Class 4: Working with RDDs, Scala Tuples

New York University

Spring 2018

Homework

1. Review the API Documentation for RDD Operations

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

2. Please read through Chapter three in the class text - you can skip Java sections. Read the Python sections only if you plan to use Python for your project.

3. Reading: If you haven't already read it, please read the paper provided in the Resources tab (also available at: http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0061):

   "Targeting Villages for Rural Development Using Satellite Image Analysis"

   by Kush R. Varshney, George H. Chen, Brian Abelson, Kendall Nowocin, Vivek Sakhrani, Ling Xu, and Brian L. Spatocco.

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 to comma delimited text files in the **`/loudacre/devstatus/devicestatus_etl`** directory on HDFS.
g. Confirm that the data in the file(s) was saved correctly.

Upload to NYU Classes the commands you used to complete this task.

Homework - Big Data Project

5. Read the material in the Application Project Description packet provided on NYU Classes in file: *BDAD_Spring2018_ProjectInfo.pdf*

6. Propose a Big Data Application!

This is an individual homework assignment. **It is not a team assignment**. Please develop your idea independently.

Write a project proposal describing a Big Data application that you would like to build using the template: *BDAD_Spring2018_IPP. pages (or .docx)*

You should identify at least two potential data sources that could be used in your project.

Some things to consider about your data sources:

- How can you obtain a copy of the data? Who owns the data? Is it open data? Is it public or private?
- How much data is there in each source – just magnitude – is it MB? GB? TB?
- If the data is very large, where can you store it?
- How do you gain access to the data? How long will it take to get access/be approved?
- Who must you ask for permission to access the data?
- Is the data static, periodic, near realtime?

  If it's static, you will copy it once, the source data will not grow.

  If it's periodic, you will copy it periodically, which means your files will be growing over time.

  If it's near realtime, you will be continually reading in the data, which means your data will be growing over time.
- If the data is being collected in near real-time, what is the velocity of the data and the volume per unit of time? Can your Hadoop environment support this?

This may, or may not, turn out to be the project that you ultimately build with your team. Use your imagination, take a risk.
You will form teams next week. Teams will select one of the projects proposed by team members, or design an entirely different project to implement as a team.