BDA
Assignment 1

Aakash Tanwar 2016215

Priya Rajpurohit 2015073

Download the log file and write a function to load it in PostgreSQL. Create your own schema for importing the data.

```
import psycopg2
conn = psycopg2.connect("host=localhost dbname=postgres user=postgres password=priya123")
cur = conn.cursor()



# try:
if(1):
   connection = psycopg2.connect("dbname=postgres user=postgres password=priya123")
   print(connection)
   cursor = connection.cursor()

   cursor.execute("""CREATE TABLE ghtorrent
    (
    sno int PRIMARY KEY,
    log_level text,
    date_time text,
    downloader_id text,
    ret_stage text,
    information text,
    repo text,
    access_key text
    );""")

   postgres_insert_query = """ INSERT INTO ghtorrent (sno,log_level,date_time,downloader_id,ret_stage,information, repo, access_key) VALUES (%s,%s,%s,%s,%s,%s, %s, %s)
   all_data=[]
   with open('ghtorrent-logs.csv', 'r') as f:
       ##next(f)
       for content in f:
           #content[len(content)-1]=""
           x=content.strip().split(",")

           all_data.append(x)
   #print(all_data)



   cursor.executemany(postgres_insert_query,all_data)
   count = cursor.rowcount
   print (count, "Record inserted successfully into table")
   connection.commit()
   cursor.close()
   connection.close()
```

How many records does the table contain?

```
ghtorrent=# select count(*) from ghtorrent ;
   count
---------
 9669634
(1 row)
```

Count the number of WARNing messages.

```
ghtorrent=# SELECT COUNT(*) FROM ghtorrent WHERE log_level='WARN';
 count
--------
 132158
(1 row)
```

How many repositories were processed in total?

```
ghtorrent=# SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE repo <>'' AND information LIKE '%URL%repos%/%?%' AND ret_stage='api_client.rb';
 count
--------
 381194
(1 row)
```

Which 10 clients did the highest HTTP requests?

```
ghtorrent=# SELECT downloader_id,COUNT(*) FROM ghtorrent WHERE information LIKE '%URL: https%' GROUP BY downloader_id ORDER BY COUNT(*) DESC LIMIT 10;
 downloader_id | count
---------------+-------
 ghtorrent-13  | 85528
 ghtorrent-4   | 19046
 ghtorrent-18  | 18948
 ghtorrent-10  | 18926
 ghtorrent-40  | 18911
 ghtorrent-39  | 18616
 ghtorrent-38  | 18614
 ghtorrent-47  | 18604
 ghtorrent-1   | 18463
 ghtorrent-24  | 18452
(10 rows)
```

Which 10 client did the highest FAILED HTTP requests? (Operation part starts with the string ”Failed”)

```
ghtorrent=# SELECT downloader_id,COUNT(*) FROM ghtorrent WHERE information LIKE '%Failed%URL: https%' GROUP BY downloader_id ORDER BY COUNT(*) DESC LIMIT 10;
 downloader_id | count
---------------+-------
 ghtorrent-13  | 79623
 ghtorrent-21  |  1378
 ghtorrent-40  |  1134
 ghtorrent-18  |   368
 ghtorrent-42  |   357
 ghtorrent-9   |   356
 ghtorrent-4   |   352
 ghtorrent-25  |   342
 ghtorrent-22  |   333
 ghtorrent-6   |   332
(10 rows)
```

What is the most active hour of day?

```
ghtorrent=# SELECT hour,COUNT(*) FROM ghtorrent GROUP BY hour ORDER BY COUNT(*) DESC LIMIT 1;
 hour |  count
------+---------
 10   | 2662487
(1 row)
```

What is the most active repository?

```
ghtorrent=# SELECT repo,COUNT(*) FROM ghtorrent WHERE repo <> '' GROUP BY repo ORDER BY COUNT(*) DESC LIMIT 1;
            repo              | count
-----------------------------+-------
 greatfakeman/Tabchi/commits | 79523
(1 row)
```

Which access keys are failing most often?

```
ghtorrent=# SELECT access_key,COUNT(*) FROM ghtorrent WHERE access_key <> '' AND information LIKE '%Failed%' GROUP BY access_key ORDER BY COUNT(*) DESC LIMIT 1;
 access_key  | count
-------------+-------
 ac6168f8776 | 79623
(1 row)
```

Compute the number of different repositories accessed by the client ghtorrent-22 .Note the time taken by your query.

```
ghtorrent=# EXPLAIN ANALYZE SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE downloader_id='ghtorrent-22';
                                                                    QUERY PLAN
-----------------------------------------------------------------------------------------------------------------------------------------------------------
 Aggregate  (cost=264080.09..264080.10 rows=1 width=8) (actual time=4888.216..4888.216 rows=1 loops=1)
   ->  Gather  (cost=5626.82..263581.30 rows=199517 width=4) (actual time=132.823..4469.485 rows=193876 loops=1)
         Workers Planned: 2
         Workers Launched: 2
         ->  Parallel Bitmap Heap Scan on ghtorrent  (cost=4626.82..242629.60 rows=83132 width=4) (actual time=122.462..4507.455 rows=64625 loops=3)
               Recheck Cond: (downloader_id = 'ghtorrent-22'::text)
               Rows Removed by Index Recheck: 1073798
               Heap Blocks: exact=16625 lossy=20778
               ->  Bitmap Index Scan on idx  (cost=0.00..4576.94 rows=199517 width=0) (actual time=107.499..107.499 rows=193876 loops=1)
                     Index Cond: (downloader_id = 'ghtorrent-22'::text)
 Planning Time: 0.828 ms
 Execution Time: 4890.962 ms
(12 rows)
```

```
ghtorrent=# CREATE INDEX idx on ghtorrent(downloader_id);
CREATE INDEX
ghtorrent=# SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE downloader_id='ghtorrent-22';
 count
-------
  9041
(1 row)
```

Now drop your index and compute the number of different repositories accessed by the client ghtorrent-22. Note the time taken by your query now.

```
ghtorrent=# DROP INDEX idx;
DROP INDEX
ghtorrent=# EXPLAIN ANALYZE SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE downloader_id='ghtorrent-22';
                                                                    QUERY PLAN
-----------------------------------------------------------------------------------------------------------------------------------
 Aggregate  (cost=264335.17..264335.18 rows=1 width=8) (actual time=5559.457..5559.458 rows=1 loops=1)
   ->  Gather  (cost=1000.00..263836.38 rows=199517 width=4) (actual time=0.732..5144.273 rows=193876 loops=1)
         Workers Planned: 2
         Workers Launched: 2
         ->  Parallel Seq Scan on ghtorrent  (cost=0.00..242884.68 rows=83132 width=4) (actual time=0.234..5191.243 rows=64625 loops=3)
               Filter: (downloader_id = 'ghtorrent-22'::text)
               Rows Removed by Filter: 3158586
 Planning Time: 32.303 ms
 Execution Time: 5559.601 ms
(9 rows)
```

```
ghtorrent=# DROP INDEX idx;
DROP INDEX
ghtorrent=# SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE downloader_id='ghtorrent-22';
 count
-------
  9041
(1 row)
```

Read in the CSV file into another table call it interesting. How many records are there?

```
lightorrent=# SELECT COUNT(*) FROM interesting;
 count
-------
  1435
(1 row)
```

How many records in the log file refer to entries in the interesting file?

```
ghtorrent=# SELECT COUNT(*) FROM ghtorrent JOIN interesting ON repo=url;
 count
-------
   535
(1 row)
```

Which of the interesting repositories has the most failed API calls?

```
ghtorrent=# DROP INDEX idx;
DROP INDEX
ghtorrent=# SELECT COUNT(DISTINCT(repo)) FROM ghtorrent WHERE downloader_id='ghtorrent-22';
 count
-------
  9041
(1 row)
```

Resources Used

PostgreSQL Documentation
Postgres.app documentation
Lecture Slides