Class 5: XML with RDDs

New York University

Spring 2018

Homework

1. Process Data Files with Spark (Provide in NYU Classes Assignment the code you used.)

   a. Start two terminal windows. In one window, start the Scala Spark Shell: $ spark-shell Use the other window for command line operations

   b. Copy the activations.zip file to the VM, unzip it, and store it in HDFS to: loudacre/activations
   (Note: It might be a good idea to delete from the VM large input files from previous homework assignments.)

   Each XML file contains data for all the devices activated by customers during a specific month. Here's an example of the XML layout:

```
<activations>
  <activation timestamp="1225499258" type="phone">
    <account-number>316</account-number>
    <device-id>
      d61b6971-33e1-42f0-bb15-aa2ae3cd8680
    </device-id>
    <phone-number>5108307062</phone-number>
    <model>iFruit 1</model>
  </activation>
  …
</activations>
```

Homework (continued)

1. Process Data Files with Spark (continued)

c. Your code should process a set of activation XML files and extract the account number and device model for each activation, and save the list to a file formatted as account_number:model.

The output will look something like:

1234:iFruit 1
987:Sorrento F00L
4566:iFruit 1
…

d. Use wholeTextFiles to create an RDD from the activations dataset. The resulting RDD will consist of tuples, in which the first value is the name of the file, and the second value is the contents of the file (XML) as a string.

e. Each XML file can contain many activation records; use flatMap to map the contents of each file to a collection of XML records. Take each XML string, parse it, and return a collection of XML records; map each record to a separate RDD element. (flatMap maps each record to a separate RDD.)

Homework (continued)

1. Process Data Files with Spark (continued)

c. Your code should process a set of activation XML files and extract the account number and device model for each activation, and save the list to a file formatted as account_number:model.

The output will look something like:

1234:iFruit 1
987:Sorrento F00L
4566:iFruit 1
…

d. Use wholeTextFiles to create an RDD from the activations dataset. The resulting RDD will consist of tuples, in which the first value is the name of the file, and the second value is the contents of the file (XML) as a string.

e. Each XML file can contain many activation records; map the contents of each file to a collection of XML records. Take each XML string, parse it, and return a collection of XML records; map each record to a separate RDD element.

f. Map each activation record to a string in the format: account-number:model

g. Save the formatted strings to a text file in the directory /loudacre/activations/account-models

Homework (continued)

1. Process Data Files with Spark (continued)

Hints:
a. Use the Scala XML library by importing it: import scala.xml._

b. Consider creating functions like the following:

```
// Given a string containing XML, parse the string, and
// return an iterator of activation XML records (Nodes) contained in the string

def getactivations(xmlstring: String): Iterator[Node] = {
    val nodes = XML.loadString(xmlstring)      \\ "activation"
    nodes.toIterator
}

// Given an activation record (XML Node), return the model name
def getmodel(activation: Node): String = {
   (activation \ "model").text
}

// Given an activation record (XML Node), return the account number
def getaccount(activation: Node): String = {
   (activation \ "account-number").text
}
```