Firstly starting docker container with Ubuntu

Then connecting to client

STEP 3

```
# jar -cvf units.jar -C units/ .
added manifest
adding: hadoop/(in = 0) (out= 0)(stored 0%)
adding: hadoop/ProcessUnits$E_EReduce.class(in = 1671) (out= 686)(deflated 58%)
adding: hadoop/ProcessUnits$E_EMapper.class(in = 1898) (out= 775)(deflated 59%)
adding: hadoop/ProcessUnits.class(in = 1567) (out= 768)(deflated 50%)
```

INSTALLING HADOOP

```
# wget https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
--2020-12-22 20:21:55--  https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 359196911 (343M) [application/x-gzip]
Saving to: 'hadoop-3.2.1.tar.gz'

hadoop-3.2.1.tar.gz          100%[=============================================>] 342.56M  4.98MB/s    in 75s

2020-12-22 20:23:10 (4.57 MB/s) - 'hadoop-3.2.1.tar.gz' saved [359196911/359196911]
```

```
# export HADOOP_HOME=/home/hadoop/hadoop-3.2.1
# export HADOOP_INSTALL=$HADOOP_HOME
# export HADOOP_MAPRED_HOME=$HADOOP_HOME
# export HADOOP_COMMON_HOME=$HADOOP_HOME
# export HADOOP_HDFS_HOME=$HADOOP_HOME
# export YARN_HOME=$HADOOP_HOME
# export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
# export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
# export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

STEP 6

```
# $HADOOP_HOME/bin/hadoop fs -ls input_dir/
Found 1 items
-rw-r--r--   1 root root        222 2020-12-22 20:41 ../input_dir/sample.txt
```

STEP 7

```
# $HADOOP_HOME/bin/hadoop jar units.jar hadoop.ProcessUnits input_dir output_dir
2020-12-22 20:45:02,298 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-12-22 20:45:02,580 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-12-22 20:45:02,580 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-12-22 20:45:02,619 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2020-12-22 20:45:02,804 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2020-12-22 20:45:02,963 INFO mapred.FileInputFormat: Total input files to process : 1
2020-12-22 20:45:03,003 INFO mapreduce.JobSubmitter: number of splits:1
2020-12-22 20:45:03,469 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1476755958_0001
2020-12-22 20:45:03,469 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-12-22 20:45:03,770 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2020-12-22 20:45:03,778 INFO mapreduce.Job: Running job: job_local1476755958_0001
2020-12-22 20:45:03,780 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2020-12-22 20:45:03,784 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2020-12-22 20:45:03,799 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-12-22 20:45:03,799 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2020-12-22 20:45:03,887 INFO mapred.LocalJobRunner: Waiting for map tasks
2020-12-22 20:45:03,898 INFO mapred.LocalJobRunner: Starting task: attempt_local1476755958_0001_m_000000_0
2020-12-22 20:45:03,962 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-12-22 20:45:03,962 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2020-12-22 20:45:04,002 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2020-12-22 20:45:04,029 INFO mapred.MapTask: Processing split: file:/input_dir/sample.txt:0+222
2020-12-22 20:45:04,053 INFO mapred.MapTask: numReduceTasks: 1
2020-12-22 20:45:05,339 INFO mapreduce.Job: Job job_local1476755958_0001 running in uber mode : false
2020-12-22 20:45:05,342 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2020-12-22 20:45:05,342 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2020-12-22 20:45:05,342 INFO mapred.MapTask: soft limit at 83886080
2020-12-22 20:45:05,343 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2020-12-22 20:45:05,343 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2020-12-22 20:45:05,343 INFO mapreduce.Job:  map 0% reduce 0%
2020-12-22 20:45:05,354 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuff
er
```

```
2020-12-22 20:45:05,379 INFO mapred.LocalJobRunner:
2020-12-22 20:45:05,379 INFO mapred.MapTask: Starting flush of map output
2020-12-22 20:45:05,380 INFO mapred.MapTask: Spilling map output
2020-12-22 20:45:05,380 INFO mapred.MapTask: bufstart = 0; bufend = 585; bufvoid = 104857600
2020-12-22 20:45:05,380 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214140(104856560); length = 257/65
53600
2020-12-22 20:45:05,411 INFO mapred.MapTask: Finished spill 0
2020-12-22 20:45:05,434 INFO mapred.Task: Task:attempt_local1476755958_0001_m_000000_0 is done. And is in the process of
 committing
2020-12-22 20:45:05,440 INFO mapred.LocalJobRunner: file:/input_dir/sample.txt:0+222
2020-12-22 20:45:05,440 INFO mapred.Task: Task 'attempt_local1476755958_0001_m_000000_0' done.
2020-12-22 20:45:05,459 INFO mapred.Task: Final Counters for attempt_local1476755958_0001_m_000000_0: Counters: 18
        File System Counters
                FILE: Number of bytes read=3496
                FILE: Number of bytes written=523220
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5
                Map output records=65
                Map output bytes=585
                Map output materialized bytes=61
                Input split bytes=78
                Combine input records=65
                Combine output records=5
                Spilled Records=5
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=229638144
        File Input Format Counters
                Bytes Read=242
2020-12-22 20:45:05,460 INFO mapred.LocalJobRunner: Finishing task: attempt_local1476755958_0001_m_000000_0
2020-12-22 20:45:05,462 INFO mapred.LocalJobRunner: map task executor complete.
2020-12-22 20:45:05,469 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2020-12-22 20:45:05,470 INFO mapred.LocalJobRunner: Starting task: attempt_local1476755958_0001_r_000000_0
2020-12-22 20:45:05,487 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-12-22 20:45:05,487 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2020-12-22 20:45:05,488 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2020-12-22 20:45:05,497 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shu
```

```
        File System Counters
                FILE: Number of bytes read=7146
                FILE: Number of bytes written=1046588
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5
                Map output records=65
                Map output bytes=585
                Map output materialized bytes=61
                Input split bytes=78
                Combine input records=65
                Combine output records=5
                Reduce input groups=5
                Reduce shuffle bytes=61
                Reduce input records=5
                Reduce output records=5
                Spilled Records=10
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=15
                Total committed heap usage (bytes)=510656512
        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=242
        File Output Format Counters
                Bytes Written=87
```

STEP 8

```
# $HADOOP_HOME/bin/hadoop fs -ls output_dir/
Found 2 items
-rw-r--r--   1 root root          0 2020-12-22 20:45 ../output_dir/_SUCCESS
-rw-r--r--   1 root root         75 2020-12-22 20:45 ../output_dir/part-00000
```

STEP 9

```
# $HADOOP_HOME/bin/hadoop fs -cat output_dir/part-00000
1979    24.615385
1980    29.153847
1981    33.615383
1984    39.615383
1985    36.923077
```

STEP 10

```
# $HADOOP_HOME/bin/hadoop fs -cat output_dir/part-00000 > data/results.txt
```

```
# cd data
# ls
ProcessUnits.java  docker-compose.yml  results.txt  sample.txt  units  units.jar
# cat results.txt
1979    24.615385
1980    29.153847
1981    33.615383
1984    39.615383
1985    36.923077
# _
```