Lab 8: Standard Operators in Streams Processing Language (SPL)

Department of Electrical and Computer Engineering, Iowa State University
Software Tools for Large-Scale Data Analysis, Spring 2013

Purpose

The purpose of this lab is to learn to use standard SPL operators, such as Filter, Functor, Custom, Union, and Split, as well as containers such as map, set, and list. For examples of their use, please see the example programs provided with the lecture, and the 50 examples provided by IBM. For documentation on these operators, consult the SPL Standard Toolkit Reference.

Submission

Create a zip (or tar) archive with the following and hand in using blackboard.

- A write-up answering questions for each experiment in the lab. For output, cut and paste results from your terminal and summarize when necessary.
- Commented Code for your program. Include all source files needed for compilation.

Resource

You will need the following before you start the lab. If you don't have these, please contact the instructors.

- IP Address of the Node
- Login id and private key file

Experiment:

City police are tracking all incoming and outgoing calls from two cell phone towers to identify potential suspects for an attack believed to be planned at midnight. The phone call information arrives as massive, continuous streams of data, one stream per tower. Each record within a stream has information about a single phone call and is of the following form. The different attributes are comma separated, with no space between them (the "csv" format).

<TimeStamp>,<Caller ID>,<Callee ID>,<Duration of Call>

where

"*TimeStamp*" is the time at which the call was started. It is of the format "*hh:mm:ss.xxx*", *hh* - hours, *mm* – minutes, ss – seconds, xxx - mseconds

"Caller ID" is the phone number from which the call is made. It is a string consisting of 10 characters.

"Callee ID" is the phone number to which the call is made. It is again a string of 10 characters.

"Duration of Call" is the duration of the call in *seconds*.

The police are interested in identifying those phone numbers with the following characteristics, which the police have deemed to be "suspicious". Any phone number which

A. made at least 35 calls during the day (not necessarily to distinct numbers)

B. made at least 30 calls to distinct numbers during the day

C. made at least 10 calls of the duration of 10sec or less during a single one hour block in the day, starting at the top of the hour. Note that here are 24 such one hour blocks, 12am-1am, 1am-2am, and so on.

   For example, if phone number P made 5 calls of duration 5 seconds each between 10am-11am, and 5 calls of duration 5 seconds each between 1pm-2pm, then P does not satisfy this criterion. If P made 15 calls of duration 3 seconds each, between 1pm-2pm, then P does satisfy criterion B.

Help the police in their investigation by writing a SPL program that helps them find a list of suspects based on the above three patterns of calls.

1. Draw an operator graph for your program.

2. Combine the list of callers of the above three types and write into an output file called "suspects.txt", with the following information, one line per record.

   <Caller ID>,<Type of Suspect>

   where "Caller ID" is the phone number deemed suspicious, and "Type of Suspect" is one of "A", "B", or "C", depending on which of the above criteria was satisfied by the caller.

   Note that there should be no more than one output line per caller, i.e. if the same caller is found suspicious according to more than one criterion among A,B, and C, the caller should not be repeatedly printed on multiple lines. Instead, such a caller should be printed once, with all the matching criteria listed next to the caller.

The input stream of call records should be read from the files "recordsX.csv" and "recordsY.csv" saved in "/datasets" folder, which are the records collected from cell phone tower "X" and "Y" respectively.