Lab 5

Introduction

In this lab, you will write a Spark application to analyze XML data.

Objectives

- Practice writing, building, and running Spark applications.
- Practice parsing XML files.

Preparation

Transfer activations.zip to the Peel cluster, unzip it, and store it in HDFS to `loudacre/activations`.

Each XML file contains data for all the devices activated by customers during a specific month. Here is 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>
```

Your task

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 file(s) formatted as *account-number:model*.

The output will look something like: (don't worry about the order)

```
9763:MeeToo 1.0
426:Titanic 1000
383:Sorrento F00L
...
```

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.

Parse each XML file and map each activation record to a string in the format: *account-number:model*. Save the formatted strings to the HDFS directory `loudacre/account-models`.

You need to provide `build.sbt` so that your code can be compiled using `sbt package` and run using `spark-submit`.

(*Scroll to the bottom of this page for some tips.*)

Submission

To receive full credit, please compress all of the following items into a single file named `YourName_NetID_lab5.zip` and upload it to NYU Brightspace.

- All the Scala source code.
- `build.sbt`.
- Output file(s) that you retrieve from HDFS.
- Screenshots showing that you can successfully build and run your Spark application and generate the desired results.

Tips

Consider creating functions like the following:

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

Please use Discord if you experience any difficulties. The graders and I will help you get your environment working.

Don’t procrastinate. The Hadoop cluster tends to get crowded near the due date.