

Bash script:

```
#!/usr/bin/env bash

apt-get update
apt-get install python3 python3-pip
pip3 install dask
pip3 install pandas
pip3 install fastparquet

gsutil cp gs://pjwstk-bigdata/0.parquet.gzip data.00.parquet.gzip
gsutil cp gs://pjwstk-bigdata/1.parquet.gzip data.01.parquet.gzip
gsutil cp gs://pjwstk-bigdata/2.parquet.gzip data.02.parquet.gzip
gsutil cp gs://pjwstk-bigdata/3.parquet.gzip data.03.parquet.gzip
gsutil cp gs://pjwstk-bigdata/4.parquet.gzip data.04.parquet.gzip
gsutil cp gs://pjwstk-bigdata/5.parquet.gzip data.05.parquet.gzip
gsutil cp gs://pjwstk-bigdata/6.parquet.gzip data.06.parquet.gzip
gsutil cp gs://pjwstk-bigdata/7.parquet.gzip data.07.parquet.gzip
gsutil cp gs://pjwstk-bigdata/8.parquet.gzip data.08.parquet.gzip
gsutil cp gs://pjwstk-bigdata/9.parquet.gzip data.09.parquet.gzip
gsutil cp gs://pjwstk-bigdata/10.parquet.gzip data.10.parquet.gzip
```

Python script:

```
from collections import defaultdict
import dask.dataframe as dd
import time
import pandas as pd

start_time = time.time()

# ./data.01.parquet is broken i guess
files = ['./data.00.parquet.gzip', './data.02.parquet.gzip',
          './data.03.parquet.gzip',
          './data.04.parquet.gzip', './data.05.parquet.gzip',
          './data.06.parquet.gzip',
          './data.07.parquet.gzip', './data.08.parquet.gzip',
          './data.09.parquet.gzip', './data.10.parquet.gzip']

df = dd.read_parquet(files)

repos = df["repo_name"].compute()

d = defaultdict(int)
for set in repos:
    for repo in set:
        d[repo] += 1


d = dict(sorted(d.items(), key=lambda x: x[1]))

df2 = pd.DataFrame.from_dict(d, orient='index')
print(df2)
print(time.time() - start_time, "seconds")
```

Machine 1:


1	38.43
2	37.67
3	37.07
4	36.93
5	36.71
6	37.02
7	37.91
8	37.19
9	38.03
10	37.72
	37.468

Machine configuration

Machine type	e2-medium
CPU platform	Intel Broadwell
vCPUs to core ratio 	—
Display device	Disabled Enable to use screen capturing and recording tools
GPUs	None

Storage

Boot disk

Name 	Image	Interface type	Size (GB)	Device name
instance-1	debian-11-bullseye-v20220317	SCSI	20	instance-1

```
s20701@instance-1:~/BGT-Labs/Lab03$ python3 main.py
0
dNG-git/py_builder 1
dmelichar-tgm/eBibliothek 1
reubano/opentag 1
jasenmh/pyFoscamLib 1
GNOME/evolution-scalix 1
...
rperier/linux 8891
mpe/powerpc 8921
scheib/chromium 9152
chromium/chromium 9492
shenzhouzd/update 9913


[988443 rows x 1 columns]
37.9153196811676 seconds
s20701@instance-1:~/BGT-Labs/Lab03$
```

Machine 2:

Note: RAM issues, had to create 1GB swap file


1	39.35
2	38.7
3	38.31
4	38.58
5	38.98
6	40.04
7	38.99
8	37.61
9	38.19
10	39.39
	38.814

Machine configuration

Machine type	e2-highcpu-4
CPU platform	Intel Broadwell
vCPUs to core ratio 	—
Display device	Disabled Enable to use screen capturing and recording tools
GPUs	None

Storage

Boot disk

Name 	Image	Interface type	Size (GB)	Device name
instance-1	debian-11-bullseye-v20220317	SCSI	20	instance-1

```
s20701@instance-1:~/BGT-Labs/Lab03$ python3 main.py
0
dNG-git/py_builder 1
dmelichar-tgm/eBibliothek 1
reubano/opentag 1
jasenmh/pyFoscamLib 1
GNOME/evolution-scalix 1
...
rperier/linux 8891
mpe/powerpc 8921
scheib/chromium 9152
chromium/chromium 9492
shenzhouzd/update 9913

[988443 rows x 1 columns]
42.927881717681885 seconds
```

Conclusion:

Results are comparable, but are inconclusive because second machine had problems executing the python script without additional swap file. Number of rows is reduced caused by TypeError in second parquet file.