Laboratorio 1. (1° giorno - mattina)

Obiettivi

- Familiarizzarsi con l'ambiente del laboratorio
  - Unix shell
  - Notebook Apache Zeppelin (https://zeppelin.apache.org/)
  - Internet browser (Firefox)
- Ripasso di alcuni comandi di HDFS
- Ripasso del linguaggio Scala

Comandi HDFS

Possiamo usare una shell unix per analizzare il contenuto di HDFS

Per elencare i file:

```
$ hdfs dfs -ls
```

o anche:

```
$ hadoop fs -ls
```

È possibile specificare un path (relativo o assoluto):

```
$ hdfs dfs -ls spark
$ hdfs dfs -ls /user/cloudera/spark
$ hdfs dfs -ls hdfs:///user/cloudera/spark
```

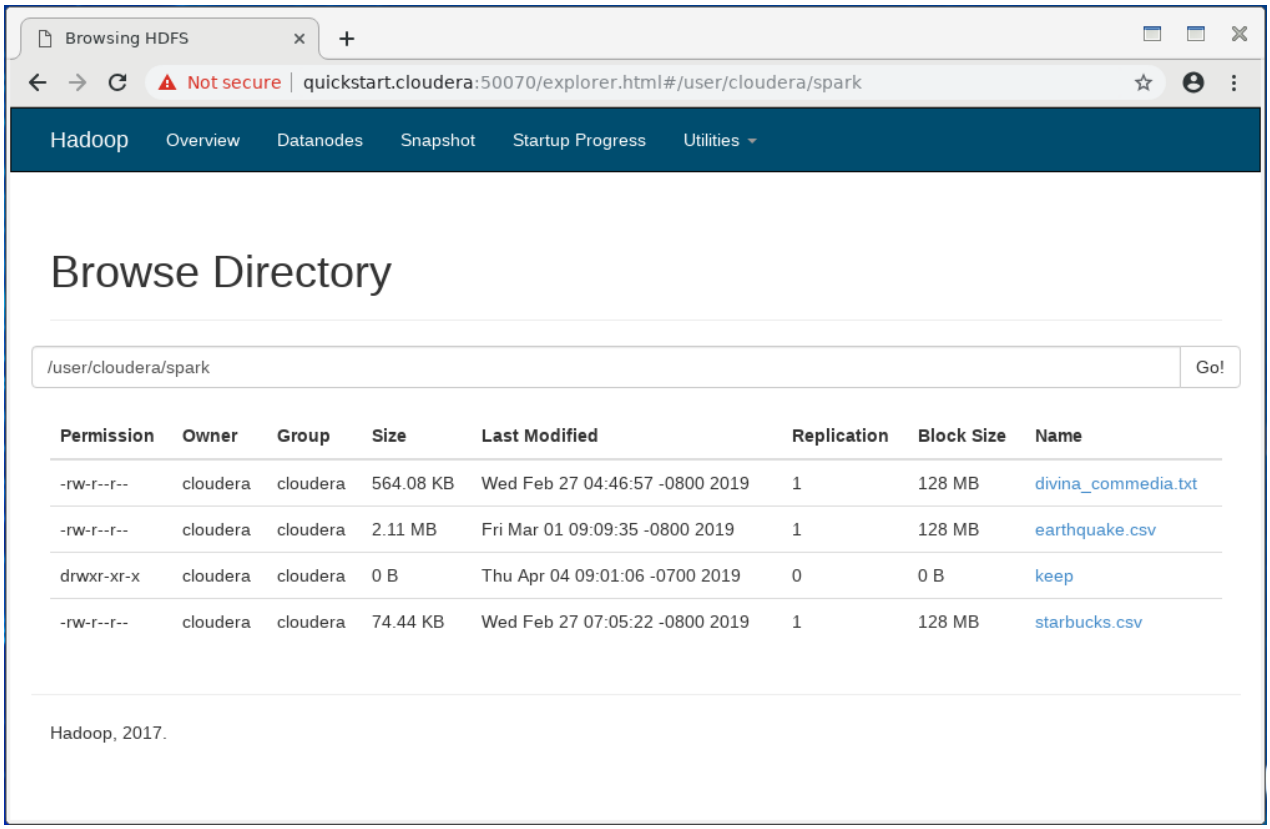
Per esaminare il contenuto di un file:

```
$ hdfs dfs -cat spark/divina_commedia.txt | less
```

Si può anche utilizzare l'interfaccia grafica. Con un Internet Browser (ad es. Firefox) si va su:

```
http://quickstart.cloudera:50070/explorer.html
```

E poi si può analizzare il file system in modo grafico:



Linguaggio SCALA

Il riferimento principale è il sito https://docs.scala-lang.org/.

Per questo corso è stato preparato un "cheat sheet" (*Scala cheat sheet.pdf*)

La shell di spark

La shell di spark si lancia con il comando:

```
$ spark-shell
```

La shell è un interprete che accetta ed esegue immediatamente comandi in Scala/Spark:

```
scala> spark.version
res0: String = 2.4.0
```

Per uscire si usa il comando *:quit*

```
scala> :quit
```

Il notebook Apache Zeppelin

Useremo per la maggior parte degli esercizi il Notebook Apache Zeppelin (https://zeppelin.apache.org/).

Il notebook è una web-application, quindi prevede un "demone" che gira in background e un Internet Browser che visualizza l'interfaccia grafica. Il demone dovrebbe partire in automatico. In ogni caso lo si può controllare con i comandi:

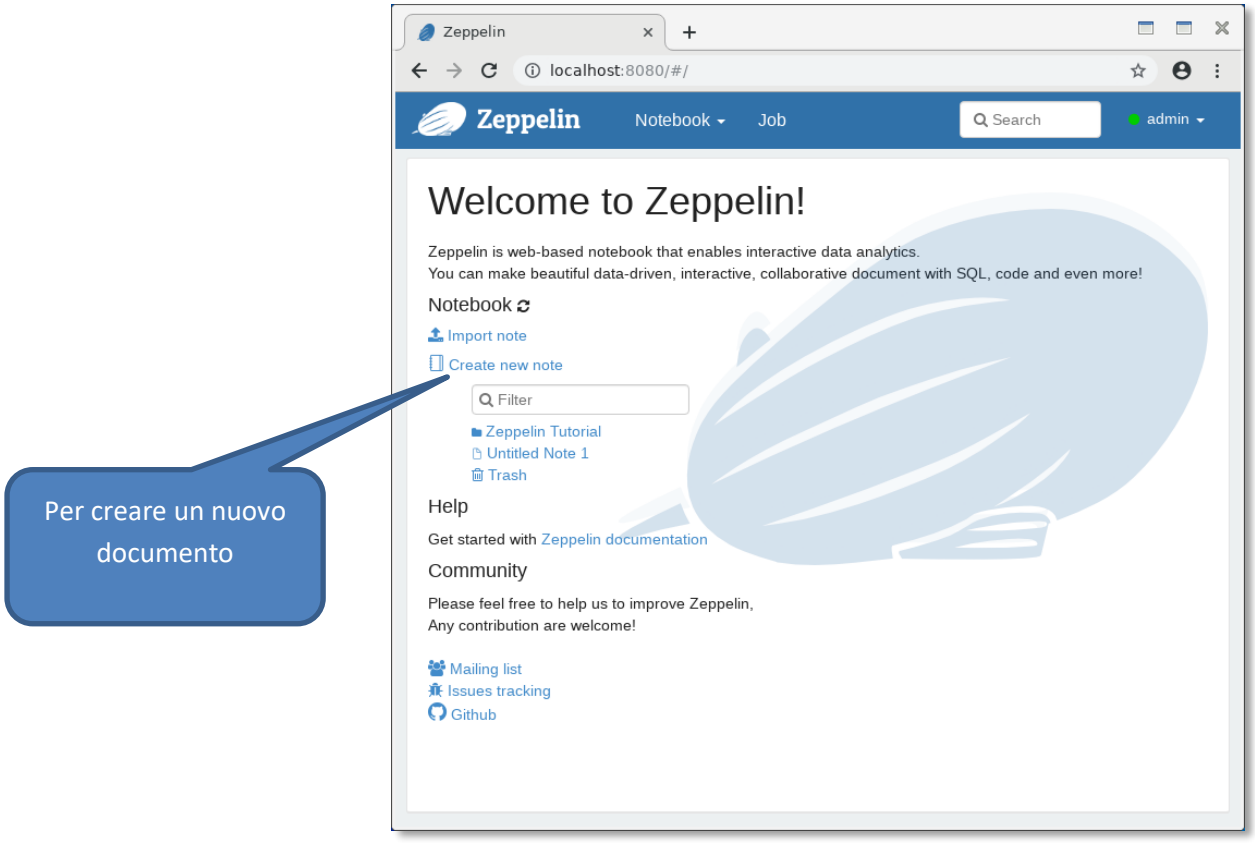
```
$ cd /opt/bigdata/zeppelin-0.8.1
$ bin/zeppelin-daemon.sh status
Zeppelin is running                              [  OK  ]
```

Ed eventualmente farlo partire con:

```
$ bin/zeppelin-daemon.sh start
```

Poi bisogna aprire un browser (ad es. Firefox) e andare su:

```
http://localhost:8080
```



Repository su github

Il materiale relativo agli esercizi da svolgere (ad es. questo stesso file o i notebook esportati) sono su un repository github pubblico. Per fare il checkhout (da shell, posizionandosi in un folder vuoto):

```
$ git clone git://github.com/gianmarco-todesco/corso-tim-spark.git
```

Per aggiornare il contenuto (ad es. nei prossimi giorni):

```
$ cd corso-tim-spark
$ git pull
```

Nel repository cè un folder *notebooks* che contiene i notebook zeppelin da importare. Per importarli:

1. Aprire Zeppelin (su Firefox! Google Chrome sembra avere un problema nella versione installata nelle macchine del laboratorio!)
2. Selezionare ***Import Note***
3. Indicare il file .json

La *nota* importata si trova nel folder ***Spark***

Esercizi

Esercizio.1

Aprire una shell unix ed esplorare i file presenti su HDFS. Prendere nota del path assoluto dei file (*/user/cloudera*). Esaminare il contenuto del file *spark/divina_commedia.txt*

Esercizio.2

Lanciare una *spark-shell*. Digitare qualche semplice espressione come ad es.:

```
scala> 10*3+10+2
res1: Int = 42
scala> print("Ciao, Spark!")
Ciao, Spark!
```

Verificare il funzionamento delle frecce in su e in giù per richiamare i comandi precedenti.

Provare a copiare ed incollare il blocco di comandi seguente:

```
if(true)
   print("è vero")
else
   print("è falso")
```

Non funziona! Perché? Digitare il comando *:paste*

```
Scala> :paste
// Entering paste mode (ctrl-D to finish)
```

Fare di nuovo la copia e premere *ctrl-D*.

Uscire dalla shell con il comando *:quit*

Esercizio.3

Verificare che il "demone" di Zeppelin sia attivo:

```
$ /opt/bigdata/zeppelin-0.8.1/bin/zeppelin-daemon.sh status
Zeppelin is running    [  OK  ]
```

Aprire un browser alla pagina: *http://localhost:8080*.
Caricare successivamente i notebook nel folder *Spark/Lab1.* Leggere le *note* e fare esperimenti.

Esercizio.4

Creare un nuovo notebook (vuoto) e provare i vari costrutti presentati nel *cheat sheet* di *Scala.*