BigData Assignment 1
Report

Input Dataset
WikiArticles Edge Relationship data, approx size of 30GB. Each line in the input is of the format 'A\tB' which can be interpreted as a directed edge from A to B.

Project Goal
In all the tasks below, we tried to find the pagerank of all nodes with non zero incoming edges from the above input data. A PySpark application was used to achieve the task of computing pageranks. We instrumented metrics like job completion time, network traffic across worker nodes, disk read/write traffic of the workers, etc. The PageRank algorithm is described in the code as well as the README files.

Hardware Used
A 3 node cluster with each node having 1 cpu with 5 cores was used. A HDFS file system was spun up on the nodes with one NameNode and 3 DataNodes. The bandwidth available between the nodes was around 7 Gbits/sec. An external filesystem of 96GB was mounted on each of the 3 nodes, the read/write bandwidth measured was around 280 MBps. Spark was launched in standalone mode on the cluster, with one Master and 3 Worker Nodes. The Memory for each executor and the driver was set to 16GB.

Terminology Used
In all the graphs, by read/write we mean read and write rate to disk of the workers. By send/recv we mean the input and output network traffic of the worker nodes. In the plots, each unit on the x-axis is equivalent to 10 secs and each unit on the y-axis is 1 byte, unless specified explicitly.

TASK 1

A PySpark application with no custom optimisations was run to calculate the pageranks of the wiki data. The computed ranks were written back to hdfs. The job took around 33 minutes for completion.

Spark split the job into 13 stages, and a total of 1157 tasks over all stages. Thereafter, each iteration of the pagerank algorithm took around 1.5 mins. PFB the related job charts

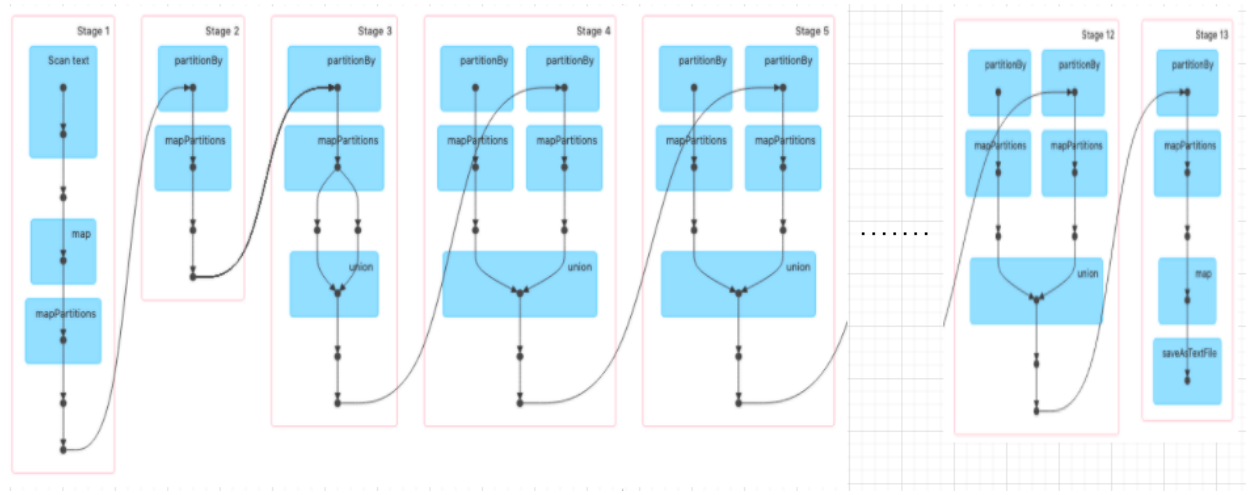


Fig 1.1 A directed graph showing the execution of all the stages in the job

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	runJob at SparkHadoopWriter.scala:83 +details	2021/09/26 22:19:02	18 s	89/89		733.0 MiB	1952.6 MiB	
12	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:17:39	1.4 min	89/89			5.7 GiB	1952.6 MiB
11	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:16:17	1.4 min	89/89			5.7 GiB	1952.6 MiB
10	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:14:54	1.4 min	89/89			5.7 GiB	1952.6 MiB
9	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:13:33	1.4 min	89/89			5.7 GiB	1952.7 MiB
8	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:12:10	1.4 min	89/89			5.7 GiB	1952.8 MiB
7	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:10:46	1.4 min	89/89			5.7 GiB	1953.3 MiB
6	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:09:23	1.4 min	89/89			5.7 GiB	1955.6 MiB
5	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:08:00	1.4 min	89/89			5.8 GiB	1970.2 MiB
4	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:06:30	1.5 min	89/89			6.4 GiB	2.0 GiB
3	reduceByKey at /mnt/data/part3/task1/task.py:61 +details	2021/09/26 22:04:58	1.5 min	89/89			7.5 GiB	2.6 GiB
2	groupByKey at /mnt/data/part3/task1/task.py:52 +details	2021/09/26 21:59:04	5.9 min	89/89			4.9 GiB	3.8 GiB
1	distinct at /mnt/data/part3/task1/task.py:52 +details	2021/09/26 21:46:50	12 min	89/89	9.9 GiB			4.9 GiB

Fig 1.2 Time taken and Shuffle read/write traffic for each stage

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1211	1211	7.6 h (1.6 min)	9.9 GiB	66.3 GiB	28.6 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1211	1211	7.6 h (1.6 min)	9.9 GiB	66.3 GiB	28.6 GiB	0

Executors

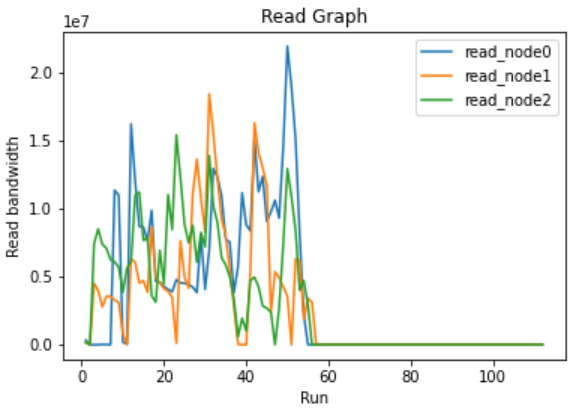
Show 20 entries Search:

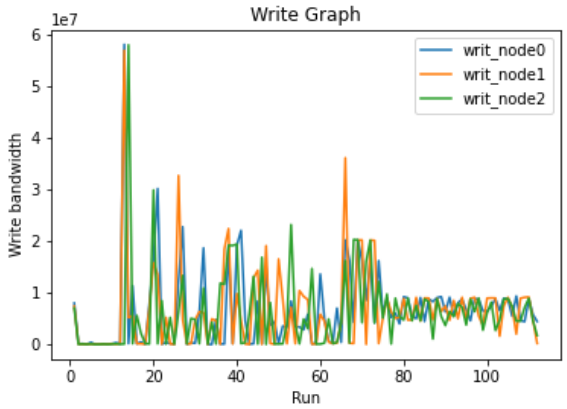
Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
0	172.17.229.2:38943	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	408	408	2.5 h (35 s)	3.4 GiB	22.4 GiB	9.7 GiB	stdout stderr
driver	c220g1-030824vm-1.wisc.cloudlab.us:43433	Active	0	0.0 B / 15.8 GiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	
1	172.17.229.3:43799	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	378	378	2.5 h (26 s)	3.1 GiB	21.7 GiB	9.3 GiB	stdout stderr
2	172.17.229.1:45263	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	425	425	2.5 h (37 s)	3.4 GiB	22.2 GiB	9.5 GiB	stdout stderr

Fig 1.3 Summary of the driver and executor nodes

Observations

- Total data in the Shuffle read operations was around 66GB. For Shuffle write operations, it was 29GB. The Shuffle read arises from the join and map operation (*calc_contribution*) in the for loop.
- Significant time was taken for finding out the distinct edges (12 mins) and then grouping them by the key (6 mins). Thereafter, each iteration of the pagerank algorithm took around 1.5 mins.





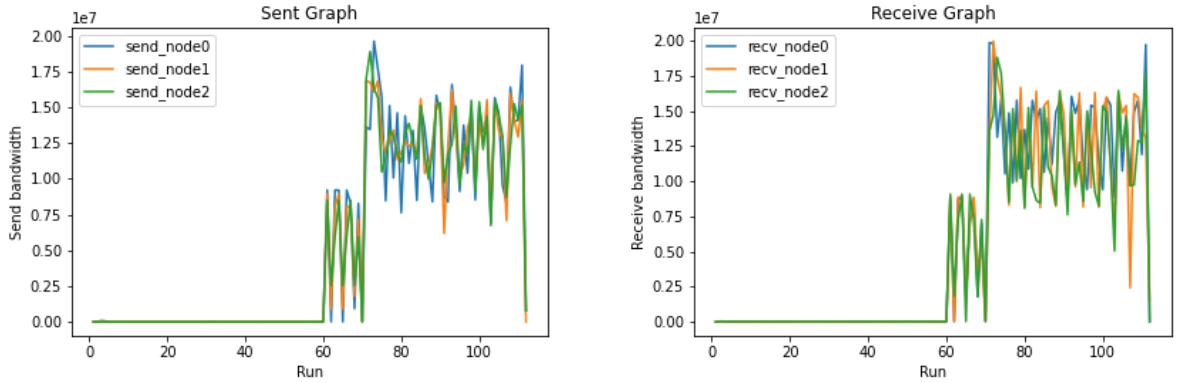


Fig 1.4 Disk and Network stats for the worker nodes

- Send/Receive network traffic increased after all the nodes read the data. The peaks in the recv/send graph correspond to the data being shuffled between stages.

- Data started to be written on disk after some delay. This delay indicates that the data is read and processed and then written back to the disk by the nodes.

TASK 2

In this task, we tried to observe the changes when custom partitioning is introduced in the code. In specific, we ran the code with partitions of 50, 150 & 300 and noted the below observations. Both *ranks* and *edges* RDDs were partitioned with the same custom partitions. Following are the statistics and observations from the experiments with different number of partitions:

Partitions	Job Time	Total tasks over Stages	Total Stages
50	42 mins	817	14
150	31 mins	1917	14
300	35 mins	3567	14

Observations

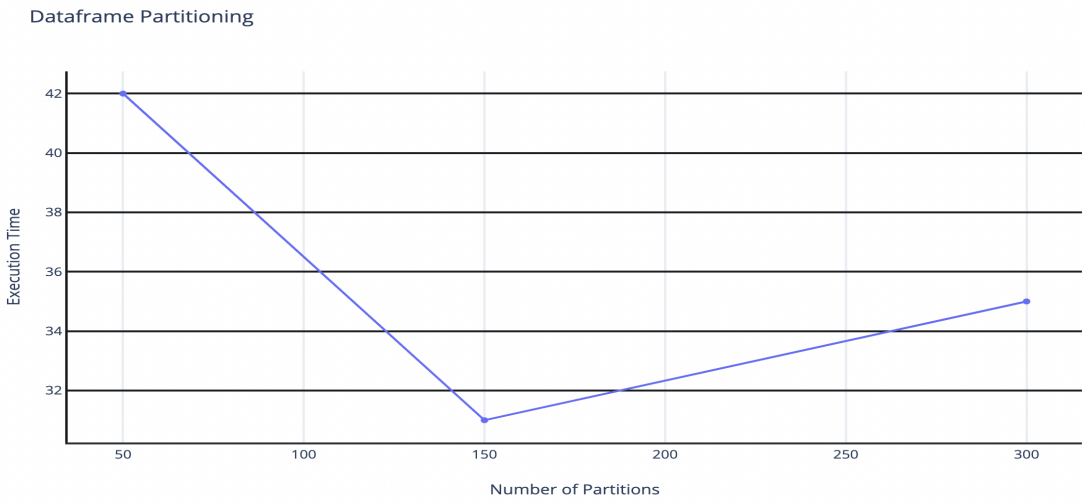


Fig 2.1 Trend of Job Completion time against number of partitions. y-axis in mins scale

- Less number of partitions lead to lower cluster usage. If one of the tasks takes longer to compute this would not be distributed among different nodes as the tasks are not split into sub-tasks that can be parallelized, leading to higher job completion time. This can be deduced from the 50 partition experiment
- Higher number of partitions lead to higher job completion time as more time is spent in scheduling and execution of jobs (sequence of parallel tasks). This can be deduced from the 300 partition experiment
- The optimal value for the number of partitions could be around 100-150 partitions because it has lower completion time compared to the other experiments.

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	871	871	9.4 h (1.5 min)	9.9 GiB	57.3 GiB	30.5 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	871	871	9.4 h (1.5 min)	9.9 GiB	57.3 GiB	30.5 GiB	0

Fig 2.2 50 partitions

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1971	1971	7.2 h (2.1 min)	9.9 GiB	59.6 GiB	32.8 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1971	1971	7.2 h (2.1 min)	9.9 GiB	59.6 GiB	32.8 GiB	0

Fig 2.3 150 partitions

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	3621	3621	7.2 h (1.6 min)	9.9 GiB	62.9 GiB	36 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	3621	3621	7.2 h (1.6 min)	9.9 GiB	62.9 GiB	36 GiB	0

Fig 2.4 300 partitions

- The shuffle read/write is increasing with the number of partitions. This is expected behavior, as the number of partitions increases, the probability that a generated tuple (from the *calc_contribution* function) lies in the same partition decreases. Hence it will contribute to both shuffle read and write increase. We can infer this from the above job statistics.

- We believe that wide dependencies (child record depends on multiple parent records) contribute to the shuffle read/write data.

TASK 3

In this task appropriate Data Frames (both *ranks* and *edges*) were persisted to cache. The computed ranks are written back to HDFS. The job took 31 mins to complete which is less than the time in Task 1. The execution contained in total 13 stages which consisted of 1211 tasks.

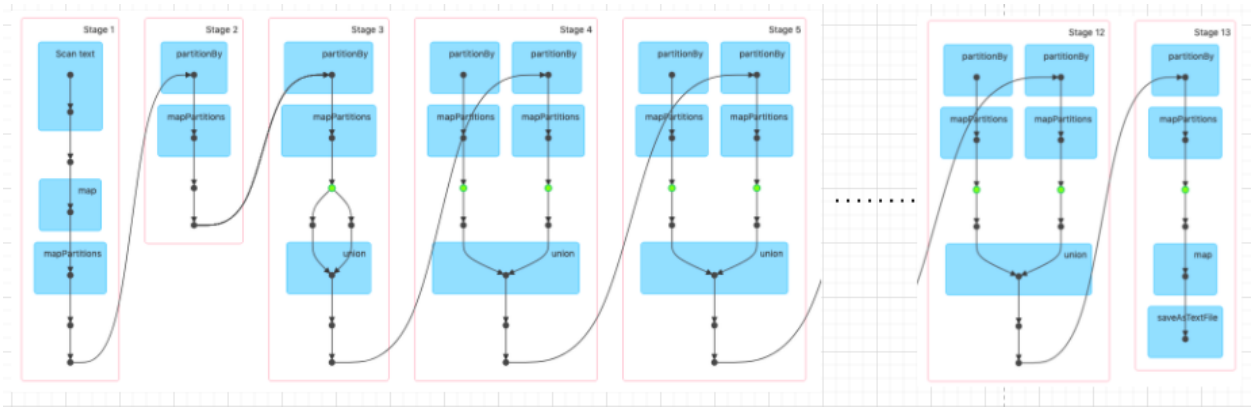


Fig 3.1 A directed graph showing the execution of all the stages in the job

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	runJob at SparkHadoopWriter.scala:83 +details	2021/09/26 23:40:13	20 s	89/89		733.0 MiB	1952.5 MiB	
12	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:39:06	1.1 min	89/89	2.7 GiB		1952.5 MiB	1952.5 MiB
11	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:38:00	1.1 min	89/89	2.7 GiB		1952.5 MiB	1952.5 MiB
10	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:36:53	1.1 min	89/89	2.7 GiB		1952.6 MiB	1952.5 MiB
9	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:35:47	1.1 min	89/89	2.7 GiB		1952.7 MiB	1952.6 MiB
8	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:34:40	1.1 min	89/89	2.7 GiB		1953.2 MiB	1952.7 MiB
7	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:33:33	1.1 min	89/89	2.7 GiB		1955.5 MiB	1953.2 MiB
6	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:32:27	1.1 min	89/89	2.7 GiB		1970.0 MiB	1955.5 MiB
5	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:31:20	1.1 min	89/89	2.7 GiB		2.0 GiB	1970.0 MiB
4	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:30:06	1.2 min	89/89	2.7 GiB		2.6 GiB	2.0 GiB
3	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/26 23:28:39	1.4 min	89/89	2.7 GiB		3.8 GiB	2.6 GiB
2	groupByKey at /mnt/data/part3/task3/task.py:52 +details	2021/09/26 23:22:42	5.9 min	89/89			4.9 GiB	3.8 GiB
1	distinct at /mnt/data/part3/task3/task.py:52 +details	2021/09/26 23:10:11	13 min	89/89	9.9 GiB			4.9 GiB

Fig 3.2 Time taken and Shuffle read/write traffic for each stage

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1211	1211	7.2 h (2.7 min)	36.6 GiB	28.6 GiB	28.6 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	0	0	1211	1211	7.2 h (2.7 min)	36.6 GiB	28.6 GiB	28.6 GiB	0

Executors

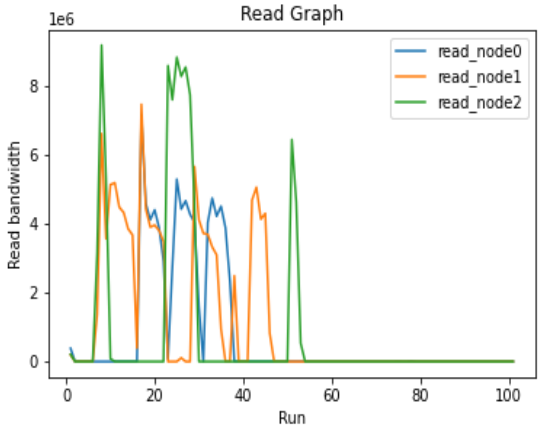
Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
0	172.17.229.2:35801	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	413	413	2.4 h (54 s)	12.5 GiB	9.7 GiB	9.7 GiB	stdout stderr
driver	c220g1-030824vm-1.wisc.cloudlab.us:33847	Active	0	0.0 B / 15.8 GiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	
1	172.17.229.3:35455	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	386	386	2.4 h (50 s)	11.9 GiB	9.3 GiB	9.3 GiB	stdout stderr
2	172.17.229.1:46779	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	412	412	2.4 h (56 s)	12.3 GiB	9.6 GiB	9.6 GiB	stdout stderr

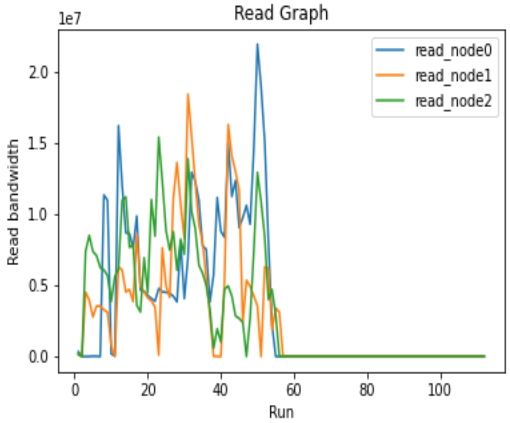
Fig 3.3 Summary of the driver and executor nodes

Observations

- The job took 31 mins to complete which is less than the time in Task 1.
- The reduce stages took an average of 1.1 mins whereas in task 1 they took 1.5 mins on average.
- Read size from disk is lower for all nodes in this task compared to Task 1 which is consistent with the fact the machine is using in-memory (cache) resources to store data and it is only going to disk for data which is not available in cache. *(notice the magnitudes on both y axes)*

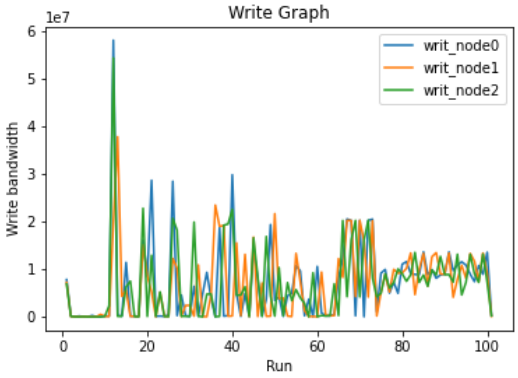


In-Memory Execution

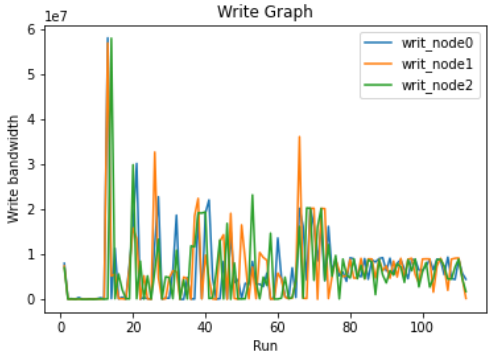


Normal Execution

- Write bandwidth remains the same when compared to Task 1. This is kind of expected as caching shouldn't have much effect on the write part. We see that the graph is similar to that of task 1.



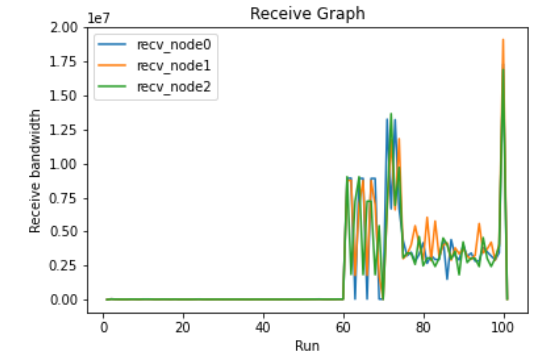
In-Memory Execution

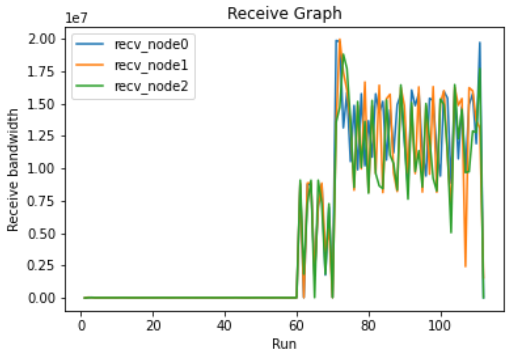


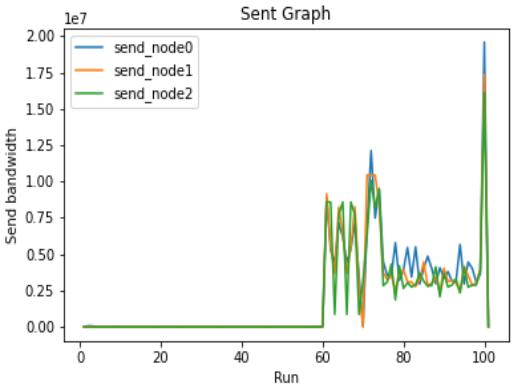
Normal Execution

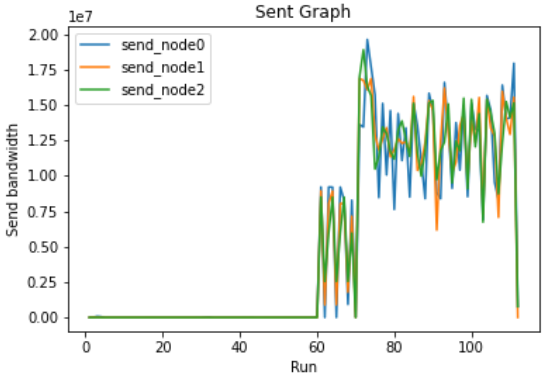
- Along with the decrease in disk reads, the total Shuffle Read data also decreases (almost by 40GB). This follows from the fact that we are caching things and they need not read from disk repeatedly and shuffled across nodes.

- Network data i.e receive and send data in each node has comparatively lower average value as compared to normal execution. This is expected since as per code, we are caching the data in each node for the data reuse and will lessen the burden of network communication with other nodes.









In-Memory Execution *Normal Execution*

TASK 4

In this task we tried to observe the failure recovery of Spark. We replicated task3 with one change. A worker process was killed when the application reached 25% and 75% of its lifetime (lifetime value derived from task 3). The computed ranks are written back to HDFS. In the case where the worker node was killed at 25% of application lifetime, the job took 43 minutes and in the second case where the worker was killed at 75% of application lifetime, the job took 44 minutes. Both of these jobs took a higher time compared to Task 3. The execution consisted of total of 13 stages and 1157 tasks.

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	runJob at SparkHadoopWriter.scala:83 +details	2021/09/27 01:29:43	26 s	89/89		733.0 MiB	1952.6 MiB	
12	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:28:10	1.5 min	89/89	2.7 GiB		1952.6 MiB	1952.6 MiB
11	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:26:36	1.6 min	89/89	2.7 GiB		1952.6 MiB	1952.6 MiB
10	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:25:00	1.6 min	89/89	2.7 GiB		1952.7 MiB	1952.6 MiB
9	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:23:27	1.5 min	89/89	2.7 GiB		1952.8 MiB	1952.7 MiB
8	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:21:52	1.6 min	89/89	2.7 GiB		1953.3 MiB	1952.8 MiB
7	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:20:17	1.6 min	89/89	2.7 GiB		1955.6 MiB	1953.3 MiB
6	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:18:41	1.6 min	89/89	2.7 GiB		1970.1 MiB	1955.6 MiB
5	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:17:03	1.6 min	89/89	2.7 GiB		2.0 GiB	1970.1 MiB
4	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:15:16	1.8 min	89/89	2.7 GiB		2.6 GiB	2.0 GiB
3	reduceByKey at /mnt/data/part3/task3/task.py:62 +details	2021/09/27 01:13:16	2.0 min	89/89	2.7 GiB		3.8 GiB	2.6 GiB
2	groupByKey at /mnt/data/part3/task3/task.py:52 +details	2021/09/27 01:04:52	8.4 min	89/89			4.9 GiB	3.8 GiB
1	distinct at /mnt/data/part3/task3/task.py:52 +details	2021/09/27 00:47:07	18 min	104/89 (20 failed)	11.7 GiB			5.8 GiB

Fig 4.1 Time taken and Shuffle read/write traffic for each stage when worker killed at 25%

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(3)	0	0.0 B / 47.5 GiB	0.0 B	10	0	0	1193	1193	6.9 h (2.6 min)	36.6 GiB	28.6 GiB	28.6 GiB	0
Dead(1)	0	0.0 B / 15.8 GiB	0.0 B	5	-15	20	33	38	1.2 h (11 s)	1.9 GiB	0.0 B	922.4 MiB	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	-15	20	1226	1231	8.1 h (2.8 min)	38.5 GiB	28.6 GiB	29.5 GiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
0	172.17.229.2:40385	Dead	0	0.0 B / 15.8 GiB	0.0 B	5	-15	20	33	38	1.2 h (11 s)	1.9 GiB	0.0 B	922.4 MiB	stdout stderr
driver	c220g1-030824vm-1.wisc.cloudlab.us:45943	Active	0	0.0 B / 15.8 GiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	
1	172.17.229.3:38849	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	593	593	3.4 h (1.2 min)	18.1 GiB	14.3 GiB	14.2 GiB	stdout stderr
2	172.17.229.1:43671	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	0	600	600	3.5 h (1.4 min)	18.5 GiB	14.3 GiB	14.4 GiB	stdout stderr

Fig 4.2 Summary of the driver and executor nodes when worker killed at 25%

Stage Id	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	runJob at SparkHadoopWriter.scala:83	+details	2021/09/27 02:16:56	26 s	89/89		733.0 MiB	1952.6 MiB	
12	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:15:26	1.5 min	89/89	2.7 GiB		1952.6 MiB	1952.6 MiB
11	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:13:56	1.5 min	89/89	2.7 GiB		1952.6 MiB	1952.6 MiB
10	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:12:25	1.5 min	89/89	2.7 GiB		1952.7 MiB	1952.6 MiB
9	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:10:55	1.5 min	89/89	2.7 GiB		1952.8 MiB	1952.7 MiB
8	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:09:25	1.5 min	89/89	2.7 GiB		1953.3 MiB	1952.8 MiB
7	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:07:55	1.5 min	89/89	2.7 GiB		1955.5 MiB	1953.3 MiB
6	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:06:24	1.5 min	89/89	2.7 GiB		1970.1 MiB	1955.5 MiB
5 (retry 1)	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:05:53	31 s	29/29	895.4 MiB		670.6 MiB	643.2 MiB
4 (retry 1)	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:05:19	34 s	29/29	895.4 MiB		868.3 MiB	670.9 MiB
4	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 01:53:49	1.3 min	89/89	2.7 GiB		2.6 GiB	2.0 GiB
3 (retry 1)	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 02:04:38	40 s	29/29	895.4 MiB		1262.7 MiB	870.8 MiB
3	reduceByKey at /mnt/data/part3/task3/task.py:62	+details	2021/09/27 01:52:22	1.5 min	89/89	2.7 GiB		3.8 GiB	2.6 GiB
2 (retry 1)	groupByKey at /mnt/data/part3/task3/task.py:52	+details	2021/09/27 02:01:56	2.7 min	29/29			1641.4 MiB	1259.0 MiB
2	groupByKey at /mnt/data/part3/task3/task.py:52	+details	2021/09/27 01:46:32	5.8 min	89/89			4.9 GiB	3.8 GiB
1 (retry 1)	distinct at /mnt/data/part3/task3/task.py:52	+details	2021/09/27 01:56:09	5.8 min	29/29	3.2 GiB			1663.1 MiB
1	distinct at /mnt/data/part3/task3/task.py:52	+details	2021/09/27 01:33:53	13 min	89/89	9.9 GiB			4.9 GiB

Fig 4.3 Time taken and Shuffle read/write traffic for each stage when worker killed at 75%

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(3)	0	0.0 B / 47.5 GiB	0.0 B	10	0	3	1184	1187	6.8 h (2.1 min)	36.6 GiB	28.6 GiB	28.6 GiB	0
Dead(1)	0	0.0 B / 15.8 GiB	0.0 B	5	-25	29	168	172	1.8 h (29 s)	5.7 GiB	4.2 GiB	4.9 GiB	0
Total(4)	0	0.0 B / 63.3 GiB	0.0 B	15	-25	32	1352	1359	8.6 h (2.6 min)	42.3 GiB	32.8 GiB	33.5 GiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
0	172.17.229.3:36699	Dead	0	0.0 B / 15.8 GiB	0.0 B	5	-25	29	168	172	1.8 h (29 s)	5.7 GiB	4.2 GiB	4.9 GiB	stdout stderr
driver	c220g1-030824vm-1.wisc.cloudlab.us:36763	Active	0	0.0 B / 15.8 GiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	
1	172.17.229.2:33961	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	1	577	578	3.4 h (1.1 min)	18.1 GiB	14.2 GiB	14.2 GiB	stdout stderr
2	172.17.229.1:33627	Active	0	0.0 B / 15.8 GiB	0.0 B	5	0	2	607	609	3.4 h (59 s)	18.5 GiB	14.4 GiB	14.4 GiB	stdout stderr

Fig 4.4 Summary of the driver and executor nodes when worker killed at 75%

Note

In the 25% failure case, we killed node 1. And in the 75% failure case, we killed node 2.

Observations

- The most obvious observation we expected from task4 is the total inactivity of the killed node and an increase in load among other node activities in all parameters like disk stats or network stats. This is very evident in all the graphs of different system parameters over time. PFA some graphs from the second case where NODE 2 WORKER WAS KILLED at 75 percent of application time.

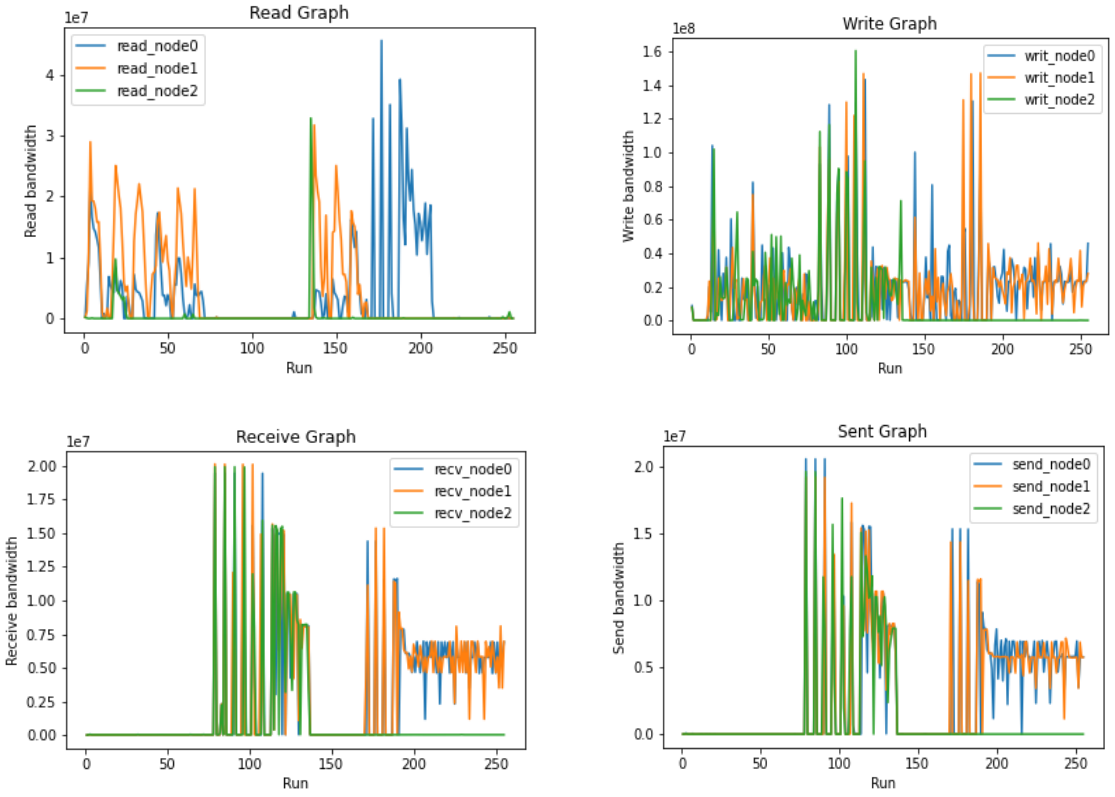


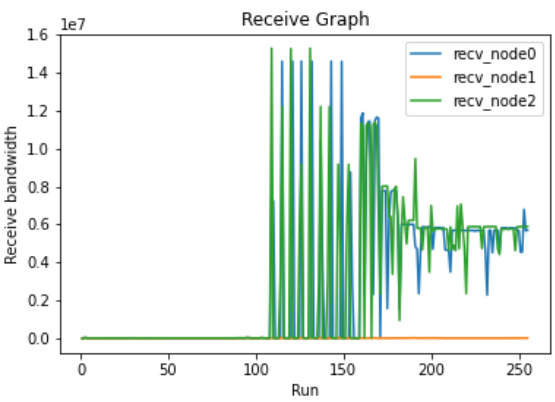
Fig 4.5 Disk Stats and Network Stats (when node 2 was killed)

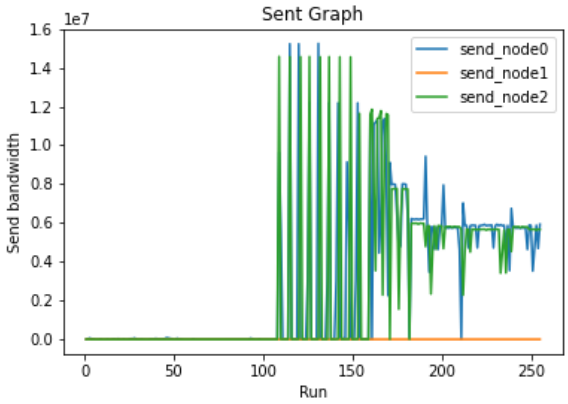
- The second inactivity plateau in the network graph corresponds to the phase when we kill the worker node and the remaining workers try to read the input data and compute the lost records. This can be further supported by the peaks in the disk statistics (read/write) plot, observed in the same time interval.

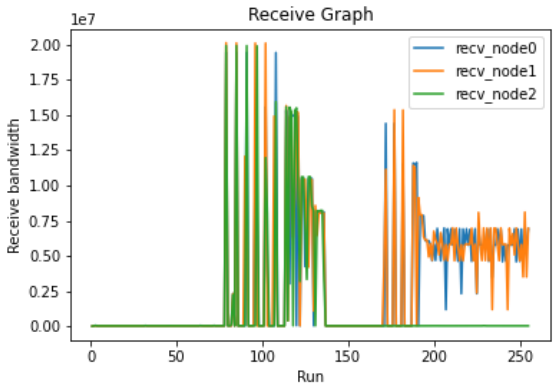
- Fault Tolerance
  We get to see Spark's fault tolerance in this setup. Whenever a worker process is killed, the other two workers try to recover the lost data and complete the execution. We can see this behaviour in the list of stages of both cases: in 75% failure and 25% failure case, we can see "retry" stages and "distinct" stage taking more time respectively. We believe that this is also the sole reason for the increase in the execution time (43 mins) as compared to the faultless execution (31 mins) in task3.

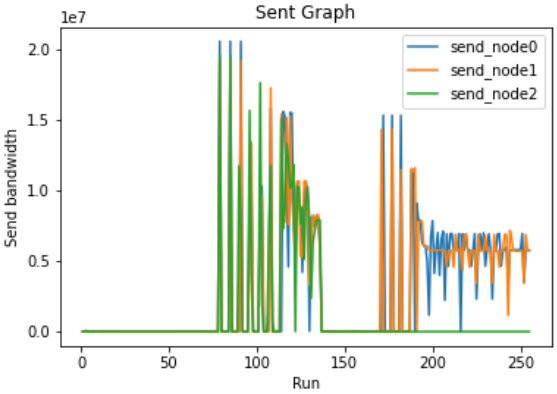
- Longer delay in network stats (receive and send graphs) in 25% failure case
  Comparing between 25% and 75% failure cases, we can see a considerable delay in the start of shuffle operations for the first case, which tells us that the node was killed in the earlier input stage

itself, after which two alive nodes took more time to recover the lost data read from the file. The delay in start of shuffle operation can be seen in the below graphs.









Recv throughput 25% fail vs 75% fail

Sent throughput, 25% fail vs 75% fail

- Repeated Shuffling in 75% failure case
  Shuffled read and write data is more in 75% failure case **approx. 4 GB more than 25% failure case**) as per the job data from SparkUI in above job diagrams. We believe that this is because of repeated shuffling, first time from the killed node and second time when the remaining two alive nodes are re-executing the tasks of the killed node again.

- **If** cost of shuffling was high
  As per the recorded metrics, both cases are taking approximately equal amounts of time to complete. We believe that if shuffle operations become costly, the completion time for the 75% failure case would have been more in comparison to the 25% failure case, which might not have been the case here

CONTRIBUTIONS

- Pankaj Kumar - Spark setup & Task 1, Task 2, Graphs, Report
- Muhammad Salman Munaf - Spark setup & Task 1, Task 3, Report
- Basava Kolagani - Spark setup & Task 1, Collection of disk & network stats, Task 4, Report

All three of us did the Spark setup part and tried to debug each others' issues. Again, all three of us spent time doing Task 1 and understanding the implementation in Spark.