**LOG FILE**

1. **Changing the number of Nodes:**

The experiment was conducted by changing the number of nodes from 10 to 16

Note: Initial experiment with 0 task nodes is not recording here.

When I increased the number of nodes from 10 to 16, I could save around 1 min 30 seconds with same data file (same number of input same number of output, same data size)

* **Input file:** fhv\_tripdata\_2018-01.csv

**Size:**1.3 GB

**Number of Instances:** 19808095

**Number of OutPuts:** 1385

**Time Elapsed:** 2 MIN 35 SEC

**Number of Task Nodes:** 16

Map-Reduce Framework

Map input records=19808095

Map output records=19808095

Map output bytes=297569319

Map output materialized bytes=114503

Input split bytes=1595

Combine input records=19808095

Combine output records=13291

Reduce input groups=1385

Reduce shuffle bytes=114503

Reduce input records=13291

Reduce output records=1385

Spilled Records=26582

Shuffled Maps =22

Failed Shuffles=0

Merged Map outputs=22

GC time elapsed (ms)=23802

CPU time spent (ms)=928670



--------------------------------------------------------------------------------------------------------------------------------

* **Input file:** fhv\_tripdata\_2018-01.csv

**Size:**1.3 GB

**Number of Instances:** 19808095

**Number of OutPuts:** 1385

**Time Elapsed:** 4 min 1 sec

**Number of Task Nodes:** 10



1. **Changing the Size of data:**

Experiment conducted by inputting 3 sized files (1.3,2.6 and 4) compared the efficiency.

* **Input file:** fhv\_tripdata\_2018-02.csv& fhv\_tripdata\_2018-03.csv

(Merge two file and create new file)

**Size:**2.6 GB

**Number of Instances:** 39158789

**Number of OutPuts:** 1419

**Time Elapsed:** 4 MIN 3 SEC

**Number of Task Nodes:** 16

Map-Reduce Framework

Map input records=39158789

Map output records=39158789

Map output bytes=588251856

Map output materialized bytes=142649

Input split bytes=2730

Combine input records=39158789

Combine output records=16392

Reduce input groups=1419

Reduce shuffle bytes=142649

Reduce input records=16392

Reduce output records=1419

Spilled Records=32784

Shuffled Maps =42

Failed Shuffles=0

Merged Map outputs=42

GC time elapsed (ms)=49226

CPU time spent (ms)=1819390



* **Input file:** fhv\_tripdata\_2018-03.csv,fhv\_tripdata\_2018-04.csv & fhv\_tripdata\_2018-05.csv

(Merge two file and create new file)

**Size:**4 GB

**Number of Instances:** 61144039

**Number of OutPuts:** 1448

**Time Elapsed:** 4 MIN 8 SEC

**Number of Task Nodes:** 16

Map-Reduce Framework

Map input records=61144039

Map output records=61144039

Map output bytes=918354291

Map output materialized bytes=161963

Input split bytes=4323

Combine input records=61144039

Combine output records=18414

Reduce input groups=1448

Reduce shuffle bytes=161963

Reduce input records=18414

Reduce output records=1448

Spilled Records=36828

Shuffled Maps =66

Failed Shuffles=0

Merged Map outputs=66

GC time elapsed (ms)=85144

CPU time spent (ms)=2958770

Physical memory (bytes) snapshot=12360343552

Virtual memory (bytes) snapshot=86643335168

Total committed heap usage (bytes)=9051639808

Shuffle Errors



1. **With and without partitioner and combiner**

**-**with partitioner and combiner

**Input file:** fhv\_tripdata\_2018-03.csv,fhv\_tripdata\_2018-04.csv & fhv\_tripdata\_2018-05.csv

(Merge two file and create new file)

**Size:**4 GB

**Number of Instances:** 61144039

**Number of OutPuts:** 1448

**Time Elapsed:** 4 MIN 8 SEC

**Number of Task Nodes:** 16

**-**without partitioner and combiner

**Input file:** fhv\_tripdata\_2018-03.csv,fhv\_tripdata\_2018-04.csv & fhv\_tripdata\_2018- 05.csv

(Merge two file and create new file)

**Size:**4 GB

**Number of Instances:** 61144039

**Number of OutPuts:** 1448

**Time Elapsed:** 6 MIN 6SEC

**Number of Task Nodes:** 16