IT350 Assignment 3

NAME: SUYASH CHINTAWAR
ROLL NO.: 191IT109
TOPIC: STREAMING DATA
VISUALIZATIONS

Note:
1) The colab link has been attached below. After opening the link, if it opens in drive, click on “Open with Google Colaboratory” to view the complete code.
2) Only output screenshots have been attached along with the explanation. Code for the same can be found in the colab notebook.

Colab notebook link:
https://colab.research.google.com/drive/1ol9pYX1ADGWYrdYVvAV1-BiDfEzFWm47

Q1. Use the allocated data source (WikiLogs API) and build a utility to extract and curate data for the analytics tasks specified in part 2. Ensure that this module provides a Live Stream of data.

The EventStreams API has been used in this assignment to fulfill the task of obtaining live stream data. The link for the same can be found below,

https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams

This API gives the recent changes made by users across the world across all wiki servers as live stream data in the json object format.
Here, the data is extracted every 10 seconds. A total of 200 recent changes are extracted from the API call in one go. 200 logs are extracted because it takes around 7-8 seconds to get the required amount of data. A 2-3 second gap favors non-overlapping data in the next API call.

An example of the obtained json object is shown below,

```
Example of the data:
{
'$schema': '/mediawiki/recentchange/1.0.0',
'meta': {'uri': 'https://www.wikidata.org/wiki/Q1408',
         'request_id':
         '65cf549f-8edf-48e1-8135-37d0c883d445', 'id':
         '7d4ef3b9-d35c-42ea-9a45-08918deeface', 'dt':
         '2022-03-07T18:28:20Z', 'domain':
         'www.wikidata.org',
          'stream': 'mediawiki.recentchange', 'topic':
          'eqiad.mediawiki.recentchange', 'partition': 0,
          'offset': 3694999547},
'id': 1639394709,
'type': 'edit',
'namespace': 0,
'title': 'Q1408',
'comment': '/* wbeditentity-update:0| */',
'timestamp': 1646677700,
'user': 'Xatnys42',
'bot': False,
'minor': False,
'patrolled': True,
'length': {'old': 308692, 'new': 309040},
'revision': {'old': 1589194897, 'new': 1589213472},
'server_url': 'https://www.wikidata.org',
'server_name': 'www.wikidata.org',
'server_script_path': '/w',
'wiki': 'wikidatawiki',
'parsedcomment': '\u200e<span dir="auto"><span
                   class="autocomment">Updated
                   Item</span></span>'
}
```

Q2. Using the data collected as part of question 1, perform a data analytics task.

The count of the type of log and the count of the server on which an action was made are the two statistics that are generated from the data.

<u>Output Screenshot:</u>

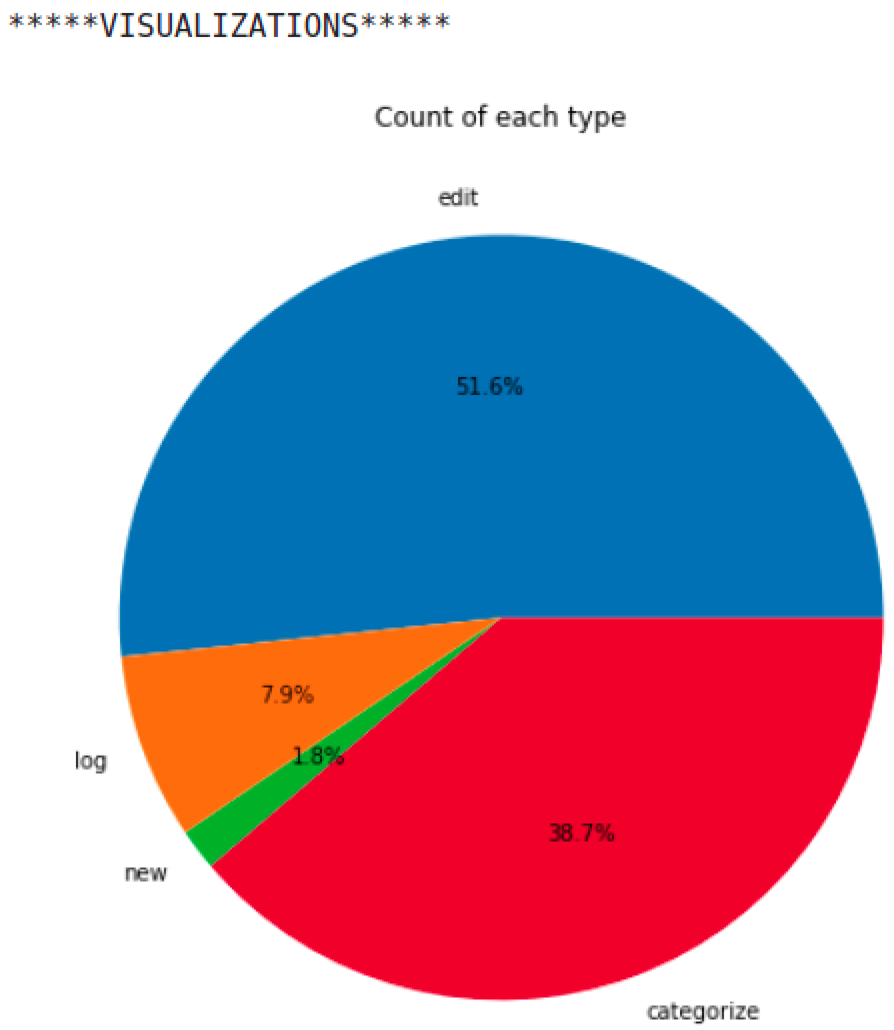
```
*****STATISTICS*****

1. COUNT OF LOG TYPES
edit : 413
log : 63
new : 14
categorize : 310

2. COUNT OF SERVERS
www.wikidata.org : 164
fr.wikipedia.org : 50
en.wikipedia.org : 106
fr.wiktionary.org : 10
zh.wikipedia.org : 2
ko.wikipedia.org : 15
de.wikipedia.org : 14
commons.wikimedia.org : 276
meta.wikimedia.org : 1
it.wikipedia.org : 20
en.wiktionary.org : 3
simple.wikipedia.org : 4
th.wikipedia.org : 7
de.wiktionary.org : 6
es.wikipedia.org : 13
uk.wikipedia.org : 4
ca.wikipedia.org : 3
beta.wikiversity.org : 1
ru.wikipedia.org : 24
lt.wikipedia.org : 2
tr.wikipedia.org : 2
ja.wikipedia.org : 1
sr.wikipedia.org : 1
en.wikisource.org : 2
cs.wikipedia.org : 1
no.wikipedia.org : 1
is.wikipedia.org : 1
hr.wikipedia.org : 1
hy.wikipedia.org : 3
ar.wikipedia.org : 10
```

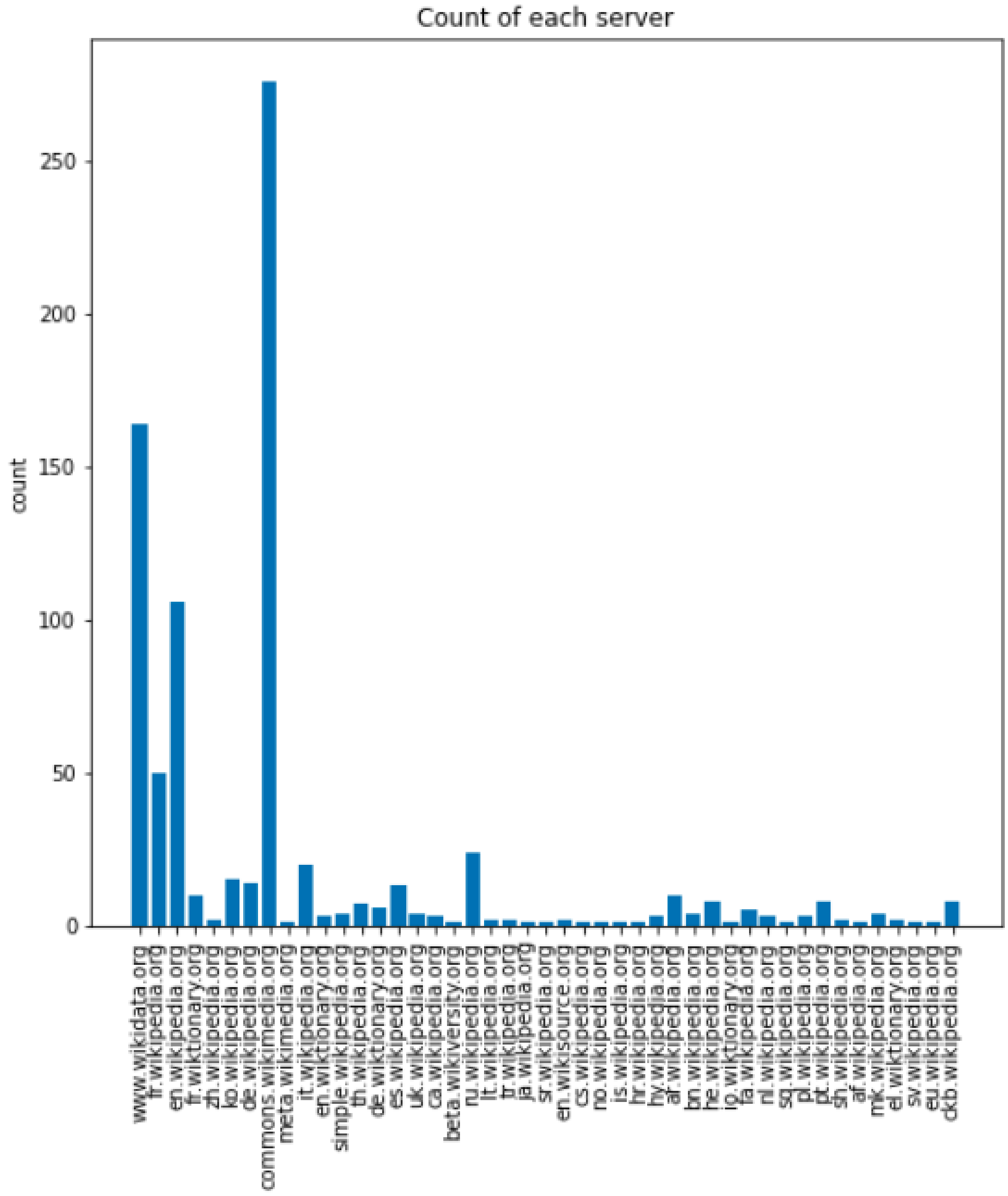
Q3. Finally, build a visualization module for data obtained from the task carried out in part 2 so that any changes in the stream are reflected in the visualization.

Output Screenshots:



```
*****VISUALIZATIONS*****
```

The pie chart shows the percentage of each type of log obtained through the data. We can see that 51% of the actions are edit actions followed by categorize action in the data obtained.



The count of each server on which an action was made is plotted in the plot above. Here, we see that the most number of actions made was on the server with the name `‘commons.wikimedia.org’` with 250+ actions in 800 logs.

THANK YOU