CS 553
CLOUD COMPUTING
PROGRAMMING ASSIGNMENT-2

SUBMITTED BY :

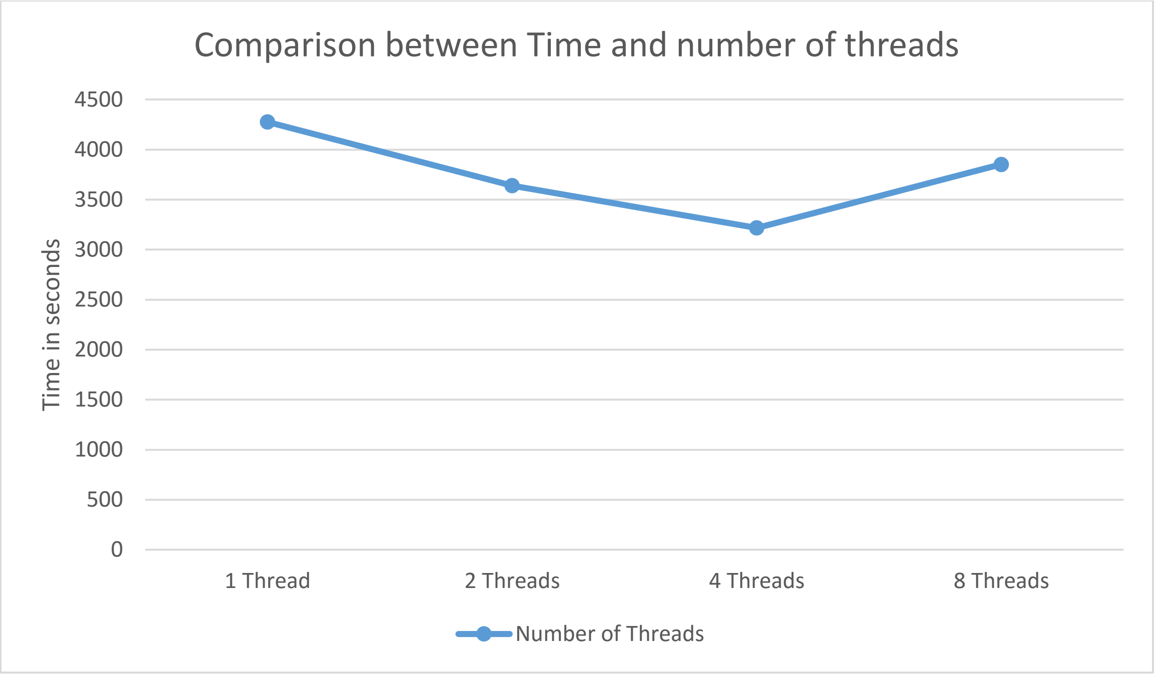
SACHIN KRISHNA MURTHY

CWID : A20354077

PERFORMANCE EVALUATION

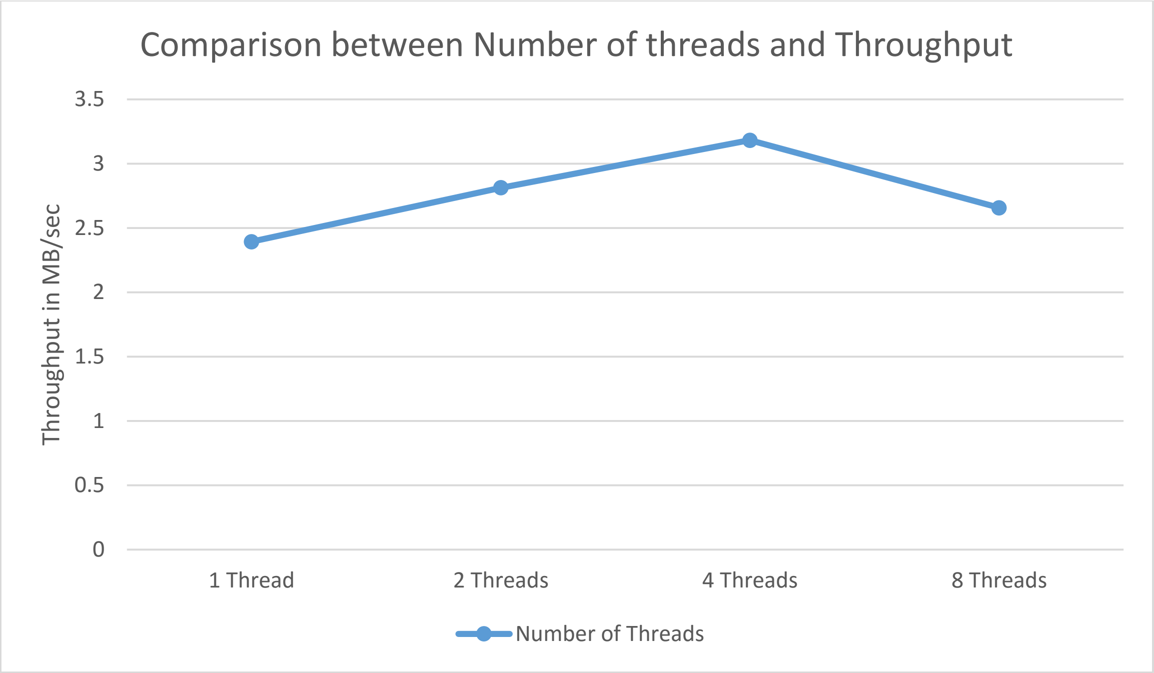
SHARED MEMORY

- **Comparison between Number of Threads and Time taken :**



From the above graph we can notice that the time taken for four threads is less when compared with other threads.

- **Comparison between number of threads and throughput value :**



From the above graph we can notice that the throughput value for 4 threads is greater than the other threads.

HADOOP

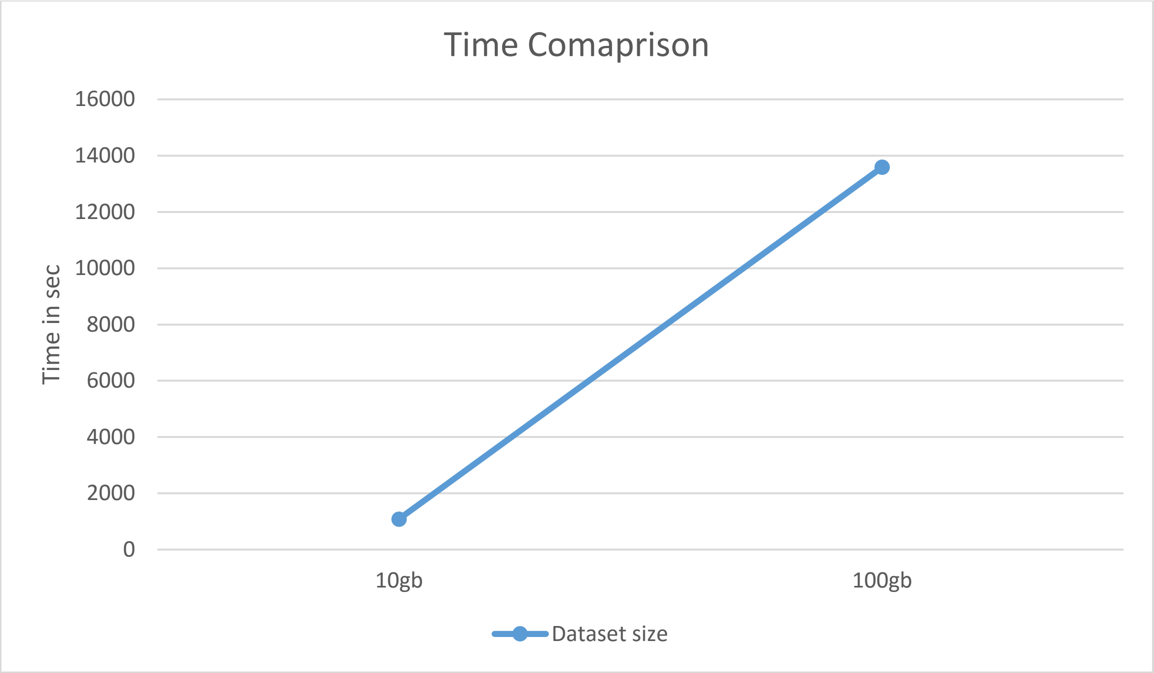
10gb :

Start Time	End Time	Total Time(end time-start time) sec	Throughput(Dataset size/Total Time)
20:49:12	21:07:14	1082	9.4639

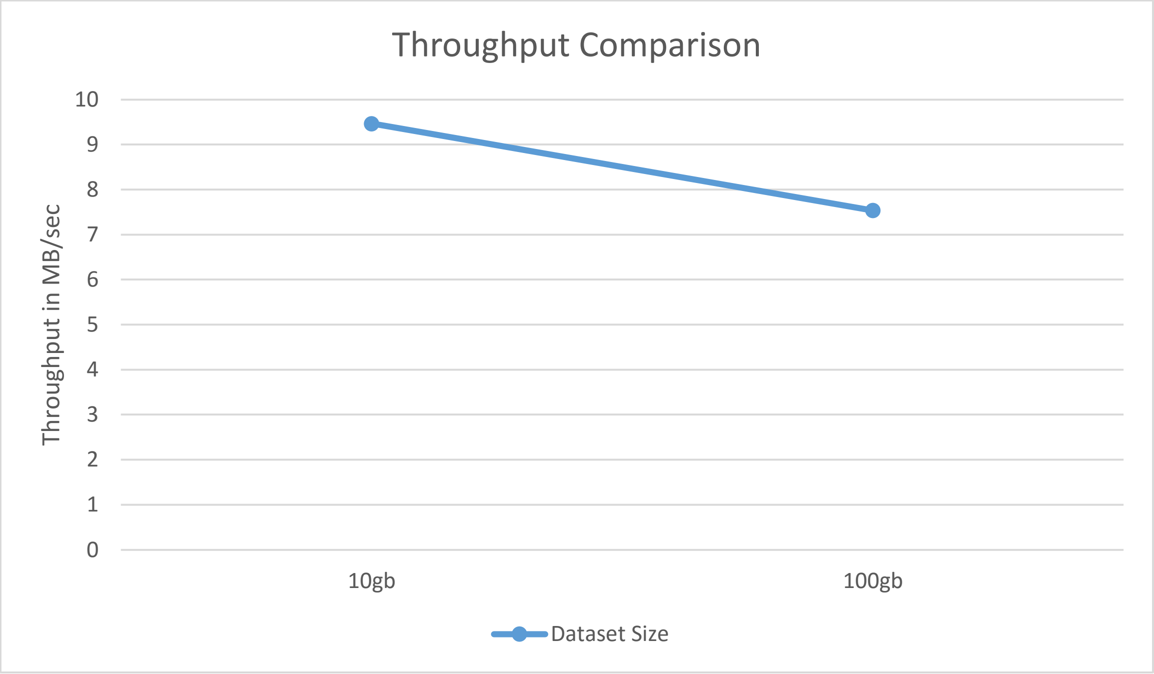
100gb :

Start Time	End Time	Total Time(end time-start time) sec	Throughput(Dataset size/Total Time)
5:40:06	9:26:33	13587	7.5366

- **Time Comparison :**



- **Throughput Comparison :**



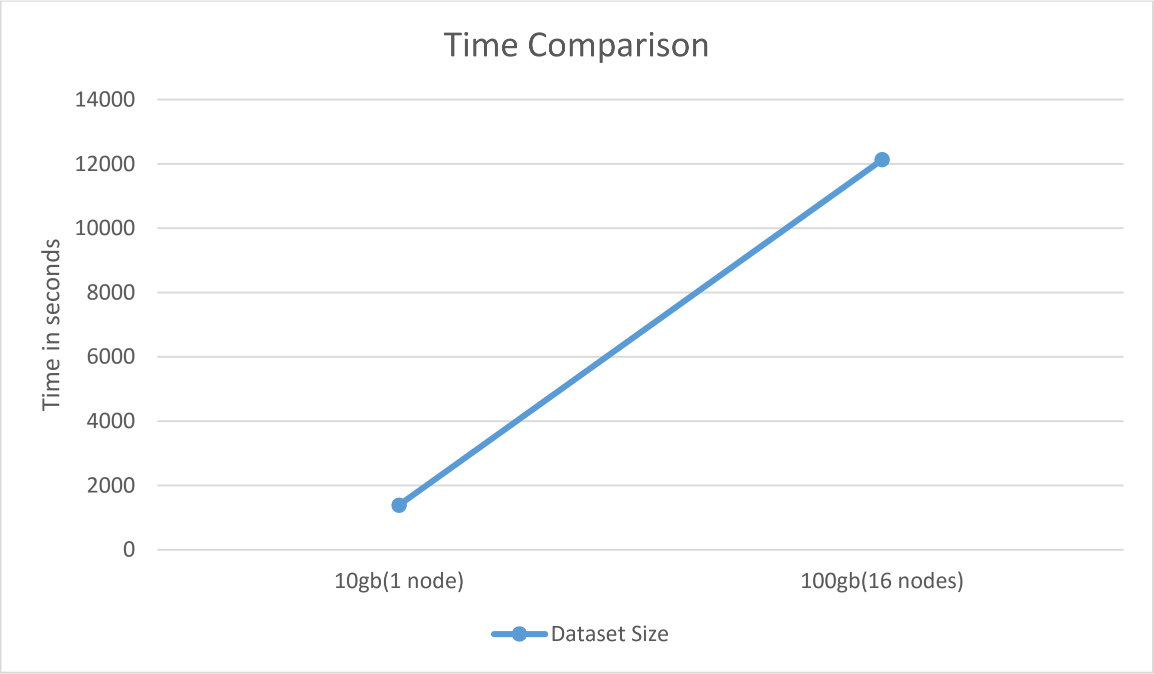
- **Comparison between 1 node (10gb) to 16 nodes (100gb) :**

We can observe that the total time required for the completion on 1 node (10gb) is 1082 seconds and on 16 nodes (100gb) is 13587 seconds. Also the throughput value for 1 node is 9.4639 whereas the throughput value for 16 nodes is 7.5366. Hence we can say that the time required for the completion of the program on 1 node (10gb) is less than the time required for the completion of the 16 nodes (100gb).

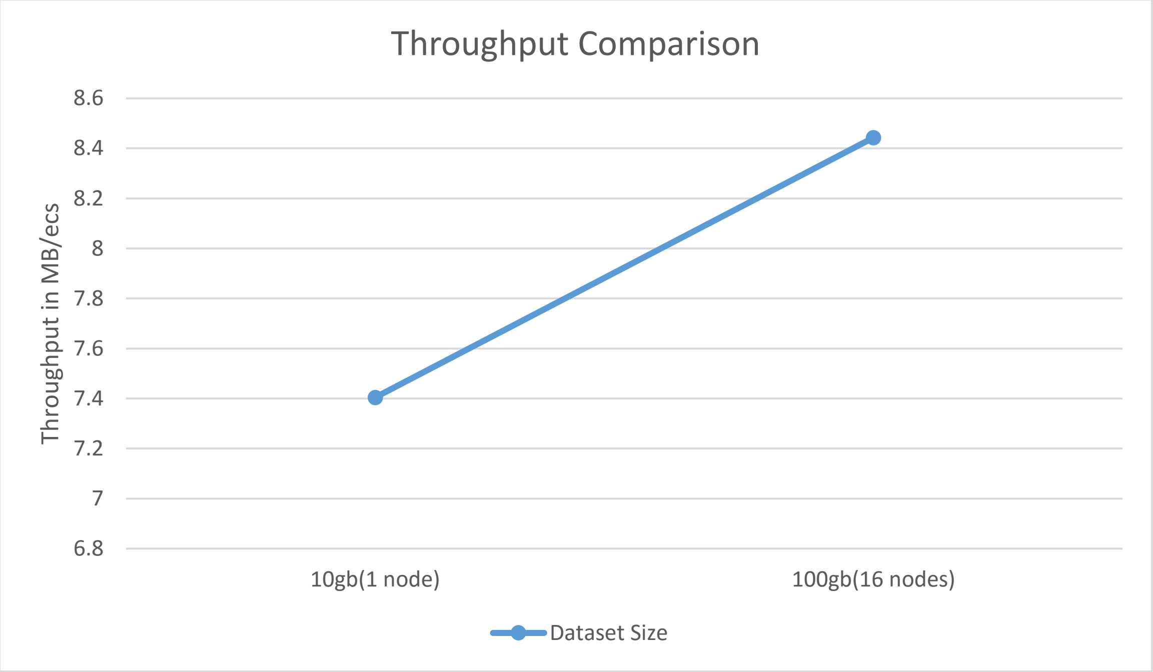
SPARK

	Total time (sec)	Throughput(MB/sec)
10gb (1 node)	1383	7.4041
100gb (16 nodes)	12127	8.443

- **Time Comparison :**



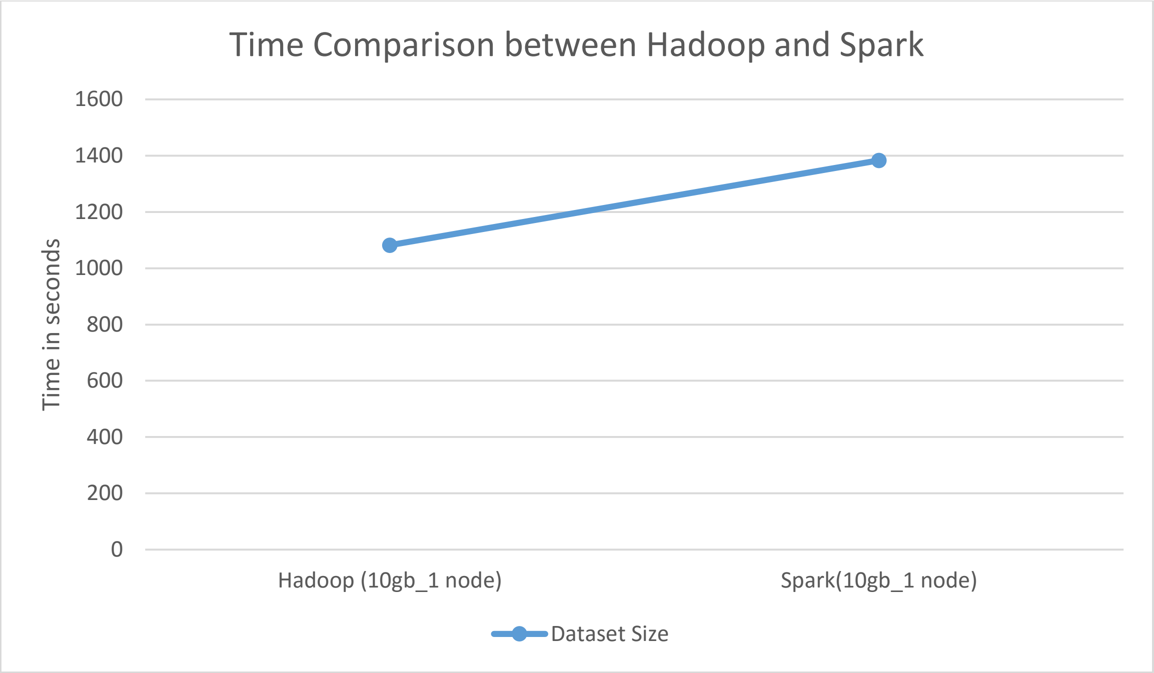
- **Throughput Comparison :**



- **Comparison between 1 node and 16 nodes of Spark :**

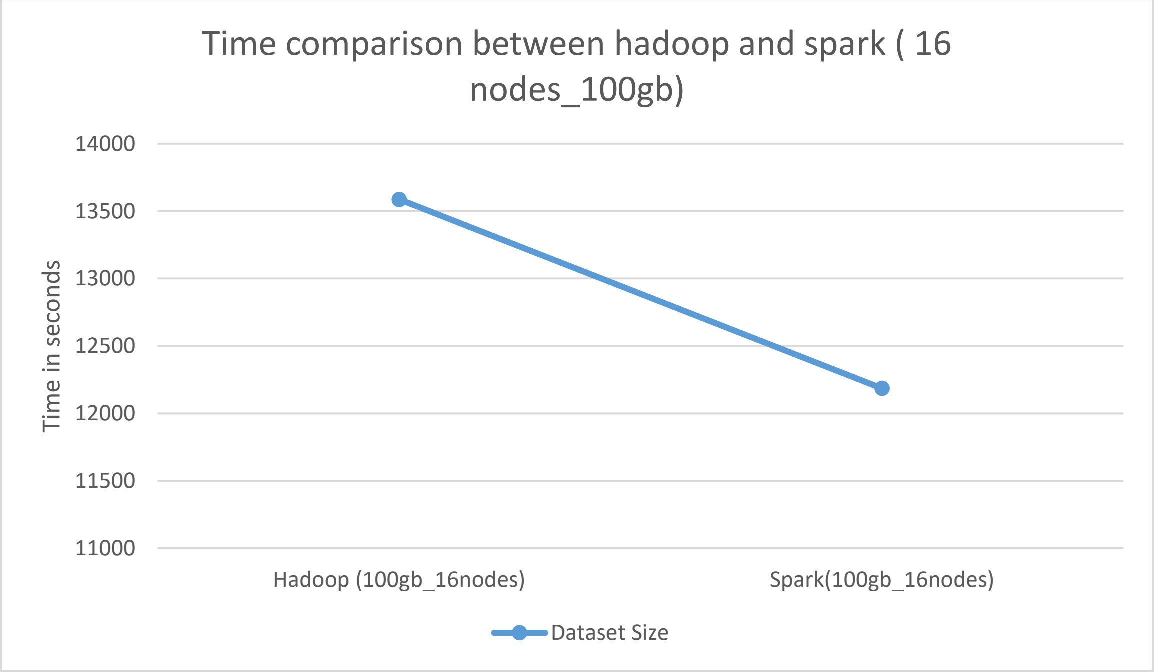
We can observe that the total time required for the completion on 1 node (10gb) is 1383 seconds and on 16 nodes (100gb) is 12127 seconds. Also the throughput value for 1 node is 7.4041 whereas the throughput value for 16 nodes is 8.443. Hence we can say that the time required for the completion of the program on 1 node (10gb) is less than the time required for the completion of the 16 nodes (100gb).

- **Time Comparison between Hadoop 10gb (1 node) to Spark 10gb (1 node) :**



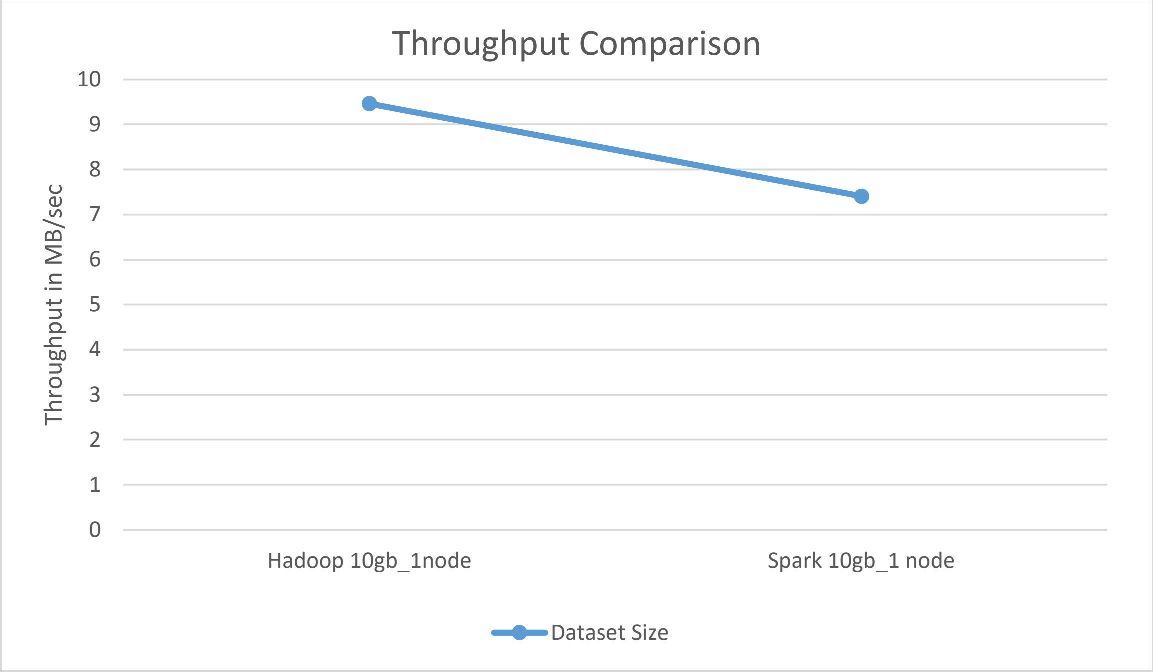
We can observe that the total time required for the completion on Hadoop 1 node (10gb) is 1082 seconds and on Spark 1 node (10gb) is 1383 seconds. As a result we can say that Hadoop performs better for sorting 10gb file on 1 node when compared with Spark.

- **Time Comparison between Hadoop 100gb (16 nodes) and Spark 100gb (16 nodes) :**



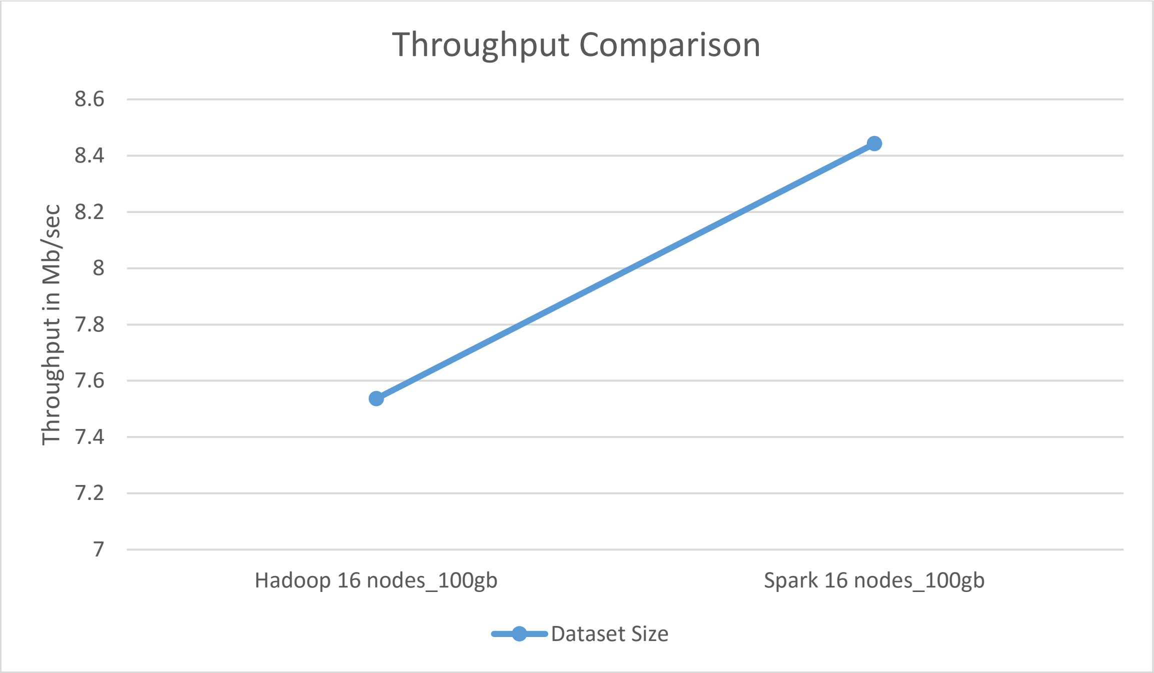
We can observe that the total time required for the completion on Hadoop 16 nodes (100gb) is 13587 seconds and on Spark 16 nodes (100gb) is 12127 seconds. Hence we can say that Spark performs better for sorting 100gb file on 16 nodes when compared with Hadoop.

- **Throughput comparison between Hadoop 1 node (10gb) and Spark 1 node (10gb) :**



We could notice that the throughput value for Hadoop 1 node (10gb) is 9.4639 MB/sec and for Spark 1 node (10gb) is 7.4041 MB/sec. Here we can say that the throughput value for Spark 1 node (10gb) is less than the Hadoop 1 node (10gb).

- **Throughput comparison between Hadoop 16 nodes (100gb) and Spark 16 nodes (100gb) :**



Here the throughput value for Hadoop 16 nodes (100gb) is 7.5366 MB/sec and for Spark 16 nodes (100gb) is 8.443 MB/sec. Here we can say that the throughput value for Spark 16 nodes (100gb) is more than the Hadoop 16 nodes (100gb).

- **Comparison between Shared Memory and Hadoop and Spark:**

  From the above analysis we can notice that the time taken for shared memory on a single node to sort 10gb for all the threads (1,2,4,8) is more than the time taken by Hadoop or Spark on a single node to sort 10gb.

Conclusions :

- Spark processes data in memory and does not deal with Hadoop's MapReduce. This makes it much faster when compared to Hadoop.
- Spark can run as a standalone or on top of Hadoop YARN, where it can read data directly from HDFS.
- But Spark performs better when all the data fits in memory.
- Because of the above reasons we can say that :
- ○ Hadoop seems to be best at one node scale.
- ○ Spark seems to deliver best performance on 16 node scales.
- ○ Spark performs better on 100 and 1000 node scales.