Explanation

yw3167 Yuting Wang hw1
email: yw3167@columbia.edu

Part 1

First read in the data through "textFile" function. In the epa-http.txt file, we select the first (IP address) and the last element (bytes) with "split" function.
The data contains: IP/Name [Timestamp] "Request" ReturnCode Bytes. Each of these elements are separated by space.
Use the command "reduceByKey (lambda x,y: x+y)" to apply on multiple partitions and returns final RDD with total counts paired with keys.

Part 2

Do part 1 again, and run "sortBy(lambda x:-x[1])" function to sort bytes from the largest one. In this way, the top-K IPs can be returned. To explain it clearer, we set K = 25 to see the result:

Part 3

The time stamp appears in the form of [DD:HH:MM:SS], which are day hour, month and second. In order to get the hour, we split by ':'. Here we compute counts of bytes per IP address per hour window from 00:00:00 to 00:59:59.

Part 4

With reference to announcement, we can aggregate IP address by the first two words. For example, "query2.lycos.cs.cmu.edu", can be aggregated as "query2.lycos.*.*". This is the situation for Name IP, which is not composed of digits. As for digit IP address, first three digits are aggregated.

Results:

Part 1:

For the sake of convenience, I just take first 25 elements:

```
[('141.243.1.172', 1634402), ('query2.lycos.cs.cmu.edu', 1325), ('140.112.68.165', 7811), ('dd15-
032.compuserve.com', 12898), ('freenet2.carleton.ca', 15173), ('ix-mia5-17.ix.netcom.com', 42461),
('hmu4.cs.auckland.ac.nz', 257009), ('131.215.67.47', 32988), ('www-c1.proxy.aol.com', 205533),
('bettong.client.uq.oz.au', 177058), ('flaxman-q950.uoregon.edu', 32988), ('161.122.12.78', 3008684),
('137.132.52.66', 2235), ('playful.mnsinc.com', 2121), ('archives.math.utk.edu', 0), ('ix-pen-nj1-
13.ix.netcom.com', 53615), ('141.243.1.174', 10739), ('www-d4.proxy.aol.com', 499876),
('butterfly.europa.com', 624), ('piweba4y.prodigy.com', 347370), ('cnts4p16.uwaterloo.ca', 3636398),
('138.25.148.25', 45491), ('204.188.47.212', 162121), ('s850.mwc.edu', 6574), ('piweba1y.prodigy.com',
38586)]
```

Part 2:

For the sake of convenience, I just take first 25 elements:

```
[('piankhi.cs.hamptonu.edu', 7267751), ('e659229.boeing.com', 5260561), ('139.121.98.45', 5041738),
('ws13dgadrv.er.usgs.gov', 4716720), ('slcmodem1-p1-14.intele.net', 4453807), ('www-c5.proxy.aol.com',
4435337), ('so.scsnet.com', 4420061), ('keyhole.es.dupont.com', 4005003), ('203.251.228.110', 3785626),
('cnts4p16.uwaterloo.ca', 3636398), ('rac3.wam.umd.edu', 3590760), ('dialin30.annex1.radix.net', 3405626),
('155.84.92.3', 3353172), ('198.102.67.27', 3182052), ('piweba5y.prodigy.com', 3084168), ('161.122.12.78',
3008684), ('net103-node5.dnr.state.mi.us', 2554562), ('sandy.rtptok1.epa.gov', 2486602),
('kenney.calvertgroup.com', 2486155), ('epsongw3.epson.co.jp', 2476446), ('svinet00.miti.go.jp', 2349975),
('wwwproxy.ac.il', 2114597), ('156.98.205.46', 1938034), ('yyj-ppp-13.cyberstore.ca', 1846850), ('yyj-ppp-
14.cyberstore.ca', 1814096)]
```

Part 3:

Total number of bytes in a time window of 1 hour:

```
[('ix-stp-fl1-20.ix.netcom.com', 684698), ('203.249.9.64', 11182), ('129.132.182.34', 97879), ('pc38-
c719.uibk.ac.at', 28796), ('ac203.du.pipex.com', 20231), ('gisws4.rtpnc.epa.gov', 0), ('202.244.226.77',
13205), ('164.155.1.107', 87349), ('dcoerr402.dcoerr4.epa.gov', 4184), ('nameless.house.gov', 52177),
('192.239.68.110', 11747), ('193.145.151.115', '4889'), ('spanky.comm.hq.af.mil', 11747), ('141.243.1.188',
11747), ('147.46.54.150', 53504), ('wicdgserv.wic.epa.gov', 24191), ('jmozingo.ehsg.saic.com', 4150),
('bamwoyaj.dart.ns.doe.ca', 49950), ('wwwproxy.ac.il', 51045), ('dcoerr435.dcoerr4.epa.gov', '3217'), ('h-
ecru.richmond.infi.net', 92453), ('ece04p.kuec.kyoto-u.ac.jp', 10792), ('194.20.235.12', 57182),
('poppy.hensa.ac.uk', 9873), ('141.243.1.184', 7513), ('pc04.phs.ed.ac.uk', '4889'), ('157.190.64.20', 18945),
('193.78.67.34', 6085), ('acquinas.netmind.com', '3495')]
```

Part 4:

For the sake of convenience, I just take first 25 elements:

```
[('140.112.68', 7811), ('202.32.50', 159922), ('149.159.22', 67951), ('198.69.241', 34462), ('202.96.29',
10296)]
[('query2.lycos', '1325'), ('tanuki.twics', 60936), ('dd15-032.compuserve', 12898), ('freenet2.carleton',
'15173'), ('ix-knx-tn1-22.ix', 14450), ('suburbia.apana', 14005), ('port11.annex1', 121697), ('www-c1.proxy',
205533), ('grenada.assist', 15571), ('ix-tf8-26.ix', 90384), ('charlotte.anu', 12498), ('p33.denver1', 14964),
('bettong.client', 177058), ('cragateway.cra', 48978), ('systems61.fisher', '4889'), ('ch-a1413.scu', '10788'),
('ad21-005.compuserve', 29979), ('archives.math', 0), ('www-d4.proxy', 499876), ('piweba4y.prodigy',
347370), ('sfsp03.slip', 191004), ('cnts4p16.uwaterloo', 3636398), ('dial01.wwrdc', '5616'), ('mizzou-ts7-
16.missouri', 12275), ('s850.mwc', 6574)]
```