HY543 | Assignment 1

Papageorgiou Efthymios - csdp1344

2 Fun with Apache Spark

2.2 Data Exploration

2.2.1. Which field of Table 1 produces exceptions and requires an update of regular expressions of the attached WebLogger.scala file.

Answer: the bytes field seems problematic. Also there are some ill-formated data in the request field.

2.2.2 what kind of patterns that field is expected to contain, and what happens in the problematic lines.

Answer: Let's discuss about the Bytes field case. If we use the patternBytes regex and print the first 10 logs that mismatch with this regex, we will see that those logs may fail because of the symbol '-' which is not a number. If we take a closer look to the regex **"""^.*\s+(\d+)$""".r** we will notice that this regex requires at least one digit. So, the string "-" doesnt match the regex. A possible fix is to eliminate the force of including at least one digit.

```
dd15-062.compuserve.com - - [01/Jul/1995:00:01:12 -0400] "GET /news/sci.space.shuttle/archive/sci-space-shuttle-22-apr-1995-40.txt HTTP/1.0" 404 -
dynip42.efn.org - - [01/Jul/1995:00:02:14 -0400] "GET /software HTTP/1.0" 302 -
ix-or10-06.ix.netcom.com - - [01/Jul/1995:00:02:40 -0400] "GET /software/winvn HTTP/1.0" 302 -
ix-or10-06.ix.netcom.com - - [01/Jul/1995:00:03:24 -0400] "GET /software HTTP/1.0" 302 -
link097.txdirect.net - - [01/Jul/1995:00:05:06 -0400] "GET /shuttle HTTP/1.0" 302 -
ix-war-mi1-20.ix.netcom.com - - [01/Jul/1995:00:05:13 -0400] "GET /shuttle/missions/sts-78/news HTTP/1.0" 302 -
ix-war-mi1-20.ix.netcom.com - - [01/Jul/1995:00:05:58 -0400] "GET /shuttle/missions/sts-72/news HTTP/1.0" 302 -
netport-27.iu.net - - [01/Jul/1995:00:10:19 -0400] "GET /pub/winvn/readme.txt HTTP/1.0" 404 -
netport-27.iu.net - - [01/Jul/1995:00:10:28 -0400] "GET /pub/winvn/readme.txt HTTP/1.0" 404 -
dynip38.efn.org - - [01/Jul/1995:00:10:50 -0400] "GET /software HTTP/1.0" 302 -
```

Now, let's talk about the Request field case. These are the ill-formatted lines we talked about before (2.2.1). The exact lines cannot match the regex **"""^.*"\w+\s+([^\s]+)\s*.*""".r.** The first 2 lines cannot be matched because the URL doesnt contain the http version. The pattern expectes a URI, a single space and the http version. Here in those 5 lines it doesn't exists, so thats why it doesn't match the pattern.

```
klothos.crl.research.digital.com - - [10/Jul/1995:16:45:50 -0400] "" 400 -
firewall.dfw.ibm.com - - [20/Jul/1995:07:34:34 -0400] "1/history/apollo/images/" 400 -
firewall.dfw.ibm.com - - [20/Jul/1995:07:53:24 -0400] "1/history/apollo/images/" 400 -
128.159.122.20 - - [20/Jul/1995:15:28:50 -0400] "k��tx��tG���t�" 400 -
128.159.122.20 - - [24/Jul/1995:13:52:50 -0400] "k��tx��tG���t�" 400 -
alyssa.p
```

2.3 Walk through on the Web Server Log File

2.3.1 Explore content size: Write in your report the min, max, and average content size.

Min: 0
Max: 6823936
Average: 20455.49857721692

2.3.2 HTTP status analysis: Write in your report the 100 most frequent status values and their frequencies.

(0,1)
(302,46573)
(400,5)
(304,132627)
(500,62)
(403,54)
(200,1701534)
(501,14)
(404,10845)s2qwda

2.3.3 Frequent hosts: Write in your report 10 hosts that accessed the server more than 10 times.

(piweba3y.prodigy.com ,17572)
(piweba4y.prodigy.com ,11591)
(piweba1y.prodigy.com ,9868)
(alyssa.prodigy.com ,7852)
(siltb10.orl.mmc.com ,7573)
(piweba2y.prodigy.com ,5922)
(edams.ksc.nasa.gov ,5434)
(163.206.89.4 ,4906)
(news.ti.com ,4863)
(disarray.demon.co.uk ,4353)

2.3.4 Top-10 error paths: Write in your report the top 10 requestURIs that did not have a return code of 200.

(/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)

2.3.5 Unique hosts: Write in your report how many unique there are in the entire log.

81983

2.3.6 Write in your report the count of 404 Response codes.

10845

2.3.7 Write in your report 40 distinct requestURIs that generate 404 errors.

(/pub/winvn/readme.txt,667)
(/pub/winvn/release.txt,547)
(/history/apollo/apollo-13.html,286)
(/shuttle/resources/orbiters/atlantis.gif,232)
(/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)
(/images/lf-logo.gif,66)
(/history/apollo/images/little-joe.jpg,66)
(/history/apollo/sa-2/sa-2-patch-small.gif,60)
(/robots.txt,59)

(/shuttle/missions/51-L/mission-51-1.html,58)
(/history/apollo/sa-9/sa-9-patch-small.gif,57)
(/persons/astronauts/a-to-d/conradCC.txt,52)
(/persons/astronauts/q-to-t/TrulyRH.txt,48)
(/history/apollo/pad-abort-test-1/images/,47)
(/pub,47)
(/shuttle/missions/mission.html,44)
(/history/apollo/sa-3/sa-3-patch-small.gif,42)
(/shuttle/missions/sts-XX/mission-sts-XX.html,41)
(/history/apollo/a-001/images/,40)
(/history/apollo/sa-5/sa-5-patch-small.gif,39)
(/history/apollo/apollo-13/apollo-13.html.,39)
(/:/spacelink.msfc.nasa.gov,38)
(/history/apollo/apollo13/apollo13.html,38)
(/history/apollo/apollo-13/apollo13.html,38)
(/people/nasa-cm/jmd.html,37)

2.3.8 Write in your report a list of the top twenty paths (in sorted order) that generate the most 404 errors.

(/pub/winvn/readme.txt,667)
(/pub/winvn/release.txt,547)
(/history/apollo/apollo-13.html,286)
(/shuttle/resources/orbiters/atlantis.gif,232)
(/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)

END