Collège de Bois de Boulogne

420-BD6-BB Traitement Big data

TP 1 – Spark – Été 2018

Alejandra Jimenez Nieto (1695997)

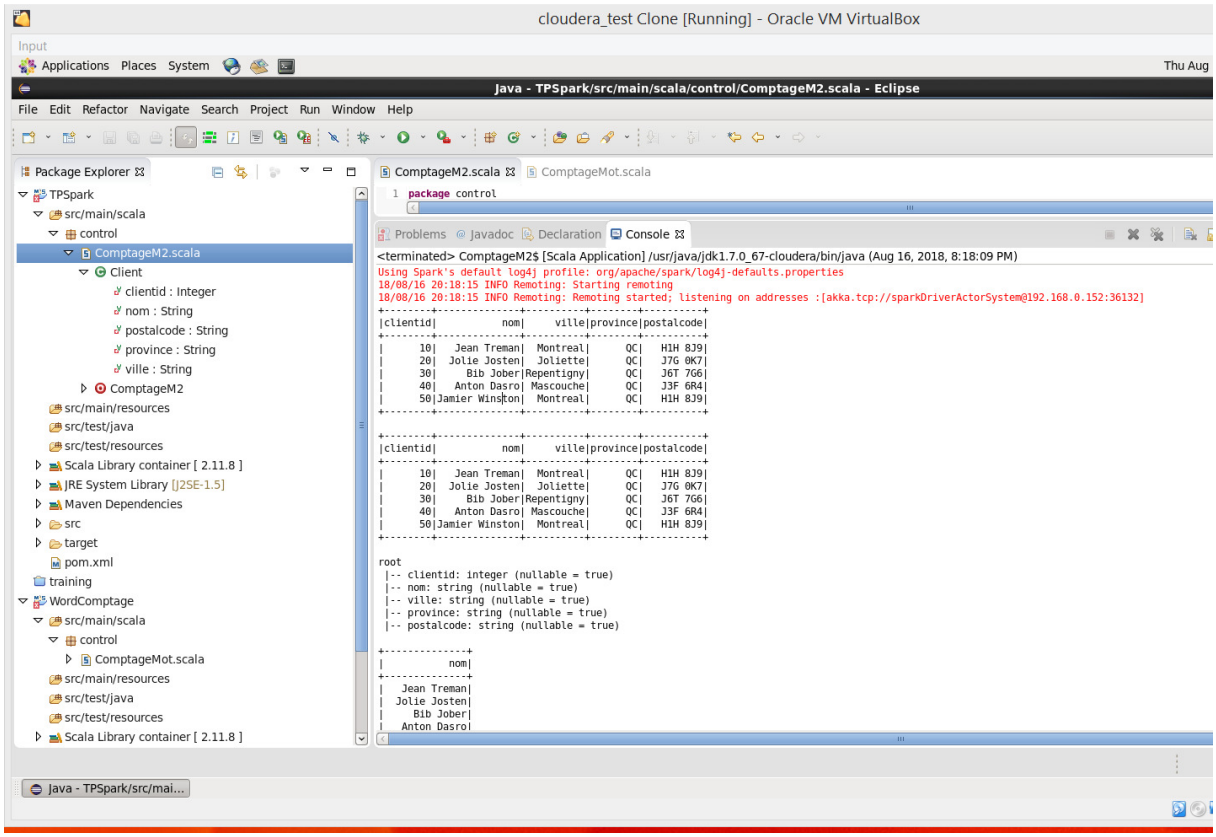
Travail présenté à: Hafed Betenfita
16 août 2018

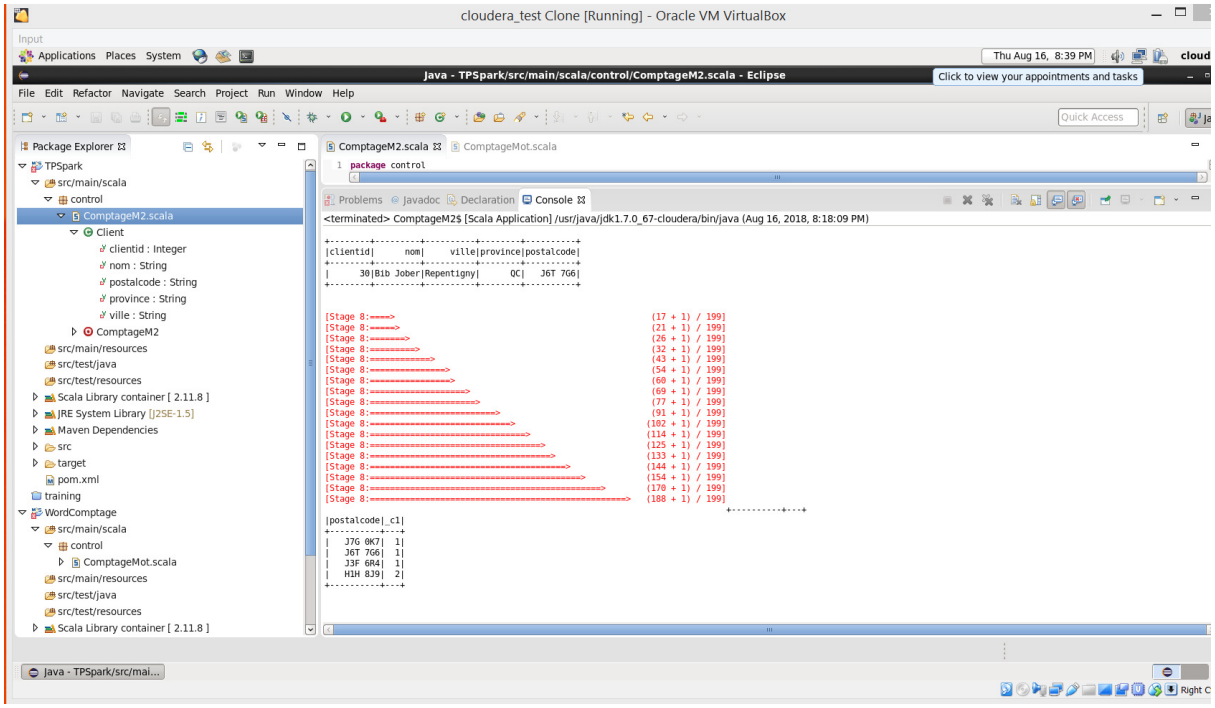
1. Code Java :

```
package control
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._

import org.apache.log4j.Logger
import org.apache.log4j.Level

//2.	Créer une classe pour représenter un Client. On devra très probablement utiliser une case class.
case class Client(clientid: Integer, nom: String, ville: String, province: String, postalcode: String)

object ComptageM2 {
  def main(args: Array[String]){

    Logger.getLogger("org").setLevel(Level.OFF)
    Logger.getLogger("akka").setLevel(Level.OFF)

    val conf = new SparkConf()
    conf.setAppName("SparkSQL").setMaster("local")

    //Creation du Context Spark
    val sc = new SparkContext(conf)


    // 1. Créer un SQLContext à partir du context Spark existant
    val sqlContext = new org.apache.spark.sql.SQLContext(sc)

    // conversion RDD vers DataFrame.
    import sqlContext.implicits._

    //au chargement du data au niveau du DataFrame
    val clientText = sc.textFile("file:///home/cloudera/client.csv")
    clientText.first()
```

```
    // creation RDD d'objets Client
    val client = clientText.map(_.split(",")).map(p =>
Client(p(0).toInt,p(1),p(2),p(3),p(4)))

    // 3. Créer un dataframe qui va contenir les objets clients lus à partir d'un
dataset représentant le fichier des clients.
    //On appelera le dataframe dfClients.
    val dfClients = client.toDF()

    // 4. Faire en sorte d'enregistrer le dataframe comme une table –clients-
    dfClients.registerTempTable("client")
    var results = sqlContext.sql("SELECT * FROM client")
    results.show()

    //5.  Afficher le contenu du dataframe dfClients.
    dfClients.show()

    // 6. Afficher le schéma du dataframe dfClients.
    dfClients.printSchema()

    // 7. Procéder à un SELECT de la colonne –nom- seulement
    dfClients.select("nom").show()

    //8.  Procéder à un SELECT des colonnes –nom- et -ville-
    dfClients.select("nom","ville")

    // 9. Afficher le détail du client ayant un ID 30
    val cid= dfClients.filter("clientid= 30")
    cid.show()

    //10.          Procéder au groupage des clients par code postal
    results = sqlContext.sql("SELECT postalcode, count(clientid) FROM client
GROUP BY postalcode")
    results.show()
  }
}
```

2. Résultat :


cloudera_test Clone [Running] - Oracle VM VirtualBox
Java - TPSpark/src/main/scala/control/ComptageM2.scala - Eclipse
File Edit Refactor Navigate Search Project Run Window Help
Package Explorer
TPSpark
src/main/scala
control
ComptageM2.scala
Client
clientid : Integer
nom : String
postalcode : String
province : String
ville : String
ComptageM2
src/main/resources
src/test/java
src/test/resources
Scala Library container [2.11.8]
JRE System Library [J2SE-1.5]
Maven Dependencies
src
target
pom.xml
training
WordComptage
ComptageMot.scala
ComptageM2.scala
ComptageMot.scala
Problems Javadoc Declaration Console
<terminated> ComptageM2$ [Scala Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Aug 16, 2018, 8:18:09 PM)
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
18/08/16 20:18:15 INFO Remoting: Starting remoting
18/08/16 20:18:15 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.0.152:36132]
|clientid| nom| ville|province|postalcode|
10	Jean Treman	Montreal	QC	H1H 8J9
20	Jolie Josten	Joliette	QC	J7G 0K7
30	Bib Jober	Repentigny	QC	J6T 7G6
40	Anton Dasro	Mascouche	QC	J3F 6R4
50	Jamier Winston	Montreal	QC	H1H 8J9
root
-- clientid: integer (nullable = true)
-- nom: string (nullable = true)
-- ville: string (nullable = true)
-- province: string (nullable = true)
-- postalcode: string (nullable = true)
nom
Jean Treman
Jolie Josten
Bib Jober
Anton Dasro
Java - TPSpark/src/mai...


cloudera_test Clone [Running] - Oracle VM VirtualBox
Java - TPSpark/src/main/scala/control/ComptageM2.scala - Eclipse
Click to view your appointments and tasks
File Edit Refactor Navigate Search Project Run Window Help
Package Explorer
ComptageM2.scala
ComptageMot.scala
Problems Javadoc Declaration Console
<terminated> ComptageM2$ [Scala Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Aug 16, 2018, 8:18:09 PM)
|clientid| nom| ville|province|postalcode|
| 30|Bib Jober|Repentigny| QC| J6T 7G6|
[Stage 8:===> (17 + 1) / 199]
[Stage 8:====> (21 + 1) / 199]
[Stage 8:======> (26 + 1) / 199]
[Stage 8:=========> (32 + 1) / 199]
[Stage 8:============> (43 + 1) / 199]
[Stage 8:==============> (54 + 1) / 199]
[Stage 8:================> (60 + 1) / 199]
[Stage 8:====================> (69 + 1) / 199]
[Stage 8:======================> (77 + 1) / 199]
[Stage 8:========================> (91 + 1) / 199]
[Stage 8:============================> (102 + 1) / 199]
[Stage 8:================================> (114 + 1) / 199]
[Stage 8:====================================> (125 + 1) / 199]
[Stage 8:======================================> (133 + 1) / 199]
[Stage 8:==> (144 + 1) / 199]
[Stage 8:==> (154 + 1) / 199]
[Stage 8:==> (170 + 1) / 199]
[Stage 8:==> (188 + 1) / 199]
postalcode	_c1
J7G 0K7	1
J6T 7G6	1
J3F 6R4	1
H1H 8J9	2
Java - TPSpark/src/mai...