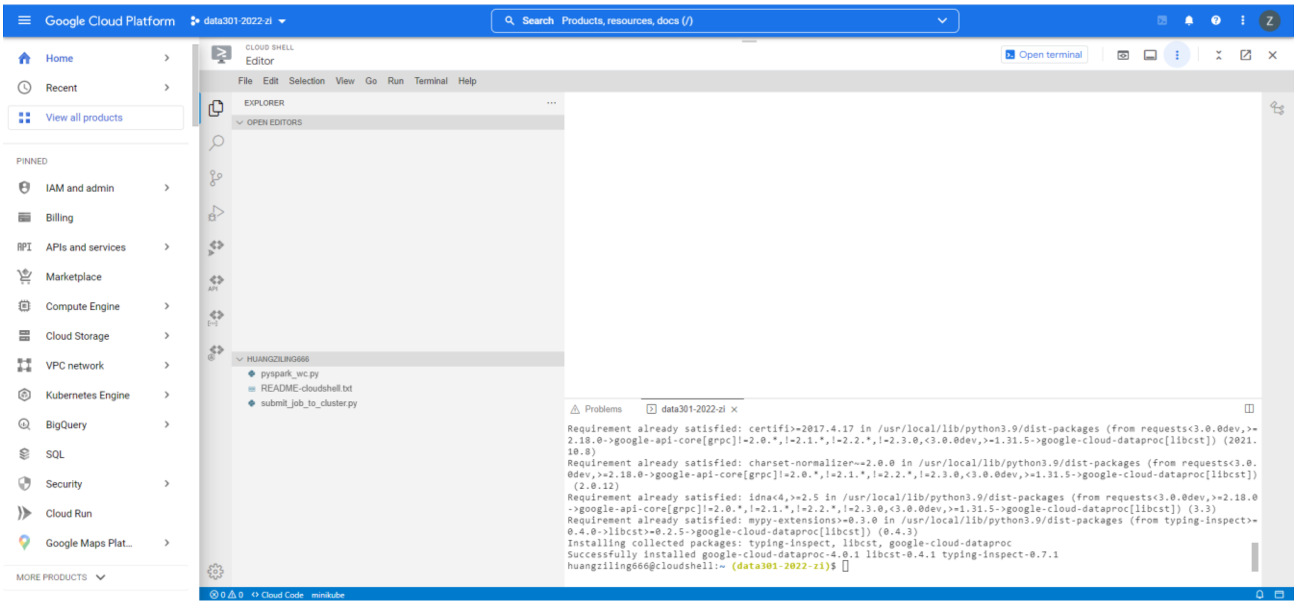
Name: Ziling Huang

Username: zhu51

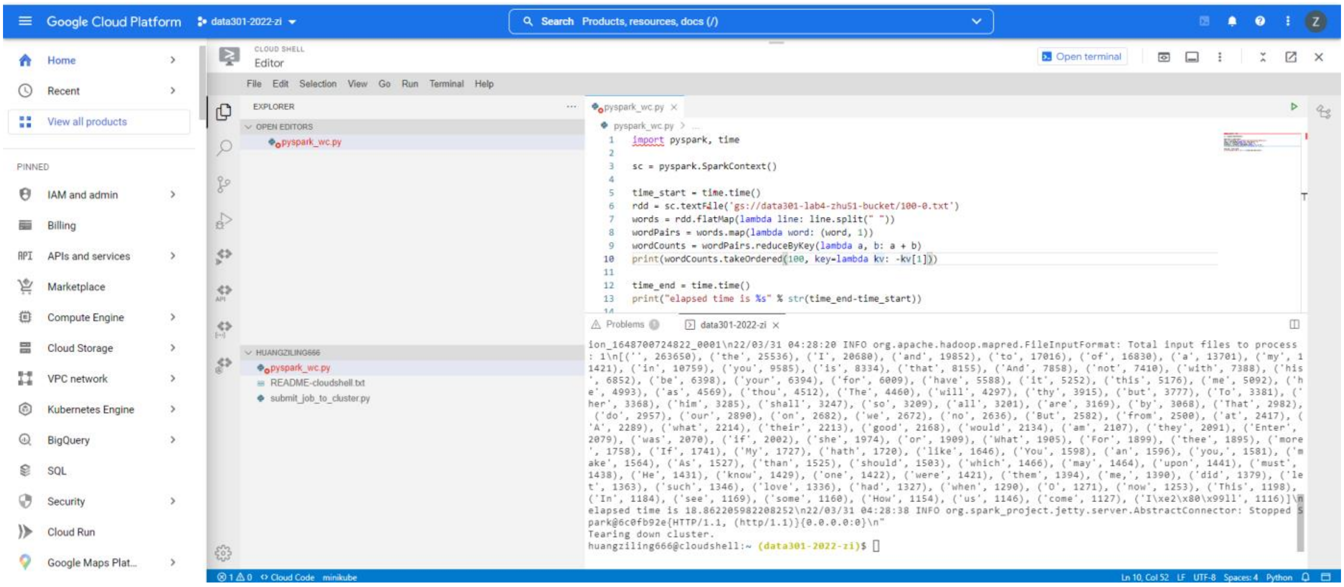
Part2



```
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0dev,>=
2.18.0->google-api-core[grpc]!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0dev,>=1.31.5->google-cloud-dataproc[libcst]) (2021.
10.8)
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.
0dev,>=2.18.0->google-api-core[grpc]!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0dev,>=1.31.5->google-cloud-dataproc[libcst])
 (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0dev,>=2.18.0
->google-api-core[grpc]!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0dev,>=1.31.5->google-cloud-dataproc[libcst]) (3.3)
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.9/dist-packages (from typing-inspect>=
0.4.0->libcst>=0.2.5->google-cloud-dataproc[libcst]) (0.4.3)
Installing collected packages: typing-inspect, libcst, google-cloud-dataproc
Successfully installed google-cloud-dataproc-4.0.1 libcst-0.4.1 typing-inspect-0.7.1
huangziling666@cloudshell:~ (data301-2022-zi)$ 
```

Part3

REGION=australia-southeast1
ZONE=australia-southeast1-a
PROJECT=**data301-2022-zi**
CLUSTER=data301-lab4-**zhu51**-cluster
BUCKET=data301-lab4-**zhu51**-bucket
When 2 cores:



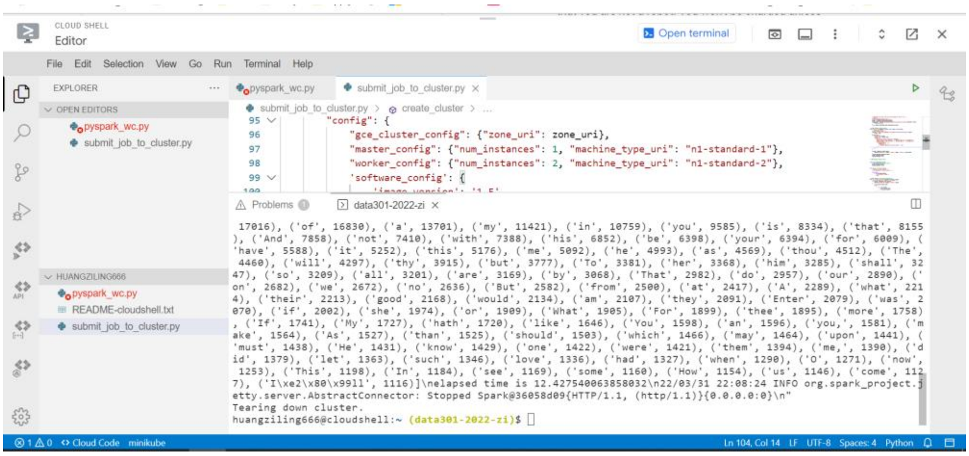
```python
import pyspark, time

sc = pyspark.SparkContext()

time_start = time.time()
rdd = sc.textFile('gs://data301-lab4-zhu51-bucket/100-0.txt')
words = rdd.flatMap(lambda line: line.split(" "))
wordPairs = words.map(lambda word: (word, 1))
wordCounts = wordPairs.reduceByKey(lambda a, b: a + b)
print(wordCounts.takeOrdered(100, key=lambda kv: -kv[1]))

time_end = time.time()
print("elapsed time is %s" % str(time_end-time_start))
```

```
ion_1648700724822_0001\n22/03/31 04:28:20 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process
: 1\n[('', 263650), ('the', 25536), ('I', 20680), ('and', 19852), ('to', 17016), ('of', 16830), ('a', 13701), ('my', 1
1421), ('in', 10759), ('you', 9585), ('is', 8334), ('that', 8155), ('And', 7858), ('not', 7410), ('with', 7388), ('his
', 6852), ('be', 6398), ('your', 6394), ('for', 6009), ('have', 5588), ('it', 5252), ('this', 5176), ('me', 5092), ('h
e', 4993), ('as', 4569), ('thou', 4512), ('The', 4460), ('will', 4297), ('thy', 3915), ('but', 3777), ('To', 3381), ('
her', 3368), ('him', 3285), ('shall', 3247), ('so', 3209), ('all', 3201), ('are', 3169), ('by', 3068), ('That', 2982),
 ('do', 2957), ('our', 2890), ('on', 2682), ('we', 2672), ('no', 2636), ('But', 2582), ('from', 2500), ('at', 2417), (
'A', 2289), ('what', 2214), ('their', 2213), ('good', 2168), ('would', 2134), ('am', 2107), ('they', 2091), ('Enter',
2079), ('was', 2070), ('if', 2002), ('she', 1974), ('or', 1909), ('What', 1905), ('For', 1899), ('thee', 1895), ('more
', 1758), ('If', 1741), ('My', 1727), ('hath', 1720), ('like', 1646), ('You', 1598), ('an', 1596), ('you,', 1581), ('m
ake', 1564), ('As', 1527), ('than', 1525), ('should', 1503), ('which', 1466), ('may', 1464), ('upon', 1441), ('must',
1438), ('He', 1431), ('know', 1429), ('one', 1422), ('were', 1421), ('them', 1394), ('me,', 1390), ('did', 1379), ('le
t', 1363), ('such', 1346), ('love', 1336), ('had', 1327), ('when', 1290), ('O', 1271), ('now', 1253), ('This', 1198),
('In', 1184), ('see', 1169), ('some', 1160), ('How', 1154), ('us', 1146), ('come', 1127), ('I\xe2\x80\x99ll', 1116)]\n
elapsed time is 18.862205982208252\n22/03/31 04:28:38 INFO org.spark_project.jetty.server.AbstractConnector: Stopped S
park@6c0fb92e{HTTP/1.1, (http/1.1)}{0.0.0.0:0}\n"
Tearing down cluster.
huangziling666@cloudshell:~ (data301-2022-zi)$
```

Part4

4b)

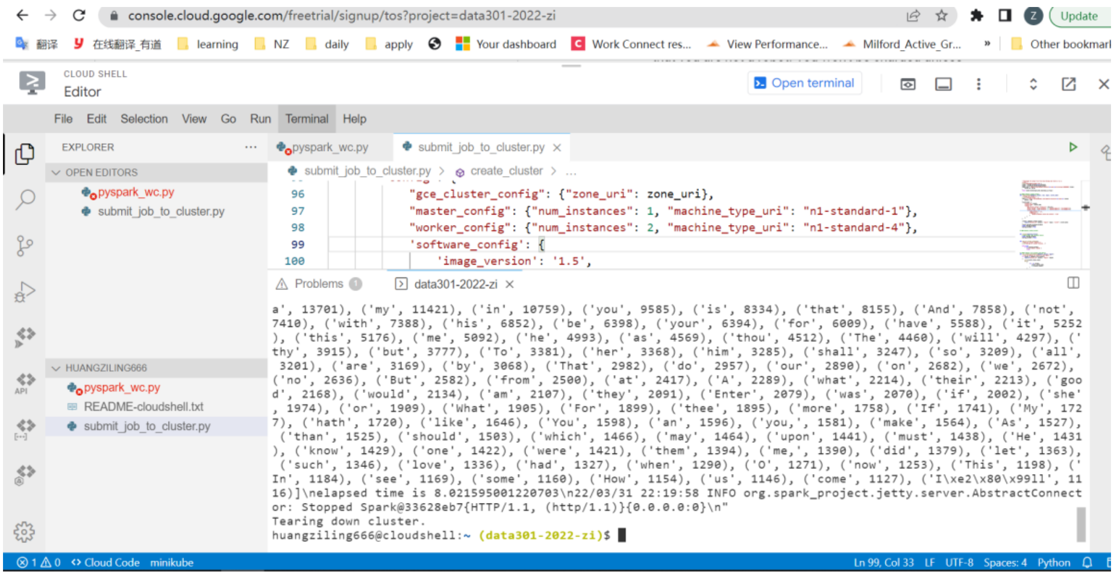
4 cores:



```python
        "config": {
            "gce_cluster_config": {"zone_uri": zone_uri},
            "master_config": {"num_instances": 1, "machine_type_uri": "n1-standard-1"},
            "worker_config": {"num_instances": 2, "machine_type_uri": "n1-standard-2"},
            'software_config': {
```

```
 17016), ('of', 16830), ('a', 13701), ('my', 11421), ('in', 10759), ('you', 9585), ('is', 8334), ('that', 8155
), ('And', 7858), ('not', 7410), ('with', 7388), ('his', 6852), ('be', 6398), ('your', 6394), ('for', 6009), (
'have', 5588), ('it', 5252), ('this', 5176), ('me', 5092), ('he', 4993), ('as', 4569), ('thou', 4512), ('The',
 4460), ('will', 4297), ('thy', 3915), ('but', 3777), ('To', 3381), ('her', 3368), ('him', 3285), ('shall', 32
47), ('so', 3209), ('all', 3201), ('are', 3169), ('by', 3068), ('That', 2982), ('do', 2957), ('our', 2890), ('
on', 2682), ('we', 2672), ('no', 2636), ('But', 2582), ('from', 2500), ('at', 2417), ('A', 2289), ('what', 221
4), ('their', 2213), ('good', 2168), ('would', 2134), ('am', 2107), ('they', 2091), ('Enter', 2079), ('was', 2
070), ('if', 2002), ('she', 1974), ('or', 1909), ('What', 1905), ('For', 1899), ('thee', 1895), ('more', 1758)
, ('If', 1741), ('My', 1727), ('hath', 1720), ('like', 1646), ('You', 1598), ('an', 1596), ('you,', 1581), ('m
ake', 1564), ('As', 1527), ('than', 1525), ('should', 1503), ('which', 1466), ('may', 1464), ('upon', 1441), (
'must', 1438), ('He', 1431), ('know', 1429), ('one', 1422), ('were', 1421), ('them', 1394), ('me,', 1390), ('d
id', 1379), ('let', 1363), ('such', 1346), ('love', 1336), ('had', 1327), ('when', 1290), ('O', 1271), ('now',
 1253), ('This', 1198), ('In', 1184), ('see', 1169), ('some', 1160), ('How', 1154), ('us', 1146), ('come', 112
7), ('I\xe2\x80\x99ll', 1116)]\nelapsed time is 12.427540063858032\n22/03/31 22:08:24 INFO org.spark_project.j
etty.server.AbstractConnector: Stopped Spark@36058d09{HTTP/1.1, (http/1.1)}{0.0.0.0:0}\n"
Tearing down cluster.
huangziling666@cloudshell:~ (data301-2022-zi)$
```

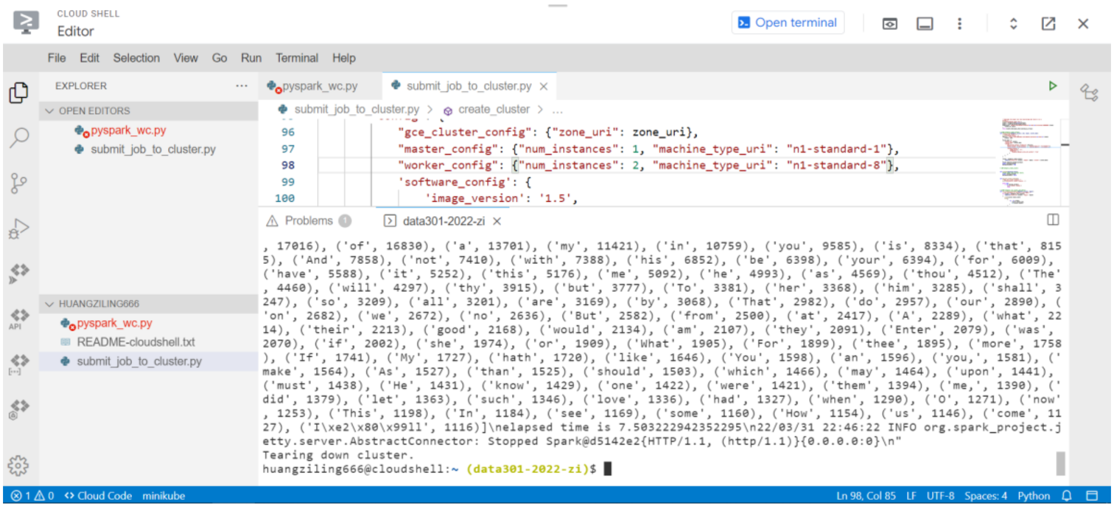
8 cores:



```python
            "gce_cluster_config": {"zone_uri": zone_uri},
            "master_config": {"num_instances": 1, "machine_type_uri": "n1-standard-1"},
            "worker_config": {"num_instances": 2, "machine_type_uri": "n1-standard-4"},
            'software_config': {
                'image_version': '1.5',
```

```
a', 13701), ('my', 11421), ('in', 10759), ('you', 9585), ('is', 8334), ('that', 8155), ('And', 7858), ('not',
7410), ('with', 7388), ('his', 6852), ('be', 6398), ('your', 6394), ('for', 6009), ('have', 5588), ('it', 5252
), ('this', 5176), ('me', 5092), ('he', 4993), ('as', 4569), ('thou', 4512), ('The', 4460), ('will', 4297), ('
thy', 3915), ('but', 3777), ('To', 3381), ('her', 3368), ('him', 3285), ('shall', 3247), ('so', 3209), ('all',
 3201), ('are', 3169), ('by', 3068), ('That', 2982), ('do', 2957), ('our', 2890), ('on', 2682), ('we', 2672),
('no', 2636), ('But', 2582), ('from', 2500), ('at', 2417), ('A', 2289), ('what', 2214), ('their', 2213), ('goo
d', 2168), ('would', 2134), ('am', 2107), ('they', 2091), ('Enter', 2079), ('was', 2070), ('if', 2002), ('she'
, 1974), ('or', 1909), ('What', 1905), ('For', 1899), ('thee', 1895), ('more', 1758), ('If', 1741), ('My', 172
7), ('hath', 1720), ('like', 1646), ('You', 1598), ('an', 1596), ('you,', 1581), ('make', 1564), ('As', 1527),
 ('than', 1525), ('should', 1503), ('which', 1466), ('may', 1464), ('upon', 1441), ('must', 1438), ('He', 1431
), ('know', 1429), ('one', 1422), ('were', 1421), ('them', 1394), ('me,', 1390), ('did', 1379), ('let', 1363),
 ('such', 1346), ('love', 1336), ('had', 1327), ('when', 1290), ('O', 1271), ('now', 1253), ('This', 1198), ('
In', 1184), ('see', 1169), ('some', 1160), ('How', 1154), ('us', 1146), ('come', 1127), ('I\xe2\x80\x99ll', 11
16)]\nelapsed time is 8.021595001220703\n22/03/31 22:19:58 INFO org.spark_project.jetty.server.AbstractConnect
or: Stopped Spark@33628eb7{HTTP/1.1, (http/1.1)}{0.0.0.0:0}\n"
Tearing down cluster.
huangziling666@cloudshell:~ (data301-2022-zi)$
```

16 cores:

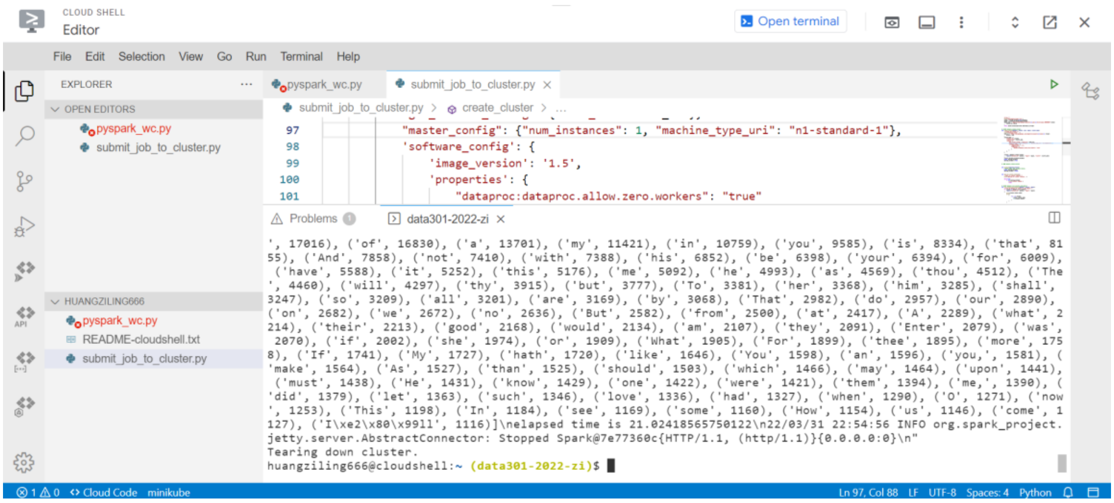


```
            "gce_cluster_config": {"zone_uri": zone_uri},
            "master_config": {"num_instances": 1, "machine_type_uri": "n1-standard-1"},
            "worker_config": {"num_instances": 2, "machine_type_uri": "n1-standard-8"},
            'software_config': {
                'image_version': '1.5',
```

```
, 17016), ('of', 16830), ('a', 13701), ('my', 11421), ('in', 10759), ('you', 9585), ('is', 8334), ('that', 815
5), ('And', 7858), ('not', 7410), ('with', 7388), ('his', 6852), ('be', 6398), ('your', 6394), ('for', 6009),
('have', 5588), ('it', 5252), ('this', 5176), ('me', 5092), ('he', 4993), ('as', 4569), ('thou', 4512), ('The'
, 4460), ('will', 4297), ('thy', 3915), ('but', 3777), ('To', 3381), ('her', 3368), ('him', 3285), ('shall', 3
247), ('so', 3209), ('all', 3201), ('are', 3169), ('by', 3068), ('That', 2982), ('do', 2957), ('our', 2890), (
'on', 2682), ('we', 2672), ('no', 2636), ('But', 2582), ('from', 2500), ('at', 2417), ('A', 2289), ('what', 22
14), ('their', 2213), ('good', 2168), ('would', 2134), ('am', 2107), ('they', 2091), ('Enter', 2079), ('was',
2070), ('if', 2002), ('she', 1974), ('or', 1909), ('What', 1905), ('For', 1899), ('thee', 1895), ('more', 1758
), ('If', 1741), ('My', 1727), ('hath', 1720), ('like', 1646), ('You', 1598), ('an', 1596), ('you,', 1581), ('
make', 1564), ('As', 1527), ('than', 1525), ('should', 1503), ('which', 1466), ('may', 1464), ('upon', 1441),
('must', 1438), ('He', 1431), ('know', 1429), ('one', 1422), ('were', 1421), ('them', 1394), ('me,', 1390), ('
did', 1379), ('let', 1363), ('such', 1346), ('love', 1336), ('had', 1327), ('when', 1290), ('O', 1271), ('now'
, 1253), ('This', 1198), ('In', 1184), ('see', 1169), ('some', 1160), ('How', 1154), ('us', 1146), ('come', 11
27), ('I\xe2\x80\x99ll', 1116)]\nelapsed time is 7.503222942352295\n22/03/31 22:46:22 INFO org.spark_project.j
etty.server.AbstractConnector: Stopped Spark@d5142e2{HTTP/1.1, (http/1.1)}{0.0.0.0:0}\n"
Tearing down cluster.
huangziling666@cloudshell:~ (data301-2022-zi)$
```

4c)

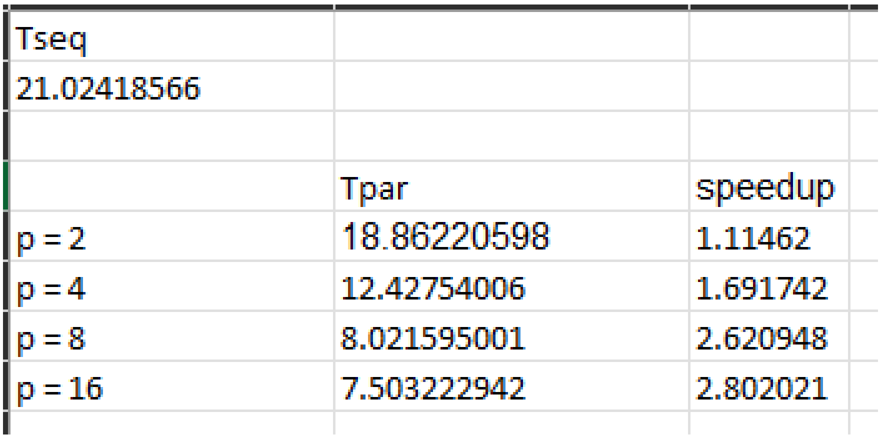
One core:



```
            "master_config": {"num_instances": 1, "machine_type_uri": "n1-standard-1"},
            'software_config': {
                'image_version': '1.5',
                'properties': {
                    "dataproc:dataproc.allow.zero.workers": "true"
```

```
', 17016), ('of', 16830), ('a', 13701), ('my', 11421), ('in', 10759), ('you', 9585), ('is', 8334), ('that', 81
55), ('And', 7858), ('not', 7410), ('with', 7388), ('his', 6852), ('be', 6398), ('your', 6394), ('for', 6009),
 ('have', 5588), ('it', 5252), ('this', 5176), ('me', 5092), ('he', 4993), ('as', 4569), ('thou', 4512), ('The
', 4460), ('will', 4297), ('thy', 3915), ('but', 3777), ('To', 3381), ('her', 3368), ('him', 3285), ('shall',
3247), ('so', 3209), ('all', 3201), ('are', 3169), ('by', 3068), ('That', 2982), ('do', 2957), ('our', 2890),
('on', 2682), ('we', 2672), ('no', 2636), ('But', 2582), ('from', 2500), ('at', 2417), ('A', 2289), ('what', 2
214), ('their', 2213), ('good', 2168), ('would', 2134), ('am', 2107), ('they', 2091), ('Enter', 2079), ('was',
 2070), ('if', 2002), ('she', 1974), ('or', 1909), ('What', 1905), ('For', 1899), ('thee', 1895), ('more', 175
8), ('If', 1741), ('My', 1727), ('hath', 1720), ('like', 1646), ('You', 1598), ('an', 1596), ('you,', 1581), (
'make', 1564), ('As', 1527), ('than', 1525), ('should', 1503), ('which', 1466), ('may', 1464), ('upon', 1441),
 ('must', 1438), ('He', 1431), ('know', 1429), ('one', 1422), ('were', 1421), ('them', 1394), ('me,', 1390), (
'did', 1379), ('let', 1363), ('such', 1346), ('love', 1336), ('had', 1327), ('when', 1290), ('O', 1271), ('now
', 1253), ('This', 1198), ('In', 1184), ('see', 1169), ('some', 1160), ('How', 1154), ('us', 1146), ('come', 1
127), ('I\xe2\x80\x99ll', 1116)]\nelapsed time is 21.02418565750122\n22/03/31 22:54:56 INFO org.spark_project.
jetty.server.AbstractConnector: Stopped Spark@7e77360c{HTTP/1.1, (http/1.1)}{0.0.0.0:0}\n"
Tearing down cluster.
huangziling666@cloudshell:~ (data301-2022-zi)$
```

4d)

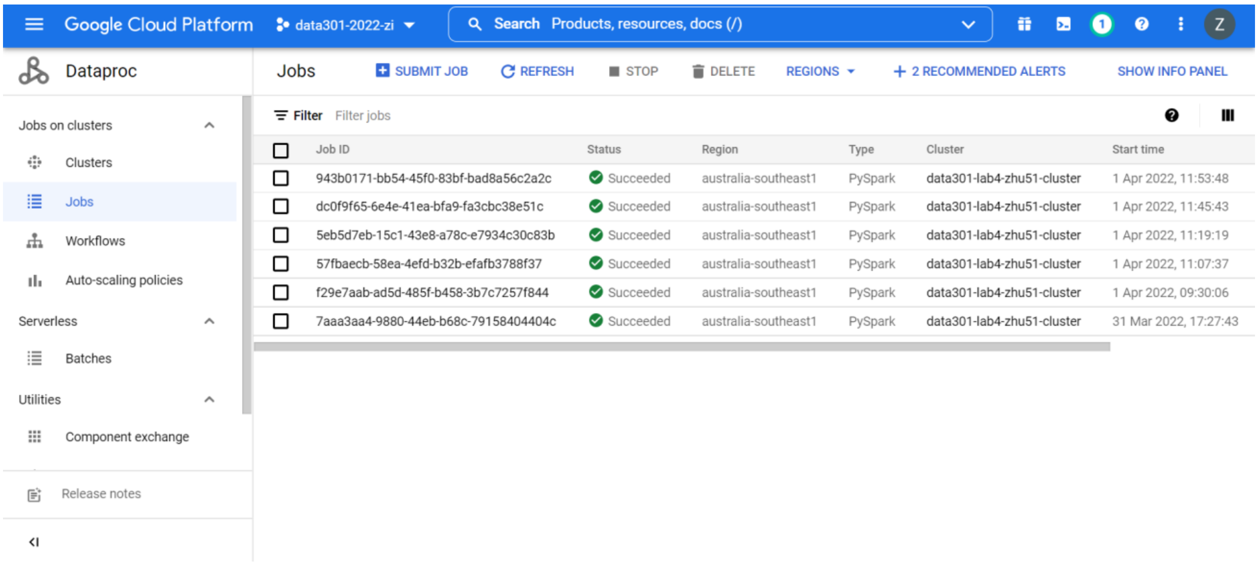


Tseq		
21.02418566		
	Tpar	speedup
p = 2	18.86220598	1.11462
p = 4	12.42754006	1.691742
p = 8	8.021595001	2.620948
p = 16	7.503222942	2.802021

4e)

As we add more processors, the processing time is decreasing as well as the speedup time is increasing, which means the more processors we use, the less time we spend. However, if we the add processors more than 16, the speedup time and the processing time will be the same since the ability has already reached a limit so more processors won’t increase the efficiency anymore.

4f)



□	Job ID	Status	Region	Type	Cluster	Start time
□	943b0171-bb54-45f0-83bf-bad8a56c2a2c	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	1 Apr 2022, 11:53:48
□	dc0f9f65-6e4e-41ea-bfa9-fa3cbc38e51c	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	1 Apr 2022, 11:45:43
□	5eb5d7eb-15c1-43e8-a78c-e7934c30c83b	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	1 Apr 2022, 11:19:19
□	57fbaecb-58ea-4efd-b32b-efafb3788f37	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	1 Apr 2022, 11:07:37
□	f29e7aab-ad5d-485f-b458-3b7c7257f844	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	1 Apr 2022, 09:30:06
□	7aaa3aa4-9880-44eb-b68c-79158404404c	✓ Succeeded	australia-southeast1	PySpark	data301-lab4-zhu51-cluster	31 Mar 2022, 17:27:43