Karathanasis Dimitris

CS-543

Ferbuary 2019

Assignment 1

For the purpose of this assignment i worked locally. Downloaded Apache Spark and used VS code with the included terminal for deploying Spark as well as running Scala.

2.1

Even though I did not use AWS for this assignment I set up the enviroment on the cluster, uploaded the nasa_log.txt and checked that everything work properly.

2.2

I successfully loaded baseRdd from nasa.txt file. In order to split the initial Rdd from Rdd[String] to Rdd[Log] I map baseRdd with getLogFields function. When i trigger an action though a ton of Exceptions occur such as “java.util.NoSuchElementException”.

1. The field of Table 1 that produces exception is the bytes field. Some data instead of the number of bytes has a ‘-’ so we are not able to read data properly.
2. The pattern expects to read “ number” and when it tries to read a problematic line an exception occurs. At this point i would like to add that not only lines containing 404 errors were problematic. I also found some lines with status code 302 that had the same issue. I corrected both in my implementation.

3. I solved this problem giving the value of bytes variable to either “0” if the pattern fails or “ patternBytes.findAllIn(str).matchData.next().group(1).toInt” if pattern succeeds.
4. I used the convertToLog function where inside i mapped the initial rdd to the new getImprovedLogFields and created a new cleanRdd[Log]. I also cached it for easiest further use.

2.3

1. Explore content size:

RESULT: Min = 0 Max = 6823936 Average = 20455

2. HTTP status analysis:

RESULT: (0,1) (302,46573) (400,5) (404,10845) (304,132627)
(500,62) (403,54) (200,1701534) (501,14)

3. Frequent hosts:

RESULT: (192.188.119.226 ,22) (192.156.154.71 ,12) (n1122099.ksc.nasa.gov ,158) (line115.worldweb.net ,53) (194.108.167.34 ,12) (129.108.37.35 ,18) (ix-phx4-29.ix.netcom.com ,33) (ad10-013.compuserve.com ,44) (ac_anx_1_5.mur.csu.edu.au ,12) (sbd0114.deltanet.com ,22)

4. Top-10 error paths:

RESULT: (/images/NASA-logosmall.gif,21010) (/images/KSC-logosmall.gif,12435) (/images/MOSAIC-logosmall.gif,6628) (/images/USA-logosmall.gif,6577) (/images/WORLD-logosmall.gif,6413) (/images/ksclogo-medium.gif,5837) (/images/launch-logo.gif,4628) (/shuttle/countdown/liftoff.html,3509) (/shuttle/countdown/,3345) (/shuttle/countdown/images/cdtclock.gif,3251)

5. Unique hosts:

RESULT: Long = 81983

6. Count 404 Response codes:

RESULT: Long = 10845

7. Print 40 distinct requestURIs that generate 404 errors:

RESULT:

/history/apollo/apollo15.html

/shuttle/missions/sts-63/sts-63-press-kit.tx"

/history/apollo/publications/sp-350/sp-350.txt~

/de/systems.html

/shuttle/technology/ml

/history/skylab/skylab-overview.txt

/history/apollo/appollo-13/

/shuttle/technology/sts-newsref/et.html#et-safety

/history/apollo/apollo-13/apollo-13.htmlhppt:

/shuttle/missions/sts-70/tmp/images.html

/imagemap.cgi//sts-70/landing/landing.map?384,207

/history/apollo/a-001/sounds/

/elv/TITAN/elvhead2.gif

/cwis/organizations/kucia/uroulette/sig.gif

/news/sci.space.shuttle/archive/sci-space-shuttle-7-feb-1994-87.txt

/Government/Research_Labs/NASA/Kennedy_Space_Center/

/history/apollo/sa-3/news/

/finance/main.html

/history/apollo/apollo/-13/apollo-13-info.html

/history/appolo/appolo-13.html

/http//www.cs.tut.fi/~p116711/http//www.cs.tut.fi/~p116711/

/software///icons/menu.xbm

/ismap/image-maps/pccomp/webmap/overview.map?320,134

/shuttle/missions/weathico.gif

/shuttle/missions/missions.htnl

/shuttle/missions/sts-71/palace.com

/wxworld/images/IMAGE6pNgmSfc/95072712_00.GIF

//spacelink.msfc.nasa.gov:70/00/About.Spacelink/File.Transfer.Protocols

/apollo.

/ksc.htlm

/htbin.cdt_mail.pl

/shuttle/missions/sts-71/images/KSC-9

/history/apollo/a-003/movies/

/shuttle/technology/sts-newsref/sts_coord.html#sts_body

/jistory/apollo/apollo-13/apollo-13.html

/shuttle/count.gif

/history/apollo/apollo-13.htm

/history/apollo/apollo-11/apollo-13.html

/shuttle/missions/delta/

/shuttle/technology/st-newsref/stsref-toc.hyml

8. Print a list of the top twenty paths (in sorted order) that generate the most 404 errors:

RESULT:

(/pub/winvn/readme.txt,667)

(/pub/winvn/release.txt,547)

(/history/apollo/apollo-13.html,286)

(/shuttle/resources/orbiters/atlantis.gif,230)

(/history/apollo/a-001/a-001-patch-small.gif,230)

(/://spacelink.msfc.nasa.gov,215)

(/history/apollo/pad-abort-test-1/pad-abort-test-1-patch-small.gif,215)

(/images/crawlerway-logo.gif,214)

(/history/apollo/sa-1/sa-1-patch-small.gif,183)

(/shuttle/resources/orbiters/discovery.gif,180)

(/shuttle/missions/sts-68/ksc-upclose.gif,175)

(/shuttle/missions/sts-71/images/KSC-95EC-0916.txt,168)

(/elv/DELTA/uncons.htm,163)

(/history/apollo/publications/sp-350/sp-350.txt~,140)

(/shuttle/missions/technology/sts-newsref/stsref-toc.html,107)

(/shuttle/resources/orbiters/challenger.gif,92)

(/procurement/procurement.htm,86)

(/history/apollo-13/apollo-13.html,73)

(/history/apollo/pad-abort-test-2/pad-abort-test-2-patch-small.gif,71)

(/shuttle/countdown/video/livevideo.jpeg,68)