu/~tinelli/classes/022/Fall13/Notes/scala-quick-reference.pdf)

Last updated 2018-12-10 17:10:59 CET