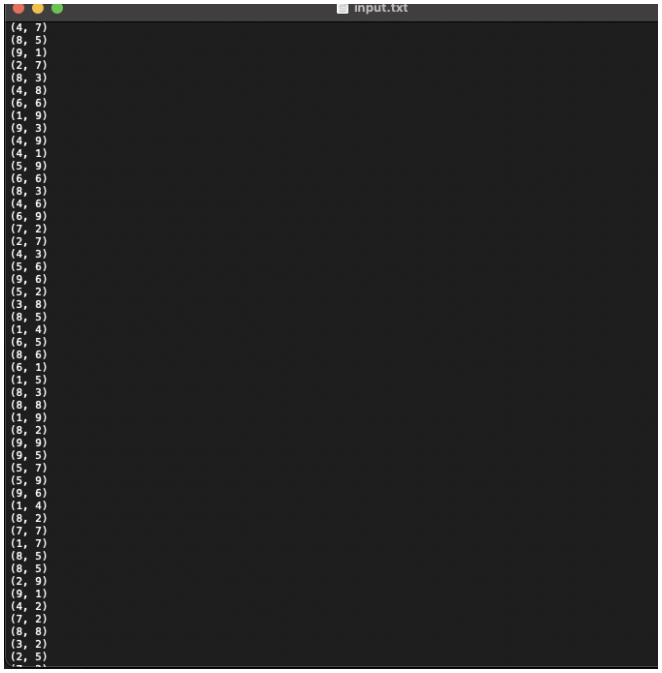
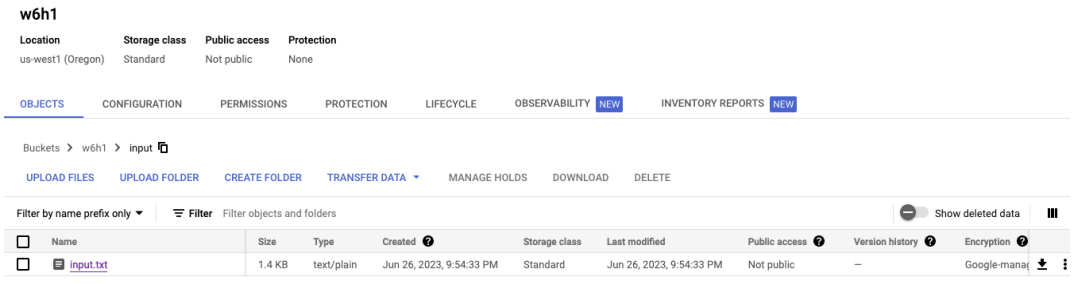
1. Randomly generate 20 coordinates in (x,y) format and write them into the input.txt

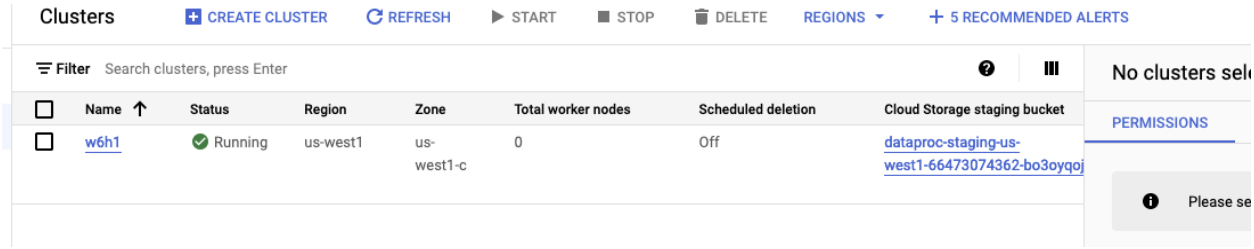
```
import random

# Generate 20 random coordinates
coordinates = [(random.randint(1, 9), random.randint(1, 9)) for _ in range(200)]

# Write the coordinates to the input file
with open("input.txt", "w") as file:
    for x, y in coordinates:
        print(f"({x}, {y})")
        file.write(f"({x}, {y})\n")
file.close()
```



2. Create a Bucket in GCP and upload the input.txt



3. Create dataproc Clusters with the same region as the Bucket



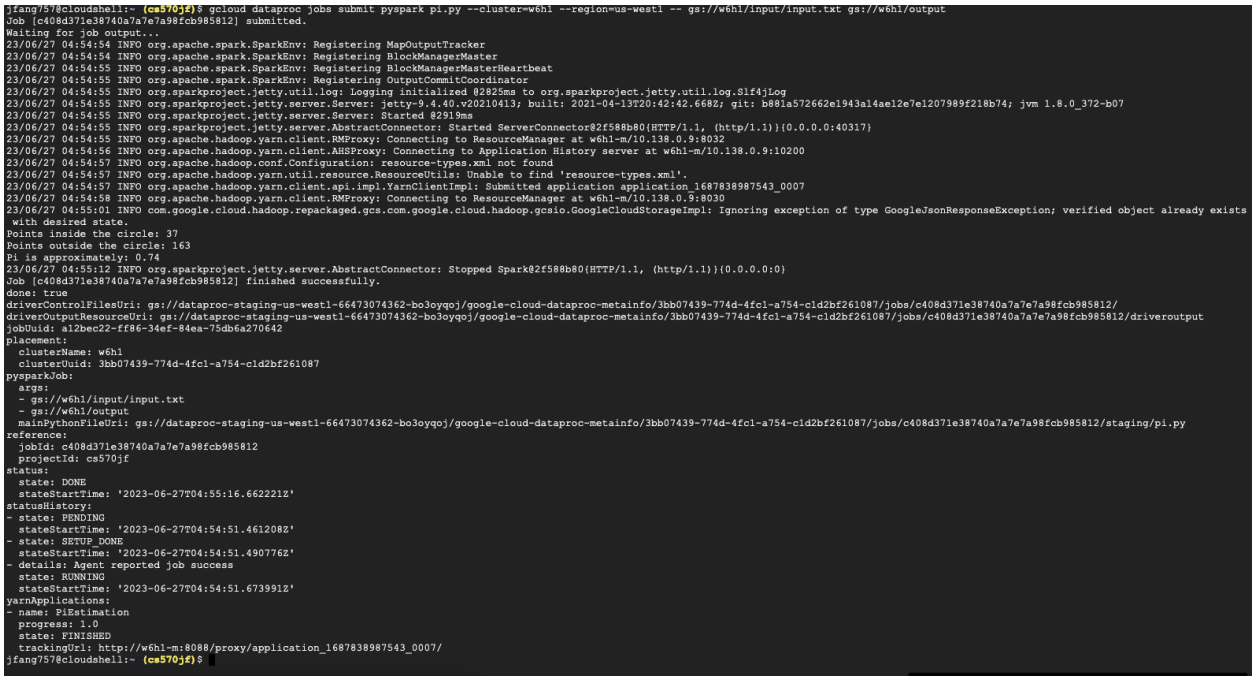
4. Pyspark Calculating Pi program

```
from pyspark.sql import SparkSession
import sys

# Create a SparkSession
spark = SparkSession.builder.appName("PiEstimation").getOrCreate()

if len(sys.argv) != 3:
   raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")
inputUri = sys.argv[1]
outputUri = sys.argv[2]

# Read the input file containing 20 coordinates (x, y)
coordinates = spark.read.text(inputUri)

# Define the function to calculate if a point is inside the circle
# Radius = 5
def points(row):
   x, y = map(float, row.value[1:-1].split(','))
   if x**2 + y**2 <= 5**2:
       return "inside"
   else:
       return "outside"

# Calculate the number of points inside and outside the unit circle
point_counts = coordinates.rdd.map(points).countByValue()

# Get the count of points inside the circle
inside_circle_count = point_counts.get("inside", 0)

# Get the count of points outside the circle
outside_circle_count = point_counts.get("outside", 0)

# Calculate the total number of points
```

```
total_count = coordinates.count()

# Estimate the value of pi
pi_estimate = 4.0 * inside_circle_count / total_count

# Print
print("Points inside the circle:", inside_circle_count)
print("Points outside the circle:", outside_circle_count)
print("Pi is approximately:", pi_estimate)

# Stop the SparkSession
spark.stop()
```

5. Run pi.py

gcloud dataproc jobs submit pyspark pi.py --cluster=w6h1 --region=us-west1 -- gs://w6h1/input/input.txt gs://w6h1/output



```
jfang757@cloudshell:~ (cs570jf)$ gcloud dataproc jobs submit pyspark pi.py --cluster=w6h1 --region=us-west1 -- gs://w6h1/input/input.txt gs://w6h1/output
Job [c408d371e38740a7a7e7a98fcb985812] submitted.
Waiting for job output...
23/06/27 04:54:54 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/06/27 04:54:54 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/06/27 04:54:55 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/06/27 04:54:55 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
23/06/27 04:54:55 INFO org.sparkproject.jetty.util.log: Logging initialized @2825ms to org.sparkproject.jetty.util.log.Slf4jLog
23/06/27 04:54:55 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_372-b07
23/06/27 04:54:55 INFO org.sparkproject.jetty.server.Server: Started @2919ms
23/06/27 04:54:55 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@2f588b80{HTTP/1.1, (http/1.1)}{0.0.0.0:40317}
23/06/27 04:54:55 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at w6h1-m/10.138.0.9:8032
23/06/27 04:54:56 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at w6h1-m/10.138.0.9:10200
23/06/27 04:54:57 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
23/06/27 04:54:57 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/06/27 04:54:57 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1687838987543_0007
23/06/27 04:54:58 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at w6h1-m/10.138.0.9:8030
23/06/27 04:55:01 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists
 with desired state.
Points inside the circle: 37
Points outside the circle: 163
Pi is approximately: 0.74
23/06/27 04:55:12 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@2f588b80{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [c408d371e38740a7a7e7a98fcb985812] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west1-66473074362-bo3oyqoj/google-cloud-dataproc-metainfo/3bb07439-774d-4fc1-a754-c1d2bf261087/jobs/c408d371e38740a7a7e7a98fcb985812/
driverOutputResourceUri: gs://dataproc-staging-us-west1-66473074362-bo3oyqoj/google-cloud-dataproc-metainfo/3bb07439-774d-4fc1-a754-c1d2bf261087/jobs/c408d371e38740a7a7e7a98fcb985812/driveroutput
jobUuid: a12bec22-ff86-34ef-84ea-75db6a270642
placement:
  clusterName: w6h1
  clusterUuid: 3bb07439-774d-4fc1-a754-c1d2bf261087
pysparkJob:
  args:
  - gs://w6h1/input/input.txt
  - gs://w6h1/output
  mainPythonFileUri: gs://dataproc-staging-us-west1-66473074362-bo3oyqoj/google-cloud-dataproc-metainfo/3bb07439-774d-4fc1-a754-c1d2bf261087/jobs/c408d371e38740a7a7e7a98fcb985812/staging/pi.py
reference:
  jobId: c408d371e38740a7a7e7a98fcb985812
  projectId: cs570jf
status:
  state: DONE
  stateStartTime: '2023-06-27T04:55:16.662221Z'
statusHistory:
- state: PENDING
  stateStartTime: '2023-06-27T04:54:51.461208Z'
- state: SETUP_DONE
  stateStartTime: '2023-06-27T04:54:51.490776Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2023-06-27T04:54:51.673991Z'
yarnApplications:
- name: PiEstimation
  progress: 1.0
  state: FINISHED
  trackingUrl: http://w6h1-m:8088/proxy/application_1687838987543_0007/
jfang757@cloudshell:~ (cs570jf)$
```

```
Points inside the circle: 37
Points outside the circle: 163
Pi is approximately: 0.74
```