Spark Streaming Advanced

# Exercise 1

Let’s consider very simplified log containing only id of news articles. A main task of exercise to detect hot news at current time frame. Rule is simple -- news is hot if it’s visit rate equals or greater than 30% among other news.

It is just enough to print on a screen the id of news which is detected as hot.

Develop your solution using this code template:

import org.apache.log4j.{Level, Logger}  
import org.apache.spark.\_  
import org.apache.spark.streaming.\_  
  
Logger.getRootLogger().setLevel(Level.ERROR)  
  
val stc = new StreamingContext(sc, Seconds(5))  
stc.checkpoint("checkpoint")  
val stream = stc.socketTextStream("localhost", 33001)  
  
🡪 Insert your code here 🡨

stc.start()  
stc.awaitTermination()

When code is ready run it in spark shell. In second terminal run command

nc -kl 33001

Paste into console contents of files part-0.log, part-1.log and part-2.log with interval of 5 seconds.

What is hot news after first insert? After second? After third?