Big Data Coursework - Questions

Data Processing and Machine Learning in the Cloud

This is the **INM432 Big Data coursework 2024**. This coursework contains extended elements of **theory** and **practice**, mainly around parallelisation of tasks withSpark and a bit about parallel training using TensorFlow.

Code and Report

Your tasks parallelization of tasks in PySpark, extension, evaluation, and theoretical reflection. Please complete and submit the **coding tasks** in a copy of **this notebook**. Write your code in the **indicated cells** and **include** the **output** in the submitted notebook.
Make sure that **your code contains comments** on its **stucture** and explanations of its **purpose**.

Provide also a **report** with the **textual answers in a separate document**.
Include **screenshots** from the Google Cloud web interface (don't use the SCREENSHOT function that Google provides, but take a picture of the graphs you see for the VMs) and result tables, as well as written text about the analysis.

Submission

Download and submit **your version of this notebook** as an **.ipynb** file and also submit a **shareable link** to your notebook on Colab in your report (created with the Colab 'Share' function) (**and don’t change the online version after submission**).

Further, provide your **report as a PDF document**. **State the number of words** in the document at the end. The report should **not have more than 2000 words**.

Please also submit **a PDF of your Jupyter notebook**.

Introduction and Description

This coursework focuses on parallelisation and scalability in the cloud with Spark and TesorFlow/Keras. We start with code based on **lessons 3 and 4** of the *Fast and Lean Data Science* course by Martin Gorner. The course is based on Tensorflow for data processing and MachineLearning. Tensorflow's data processing approach is somewhat similar to that of Spark, but you don't need to study Tensorflow, just make sure you understand the high-level structure.
What we will do here is **parallelising pre-processing**, and **measuring** performance, and we will perform **evaluation** and **analysis** on the cloud performance, as well as **theoretical discussion**.

This coursework contains **3 sections**.

Section 0

This section just contains some necessary code for setting up the environment. It has no tasks for you (but do read the code and comments).

Section 1

Section 1 is about preprocessing a set of image files. We will work with a public dataset “Flowers” (3600 images, 5 classes). This is not a vast dataset, but it keeps the tasks more manageable for development and you can scale up later, if you like.

In **'Getting Started'** we will work through the data preprocessing code from *Fast and Lean Data Science* which uses TensorFlow's `tf.data` package. There is no task for you here, but you will need to re-use some of this code later.

In **Task 1** you will **parallelise the data preprocessing in Spark**, using Google Cloud (GC) Dataproc. This involves adapting the code from 'Getting Started' to use Spark and running it in the cloud.

Section 2

In **Section 2** we are going to **measure the speed of reading data** in the cloud. In **Task 2** we will **paralellize the measuring** of different configurations **using Spark**.

Section 3

This section is about the theoretical discussion, based on one paper, in **Task 3**. The answers should be given in the PDF report.

General points

For **all coding tasks**, take the **time of the operations** and for the cloud operations, get performance **information from the web interfaces** for your reporting and analysis.

The **tasks** are **mostly independent** of each other. The later tasks can mostly be addressed without needing the solution to the earlier ones.

Section 0: Set-up

As usual, you need to run the **imports and authentication every time you work with this notebook**. Use the **local Spark** installation for development before you send jobs to the cloud.

Read through this section once and **fill in the project ID the first time**, then you can just step straight throught this at the beginning of each session - except for the two authentication cells.

Imports

We import some **packages that will be needed throughout**. For the **code that runs in the cloud**, we will need **separate import sections** that will need to be partly different from the one below.

In [1]:
```
import os, sys, math
import numpy as np
import scipy as sp
import scipy.stats
import time
import datetime
import string
import random
from matplotlib import pyplot as plt
import tensorflow as tf
print("Tensorflow version " + tf.__version__)
import pickle
```

```
Tensorflow version 2.15.0
```

Cloud and Drive authentication

This is for **authenticating with with GCS Google Drive**, so that we can create and use our own buckets and access Dataproc and AI-Platform.

This section **starts with the two interactive authentications**.

First, we mount Google Drive for persistent local storage and create a directory `DB-CW` thay you can use for this work. Then we'll set up the cloud environment, including a storage bucket.

In [2]:
```
print('Mounting google drive...')
from google.colab import drive
drive.mount('/content/drive')
%cd "/content/drive/MyDrive"
!mkdir BD-CW
%cd "/content/drive/MyDrive/BD-CW"
```

```
Mounting google drive...
Mounted at /content/drive
/content/drive/MyDrive
mkdir: cannot create directory ‘BD-CW’: File exists
/content/drive/MyDrive/BD-CW
```

Next, we authenticate with the GCS to enable access to Dataproc and AI-Platform.

In [3]:
```
import sys
if 'google.colab' in sys.modules:
    from google.colab import auth
    auth.authenticate_user()
```

It is useful to **create a new Google Cloud project** for this coursework. You can do this on the GC Console page by clicking on the entry at the top, right of the *Google Cloud Platform* and choosing *New Project*. **Copy** the **generated project ID** to the next cell. Also **enable billing** and the **Compute, Storage and Dataproc** APIs like we did during the labs.

We also specify the **default project and region**. The REGION should be `us-central1` as that seems to be the only one that reliably works with the free credit. This way we don't have to specify this information every time we access the cloud.

In [4]:
```
PROJECT = 'earnest-crow-421721'  ### USE YOUR GOOGLE CLOUD PROJECT ID HERE. ###
!gcloud config set project $PROJECT
REGION = 'us-central1'
CLUSTER = '{}-cluster'.format(PROJECT)
!gcloud config set compute/region $REGION
!gcloud config set dataproc/region $REGION

!gcloud config list # show some information
```

```
Updated property [core/project].
WARNING: Property validation for compute/region was skipped.
Updated property [compute/region].
Updated property [dataproc/region].
[component_manager]
disable_update_check = True
[compute]
region = us-central1
[core]
account = SiddhiBhosale00@gmail.com
project = earnest-crow-421721
[dataproc]
region = us-central1

Your active configuration is: [default]
```

With the cell below, we **create a storage bucket** that we will use later for **global storage**. If the bucket exists you will see a "ServiceException: 409 ...", which does not cause any problems. **You must create your own bucket to have write access.**

In [5]:
```
BUCKET = 'gs://{}-storage'.format(PROJECT)
!gsutil mb $BUCKET
```

```
Creating gs://earnest-crow-421721-storage/...
ServiceException: 409 A Cloud Storage bucket named 'earnest-crow-421721-storage' a
lready exists. Try another name. Bucket names must be globally unique across all G
oogle Cloud projects, including those outside of your organization.
```

The cell below just **defines some routines for displaying images** that will be **used later**. You can see the code by double-clicking, but you don't need to study this.

In [9]:
```
#@title Utility functions for image display **[RUN THIS TO ACTIVATE]** { display-mo
def display_9_images_from_dataset(dataset):
  plt.figure(figsize=(13,13))
  subplot=331
  for i, (image, label) in enumerate(dataset):
    plt.subplot(subplot)
    plt.axis('off')
    plt.imshow(image.numpy().astype(np.uint8))
    plt.title(str(label.numpy()), fontsize=16)
    # plt.title(label.numpy().decode(), fontsize=16)
    subplot += 1
    if i==8:
      break
  plt.tight_layout()
  plt.subplots_adjust(wspace=0.1, hspace=0.1)
  plt.show()
```

```
def display_training_curves(training, validation, title, subplot):
  if subplot%10==1: # set up the subplots on the first call
    plt.subplots(figsize=(10,10), facecolor='#F0F0F0')
    plt.tight_layout()
  ax = plt.subplot(subplot)
  ax.set_facecolor('#F8F8F8')
  ax.plot(training)
  ax.plot(validation)
  ax.set_title('model '+ title)
  ax.set_ylabel(title)
  ax.set_xlabel('epoch')
  ax.legend(['train', 'valid.'])

def dataset_to_numpy_util(dataset, N):
    dataset = dataset.batch(N)
    for images, labels in dataset:
        numpy_images = images.numpy()
        numpy_labels = labels.numpy()
        break;
    return numpy_images, numpy_labels

def title_from_label_and_target(label, correct_label):
  correct = (label == correct_label)
  return "{} [{}{}{}]".format(CLASSES[label], str(correct), ', shoud be ' if not co
                              CLASSES[correct_label] if not correct else ''), corre

def display_one_flower(image, title, subplot, red=False):
    plt.subplot(subplot)
    plt.axis('off')
    plt.imshow(image)
    plt.title(title, fontsize=16, color='red' if red else 'black')
    return subplot+1

def display_9_images_with_predictions(images, predictions, labels):
  subplot=331
  plt.figure(figsize=(13,13))
  classes = np.argmax(predictions, axis=-1)
  for i, image in enumerate(images):
    title, correct = title_from_label_and_target(classes[i], labels[i])
    subplot = display_one_flower(image, title, subplot, not correct)
    if i >= 8:
      break;

  plt.tight_layout()
  plt.subplots_adjust(wspace=0.1, hspace=0.1)
  plt.show()
```

Install Spark locally for quick testing

You can use the cell below to **install Spark locally on this Colab VM** (like in the labs), to do quicker small-scale interactive testing. Using Spark in the cloud with **Dataproc is still required for the final version**.

In [10]:

```
%cd
!apt-get update -qq
!apt-get install openjdk-8-jdk-headless -qq >> /dev/null # send any output to null
!tar -xzf "/content/drive/My Drive/Big_Data/data/spark/spark-3.5.0-bin-hadoop3.tgz"

!pip install -q findspark
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
```

```
os.environ["SPARK_HOME"] = "/root/spark-3.5.0-bin-hadoop3"
import findspark
findspark.init()
import pyspark
print(pyspark.__version__)
sc = pyspark.SparkContext.getOrCreate()
print(sc)
```

```
/root
3.5.0
/usr/lib/python3.10/subprocess.py:1796: RuntimeWarning: os.fork() was called. os.f
ork() is incompatible with multithreaded code, and JAX is multithreaded, so this w
ill likely lead to a deadlock.
  self.pid = _posixsubprocess.fork_exec(
<SparkContext master=local[*] appName=pyspark-shell>
```

Section 1: Data pre-processing

This section is about the **pre-processing of a dataset** for deep learning. We first look at a ready-made solution using Tensorflow and then we build a implement the same process with Spark. The tasks are about **parallelisation** and **analysis** the performance of the cloud implementations.

1.1 Getting started

In this section, we get started with the data pre-processing. The code is based on lecture 3 of the 'Fast and Lean Data Science' course.

This code is using the TensorFlow `tf.data` package, which supports map functions, similar to Spark. Your **task** will be to **re-implement the same approach in Spark**.

We start by **setting some variables for the *Flowers* dataset**.

In [11]:
```
GCS_PATTERN = 'gs://flowers-public/*/*.jpg' # glob  pattern for input files
PARTITIONS = 16 # no of partitions we will use later
TARGET_SIZE = [192, 192] # target resolution for the images
CLASSES = [b'daisy', b'dandelion', b'roses', b'sunflowers', b'tulips']
    # labels for the data
```

We **read the image files** from the public GCS bucket that contains the *Flowers* dataset. **TensorFlow** has **functions** to execute glob patterns that we use to calculate the the number of images in total and per partition (rounded up as we cannont deal with parts of images).

In [12]:
```
nb_images = len(tf.io.gfile.glob(GCS_PATTERN)) # number of images
partition_size = math.ceil(1.0 * nb_images / PARTITIONS) # images per partition (fl
print("GCS_PATTERN matches {} images, to be divided into {} partitions with up to {
```

```
GCS_PATTERN matches 3670 images, to be divided into 16 partitions with up to 230 i
mages each.
```

Map functions

In order to read use the images for learning, they need to be **preprocessed** (decoded, resized, cropped, and potentially recompressed). Below are **map functions** for these steps.

You **don't need to study** the **internals of these functions** in detail.

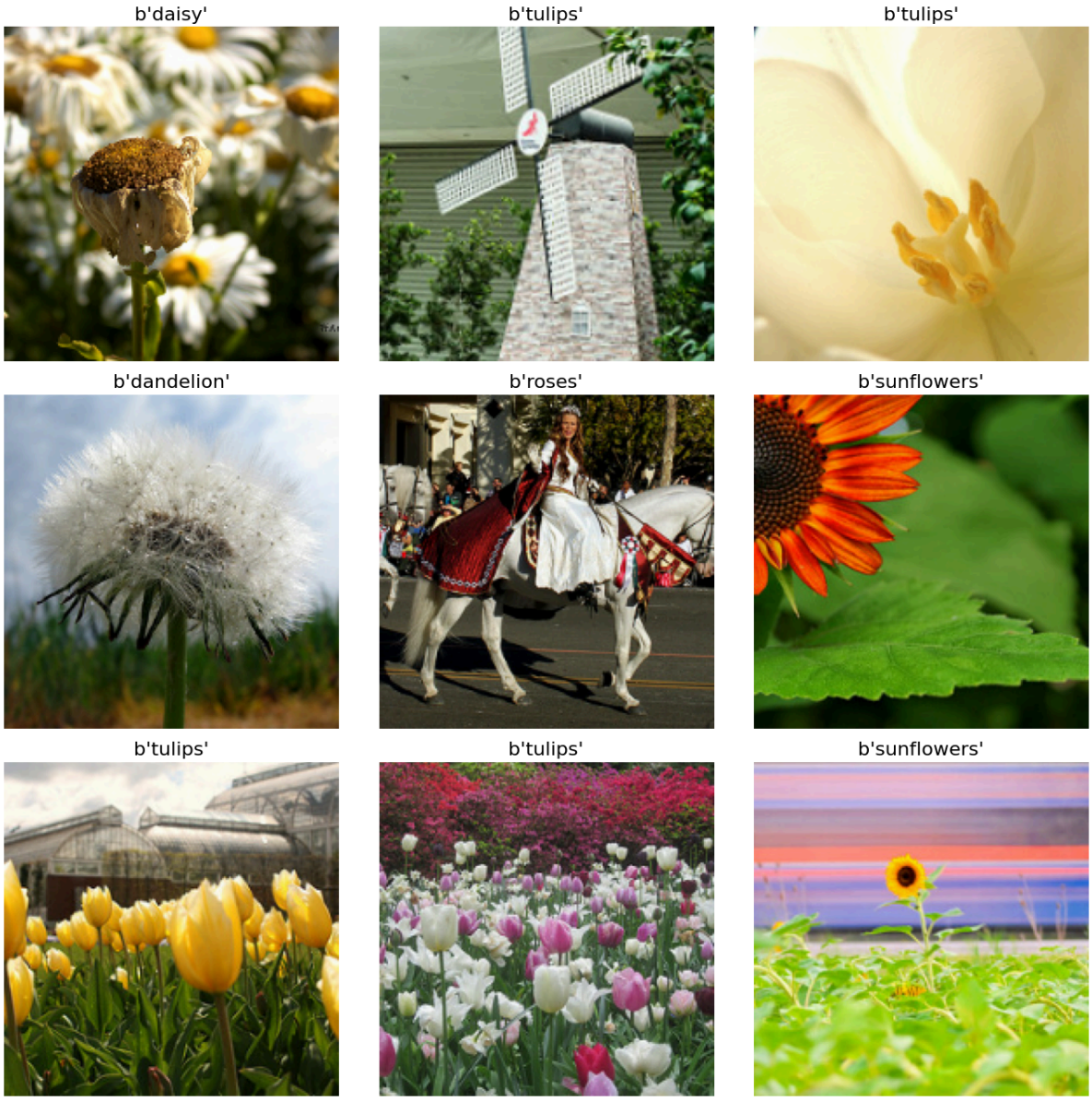
In [29]:
```
def decoding_jpeg(filepath):
    # extracts the image data and creates a class label, based on the filepath
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    # parse flower name from containing directory
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/')
    label2 = label.values[-2]
    return image, label2

def resizing_and_cropping_image(image, label):
    # Resizes and cropd using "fill" algorithm:
    # always make sure the resulting image is cut out from the source image
    # so that it fills the TARGET_SIZE entirely with no black bars
    # and a preserved aspect ratio.
    w = tf.shape(image)[0]
    h = tf.shape(image)[1]
    tw = TARGET_SIZE[1]
    th = TARGET_SIZE[0]
    resize_crit = (w * th) / (h * tw)
    image = tf.cond(resize_crit < 1,
                    lambda: tf.image.resize(image, [w*tw/w, h*tw/w]), # if true
                    lambda: tf.image.resize(image, [w*th/h, h*th/h])  # if false
                    )
    nw = tf.shape(image)[0]
    nh = tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (nw - tw) // 2, (nh - th) // 2, tw
    return image, label

def optimizing_image_encoding(image, label):
    # this reduces the amount of data, but takes some time
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, optimize_size=True, chroma_downsampling=Fal
    return image, label
```

With `tf.data`, we can apply decoding and resizing as map functions.

In [30]:
```
dsetFiles = tf.data.Dataset.list_files(GCS_PATTERN) # This also shuffles the images
dsetDecoded = dsetFiles.map(decoding_jpeg)
dsetResized = dsetDecoded.map(resizing_and_cropping_image)
```

We can also look at some images using the image display function defined above (the one with the hidden code).

In [31]:
```
display_9_images_from_dataset(dsetResized)
```



Now, let's test continuous reading from the dataset. We can see that reading the first 100 files already takes some time.

In [32]:
```
### CODING TASK ###
sample_set = dsetResized.batch(10).take(10) # take 10 batches of 10 images for test
for image, label in sample_set:
    print("Image batch shape {}, {})".format(image.numpy().shape,
        [lbl.decode('utf8') for lbl in label.numpy()]))
```

```
Image batch shape (10, 192, 192, 3), ['roses', 'tulips', 'dandelion', 'tulips', 's
unflowers', 'tulips', 'tulips', 'daisy', 'dandelion', 'daisy'])
Image batch shape (10, 192, 192, 3), ['roses', 'sunflowers', 'roses', 'roses', 'ro
ses', 'sunflowers', 'dandelion', 'sunflowers', 'sunflowers', 'daisy'])
Image batch shape (10, 192, 192, 3), ['daisy', 'daisy', 'roses', 'sunflowers', 'da
ndelion', 'roses', 'daisy', 'daisy', 'dandelion', 'dandelion'])
Image batch shape (10, 192, 192, 3), ['tulips', 'sunflowers', 'roses', 'sunflower
s', 'sunflowers', 'roses', 'dandelion', 'tulips', 'tulips', 'dandelion'])
Image batch shape (10, 192, 192, 3), ['daisy', 'roses', 'daisy', 'dandelion', 'sun
flowers', 'tulips', 'daisy', 'roses', 'dandelion', 'sunflowers'])
Image batch shape (10, 192, 192, 3), ['tulips', 'tulips', 'dandelion', 'sunflower
s', 'dandelion', 'roses', 'roses', 'dandelion', 'tulips', 'dandelion'])
Image batch shape (10, 192, 192, 3), ['tulips', 'roses', 'sunflowers', 'sunflower
s', 'daisy', 'tulips', 'roses', 'dandelion', 'tulips', 'dandelion'])
Image batch shape (10, 192, 192, 3), ['roses', 'daisy', 'daisy', 'daisy', 'daisy',
'daisy', 'tulips', 'tulips', 'tulips', 'dandelion'])
Image batch shape (10, 192, 192, 3), ['daisy', 'dandelion', 'dandelion', 'sunflowe
rs', 'daisy', 'roses', 'dandelion', 'tulips', 'tulips', 'tulips'])
Image batch shape (10, 192, 192, 3), ['tulips', 'daisy', 'tulips', 'daisy', 'dande
lion', 'roses', 'tulips', 'dandelion', 'dandelion', 'tulips'])
```

1.2 Improving Speed

Using individual image files didn't look very fast. The 'Lean and Fast Data Science' course introduced **two techniques to improve the speed**.

Recompress the images

By **compressing** the images in the **reduced resolution** we save on the size. This **costs some CPU time** upfront, but **saves network and disk bandwith**, especially when the data are **read multiple times**.

In [33]:
```
### CODING TASK ###
# This is a quick test to get an idea how long recompressions takes.
dataset4 = dsetResized.map(optimizing_image_encoding)
test_set = dataset4.batch(10).take(10)
for image, label in test_set:
    print("Image batch shape {}, {})".format(image.numpy().shape, [lbl.decode('utf8
```

```
Image batch shape (10,), ['dandelion', 'sunflowers', 'tulips', 'sunflowers', 'dais
y', 'daisy', 'daisy', 'dandelion', 'dandelion', 'roses'])
Image batch shape (10,), ['tulips', 'daisy', 'daisy', 'daisy', 'dandelion', 'dande
lion', 'sunflowers', 'dandelion', 'sunflowers', 'tulips'])
Image batch shape (10,), ['tulips', 'dandelion', 'daisy', 'roses', 'sunflowers',
'sunflowers', 'dandelion', 'tulips', 'roses', 'roses'])
Image batch shape (10,), ['roses', 'sunflowers', 'sunflowers', 'tulips', 'dandelio
n', 'sunflowers', 'sunflowers', 'daisy', 'sunflowers', 'roses'])
Image batch shape (10,), ['roses', 'sunflowers', 'daisy', 'daisy', 'dandelion', 'r
oses', 'tulips', 'dandelion', 'dandelion', 'tulips'])
Image batch shape (10,), ['tulips', 'sunflowers', 'daisy', 'sunflowers', 'daisy',
'sunflowers', 'tulips', 'sunflowers', 'roses', 'daisy'])
Image batch shape (10,), ['tulips', 'sunflowers', 'dandelion', 'daisy', 'dandelio
n', 'dandelion', 'tulips', 'dandelion', 'roses', 'dandelion'])
Image batch shape (10,), ['roses', 'daisy', 'sunflowers', 'roses', 'sunflowers',
'sunflowers', 'daisy', 'roses', 'sunflowers', 'daisy'])
Image batch shape (10,), ['sunflowers', 'tulips', 'tulips', 'sunflowers', 'roses',
'daisy', 'roses', 'dandelion', 'daisy', 'sunflowers'])
Image batch shape (10,), ['dandelion', 'dandelion', 'sunflowers', 'tulips', 'dande
lion', 'dandelion', 'dandelion', 'tulips', 'tulips', 'dandelion'])
```

Write the dataset to TFRecord files

By writing **multiple preprocessed samples into a single file**, we can make further speed gains. We distribute the data over **partitions** to facilitate **parallelisation** when the data are used. First we need to **define a location** where we want to put the file.

In [34]:
```
GCS_OUTPUT = BUCKET + '/tfrecords-jpeg-192x192-2/flowers'  # prefix for output file
```

Now we can **write the TFRecord files** to the bucket.

Running the cell takes some time and **only needs to be done once** or not at all, as you can use the publicly available data for the next few cells. For convenience I have commented out the call to `write_tfrecords` at the end of the next cell. You don't need to run it (it takes some time), but you'll need to use the code below later (but there is no need to study it in detail).

There is a **ready-made pre-processed data** versions available here: `gs://flowers-public/tfrecords-jpeg-192x192-2/`, that we can use for testing.

In [35]:
```
### CODING TASK ###
# functions for writing TFRecord entries
# Feature values are always stored as lists, a single data element will be a list c
def _bytestring_feature(list_of_bytestrings):
    return tf.train.Feature(bytes_list=tf.train.BytesList(value=list_of_bytestrings

def _int_feature(list_of_ints): # int64
    return tf.train.Feature(int64_list=tf.train.Int64List(value=list_of_ints))

def to_tfrecord(tfrec_filewriter, img_bytes, label): # Create tf data records
    class_num = np.argmax(np.array(CLASSES)==label) # 'roses' => 2 (order defined i
    one_hot_class = np.eye(len(CLASSES))[class_num]     # [0, 0, 1, 0, 0] for class
    feature = {
        "image": _bytestring_feature([img_bytes]), # one image in the list
        "class": _int_feature([class_num]) #,      # one class in the list
    }
    return tf.train.Example(features=tf.train.Features(feature=feature))

def write_tfrecords(GCS_PATTERN,GCS_OUTPUT,partition_size): # write the images to f
    print("Writing TFRecords")
    tt0 = time.time()
    filenames = tf.data.Dataset.list_files(GCS_PATTERN)
    dataset1 = filenames.map(extracting_jpeg_and_label)
    dataset2 = dataset1.map(resizing_and_cropping_image)
    dataset3 = dataset2.map(optimizing_image_encoding)
    dataset4 = dataset3.batch(partition_size) # partitioning: there will be one "ba
    for partition, (image, label) in enumerate(dataset4):
        # batch size used as partition size here
        partition_size = image.numpy().shape[0]
        # good practice to have the number of records in the filename
        filename = GCS_OUTPUT + "{:02d}-{}.tfrec".format(partition, partition_size)
        # You need to change GCS_OUTPUT to your own bucket to actually create new f
        with tf.io.TFRecordWriter(filename) as out_file:
            for i in range(partition_size):
                example = to_tfrecord(out_file,
                                      image.numpy()[i], # re-compressed image: alread
                                      label.numpy()[i] #
                                      )
                out_file.write(example.SerializeToString())
```

```
        print("Wrote file {} containing {} records".format(filename, partition_size
    print("Total time: "+str(time.time()-tt0))

write_tfrecords(GCS_PATTERN,GCS_OUTPUT,partition_size) # uncomment to run this cell
```

```
Writing TFRecords
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers00-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers01-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers02-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers03-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers04-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers05-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers06-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers07-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers08-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers09-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers10-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers11-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers12-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers13-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers14-23
0.tfrec containing 230 records
Wrote file gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers15-22
0.tfrec containing 220 records
Total time: 863.5059428215027
```

Test the TFRecord files

We can now **read from the TFRecord files**. By default, we use the files in the public bucket. Comment out the 1st line of the cell below to use the files written in the cell above.

In [36]:

```
#GCS_OUTPUT = 'gs://flowers-public/tfrecords-jpeg-192x192-2/'
# remove the line above to use your own files that you generated above

def read_tfrecord(example):
    features = {
        "image": tf.io.FixedLenFeature([], tf.string),  # tf.string = bytestring (r
        "class": tf.io.FixedLenFeature([], tf.int64) #,   # shape [] means scalar
    }
    # decode the TFRecord
    example = tf.io.parse_single_example(example, features)
    image = tf.image.decode_jpeg(example['image'], channels=3)
    image = tf.reshape(image, [*TARGET_SIZE, 3])
    class_num = example['class']
    return image, class_num

def load_dataset(filenames):
    # read from TFRecords. For optimal performance, read from multiple
    # TFRecord files at once and set the option experimental_deterministic = False
```

```
    # to allow order-altering optimizations.
    option_no_order = tf.data.Options()
    option_no_order.experimental_deterministic = False

    dataset = tf.data.TFRecordDataset(filenames)
    dataset = dataset.with_options(option_no_order)
    dataset = dataset.map(read_tfrecord)
    return dataset


filenames = tf.io.gfile.glob(GCS_OUTPUT + "*.tfrec")
datasetTfrec = load_dataset(filenames)
```

Let's have a look **if reading from the TFRecord** files is **quicker**.

In [37]:
```
### CODING TASK ###
batched_dataset = datasetTfrec.batch(10)
sample_set = batched_dataset.take(10)
for image, label in sample_set:
    print("Image batch shape {}, {})".format(image.numpy().shape, \
                        [str(lbl) for lbl in label.numpy()]))
```

```
Image batch shape (10, 192, 192, 3), ['0', '2', '0', '1', '4', '4', '3', '1', '4',
'3'])
Image batch shape (10, 192, 192, 3), ['3', '1', '4', '0', '4', '3', '1', '3', '1',
'0'])
Image batch shape (10, 192, 192, 3), ['4', '3', '0', '0', '2', '4', '4', '1', '3',
'1'])
Image batch shape (10, 192, 192, 3), ['1', '2', '2', '4', '3', '4', '4', '1', '3',
'1'])
Image batch shape (10, 192, 192, 3), ['0', '0', '0', '3', '4', '4', '1', '4', '4',
'1'])
Image batch shape (10, 192, 192, 3), ['1', '4', '3', '4', '1', '0', '4', '2', '4',
'2'])
Image batch shape (10, 192, 192, 3), ['1', '1', '3', '0', '0', '4', '4', '3', '1',
'2'])
Image batch shape (10, 192, 192, 3), ['3', '2', '1', '0', '2', '2', '2', '4', '0',
'1'])
Image batch shape (10, 192, 192, 3), ['0', '2', '4', '4', '4', '2', '3', '0', '1',
'3'])
Image batch shape (10, 192, 192, 3), ['4', '3', '0', '1', '3', '3', '4', '3', '1',
'4'])
```

Wow, we have a **massive speed-up**! The repackageing is worthwhile :-)

Task 1: Write TFRecord files to the cloud with Spark (40%)

Since recompressing and repackaging is very effective, we would like to be able to do it inparallel for large datasets. This is a relatively straightforward case of **parallelisation**. We will **use Spark to implement** the same process as above, but in parallel.

1a) Create the script (14%)

Re-implement the pre-processing in Spark, using Spark mechanisms for **distributing** the workload **over multiple machines**.

You need to:

i) **Copy** over the **mapping functions** (see section 1.1) and **adapt** the resizing and recompression functions **to Spark** (only one argument). (3%)

ii) **Replace** the TensorFlow **Dataset objects with RDDs**, starting with an RDD that contains the list of image filenames. (3%)

iii) **Sample** the the RDD to a smaller number at an appropriate position in the code. Specify a sampling factor of 0.02 for short tests. (1%)

iv) Then **use the functions from above** to write the TFRecord files. (3%)

v) The code for **writing to the TFRecord files** needs to be put into a function, that can be applied to every partition with the 'RDD.mapPartitionsWithIndex' function. The return value of that function is not used here, but you should return the filename, so that you have a list of the created TFRecord files. (4%)

In [ ]:
```
# Defining necessary libraries
import os
import numpy as np
import tensorflow as tf
import pyspark

# File glob pattern for input files
input_glob_pattern = 'gs://flowers-public/*/*.jpg'
# Project and storage bucket IDs
project_id = 'earnest-crow-421721'
storage_bucket = 'earnest-crow-421721-storage'
# Output path for TFRecord files
output_path = storage_bucket + '/tfrecords-jpeg-192x192-2/flowers'
# Number of partitions for RDD
num_partitions = 16
# Target image size
target_size = [192, 192]
# Labels for the data
data_labels = [b'daisy', b'dandelion', b'roses', b'sunflowers', b'tulips']

# i) Adapting the resizing and recompression functions to spark

def extracting_jpeg_and_label(filepath):
    # extracts the image data and creates a class label, based on the filepath
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    # parse flower name from containing directory
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/')
    label2 = label.values[-2]
    return image, label2

def resizing_and_cropping_image(image_label):
    image, label2 = image_label
    width = tf.shape(image)[0]
    height = tf.shape(image)[1]
    twidth = TARGET_SIZE[1]
    theight = TARGET_SIZE[0]
    resizecrit = (width * theight) / (height * tweight)
    image = tf.cond(resizecrit < 1,
                    lambda: tf.image.resize(image, [width*twidth/width, height*twic
                    lambda: tf.image.resize(image, [width*theight/height, height*th
                    )
    newwidth = tf.shape(image)[0]
    newheight = tf.shape(image)[1]
```

```
    image = tf.image.crop_to_bounding_box(image, (newwidth - twidth) // 2, (newheig
    return image, label2

def optimizing_image_encoding(image_label):
    image, label2 = image_label
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, optimize_size=True, chroma_downsampling=Fal
    return (image, class_name)

# ii) & iii) Replacing TensorFlow Dataset objects with RDDs AND Sampling RDDs
file_paths = tf.io.gfile.glob(input_glob_pattern)
spark_ctx = pyspark.SparkContext.getOrCreate()
image_files_rdd = spark_ctx.parallelize(file_paths)
sampling_files_rdd = image_files_rdd.sample(False, 0.02)
filesrdd = spark_ctx.parallelize(file_paths)
decoded_images_rdd = image_files_rdd.map(extracting_jpeg_and_label)
resized_images_rdd = decoded_images_rdd.map(resizing_and_cropping_image)
recompressed_images_rdd = resized_images_rdd.map(optimizing_image_encoding)

# iv) Writing TFRecords files using the functions above

def making_tfrecord(tf_writer, image_data, label_data):
    class_index = np.argmax(np.array(CLASSES) == label_data)
    example_features = {
        "image": _bytestring_feature([image_data]),
        "class": _int_feature([class_index])
    }
    return tf.train.Example(features=tf.train.Features(feature=example_features))

print("Writing TFRecord files")

# v) writing to the TFRecord files in a function

def writing_tfrecord_files(partition_index, partition_data):
    output_file_name = OUTPUT_PATH + "{}.tfrec".format(partition_index)
    with tf.io.TFRecordWriter(output_file_name) as tf_writer:
        for img, label in partition_data:
            tfrecord_ex = making_tfrecord(tf_writer, img.numpy(), label.numpy())
            tf_writer.write(tfrecord_exe.SerializeToString())
    return [output_file_name]

final_rdd = recompressed_images_rdd.repartition(num_partitions)
resulting_filenames_rdd = final_rdd.mapPartitionsWithIndex(writing_tfrecord_files)
```

```
Writing TFRecord files
```

1b) Testing (3%)

i) Read from the TFRecord Dataset, using `load_dataset` and `display_9_images_from_dataset` to test.

In [ ]:
```
### CODING TASK ###

# Executing the workflow
file_paths = tf.io.gfile.glob(GCS_OUTPUT + "*.tfrec")
decodeddataset = load_dataset(file_paths)
display_9_images_from_dataset(decodeddataset)
```

ii) Write your code above into a file using the *cell magic* `%%writefile spark_write_tfrec.py` at the beginning of the file. Then, run the file locally in Spark.

In [ ]:
```
### CODING TASK ###
%%writefile spark_write_tfrec.py

# Defining necessary libraries
import os
import numpy as np
import tensorflow as tf
import pyspark

# File glob pattern for input files
input_glob_pattern = 'gs://flowers-public/*/*.jpg'
# Project and storage bucket IDs
project_id = 'earnest-crow-421721'
storage_bucket = 'earnest-crow-421721-storage'
# Output path for TFRecord files
output_path = storage_bucket + '/tfrecords-jpeg-192x192-2/flowers'
# Number of partitions for RDD
num_partitions = 16
# Target image size
target_size = [192, 192]
# Labels for the data
data_labels = [b'daisy', b'dandelion', b'roses', b'sunflowers', b'tulips']
```

```
# i) Adapting the resizing and recompression functions to spark

def extracting_jpeg_and_label(filepath):
    # extracts the image data and creates a class label, based on the filepath
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    # parse flower name from containing directory
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/')
    label2 = label.values[-2]
    return image, label2

def resizing_and_cropping_image(image_label):
    image, label2 = image_label
    width = tf.shape(image)[0]
    height = tf.shape(image)[1]
    twidth = TARGET_SIZE[1]
    theight = TARGET_SIZE[0]
    resizecrit = (width * theight) / (height * tweight)
    image = tf.cond(resizecrit < 1,
                    lambda: tf.image.resize(image, [width*twidth/width, height*twic
                    lambda: tf.image.resize(image, [width*theight/height, height*th
                    )
    newwidth = tf.shape(image)[0]
    newheight = tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (newwidth - twidth) // 2, (newheig
    return image, label2

def optimizing_image_encoding(image_label):
    image, label2 = image_label
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, optimize_size=True, chroma_downsampling=Fal
    return (image, class_name)

# ii) & iii) Replacing TensorFlow Dataset objects with RDDs AND Sampling RDDs
file_paths = tf.io.gfile.glob(input_glob_pattern)
spark_ctx = pyspark.SparkContext.getOrCreate()

##Task 1D##
filesrdd = spark_ctx.parallelize(file_paths,16)

image_files_rdd = spark_ctx.parallelize(file_paths,16)
sampling_files_rdd = image_files_rdd.sample(False, 0.02)
decoded_images_rdd = image_files_rdd.map(extracting_jpeg_and_label)
resized_images_rdd = decoded_images_rdd.map(resizing_and_cropping_image)
recompressed_images_rdd = resized_images_rdd.map(optimizing_image_encoding)

# iv) Writing TFRecords files using the functions above

def making_tfrecord(tf_writer, image_data, label_data):
    class_index = np.argmax(np.array(CLASSES) == label_data)
    example_features = {
        "image": _bytestring_feature([image_data]),
        "class": _int_feature([class_index])
    }
    return tf.train.Example(features=tf.train.Features(feature=example_features))

print("Writing TFRecord files")

# v) writing to the TFRecord files in a function

def writing_tfrecord_files(partition_index, partition_data):
    output_file_name = OUTPUT_PATH + "{}.tfrec".format(partition_index)
    with tf.io.TFRecordWriter(output_file_name) as tf_writer:
```

```
        for img, label in partition_data:
            tfrecord_ex = making_tfrecord(tf_writer, img.numpy(), label.numpy())
            tf_writer.write(tfrecord_exe.SerializeToString())
    return [output_file_name]

final_rdd = recompressed_images_rdd.repartition(num_partitions)
resulting_filenames_rdd = final_rdd.mapPartitionsWithIndex(writing_tfrecord_files)
```

```
Overwriting spark_write_tfrec.py
```

1c) Set up a cluster and run the script. (6%)

Following the example from the labs, set up a cluster to run PySpark jobs in the cloud. You need to set up so that TensorFlow is installed on all nodes in the cluster.

i) Single machine cluster

Set up a cluster with a single machine using the maximal SSD size (100) and 8 vCPUs.

Enable **package installation** by passing a flag `--initialization-actions` with argument `gs://goog-dataproc-initialization-actions-$REGION/python/pip-install.sh` (this is a public script that will read metadata to determine which packages to install). Then, the **packages are specified** by providing a `--metadata` flag with the argument `PIP_PACKAGES=tensorflow==2.4.0`.

Note: consider using `PIP_PACKAGES="tensorflow numpy"` or `PIP_PACKAGES=tensorflow` in case an older version of tensorflow is causing issues.

When the cluster is running, run your script to check that it works and keep the output cell output. (3%)

In [ ]:

```
### CODING TASK ###

# Setting up a single-machine cluster with maximal SSD size (100) and 8 vCPUs.
!gcloud dataproc clusters create 'earnest-crow-421721-single' \
    --bucket 'earnest-crow-421721-storage' \
    --region us-central1 \
    --zone us-central1-c \
    --image-version 1.5-ubuntu18 --single-node \
    --master-machine-type n1-standard-8 --master-boot-disk-type pd-ssd --master-boo
    --initialization-actions gs://goog-dataproc-initialization-actions-us-central1/
    --metadata PIP_PACKAGES='tensorflow==2.4.0 numpy protobuf==3.20.0'
```

```
Waiting on operation [projects/earnest-crow-421721/regions/us-central1/operations/
38d70477-0b24-3a29-8873-1be7cefcf6f2].

WARNING: Consider using Auto Zone rather than selecting a zone manually. See http
s://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Don't create production clusters that reference initialization actions lo
cated in the gs://goog-dataproc-initialization-actions-REGION public buckets. Thes
e scripts are provided as reference implementations, and they are synchronized wit
h ongoing GitHub repository changes—a new version of a initialization action in pu
blic buckets may break your cluster creation. Instead, copy the following initiali
zation actions from public buckets into your bucket : gs://goog-dataproc-initializ
ation-actions-us-central1/python/pip-install.sh
WARNING: Failed to validate permissions required for default service account: '972
034511549-compute@developer.gserviceaccount.com'. Cluster creation could still be
successful if required permissions have been granted to the respective service acc
ounts as mentioned in the document https://cloud.google.com/dataproc/docs/concept
s/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be
due to Cloud Resource Manager API hasn't been enabled in your project '97203451154
9' before or it is disabled. Enable it by visiting 'https://console.developers.goo
gle.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=97203451154
9'.
WARNING: The firewall rules for specified network or subnetwork would allow ingres
s traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'earnest-crow-421721-storage' is not
using uniform bucket level access IAM configuration. It is recommended to update b
ucket to enable the same. See https://cloud.google.com/storage/docs/uniform-bucket
-level-access.
Created [https://dataproc.googleapis.com/v1/projects/earnest-crow-421721/regions/u
s-central1/clusters/earnest-crow-421721-single] Cluster placed in zone [us-central
1-c].
```

Run the script in the cloud and test the output.

In [ ]:
```
### CODING TASK ###
%time
!gcloud dataproc jobs submit pyspark --cluster 'earnest-crow-421721-single' \
    --region "us-central1" spark_write_tfrec.py
```

```
CPU times: user 4 µs, sys: 0 ns, total: 4 µs
Wall time: 7.63 µs
Job [e2133a84abdf451193b7041ed2013afd] submitted.
Waiting for job output...
2024-05-04 08:44:50.240747: W tensorflow/stream_executor/platform/default/dso_load
er.cc:60] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.s
o.11.0: cannot open shared object file: No such file or directory; LD_LIBRARY_PAT
H: :/usr/lib/hadoop/lib/native
2024-05-04 08:44:50.240794: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] I
gnore above cudart dlerror if you do not have a GPU set up on your machine.
24/05/04 08:44:53 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/05/04 08:44:53 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/05/04 08:44:53 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordina
tor
24/05/04 08:44:53 INFO org.spark_project.jetty.util.log: Logging initialized @5322
ms to org.spark_project.jetty.util.log.Slf4jLog
24/05/04 08:44:53 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHO
T; built: unknown; git: unknown; jvm 1.8.0_382-b05
24/05/04 08:44:53 INFO org.spark_project.jetty.server.Server: Started @5460ms
24/05/04 08:44:53 INFO org.spark_project.jetty.server.AbstractConnector: Started S
erverConnector@4a63dd5d{HTTP/1.1, (http/1.1)}{0.0.0.0:39025}
24/05/04 08:44:54 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to Resour
ceManager at earnest-crow-421721-single-m/10.128.0.4:8032
24/05/04 08:44:54 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Appli
cation History server at earnest-crow-421721-single-m/10.128.0.4:10200
24/05/04 08:44:54 INFO org.apache.hadoop.conf.Configuration: resource-types.xml no
t found
24/05/04 08:44:54 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable
to find 'resource-types.xml'.
24/05/04 08:44:54 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/05/04 08:44:54 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = vcores, units = , type = COUNTABLE
24/05/04 08:44:57 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Subm
itted application application_1714811729661_0002
Writing TFRecord files
24/05/04 08:45:04 INFO org.spark_project.jetty.server.AbstractConnector: Stopped S
park@4a63dd5d{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [e2133a84abdf451193b7041ed2013afd] finished successfully.
done: true
driverControlFilesUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-meta
info/208e6e9f-f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/
driverOutputResourceUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-me
tainfo/208e6e9f-f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/
driveroutput
jobUuid: 158e252e-9ee3-3b6f-bc68-3053e3afc2da
placement:
  clusterName: earnest-crow-421721-single
  clusterUuid: 208e6e9f-f94e-41ec-978b-94e9aa2fa7dc
pysparkJob:
  mainPythonFileUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-metain
fo/208e6e9f-f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/stag
ing/spark_write_tfrec.py
reference:
  jobId: e2133a84abdf451193b7041ed2013afd
  projectId: earnest-crow-421721
status:
  state: DONE
  stateStartTime: '2024-05-04T08:45:05.444487Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-05-04T08:44:47.111456Z'
- state: SETUP_DONE
  stateStartTime: '2024-05-04T08:44:47.139177Z'
```

```
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-05-04T08:44:47.369153Z'
yarnApplications:
- name: spark_write_tfrec.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://earnest-crow-421721-single-m:8088/proxy/application_17148117
29661_0002/
```

In the free credit tier on Google Cloud, there are normally the following **restrictions** on compute machines:

- max 100GB of *SSD persistent disk*
- max 2000GB of *standard persistent disk*
- max 8 *vCPU*s
- no GPUs

See here for details The **disks are virtual** disks, where **I/O speed is limited in proportion to the size**, so we should allocate them evenly. This has mainly an effect on the **time the cluster needs to start**, as we are reading the data mainly from the bucket and we are not writing much to disk at all.

ii) Maximal cluster

Use the **largest possible cluster** within these constraints, i.e. **1 master and 7 worker nodes**. Each of them with 1 (virtual) CPU. The master should get the full *SSD* capacity and the 7 worker nodes should get equal shares of the *standard* disk capacity to maximise throughput.

Once the cluster is running, test your script. (3%)

In [ ]:
```
### CODING TASK ###
!gcloud dataproc clusters create 'earnest-crow-421721-maximal' \
  --bucket 'earnest-crow-421721-storage' \
  --region us-central1 \
  --zone us-central1-c \
  --image-version 1.5-ubuntu18 \
  --master-machine-type n1-standard-1 \
  --master-boot-disk-type pd-ssd \
  --master-boot-disk-size 100 \
  --num-workers 7 --worker-machine-type n1-standard-1 \
  --worker-boot-disk-type pd-standard --worker-boot-disk-size 100 \
  --initialization-actions gs://goog-dataproc-initialization-actions-us-central1/py
  --metadata PIP_PACKAGES='tensorflow==2.4.0 numpy  protobuf==3.20.0'
 #ERROR: (gcloud.dataproc.clusters.create) INVALID_ARGUMENT: Insufficient 'IN_USE_A
#Had to decrease the ssd size also due to space errors
```

```
Waiting on operation [projects/earnest-crow-421721/regions/us-central1/operations/
77abcd26-2d7c-3793-ac13-66b8e749c698].

WARNING: Consider using Auto Zone rather than selecting a zone manually. See http
s://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Creating clusters using the n1-standard-1 machine type is not recommende
d. Consider using a machine type with higher memory.
WARNING: Don't create production clusters that reference initialization actions lo
cated in the gs://goog-dataproc-initialization-actions-REGION public buckets. Thes
e scripts are provided as reference implementations, and they are synchronized wit
h ongoing GitHub repository changes—a new version of a initialization action in pu
blic buckets may break your cluster creation. Instead, copy the following initiali
zation actions from public buckets into your bucket : gs://goog-dataproc-initializ
ation-actions-us-central1/python/pip-install.sh
WARNING: Failed to validate permissions required for default service account: '972
034511549-compute@developer.gserviceaccount.com'. Cluster creation could still be
successful if required permissions have been granted to the respective service acc
ounts as mentioned in the document https://cloud.google.com/dataproc/docs/concept
s/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be
due to Cloud Resource Manager API hasn't been enabled in your project '97203451154
9' before or it is disabled. Enable it by visiting 'https://console.developers.goo
gle.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=97203451154
9'.
WARNING: For PD-Standard without local SSDs, we strongly recommend provisioning 1T
B or larger to ensure consistently high I/O performance. See https://cloud.google.
com/compute/docs/disks/performance for information on disk I/O performance.
WARNING: The firewall rules for specified network or subnetwork would allow ingres
s traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'earnest-crow-421721-storage' is not
using uniform bucket level access IAM configuration. It is recommended to update b
ucket to enable the same. See https://cloud.google.com/storage/docs/uniform-bucket
-level-access.
Created [https://dataproc.googleapis.com/v1/projects/earnest-crow-421721/regions/u
s-central1/clusters/earnest-crow-421721-maximal] Cluster placed in zone [us-centra
l1-c].
```

In [ ]:

```
### CODING TASK ###
%time
!gcloud dataproc jobs submit pyspark --cluster 'earnest-crow-421721-maximal' \
    --region "us-central1" spark_write_tfrec.py
```

```
CPU times: user 4 µs, sys: 0 ns, total: 4 µs
Wall time: 8.34 µs
Job [482e7379032f4d96b79840df50e3abd3] submitted.
Waiting for job output...
2024-05-04 08:57:17.738883: W tensorflow/stream_executor/platform/default/dso_load
er.cc:60] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.s
o.11.0: cannot open shared object file: No such file or directory; LD_LIBRARY_PAT
H: :/usr/lib/hadoop/lib/native
2024-05-04 08:57:17.739055: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] I
gnore above cudart dlerror if you do not have a GPU set up on your machine.
24/05/04 08:57:21 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/05/04 08:57:21 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/05/04 08:57:21 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordina
tor
24/05/04 08:57:22 INFO org.spark_project.jetty.util.log: Logging initialized @8231
ms to org.spark_project.jetty.util.log.Slf4jLog
24/05/04 08:57:22 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHO
T; built: unknown; git: unknown; jvm 1.8.0_382-b05
24/05/04 08:57:22 INFO org.spark_project.jetty.server.Server: Started @8470ms
24/05/04 08:57:22 INFO org.spark_project.jetty.server.AbstractConnector: Started S
erverConnector@45915c69{HTTP/1.1, (http/1.1)}{0.0.0.0:35919}
24/05/04 08:57:24 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to Resour
ceManager at earnest-crow-421721-maximal-m/10.128.0.10:8032
24/05/04 08:57:25 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Appli
cation History server at earnest-crow-421721-maximal-m/10.128.0.10:10200
24/05/04 08:57:25 INFO org.apache.hadoop.conf.Configuration: resource-types.xml no
t found
24/05/04 08:57:25 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable
to find 'resource-types.xml'.
24/05/04 08:57:25 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/05/04 08:57:25 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = vcores, units = , type = COUNTABLE
24/05/04 08:57:28 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Subm
itted application application_1714812680222_0002
Writing TFRecord files
24/05/04 08:57:40 INFO org.spark_project.jetty.server.AbstractConnector: Stopped S
park@45915c69{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [482e7379032f4d96b79840df50e3abd3] finished successfully.
done: true
driverControlFilesUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-meta
info/ecce2160-dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/
driverOutputResourceUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-me
tainfo/ecce2160-dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/
driveroutput
jobUuid: 67e872a5-57c5-3a48-9050-115de364e2b6
placement:
  clusterName: earnest-crow-421721-maximal
  clusterUuid: ecce2160-dff6-428c-bd1c-e2aae8e1542f
pysparkJob:
  mainPythonFileUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-metain
fo/ecce2160-dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/stag
ing/spark_write_tfrec.py
reference:
  jobId: 482e7379032f4d96b79840df50e3abd3
  projectId: earnest-crow-421721
status:
  state: DONE
  stateStartTime: '2024-05-04T08:57:43.805127Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-05-04T08:57:12.627675Z'
- state: SETUP_DONE
  stateStartTime: '2024-05-04T08:57:12.753787Z'
```

```
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-05-04T08:57:12.967398Z'
yarnApplications:
- name: spark_write_tfrec.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://earnest-crow-421721-maximal-m:8088/proxy/application_1714812
680222_0002/
```

1d) Optimisation, experiments, and discussion (17%)

i) Improve parallelisation

If you implemented a straightfoward version, you will **probably** observe that **all the computation** is done on only **two nodes**. This can be adressed by using the **second parameter** in the initial call to **parallelize**. Make the **suitable change** in the code you have written above and mark it up in comments as `### TASK 1d ###`.

Demonstrate the difference in cluster utilisation before and after the change based on different parameter values with **screenshots from Google Cloud** and measure the **difference in the processing time**. (6%)

ii) Experiment with cluster configurations.

In addition to the experiments above (using 8 VMs),test your program with 4 machines with double the resources each (2 vCPUs, memory, disk) and 1 machine with eightfold resources. Discuss the results in terms of disk I/O and network bandwidth allocation in the cloud. (7%)

iii) Explain the difference between this use of Spark and most standard applications like e.g. in our labs in terms of where the data is stored. What kind of parallelisation approach is used here? (4%)

Write the code below and your answers in the report.

Section 2: Speed tests

We have seen that **reading from the pre-processed TFRecord files** is **faster** than reading individual image files and decoding on the fly. This task is about **measuring this effect** and **parallelizing the tests with PySpark**.

2.1 Speed test implementation

Here is **code for time measurement** to determine the **throughput in images per second**. It doesn't render the images but extracts and prints some basic information in order to make sure the image data are read. We write the information to the null device for longer measurements `null_file=open("/dev/null", mode='w')`. That way it will not clutter our cell output.

We use batches (`dset2 = dset1.batch(batch_size)`) and select a number of batches with (`dset3 = dset2.take(batch_number)`). Then we use the `time.time()` to take the

time measurement and take it multiple times, reading from the same dataset to see if reading speed changes with mutiple readings.

We then **vary** the size of the batch (`batch_size`) and the number of batches (`batch_number`) and **store the results for different values**. Store also the **results for each repetition** over the same dataset (repeat 2 or 3 times).

The speed test should be combined in a **function** `time_configs()` that takes a configuration, i.e. a dataset and arrays of `batch_sizes` , `batch_numbers` , and `repetitions` (an array of integers starting from 1), as **arguments** and runs the time measurement for each combination of batch_size and batch_number for the requested number of repetitions.

In [38]:
```
# Here are some useful values for testing your code, use higher values later for ac
batch_sizes = [2,4]
batch_numbers = [3,6]
repetitions = [1]

def time_configs(dataset, batch_sizes, batch_numbers, repetitions):
    dims = [len(batch_sizes),len(batch_numbers),len(repetitions)]
    print(dims)
    results = np.zeros(dims)
    params = np.zeros(dims + [3])
    print( results.shape )
    with open("/dev/null",mode='w') as null_file: # for printing the output without
        tt = time.time() # for overall time taking
        for bsi,bs in enumerate(batch_sizes):
            for dsi, ds in enumerate(batch_numbers):
                batched_dataset = dataset.batch(bs)
                timing_set = batched_dataset.take(ds)
                for ri,rep in enumerate(repetitions):
                    print("bs: {}, ds: {}, rep: {}".format(bs,ds,rep))
                    t0 = time.time()
                    for image, label in timing_set:
                        #print("Image batch shape {}".format(image.numpy().shape),
                        print("Image batch shape {}, {})".format(image.numpy().shap
                            [str(lbl) for lbl in label.numpy()]), null_file)
                    td = time.time() - t0 # duration for reading images
                    results[bsi,dsi,ri] = ( bs * ds) / td
                    params[bsi,dsi,ri] = [ bs, ds, rep ]
    print("total time: "+str(time.time()-tt))
    return results, params
```

Let's try this function with a **small number** of configurations of batch_sizes batch_numbers and repetions, so that we get a set of parameter combinations and corresponding reading speeds. Try reading from the image files (dataset4) and the TFRecord files (datasetTfrec).

In [39]:
```
[res,par] = time_configs(dataset4, batch_sizes, batch_numbers, repetitions)
print(res)
print(par)

print("=============")

[res,par] = time_configs(datasetTfrec, batch_sizes, batch_numbers, repetitions)
print(res)
print(par)
```

```
[2, 2, 1]
(2, 2, 1)
bs: 2, ds: 3, rep: 1
Image batch shape (2,), ["b'tulips'", "b'tulips'"]) <_io.TextIOWrapper name='/dev/
null' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'roses'", "b'sunflowers'"]) <_io.TextIOWrapper name='/d
ev/null' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'daisy'", "b'sunflowers'"]) <_io.TextIOWrapper name='/d
ev/null' mode='w' encoding='UTF-8'>
bs: 2, ds: 6, rep: 1
Image batch shape (2,), ["b'sunflowers'", "b'sunflowers'"]) <_io.TextIOWrapper nam
e='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'daisy'", "b'daisy'"]) <_io.TextIOWrapper name='/dev/nu
ll' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'sunflowers'", "b'roses'"]) <_io.TextIOWrapper name='/d
ev/null' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'roses'", "b'tulips'"]) <_io.TextIOWrapper name='/dev/n
ull' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'dandelion'", "b'tulips'"]) <_io.TextIOWrapper name='/d
ev/null' mode='w' encoding='UTF-8'>
Image batch shape (2,), ["b'daisy'", "b'roses'"]) <_io.TextIOWrapper name='/dev/nu
ll' mode='w' encoding='UTF-8'>
bs: 4, ds: 3, rep: 1
Image batch shape (4,), ["b'dandelion'", "b'roses'", "b'roses'", "b'roses'"]) <_i
o.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'sunflowers'", "b'sunflowers'", "b'tulips'", "b'tulip
s'"]) <_io.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'roses'", "b'dandelion'", "b'tulips'", "b'daisy'"]) <_i
o.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
bs: 4, ds: 6, rep: 1
Image batch shape (4,), ["b'daisy'", "b'sunflowers'", "b'roses'", "b'roses'"]) <_i
o.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'daisy'", "b'dandelion'", "b'tulips'", "b'dandelion'"])
<_io.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'sunflowers'", "b'dandelion'", "b'tulips'", "b'rose
s'"]) <_io.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'dandelion'", "b'sunflowers'", "b'daisy'", "b'sunflower
s'"]) <_io.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'tulips'", "b'tulips'", "b'daisy'", "b'dandelion'"]) <_
io.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4,), ["b'roses'", "b'dandelion'", "b'tulips'", "b'roses'"]) <_i
o.TextIOWrapper name='/dev/null' mode='w' encoding='UTF-8'>
total time: 32.133134603500366
[[[1.2689509 ]
  [1.60771889]]

 [[1.6569954 ]
  [1.89360162]]]
[[[[2. 3. 1.]]

  [[2. 6. 1.]]]


 [[[4. 3. 1.]]

  [[4. 6. 1.]]]]
=============
[2, 2, 1]
(2, 2, 1)
bs: 2, ds: 3, rep: 1
Image batch shape (2, 192, 192, 3), ['0', '2']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['0', '1']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
```

```
Image batch shape (2, 192, 192, 3), ['4', '4']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
bs: 2, ds: 6, rep: 1
Image batch shape (2, 192, 192, 3), ['0', '2']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['0', '1']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['4', '4']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['3', '1']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['4', '3']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
Image batch shape (2, 192, 192, 3), ['3', '1']) <_io.TextIOWrapper name='/dev/nul
l' mode='w' encoding='UTF-8'>
bs: 4, ds: 3, rep: 1
Image batch shape (4, 192, 192, 3), ['0', '2', '0', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '4', '3', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '3', '3', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
bs: 4, ds: 6, rep: 1
Image batch shape (4, 192, 192, 3), ['0', '2', '0', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '4', '3', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '3', '3', '1']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '0', '4', '3']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['1', '3', '1', '0']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
Image batch shape (4, 192, 192, 3), ['4', '3', '0', '0']) <_io.TextIOWrapper name
='/dev/null' mode='w' encoding='UTF-8'>
total time: 6.201369762420654
[[[ 3.27734665]
  [ 9.88192109]]

 [[ 8.95971358]
  [13.33128713]]]
[[[[2. 3. 1.]]

  [[2. 6. 1.]]]


 [[[4. 3. 1.]]

  [[4. 6. 1.]]]]
```

Task 2: Parallelising the speed test with Spark in the cloud. (36%)

As an exercise in **Spark programming and optimisation** as well as **performance analysis**, we will now implement the **speed test** with multiple parameters in parallel with Spark. Runing multiple tests in parallel would **not be a useful approach on a single machine, but it can be in the cloud** (you will be asked to reason about this later).

2a) Create the script (14%)

Your task is now to **port the speed test above to Spark** for running it in the cloud in Dataproc. **Adapt the speed testing** as a Spark program that performs the same actions as above, but **with Spark RDDs in a distributed way**. The distribution should be such that **each parameter combination (except repetition)** is processed in a separate Spark task.

More specifically:

- i) combine the previous cells to have the code to create a dataset and create a list of parameter combinations in an RDD (2%)
- ii) get a Spark context and create the dataset and run timing test for each combination in parallel (2%)
- iii) transform the resulting RDD to the structure (parameter_combination, images_per_second) and save these values in an array (2%)
- iv) create an RDD with all results for each parameter as (parameter_value,images_per_second) and collect the result for each parameter (2%)
- v) create an RDD with the average reading speeds for each parameter value and collect the results. Keep associativity in mind when implementing the average. (3%)
- vi) write the results to a pickle file in your bucket (2%)
- vii) Write your code it into a file using the *cell magic* `%%writefile spark_job.py` (1%)

Important: The task here is not to parallelize the pre-processing, but to run multiple speed tests in parallel using Spark.

In [47]:
```
### CODING TASK
# Importing necessary libraries
import os
import time
import pyspark
import tensorflow as tf
from pyspark.sql import SparkSession

# i)  Creating a Dataset and a list of Parameter Combinations

# Constants for config
TFRECORDS_GCS_PATTERN = 'gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/
JPEG_GCS_PATTERN = 'gs://flowers-public/**/*.jpg'
TARGET_SIZE = [192, 192]
BATCH_SIZES = [2, 4, 6, 8]
BATCH_NUMBERS = [6, 9, 12, 15]
REPETITIONS = [1, 2, 3]

# JPEG images functions
def decode_jpeg_and_label(filepath):
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/').values[-2]
    return image, label

def resize_and_crop_image(image, label):
    w, h = tf.shape(image)[0], tf.shape(image)[1]
    resize_crit = (w * TARGET_SIZE[0]) / (h * TARGET_SIZE[1])
    image = tf.cond(
        resize_crit < 1,
        lambda: tf.image.resize(image, [w * TARGET_SIZE[1] / w, h * TARGET_SIZE[1]
        lambda: tf.image.resize(image, [w * TARGET_SIZE[0] / h, h * TARGET_SIZE[0]
```

```
    )
    nw, nh = tf.shape(image)[0], tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (nw - TARGET_SIZE[1]) // 2, (nh -
    return image, label

def recompress_image(image, label):
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, format='rgb', quality=70)
    image = tf.image.decode_jpeg(image)
    return image, label

# TFrecords functions
def decode_tfrecord(record):
    features = {
        "image": tf.io.FixedLenFeature([], tf.string),
        "class": tf.io.FixedLenFeature([], tf.int64)
    }
    example = tf.io.parse_single_example(record, features)
    image = tf.image.decode_jpeg(example['image'], channels=3)
    image = tf.reshape(image, [*TARGET_SIZE, 3])
    return image, example['class']

def load_tfrecord_dataset():
    dataset = tf.data.TFRecordDataset(tf.io.gfile.glob(TFRECORDS_GCS_PATTERN))
    dataset = dataset.map(decode_tfrecord)
    return dataset

# ii) Creating dataset, running time configs and creating a spark context

# Recreating dataset4
def load_jpeg_dataset():
    dataset = tf.data.Dataset.list_files(JPEG_GCS_PATTERN)
    dataset4 = dataset4.map(decode_jpeg_and_label)
    dataset4 = dataset4.map(resize_and_crop_image)
    dataset4 = dataset4.map(recompress_image)
    return dataset4

#adaption of time configs
def time_configs(dataset_loader, batch_sizes, batch_numbers, repetitions):
    results = []
    for bsize in batch_sizes:
        dataset = dataset_loader().batch(bsize)
        for bnumber in batch_numbers:
            for rep in repetitions:
                total_time = 0
                for _ in range(rep):
                    start_time = time.time()
                    for _ in range(bnumber):
                        for _ in dataset.take(1):
                            pass
                    end_time = time.time()
                    total_time += (end_time - start_time)
                avg_time = total_time / rep
                throughput = (bsize * bnumber) / avg_time if avg_time > 0 else 0
                results.append(((bsize, bnumber, rep), throughput))
    return results

# Creating a Spark context and session
spark_ctx = pyspark.SparkContext.getOrCreate()
spark_s = SparkSession(spark_ctx)

# Generating all parameter combinations
parameter_combinations = [(bsize, bnumber, rep) for bsize in BATCH_SIZES for bnumbe
```

```
# Parallelizing the parameter combinations into an RDD for both TFRecords and JPEGs
params_rdd = spark_ctx.parallelize(parameter_combinations)
processed_tfrecord_images_rdd = params_rdd.flatMap(lambda params: time_configs(loac
processed_jpeg_images_rdd = params_rdd.flatMap(lambda params: time_configs(load_jpe

# iii) Transforming the resulting RDDs to the structure and saving them to an array

tfrecord_array = processed_tfrecord_images_rdd.collect()
jpeg_array = processed_jpeg_images_rdd.collect()

# Printing the results for verification
print("TFRecord Results Array:")
print(tfrecord_array)
print("JPEG Results Array:")
print(jpeg_array)

# iv) Creating an RDD with all results for each parameter

#Expanding the tuple results
def expand_results(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), throughput),
        (('num_batches', num_batches), throughput),
        (('repetitions', repetitions), throughput)
    ]

# Transformation
tfrecord_results = processed_tfrecord_images_rdd.flatMap(expand_results)
jpeg_results = processed_jpeg_images_rdd.flatMap(expand_results)

# Collecting results by parameter
tfrecord_results_param = tfrecord_results.groupByKey().mapValues(list).collect()
jpeg_results_param = jpeg_results.groupByKey().mapValues(list).collect()

# Printing results(parameters)
print("TFRecord Results by Parameter:")
for key, values in tfrecord_results_param:
    print(f"{key}: {values}")

print("JPEG Results by Parameter:")
for key, values in jpeg_results_param:
    print(f"{key}: {values}")


#v) Creating an RDD with the average reading speeds for each parameter value and co

# Summing the throughputs and counting occurrences
def throughputs_counting_avg(a, b):
    throughput_sum_a, count_a = a
    throughput_sum_b, count_b = b
    return (throughput_sum_a + throughput_sum_b, count_a + count_b)

# Mapping results for averaging
def avg_mapping(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), (throughput, 1)),
        (('num_batches', num_batches), (throughput, 1)),
        (('repetitions', repetitions), (throughput, 1))
    ]

# Calculating averages from sums and counts
def cal_avg(a):
```

```
    throughput_sum, count = a
    return throughput_sum / count if count != 0 else 0

# Mapping the results for averaging and reducing to calculate sums and counts
mapped_tfrecord_results = processed_tfrecord_images_rdd.flatMap(avg_mapping)
mapped_jpeg_results = processed_jpeg_images_rdd.flatMap(avg_mapping)

# Reducing to sum the throughputs and count occurrences for each parameter(by key)
reduced_tfrecord_results = mapped_tfrecord_results.reduceByKey(throughputs_counting
reduced_jpeg_results = mapped_jpeg_results.reduceByKey(throughputs_counting_avg)

# Calculating the throughput for each parameter(average)
avg_tfrecord_results = reduced_tfrecord_results.mapValues(cal_avg).collect()
avg_jpeg_results = reduced_jpeg_results.mapValues(cal_avg).collect()

# Printing the results
print("Average TFRecord Results by Parameter:")
for param, avg_throughput in avg_tfrecord_results:
    print(f"{param}: {avg_throughput}")

print("Average JPEG Results by Parameter:")
for param, avg_throughput in avg_jpeg_results:
    print(f"{param}: {avg_throughput}")

# vi) Writing the results to a pickle file in the bucket.

import pickle
import gcsfs

bucket_path = 'gs://earnest-crow-421721-storage'

#Saving the average speeds for images and tfrecords
def save_data_gcs(data, file_path):
    fs = gcsfs.GCSFileSystem(project='earnest-crow-421721')
    with fs.open(file_path, 'wb') as file:
        pickle.dump(data, file)

# Defining pickle files paths
picklepath_tfrecord = f'{bucket_path}/tfrecord_results.pickle'
picklepath_jpeg = f'{bucket_path}/jpeg_results.pickle'

# Saving the average results in Google Cloud Storage
save_data_gcs(avg_tfrecord_results, picklepath_tfrecord)
save_data_gcs(avg_jpeg_results, picklepath_jpeg)
```

```
TFRecord Results Array:
[((2, 6, 1), 73.94622802841981), ((2, 6, 2), 76.60152984972407), ((2, 6, 3), 80.21
729734094662), ((2, 9, 1), 83.21609794917806), ((2, 9, 2), 60.53041815492297),
((2, 9, 3), 43.687978305347336), ((2, 12, 1), 50.69755762213778), ((2, 12, 2), 50.
10701823724244), ((2, 12, 3), 49.13291141608883), ((2, 15, 1), 76.06419589795732),
((2, 15, 2), 74.62683061791306), ((2, 15, 3), 83.21191555173661), ((4, 6, 1), 155.
08300967657948), ((4, 6, 2), 134.9178617572791), ((4, 6, 3), 160.36345736074333),
((4, 9, 1), 156.13393638207978), ((4, 9, 2), 141.52355110413154), ((4, 9, 3), 168.
45521470019918), ((4, 12, 1), 147.44847638389544), ((4, 12, 2), 152.389459947984),
((4, 12, 3), 150.25856434080677), ((4, 15, 1), 163.79542546364337), ((4, 15, 2), 1
23.43258381087402), ((4, 15, 3), 91.34510841454548), ((6, 6, 1), 157.0712451173652
7), ((6, 6, 2), 160.08199865356988), ((6, 6, 3), 207.94086753264162), ((6, 9, 1),
236.9574027343603), ((6, 9, 2), 221.33238347779053), ((6, 9, 3), 233.9818634739842
5), ((6, 12, 1), 235.29283875511896), ((6, 12, 2), 228.09748094906837), ((6, 12,
3), 233.05956852180748), ((6, 15, 1), 241.1665312255631), ((6, 15, 2), 219.5868487
7957978), ((6, 15, 3), 226.744101953168), ((8, 6, 1), 354.27794739500206), ((8, 6,
2), 314.20385594103175), ((8, 6, 3), 311.18140886433014), ((8, 9, 1), 271.27187472
20942), ((8, 9, 2), 224.54666302819493), ((8, 9, 3), 183.05153982615585), ((8, 12,
1), 203.47611366475448), ((8, 12, 2), 167.13426422487507), ((8, 12, 3), 200.378434
6436719), ((8, 15, 1), 316.28091461448986), ((8, 15, 2), 288.69383366605933), ((8,
15, 3), 320.30825292301677)]
JPEG Results Array:
[((2, 6, 1), 1.66550065583766), ((2, 6, 2), 1.9880115335651998), ((2, 6, 3), 1.902
2265160509704), ((2, 9, 1), 1.5630209083923348), ((2, 9, 2), 1.6496164611627298),
((2, 9, 3), 1.6834524212022748), ((2, 12, 1), 1.8754705244225924), ((2, 12, 2), 1.
9431209444216018), ((2, 12, 3), 1.6459548811110143), ((2, 15, 1), 1.88132459275048
18), ((2, 15, 2), 2.146437001816791), ((2, 15, 3), 1.735055315346659), ((4, 6, 1),
2.7391173310627286), ((4, 6, 2), 2.9568235648948358), ((4, 6, 3), 2.76144848236088
16), ((4, 9, 1), 2.461863102230397), ((4, 9, 2), 2.8735473054011185), ((4, 9, 3),
2.9848335332250375), ((4, 12, 1), 2.640344938518777), ((4, 12, 2), 3.0274718738320
003), ((4, 12, 3), 2.755750872035006), ((4, 15, 1), 2.7447514169362948), ((4, 15,
2), 3.00834029508503), ((4, 15, 3), 2.9135471260320167), ((6, 6, 1), 3.51525791752
22107), ((6, 6, 2), 3.3740515064421013), ((6, 6, 3), 3.3293277802598307), ((6, 9,
1), 3.1854607384584046), ((6, 9, 2), 3.008640047502239), ((6, 9, 3), 3.51970454336
2407), ((6, 12, 1), 3.4275139937484136), ((6, 12, 2), 3.4373825708965398), ((6, 1
2, 3), 3.487622544377441), ((6, 15, 1), 3.3832520886477937), ((6, 15, 2), 3.259458
295466586), ((6, 15, 3), 3.5879337483100624), ((8, 6, 1), 3.878795725380313), ((8,
6, 2), 3.9692651479992858), ((8, 6, 3), 3.573904612002001), ((8, 9, 1), 3.70095216
92288856), ((8, 9, 2), 3.863089469308095), ((8, 9, 3), 3.83633187643117), ((8, 12,
1), 4.354400331954838), ((8, 12, 2), 4.051502505909698), ((8, 12, 3), 2.7241733539
39358), ((8, 15, 1), 2.2636374812919837), ((8, 15, 2), 3.019785722589403), ((8, 1
5, 3), 3.8837067232138898)]
TFRecord Results by Parameter:
('num_batches', 9): [25.026318956032753, 27.722982681552143, 27.358683260394532, 1
56.99122379276466, 117.403647254762, 142.62445722964912, 182.5399090735447, 185.90
910833282373, 181.4394324480644, 265.08431894749225, 356.5424940274181, 264.018331
52895335]
('num_batches', 15): [54.23788195343949, 60.577203738067595, 68.43180592053741, 11
2.75904944074865, 156.1655219660884, 71.10320170073253, 128.20497180245243, 149.03
622056756868, 225.2084127555157, 339.6894369695268, 323.29094654408607, 325.117582
3877693]
('repetitions', 3): [50.262453355880076, 27.358683260394532, 59.47382571694946, 6
8.43180592053741, 71.40513651842431, 142.62445722964912, 148.83387611875568, 71.10
320170073253, 182.36559527653264, 181.4394324480644, 128.48000147487502, 225.20841
27555157, 234.05605442683023, 264.01833152895335, 205.87768989998065, 325.11758238
77693]
('repetitions', 1): [50.001090788795274, 25.026318956032753, 32.22812642770312, 5
4.23788195343949, 61.26649138182881, 156.99122379276466, 164.86814501566982, 112.7
5904944074865, 181.22967129117862, 182.5399090735447, 210.7541888797621, 128.20497
180245243, 224.257355070688, 265.08431894749225, 338.7833713495781, 339.6894369695
268]
('batch_size', 2): [50.001090788795274, 51.24370915743867, 50.262453355880076, 25.
026318956032753, 27.722982681552143, 27.358683260394532, 32.22812642770312, 66.830
91422772688, 59.47382571694946, 54.23788195343949, 60.577203738067595, 68.43180592
```

```
053741]
('batch_size', 6): [181.22967129117862, 255.89153902730843, 182.36559527653264, 18
2.5399090735447, 185.90910833282373, 181.4394324480644, 210.7541888797621, 207.690
86311854147, 128.48000147487502, 128.20497180245243, 149.03622056756868, 225.20841
27555157]
('batch_size', 8): [224.257355070688, 361.5032671053931, 234.05605442683023, 265.0
8431894749225, 356.5424940274181, 264.01833152895335, 338.7833713495781, 114.52135
316439352, 205.87768989998065, 339.6894369695268, 323.29094654408607, 325.11758238
77693]
('num_batches', 6): [50.001090788795274, 51.24370915743867, 50.262453355880076, 6
1.26649138182881, 78.05334840692176, 71.40513651842431, 181.22967129117862, 255.89
153902730843, 182.36559527653264, 224.257355070688, 361.5032671053931, 234.0560544
2683023]
('num_batches', 12): [32.22812642770312, 66.83091422772688, 59.47382571694946, 16
4.86814501566982, 118.55729703866736, 148.83387611875568, 210.7541888797621, 207.6
9086311854147, 128.48000147487502, 338.7833713495781, 114.52135316439352, 205.8776
8989998065]
('batch_size', 4): [61.26649138182881, 78.05334840692176, 71.40513651842431, 156.9
9122379276466, 117.403647254762, 142.62445722964912, 164.86814501566982, 118.55729
703866736, 148.83387611875568, 112.75904944074865, 156.1655219660884, 71.103201700
73253]
('repetitions', 2): [51.24370915743867, 27.722982681552143, 66.83091422772688, 60.
577203738067595, 78.05334840692176, 117.403647254762, 118.55729703866736, 156.1655
219660884, 255.89153902730843, 185.90910833282373, 207.69086311854147, 149.0362205
6756868, 361.5032671053931, 356.5424940274181, 114.52135316439352, 323.29094654408
607]
JPEG Results by Parameter:
('num_batches', 15): [1.6314037843490818, 1.120514366977175, 1.7062500168653205,
2.62541794575748, 2.823735673916817, 2.760913266380858, 3.3618269657545445, 2.1673
46184025169, 3.2319593139290674, 3.7730680548783586, 3.465221307050728, 3.93542336
78397943]
('num_batches', 9): [1.6634618138783455, 1.6705349770900584, 1.1815960826915826,
1.2686576127971005, 1.8933070695682113, 2.30983811913568, 3.664313683261068, 1.897
7728617437488, 2.71477794086879, 3.8819441211168044, 3.6472050782243897, 3.6559915
107778647]
('repetitions', 3): [1.240436022768366, 1.1815960826915826, 1.6668307501227977, 1.
7062500168653205, 2.7269169339653963, 2.30983811913568, 2.7670904412103936, 2.7609
13266380858, 2.8429542240738543, 2.71477794086879, 3.0151805044334257, 3.231959313
9290674, 3.7741941978669566, 3.6559915107778647, 3.992124266537034, 3.935423367839
7943]
('repetitions', 1): [0.8466648200905419, 1.6634618138783455, 1.021016944614028, 1.
6314037843490818, 3.3786984435168823, 1.2686576127971005, 2.599097340276581, 2.625
41794575748, 1.372536518965763, 3.664313683261068, 3.1001041228443635, 3.361826965
7545445, 3.604515135809161, 3.8819441211168044, 4.184673323069732, 3.7730680548783
586]
('batch_size', 4): [3.3786984435168823, 2.9807411370456283, 2.7269169339653963, 1.
2686576127971005, 1.8933070695682113, 2.30983811913568, 2.599097340276581, 3.01964
68678215287, 2.7670904412103936, 2.62541794575748, 2.823735673916817, 2.7609132663
80858]
('repetitions', 2): [0.9373009842799958, 1.6705349770900584, 1.8229729391642906,
1.120514366977175, 2.9807411370456283, 1.8933070695682113, 3.0196468678215287, 2.8
23735673916817, 2.1687766002713933, 1.8977728617437488, 2.3472254481460055, 2.1673
46184025169, 3.5231528196335105, 3.6472050782243897, 4.254602531654986, 3.46522130
7050728]
('num_batches', 12): [1.021016944614028, 1.8229729391642906, 1.6668307501227977,
2.599097340276581, 3.0196468678215287, 2.7670904412103936, 3.1001041228443635, 2.3
472254481460055, 3.0151805044334257, 4.184673323069732, 4.254602531654986, 3.99212
4266537034]
('batch_size', 2): [0.8466648200905419, 0.9373009842799958, 1.240436022768366, 1.6
634618138783455, 1.6705349770900584, 1.1815960826915826, 1.021016944614028, 1.8229
729391642906, 1.6668307501227977, 1.6314037843490818, 1.120514366977175, 1.7062500
168653205]
('num_batches', 6): [0.8466648200905419, 0.9373009842799958, 1.240436022768366, 3.
3786984435168823, 2.9807411370456283, 2.7269169339653963, 1.372536518965763, 2.168
```

```
7766002713933, 2.8429542240738543, 3.604515135809161, 3.5231528196335105, 3.774194
1978669566]
('batch_size', 6): [1.372536518965763, 2.1687766002713933, 2.8429542240738543, 3.6
64313683261068, 1.8977728617437488, 2.71477794086879, 3.1001041228443635, 2.347225
4481460055, 3.0151805044334257, 3.3618269657545445, 2.167346184025169, 3.231959313
9290674]
('batch_size', 8): [3.604515135809161, 3.5231528196335105, 3.7741941978669566, 3.8
819441211168044, 3.6472050782243897, 3.6559915107778647, 4.184673323069732, 4.2546
02531654986, 3.992124266537034, 3.7730680548783586, 3.465221307050728, 3.935423367
8397943]
Average TFRecord Results by Parameter:
('repetitions', 3): 132.1426551370059
('repetitions', 1): 133.61532723620144
('num_batches', 15): 129.51207815347155
('num_batches', 9): 133.3790702530305
('batch_size', 6): 134.2376497583326
('batch_size', 8): 223.91551032208608
('num_batches', 6): 120.40067403331068
('repetitions', 2): 130.80733078358645
('num_batches', 12): 145.46192843591237
('batch_size', 4): 121.27231042460484
('batch_size', 2): 49.328280370701584
Average JPEG Results by Parameter:
('num_batches', 15): 2.802969035576769
('num_batches', 9): 2.942273894350001
('repetitions', 3): 2.8050228505674726
('repetitions', 1): 2.987737896289091
('num_batches', 6): 3.028042469861689
('batch_size', 4): 2.882161284148593
('repetitions', 2): 2.924392029542349
('num_batches', 12): 2.8495849687434247
('batch_size', 2): 1.897810610657784
('batch_size', 8): 3.442422189426671
('batch_size', 6): 3.4004762842988367
```

In [7]:

```
# vii)

%%writefile spark_job.py

### CODING TASK
# Importing necessary libraries
import os
import time
import pyspark
import tensorflow as tf
from pyspark.sql import SparkSession

# i)  Creating a Dataset and a list of Parameter Combinations

# Constants for config
TFRECORDS_GCS_PATTERN = 'gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/
JPEG_GCS_PATTERN = 'gs://flowers-public/**/*.jpg'
TARGET_SIZE = [192, 192]
BATCH_SIZES = [2, 4, 6, 8]
BATCH_NUMBERS = [6, 9, 12, 15]
REPETITIONS = [1, 2, 3]

# JPEG images functions
def decode_jpeg_and_label(filepath):
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/').values[-2]
    return image, label
```

```
def resize_and_crop_image(image, label):
    w, h = tf.shape(image)[0], tf.shape(image)[1]
    resize_crit = (w * TARGET_SIZE[0]) / (h * TARGET_SIZE[1])
    image = tf.cond(
        resize_crit < 1,
        lambda: tf.image.resize(image, [w * TARGET_SIZE[1] / w, h * TARGET_SIZE[1]
        lambda: tf.image.resize(image, [w * TARGET_SIZE[0] / h, h * TARGET_SIZE[0]
    )
    nw, nh = tf.shape(image)[0], tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (nw - TARGET_SIZE[1]) // 2, (nh -
    return image, label

def recompress_image(image, label):
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, format='rgb', quality=70)
    image = tf.image.decode_jpeg(image)
    return image, label

# TFrecords functions
def decode_tfrecord(record):
    features = {
        "image": tf.io.FixedLenFeature([], tf.string),
        "class": tf.io.FixedLenFeature([], tf.int64)
    }
    example = tf.io.parse_single_example(record, features)
    image = tf.image.decode_jpeg(example['image'], channels=3)
    image = tf.reshape(image, [*TARGET_SIZE, 3])
    return image, example['class']

def load_tfrecord_dataset():
    dataset = tf.data.TFRecordDataset(tf.io.gfile.glob(TFRECORDS_GCS_PATTERN))
    dataset = dataset.map(decode_tfrecord)
    return dataset

# ii) Creating dataset, running time configs and creating a spark context

# Recreating dataset4
def load_jpeg_dataset():
    dataset = tf.data.Dataset.list_files(JPEG_GCS_PATTERN)
    dataset4 = dataset4.map(decode_jpeg_and_label)
    dataset4 = dataset4.map(resize_and_crop_image)
    dataset4 = dataset4.map(recompress_image)
    return dataset4

#adaption of time configs
def time_configs(dataset_loader, batch_sizes, batch_numbers, repetitions):
    results = []
    for bsize in batch_sizes:
        dataset = dataset_loader().batch(bsize)
        for bnumber in batch_numbers:
            for rep in repetitions:
                total_time = 0
                for _ in range(rep):
                    start_time = time.time()
                    for _ in range(bnumber):
                        for _ in dataset.take(1):
                            pass
                    end_time = time.time()
                    total_time += (end_time - start_time)
                avg_time = total_time / rep
                throughput = (bsize * bnumber) / avg_time if avg_time > 0 else 0
                results.append(((bsize, bnumber, rep), throughput))
    return results
```

```
# Creating a Spark context and session
spark_ctx = pyspark.SparkContext.getOrCreate()
spark_s = SparkSession(spark_ctx)

# Generating all parameter combinations
parameter_combinations = [(bsize, bnumber, rep) for bsize in BATCH_SIZES for bnumbe

# Parallelizing the parameter combinations into an RDD for both TFRecords and JPEGs
params_rdd = spark_ctx.parallelize(parameter_combinations)
processed_tfrecord_images_rdd = params_rdd.flatMap(lambda params: time_configs(loac
processed_jpeg_images_rdd = params_rdd.flatMap(lambda params: time_configs(load_jpe

# iii) Transforming the resulting RDDs to the structure and saving them to an array

tfrecord_array = processed_tfrecord_images_rdd.collect()
jpeg_array = processed_jpeg_images_rdd.collect()

# Printing the results for verification
print("TFRecord Results Array:")
print(tfrecord_array)
print("JPEG Results Array:")
print(jpeg_array)

# iv) Creating an RDD with all results for each parameter

#Expanding the tuple results
def expand_results(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), throughput),
        (('num_batches', num_batches), throughput),
        (('repetitions', repetitions), throughput)
    ]

# Transformation
tfrecord_results = processed_tfrecord_images_rdd.flatMap(expand_results)
jpeg_results = processed_jpeg_images_rdd.flatMap(expand_results)

# Collecting results by parameter
tfrecord_results_param = tfrecord_results.groupByKey().mapValues(list).collect()
jpeg_results_param = jpeg_results.groupByKey().mapValues(list).collect()

# Printing results(parameters)
print("TFRecord Results by Parameter:")
for key, values in tfrecord_results_param:
    print(f"{key}: {values}")

print("JPEG Results by Parameter:")
for key, values in jpeg_results_param:
    print(f"{key}: {values}")


#v) Creating an RDD with the average reading speeds for each parameter value and co

# Summing the throughputs and counting occurrences
def throughputs_counting_avg(a, b):
    throughput_sum_a, count_a = a
    throughput_sum_b, count_b = b
    return (throughput_sum_a + throughput_sum_b, count_a + count_b)

# Mapping results for averaging
def avg_mapping(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
```

```
        (('batch_size', batch_size), (throughput, 1)),
        (('num_batches', num_batches), (throughput, 1)),
        (('repetitions', repetitions), (throughput, 1))
    ]

# Calculating averages from sums and counts
def cal_avg(a):
    throughput_sum, count = a
    return throughput_sum / count if count != 0 else 0

# Mapping the results for averaging and reducing to calculate sums and counts
mapped_tfrecord_results = processed_tfrecord_images_rdd.flatMap(avg_mapping)
mapped_jpeg_results = processed_jpeg_images_rdd.flatMap(avg_mapping)

# Reducing to sum the throughputs and count occurrences for each parameter(by key)
reduced_tfrecord_results = mapped_tfrecord_results.reduceByKey(throughputs_counting
reduced_jpeg_results = mapped_jpeg_results.reduceByKey(throughputs_counting_avg)

# Calculating the throughput for each parameter(average)
avg_tfrecord_results = reduced_tfrecord_results.mapValues(cal_avg).collect()
avg_jpeg_results = reduced_jpeg_results.mapValues(cal_avg).collect()

# Printing the results
print("Average TFRecord Results by Parameter:")
for param, avg_throughput in avg_tfrecord_results:
    print(f"{param}: {avg_throughput}")

print("Average JPEG Results by Parameter:")
for param, avg_throughput in avg_jpeg_results:
    print(f"{param}: {avg_throughput}")

# vi) Writing the results to a pickle file in the bucket.

import pickle
import gcsfs

bucket_path = 'gs://earnest-crow-421721-storage'

#Saving the average speeds for images and tfrecords
def saving_data_gcs(data, file_path):
    fs = gcsfs.GCSFileSystem(project='earnest-crow-421721')
    with fs.open(file_path, 'wb') as file:
        pickle.dump(data, file)

# Defining pickle files paths
picklepath_tfrecord = f'{bucket_path}/tfrecord_results.pickle'
picklepath_jpeg = f'{bucket_path}/jpeg_results.pickle'

# Saving the average results in Google Cloud Storage
saving_data_gcs(avg_tfrecord_results, picklepath_tfrecord)
saving_data_gcs(avg_jpeg_results, picklepath_jpeg)
```

```
Writing spark_job.py
```

2b) Testing the code and collecting results (4%)

i) First, test locally with `%run`.

It is useful to create a **new filename argument**, so that old results don't get overwritten.

You can for instance use `datetime.datetime.now().strftime("%y%m%d-%H%M")` to get a string with the current date and time and use that in the file name.

In [9]:
```
%%writefile spark_job2.py

### CODING TASK
# Importing necessary libraries
import os
import time
import pyspark
import tensorflow as tf
from pyspark.sql import SparkSession

# i)  Creating a Dataset and a list of Parameter Combinations

# Constants for config
TFRECORDS_GCS_PATTERN = 'gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/
JPEG_GCS_PATTERN = 'gs://flowers-public/**/*.jpg'
TARGET_SIZE = [192, 192]
BATCH_SIZES = [2, 4, 6, 8]
BATCH_NUMBERS = [6, 9, 12, 15]
REPETITIONS = [1, 2, 3]

# JPEG images functions
def decode_jpeg_and_label(filepath):
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/').values[-2]
    return image, label

def resize_and_crop_image(image, label):
    w, h = tf.shape(image)[0], tf.shape(image)[1]
    resize_crit = (w * TARGET_SIZE[0]) / (h * TARGET_SIZE[1])
    image = tf.cond(
        resize_crit < 1,
        lambda: tf.image.resize(image, [w * TARGET_SIZE[1] / w, h * TARGET_SIZE[1]
        lambda: tf.image.resize(image, [w * TARGET_SIZE[0] / h, h * TARGET_SIZE[0]
    )
    nw, nh = tf.shape(image)[0], tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (nw - TARGET_SIZE[1]) // 2, (nh -
    return image, label

def recompress_image(image, label):
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, format='rgb', quality=70)
    image = tf.image.decode_jpeg(image)
    return image, label

# TFrecords functions
def decode_tfrecord(record):
    features = {
        "image": tf.io.FixedLenFeature([], tf.string),
        "class": tf.io.FixedLenFeature([], tf.int64)
    }
    example = tf.io.parse_single_example(record, features)
    image = tf.image.decode_jpeg(example['image'], channels=3)
    image = tf.reshape(image, [*TARGET_SIZE, 3])
    return image, example['class']

def load_tfrecord_dataset():
    dataset = tf.data.TFRecordDataset(tf.io.gfile.glob(TFRECORDS_GCS_PATTERN))
    dataset = dataset.map(decode_tfrecord)
    return dataset

# ii) Creating dataset, running time configs and creating a spark context
```

```
# Recreating dataset4
def load_jpeg_dataset():
    dataset = tf.data.Dataset.list_files(JPEG_GCS_PATTERN)
    dataset4 = dataset4.map(decode_jpeg_and_label)
    dataset4 = dataset4.map(resize_and_crop_image)
    dataset4 = dataset4.map(recompress_image)
    return dataset4

#adaption of time configs
def time_configs(dataset_loader, batch_sizes, batch_numbers, repetitions):
    results = []
    for bsize in batch_sizes:
        dataset = dataset_loader().batch(bsize)
        for bnumber in batch_numbers:
            for rep in repetitions:
                total_time = 0
                for _ in range(rep):
                    start_time = time.time()
                    for _ in range(bnumber):
                        for _ in dataset.take(1):
                            pass
                    end_time = time.time()
                    total_time += (end_time - start_time)
                avg_time = total_time / rep
                throughput = (bsize * bnumber) / avg_time if avg_time > 0 else 0
                results.append(((bsize, bnumber, rep), throughput))
    return results

# Creating a Spark context and session
spark_ctx = pyspark.SparkContext.getOrCreate()
spark_s = SparkSession(spark_ctx)

# Generating all parameter combinations
parameter_combinations = [(bsize, bnumber, rep) for bsize in BATCH_SIZES for bnumbe

# Parallelizing the parameter combinations into an RDD for both TFRecords and JPEGs
params_rdd = spark_ctx.parallelize(parameter_combinations)
processed_tfrecord_images_rdd = params_rdd.flatMap(lambda params: time_configs(loac
processed_jpeg_images_rdd = params_rdd.flatMap(lambda params: time_configs(load_jpe

# iii) Transforming the resulting RDDs to the structure and saving them to an array

tfrecord_array = processed_tfrecord_images_rdd.collect()
jpeg_array = processed_jpeg_images_rdd.collect()

# Printing the results for verification
print("TFRecord Results Array:")
print(tfrecord_array)
print("JPEG Results Array:")
print(jpeg_array)

# iv) Creating an RDD with all results for each parameter

#Expanding the tuple results
def expand_results(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), throughput),
        (('num_batches', num_batches), throughput),
        (('repetitions', repetitions), throughput)
    ]

# Transformation
tfrecord_results = processed_tfrecord_images_rdd.flatMap(expand_results)
```

```
jpeg_results = processed_jpeg_images_rdd.flatMap(expand_results)

# Collecting results by parameter
tfrecord_results_param = tfrecord_results.groupByKey().mapValues(list).collect()
jpeg_results_param = jpeg_results.groupByKey().mapValues(list).collect()

# Printing results(parameters)
print("TFRecord Results by Parameter:")
for key, values in tfrecord_results_param:
    print(f"{key}: {values}")

print("JPEG Results by Parameter:")
for key, values in jpeg_results_param:
    print(f"{key}: {values}")


#v) Creating an RDD with the average reading speeds for each parameter value and co

# Summing the throughputs and counting occurrences
def throughputs_counting_avg(a, b):
    throughput_sum_a, count_a = a
    throughput_sum_b, count_b = b
    return (throughput_sum_a + throughput_sum_b, count_a + count_b)

# Mapping results for averaging
def avg_mapping(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), (throughput, 1)),
        (('num_batches', num_batches), (throughput, 1)),
        (('repetitions', repetitions), (throughput, 1))
    ]

# Calculating averages from sums and counts
def cal_avg(a):
    throughput_sum, count = a
    return throughput_sum / count if count != 0 else 0

# Mapping the results for averaging and reducing to calculate sums and counts
mapped_tfrecord_results = processed_tfrecord_images_rdd.flatMap(avg_mapping)
mapped_jpeg_results = processed_jpeg_images_rdd.flatMap(avg_mapping)

# Reducing to sum the throughputs and count occurrences for each parameter(by key)
reduced_tfrecord_results = mapped_tfrecord_results.reduceByKey(throughputs_counting
reduced_jpeg_results = mapped_jpeg_results.reduceByKey(throughputs_counting_avg)

# Calculating the throughput for each parameter(average)
avg_tfrecord_results = reduced_tfrecord_results.mapValues(cal_avg).collect()
avg_jpeg_results = reduced_jpeg_results.mapValues(cal_avg).collect()

# Printing the results
print("Average TFRecord Results by Parameter:")
for param, avg_throughput in avg_tfrecord_results:
    print(f"{param}: {avg_throughput}")

print("Average JPEG Results by Parameter:")
for param, avg_throughput in avg_jpeg_results:
    print(f"{param}: {avg_throughput}")

# New saving code that includes the time

import pickle
from datetime import datetime
import gcsfs
```

```
def saving_data_gcs(data, b_path, filename_b):
    # Creating a timestamp
    timestamp = datetime.now().strftime("%y%m%d-%H%M")
    file_path = f'{b_path}/{filename_b}_{timestamp}.pickle'

    # Saving data to GCS using gcsfs
    fs = gcsfs.GCSFileSystem(project='earnest-crow-421721')
    with fs.open(file_path, 'wb') as file:
        pickle.dump(data, file)
    print(f"Saving file as {file_path}")

# Defining the base names for the pickle files
tfrecord_filename = 'tfrecord_base_results'
jpeg_filename = 'jpeg_base_results'

# Calling the function to save data to GCS
saving_data_gcs(avg_tfrecord_results, 'gs://earnest-crow-421721-storage', tfrecord_
saving_data_gcs(avg_jpeg_results, 'gs://earnest-crow-421721-storage', jpeg_filename
```

```
Overwriting spark_job2.py
```

In [10]:
```
%run ./spark_job2.py
```

```
TFRecord Results Array:
[((2, 6, 1), 34.00265499818609), ((2, 6, 2), 36.96176458924803), ((2, 6, 3), 65.54
0096256016), ((2, 9, 1), 60.777080718690904), ((2, 9, 2), 69.68967395217375), ((2,
9, 3), 68.33242954509593), ((2, 12, 1), 65.66831995459602), ((2, 12, 2), 65.293995
6268855), ((2, 12, 3), 65.88867887684405), ((2, 15, 1), 69.84360429488717), ((2, 1
5, 2), 64.58984290283146), ((2, 15, 3), 64.13251120912435), ((4, 6, 1), 146.832392
50487918), ((4, 6, 2), 123.34238337321673), ((4, 6, 3), 121.90263609149301), ((4,
9, 1), 125.47934632212707), ((4, 9, 2), 83.94327801693765), ((4, 9, 3), 71.5272223
8810708), ((4, 12, 1), 76.60380327789011), ((4, 12, 2), 79.44835946798334), ((4, 1
2, 3), 135.2764314136726), ((4, 15, 1), 137.24255388768458), ((4, 15, 2), 122.5717
4522763918), ((4, 15, 3), 131.61234178097857), ((6, 6, 1), 103.23995889407841),
((6, 6, 2), 106.76162688955847), ((6, 6, 3), 195.65545346289355), ((6, 9, 1), 208.
25395144072425), ((6, 9, 2), 198.62842349524064), ((6, 9, 3), 200.9000291529156),
((6, 12, 1), 209.15020396290575), ((6, 12, 2), 194.83463695188422), ((6, 12, 3), 2
01.55398969014595), ((6, 15, 1), 200.11119663611092), ((6, 15, 2), 199.88359236996
277), ((6, 15, 3), 195.65692392137595), ((8, 6, 1), 292.1180505980882), ((8, 6,
2), 254.84878753950235), ((8, 6, 3), 267.4975501452906), ((8, 9, 1), 258.946283138
9456), ((8, 9, 2), 142.79386400840525), ((8, 9, 3), 140.37390352457206), ((8, 12,
1), 129.31322231293007), ((8, 12, 2), 154.0905558524577), ((8, 12, 3), 259.3816529
4345765), ((8, 15, 1), 262.46004563828944), ((8, 15, 2), 231.81296396769002), ((8,
15, 3), 226.5384826004368)]
JPEG Results Array:
[((2, 6, 1), 4.610972197105324), ((2, 6, 2), 3.8353559163095285), ((2, 6, 3), 5.29
285372817615), ((2, 9, 1), 3.82983537598905), ((2, 9, 2), 2.4734711176047455),
((2, 9, 3), 4.236057722735819), ((2, 12, 1), 3.97439791382922), ((2, 12, 2), 3.656
3418527740397), ((2, 12, 3), 3.8534009992790326), ((2, 15, 1), 4.30551413854436),
((2, 15, 2), 3.985619109110499), ((2, 15, 3), 3.9944666477797024), ((4, 6, 1), 5.4
2729642306725), ((4, 6, 2), 6.232113619362876), ((4, 6, 3), 7.066861487684996),
((4, 9, 1), 5.306682911986174), ((4, 9, 2), 6.752619086536817), ((4, 9, 3), 6.4580
77631942944), ((4, 12, 1), 7.248698512009593), ((4, 12, 2), 6.7476115127200895),
((4, 12, 3), 6.739422913796458), ((4, 15, 1), 7.272798523142825), ((4, 15, 2), 7.1
62179636183717), ((4, 15, 3), 6.2971861770383475), ((6, 6, 1), 9.368952844233213),
((6, 6, 2), 8.360150049232413), ((6, 6, 3), 7.4940484344160465), ((6, 9, 1), 5.411
961063829803), ((6, 9, 2), 8.847109447095736), ((6, 9, 3), 7.668849546015992),
((6, 12, 1), 7.1065535607461126), ((6, 12, 2), 8.667574237998922), ((6, 12, 3), 8.
235072919240308), ((6, 15, 1), 9.906398587832992), ((6, 15, 2), 8.63727416346914
7), ((6, 15, 3), 7.52836093339485), ((8, 6, 1), 11.57396332141865), ((8, 6, 2), 9.
028740429700516), ((8, 6, 3), 10.450437340358386), ((8, 9, 1), 8.099162167257107),
((8, 9, 2), 10.609832225356307), ((8, 9, 3), 9.126775703387185), ((8, 12, 1), 8.41
371505157108), ((8, 12, 2), 10.558306702762076), ((8, 12, 3), 8.572888915031083),
((8, 15, 1), 13.33522046268547), ((8, 15, 2), 10.030836188306212), ((8, 15, 3), 8.
75462044188252)]
TFRecord Results by Parameter:
('num_batches', 15): [65.08871846281336, 64.8898726121876, 65.18103494718346, 152.
57026247409755, 152.87616160665064, 151.5733783263691, 122.91517006330913, 184.818
86654994537, 193.8503110524396, 130.64308921205821, 134.13352791341114, 139.469659
4840466]
('num_batches', 9): [33.64938308141507, 40.86307434838805, 63.614067352841545, 87.
55934025597178, 68.17118312115282, 62.38259244294281, 38.32496461802014, 62.866851
045262315, 39.14091304205067, 269.94667748873246, 241.3734191809574, 250.382471755
71803]
('repetitions', 3): [22.90338558260959, 63.614067352841545, 66.44022434666941, 65.
18103494718346, 122.1715580753047, 62.38259244294281, 80.34761564457708, 151.57337
83263691, 82.08282670104234, 39.14091304205067, 79.55143290889555, 193.85031105243
96, 245.77880143826104, 250.38247175571803, 203.42351550171088, 139.4696594840466]
('repetitions', 1): [20.069584509802592, 33.64938308141507, 68.36022062528608, 65.
08871846281336, 117.45323610057756, 87.55934025597178, 73.68295553694713, 152.5702
6247409755, 83.90906175969307, 38.32496461802014, 62.92390682199252, 122.915170063
30913, 232.56924692403396, 269.94667748873246, 255.36858594846498, 130.64308921205
821]
('batch_size', 2): [20.069584509802592, 18.98942373155048, 22.90338558260959, 33.6
4938308141507, 40.86307434838805, 63.614067352841545, 68.36022062528608, 63.630804
240229075, 66.44022434666941, 65.08871846281336, 64.8898726121876, 65.181034947183
46]
```

```
('batch_size', 8): [232.56924692403396, 241.51129393008614, 245.77880143826104, 26
9.94667748873246, 241.3734191809574, 250.38247175571803, 255.36858594846498, 254.8
791155229744, 203.42351550171088, 130.64308921205821, 134.13352791341114, 139.4696
594840466]
('batch_size', 6): [83.90906175969307, 105.36746143840473, 82.08282670104234, 38.3
2496461802014, 62.866851045262315, 39.14091304205067, 62.92390682199252, 59.061943
28709691, 79.55143290889555, 122.91517006330913, 184.81886654994537, 193.850311052
4396]
('num_batches', 6): [20.069584509802592, 18.98942373155048, 22.90338558260959, 11
7.45323610057756, 133.34370446513253, 122.1715580753047, 83.90906175969307, 105.36
746143840473, 82.08282670104234, 232.56924692403396, 241.51129393008614, 245.77880
143826104]
('num_batches', 12): [68.36022062528608, 63.630804240229075, 66.44022434666941, 7
3.68295553694713, 72.57608063424486, 80.34761564457708, 62.92390682199252, 59.0619
4328709691, 79.55143290889555, 255.36858594846498, 254.8791155229744, 203.42351550
171088]
('repetitions', 2): [18.98942373155048, 40.86307434838805, 63.630804240229075, 64.
8898726121876, 133.34370446513253, 68.17118312115282, 72.57608063424486, 152.87616
160665064, 105.36746143840473, 62.866851045262315, 59.06194328709691, 184.81886654
994537, 241.51129393008614, 241.3734191809574, 254.8791155229744, 134.133527913411
14]
('batch_size', 4): [117.45323610057756, 133.34370446513253, 122.1715580753047, 87.
55934025597178, 68.17118312115282, 62.38259244294281, 73.68295553694713, 72.576080
63424486, 80.34761564457708, 152.57026247409755, 152.87616160665064, 151.573378326
3691]
JPEG Results by Parameter:
('num_batches', 15): [2.4992384350324057, 3.068472305642756, 2.907421734765757, 4.
778447097126253, 6.659708053145403, 6.43841134393622, 5.951504383362745, 5.9987254
94326231, 7.792096303929812, 10.70269606822657, 12.39756636819101, 11.431981972296
427]
('num_batches', 9): [3.2316577803803126, 4.562317512519448, 3.0226591723561955, 6.
623362636772632, 6.0166133421473695, 7.2142402978596785, 9.395855936170756, 6.1398
56546190651, 8.107362203875093, 7.815482596071414, 9.786852647291312, 8.6089061622
22862]
('repetitions', 3): [4.615981331490466, 3.0226591723561955, 2.969243584449295, 2.9
07421734765757, 6.798831093212327, 7.2142402978596785, 6.2728557412245705, 6.43841
134393622, 8.466421943367946, 8.107362203875093, 7.037076037702095, 7.792096303929
812, 10.485053897151614, 8.608906162222862, 12.632600623889402, 11.43198197229642
7]
('repetitions', 1): [4.549871531130345, 3.2316577803803126, 4.317175491426102, 2.4
992384350324057, 2.7773553787688203, 6.623362636772632, 5.872064698164115, 4.77844
7097126253, 9.289207505020492, 9.395855936170756, 8.981629306051028, 5.95150438336
2745, 10.618350569304171, 7.815482596071414, 9.22898410305447, 10.70269606822657]
('batch_size', 2): [4.549871531130345, 3.440318259689317, 4.615981331490466, 3.231
6577803803126, 4.562317512519448, 3.0226591723561955, 4.317175491426102, 3.4958034
975469663, 2.969243584449295, 2.4992384350324057, 3.068472305642756, 2.90742173476
5757]
('num_batches', 6): [4.549871531130345, 3.440318259689317, 4.615981331490466, 2.77
73553787688203, 6.103311875027473, 6.798831093212327, 9.289207505020492, 7.6235465
613094275, 8.466421943367946, 10.618350569304171, 8.417259474453177, 10.4850538971
51614]
('repetitions', 2): [3.440318259689317, 4.562317512519448, 3.4958034975469663, 3.0
68472305642756, 6.103311875027473, 6.0166133421473695, 7.48469801698988, 6.6597080
53145403, 7.6235465613094275, 6.139856546190651, 5.453977624253401, 5.998725494326
231, 8.417259474453177, 9.786852647291312, 9.517829348832093, 12.39756636819101]
('batch_size', 4): [2.7773553787688203, 6.103311875027473, 6.798831093212327, 6.62
3362636772632, 6.0166133421473695, 7.2142402978596785, 5.872064698164115, 7.484698
01698988, 6.2728557412245705, 4.778447097126253, 6.659708053145403, 6.438411343936
22]
('num_batches', 12): [4.317175491426102, 3.4958034975469663, 2.969243584449295, 5.
872064698164115, 7.48469801698988, 6.2728557412245705, 8.981629306051028, 5.453977
624253401, 7.037076037702095, 9.22898410305447, 9.517829348832093, 12.632600623889
402]
('batch_size', 6): [9.289207505020492, 7.6235465613094275, 8.466421943367946, 9.39
```

```
5855936170756, 6.139856546190651, 8.107362203875093, 8.981629306051028, 5.45397762
4253401, 7.037076037702095, 5.951504383362745, 5.998725494326231, 7.79209630392981
2]
('batch_size', 8): [10.618350569304171, 8.417259474453177, 10.485053897151614, 7.8
15482596071414, 9.786852647291312, 8.608906162222862, 9.22898410305447, 9.51782934
8832093, 12.632600623889402, 10.70269606822657, 12.39756636819101, 11.431981972296
427]
Average TFRecord Results by Parameter:
('num_batches', 15): 148.81490219620105
('num_batches', 9): 138.70663477194705
('repetitions', 3): 142.12944798836
('repetitions', 1): 140.5339277207717
('num_batches', 12): 147.90957764011807
('batch_size', 4): 137.9651861707565
('repetitions', 2): 140.05573710152188
('batch_size', 6): 161.8095815577036
('batch_size', 2): 51.058168455842974
('batch_size', 8): 212.79254756323505
('num_batches', 6): 128.19436913927197
Average JPEG Results by Parameter:
('num_batches', 9): 7.051635271950214
('repetitions', 3): 7.191135704647024
('num_batches', 15): 7.600200413722561
('repetitions', 1): 7.204904746982785
('num_batches', 12): 6.965847749183683
('batch_size', 4): 6.257313062185006
('repetitions', 2): 7.329472260439273
('batch_size', 2): 4.047188029664985
('num_batches', 6): 7.349666847902323
('batch_size', 8): 10.259220266357834
('batch_size', 6): 8.403628924550953
Saving file as gs://earnest-crow-421721-storage/tfrecord_results_240504-2257.pickl
e
Saving file as gs://earnest-crow-421721-storage/jpeg_results_240504-2257.pickle
```

ii) Cloud

If you have a cluster running, you can run the speed test job in the cloud.

While you run this job, switch to the Dataproc web page and take **screenshots of the CPU and network load** over time. They are displayed with some delay, so you may need to wait a little. These images will be useful in the next task. Again, don't use the SCREENSHOT function that Google provides, but just take a picture of the graphs you see for the VMs.

In [11]:
```
### CODING TASK ###
#Using the max SSD size of 100 with 8 vCPUs for the master machine
!gcloud dataproc clusters create 'jobs2b' \
    --bucket 'earnest-crow-421721-storage' \
    --region us-central1 \
    --zone us-central1-c \
    --image-version 1.5-ubuntu18 \
    --single-node \
    --master-machine-type n1-standard-8 \
    --master-boot-disk-type pd-ssd \
    --master-boot-disk-size 100 \
    --initialization-actions gs://goog-dataproc-initialization-actions-us-central1/
    --metadata PIP_PACKAGES='tensorflow==2.4.0 numpy protobuf==3.20.0 gcsfs'
```

```
Waiting on operation [projects/earnest-crow-421721/regions/us-central1/operations/
7f854b42-73d7-3bf4-8207-0c9c332f7412].

WARNING: Consider using Auto Zone rather than selecting a zone manually. See http
s://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Don't create production clusters that reference initialization actions lo
cated in the gs://goog-dataproc-initialization-actions-REGION public buckets. Thes
e scripts are provided as reference implementations, and they are synchronized wit
h ongoing GitHub repository changes—a new version of a initialization action in pu
blic buckets may break your cluster creation. Instead, copy the following initiali
zation actions from public buckets into your bucket : gs://goog-dataproc-initializ
ation-actions-us-central1/python/pip-install.sh
WARNING: Failed to validate permissions required for default service account: '972
034511549-compute@developer.gserviceaccount.com'. Cluster creation could still be
successful if required permissions have been granted to the respective service acc
ounts as mentioned in the document https://cloud.google.com/dataproc/docs/concept
s/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be
due to Cloud Resource Manager API hasn't been enabled in your project '97203451154
9' before or it is disabled. Enable it by visiting 'https://console.developers.goo
gle.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=97203451154
9'.
WARNING: The firewall rules for specified network or subnetwork would allow ingres
s traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'earnest-crow-421721-storage' is not
using uniform bucket level access IAM configuration. It is recommended to update b
ucket to enable the same. See https://cloud.google.com/storage/docs/uniform-bucket
-level-access.
Created [https://dataproc.googleapis.com/v1/projects/earnest-crow-421721/regions/u
s-central1/clusters/jobs2b] Cluster placed in zone [us-central1-c].
```

In [12]:
```
!gcloud dataproc jobs submit pyspark --cluster 'jobs2b'  \
    --region "us-central1" spark_job2.py
```

```
Job [42f4fda3b85d4cc882d985e4d0208ccc] submitted.
Waiting for job output...
2024-05-04 23:01:12.118141: W tensorflow/stream_executor/platform/default/dso_load
er.cc:60] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.s
o.11.0: cannot open shared object file: No such file or directory; LD_LIBRARY_PAT
H: :/usr/lib/hadoop/lib/native
2024-05-04 23:01:12.118187: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] I
gnore above cudart dlerror if you do not have a GPU set up on your machine.
24/05/04 23:01:15 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/05/04 23:01:15 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/05/04 23:01:15 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordina
tor
24/05/04 23:01:15 INFO org.spark_project.jetty.util.log: Logging initialized @6052
ms to org.spark_project.jetty.util.log.Slf4jLog
24/05/04 23:01:15 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHO
T; built: unknown; git: unknown; jvm 1.8.0_382-b05
24/05/04 23:01:15 INFO org.spark_project.jetty.server.Server: Started @6180ms
24/05/04 23:01:15 INFO org.spark_project.jetty.server.AbstractConnector: Started S
erverConnector@600eaa78{HTTP/1.1, (http/1.1)}{0.0.0.0:44643}
24/05/04 23:01:17 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to Resour
ceManager at jobs2b-m/10.128.0.14:8032
24/05/04 23:01:17 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Appli
cation History server at jobs2b-m/10.128.0.14:10200
24/05/04 23:01:17 INFO org.apache.hadoop.conf.Configuration: resource-types.xml no
t found
24/05/04 23:01:17 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable
to find 'resource-types.xml'.
24/05/04 23:01:17 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/05/04 23:01:17 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = vcores, units = , type = COUNTABLE
24/05/04 23:01:20 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Subm
itted application application_1714863560718_0001
TFRecord Results Array:
[((2, 6, 1), 122.1944461978451), ((2, 6, 2), 134.08721644266248), ((2, 6, 3), 125.
91201207461579), ((2, 9, 1), 134.89592475592846), ((2, 9, 2), 132.75202761334634),
((2, 9, 3), 130.318370672379), ((2, 12, 1), 121.16166958347075), ((2, 12, 2), 132.
21068171116997), ((2, 12, 3), 127.85828424525734), ((2, 15, 1), 135.4067072147109
3), ((2, 15, 2), 133.05036186504435), ((2, 15, 3), 135.1205010972868), ((4, 6, 1),
276.6371955743895), ((4, 6, 2), 278.0396882720866), ((4, 6, 3), 268.135382713922
8), ((4, 9, 1), 270.29121542036165), ((4, 9, 2), 260.43624359020794), ((4, 9, 3),
270.21043834611356), ((4, 12, 1), 271.3300044878767), ((4, 12, 2), 268.36836390902
27), ((4, 12, 3), 270.87683885113876), ((4, 15, 1), 271.1767273009397), ((4, 15,
2), 258.7402185211422), ((4, 15, 3), 273.4789425030211), ((6, 6, 1), 370.069320471
15566), ((6, 6, 2), 399.8494397286229), ((6, 6, 3), 383.8077135224306), ((6, 9,
1), 403.29558906275594), ((6, 9, 2), 393.61131970216616), ((6, 9, 3), 410.42908218
460684), ((6, 12, 1), 364.53717679437966), ((6, 12, 2), 401.4096170065138), ((6, 1
2, 3), 396.2028259041358), ((6, 15, 1), 416.9467869722685), ((6, 15, 2), 408.53300
89490664), ((6, 15, 3), 406.36530396066036), ((8, 6, 1), 562.344361796918), ((8,
6, 2), 577.6159404385342), ((8, 6, 3), 543.8309817558405), ((8, 9, 1), 560.7191705
534595), ((8, 9, 2), 523.7805875558054), ((8, 9, 3), 543.7774495263396), ((8, 12,
1), 485.2266248907902), ((8, 12, 2), 557.826211459695), ((8, 12, 3), 560.452771921
4894), ((8, 15, 1), 563.0549175917298), ((8, 15, 2), 523.3971344607199), ((8, 15,
3), 553.719336306649)]
JPEG Results Array:
[((2, 6, 1), 4.822273286802906), ((2, 6, 2), 5.104173810847825), ((2, 6, 3), 4.928
528525156672), ((2, 9, 1), 4.839244652583801), ((2, 9, 2), 5.0279826259068985),
((2, 9, 3), 4.913130791465496), ((2, 12, 1), 5.070603346277069), ((2, 12, 2), 5.08
9586942931955), ((2, 12, 3), 5.0012764115362245), ((2, 15, 1), 4.883076225263706),
((2, 15, 2), 5.008744166939669), ((2, 15, 3), 5.09027956926783), ((4, 6, 1), 6.821
168320945722), ((4, 6, 2), 6.391708753350918), ((4, 6, 3), 6.55165185039813), ((4,
9, 1), 6.754235080748242), ((4, 9, 2), 6.705603576489029), ((4, 9, 3), 6.561491921
74355), ((4, 12, 1), 6.553757017095201), ((4, 12, 2), 6.5616612101204606), ((4, 1
2, 3), 6.635768104176346), ((4, 15, 1), 6.693475331966488), ((4, 15, 2), 6.4465764
```

```
22845909), ((4, 15, 3), 6.535652032935457), ((6, 6, 1), 7.28524789706208), ((6, 6,
2), 7.3215871424357735), ((6, 6, 3), 7.304707382007127), ((6, 9, 1), 7.42268882912
1589), ((6, 9, 2), 7.228615821646427), ((6, 9, 3), 7.386922760738648), ((6, 12,
1), 7.340408654719095), ((6, 12, 2), 7.363879476225803), ((6, 12, 3), 7.3338540004
36015), ((6, 15, 1), 7.283466294327261), ((6, 15, 2), 7.3391490697871165), ((6, 1
5, 3), 7.28095614210202), ((8, 6, 1), 7.8083227355073515), ((8, 6, 2), 7.617019622
358096), ((8, 6, 3), 7.70427148424941), ((8, 9, 1), 7.8019235072256885), ((8, 9,
2), 7.801452381989104), ((8, 9, 3), 7.910846904426487), ((8, 12, 1), 7.78319915080
3921), ((8, 12, 2), 7.874082016212131), ((8, 12, 3), 7.765019754537449), ((8, 15,
1), 7.7136171336554895), ((8, 15, 2), 7.905959204764142), ((8, 15, 3), 7.791537462
997472)]
TFRecord Results by Parameter:
('repetitions', 1): [132.09920895295187, 142.98194958155074, 133.2150626286153, 13
6.38534576197702, 280.8058915420665, 279.11321100001663, 280.01189441343655, 230.9
7998759091126, 411.9770485630796, 371.2740881350598, 396.9122452842092, 410.004398
8699792, 547.8632400483297, 556.6798062246997, 555.7923052395761, 568.610604302681
2]
('repetitions', 3): [142.2639225455354, 137.98439536137587, 143.11258883229485, 14
0.10726451726885, 276.65949770877955, 277.21279333981613, 267.21872009450306, 284.
9019263034039, 411.09842461445754, 410.6907984048102, 387.9540366127824, 400.98096
86112564, 546.149630115916, 554.6081855808062, 518.1040325970406, 518.382578466353
3]
('num_batches', 9): [142.98194958155074, 126.22641764955196, 137.98439536137587, 2
79.11321100001663, 277.9850453765949, 277.21279333981613, 371.2740881350598, 416.8
954440913446, 410.6907984048102, 556.6798062246997, 557.6409299989475, 554.6081855
808062]
('num_batches', 15): [136.38534576197702, 143.5670435855615, 140.10726451726885, 2
30.97998759091126, 258.53209124194075, 284.9019263034039, 410.0043988699792, 406.5
351064242236, 400.9809686112564, 568.6106043026812, 525.1284790709296, 518.3825784
663533]
('batch_size', 2): [132.09920895295187, 136.81892948741472, 142.2639225455354, 14
2.98194958155074, 126.22641764955196, 137.98439536137587, 133.2150626286153, 135.7
9585691853595, 143.11258883229485, 136.38534576197702, 143.5670435855615, 140.1072
6451726885]
('num_batches', 6): [132.09920895295187, 136.81892948741472, 142.2639225455354, 28
0.8058915420665, 276.52814431956006, 276.65949770877955, 411.9770485630796, 415.57
936455512254, 411.09842461445754, 547.8632400483297, 530.7691893986729, 546.149630
115916]
('repetitions', 2): [136.81892948741472, 126.22641764955196, 135.79585691853595, 1
43.5670435855615, 276.52814431956006, 277.9850453765949, 262.1602143361731, 258.53
209124194075, 415.57936455512254, 416.8954440913446, 401.7893369256099, 406.535106
4242236, 530.7691893986729, 557.6409299989475, 549.9434340564785, 525.128479070929
6]
('num_batches', 12): [133.2150626286153, 135.79585691853595, 143.11258883229485, 2
80.01189441343655, 262.1602143361731, 267.21872009450306, 396.9122452842092, 401.7
893369256099, 387.9540366127824, 555.7923052395761, 549.9434340564785, 518.1040325
970406]
('batch_size', 4): [280.8058915420665, 276.52814431956006, 276.65949770877955, 27
9.11321100001663, 277.9850453765949, 277.21279333981613, 280.01189441343655, 262.1
602143361731, 267.21872009450306, 230.97998759091126, 258.53209124194075, 284.9019
263034039]
('batch_size', 6): [411.9770485630796, 415.57936455512254, 411.09842461445754, 37
1.2740881350598, 416.8954440913446, 410.6907984048102, 396.9122452842092, 401.7893
369256099, 387.9540366127824, 410.0043988699792, 406.5351064242236, 400.9809686112
564]
('batch_size', 8): [547.8632400483297, 530.7691893986729, 546.149630115916, 556.67
98062246997, 557.6409299989475, 554.6081855808062, 555.7923052395761, 549.94343405
64785, 518.1040325970406, 568.6106043026812, 525.1284790709296, 518.3825784663533]
JPEG Results by Parameter:
('repetitions', 1): [4.890089581020583, 5.165940844780723, 4.972648660333474, 4.92
9227889400562, 6.535107687465625, 6.478244853419868, 6.568513433429519, 6.77102010
0352031, 7.170622410446018, 7.381458316046889, 7.194577184602879, 7.2852808679821
8, 7.88801763861595, 7.827904547275737, 8.036807308385038, 7.6596538108109495]
('repetitions', 3): [5.021279431585437, 5.011347619729307, 5.052079786721485, 5.02
```

```
20344984213785, 6.680118959648563, 6.5133931782273295, 6.579327268493744, 6.624223
305745517, 7.329444973028924, 7.416954459411718, 7.551541886858928, 7.421977705362
628, 7.7764844061202565, 7.897718643174275, 7.77407743613668, 7.810924277547303]
('num_batches', 9): [5.165940844780723, 5.078830084942746, 5.011347619729307, 6.47
8244853419868, 6.636397556642725, 6.5133931782273295, 7.381458316046889, 7.4567111
81650267, 7.416954459411718, 7.827904547275737, 7.929770758406088, 7.8977186431742
75]
('num_batches', 15): [4.929227889400562, 5.013907084834144, 5.0220344984213785, 6.
771020100352031, 6.599814254027935, 6.624223305745517, 7.28528086798218, 7.2415324
44365016, 7.421977705362628, 7.6596538108109495, 7.734216936252593, 7.810924277547
303]
('batch_size', 2): [4.890089581020583, 5.122798142967597, 5.021279431585437, 5.165
940844780723, 5.078830084942746, 5.011347619729307, 4.972648660333474, 5.143333853
9004305, 5.052079786721485, 4.929227889400562, 5.013907084834144, 5.02203449842137
85]
('num_batches', 6): [4.890089581020583, 5.122798142967597, 5.021279431585437, 6.53
5107687465625, 6.701699486953301, 6.680118959648563, 7.170622410446018, 7.24101199
0726111, 7.329444973028924, 7.88801763861595, 7.995665833335501, 7.776484406120256
5]
('repetitions', 2): [5.122798142967597, 5.078830084942746, 5.1433338539004305, 5.0
13907084834144, 6.701699486953301, 6.636397556642725, 6.632107354925013, 6.5998142
54027935, 7.241011990726111, 7.456711181650267, 7.369310192631678, 7.2415324443650
16, 7.995665833335501, 7.929770758406088, 7.6944684413628766, 7.734216936252593]
('num_batches', 12): [4.972648660333474, 5.1433338539004305, 5.052079786721485, 6.
568513433429519, 6.632107354925013, 6.579327268493744, 7.194577184602879, 7.369310
192631678, 7.551541886858928, 8.036807308385038, 7.6944684413628766, 7.77407743613
668]
('batch_size', 4): [6.535107687465625, 6.701699486953301, 6.680118959648563, 6.478
244853419868, 6.636397556642725, 6.5133931782273295, 6.568513433429519, 6.63210735
4925013, 6.579327268493744, 6.771020100352031, 6.599814254027935, 6.62422330574551
7]
('batch_size', 6): [7.170622410446018, 7.241011990726111, 7.329444973028924, 7.381
458316046889, 7.456711181650267, 7.416954459411718, 7.194577184602879, 7.369310192
631678, 7.551541886858928, 7.28528086798218, 7.241532444365016, 7.421977705362628]
('batch_size', 8): [7.88801763861595, 7.995665833335501, 7.7764844061202565, 7.827
904547275737, 7.929770758406088, 7.897718643174275, 8.036807308385038, 7.694468441
3628766, 7.77407743613668, 7.6596538108109495, 7.734216936252593, 7.81092427754730
3]
Average TFRecord Results by Parameter:
('repetitions', 1): 327.52743544158346
('repetitions', 3): 328.42550774376605
('num_batches', 9): 324.96211176675223
('num_batches', 15): 333.72569907831115
('batch_size', 2): 136.11590964861153
('num_batches', 6): 328.2383452728232
('repetitions', 2): 333.8954787593579
('num_batches', 12): 332.87173980838986
('batch_size', 4): 267.82630668647965
('batch_size', 6): 387.78355003355614
('batch_size', 8): 528.0721295576293
Average JPEG Results by Parameter:
('repetitions', 1): 6.760415860678024
('repetitions', 3): 6.741127620441647
('num_batches', 9): 6.744885500457819
('num_batches', 15): 6.750361098975987
('batch_size', 2): 5.077287022675348
('num_batches', 6): 6.70702677379877
('repetitions', 2): 6.743083825681221
('num_batches', 12): 6.7905630358352775
('batch_size', 4): 6.650846952942704
('batch_size', 6): 7.4027552469087174
('batch_size', 8): 7.86194718654109
Saving file as gs://earnest-crow-421721-storage/tfrecord_results_240504-2325.pickl
e
```

```
Saving file as gs://earnest-crow-421721-storage/jpeg_results_240504-2325.pickle
24/05/04 23:25:52 INFO org.spark_project.jetty.server.AbstractConnector: Stopped S
park@600eaa78{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [42f4fda3b85d4cc882d985e4d0208ccc] finished successfully.
done: true
driverControlFilesUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-meta
info/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/
driverOutputResourceUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-me
tainfo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/
driveroutput
jobUuid: b437c107-3ba6-35a0-a352-966040eb6ff1
placement:
  clusterName: jobs2b
  clusterUuid: 90181cc6-cd42-4d7f-9f64-82c18db1a9c8
pysparkJob:
  mainPythonFileUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-metain
fo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/stag
ing/spark_job2.py
reference:
  jobId: 42f4fda3b85d4cc882d985e4d0208ccc
  projectId: earnest-crow-421721
status:
  state: DONE
  stateStartTime: '2024-05-04T23:25:55.290753Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-05-04T23:01:07.501692Z'
- state: SETUP_DONE
  stateStartTime: '2024-05-04T23:01:07.527399Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-05-04T23:01:07.846450Z'
yarnApplications:
- name: spark_job2.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://jobs2b-m:8088/proxy/application_1714863560718_0001/
```

2c) Improve efficiency (6%)

If you implemented a straightfoward version of 2a), you will **probably have an inefficiency** in your code.

Because we are reading multiple times from an RDD to read the values for the different parameters and their averages, caching existing results is important. Explain **where in the process caching can help**, and **add a call to** `RDD.cache()` to your code, if you haven't yet. Measure the the effect of using caching or not using it.

Make the **suitable change** in the code you have written above and mark them up in comments as `### TASK 2c ###`.

Explain in your report what the **reasons for this change** are and **demonstrate and interpret its effect**

In [13]:
```
%%writefile spark_job2.py

### CODING TASK
# Importing necessary libraries
import os
import time
```

```
import pyspark
import tensorflow as tf
from pyspark.sql import SparkSession

# i)  Creating a Dataset and a List of Parameter Combinations

# Constants for config
TFRECORDS_GCS_PATTERN = 'gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/
JPEG_GCS_PATTERN = 'gs://flowers-public/**/*.jpg'
TARGET_SIZE = [192, 192]
BATCH_SIZES = [2, 4, 6, 8]
BATCH_NUMBERS = [6, 9, 12, 15]
REPETITIONS = [1, 2, 3]

# JPEG images functions
def decode_jpeg_and_label(filepath):
    bits = tf.io.read_file(filepath)
    image = tf.image.decode_jpeg(bits)
    label = tf.strings.split(tf.expand_dims(filepath, axis=-1), sep='/').values[-2]
    return image, label

def resize_and_crop_image(image, label):
    w, h = tf.shape(image)[0], tf.shape(image)[1]
    resize_crit = (w * TARGET_SIZE[0]) / (h * TARGET_SIZE[1])
    image = tf.cond(
        resize_crit < 1,
        lambda: tf.image.resize(image, [w * TARGET_SIZE[1] / w, h * TARGET_SIZE[1]
        lambda: tf.image.resize(image, [w * TARGET_SIZE[0] / h, h * TARGET_SIZE[0]
    )
    nw, nh = tf.shape(image)[0], tf.shape(image)[1]
    image = tf.image.crop_to_bounding_box(image, (nw - TARGET_SIZE[1]) // 2, (nh -
    return image, label

def recompress_image(image, label):
    image = tf.cast(image, tf.uint8)
    image = tf.image.encode_jpeg(image, format='rgb', quality=70)
    image = tf.image.decode_jpeg(image)
    return image, label

# TFrecords functions
def decode_tfrecord(record):
    features = {
        "image": tf.io.FixedLenFeature([], tf.string),
        "class": tf.io.FixedLenFeature([], tf.int64)
    }
    example = tf.io.parse_single_example(record, features)
    image = tf.image.decode_jpeg(example['image'], channels=3)
    image = tf.reshape(image, [*TARGET_SIZE, 3])
    return image, example['class']

def load_tfrecord_dataset():
    dataset = tf.data.TFRecordDataset(tf.io.gfile.glob(TFRECORDS_GCS_PATTERN))
    dataset = dataset.map(decode_tfrecord).cache() ### TASK 2C ###
    return dataset

# ii) Creating dataset, running time configs and creating a spark context

# Recreating dataset4
def load_jpeg_dataset():
    dataset = tf.data.Dataset.list_files(JPEG_GCS_PATTERN)
    dataset4 = dataset4.map(decode_jpeg_and_label)
    dataset4 = dataset4.map(resize_and_crop_image)
    dataset4 = dataset4.map(recompress_image).cache() ### TASK 2C ###
    return dataset4
```

```
#adaption of time configs
def time_configs(dataset_loader, batch_sizes, batch_numbers, repetitions):
    results = []
    for bsize in batch_sizes:
        dataset = dataset_loader().batch(bsize)
        for bnumber in batch_numbers:
            for rep in repetitions:
                total_time = 0
                for _ in range(rep):
                    start_time = time.time()
                    for _ in range(bnumber):
                        for _ in dataset.take(1):
                            pass
                    end_time = time.time()
                    total_time += (end_time - start_time)
                avg_time = total_time / rep
                throughput = (bsize * bnumber) / avg_time if avg_time > 0 else 0
                results.append(((bsize, bnumber, rep), throughput))
    return results

# Creating a Spark context and session
spark_ctx = pyspark.SparkContext.getOrCreate()
spark_s = SparkSession(spark_ctx)

# Generating all parameter combinations
parameter_combinations = [(bsize, bnumber, rep) for bsize in BATCH_SIZES for bnumbe

# Parallelizing the parameter combinations into an RDD for both TFRecords and JPEGs
params_rdd = spark_ctx.parallelize(parameter_combinations)
processed_tfrecord_images_rdd = params_rdd.flatMap(lambda params: time_configs(loac
processed_jpeg_images_rdd = params_rdd.flatMap(lambda params: time_configs(load_jpe

# iii) Transforming the resulting RDDs to the structure and saving them to an array

tfrecord_array = processed_tfrecord_images_rdd.collect()
jpeg_array = processed_jpeg_images_rdd.collect()

# Printing the results for verification
print("TFRecord Results Array:")
print(tfrecord_array)
print("JPEG Results Array:")
print(jpeg_array)

# iv) Creating an RDD with all results for each parameter

#Expanding the tuple results
def expand_results(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), throughput),
        (('num_batches', num_batches), throughput),
        (('repetitions', repetitions), throughput)
    ]

# Transformation
tfrecord_results = processed_tfrecord_images_rdd.flatMap(expand_results)
jpeg_results = processed_jpeg_images_rdd.flatMap(expand_results)

# Collecting results by parameter
tfrecord_results_param = tfrecord_results.groupByKey().mapValues(list).collect()
jpeg_results_param = jpeg_results.groupByKey().mapValues(list).collect()

# Printing results(parameters)
```

```
print("TFRecord Results by Parameter:")
for key, values in tfrecord_results_param:
    print(f"{key}: {values}")

print("JPEG Results by Parameter:")
for key, values in jpeg_results_param:
    print(f"{key}: {values}")


#v) Creating an RDD with the average reading speeds for each parameter value and co

# Summing the throughputs and counting occurrences
def throughputs_counting_avg(a, b):
    throughput_sum_a, count_a = a
    throughput_sum_b, count_b = b
    return (throughput_sum_a + throughput_sum_b, count_a + count_b)

# Mapping results for averaging
def avg_mapping(result):
    (batch_size, num_batches, repetitions), throughput = result
    return [
        (('batch_size', batch_size), (throughput, 1)),
        (('num_batches', num_batches), (throughput, 1)),
        (('repetitions', repetitions), (throughput, 1))
    ]

# Calculating averages from sums and counts
def cal_avg(a):
    throughput_sum, count = a
    return throughput_sum / count if count != 0 else 0

# Mapping the results for averaging and reducing to calculate sums and counts
mapped_tfrecord_results = processed_tfrecord_images_rdd.flatMap(avg_mapping)
mapped_jpeg_results = processed_jpeg_images_rdd.flatMap(avg_mapping)

# Reducing to sum the throughputs and count occurrences for each parameter(by key)
reduced_tfrecord_results = mapped_tfrecord_results.reduceByKey(throughputs_counting
reduced_jpeg_results = mapped_jpeg_results.reduceByKey(throughputs_counting_avg)

# Calculating the throughput for each parameter(average)
avg_tfrecord_results = reduced_tfrecord_results.mapValues(cal_avg).collect()
avg_jpeg_results = reduced_jpeg_results.mapValues(cal_avg).collect()

# Printing the results
print("Average TFRecord Results by Parameter:")
for param, avg_throughput in avg_tfrecord_results:
    print(f"{param}: {avg_throughput}")

print("Average JPEG Results by Parameter:")
for param, avg_throughput in avg_jpeg_results:
    print(f"{param}: {avg_throughput}")

# New saving code that includes the time

import pickle
from datetime import datetime
import gcsfs

def saving_data_gcs(data, b_path, filename_b):
    # Creating a timestamp
    timestamp = datetime.now().strftime("%y%m%d-%H%M")
    file_path = f'{b_path}/{filename_b}_{timestamp}.pickle'

    # Saving data to GCS using gcsfs
```

```
    fs = gcsfs.GCSFileSystem(project='earnest-crow-421721')
    with fs.open(file_path, 'wb') as file:
        pickle.dump(data, file)
    print(f"Saving file as {file_path}")

# Defining the base names for the pickle files
tfrecord_filename = 'tfrecord_base_results'
jpeg_filename = 'jpeg_base_results'

# Calling the function to save data to GCS
saving_data_gcs(avg_tfrecord_results, 'gs://earnest-crow-421721-storage', tfrecord_
saving_data_gcs(avg_jpeg_results, 'gs://earnest-crow-421721-storage', jpeg_filename
```

```
Overwriting spark_job2.py
```

In [14]:

```
### CODING TASK ###
#Using the max SSD size of 100 with 8 vCPUs for the master machine
!gcloud dataproc clusters create 'jobs2c' \
    --bucket 'earnest-crow-421721-storage' \
    --region us-central1 \
    --zone us-central1-c \
    --image-version 1.5-ubuntu18 \
    --single-node \
    --master-machine-type n1-standard-8 \
    --master-boot-disk-type pd-ssd \
    --master-boot-disk-size 100 \
    --initialization-actions gs://goog-dataproc-initialization-actions-us-central1/
    --metadata PIP_PACKAGES='tensorflow==2.4.0 numpy protobuf==3.20.0 gcsfs'
```

```
Waiting on operation [projects/earnest-crow-421721/regions/us-central1/operations/
562e1f7b-39ff-34fc-a366-cceb822d31c0].

WARNING: Consider using Auto Zone rather than selecting a zone manually. See http
s://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Don't create production clusters that reference initialization actions lo
cated in the gs://goog-dataproc-initialization-actions-REGION public buckets. Thes
e scripts are provided as reference implementations, and they are synchronized wit
h ongoing GitHub repository changes—a new version of a initialization action in pu
blic buckets may break your cluster creation. Instead, copy the following initiali
zation actions from public buckets into your bucket : gs://goog-dataproc-initializ
ation-actions-us-central1/python/pip-install.sh
WARNING: Failed to validate permissions required for default service account: '972
034511549-compute@developer.gserviceaccount.com'. Cluster creation could still be
successful if required permissions have been granted to the respective service acc
ounts as mentioned in the document https://cloud.google.com/dataproc/docs/concept
s/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be
due to Cloud Resource Manager API hasn't been enabled in your project '97203451154
9' before or it is disabled. Enable it by visiting 'https://console.developers.goo
gle.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=97203451154
9'.
WARNING: The firewall rules for specified network or subnetwork would allow ingres
s traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'earnest-crow-421721-storage' is not
using uniform bucket level access IAM configuration. It is recommended to update b
ucket to enable the same. See https://cloud.google.com/storage/docs/uniform-bucket
-level-access.
Created [https://dataproc.googleapis.com/v1/projects/earnest-crow-421721/regions/u
s-central1/clusters/jobs2c] Cluster placed in zone [us-central1-c].
```

In [15]:

```
!gcloud dataproc jobs submit pyspark --cluster 'jobs2c'  \
    --region "us-central1" spark_job2.py
```

```
Job [5edef5349c304554b9ec52f9b9a6b1a5] submitted.
Waiting for job output...
2024-05-04 23:34:56.022358: W tensorflow/stream_executor/platform/default/dso_load
er.cc:60] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.s
o.11.0: cannot open shared object file: No such file or directory; LD_LIBRARY_PAT
H: :/usr/lib/hadoop/lib/native
2024-05-04 23:34:56.022403: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] I
gnore above cudart dlerror if you do not have a GPU set up on your machine.
24/05/04 23:34:59 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/05/04 23:34:59 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/05/04 23:34:59 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordina
tor
24/05/04 23:34:59 INFO org.spark_project.jetty.util.log: Logging initialized @5651
ms to org.spark_project.jetty.util.log.Slf4jLog
24/05/04 23:34:59 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHO
T; built: unknown; git: unknown; jvm 1.8.0_382-b05
24/05/04 23:34:59 INFO org.spark_project.jetty.server.Server: Started @5785ms
24/05/04 23:34:59 INFO org.spark_project.jetty.server.AbstractConnector: Started S
erverConnector@4a63dd5d{HTTP/1.1, (http/1.1)}{0.0.0.0:43527}
24/05/04 23:35:00 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to Resour
ceManager at jobs2c-m/10.128.0.15:8032
24/05/04 23:35:00 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Appli
cation History server at jobs2c-m/10.128.0.15:10200
24/05/04 23:35:01 INFO org.apache.hadoop.conf.Configuration: resource-types.xml no
t found
24/05/04 23:35:01 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable
to find 'resource-types.xml'.
24/05/04 23:35:01 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = memory-mb, units = Mi, type = COUNTABLE
24/05/04 23:35:01 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding
resource type - name = vcores, units = , type = COUNTABLE
24/05/04 23:35:03 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Subm
itted application application_1714865587157_0001
TFRecords Processing Results:
+----------+-----------+-----------+------------+------------------+-------------
-----+
|batch_size|num_batches|repetitions|total_images|    processing_time|        throu
ghput|
+----------+-----------+-----------+------------+------------------+-------------
-----+
|         2|          6|          1|          12|0.08843111991882324|135.698835557
16123|
|         2|          6|          2|          12|0.08448481559753418|142.037358016
67818|
|         2|          6|          2|          12|0.08481979370117188| 141.47641106
36384|
|         2|          6|          3|          12|0.08427882194519043|142.384524641
36287|
|         2|          6|          3|          12|0.08156085014343262| 147.12941293
39854|
|         2|          6|          3|          12| 0.0823049545288086| 145.79924220
47901|
|         2|          9|          1|          18|0.12089157104492188|148.893755324
85565|
|         2|          9|          2|          18|0.12245535850524902|146.992342513
36117|
|         2|          9|          2|          18|0.12192940711975098|147.626404697
60993|
|         2|          9|          3|          18|0.12286090850830078|146.507137368
13915|
|         2|          9|          3|          18| 0.1218113899230957| 147.76943282
03964|
|         2|          9|          3|          18| 0.1236429214477539|145.580513540
40524|
```

```
|         2|         12|          1|          24|0.16709065437316895|143.634604161
64284|
|         2|         12|          2|          24|0.16608786582946777|144.501826669
51852|
|         2|         12|          2|          24| 0.1650528907775879| 145.40793491
66958|
|         2|         12|          3|          24|0.16941523551940918|141.663764338
65904|
|         2|         12|          3|          24|0.16260433197021484|147.597543738
23336|
|         2|         12|          3|          24|0.16436052322387695| 146.02046482
48131|
|         2|         15|          1|          30| 0.2097020149230957| 143.06014184
46167|
|         2|         15|          2|          30|  0.210982084274292|142.192168132
14257|
+----------+-----------+-----------+------------+-------------------+-------------
-----+
only showing top 20 rows

JPEG Processing Results:
+----------+-----------+-----------+------------+------------------+-------------
----+
|batch_size|num_batches|repetitions|total_images|   processing_time|        throug
hput|
+----------+-----------+-----------+------------+------------------+-------------
----+
|         2|          6|          1|          12|2.2435944080352783| 5.34856030886
1009|
|         2|          6|          2|          12|2.2882943153381348| 5.24408067597
1436|
|         2|          6|          2|          12|2.2194628715515137| 5.40671355840
7675|
|         2|          6|          3|          12|2.1602602005004883| 5.55488639619
4238|
|         2|          6|          3|          12| 2.384901523590088|5.031654297379
9265|
|         2|          6|          3|          12|2.2667689323425293| 5.29387880201
7519|
|         2|          9|          1|          18|3.3041653633117676| 5.44766923588
7935|
|         2|          9|          2|          18| 3.490770101547241| 5.15645530251
9555|
|         2|          9|          2|          18| 3.287404775619507| 5.47544377056
7598|
|         2|          9|          3|          18|3.2856204509735107| 5.47841732439
5063|
|         2|          9|          3|          18|3.5041677951812744| 5.13674031955
6769|
|         2|          9|          3|          18| 3.303894281387329| 5.44811621285
9471|
|         2|         12|          1|          24| 4.456442356109619|5.385461783679
7285|
|         2|         12|          2|          24| 4.511239290237427|  5.3200458800
6168|
|         2|         12|          2|          24|4.6061131954193115| 5.21046682566
7143|
|         2|         12|          3|          24|  4.52794623374939| 5.30041629494
4976|
|         2|         12|          3|          24| 4.632786750793457| 5.18046724164
2305|
|         2|         12|          3|          24| 4.484798192977905| 5.35141136062
2228|
|         2|         15|          1|          30| 5.752248764038086|  5.2153516356
```

```
1667|
|         2|         15|          2|          30| 5.710488557815552| 5.25349095725
6375|
+----------+-----------+-----------+------------+------------------+-------------
----+
only showing top 20 rows

Sample TFRecord Performance Data: [(2, 6, 1, 12, 0.08843111991882324, 135.69883555
716123), (2, 6, 2, 12, 0.08448481559753418, 142.03735801667818), (2, 6, 2, 12, 0.0
8481979370117188, 141.4764110636384), (2, 6, 3, 12, 0.08427882194519043, 142.38452
464136287), (2, 6, 3, 12, 0.08156085014343262, 147.1294129339854)]
Sample Image Performance Data: [(2, 6, 1, 12, 2.2435944080352783, 5.34856030886100
9), (2, 6, 2, 12, 2.2882943153381348, 5.244080675971436), (2, 6, 2, 12, 2.21946287
15515137, 5.406713558407675), (2, 6, 3, 12, 2.1602602005004883, 5.55488639619423
8), (2, 6, 3, 12, 2.384901523590088, 5.0316542973799265)]
Collected TFRecord Performance Data: [((2, 6, 1), 135.69883555716123), ((2, 6, 2),
142.03735801667818), ((2, 6, 2), 141.4764110636384), ((2, 6, 3), 142.3845246413628
7), ((2, 6, 3), 147.1294129339854), ((2, 6, 3), 145.7992422047901), ((2, 9, 1), 14
8.89375532485565), ((2, 9, 2), 146.99234251336117), ((2, 9, 2), 147.6264046976099
3), ((2, 9, 3), 146.50713736813915), ((2, 9, 3), 147.7694328203964), ((2, 9, 3), 1
45.58051354040524), ((2, 12, 1), 143.63460416164284), ((2, 12, 2), 144.50182666951
852), ((2, 12, 2), 145.4079349166958), ((2, 12, 3), 141.66376433865904), ((2, 12,
3), 147.59754373823336), ((2, 12, 3), 146.0204648248131), ((2, 15, 1), 143.0601418
446167), ((2, 15, 2), 142.19216813214257), ((2, 15, 2), 141.97942113333838), ((2,
15, 3), 141.20238888776672), ((2, 15, 3), 143.48886907219082), ((2, 15, 3), 139.47
338259956794), ((4, 6, 1), 273.38778458856024), ((4, 6, 2), 271.2433673386901),
((4, 6, 2), 285.72769617235065), ((4, 6, 3), 286.11263930693997), ((4, 6, 3), 282.
6874252304167), ((4, 6, 3), 285.45302344574134), ((4, 9, 1), 283.33988977521574),
((4, 9, 2), 289.9633673747643), ((4, 9, 2), 285.508077691051), ((4, 9, 3), 286.045
15876037897), ((4, 9, 3), 286.85757709318057), ((4, 9, 3), 288.81584217502825),
((4, 12, 1), 289.5785512916403), ((4, 12, 2), 258.7898410445862), ((4, 12, 2), 27
5.7970303336662), ((4, 12, 3), 280.48236036595813), ((4, 12, 3), 280.6806156591568
4), ((4, 12, 3), 280.58282162418294), ((4, 15, 1), 278.07847600525974), ((4, 15,
2), 277.33594294990263), ((4, 15, 2), 279.39861043572097), ((4, 15, 3), 282.651230
023283), ((4, 15, 3), 282.36612031667954), ((4, 15, 3), 283.7923292311752), ((6,
6, 1), 411.58625201370546), ((6, 6, 2), 446.10763667304826), ((6, 6, 2), 447.58604
907604473), ((6, 6, 3), 443.4258998763652), ((6, 6, 3), 446.96466492808366), ((6,
6, 3), 454.9164070643954), ((6, 9, 1), 424.6807119124585), ((6, 9, 2), 447.2722705
6040255), ((6, 9, 2), 452.9033094043072), ((6, 9, 3), 449.32017663932965), ((6, 9,
3), 437.84756594997737), ((6, 9, 3), 286.37590973230226), ((6, 12, 1), 426.7618123
0047865), ((6, 12, 2), 441.6283587546248), ((6, 12, 2), 450.5173484823621), ((6, 1
2, 3), 447.83893867515314), ((6, 12, 3), 431.57627508621096), ((6, 12, 3), 444.409
9421659088), ((6, 15, 1), 438.5705953838754), ((6, 15, 2), 450.71933632788154),
((6, 15, 2), 446.9633418584825), ((6, 15, 3), 445.52058019934026), ((6, 15, 3), 43
5.8442961946921), ((6, 15, 3), 438.6582534425658), ((8, 6, 1), 600.4049648393466),
((8, 6, 2), 594.5818158191623), ((8, 6, 2), 593.6385917320281), ((8, 6, 3), 568.14
78287372022), ((8, 6, 3), 599.8200241326401), ((8, 6, 3), 598.2639621060389), ((8,
9, 1), 602.5360995055846), ((8, 9, 2), 594.4296689388287), ((8, 9, 2), 600.9868614
300782), ((8, 9, 3), 573.1761436860256), ((8, 9, 3), 571.6733704870366), ((8, 9,
3), 587.7805442017984), ((8, 12, 1), 579.1970908756142), ((8, 12, 2), 585.59729142
48815), ((8, 12, 2), 593.3534097841753), ((8, 12, 3), 587.1129029671094), ((8, 12,
3), 571.7946014723258), ((8, 12, 3), 594.0879151684871), ((8, 15, 1), 589.12004756
81597), ((8, 15, 2), 576.654869669726), ((8, 15, 2), 588.932549363756), ((8, 15,
3), 593.604985505332), ((8, 15, 3), 570.0486558482155), ((8, 15, 3), 593.133499573
401)]
Collected JPEG Performance Data: [((2, 6, 1), 5.348560308861009), ((2, 6, 2), 5.24
4080675971436), ((2, 6, 2), 5.406713558407675), ((2, 6, 3), 5.554886396194238),
((2, 6, 3), 5.0316542973799265), ((2, 6, 3), 5.293878802017519), ((2, 9, 1), 5.447
669235887935), ((2, 9, 2), 5.156455302519555), ((2, 9, 2), 5.475443770567598),
((2, 9, 3), 5.478417324395063), ((2, 9, 3), 5.136740319556769), ((2, 9, 3), 5.4481
16212859471), ((2, 12, 1), 5.3854617836797285), ((2, 12, 2), 5.32004588006168),
((2, 12, 2), 5.210466825667143), ((2, 12, 3), 5.300416294944976), ((2, 12, 3), 5.1
80467241642305), ((2, 12, 3), 5.351411360622228), ((2, 15, 1), 5.21535163561667),
((2, 15, 2), 5.253490957256375), ((2, 15, 2), 5.492658891973272), ((2, 15, 3), 5.3
```

```
751425492818115), ((2, 15, 3), 5.10463996884375), ((2, 15, 3), 5.216591205975454),
((4, 6, 1), 7.059362413036321), ((4, 6, 2), 6.925264423819461), ((4, 6, 2), 6.6832
69348845411), ((4, 6, 3), 7.1098569719626115), ((4, 6, 3), 6.8411821877107), ((4,
6, 3), 6.833989327757336), ((4, 9, 1), 6.944304536806277), ((4, 9, 2), 7.022171784
908419), ((4, 9, 2), 6.82559954954959), ((4, 9, 3), 6.99895596183615), ((4, 9, 3),
6.7520951871600365), ((4, 9, 3), 6.68555468473167), ((4, 12, 1), 6.95080957949842
1), ((4, 12, 2), 6.928848813869275), ((4, 12, 2), 6.908317596654002), ((4, 12, 3),
6.767285635212424), ((4, 12, 3), 6.861966312297862), ((4, 12, 3), 6.80526361048719
95), ((4, 15, 1), 6.955257614279101), ((4, 15, 2), 7.058041735065355), ((4, 15,
2), 6.678793638907486), ((4, 15, 3), 6.744161734308452), ((4, 15, 3), 6.5897890967
31978), ((4, 15, 3), 6.864777503929148), ((6, 6, 1), 7.754005141950547), ((6, 6,
2), 7.674072157091789), ((6, 6, 2), 7.7477613864092945), ((6, 6, 3), 7.83607543746
6154), ((6, 6, 3), 7.6804605744942895), ((6, 6, 3), 7.803241347608059), ((6, 9,
1), 7.655392897889694), ((6, 9, 2), 7.520216812620203), ((6, 9, 2), 7.701789510355
513), ((6, 9, 3), 7.538710778710327), ((6, 9, 3), 7.702222711462934), ((6, 9, 3),
7.758201807142663), ((6, 12, 1), 7.511904478807113), ((6, 12, 2), 7.44190689921876
6), ((6, 12, 2), 7.630176316286755), ((6, 12, 3), 7.6766157497654675), ((6, 12,
3), 7.735024433634457), ((6, 12, 3), 7.854389923905579), ((6, 15, 1), 7.7175317423
50928), ((6, 15, 2), 7.73242306590824), ((6, 15, 2), 7.441650381884773), ((6, 15,
3), 7.672576089553973), ((6, 15, 3), 7.547165163854142), ((6, 15, 3), 7.5436985116
593815), ((8, 6, 1), 8.065664829476388), ((8, 6, 2), 7.646174347195818), ((8, 6,
2), 7.938517853272982), ((8, 6, 3), 8.000122707977505), ((8, 6, 3), 8.031559020092
521), ((8, 6, 3), 8.238402695913866), ((8, 9, 1), 7.972136923660537), ((8, 9, 2),
7.875882985809661), ((8, 9, 2), 7.53327174635892), ((8, 9, 3), 7.950785831172573),
((8, 9, 3), 7.9848268403236045), ((8, 9, 3), 8.161529757555533), ((8, 12, 1), 8.25
6150111677055), ((8, 12, 2), 8.09521790821344), ((8, 12, 2), 8.090176718481898),
((8, 12, 3), 7.868775232914589), ((8, 12, 3), 7.910684024650727), ((8, 12, 3), 7.8
68395583024645), ((8, 15, 1), 8.13959065425757), ((8, 15, 2), 8.005730725748775),
((8, 15, 2), 7.969893947351144), ((8, 15, 3), 8.04051301809569), ((8, 15, 3), 7.83
227802820726), ((8, 15, 3), 8.07920579240023)]
Collected Average TFRecord Reading Speeds: [((2, 6, 1), 0.08843111991882324), ((2,
6, 3), 0.08271487553914388), ((2, 9, 2), 0.1221923828125), ((2, 12, 1), 0.16709065
437316895), ((2, 12, 3), 0.16546003023783365), ((2, 15, 2), 0.21114015579223633),
((4, 6, 1), 0.08778738975524902), ((4, 6, 3), 0.08428645133972168), ((4, 9, 2), 0.
1251223087310791), ((4, 12, 1), 0.1657581329345703), ((4, 12, 3), 0.17107303937276
205), ((4, 15, 2), 0.21554553508758545), ((6, 6, 1), 0.08746647834777832), ((6, 6,
3), 0.08028825124104817), ((6, 9, 2), 0.11998128890991211), ((6, 12, 1), 0.1687123
7754821777), ((6, 12, 3), 0.1632049878438314), ((6, 15, 2), 0.20051980018615723),
((8, 6, 1), 0.07994604110717773), ((8, 6, 3), 0.08158040046691895), ((8, 9, 2), 0.
12046372890472412), ((8, 12, 1), 0.16574668884277344), ((8, 12, 3), 0.164332230885
82358), ((8, 15, 2), 0.20592761039733887), ((2, 6, 2), 0.08465230464935303), ((2,
9, 1), 0.12089157104492188), ((2, 9, 3), 0.1227717399597168), ((2, 12, 2), 0.16557
037830352783), ((2, 15, 1), 0.2097020149230957), ((2, 15, 3), 0.2122104167938232
4), ((4, 6, 2), 0.08623874187469482), ((4, 9, 1), 0.12705588340759277), ((4, 9,
3), 0.12533299128214517), ((4, 12, 2), 0.17975986003875732), ((4, 15, 1), 0.215766
42990112305), ((4, 15, 3), 0.21206267674763998), ((6, 6, 2), 0.08056473731994629),
((6, 9, 1), 0.12715435028076172), ((6, 9, 3), 0.1440251668294271), ((6, 12, 2), 0.
1614246368408203), ((6, 15, 1), 0.20521211624145508), ((6, 15, 3), 0.2045592466990
153), ((8, 6, 2), 0.08079314231872559), ((8, 9, 1), 0.11949491500854492), ((8, 9,
3), 0.12468552589416504), ((8, 12, 2), 0.16286373138427734), ((8, 15, 1), 0.203693
62831115723), ((8, 15, 3), 0.20499277114868164)]
Collected Average JPEG Reading Speeds: [((2, 6, 1), 2.2435944080352783), ((2, 6,
3), 2.2706435521443686), ((2, 9, 2), 3.389087438583374), ((2, 12, 1), 4.4564423561
09619), ((2, 12, 3), 4.548510392506917), ((2, 15, 2), 5.586162090301514), ((4, 6,
1), 3.39974045753479), ((4, 6, 3), 3.4652063051859536), ((4, 9, 2), 5.200440526008
606), ((4, 12, 1), 6.905670404434204), ((4, 12, 3), 7.047130187352498), ((4, 15,
2), 8.742299914360046), ((6, 6, 1), 4.642761945724487), ((6, 6, 3), 4.631607611974
08), ((6, 9, 2), 7.096000790596008), ((6, 12, 1), 9.584786415100098), ((6, 12, 3),
9.28476349512736), ((6, 15, 2), 11.866696238517761), ((8, 6, 1), 5.95115232467651
4), ((8, 6, 3), 5.934234619140625), ((8, 9, 2), 9.349716186523438), ((8, 12, 1), 1
1.627695560455322), ((8, 12, 3), 12.178771575291952), ((8, 15, 2), 15.022962331771
85), ((2, 6, 2), 2.253878593444824), ((2, 9, 1), 3.3041653633117676), ((2, 9, 3),
3.364560842514038), ((2, 12, 2), 4.558676242828369), ((2, 15, 1), 5.75224876403808
6), ((2, 15, 3), 5.73637843132019), ((4, 6, 2), 3.528314232826233), ((4, 9, 1), 5.
```

```
1841044425964355), ((4, 9, 3), 5.286682208379109), ((4, 12, 2), 6.93785190582275
4), ((4, 15, 1), 8.626567602157593), ((4, 15, 3), 8.913949330647787), ((6, 6, 2),
4.668812155723572), ((6, 9, 1), 7.053850889205933), ((6, 9, 3), 7.0447891553243),
((6, 12, 2), 9.555578351020813), ((6, 15, 1), 11.661759614944458), ((6, 15, 3), 1
1.861861228942871), ((8, 6, 2), 6.16205894947052), ((8, 9, 1), 9.031455516815186),
((8, 9, 3), 8.964895486831665), ((8, 12, 2), 11.862547874450684), ((8, 15, 1), 14.
742756128311157), ((8, 15, 3), 15.032859643300375)]
WARNING:  Python 3.5-3.7 will be deprecated on August 8th, 2023. Please use Python
version 3.8 and up.

If you have a compatible Python interpreter installed, you can use it by setting
the CLOUDSDK_PYTHON environment variable to point to it.

Copying file:///tmp/avg_tfrecord_speeds_240504-2347.pkl [Content-Type=application/
octet-stream]...
/ [1 files][   1016 B/   1016 B]
Operation completed over 1 objects/1016.0 B.
File successfully saved and uploaded to: gs://earnest-crow-421721-storage/avg_tfre
cord_speeds_240504-2347.pkl
24/05/04 23:47:41 INFO org.spark_project.jetty.server.AbstractConnector: Stopped S
park@4a63dd5d{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [5edef5349c304554b9ec52f9b9a6b1a5] finished successfully.
done: true
driverControlFilesUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-meta
info/dac9d657-95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/
driverOutputResourceUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-me
tainfo/dac9d657-95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/
driveroutput
jobUuid: b40f45dc-ebc5-394d-9758-03bfa215097d
placement:
  clusterName: jobs2c
  clusterUuid: dac9d657-95e8-4f9e-a526-b8143e1e5a04
pysparkJob:
  mainPythonFileUri: gs://earnest-crow-421721-storage/google-cloud-dataproc-metain
fo/dac9d657-95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/stag
ing/spark_job2.py
reference:
  jobId: 5edef5349c304554b9ec52f9b9a6b1a5
  projectId: earnest-crow-421721
status:
  state: DONE
  stateStartTime: '2024-05-04T23:47:46.292032Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-05-04T23:34:51.682542Z'
- state: SETUP_DONE
  stateStartTime: '2024-05-04T23:34:51.707941Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-05-04T23:34:52.024971Z'
yarnApplications:
- name: spark_job2.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://jobs2c-m:8088/proxy/application_1714865587157_0001/
```

2d) Retrieve, analyse and discuss the output (12%)

Run the tests over a wide range of different paramters and list the results in a table.

Perform a **linear regression** (e.g. using scikit-learn) over **the values for each parameter** and for the **two cases** (reading from image files/reading TFRecord files). List a **table** with the

output and interpret the results in terms of the effects of overall.
Also, **plot** the output values, the averages per parameter value and the regression lines for each parameter and for the product of batch_size and batch_number

Discuss the **implications** of this result for **applications** like large-scale machine learning. Keep in mind that cloud data may be stored in distant physical locations. Use the numbers provided in the PDF latency-numbers document available on Moodle or here for your arguments.

How is the **observed** behaviour **similar or different** from what you'd expect from a **single machine**? Why would cloud providers tie throughput to capacity of disk resources?

By **parallelising** the speed test we are making **assumptions** about the limits of the bucket reading speeds. See here for more information. Discuss, **what we need to consider** in **speed tests** in parallel on the cloud, which bottlenecks we might be identifying, and how this relates to your results.

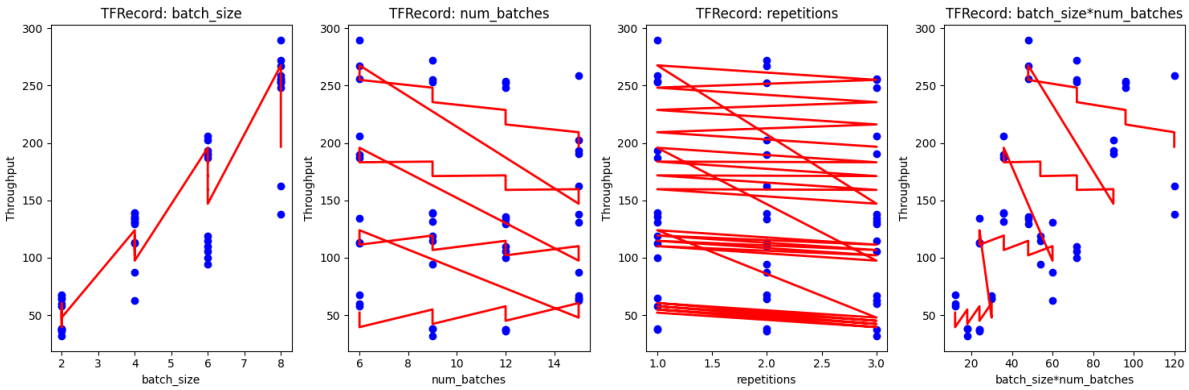
Discuss to what extent **linear modelling** reflects the **effects** we are observing. Discuss what could be expected from a theoretical perspective and what can be useful in practice.

Write your **code below** and **include the output** in your submitted `ipynb` file. Provide the answer **text in your report**.

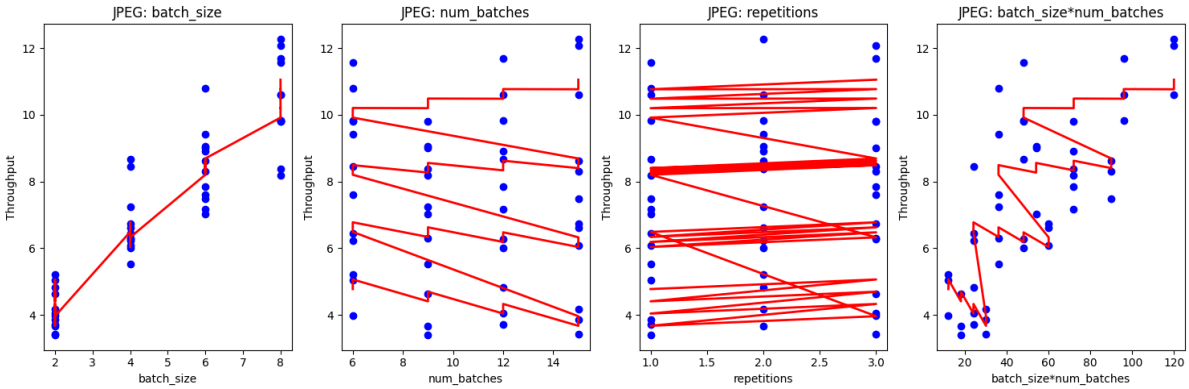
In [18]:
```
### CODING TASK ###

# Importing libraries
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np
from pyspark.sql import SparkSession

# Spark session initialisation
spark = SparkSession.builder.appName("Image Processing Analysis").getOrCreate()

# Converting RDD to DataFrame
def rdd_to_df(rdd, schema):
    return spark.createDataFrame(rdd, schema)

# Schema defining for TFRecord & JPEG results
schema = ["params", "throughput"]
tfrecord_df = rdd_to_df(processed_tfrecord_images_rdd, schema)
jpeg_df = rdd_to_df(processed_jpeg_images_rdd, schema)

# Spark to Pandas DataFrame conversion
tfrecord_s_pd = tfrecord_df.toPandas()
jpeg_s_pd = jpeg_df.toPandas()

# Expanding 'batch_size', 'num_batches', 'repetitions' into different columns &
# Adding a new column for product of batch_size and num_batches

# For TFRecords
tfrecord_s_pd[['batch_size', 'num_batches', 'repetitions']] = pd.DataFrame(tfrecord
tfrecord_s_pd['batch_size*num_batches'] = tfrecord_s_pd['batch_size'] * tfrecord_s_

# For JPEG images
jpeg_s_pd[['batch_size', 'num_batches', 'repetitions']] = pd.DataFrame(jpeg_s_pd['p
jpeg_s_pd['batch_size*num_batches'] = jpeg_s_pd['batch_size'] * jpeg_s_pd['num_batc
```

```
# Regression and plotting of results

def reg_plot(df, title):
    features = ['batch_size', 'num_batches', 'repetitions', 'batch_size*num_batches
    X = df[features]
    y = df['throughput']

    model = LinearRegression()
    model.fit(X, y)

    # Predicting the values for plotting
    predicted = model.predict(X)

    # Plotting
    plt.figure(figsize=(15, 5))
    for i, feature in enumerate(features):
        plt.subplot(1, len(features), i + 1)
        plt.scatter(df[feature], y, color='blue')
        plt.plot(df[feature], predicted, color='red', linewidth=2)
        plt.title(f'{title}: {feature}')
        plt.xlabel(feature)
        plt.ylabel('Throughput')
    plt.tight_layout()
    plt.show()

    return model.intercept_, model.coef_

# Performing regression and plotting for TFRecord and JPEG
print("TFRecord Results for regression:")
tfrecord_intcpt, tfrecord_co = reg_plot(tfrecord_s_pd, "TFRecord")
print("Intercept:", tfrecord_intcpt)
print("Coefficients:", tfrecord_co)

print("JPEG Results for regression:")
jpeg_intcpt, jpeg_co = reg_plot(jpeg_s_pd, "JPEG")
print("Intercept:", jpeg_intcpt)
print("Coefficients:", jpeg_co)
```

```
TFRecord Regression Results:
```




```
Intercept: -33.582855010889915
Coefficients: [43.28031465  3.3896605  -6.31060232 -1.23251214]
JPEG Regression Results:
```




```
Intercept: 4.085645831426995
Coefficients: [ 0.64016776 -0.19461936  0.14424194  0.03612301]
```

Section 3. Theoretical discussion

Task 3: Discussion in context. (24%)

In this task we refer an idea that is introduced in this paper:

- Alipourfard, O., Liu, H. H., Chen, J., Venkataraman, S., Yu, M., & Zhang, M. (2017). Cherrypick: Adaptively unearthing the best cloud configurations for big data analytics.. In USENIX NSDI 17 (pp. 469-482).

Alipourfard et al (2017) introduce the prediction an optimal or near-optimal cloud configuration for a given compute task.

3a) Contextualise

Relate the previous tasks and the results to this concept. (It is not necessary to work through the full details of the paper, focus just on the main ideas). To what extent and under what conditions do the concepts and techniques in the paper apply to the task in this coursework? (12%)

3b) Strategise

Define - as far as possible - concrete strategies for different application scenarios (batch, stream) and discuss the general relationship with the concepts above. (12%)

Provide the answers to these questions in your report.

Final cleanup

Once you have finshed the work, you can delete the buckets, to stop incurring cost that depletes your credit.

In [48]:
```
!gsutil -m rm -r $BUCKET/* # Empty your bucket
!gsutil rb $BUCKET # delete the bucket
```

```
Removing gs://earnest-crow-421721-storage/jpeg_results_240504-1038.pickle#17148190
92797024...
Removing gs://earnest-crow-421721-storage/avg_tfrecord_speeds.pkl#171456598165573
9...
Removing gs://earnest-crow-421721-storage/avg_tfrecord_speeds_240504-2347.pkl#1714
866461212171...
Removing gs://earnest-crow-421721-storage/jpeg_results_240504-2257.pickle#17148634
68232104...
Removing gs://earnest-crow-421721-storage/avg_tfrecord_speeds_240501-1220.pkl#1714
566002873032...
Removing gs://earnest-crow-421721-storage/avg_tfrecord_speeds_240501-1238.pkl#1714
567121736339...
Removing gs://earnest-crow-421721-storage/tfrecord_results.pickle#171492079370581
1...
Removing gs://earnest-crow-421721-storage/jpeg_results.pickle#1714920794871306...
Removing gs://earnest-crow-421721-storage/jpeg_results_240504-2325.pickle#17148651
52067906...
Removing gs://earnest-crow-421721-storage/tfrecord_results_240504-1038.pickle#1714
819092346221...
Removing gs://earnest-crow-421721-storage/tfrecord_results_240504-2257.pickle#1714
863467809939...
Removing gs://earnest-crow-421721-storage/tfrecord_results_240504-2325.pickle#1714
865151847827...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/cluster.properties#1714347522924375...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714347805800079...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714347805836061...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714347746085279...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714347746086322...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714347719360052...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714347719717993...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714347735380114...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714347735478198...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714347664640288...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714347664830398...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714347736588388...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714347736584539...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714347663949220...
```

```
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714347664274844...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714347716609636...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714347716679842...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714347662772373...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714347662911924...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714347719498700...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714347719466133...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714347663534867...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714347663702904...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714347722305494...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714347721282013...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714347662440244...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714347662602604...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714347723386908...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714347723458082...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714347665141013...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714347665354839...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714347725584960...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714347725704977...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714347656925199...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/11057c83-
adc6-4e58-9d43-e39b3db9039d/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714347657088145...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
```

```
f94e-41ec-978b-94e9aa2fa7dc/cluster.properties#1714811635732943...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-initialization-s
cript-0_output#1714811824305974...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-initialization-s
cripts_component-stats#1714811824307554...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-post-hdfs-startu
p-script_component-stats#1714811780785526...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-post-hdfs-startu
p-script_output#1714811780772405...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-startup-script_c
omponent-stats#1714811769189237...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/earnest-crow-421721-single-m/dataproc-startup-script_o
utput#1714811769260202...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/dc23433cbfb04eac9f35f8cf03bd970a/driveroutput.000
000000#1714812249355153...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/dc23433cbfb04eac9f35f8cf03bd970a/driveroutput.000
000001#1714812249543381...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/dc23433cbfb04eac9f35f8cf03bd970a/staging/spark_wr
ite_tfrec.py#1714812223750477...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/driveroutput.000
000000#1714812304643422...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/driveroutput.000
000001#1714812305289632...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/208e6e9f-
f94e-41ec-978b-94e9aa2fa7dc/jobs/e2133a84abdf451193b7041ed2013afd/staging/spark_wr
ite_tfrec.py#1714812286597123...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/cluster.properties#1714351662066328...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714352070472801...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714352070453708...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714351922590878...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714351922520138...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714351890032834...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714351890320657...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714351876098724...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714351876111075...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
```

```
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714351803729564...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714351803895365...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714351894834754...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714351894830265...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714351814928854...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714351815460984...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714351900013450...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714351900007473...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714351798588882...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714351798757958...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714351856071170...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714351856009544...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714351800475252...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714351800639655...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714351856501338...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714351856429116...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714351799562557...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714351799723490...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714351860004480...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714351860055426...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714351805110522...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
```

```
ut#1714351805300822...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714351874183708...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714351874178595...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714351802923431...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/213a771d-
7463-41c9-ab86-e821cb2d544d/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714351803086289...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/cluster.properties#1714350928790756...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714351434421712...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714351434426994...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714351245755567...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714351245758858...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714351200090098...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714351200465999...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714351224279644...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714351224275343...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714351148777785...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714351148915513...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714351253910365...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714351254003428...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714351136318875...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714351136488763...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714351258986644...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714351258998504...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
```

```
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714351182004417...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/26109405-
8ac2-4c85-a2b0-ce610b1d9af3/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714351182267424...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/cluster.properties#1714808313187527...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714808543743566...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714808543664699...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714808449668539...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714808449657052...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714808439044866...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/earnest-crow-421721-m/dataproc-startup-script_output#1
714808439044847...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/29f45e51-
9b98-4453-8c06-b21a4c91734a/jobs/4e883283dd2e4616ba3ce0edd0fdef39/staging/spark_wr
ite_tfrec.py#1714811600548431...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/cluster.properties#1714766160922292...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714766488148737...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714766488132465...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714766424811426...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714766424693147...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714766373381349...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714766389908638...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714766373502614...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714766389910181...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714766309068035...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714766309725272...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714766404020843...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
```

```
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714766404091430...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714766318188184...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714766317806668...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714766376385281...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714766376411362...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714766310918845...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714766311273633...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714766377879223...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714766377950209...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714766311867607...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714766312049020...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714766392770838...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714766392719155...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714766322606515...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714766322857020...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714766391333643...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714766391260810...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714766324149404...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714766324324175...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714766394374606...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714766394345258...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
```

```
onent-stats#1714766311589774...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2dfd6b4d-
1d8a-4132-ac3e-633b85052cad/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714766311817776...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/cluster.properties#1714747797402852...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714748107220366...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714748107297171...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714748042029930...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714748041952948...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714748004832368...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714748005049312...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714748026272113...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714748026350733...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714747948705339...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714747949543105...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714748015267976...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714748015276362...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714747940105060...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714747940318378...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714748069986453...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714748070063821...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714747951858406...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714747952004114...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714748005097576...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
```

```
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714748005093862...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714747943866255...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714747944049002...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714748004389469...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714748004389208...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714747939827084...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714747940027742...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714748012477009...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714748012522946...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714747948162357...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714747948486077...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714748036622909...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714748036621806...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714747955821852...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/2ff62968-
a6bc-45e9-b863-db33524e9fe1/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714747956280085...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/cluster.properties#1714767105786828...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714767440518151...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714767440534693...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714767372970658...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714767372928161...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714767340656137...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714767340953829...
```

```
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714767342297298...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714767342286233...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714767258388721...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714767258556521...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714767349618379...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714767349603521...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714767261844333...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714767262712438...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714767313057918...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714767313137150...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714767252225791...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714767252443681...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714767374914720...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714767374925179...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714767258560582...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714767353446114...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714767258804536...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714767353398203...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714767260619082...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714767260875326...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714767360129091...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
```

```
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714767360050293...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714767258083064...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714767258382808...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714767337734818...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714767337795323...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714767257213804...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/42a96780-
e9a9-43d9-93a3-507e6e54898f/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714767257359509...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/cluster.properties#1714746698505568...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714746914440536...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714746914442394...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714746870466600...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714746890843432...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714746858945215...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/4fe1416d-
3593-4d9e-bad2-f7c05e8b3b1b/earnest-crow-421721-m/dataproc-startup-script_output#1
714746859018299...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/cluster.properties#1714746000796049...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714746179779581...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714746179759164...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714746136658825...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714746136667203...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714746125197974...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/5f180138-
af8b-48d0-8d5e-d4c43026c793/earnest-crow-421721-m/dataproc-startup-script_output#1
714746125271712...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/cluster.properties#1714767745538871...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-initialization-script-0
```

```
_output#1714767945542356...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714767945550545...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714767903300086...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714767903302524...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714767892876714...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/614aa826-
ecfb-4119-94ad-5899d8836890/earnest-crow-421721-m/dataproc-startup-script_output#1
714767892988392...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/cluster.properties#1714770404749453...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714770595482501...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714770595490047...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714770549137574...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714770549211163...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714770537521273...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/67f853ea-
ea0c-4a67-8236-5875c4d3b86e/earnest-crow-421721-m/dataproc-startup-script_output#1
714770537593249...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/cluster.properties#1714346471030327...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714346748787176...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714346753780000...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714346690564859...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714346690478381...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714346664730458...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714346664956918...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714346674593430...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714346674594874...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
```

```
onent-stats#1714346607732974...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714346607839783...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714346682695470...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714346682693438...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714346611039392...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714346611225887...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714346690000469...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714346690070989...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714346617437445...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714346617756674...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714346673602602...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714346673604795...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714346607036868...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714346607192519...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714346668432469...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714346668387118...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714346611525060...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714346611687037...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714346682391753...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714346682293095...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714346623086894...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714346623228971...
```

```
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714346697620110...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714346697597488...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714346619355857...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/775f4571-
a23f-41e7-8473-2d5fe38d0535/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714346619510024...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/cluster.properties#1714352825503981...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714353131372905...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714353131366026...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714353069903375...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714353069830234...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714353035963997...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714353036208990...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714353037887590...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714353037883166...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714352966263086...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714352966426848...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714353025118621...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714353025143433...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714352956890649...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714352957223334...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714353063328654...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714353063250128...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
```

```
onent-stats#1714352962611862...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714352962806181...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714353021095365...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714353021097887...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714352964515603...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714352964731118...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714353018927368...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714353018917924...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714352963551097...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714352963732906...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714353013889602...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714353013926925...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714352957771147...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714352957928412...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714353039077613...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714353039077208...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714352967788124...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/7d0c0b32-
a2e5-4a19-944f-afde9b2edf69/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714352968101592...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/cluster.properties#1714349926447710...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714350204760325...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714350204824488...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714350145825072...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
```

```
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714350145735558...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714350121013578...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714350121254913...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714350174471895...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714350174497984...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714350071790663...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714350072064725...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714350143412112...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714350143475355...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714350067558593...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714350067792118...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714350141639464...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714350141616278...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714350083312375...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714350083484295...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714350139836296...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714350139901032...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714350078679210...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714350079019560...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714350129077356...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714350129011593...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
```

```
onent-stats#1714350069266804...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714350069509511...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714350142502527...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714350142489313...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714350081904068...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714350082154345...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714350137736729...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714350137728345...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714350066016599...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8be08a8d-
b1e3-4a34-8594-cc7419942b7e/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714350066138539...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/cluster.properties#1714768099909829...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714768324209294...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714768324214824...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714768278770553...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714768278738836...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714768267107499...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/8d4ffe92-
eca4-4ff1-9d4c-cd07d6311cd1/earnest-crow-421721-m/dataproc-startup-script_output#1
714768267156861...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/cluster.properties#1714863471060247...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/driveroutput.000
000000#1714865153469743...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/driveroutput.000
000001#1714865153669383...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/jobs/42f4fda3b85d4cc882d985e4d0208ccc/staging/spark_jo
b2.py#1714863667172843...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-initialization-script-0_output#17148
63659384505...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-
cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-initialization-scripts_component-sta
```

```
ts#1714863659377410...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-post-hdfs-startup-script_component-stats#1714863609074987...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-post-hdfs-startup-script_output#1714863609003945...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-startup-script_component-stats#1714863598436835...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/90181cc6-cd42-4d7f-9f64-82c18db1a9c8/jobs2b-m/dataproc-startup-script_output#1714863598508880...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/cluster.properties#1714768485906978...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-initialization-script-0_output#1714768793218837...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-initialization-scripts_component-stats#1714768793229278...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-script_component-stats#1714768730600650...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-script_output#1714768730541663...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-startup-script_component-stats#1714768697423513...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-0/dataproc-initialization-script-0_output#1714768701684312...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-m/dataproc-startup-script_output#1714768697637794...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-0/dataproc-initialization-scripts_component-stats#1714768701683696...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-0/dataproc-startup-script_component-stats#1714768624572419...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-0/dataproc-startup-script_output#1714768624726462...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-1/dataproc-initialization-script-0_output#1714768714494694...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-1/dataproc-initialization-scripts_component-stats#1714768714509383...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-1/dataproc-startup-script_component-stats#1714768630769066...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-1/dataproc-startup-script_output#1714768630987927...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-2/dataproc-initialization-script-0_output#1714768689811101...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-2/dataproc-initialization-scripts_component-stats#1714768689815050...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
```

```
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714768625901186...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714768626192730...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714768724088750...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714768724097945...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714768634758446...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714768634913664...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714768692513286...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714768692507619...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714768628970402...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714768629384522...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714768689183806...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714768689157076...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714768630777808...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714768630555563...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714768713460603...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714768713550559...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714768631537297...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/9e36c6b1-
ab9e-49fe-bfe8-488737d59326/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714768632433980...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/cluster.properties#1714345732557280...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714345929758807...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714345929685167...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714345883475679...
```

```
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714345883553526...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714345871714522...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a48e9117-
5aaf-41c4-be18-469af472a835/earnest-crow-421721-m/dataproc-startup-script_output#1
714345871792956...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/cluster.properties#1714766668452930...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714766859329725...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714766859318853...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714766811547572...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714766811527588...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714766801009642...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/a9030b6d-
bccf-4918-854c-155b04499fdb/earnest-crow-421721-m/dataproc-startup-script_output#1
714766801114004...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/cluster.properties#1714826153166239...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-initialization-script-0_output#1
714826341910510...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-initialization-scripts_component
-stats#1714826341902714...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-post-hdfs-startup-script_compone
nt-stats#1714826300018452...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-post-hdfs-startup-script_output#
1714826300011032...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-startup-script_component-stats#1
714826289535087...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c4ae86d6-
7b12-4edc-ae3c-88c3500e4101/job-for-2a-m/dataproc-startup-script_output#1714826289
634680...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/cluster.properties#1714349610743126...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714349798269818...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714349798271352...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714349751778972...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714349751788525...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
```

```
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714349740123206...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/c5bfd593-
03ee-4fcf-ba7f-0a5609eb94e2/earnest-crow-421721-m/dataproc-startup-script_output#1
714349740157087...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/cluster.properties#1714347160875457...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714347383786355...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714347383776594...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714347340484249...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714347340484614...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714347328814435...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cde3c40e-
56ae-42a3-8a48-1fef94d97e33/earnest-crow-421721-m/dataproc-startup-script_output#1
714347328885939...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/cluster.properties#1714599571003140...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714599859314522...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714599859387787...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714599800510256...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714599800420566...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714599768471320...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714599768983701...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714599797310692...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714599797326901...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714599714756578...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714599715089293...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714599782583705...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714599782593276...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
```

```
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714599708748264...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714599708877845...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714599771268856...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714599771260542...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714599715821026...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714599715992319...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714599770919245...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714599770918333...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714599708996532...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714599709300223...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714599776808888...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714599776799216...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714599715309512...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714599715476670...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714599761011249...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714599761021879...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714599707373424...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714599707467586...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714599830008617...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714599830004820...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714599707101181...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/cdf8fe84-
660b-4260-88e9-979dcb932226/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
```

```
ut#1714599707301467...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/cluster.properties#1714598914368512...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714599092912569...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714599092901586...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714599048079124...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714599048047716...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714599037521924...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/d3eca7b9-
a8ae-4b12-8c8a-a74a9cb6004c/earnest-crow-421721-m/dataproc-startup-script_output#1
714599037592984...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/cluster.properties#1714865501051624...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/driveroutput.000
000000#1714866462977693...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/driveroutput.000
000001#1714866463180103...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs/5edef5349c304554b9ec52f9b9a6b1a5/staging/spark_jo
b2.py#1714865691447740...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-initialization-script-0_output#17148
65684523541...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-initialization-scripts_component-sta
ts#1714865684444139...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-post-hdfs-startup-script_component-s
tats#1714865635349647...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-post-hdfs-startup-script_output#1714
865635353296...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-startup-script_component-stats#17148
65624881962...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dac9d657-
95e8-4f9e-a526-b8143e1e5a04/jobs2c-m/dataproc-startup-script_output#17148656249609
79...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/cluster.properties#1714362458129117...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714362781004131...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714362776006659...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714362712013394...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714362711971471...
```

```
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714362683057950...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714362683230894...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714362685522784...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714362685457507...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714362613681505...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714362613957765...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714362673216649...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714362673151124...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714362604827999...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714362605020644...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714362674044175...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714362674110590...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714362614796618...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714362614990085...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pt-0_output#1714362674546388...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-3/dataproc-initialization-scri
pts_component-stats#1714362674534174...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-3/dataproc-startup-script_comp
onent-stats#1714362616963186...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-3/dataproc-startup-script_outp
ut#1714362617096136...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pt-0_output#1714362671480319...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-4/dataproc-initialization-scri
pts_component-stats#1714362671496960...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-4/dataproc-startup-script_comp
onent-stats#1714362613204741...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
```

```
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-4/dataproc-startup-script_outp
ut#1714362613408217...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pt-0_output#1714362680991835...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-5/dataproc-initialization-scri
pts_component-stats#1714362680983445...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-5/dataproc-startup-script_comp
onent-stats#1714362620826350...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-5/dataproc-startup-script_outp
ut#1714362621010844...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pt-0_output#1714362672918594...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-6/dataproc-initialization-scri
pts_component-stats#1714362672921469...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-6/dataproc-startup-script_comp
onent-stats#1714362605192331...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/dc076ea0-
910f-4afb-8a45-05804da54a00/earnest-crow-421721-1-w-6/dataproc-startup-script_outp
ut#1714362605426369...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/cluster.properties#1714355820374118...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714356029639527...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714356029642540...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714355964543570...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714355964466189...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714355954084256...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e3ba66cc-
24e2-469a-a12c-dcc6401be525/earnest-crow-421721-m/dataproc-startup-script_output#1
714355954137762...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/cluster.properties#1714765859410365...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-initialization-script-0
_output#1714766035309402...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-initialization-scripts_
component-stats#1714766035313694...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_component-stats#1714765990173096...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-post-hdfs-startup-scrip
t_output#1714765990170233...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-startup-script_componen
t-stats#1714765979643044...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/e738a3a5-
```

```
3c28-4b6b-9dbc-5c8d707a6239/earnest-crow-421721-m/dataproc-startup-script_output#1
714765979716565...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/cluster.properties#1714350562431921...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-initialization-script
-0_output#1714350848836405...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-initialization-script
s_component-stats#1714350848836488...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_component-stats#1714350792182635...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-post-hdfs-startup-scr
ipt_output#1714350792091361...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-startup-script_compon
ent-stats#1714350760901895...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-m/dataproc-startup-script_output
#1714350761243048...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pt-0_output#1714350783689478...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-0/dataproc-initialization-scri
pts_component-stats#1714350783702317...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-0/dataproc-startup-script_comp
onent-stats#1714350709001958...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-0/dataproc-startup-script_outp
ut#1714350709187571...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pt-0_output#1714350818117313...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-1/dataproc-initialization-scri
pts_component-stats#1714350818137607...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-1/dataproc-startup-script_comp
onent-stats#1714350707259885...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-1/dataproc-startup-script_outp
ut#1714350707431222...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pt-0_output#1714350799570934...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-2/dataproc-initialization-scri
pts_component-stats#1714350799539825...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-2/dataproc-startup-script_comp
onent-stats#1714350705417724...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ebe23361-
adae-4093-bd82-fc0e6a4b2877/earnest-crow-421721-1-w-2/dataproc-startup-script_outp
ut#1714350705585485...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/cluster.properties#1714812511884010...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-initialization-
script-0_output#1714812814057221...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
```

```
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-initialization-
scripts_component-stats#1714812814053530...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-post-hdfs-start
up-script_component-stats#1714812753698659...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-post-hdfs-start
up-script_output#1714812753627540...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-startup-script_
component-stats#1714812719098381...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-m/dataproc-startup-script_
output#1714812719560490...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-0/dataproc-initializatio
n-script-0_output#1714812736638959...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-0/dataproc-initializatio
n-scripts_component-stats#1714812736667936...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-0/dataproc-startup-scrip
t_component-stats#1714812659351899...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-0/dataproc-startup-scrip
t_output#1714812660197237...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-1/dataproc-initializatio
n-script-0_output#1714812740755194...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-1/dataproc-initializatio
n-scripts_component-stats#1714812740684846...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-1/dataproc-startup-scrip
t_component-stats#1714812660929412...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-1/dataproc-startup-scrip
t_output#1714812661111778...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-2/dataproc-initializatio
n-script-0_output#1714812718192788...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-2/dataproc-initializatio
n-scripts_component-stats#1714812718195030...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-2/dataproc-startup-scrip
t_component-stats#1714812658393438...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-2/dataproc-startup-scrip
t_output#1714812658561525...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-3/dataproc-initializatio
n-script-0_output#1714812723060280...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-3/dataproc-initializatio
n-scripts_component-stats#1714812723022602...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-3/dataproc-startup-scrip
t_component-stats#1714812659096981...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-3/dataproc-startup-scrip
t_output#1714812659241684...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-4/dataproc-initializatio
```

```
n-script-0_output#1714812717834908...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-4/dataproc-initializatio
n-scripts_component-stats#1714812717905555...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-4/dataproc-startup-scrip
t_component-stats#1714812654357142...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-4/dataproc-startup-scrip
t_output#1714812654548483...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-5/dataproc-initializatio
n-script-0_output#1714812748480799...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-5/dataproc-initializatio
n-scripts_component-stats#1714812748480862...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-5/dataproc-startup-scrip
t_component-stats#1714812678161597...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-5/dataproc-startup-scrip
t_output#1714812678484822...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-6/dataproc-initializatio
n-script-0_output#1714812724541260...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-6/dataproc-initializatio
n-scripts_component-stats#1714812724583846...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-6/dataproc-startup-scrip
t_component-stats#1714812656156579...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/earnest-crow-421721-maximal-w-6/dataproc-startup-scrip
t_output#1714812656207781...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/2ae5b2076df846d4997dedaa780ed7be/driveroutput.000
000000#1714812863898628...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/2ae5b2076df846d4997dedaa780ed7be/driveroutput.000
000001#1714812864909053...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/2ae5b2076df846d4997dedaa780ed7be/staging/spark_wr
ite_tfrec.py#1714812819418496...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/driveroutput.000
000000#1714813061371941...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/driveroutput.000
000001#1714813062031933...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/ecce2160-
dff6-428c-bd1c-e2aae8e1542f/jobs/482e7379032f4d96b79840df50e3abd3/staging/spark_wr
ite_tfrec.py#1714813032110490...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/cluster.properties#1714830623477378...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-initialization-script-0_output#1
714830829493778...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-initialization-scripts_component
-stats#1714830829423732...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-post-hdfs-startup-script_compone
nt-stats#1714830782027749...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
```

```
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-post-hdfs-startup-script_output#
1714830782024161...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-startup-script_component-stats#1
714830771626492...
Removing gs://earnest-crow-421721-storage/google-cloud-dataproc-metainfo/f443e9b7-
91c8-4abf-8b08-f631e386703f/job-for-2b-m/dataproc-startup-script_output#1714830771
753506...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers00-230.t
frec#1714911774532402...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers01-230.t
frec#1714911826949685...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers02-230.t
frec#1714911863197799...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers03-230.t
frec#1714911902870964...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers04-230.t
frec#1714911938694418...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers05-230.t
frec#1714912007163186...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers06-230.t
frec#1714912069481402...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers07-230.t
frec#1714912112437615...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers08-230.t
frec#1714912148222930...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers09-230.t
frec#1714912187117984...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers10-230.t
frec#1714912267977266...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers11-230.t
frec#1714912347054913...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers12-230.t
frec#1714912407929517...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers13-230.t
frec#1714912450581947...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers14-230.t
frec#1714912493459891...
Removing gs://earnest-crow-421721-storage/tfrecords-jpeg-192x192-2/flowers15-220.t
frec#1714912568917073...
/ [631/631 objects] 100% Done 26.08 objects/s ETA 00:00:00
Operation completed over 631 objects.
Removing gs://earnest-crow-421721-storage/...
```