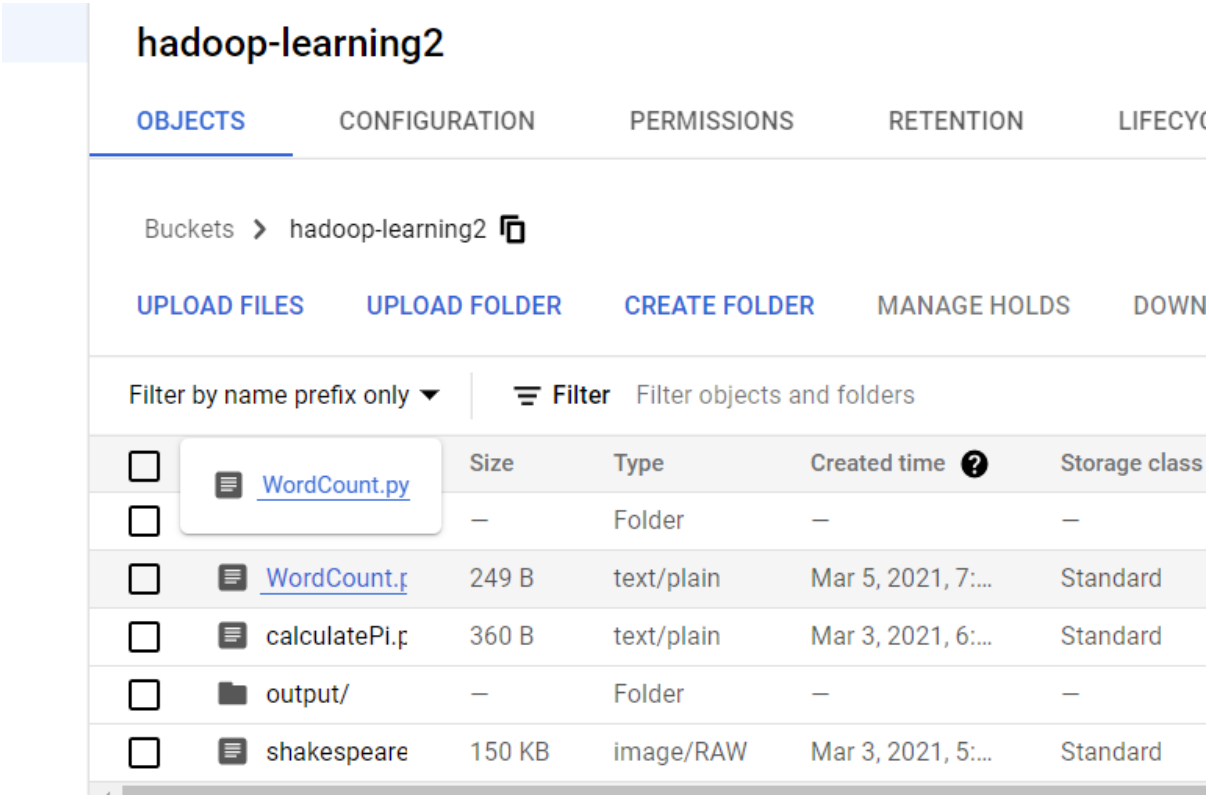
MapReduce job run in Google Dataproc

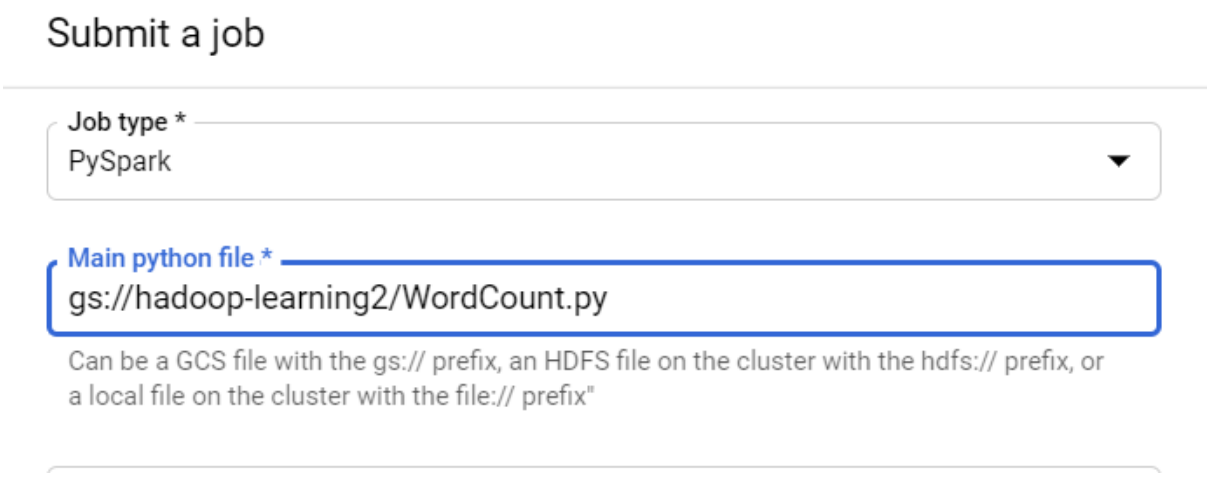
1. Prepared PySpark script WordCount.py

```
from pyspark import SparkContext
sc =SparkContext()
file = "gs://hadoop-learning2/shakespeare.raw"
words = sc.textFile(file).flatMap(lambda line: line.split(" "))
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)
wordCounts.saveAsTextFile("gs://hadoop-learning2/output2/")
```

2. Uploaded the script file and data file to Google Cloud bucket



3. Submit a job with the files in Google Dataproc



4. Job complete

Start time:	Mar 5, 2021, 8:11:02 AM
Elapsed time:	32 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-a4ef
Job type	PySpark
Main python file	gs://hadoop-learning2/WordCount.py
Labels	

Job output LINE WRAP: OFF

```
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
21/03/05 00:11:31 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired '
21/03/05 00:11:32 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired '
21/03/05 00:11:32 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired '
21/03/05 00:11:33 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@11fe2512{HTTP/1.1, (http/1.1)}{0.0.0.0:0}

Job output is complete
```

5. Results generated successfully.

Buckets > hadoop-learning2 > output2

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by name prefix only Filter Filter objects and folders

Name	Size	Type	Created time	Storage class	Last modified
_SUCC	0 B	application/octet-stream	Mar 5, 2021, 8:...	Standard	Mar 5, 202...
part-0(	44.7 KB	application/octet-stream	Mar 5, 2021, 8:...	Standard	Mar 5, 202...
part-0(	44.6 KB	application/octet-stream	Mar 5, 2021, 8:...	Standard	Mar 5, 202...