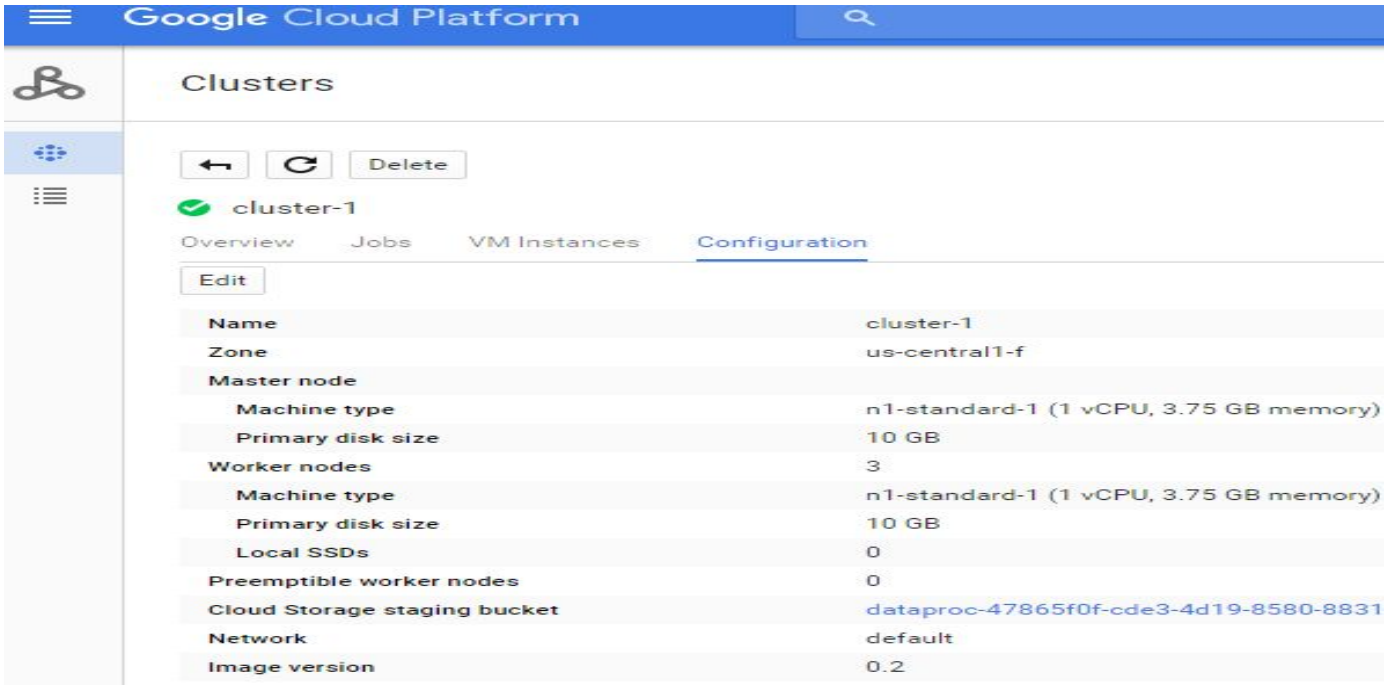
Results :

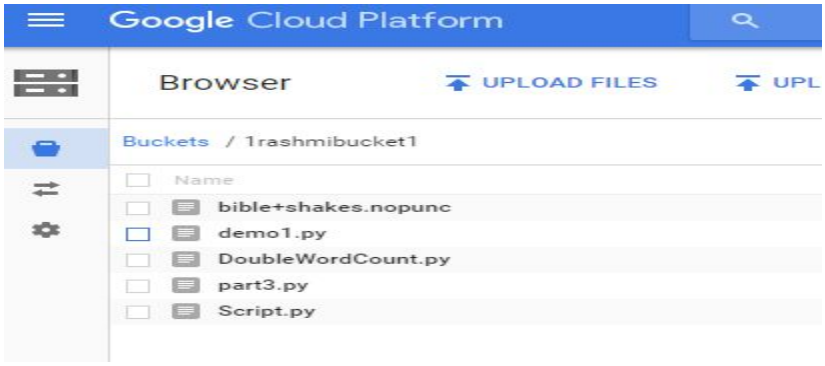
Initial Configuration:

- Programming Language : Python
- Execution Environment : Google Cloud
- Programming Framework : Apache Spark

Cluster Creation:



Bucket Creation:



1)Assignment Part 1: One Word Frequency:

The python code gives no of times each word has appeared in the input file. I have sorted the output with key(word in our case.)

File Output :

No of Files Generated : 14

Size of each File : 60KB except last one

Output is sorted wrt the word

```
, (u'longer', 136), (u'longermarried', 1), (u'longest', 6), (u'longeth', 6), (u'longexp
', 2), (u'loosen', 1), (u'looseth', 2), (u'loosewived', 1), (u'loosing', 5), (u'lop', 5
1), (u"lour'st", 1), (u'loureth', 1), (u'louring', 4), (u'lours', 1), (u'louse', 2), (u
ings', 1), (u'lovest', 50), (u'lovesuit', 3), (u'loveth', 68), (u'lovethoughts', 1), (u
cina', 2), (u'lucio', 135), (u"lucio's", 1), (u'lucius', 250), (u'luck', 26), (u'luckie
(u'lusted', 4), (u'lusteth', 6), (u'lustful', 9), (u'lustier', 3), (u'lustiest', 1), (u
3), (u'lydia', 5), (u'lydians', 1), (u'lying', 88), (u'lyingest', 2), (u'lym', 1), (u'l
6), (u'mackerel', 1), (u'macmorris', 12), (u'maculate', 1), (u'maculation', 1), (u'mad'
gnificencein', 1), (u'magnificent', 2), (u'magnifico', 1), (u'magnificoes', 3), (u'magn
u'mailed', 1), (u'mails', 1), (u'maim', 4), (u"maim'd", 4), (u'maimed', 8), (u'maims',
cham', 2), (u'malchiah', 9), (u'malchiel', 3), (u'malchielites', 1), (u'malchijah', 6),
, (u'mammocked', 1), (u'mammon', 4), (u'mamre', 10), (u'man', 4473), (u"man'", 6), (u"m
nna', 20), (u'manned', 2), (u'manner', 278), (u"manner'd", 2), (u'mannerit', 1), (u'man
bleconstant', 1), (u'marbled', 1), (u'marblehearted', 1), (u'marcellus', 48), (u'march'
, 1), (u'marketmen', 1), (u'marketplace', 25), (u'marketplaces', 1), (u'marketprice', 1
hall'st", 1), (u'marshalsea', 1), (u'marshalship', 1), (u'mart', 18), (u'marted', 1), (
, (u"master'd", 5), (u"master's", 89), (u"master'smate", 3), (u'masterbuilder', 1), (u'
1), (u'mattering', 1), (u'matternurse', 1), (u'matters', 49), (u'matterwear', 1), (u'ma
(u'mealy', 1), (u'mean', 320), (u"mean'st", 7), (u"meanapparell'd", 1), (u'meanborn', 2
u'medice', 1), (u'medicinable', 3), (u'medicinal', 2), (u'medicine', 28), (u'medicines'
1), (u'melchi', 2), (u'melchiah', 1), (u'melchisedec', 9), (u'melchishua', 2), (u'melch
), (u'menon', 1), (u'menpleasers', 2), (u'menservants', 10), (u'menstealers', 1), (u'me
, (u'mere', 57), (u'mered', 2), (u'merely', 28), (u'meremoth', 6), (u'meres', 1), (u'me
```

Output Link :

https://storage.googleapis.com/dataproc-47865f0f-cde3-4d19-8580-883107cf6808-us/google-cloud-dataproc-metainfo/c737f135-9db1-4625-9498-19cc96e16a18/jobs/232447a6-8f8c-4bee-bbc2-da360f094ece/driveroutput.000000003

2)Assignment Part 2 : Bigram Frequency

For solving this problem, i made use of glom(), join() and split() function primarily. The underlying idea of using glom() is getting an entire list rather than getting input ,line by line.I performed split operation afterwards and concatenated two adjacent words as one. Afterwards i transformed it into a pair RDD then did reduceByKey.

Bucket and Cluster Output:

6d7fbc47-6cf9-490d-8f90-d1d8e3003811

Start time: Feb 4, 2016, 5:50:24 PM Elapsed time: 50 sec Status: Succeeded

Output Configuration

Line wrapping Equivalent command line

```
16/02/04 22:50:32 INFO akka.event.slf4j.Slf4jLogger: Slf4jLogger started
16/02/04 22:50:32 INFO Remoting: Starting remoting
16/02/04 22:50:32 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@10.240.0.5:33990]
16/02/04 22:50:33 INFO org.spark-project.jetty.server.Server: jetty-8.y.z-SNAPSHOT
16/02/04 22:50:33 INFO org.spark-project.jetty.server.AbstractConnector: Started SocketConnector@0.0.0.0:56097
16/02/04 22:50:33 INFO org.spark-project.jetty.server.Server: jetty-8.y.z-SNAPSHOT
16/02/04 22:50:33 INFO org.spark-project.jetty.server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/02/04 22:50:33 WARN org.apache.spark.metrics.MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
16/02/04 22:50:33 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-1-m/10.240.0.5:8032
16/02/04 22:50:37 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1454625038807_0004
16/02/04 22:50:48 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.4.3-hadoop2
16/02/04 22:50:48 INFO org.apache.hadoop.mapred.FileInputFormat: Total input paths to process : 1
[((u'holy', u'bible'), 1), ((u'bible', u'authorized'), 1), ((u'authorized', u'king'), 1), ((u'king', u'james'), 1), ((u'james', u'version'), 1), ((u'version', u'textfile'), 1), ((u'textfile', u'890904'), 1),
16/02/04 22:51:03 INFO akka.remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/02/04 22:51:03 INFO akka.remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
```

Job output is complete

File Output:

No of Files Generated : 48

Size of each File : 60KB except last one

Output is sorted wrt the word

Output Link:

https://storage.googleapis.com/dataproc-47865f0f-cde3-4d19-8580-883107cf6808-us/google-cloud-dataproc-metainfo/c5ace7d7-ce36-4939-a7fc-855eecc49620/jobs/dfbbfd73-4522-4abf-ae4e-33c06a1f0362/driveroutput.000000005

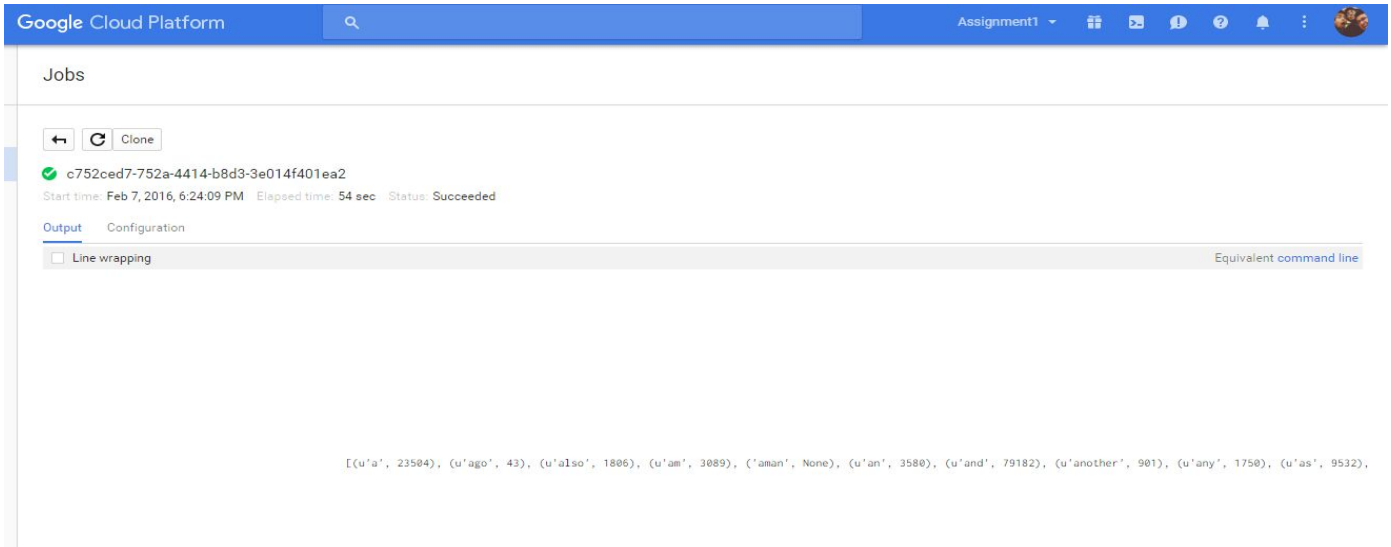
```
(27, (u'naked', u'and')), (27, (u'slew', u'them')), (27, (u'says',
u'he')), (27, (u'say', u'ye')), (27, (u'ends', u'well')), (27, (u'left',
u'and')), (27, (u'worth', u'the')), (27, (u'say', u'if')), (27,
(u'buckingham', u'and')), (27, (u'to', u'understand')), (27, (u'which',
u'being')), (27, (u'but', u'o')), (27, (u'you', u'exit')), (27, (u'the',
u'chain')), (27, (u'of', u'trouble')), (27, (u'show', u'the')), (27,
(u'the', u'portion')), (27, (u'good', u'friends')), (27, (u'valour',
u'and')), (27, (u'itself', u'and')), (27, (u'of', u'three')), (27,
(u'his', u'kind')), (27, (u'cried', u'with')), (27, (u'stand', u'to')),
(27, (u'and', u'bound')), (27, (u'the', u'increase')), (27, (u'who',
u'have')), (27, (u'they', u'buried')), (27, (u'slew', u'the')), (27,
(u'how', u'say')), (27, (u'out', u'against')), (27, (u'your', u'lord')),
(27, (u'heart', u'be')), (27, (u'word', u'which')), (27, (u'he',
u'knoweth')), (27, (u'when', u'there')), (27, (u'shall', u'choose')),
(27, (u'ye', u'in')), (27, (u'not', u'they')), (27, (u'his', u'part')),
(27, (u'how', u'fares')), (27, (u'talents', u'of')), (27, (u'care',
u'of')), (27, (u'laugh', u'at')), (27, (u'else', u'i')), (27, (u'so',
u'should')), (27, (u'will', u'lay')), (27, (u'than', u'for')), (27,
(u'or', u'four')), (27, (u'john', u'the')), (27, (u'wherefore', u'the')),
(27, (u'for', u'thine')), (27, (u'reign', u'of')), (27, (u'bold',
u'to')), (27, (u'hath', u'left')), (27, (u'your', u'generations')), (27,
(u'side', u'the')), (27, (u'man', u'by')), (27, (u'the', u'measure')),
(27, (u'be', u'thought')), (27, (u'v', u'act')), (27, (u'send',
u'thee')), (27, (u'unto', u'abraham')), (27, (u'war', u'with')), (27,
(u'of', u'glory')), (27, (u'when', u'thy')), (27, (u'to', u'woo')), (27,
(u'own', u'part')), (27, (u'again', u'but')), (27, (u'that', u'these')),
(27, (u'place', u'in')), (27, (u'name', u'shall')), (27, (u'be', u'if')),
(27, (u'of', u'spirit')), (27, (u'things', u'but')), (27, (u'is',
u'some')), (27, (u'his', u'course')), (27, (u'about', u'with')), (27,
(u'my', u'best')), (27, (u'good', u'madam')), (27, (u'eternal',
u'life')), (27, (u'thy', u'judgments')), (27, (u'their', u'blood')), (27,
(u'does', u'he')), (27, (u'not', u'return')), (27, (u'a', u'priest')),
```

3)Assignment Part 3 : One Word frequency in Another dataset

In order to solve this problem, i made use of pair RDDs. Two pair RDDs then joined using right Outer Join. The key for join operation was the word from smaller list which we wanted to search in the bigger list.

Bucket and cluster Output:

https://storage.googleapis.com/dataproc-47865f0f-cde3-4d19-8580-883107cf6808-us/google-cloud-dataproc-metainfo/c5ace7d7-ce36-4939-a7fc-855eecc49620/jobs/9aaf7f73-c681-4887-aa9e-0e5ce799cf40/driveroutput.000000000



File Output:

```
1.4.3-hadoop2
16/02/08 00:27:40 INFO org.apache.hadoop.mapred.FileInputFormat: Total
input paths to process : 1
□[Stage 1:>                                                          (0 +
0) / 2]□[Stage 0:>                    (0 + 0) / 2][Stage 1:>
(0 + 0) / 2]□[Stage 0:>                    (0 + 1) / 2][Stage 1:>
(0 + 0) / 2]□[Stage 0:=========>           (1 + 1) / 2][Stage 1:>
(0 + 1) / 2]□[Stage 1:>
(0 + 1) / 2]□[Stage 1:>
(0 + 2) / 2]□[Stage 1:==============================>
(1 + 1) / 2]□[Stage 2:==============================>
(1 + 1) / 2]□
□[(u'four', 478), (u'unto', 9458), (u'for', 16941), (u'intent', 64),
(u'bright', 111), (u'come', 4562), (u'went', 1504), (u'with', 14263),
(u'am', 3089), (u'company', 293), (u'ago', 43), (u'came', 2452),
(u'sent', 954), (u'this', 9680), (u'you', 16603), (u'behold', 1499),
(u'them', 8507), (u'another', 901), (u'what', 5630), (u'the', 93739),
(u'together', 754), (u'clothing', 20), (u'a', 23504), (u'me', 12123),
(u'ask', 278), (u'ye', 4289), (u'unlawful', 13), (u'is', 16529),
(u'gainsaying', 4), (u'stood', 447), (u'hour', 402), (u'have', 9990),
(u'shewed', 135), (u'as', 9532), (u'but', 10571), (u'my', 17312),
(u'was', 6882), (u'and', 79182), (u'myself', 736), (u'ninth', 40),
(u'call', 782), (u'i', 30240), (u'not', 15440), (u'therefore', 1888),
(u'how', 2801), (u'soon', 223), (u'until', 445), (u'in', 24350),
(u'nation', 175), (u'up', 3505), (u'know', 2476), (u'an', 3580), (u'at',
4207), (u'before', 2655), (u'found', 636), (u'jew', 99), (u'also', 1806),
(u'saying', 1504), (u'any', 1750), (u'prayed', 67), (u'should', 2437),
(u'that', 24407), (u'fasting', 30), (u'one', 3872), (u'keep', 867),
(u'peter', 250), (u'he', 17087), (u'said', 4405), (u'it', 14141),
(u'were', 4432), (u'stand', 850), (u'house', 2550), (u'him', 12069),
(u'common', 177), (u'to', 33929), (u'hath', 4291), (u'talked', 50),
(u'god', 5229), (u'days', 1069), (u'took', 917), (u'cornelius', 33),
```

Output Link: