Lab 3

Introduction

A common part of the ETL (extract, transform, and load) process is data scrubbing. In this lab, you will process data in order to get it into a standardized format for later processing.

Objectives

- Practice data scrubbing with Spark.

Your task

Review the contents of the file devicestatus.txt. This file contains data collected from mobile devices on Loudacre's network (Loudacre is a fictional telephone company), 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 the 20th character (*i.e.*, position 19)**.

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

1. Put the dataset to `/user/yourNetID/loudacre/devstatus/devicestatus.txt` on HDFS and load the dataset in Spark.

2. Determine which delimiter to use. *Hint: the 20th character (at position 19) is the first use of the delimiter.*
3. Filter out any records which do not parse correctly. *Hint: each record should have exactly 14 values.*
4. Extract the **date** (first field), **manufacturer and model** (second field), **device ID** (third field), and **latitude and longitude** (13th and 14th fields respectively).
5. The second field 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 (discard the model).
6. Save the extracted data, comma delimited, to text files in the `/user/yourNetID/loudacre/devstatus/devicestatus_etl` directory on HDFS. (Note: If you represented the data as a tuple, 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.) Here is a sample output record:

    ```
    2014-03-15:10:10:20,Sorrento,8cc3b47e-bd01-4482-b500-28f2342679af,33
    ```

7. Confirm that the data in the file(s) was saved correctly. Retrieve the output file(s) from HDFS.

Please take screenshots showing the commands you used to complete this task.

Submission

To receive full credit, please compress all of the following items into a single file named `YourName_NetID_lab3.zip` and upload it to NYU Brightspace.

- The output file(s) that you retrieved from HDFS.

- Screenshot(s) showing the commands executing.

Tips

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.