Ask Bigger Questions

# CAP - Developing with Spark and Hadoop:

# Homework Assignment Guide for Students

# Homework: Process Data Files with Spark

## Files and Data Used in This Homework:

Exercise Directory: `$DEV1/exercises/spark-etl`

Data files (local):

`$DEV1DATA/activations/*`

`$DEV1DATA/devicestatus.txt` (Bonus)

Stubs (professor VM only):

`ActivationModels.pyspark`

`ActivationModels.scalaspark`

**In this exercise you will parse a set of activation records in XML format to extract the account numbers and model names.**

Spark is commonly used for ETL (Extract/Transform/Load) operations. Sometimes data is stored in line-oriented records, like the web logs in the previous exercise, but sometimes the data is in a multi-line format that must be processed as a whole file. In this exercise you will practice working with file-based instead of line-based formats.

## Review the API Documentation for RDD Operations

Visit the Spark API page you bookmarked previously. Follow the link at the top for the RDD class and review the list of available methods.

## The Data

Review the data in `$DEV1DATA/activations`. Each XML file contains data for all the devices activated by customers during a specific month.

Copy this data to `/loudacre` in HDFS.

Sample input data:

```
<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>
```

## The 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 a file formatted as `account_number:model`.

The output will look something like:

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

1. Start with the `ActivationModels` (professor VM only) script in the exercise directory. (A stub is provided for Scala and Python; use whichever language you prefer.) Note that for convenience you have been provided with functions to parse the XML, as that is not the focus of this Exercise. Copy the stub code into the Spark Shell.

2. 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.

3. Each XML file can contain many activation records; use `flatMap` to map the contents of each file to a collection of XML records by calling the provided `getactivations` function. `getactivations` takes an XML string, parses it, and returns a collection of XML records; `flatMap` maps each record to a separate RDD element.

4. Map each activation record to a string in the format `account-number:model`. Use the provided `getaccount` and `getmodel` functions to find the values from the activation record.

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

## Optional Homework

Another common part of the ETL process is data scrubbing. In this optional homework, you will process data in order to get it into a standardized format for later processing.

Review the contents of the file `$DEV1DATA/devicestatus.txt`. This file contains data collected from mobile devices on Loudacre's network, including device ID, current status, location and so on. Because Loudacre previously acquired other mobile provider's networks, the data from different subnetworks has a different format. Note that the records in this file have different field delimiters: some use commas, some use pipes (`|`) and so on. Your task is to

- Load the dataset

- Determine which delimiter to use (hint: the character at position 19 is the first use of the delimiter)

- Filter out any records which do not parse correctly (hint: each record should have exactly 14 values)

- Extract the date (first field), model (second field), device ID (third field), and latitude and longitude (13th and 14th fields respectively)

- The second field contains the device manufacturer and model name (e.g. `Ronin S2`.) Split this field by spaces to separate the manufacturer from the model (e.g. manufacturer `Ronin`, model `S2`.)

- Save the extracted data to comma delimited text files in the `/loudacre/devicestatus_etl` directory on HDFS.

- Confirm that the data in the file(s) was saved correctly.

The solutions to the optional homework are in `$DEV1/exercises/spark-etl/bonus` (professor VM only).

## This is the end of the Homework