

Part b

- 2.1 Enumerate all **HTTP status codes** and give counts of each

The code is in mapper1.py and reducer1.py.

```
[training@localhost python-example]$ hs mapper1.py reducer1.py /user/csds/input/july.txt /user/csds/outputpy2
packageJobJar: [mapper1.py, reducer1.py, /tmp/hadoop-training/hadoop-unjar4606683547414743562/] [] /tmp/streamjob7002460207554099582.jar tmpDir=null
17/03/30 04:55:24 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
17/03/30 04:55:25 WARN snappy.LoadSnappy: Snappy native library is available
17/03/30 04:55:25 INFO snappy.LoadSnappy: Snappy native library loaded
17/03/30 04:55:25 INFO mapred.FileInputFormat: Total input paths to process : 1
17/03/30 04:55:25 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local]
17/03/30 04:55:25 INFO streaming.StreamJob: Running job: job_201703300353_0005
17/03/30 04:55:25 INFO streaming.StreamJob: To kill this job, run:
17/03/30 04:55:25 INFO streaming.StreamJob: UNDEF/bin/hadoop job -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201703300353_0005
17/03/30 04:55:25 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201703300353_0005
17/03/30 04:55:26 INFO streaming.StreamJob: map 0% reduce 0%
17/03/30 04:55:39 INFO streaming.StreamJob: map 2% reduce 0%
17/03/30 04:55:42 INFO streaming.StreamJob: map 4% reduce 0%
17/03/30 04:55:45 INFO streaming.StreamJob: map 6% reduce 0%
17/03/30 04:55:48 INFO streaming.StreamJob: map 7% reduce 0%
```

```
[training@localhost python-example]$ hadoop fs -ls /user/csds/outputpy2
Found 3 items
-rw-r--r-- 1 training supergroup 0 2017-03-30 05:00 /user/csds/outputpy2/SUCCESS
drwxr-xr-x - training supergroup 0 2017-03-30 04:55 /user/csds/outputpy2/logs
-rw-r--r-- 1 training supergroup 99 2017-03-30 05:00 /user/csds/outputpy2/part-000000
[training@localhost python-example]$ hadoop fs -cat /user/csds/outputpy2/part-000000
200 - 9301566
302 - 219210
304 - 800316
400 - 45
403 - 675
404 - 62700
500 - 195
501 - 123
[training@localhost python-example]$
```

- 2.2 Find the total bandwidth that was sent by the Nasa webserver in the month of July 1995.

If we consider that all the entries in the log are served by the nasa server.

The code is in mapper2.py and reducer2.py.

The bandwidth we got is 77391946982.

```

[training@localhost python-example]$ hs mapper2.py reducer2.py /user/csds/input/july.
txt /user/csds/outputpy3
packageJobJar: [mapper2.py, reducer2.py, /tmp/hadoop-training/hadoop-unjar140927936081
0604292/] [] /tmp/streamjob1399498729676833174.jar tmpDir=null
17/03/30 05:45:03 WARN mapred.JobClient: Use GenericOptionsParser for parsing the argu
ments. Applications should implement Tool for the same.
17/03/30 05:45:03 WARN snappy.LoadSnappy: Snappy native library is available
17/03/30 05:45:03 INFO snappy.LoadSnappy: Snappy native library loaded
17/03/30 05:45:03 INFO mapred.FileInputFormat: Total input paths to process : 1
17/03/30 05:45:04 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cach
e/training/mapred/local]
17/03/30 05:45:04 INFO streaming.StreamJob: Running job: job_201703300353_0007
17/03/30 05:45:04 INFO streaming.StreamJob: To kill this job, run:
17/03/30 05:45:04 INFO streaming.StreamJob: UNDEF/bin/hadoop job -Dmapred.job.tracker
=0.0.0.0:8021 -kill job_201703300353_0007
17/03/30 05:45:04 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdeta
ils.jsp?jobid=job_201703300353_0007
17/03/30 05:45:05 INFO streaming.StreamJob: map 0% reduce 0%
17/03/30 05:45:18 INFO streaming.StreamJob: map 2% reduce 0%
17/03/30 05:45:22 INFO streaming.StreamJob: map 3% reduce 0%
17/03/30 05:45:25 INFO streaming.StreamJob: map 5% reduce 0%

17/03/30 05:48:38 INFO streaming.StreamJob: map 89% reduce 27%
17/03/30 05:48:39 INFO streaming.StreamJob: map 90% reduce 27%
17/03/30 05:48:41 INFO streaming.StreamJob: map 91% reduce 27%
17/03/30 05:48:42 INFO streaming.StreamJob: map 93% reduce 27%
17/03/30 05:48:44 INFO streaming.StreamJob: map 94% reduce 27%
17/03/30 05:48:45 INFO streaming.StreamJob: map 94% reduce 31%
17/03/30 05:48:52 INFO streaming.StreamJob: map 99% reduce 31%
17/03/30 05:48:53 INFO streaming.StreamJob: map 100% reduce 31%
17/03/30 05:48:54 INFO streaming.StreamJob: map 100% reduce 33%
17/03/30 05:48:57 INFO streaming.StreamJob: map 100% reduce 67%
17/03/30 05:49:07 INFO streaming.StreamJob: map 100% reduce 100%
17/03/30 05:49:10 INFO streaming.StreamJob: Job complete: job_201703300353_0007
17/03/30 05:49:10 INFO streaming.StreamJob: Output: /user/csds/outputpy3
[training@localhost python-example]$ hadoop fs -cat /user/csds/outputpy3/part-00000
bandwidth 77391946982
[training@localhost python-example]$

```