TP_SPARK_CORE_2

Axel Jacquin

January 17, 2017

One of the commons uses for Spark is doing data Extract/Transform/Load operations (ETL). 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.

1 The Data

1. Review the data in /home/training/training_materials/sparkdev/data/activations Each XML file contains data for all the devices activated by customers during a specific month.

2. Copy this dataset in HDFS under /loudacre directory

2 The Task

Your objective is to extract the account number and device model for each activation, and save the list to a file as account_number:model.

3. The file /home/training/training_materials/sparkdev/stubs/ActivationModels.scalaspark provides you scala functions to parse the given XML. Have a look in this file. Copy the stub code into a Spark Shell.

4. Use wholeTextFiles to create an RDD named activationFiles 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.

5. Each XML file can contains many activation records; use *flatMap* to map the contents of each file to a collection of XML records by calling the provided *getactivatons* function. *getactivations* takes an XML string, parses it and returns a collection of XML records.

6. Map each activation record to a string in the format *account-number:model* using the provided *getaccount* and *getmodel* functions

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

3 Bonus Exercise

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

Review the contents of the data file /home/training/training_materials/sparkdev/datadevicestatus.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 sub-networks 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 each delimiter to use (hint: the character at postion 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.