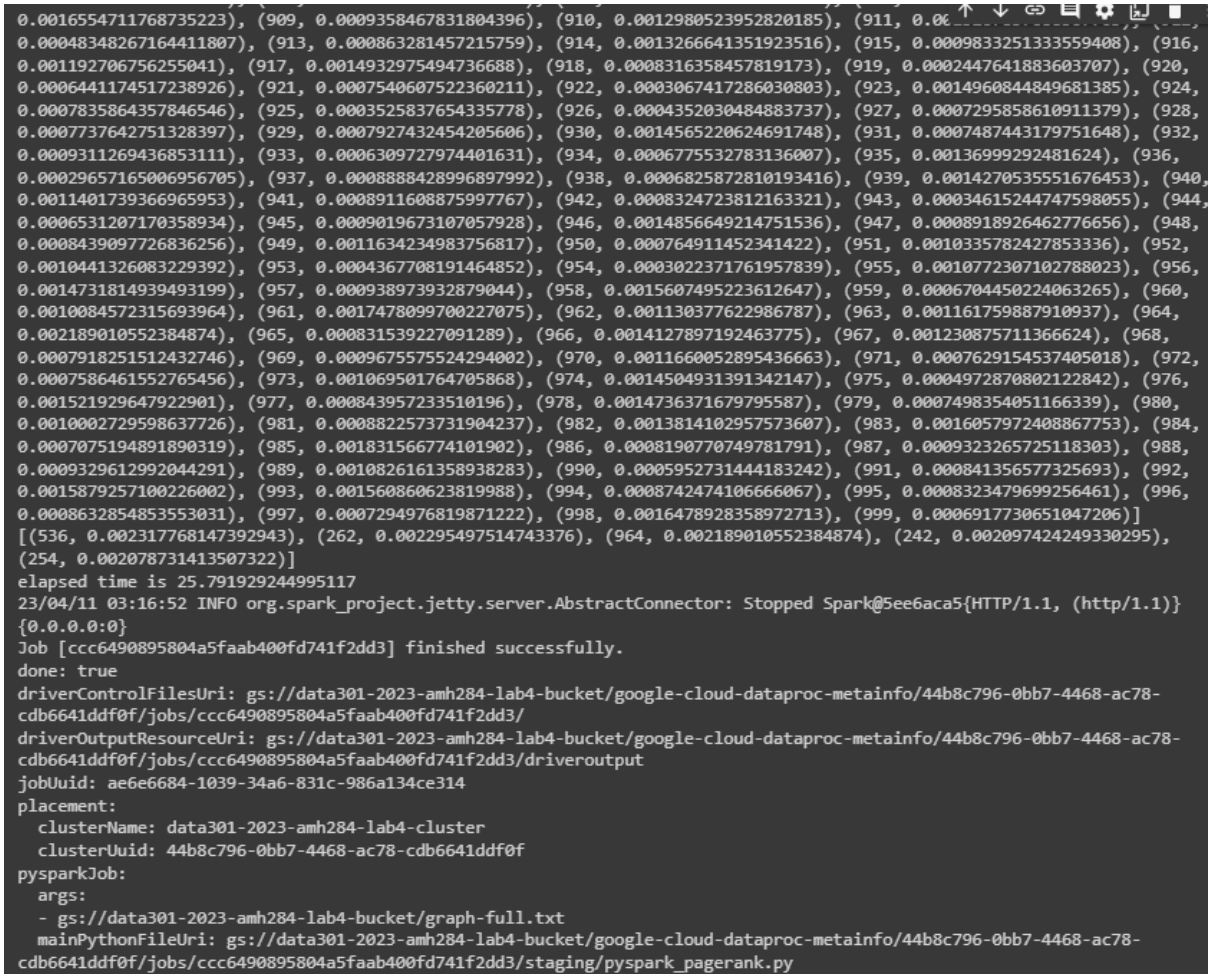
DATA301 – Lab 4 writeup

1 vCPU – graph-small.txt:

```
23/04/11 02:21:36 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application
application_1681179580290_0001
23/04/11 02:21:58 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
R: [(0, 0.035400591802442975), (1, 0.005171063533189885), (2, 0.007109590714645795), (3, 0.00699484044336462), (4,
0.016446111939280417), (5, 0.005496844981748572), (6, 0.0026390553476102517), (7, 0.005708505652182356), (8,
0.012862068917611628), (9, 0.007283518367242062), (10, 0.00388499606145024), (11, 0.003595499121159902), (12,
0.020690343993951192), (13, 0.0393777270035732 8), (14, 0.005954469046162684), (15, 0.004211316450735478), (16,
0.00475754640762407), (17, 0.0137680884958506), (18, 0.004719333295801118), (19, 0.00363988954638501 3), (20,
0.0052539718798640875), (21, 0.01706194255432705), (22, 0.0026746831662129325), (23, 0.006079163246662509), (24,
0.006484629773168008), (25, 0.01538112682133084), (26, 0.03489656776403214), (27, 0.005823318003010782), (28,
0.006223751644838693), (29, 0.006038775174000852), (30, 0.012166055454916381), (31, 0.0035243227245196113), (32,
0.003907952132011132), (33, 0.005252714771083761), (34, 0.013281350423884777), (35, 0.004025475654532078), (36,
0.0019304793257914786), (37, 0.0061845998242534665), (38, 0.011389080370852529), (39, 0.039884267610139176), (40,
0.008390858686994008), (41, 0.012067981145007832), (42, 0.0054633673736949285), (43, 0.012766987824747692), (44,
0.002950912126713505), (45, 0.0037273275574807602), (46, 0.004758362304187359), (47, 0.032040638779145 35), (48,
0.004814487946254758), (49, 0.002525039827018134), (50, 0.007685581620598552), (51, 0.013646086643850955), (52,
0.042431404065756184), (53, 0.013361483266072945), (54, 0.006959628447489829), (55, 0.0072508368788184036), (56,
0.020051832175604214), (57, 0.003910709902156528), (58, 0.0015567954201211119), (59, 0.009226981043302595), (60,
0.023101971595224743), (61, 0.007334388525680361), (62, 0.002655182657172953), (63, 0.0065259605535566215), (64,
0.021995275657399286), (65, 0.03308454606484258), (66, 0.00804099080786488), (67, 0.0034117695767682805), (68,
0.0055621205808055445), (69, 0.009521144526070478), (70, 0.006673509313876524), (71, 0.0027966203919454714), (72,
0.007529708363041359), (73, 0.01762423490407708), (74, 0.006432471709856831), (75, 0.003461084692367296), (76,
0.006336822787608032), (77, 0.016217203651570505), (78, 0.023332947897245708), (79, 0.005262472484228285), (80,
0.001814513348120128), (81, 0.007178403497634230 6), (82, 0.013421384014995534), (83, 0.0020300000327560437), (84,
0.0012246126965946577), (85, 0.002173895635096955), (86, 0.013861672934458055), (87, 0.0031380393059528755), (88,
0.002069528570438413), (89, 0.008284773697508746), (90, 0.0049404544625377512), (91, 0.022042118492069357), (92,
0.0053301963005114705), (93, 0.002737855058625406 6), (94, 0.006649798417048129), (95, 0.02167043352406266), (96,
0.00586285464184385), (97, 0.0035175133844280422), (98, 0.008899355457124027), (99, 0.013489147147690846)]
[(52, 0.042431404065756184), (39, 0.039884267610139176), (13, 0.0393777270035732 8), (0, 0.035400591802442975), (26,
0.03489656776403214)]
elapsed time is 37.680768966674805
23/04/11 02:22:34 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@2abcbfce{HTTP/1.1, (http/1.1)}
{0.0.0.0:0}
Job [1ee1836c317147dfbca3438bab93288e] finished successfully.
done: true
driverControlFilesUri: gs://data301-2023-amh284-lab4-bucket/google-cloud-dataproc-metainfo/7b547cad-1010-4f78-b992-
a78d0955a99a/jobs/1ee1836c317147dfbca3438bab93288e/
driverOutputResourceUri: gs://data301-2023-amh284-lab4-bucket/google-cloud-dataproc-metainfo/7b547cad-1010-4f78-b992-
a78d0955a99a/jobs/1ee1836c317147dfbca3438bab93288e/driveroutput
jobUuid: 25cf35ca-ac4d-32c7-b272-34617661a235
```

Time: 37.68s

4 vCPUs – graph-full.txt:



Time: 25.79s

8 vCPUs – facebook-medium.txt:

```
0.00046868997454572135), (2005, 0.00011734577974586898), (2006, 0.00023483598551833482), (2007,
0.00023483598551831693), (2008, 0.000105042536877290 3), (2009, 0.00010521184040996893), (2010, 0.00010521184040993245),
(2011, 5.260592020499398e-05), (2012, 1.3151480051240324e-05), (2013, 6.376205434359697e-05), (2014,
6.376205434361541e-05), (2015, 0.00019068541917307264), (2016, 0.00019068541917308167), (2017, 6.356180639102329e-05),
(2018, 0.00016122409793051438), (2019, 0.00016122409793051 74), (2020, 0.00025175528041995 43), (2021,
0.00025175528042003386), (2022, 0.00012587764020996424), (2023, 0.00014904048917319638), (2024,
0.00028367625516893976), (2025, 0.0002836762551690944), (2026, 0.00015246926790562 51), (2027, 0.00016983274545865693),
(2028, 0.00019135665329966 12), (2029, 0.00019135665329978528), (2030, 3.148965071936907e-05), (2031,
0.00031400919667592974), (2032, 0.00033490117929738474), (2033, 0.00033490117929739776), (2034, 9.264829147002243e-05),
(2035, 9.264829147003933e-05), (2036, 5.4430385830985656e-05), (2037, 5.443038583100696e-05), (2038,
1.0886077166195734e-05), (2039, 2.1772154332406553e-06), (2040, 0.00011540515839523935), (2041,
0.00011540515839524094), (2042, 1.0491378035930796e-05), (2043, 4.991293258252897e-05), (2044, 4.996364711302129e-05),
(2045, 7.617204177364726e-05), (2046, 8.131522102671022e-05), (2047, 0.00018380075378973795), (2048,
0.00018380075378976327), (2049, 6.12669179299099e-05), (2050, 0.00010273908547012197), (2051, 0.00032580190469225527),
(2052, 0.0003258019046922879), (2053, 8.274276778599006e-05), (2054, 0.00010912010887291967), (2055,
0.00010927820726720075), (2056, 0.00011955160354548852), (2057, 0.00020374411407799004), (2058, 0.00020374411407799635),
(2059, 0.00010187205703893 8), (2060, 5.093602851951157e-05), (2061, 2.5468014259729405e-05), (2062,
1.2734007129883798e-05), (2063, 2.9020881823711 67e-05), (2064, 2.9020881182396212e-05), (2065, 1.4510440591182453e-
05), (2066, 7.255220295602223e-06), (2067, 1.4876536560430756e-05)]
[(41, 0.0023694802364354656), (42, 0.0023694802364352756), (1290, 0.0022868474237279506),
(851, 0.0021750897793447904)]
elapsed time is 19.83701491355896
23/04/11 03:24:15 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@7016b8d7{HTTP/1.1, (http/1.1)}
{0.0.0.0:0}
Job [edda43d628e94a1cade8bafeaa6e092f] finished successfully.
done: true
driverControlFilesUri: gs://data301-2023-amh284-lab4-bucket/google-cloud-dataproc-metainfo/0cf214a6-6afc-4e29-8759-
adaea2e5c84b/jobs/edda43d628e94a1cade8bafeaa6e092f/
driverOutputResourceUri: gs://data301-2023-amh284-lab4-bucket/google-cloud-dataproc-metainfo/0cf214a6-6afc-4e29-8759-
adaea2e5c84b/jobs/edda43d628e94a1cade8bafeaa6e092f/driveroutput
jobUuid: 1357278c-b1c4-3a11-abbc-29633ed1968c
placement:
  clusterName: data301-2023-amh284-lab4-cluster
  clusterUuid: 0cf214a6-6afc-4e29-8759-adaea2e5c84b
pysparkJob:
  args:
  - gs://data301-2023-amh284-lab4-bucket/facebook-medium.txt
  mainPythonFileUri: gs://data301-2023-amh284-lab4-bucket/google-cloud-dataproc-metainfo/0cf214a6-6afc-4e29-8759-
adaea2e5c84b/jobs/edda43d628e94a1cade8bafeaa6e092f/staging/pyspark_pagerank.py
```

Time: 19.837

16 vCPUs – facebook-large.txt:

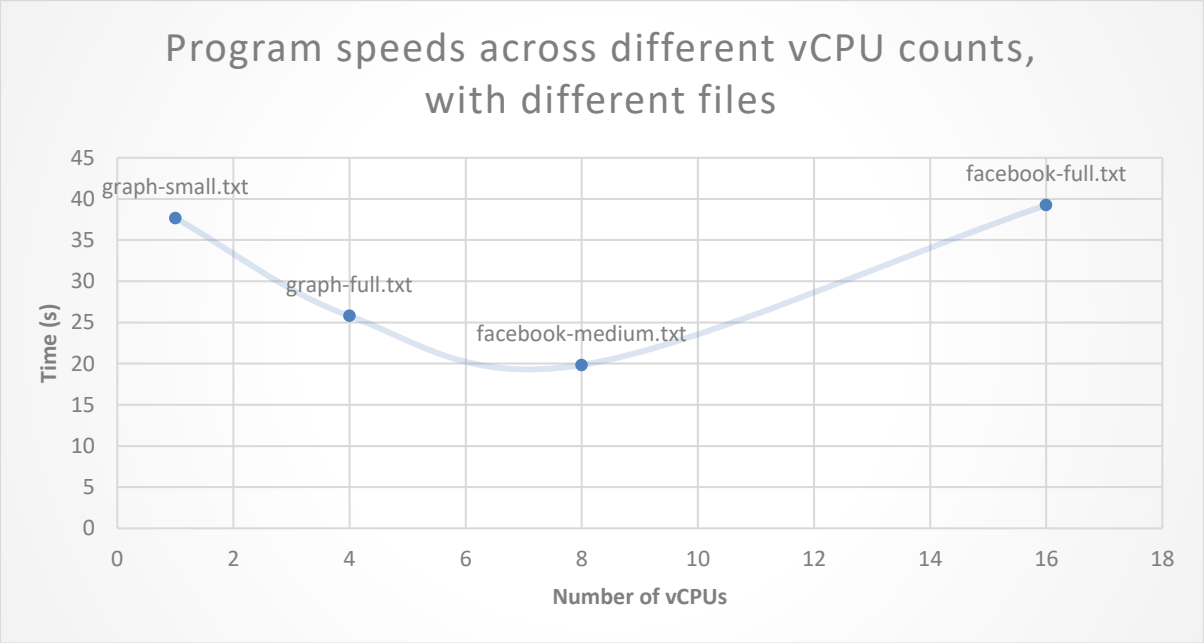
```
(36960, 2.44405711541861e-41), (36961, 2.44405711541861e-41), (36962, 1.222028557709305e-41), (36963, 1.8330428365639576e-41)
, (36964, 1.3747821274229683e-41), (36965, 1.3747821274229683e-41), (36966, 1.3747821274229683e-41), (36967, 3.
4369553185574206e-41), (36968, 8.592388296393548e-41), (36969, 5.728258864262365e-41), (36970, 3.818839242841576e-41),
(36971, 3.818839242841576e-41), (36972, 3.818839242841576e-41), (36973, 5.728258864262365e-41), (36974, 3.818839242841576e-41)
, (36975, 3.818839242841576e-41), (36976, 5.728258864262368e-41), (36977, 1.909419621420789e-41), (36978, 2.
8641294321311834e-41), (36979, 2.8641294321311834e-41), (36980, 1.2888582444590327e-40), (36981, 6.444291222295163e-41),
(36982, 3.866574733377098e-40), (36983, 3.866574733377098e-40), (36984, 3.866574733377098e-40), (36985, 3.866574733377098e-40)
, (36986, 1.933287366688549e-40), (36987, 6.444291222295163e-41), (36988, 3.6517650259672586e-40), (36989, 5.
477647538950888e-40), (36990, 3.6517650259672586e-40), (36991, 3.6517650259672586e-40), (36992, 3.6517650259672586e-40),
(36993, 3.6517650259672586e-40), (36994, 1.8258825129836293e-40), (36995, 1.8258825129836293e-40), (36996, 1.
8258825129836293e-40), (36997, 1.8258825129836293e-40), (36998, 1.8258825129836293e-40)]
[(426, 0.002799640355859079), (427, 0.0027763752183950566), (433, 0.002708151848806962), (438, 0.0026984926092680087), (434,
0.002685260079087061)]
elapsed time is 39.26077151298523
23/04/13 02:44:34 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@5ee6aca5{HTTP/1.1, (http/1.1)}{0.0.0.
0:0}
```

Time: 39.261s

Graph:

Note: The speeds are inconsistent because each run with different vCPUs had different file sizes.

```
Number of lines:
   1024 ./graph-small.txt
   8192 ./graph-full.txt
  55093 ./facebook-medium.txt
 202791 ./facebook-large.txt
```



Despite the size of the files increasing with the vCPU count, the program does still speed up. graph-full is 8x larger than graph-small with only a 4x processor increase, however the program is still significantly faster. Likewise, facebook-medium is 7x larger still, and doubling the processor count increases the speed.

However, once we get to facebook-full.txt, the program slows down significantly. It's merely a 4x file size increase against a 2x processor increase, however it takes twice as long to finish as the last file.