Kathia Teran
CSCI 636 – M02
Spring 2020

Assignment 1

A. MapReduce Exercise
For the following problem describe how you would solve it using map-reduce.

The input is a list of housing data where each input record contains information about a single house: (address, city, state, zip, value).
The output should be the average house value in each zip code.

Answer:

- Input is mapped:

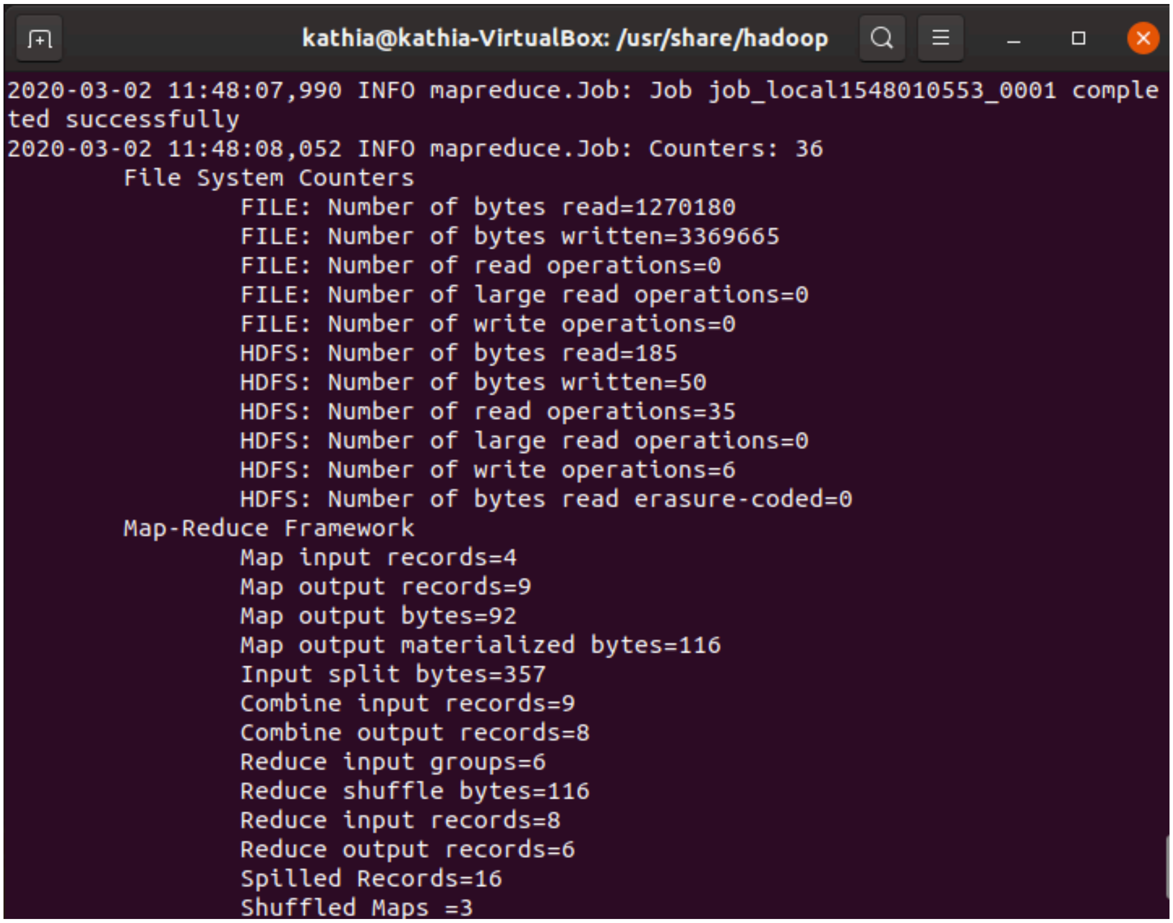
1. User function gets called for each key and value pair.
2. Input gets partitioned into *intermediate M* shards constituted of key and value pairs.
3. Unnecessary data gets discarded, and key and value sets get generated.
4. A framework groups all the intermediate values with the intermediate keys, and passes them to the Reduce function.
5. This gets written into an intermediate file.
6. Partitioning stage: the reducers (R) that will handle which keys get identified.
7.

- Input is reduced:

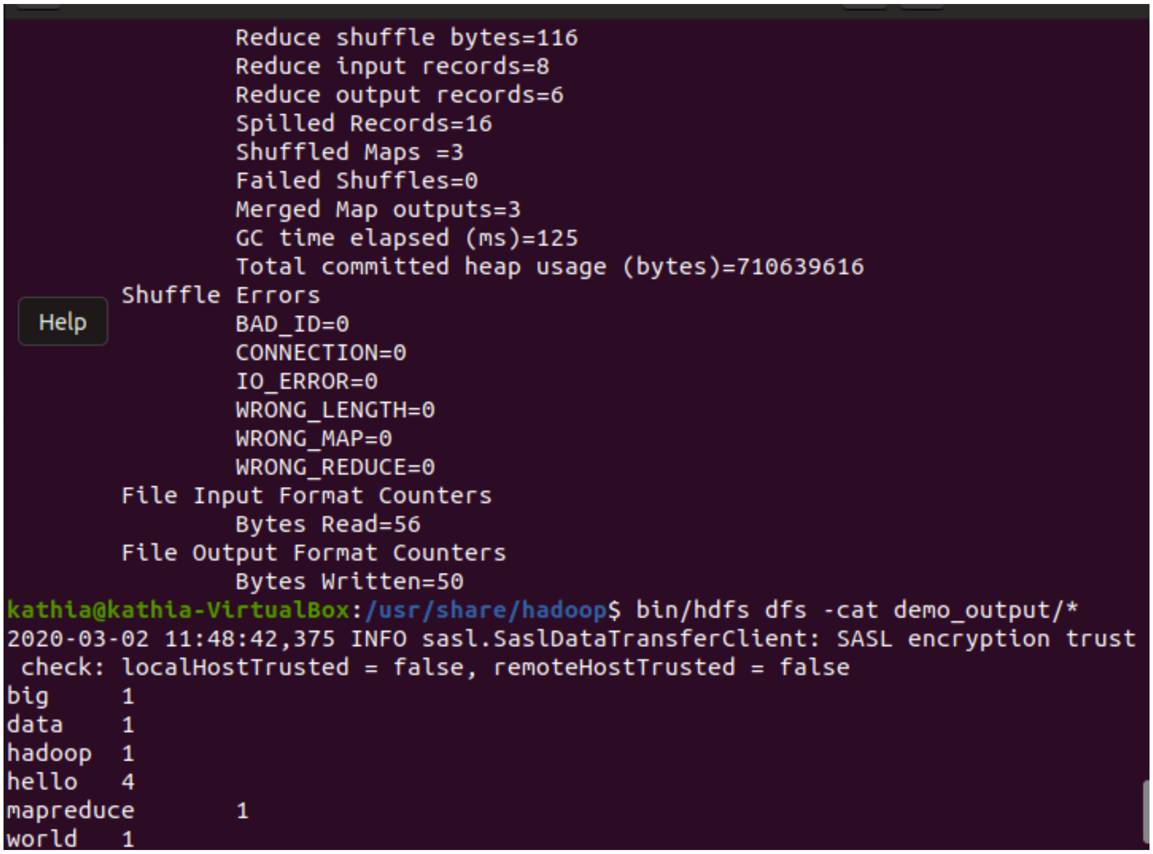
1. The input key and set of values get merged together into a smaller set of values.
2. Partitioning function partitions intermediate key space into R pieces.
3. The user specifies number of partitions (R) and the partitioning function.
4. Shuffle & sort: the relevant partition of the output from all mappers get fetched.
5. Reduce: input becomes the sorted output of mappers
6. The *Reduce* function gets called per key with the list of values for that key to add the results.

- When all 'map' and 'reduce' tasks have completed, the user program gets woken by the master.
- Then *MapReduce* gets called in the user program and resumes execution, where the output is available in *R* output files.

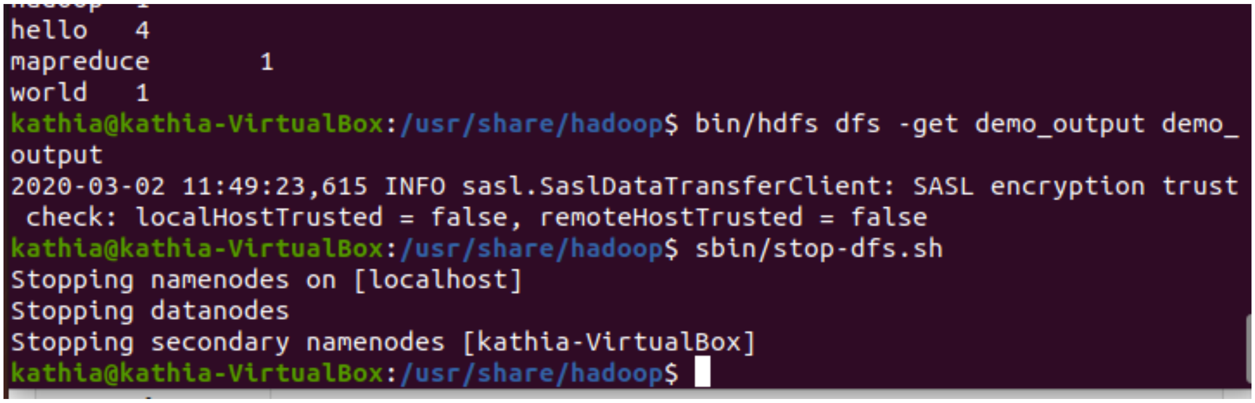
B. Run wordcount exercise you did before on **Hadoop Pseudo-Distributed Operation**. Take few screenshots on each important step. (next page)



```
2020-03-02 11:48:07,990 INFO mapreduce.Job: Job job_local1548010553_0001 comple
ted successfully
2020-03-02 11:48:08,052 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=1270180
                FILE: Number of bytes written=3369665
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=185
                HDFS: Number of bytes written=50
                HDFS: Number of read operations=35
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=4
                Map output records=9
                Map output bytes=92
                Map output materialized bytes=116
                Input split bytes=357
                Combine input records=9
                Combine output records=8
                Reduce input groups=6
                Reduce shuffle bytes=116
                Reduce input records=8
                Reduce output records=6
                Spilled Records=16
                Shuffled Maps =3
```



```
                Reduce shuffle bytes=116
                Reduce input records=8
                Reduce output records=6
                Spilled Records=16
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=125
                Total committed heap usage (bytes)=710639616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=56
        File Output Format Counters
                Bytes Written=50
kathia@kathia-VirtualBox:/usr/share/hadoop$ bin/hdfs dfs -cat demo_output/*
2020-03-02 11:48:42,375 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
big     1
data    1
hadoop  1
hello   4
mapreduce       1
world   1
```



```
hadoop  1
hello   4
mapreduce       1
world   1
kathia@kathia-VirtualBox:/usr/share/hadoop$ bin/hdfs dfs -get demo_output demo_
output
2020-03-02 11:49:23,615 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
kathia@kathia-VirtualBox:/usr/share/hadoop$ sbin/stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [kathia-VirtualBox]
kathia@kathia-VirtualBox:/usr/share/hadoop$ 
```

Overview 'localhost:9000' (active)

Started:	Mon Mar 02 11:19:21 -0500 2020
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 11:56:00 -0400 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-6d131e77-92c3-46bb-83a6-9979d942f47f
Block Pool ID:	BP-2074698690-127.0.1.1-1583165897851

Summary

Security is off.

Safemode is off.

4 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 4 total filesystem object(s).

Heap Memory used 37.29 MB of 62.3 MB Heap Memory. Max Heap Memory is 349.94 MB.

Non Heap Memory used 51.26 MB of 54.88 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	117.61 GB
Configured Remote Capacity:	0 B
DFS Used:	28 KB (0%)
Non DFS Used:	12.4 GB
DFS Remaining:	99.2 GB (84.34%)
Block Pool Used:	28 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)

Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Mon Mar 02 11:19:21 -0500 2020
Last Checkpoint Time	Mon Mar 02 11:18:22 -0500 2020
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 4

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-kathia/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-kathia/dfs/name/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/tmp/hadoop-kathia/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	117.61 GB	28 KB (0%)	99.2 GB (84.34%)	28 KB	1

Hadoop,
2019.