Lab 9: Standard Operators in Streams Processing Language (SPL)

Department of Electrical and Computer Engineering, Iowa State University
Software Tools for Large-Scale Data Analysis, Spring 2015

Purpose

The purpose of this lab is to learn to use standard SPL operators, such as TCPSource, Filter, Functor, Custom, 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 in IBM Information Center.

Submission

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

- Commented Code for your programs. Include all source files needed for compilation.
- Output of your programs.

Experiment:

City police are tracking all incoming and outgoing calls from a cell phone tower to identify potential suspects for an attack believed to be planned in a couple of hours. The phone call information arrives as massive, continuous stream of data. Each record within the 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 ended. It is of the format "MMMdd *hh:mm:ss.xxx*", MMM- month, dd – date, *hh* - hours, *mm* – minutes, ss – seconds, xxx – mseconds. The timestamp is in 24-hour format.

"Caller ID" is the phone number from which the call is made. It is 9-digit number.

"Callee ID" is the phone number to which the call is made. It is again a 9-digit number.

"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 15 calls in a period of 10 minutes (not necessarily to distinct numbers)

B. made at least 10 calls to distinct numbers in a period of 10 minutes.

C. made at least 5 calls of the duration of 10sec or less during a single one minute block, (starting at the top of the minute). Note that there are 60 such one-minute blocks in an hour. Examples of one-minute blocks are 00:00:00-00:00:59, 00:01:00-00:01:59, 00:02:00-00:02:59 (in hh:mm:ss format). For example, if phone number P made 3 calls of duration 5 seconds each between 01:03:00-01:03:59, and 2 calls of duration 5 seconds each between 01:05:00-01:05:59, then P does not satisfy this criterion. If P made 15 calls of duration 3 seconds each, between 01:07:00-01:07:59, then P does satisfy criterion B.

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

1. 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.

You will be using TCPSource operator to read the live streaming call records from server “10.24.84.229” at port number <your unique port number>. You will have an access to the streaming data 24 hours a day.

P.S. For the final results, run the program for at least 10 minutes, starting from the top of 10 minutes block in any hour. For instance, you may run the program at 1 2:00am, 12:10am, 12:20am, 12:30am, 12:40am, 12:50am, 1:00am, 1:10am, 1:20am, .., 11:10pm, 11:20pm, 11:30pm, 11:40pm, 11:50pm.