Type1

Terraform script:

```
terraform {
  required_providers {
    google = {
      source  = "hashicorp/google"
      version = "4.15.0"
    }
  }
}

provider "google" {
  project     = "bgt-labs"
  credentials = "gcloudcredentials.json"
  region      = "us-central1"
  zone        = "us-central1-a"
}

resource "google_service_account" "defaultUser" {
  account_id   = "terraformcreated"
  display_name = "terraformCreated"
}


resource "google_compute_instance" "Scheduler" {
  name         = "dask-scheduler"
  machine_type = "e2-small"
  boot_disk {
    initialize_params {
      image = "debian-cloud/debian-11"
    }
  }
  network_interface {
    network    = "default"
    network_ip = "10.128.0.2"
    access_config {
      // Ephemeral public IP
    }
  }
  service_account {
    # Google recommends custom service accounts that have cloud-platform
scope and permissions granted via IAM Roles.
    email  = google_service_account.defaultUser.email
    scopes = ["cloud-platform"]
  }
  metadata_startup_script = file("./setupScheduler.sh")

}

resource "google_compute_instance" "Worker" {
  name         = "dask-worker"
  machine_type = "e2-small"
  boot_disk {
    initialize_params {
      image = "debian-cloud/debian-11"
      size = 40
    }
  }
```

```
  network_interface {
    network    = "default"
    network_ip = "10.128.0.3"
        access_config {
      // Ephemeral public IP
    }
  }
  service_account {
    # Google recommends custom service accounts that have cloud-platform
scope and permissions granted via IAM Roles.
    email  = google_service_account.defaultUser.email
    scopes = ["cloud-platform"]
  }
    metadata_startup_script = file("./setupWorker.sh")
}
resource "google_compute_instance" "Worker2" {
  name         = "dask-worker2"
  machine_type = "e2-small"
  boot_disk {
    initialize_params {
      image = "debian-cloud/debian-11"
      size = 40
    }
  }
  network_interface {
    network    = "default"
    network_ip = "10.128.0.4"
        access_config {
      // Ephemeral public IP
    }
  }
  service_account {
    # Google recommends custom service accounts that have cloud-platform
scope and permissions granted via IAM Roles.
    email  = google_service_account.defaultUser.email
    scopes = ["cloud-platform"]
  }
    metadata_startup_script = file("./setupWorker.sh")
}

resource "google_compute_instance" "Client" {
  name         = "dask-client"
  machine_type = "e2-small"

  boot_disk {
    initialize_params {
      image = "debian-cloud/debian-11"
      size = 40
    }
  }
  network_interface {
    network    = "default"
    network_ip = "10.128.0.5"
        access_config {
      // Ephemeral public IP
    }
  }
  service_account {
    # Google recommends custom service accounts that have cloud-platform
scope and permissions granted via IAM Roles.
    email  = google_service_account.defaultUser.email
```

```
    scopes = ["cloud-platform"]
  }
    metadata_startup_script = file("./setupClient.sh")

}
```

Python script:

```
import dask.dataframe as dd
from dask.distributed import Client
import time

if __name__ == "__main__":
    start_time = time.time()
    client = Client("10.128.0.2:8786")

    df =
dd.read_parquet(["./data/new_*.parquet","./data2/new_*.parquet","./data3/new_
*.parquet"], engine="pyarrow")

    df = df['repo_name'].explode()

    df = df.value_counts(ascending=True)
    print(df.compute())
    print(client)

    print(time.time() - start_time, "seconds")
```

Scheduler bash script:

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

sudo apt-get update
sudo apt-get -y install python3 python3-pip
sudo pip3 install dask distributed --upgrade
dask-scheduler
```

Worker bash script:

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

sudo apt-get update
sudo apt-get -y install python3 python3-pip git
pip install dask[complete]
pip install pyarrow

git clone https://github.com/jansitarski/BGT-Labs
cd BGT-Labs/Lab04/

mkdir data
gsutil -m cp gs://pjwstk-bigdata/*.parquet ./data/.
sudo cp -r data data2
sudo cp -r data data3
```

Client bash script:

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

sudo apt-get update
sudo apt-get -y install python pip git
pip install dask[complete]
pip install pyarrow
export PATH="/home/s20701/.local/bin:$PATH"
git clone https://github.com/jansitarski/BGT-Labs
cd BGT-Labs/Lab04/
```

Distributed

```
s20701@dask-client:/BGT-Labs/Lab04$ python3 main.py
rainyjune/animationIcon                      4
cacerrillos/project-stable-serverside        4
cacerrillos/project-stable                   4
node-js-libs/curlrequest                     4
caceres-lab/InvFEST-code                     4
                                           ...
rperier/linux                            39788
mpe/powerpc                              39928
scheib/chromium                          40500
chromium/chromium                        41972
shenzhouzd/update                        43908
Name: repo_name, Length: 1034824, dtype: int64
<Client: 'tcp://10.128.0.2:8786' processes=1 threads=2, memory=1.94 GiB>
1055.5983526706696 seconds
s20701@dask-client:/BGT-Labs/Lab04$
```

```
s20701@dask-worker:~$ dask-worker 10.128.0.2:8786
2022-04-06 22:48:30,677 - distributed.nanny - INFO -         Start Nanny at: 'tcp://10.128.0.3:43441'
2022-04-06 22:48:34,932 - distributed.worker - INFO -       Start worker at:     tcp://10.128.0.3:38937
2022-04-06 22:48:34,933 - distributed.worker - INFO -          Listening to:     tcp://10.128.0.3:38937
2022-04-06 22:48:34,933 - distributed.worker - INFO -          dashboard at:           10.128.0.3:44203
2022-04-06 22:48:34,933 - distributed.worker - INFO - Waiting to connect to:      tcp://10.128.0.2:8786
2022-04-06 22:48:34,933 - distributed.worker - INFO - -------------------------------------------------
2022-04-06 22:48:34,933 - distributed.worker - INFO -               Threads:                          2
2022-04-06 22:48:34,933 - distributed.worker - INFO -                Memory:                   1.94 GiB
2022-04-06 22:48:34,933 - distributed.worker - INFO -       Local Directory: /home/s20701/dask-worker-space/worker-
l6m1xvty
2022-04-06 22:48:34,933 - distributed.worker - INFO - -------------------------------------------------
2022-04-06 22:48:34,985 - distributed.worker - INFO -         Registered to:      tcp://10.128.0.2:8786
2022-04-06 22:48:34,986 - distributed.worker - INFO - -------------------------------------------------
2022-04-06 22:48:34,988 - distributed.core - INFO - Starting established connection
2022-04-06 22:50:36,672 - distributed.core - INFO - Event loop was unresponsive in Worker for 3.50s.  This is often
 caused by long-running GIL-holding functions or moving large chunks of data. This can cause timeouts and instabili
ty.
```

```
s20701@dask-scheduler:~$ dask-scheduler
2022-04-05 18:03:07,572 - distributed.scheduler - INFO - -----------------------------------------------
2022-04-05 18:03:07,575 - distributed.http.proxy - INFO - To route to workers diagnostics web server please install ju
pyter-server-proxy: python -m pip install jupyter-server-proxy
2022-04-05 18:03:07,581 - distributed.scheduler - INFO - -----------------------------------------------
2022-04-05 18:03:07,582 - distributed.scheduler - INFO - Clear task state
2022-04-05 18:03:07,585 - distributed.scheduler - INFO -   Scheduler at:     tcp://10.128.0.2:8786
2022-04-05 18:03:07,586 - distributed.scheduler - INFO -   dashboard at:                     :8787
2022-04-05 18:07:36,909 - distributed.scheduler - INFO - Register worker <WorkerState 'tcp://10.128.0.4:44071', status
: undefined, memory: 0, processing: 0>
2022-04-05 18:07:36,956 - distributed.scheduler - INFO - Starting worker compute stream, tcp://10.128.0.4:44071
2022-04-05 18:07:36,956 - distributed.core - INFO - Starting established connection
2022-04-05 18:07:57,812 - distributed.scheduler - INFO - Register worker <WorkerState 'tcp://10.128.0.3:37967', status
: undefined, memory: 0, processing: 0>
2022-04-05 18:07:57,813 - distributed.scheduler - INFO - Starting worker compute stream, tcp://10.128.0.3:37967
2022-04-05 18:07:57,813 - distributed.core - INFO - Starting established connection
2022-04-05 18:18:40,302 - distributed.scheduler - INFO - Receive client connection: Client-d19fa6f4-b50c-11ec-a64e-95a
0d0f8028b
2022-04-05 18:18:40,303 - distributed.core - INFO - Starting established connection
2022-04-05 18:23:15,544 - distributed.scheduler - INFO - Remove client Client-d19fa6f4-b50c-11ec-a64e-95a0d0f8028b
2022-04-05 18:23:15,545 - distributed.scheduler - INFO - Remove client Client-d19fa6f4-b50c-11ec-a64e-95a0d0f8028b
2022-04-05 18:23:15,546 - distributed.scheduler - INFO - Close client connection: Client-d19fa6f4-b50c-11ec-a64e-95a0d
0f8028b
```

Type2

```
s20701@instance-1:~/BGT-Labs/Lab04$ python3 main2.py
Launching cluster with the following configuration:
  Source Image: projects/ubuntu-os-cloud/global/images/ubuntu-minimal-1804-bionic-v20201014
  Docker Image: daskdev/dask:latest
  Machine Type: e2-small
  Filesystem Size: 50
  Disk Type: pd-standard
  N-GPU Type:
  Zone: us-east1-c
Creating scheduler instance
dask-e9d04c9d-scheduler
        Internal IP: 10.142.0.7
        External IP: 34.73.222.159
Waiting for scheduler to run at 10.142.0.7:8786
Scheduler is running
/usr/lib/python3.9/contextlib.py:124: UserWarning: Creating your cluster is taking a surprisingly long time. This i
s likely due to pending resources. Hang tight!
  next(self.gen)
Creating worker instance
Creating worker instance
Creating worker instance
dask-e9d04c9d-worker-e59b15a5
        Internal IP: 10.142.0.10
        External IP: 34.148.228.175
dask-e9d04c9d-worker-a2ee279a
        Internal IP: 10.142.0.9
        External IP: 35.185.51.19
dask-e9d04c9d-worker-34a4a10d
        Internal IP: 10.142.0.8
        External IP: 34.148.42.187
/home/s20701/.local/lib/python3.9/site-packages/distributed/client.py:1287: VersionMismatchWarning: Mismatched vers
ions found
```

```
+-------------+-----------+-----------+---------+
| Package     | client    | scheduler | workers |
+-------------+-----------+-----------+---------+
| blosc       | 1.10.6    | 1.10.2    | None    |
| dask        | 2022.04.0 | 2022.03.0 | None    |
| distributed | 2022.4.0  | 2022.3.0  | None    |
| lz4         | 4.0.0     | 3.1.10    | None    |
+-------------+-----------+-----------+---------+
  warnings.warn(version_module.VersionMismatchWarning(msg[0]["warning"]))
rainyjune/animationIcon                      4
cacerrillos/project-stable-serverside        4
cacerrillos/project-stable                   4
node-js-libs/curlrequest                     4
caceres-lab/InvFEST-code                     4
                                           ...
rperier/linux                            39788
mpe/powerpc                              39928
scheib/chromium                          40500
chromium/chromium                        41972
shenzhouzd/update                        43908
Name: repo_name, Length: 1034824, dtype: int64
<Client: 'tls://10.142.0.7:8786' processes=3 threads=6, memory=5.81 GiB>
747.7999970912933 seconds
Closing Instance: dask-e9d04c9d-worker-34a4a10d
Closing Instance: dask-e9d04c9d-worker-e59b15a5
Closing Instance: dask-e9d04c9d-worker-a2ee279a
Closing Instance: dask-e9d04c9d-scheduler
s20701@instance-1:~/BGT-Labs/Lab04$
```