**Assignment on Spark**

Here is an extract of what the GHTorrent log looks like:

DEBUG, 2017-03-23T10:02:27+00:00, ghtorrent-40 -- ghtorrent.rb: Repo EFForg/https-everywhere exists

DEBUG, 2017-03-24T12:06:23+00:00, ghtorrent-49 -- ghtorrent.rb: Repo Shikanime/print exists

INFO, 2017-03-23T13:00:55+00:00, ghtorrent-42 -- api\_client.rb: Successful request. URL: https://api.github.com/repos/CanonicalLtd/maas-docs/issues/365/events?per\_page=100, Remaining: 4943, Total: 88 ms

WARN, 2017-03-23T20:04:28+00:00, ghtorrent-13 -- api\_client.rb: Failed request. URL: https://api.github.com/repos/greatfakeman/Tabchi/commits?sha=Tabchi&per\_page=100, Status code: 404, Status: Not Found, Access: ac6168f8776, IP: 0.0.0.0, Remaining: 3031

DEBUG, 2017-03-23T09:06:09+00:00, ghtorrent-2 -- ghtorrent.rb: Transaction committed (11 ms)

Each log line comprises of a standard part (up to .rb:) and an operation-specific part. The standard part fields are like so:

1. Logging level, one of DEBUG, INFO, WARN, ERROR (separated by ,)
2. A timestamp (separated by ,)
3. The downloader id, denoting the downloader instance (separated by --)
4. The retrieval stage, denoted by the Ruby class name, one of:
   * event\_processing
   * ght\_data\_retrieval
   * api\_client
   * retriever
   * ghtorrent

Questions:

1. Write a function to load it in an RDD.
2. How many lines does the RDD contain?
3. Count the number of WARNing messages
4. How many repositories where processed in total? Use the api\_client lines only.
5. Which client did most HTTP requests?
6. Which client did most FAILED HTTP requests? Use group\_by to provide an answer.
7. What is the most active hour of day?
8. What is the most active repository (**hint**: use messages from the ghtorrent.rb layer only)?