Homework (continued)

4. Data Scrubbing with Spark

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

Review the contents of the file **`devicestatus.txt`**. This file contains data collected from mobile devices on Loudacre's network (Loudacre is a fictional telco), 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. Though the delimiter symbol may vary, it will appear at position 19 (the 20th character).

Your task is to read in the file and drop records that do not contain 14 values.
From the remaining valid records, produce a cleaned up output file that contains the date, manufacturer (without model), device ID, latitude and longitude.

Steps:

a. Load the dataset
b. Determine which delimiter to use - hint: the character at position 19 (the 20th character) is the first use of the delimiter.
c. Filter out any records which do not parse correctly - hint: each record should have exactly 14 values.
d. Extract the date (first field), mfr_model (second field), device ID (third field), and latitude and longitude (13th and 14th fields respectively).
e. The second field (mfr_model) contains the device manufacturer and model name (e.g. "Ronin S2" or "Sorrento F41L") Split this field on the blank(s) to separate the manufacturer from the model (e.g. manufacturer "Ronin", model "S2"), and assign the value extracted for manufacturer to the second field of your output.
f. Save the extracted data, comma delimited, to text files in the **`/loudacre/devstatus/devicestatus_etl`** directory on HDFS.

   Remember to trim the '(' and ')' at the start and end of each line so that each line of the file has just the comma-separated values.
g. Confirm that the data in the file(s) was saved correctly.

Upload to NYU Classes the commands you used to complete this task and a screenshot(s) showing the commands executing.

Ⓒ 2013-2020 Suzanne McIntosh