Raj Kamal Yadav
2016076

Rajat Bansal
2016260

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Data Format

- id serial Primary Key Integer
- log_lvl VARCHAR
- timestamp VARCHAR
- dwnlder_id VARCHAR
- retrval_stage VARCHAR
- repo VARCHAR
- mssg VARCHAR

Preprocessing:

1. The Log provided is not uniformly delimited.
2. Logging level, one of DEBUG, INFO, WARN, ERROR (separated by ,)
3. Timestamp (separated by ,)
4. Downloader id, denoting the downloader instance (separated by --)
5. Retrieval stage, denoted by the Ruby class name, one of : event_processing,ght_data_retrieval,api_client retriever and ghtorrent.
6. Java JDBC with PSQL driver is used for preprocess and creating connection to PSQL Server.
7. Data insertion is done in Batches of Million to make the process fast and restrict the Heap Space from running out.

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Loading Into Database

```
public static void main(String[] args) {
        System.out.println(getrepo("INFO, 2017-03-23T11:12:36+00:00, gbtorrent-32 -- api_client.rb: Successful request. URL: https://api.github.com/repos/
    try {
        Class.forName("org.postgresql.Driver");
        Connection con=DriverManager.getConnection( s: "jdbc:postgresql://localhost:5432/postgres", s1: "postgres",
                s2: "abcdefg");
        PreparedStatement ps1=con.prepareStatement( s: "CREATE TABLE IF NOT EXISTS schema_bda_1.mytable (id serial " +
                "PRIMARY KEY,log_lvl varchar,timestamp varchar,dweldor_id varchar,retrval_stage varchar, repo " +
                "varchar, msgg " +
                "varchar);");
        ps1.executeQuery();

    }
    catch(Exception ex){
        System.out.println(ex.getMessage());
    }
///
    try {
        System.out.println("read");
        BufferedReader br = new BufferedReader(new FileReader( s: "/home/bansal/Downloads/a.txt"));
        String line;
        Class.forName("org.postgresql.Driver");
        Connection con=DriverManager.getConnection( s: "jdbc:postgresql://localhost:5432/postgres", s1: "postgres",
                s2: "abcdefg");
        PreparedStatement ps2= con.prepareStatement( s: "INSERT INTO schema_bda_1.mytable(\"log_lvl\",\"timestamp\"," +
                "\"dweldor_id\",\"retrval_stage\", \"repo\", \"msgg\") VALUES (?, ?, ?, ?, ?, ?);");
        int count= 0;
        List<String> list;
        while ((line = br.readLine()) != null) {
            list = new ArrayList<>();
            if(count>= 10000000){
                break;
            }
            if(count>= 9000000){
                list= getItems(line);
            }
            if(count%1000000== 0){
                System.out.println("whoops");
            }
            if (!list.isEmpty()) {
                ps2.setString( i: 1, list.get(0));
                ps2.setString( i: 2, list.get(1));
                ps2.setString( i: 3, list.get(2));
                ps2.setString( i: 4, list.get(3));
                ps2.setString( i: 5, getrepo(list.get(4)));
                ps2.setString( i: 6, list.get(4));

                ps2.addBatch();
            }
```

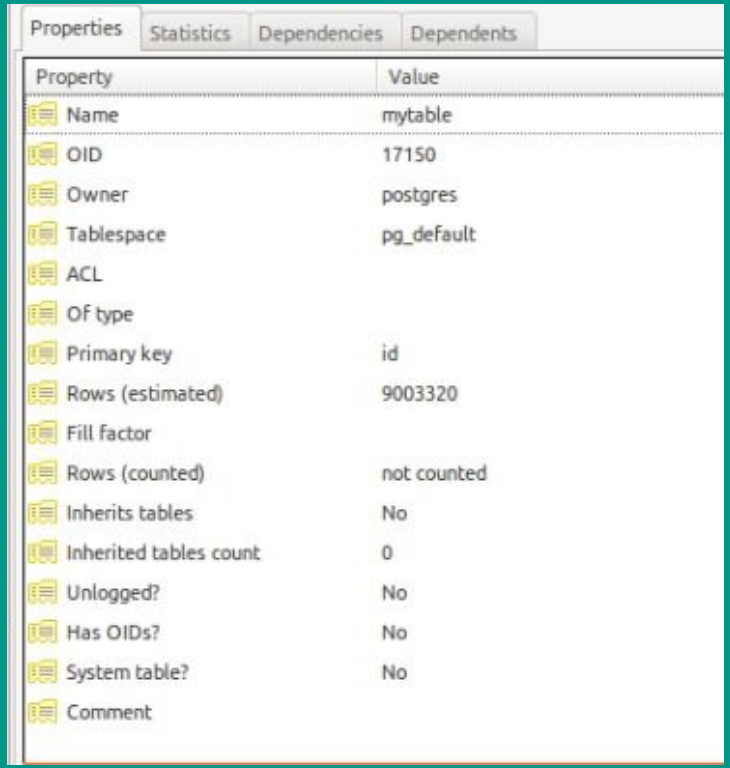
Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Schema

```
CREATE TABLE schemma bda 1.mytable
(
    id serial NOT NULL,
    log lvl character varying,
    "timestamp" character varying,
    dwnlder id character varying,
    retrval stage character varying,
    repo character varying,
    mssg character varying,
    CONSTRAINT mytable pkey PRIMARY KEY (id)
)
```



Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

How many records does the table contain?

Any records(154) with any attribute value as NULL is dropped.

Query

```
SELECT COUNT(*) FROM schemma_bda_1.mytable
```

8:16 PM

Output

Data Output | Explain | Messages | History

	count bigint
1	9669634

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

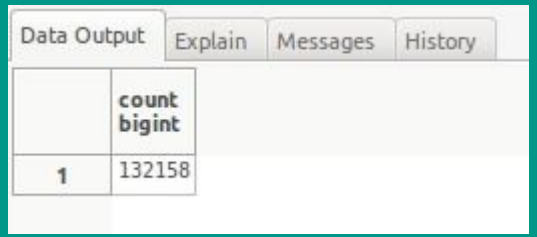
Count the number of WARNing messages.

Query

```
SELECT COUNT(*) FROM schemma_bda_1.mytable WHERE log_lvl= 'WARN'
```

8:16 PM

Output



Data Output | Explain | Messages | History

	count bigint
1	132158

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Processed Repositories with api_client.

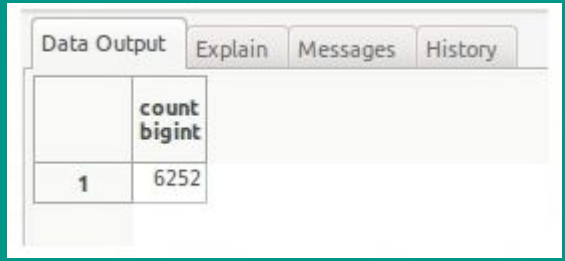
The repo column contains the name of the repositories as “Name of the repository + Name of the user” just as Github.

Query

```
Question4.
select COUNT(DISTINCT repo)
from schemma_bda_1.mytable
where retrval_stage= ' api_client.rb' and log_lvl= 'WARN'
```

8:13 PM

Output



Data Output | Explain | Messages | History

	count bigint
1	6252

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Which 10 clients did the highest HTTP requests?

For HTTP matchcase the string “https://adi.github.com” is used.

Query

```
Select dwnlder_id, count(*) as c
from schemma_bda_1.mytable
where mssg like '%https://api.github.com%'
group by dwnlder_id
order by c
desc
limit 10;
```

8:04 PM

Output

Data Output | Explain | Messages | History

	dwnlder_id character varying	c bigint
1	ghtorrent-13	85528
2	ghtorrent-4	19046
3	ghtorrent-18	18950
4	ghtorrent-10	18926
5	ghtorrent-40	18911
6	ghtorrent-39	18616
7	ghtorrent-38	18614
8	ghtorrent-47	18605
9	ghtorrent-1	18465
10	ghtorrent-24	18452

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Which 10 client did the highest FAILED HTTP requests?

Query

```
Select dwnlder_id, count(*) as c
from schemma_bda_1.mytable
where mssg like '_Failed%'
group by dwnlder_id
order by c
desc
limit 10;
```

8:02 PM

Output

Data Output | Explain | Messages | History

	dwnlder_id character varying	c bigint
1	ghtorrent-13	79623
2	ghtorrent-21	1378
3	ghtorrent-40	1134
4	ghtorrent-18	368
5	ghtorrent-42	357
6	ghtorrent-9	356
7	ghtorrent-4	352
8	ghtorrent-25	342
9	ghtorrent-22	333
10	ghtorrent-6	332

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

What is the most active hour of day?

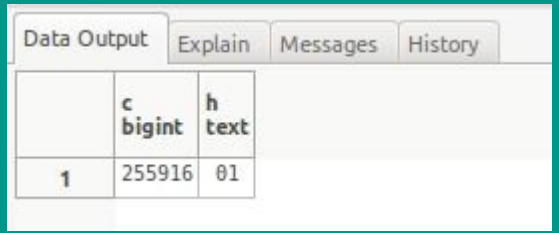
The timestamp is include without considering the time zone mentioned in the raw data but instead all the timing is brought into the same time zone.

Query

```
SELECT COUNT(*) as C, substring(tb.timestamp from 11 for 3) as H
FROM schemma_bda_1.mytable as tb
WHERE mssg like '%https://%'
GROUP BY H
ORDER BY C
DESC LIMIT 1
```

8:01 PM

Output



Data Output | Explain | Messages | History

	c bigint	h text
1	255916	01

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

What is the most active repository?

Query

```
select count(*) as c, repo
from schemma_bda_1.mytable
where (mssg like '%api.github.com/repos/%')
group by repo
order by c
desc
limit 5
```

7:58 PM

Output

Data Output | Explain | Messages | History

	c bigint	repo character varying
1	79524	greatfakeman/Tabchi
2	4084	mithro/chromium-infra
3	2575	shuhongwu/hockeyapp
4	2299	obophenotype/human-phenotype-ontology
5	1149	kubernetes/kubernetes

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Which access keys are failing most often?

Query

```
select substring(tb.mssg for 11 from position('Access' in tb.mssg)+8 ) as
fstrng,count(*) as cnt
from schemma_bda_1.mytable as tb
where mssg like '%Access:%' and mssg like '%Failed request.%'
group by fstrng
order by cnt
desc
limit 1
```

7:46 PM

Output

Data Output | Explain | Messages | History

	fstrng text	cnt bigint
1	ac6168f8776	79623

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

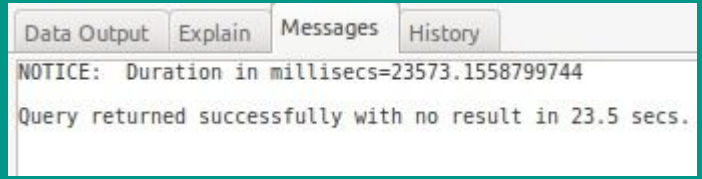
No. of different repositories accessed by ghtorrent-22. (without indexing)

Query

```
DO $proc$
DECLARE
  StartTime timestamptz;
  EndTime timestamptz;
  Delta double precision;
BEGIN
  StartTime := clock_timestamp();
  perform distinct count(*) as M
  from schemma_bda_1.mytable
  where dwnlder_id =' ghtorrent-22';
  EndTime := clock_timestamp();
  Delta := 1000 * ( extract(epoch from EndTime) - extract(epoch from StartTime)
);
  RAISE NOTICE 'Duration in millisecs=%', Delta;
END;
$proc$;
```

7:42 PM

Output



Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

No. of different repositories accessed by ghtorrent-22. (with indexing)

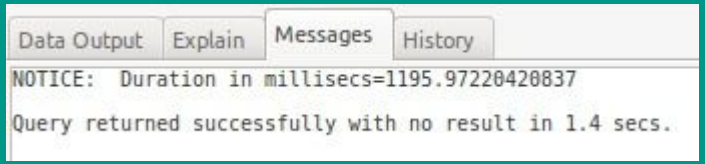
Query

```
CREATE INDEX downindex on TABLE schemma_bda_1.tab(dwnlder_id);

DO $proc$
DECLARE
  StartTime timestamptz;
  EndTime timestamptz;
  Delta double precision;
BEGIN
  StartTime := clock_timestamp();
  perform distinct count(*) as M
  from schemma_bda_1.tab
  where dwnlder_id =' ghtorrent-22';
  EndTime := clock_timestamp();
  Delta := 1000 * ( extract(epoch from EndTime) - extract(epoch from StartTime)
);
  RAISE NOTICE 'Duration in millisecs=%', Delta;
END;
$proc$;
```

7:46 PM

Output



Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Loading Interesting CSV to Database and reporting records.

Query

```
CREATE TABLE schemma_bda_1.interesting
(
  id integer,
  url character varying,
  owner_id integer,
  nme character varying,
  lang character varying,
  created_at character varying,
  forked_from character varying,
  deleted integer,
  updated_at character varying
)
```

```
COPY schemma_dw_1.interesting FROM '/tmp/important-repos.csv'
(DELIMITER(','));
```

Output

Count: 1435 Rows

Name	interesting
OID	17129
Owner	postgres
Tablespace	pg_default
ACL	
Of type	
Primary key	<no primary key>
Rows (estimated)	1435
Fill factor	
Rows (counted)	1435
Inherits tables	No
Inherited tables count	0
Unlogged?	No
Has OIDs?	No
System table?	No
Comment	

Raj Kamal Yadav

Rajat Bansal

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

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

Query

```
select count(*) from
(select substring(repo from Position('/' in repo) +1) as repp
from schemma_bda_1.mytable
where (mssg like '%api.github.com/repos/%')) as A inner join
schemma_bda_1.interesting as B
on A.repp = B.nme
```

Output

Data Output | Explain | Messages | History

	count bigint
1	87939

Raj Kamal Yadav
2016076

Rajat Bansal
2016260

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

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

Query

```
select count(*) as c, substring(M.repo from Position('/' in repo)+1) as repp
from schemma_bda_1.interesting as I, schemma_bda_1.mytable as M
where M.mssg like '_Failed%' and I.nme= substring(M.repo from Position('/' in
repo)+1)
group by repp
order by c
desc
limit 10
```

8:25 PM

Output

Data Output | Explain | Messages | History

	c bigint	repp text
1	740	hello-world
2	309	test
3	166	demo
4	88	Test
5	47	-
6	26	hello
7	24	Ruby_k59
8	20	website
9	16	TestRepo
10	15	angular

Resources Used

StackOverFlow for How to measure the time taken for the execution of the Query.