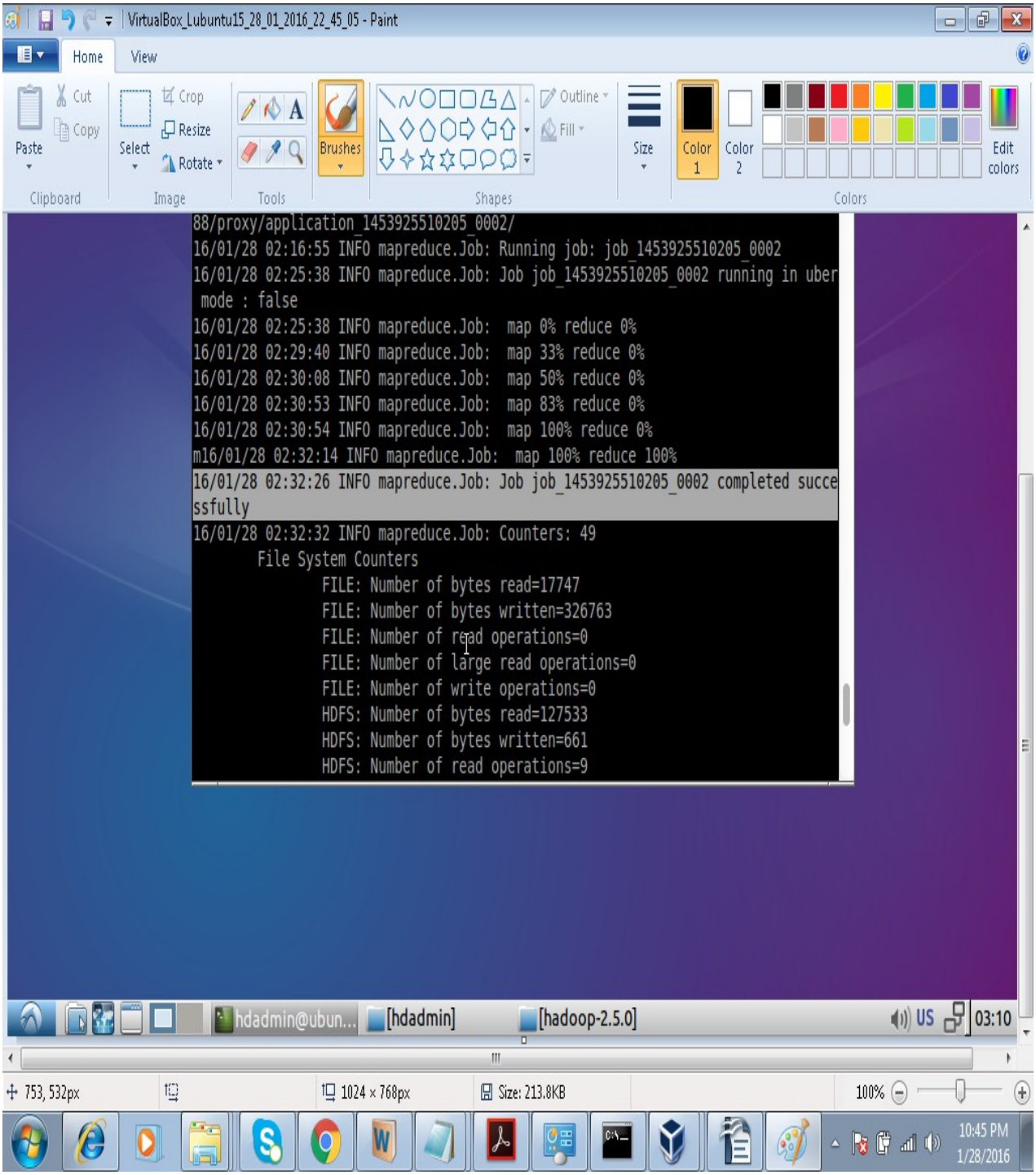
HADOOP STREAMING - UBUNTU



Time taken by job without hadoop streaming

image1 - Windows Photo Viewer

File Print E-mail Burn Open

hdadmin@ubuntu: ~/jars

File Edit Tabs Help

```
16/01/28 02:32:26 INFO mapreduce.Job: Job job_1453925510205_0002 completed succe
ssfully
16/01/28 02:32:32 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=17747
                FILE: Number of bytes written=326763
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=127533
                HDFS: Number of bytes written=661
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=640779
                Total time spent by all reduces in occupied slots (ms)=30894
                Total time spent by all map tasks (ms)=640779
                Total time spent by all reduce tasks (ms)=30894
                Total vcore-seconds taken by all map tasks=640779
                Total vcore-seconds taken by all reduce tasks=30894
                Total megabyte-seconds taken by all map tasks=656157696
                Total megabyte-seconds taken by all reduce tasks=31635456
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Map output bytes=15743
                Map output materialized bytes=17753
                Input split bytes=178
                Combine input records=0
                Combine output records=0
                Reduce input groups=58
                Reduce shuffle bytes=17753
                Reduce input records=999
                Reduce output records=58
                Spilled Records=1998
                Shuffled Maps =2
```

hdadmin... [hdadmin] [hadoop-... New Tab - ... New Tab - ... Namenod... US 03:13

VirtualBox_Lubuntu15_28_01_2016_22_48_223 - Windows Photo Viewer

File Print E-mail Burn Open

hdadmin@ubuntu: ~/jars

File Edit Tabs Help

```
            Total time spent by all map tasks (ms)=640779
            Total time spent by all reduce tasks (ms)=30894
            Total vcore-seconds taken by all map tasks=640779
            Total vcore-seconds taken by all reduce tasks=30894
            Total megabyte-seconds taken by all map tasks=656157696
            Total megabyte-seconds taken by all reduce tasks=31635456
        Map-Reduce Framework
            Map input records=999
            Map output records=999
            Map output bytes=15743
            Map output materialized bytes=17753
            Input split bytes=178
            Combine input records=0
            Combine output records=0
            Reduce input groups=58
            Reduce shuffle bytes=17753
            Reduce input records=999
            Reduce output records=58
            Spilled Records=1998
            Shuffled Maps =2
            Failed Shuffles=0
            Merged Map outputs=2
            GC time elapsed (ms)=9724
            CPU time spent (ms)=6170
            Physical memory (bytes) snapshot=350498816
            Virtual memory (bytes) snapshot=1979031552
            Total committed heap usage (bytes)=250212352
        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=127355
        File Output Format Counters
            Bytes Written=661
hdadmin@ubuntu:~/jars$ hadoop jar sales.jar mapreduce.SalesCountryDriver /input/sales.csv /output
```

hdadmin... [hdadmin] [hadoop-... New Tab - ... New Tab - ... Namenod... US 03:13

10:49 PM
1/28/2016

VirtualBox_Lubuntu15_29_01_2016_00_07_17 - Windows Photo Viewer

File Print E-mail Burn Open

Namenode informat | All Applications

localhost:8088/cluster

Apps | Namenode inforr | All Applications

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes
2	0	0	2	0	0 B	8 GB	0 B	0	8	0	1	0

Show 20 entries

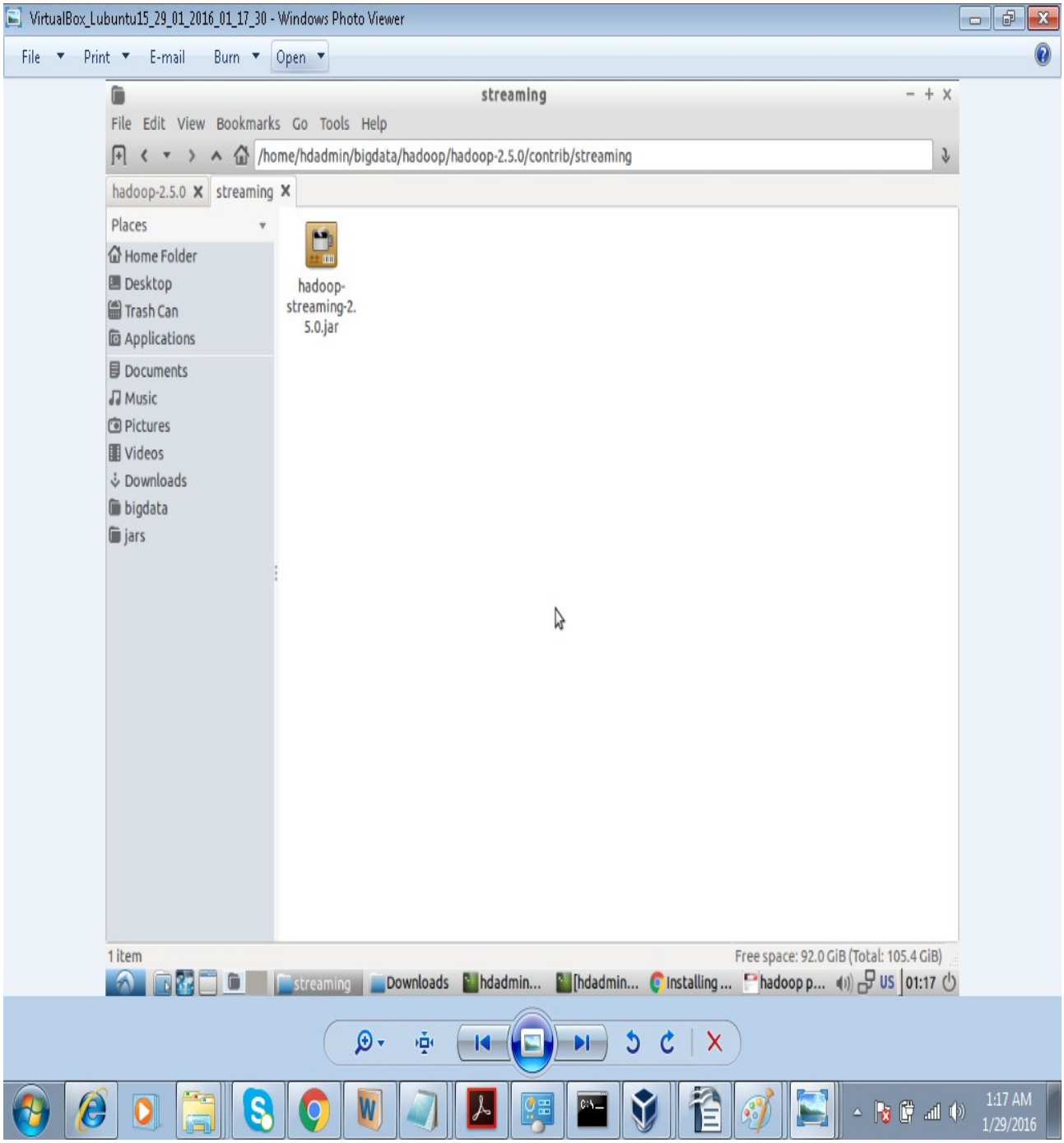
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus
application_1453925510205_0002	hdadmin	SalePerCountry	MAPREDUCE	default	Wed, 27 Jan 2016 20:46:55 GMT	Wed, 27 Jan 2016 21:02:23 GMT	FINISHED	SUCCEEDED
application_1453925510205_0001	hdadmin	SalePerCountry	MAPREDUCE	default	Wed, 27 Jan 2016 20:18:34 GMT	Wed, 27 Jan 2016 20:43:28 GMT	FINISHED	SUCCEEDED

Showing 1 to 2 of 2 entries

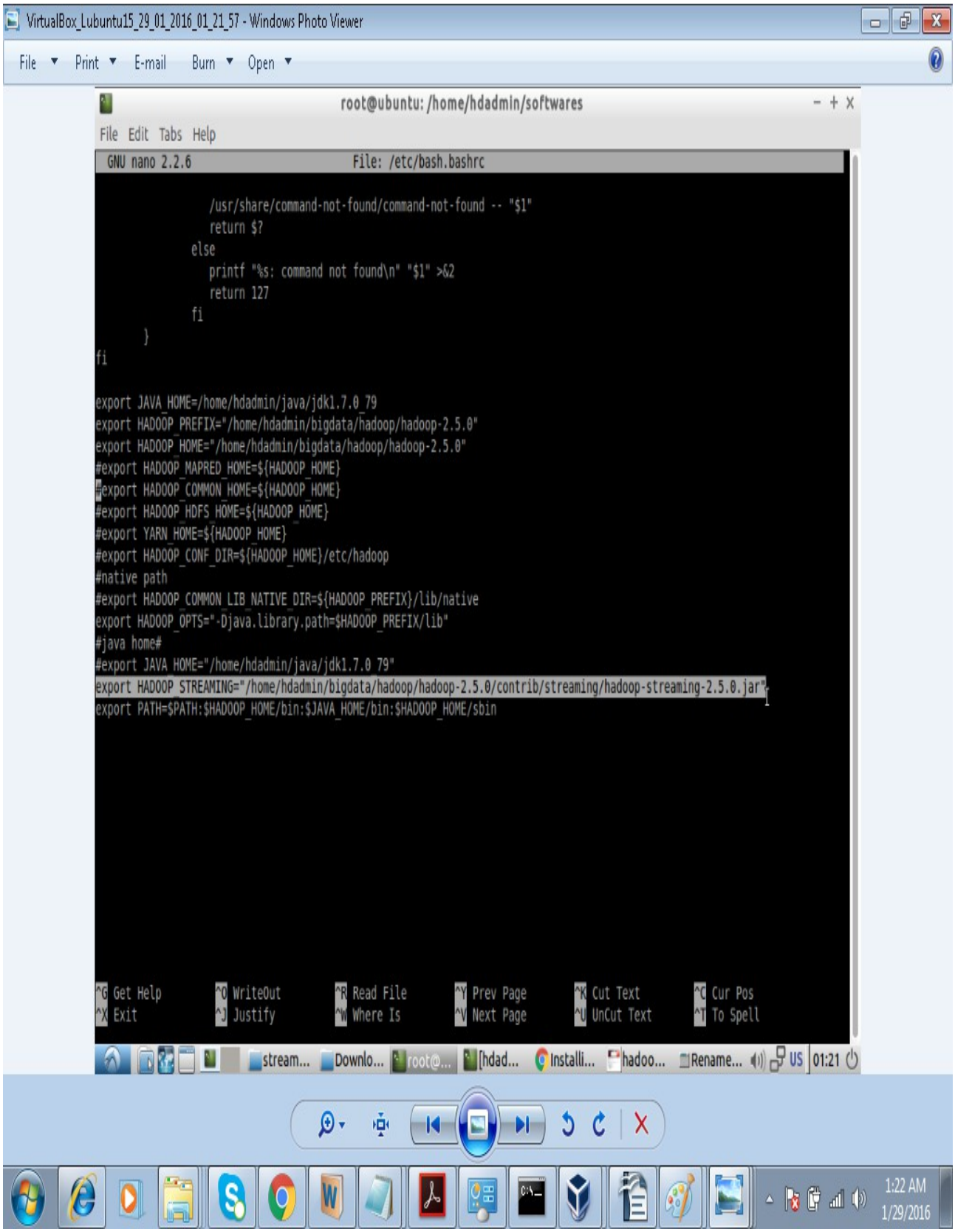
About Apache Hadoop

[hdadmin... [hdadmin] [hadoop-... New Tab - ... [New Tab ... All Applic... US 04:32

Create a folder /contrib/streaming inside /hadoop/hadoop-2.5.0



change the environment variable add HADOOP_STREAMING
export HADOOP_STREAMING="/home/hdadmin/hadoop/hadoop-2.5.0/contrib/streaming/hadoop-streaming-2.5.0.jar"

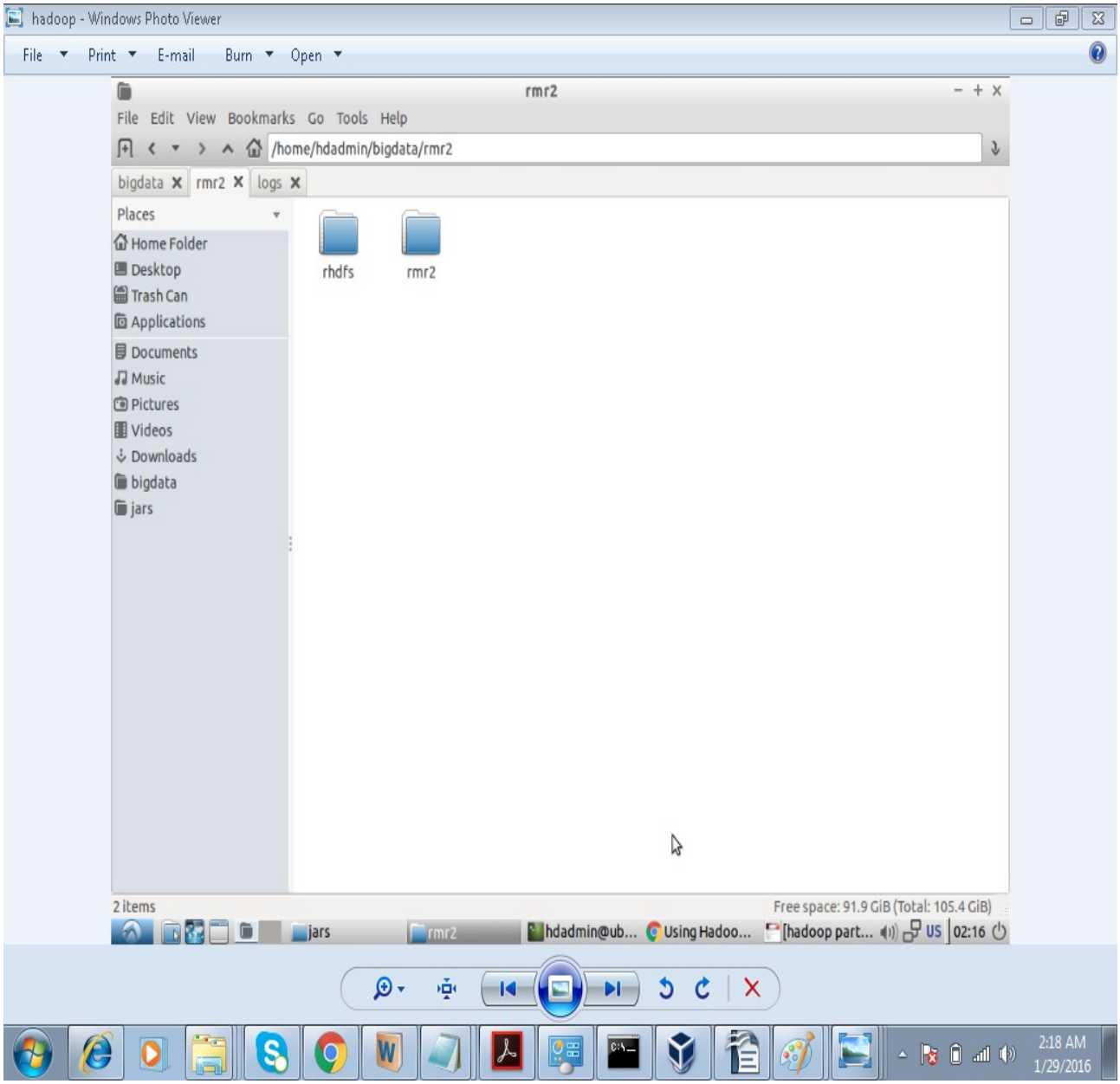


NOTE:

Always format namenode (hadoop namenode -format) only after stopping hadoop.
If format namenode after starting it shows error.

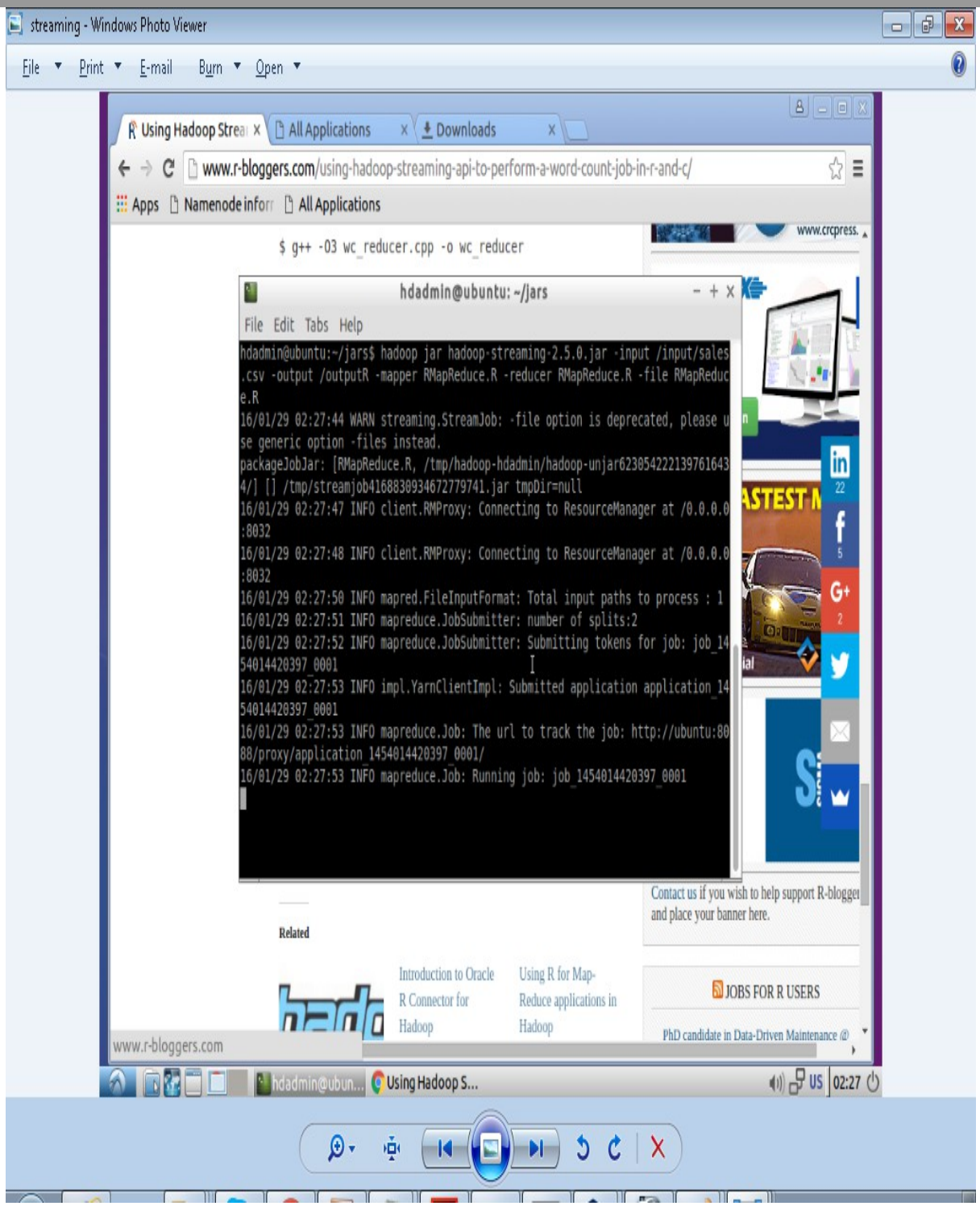
Also dont start in more than 1 terminal in such case it starts 2 datanodes or namenode and show port already in use 50070 error.

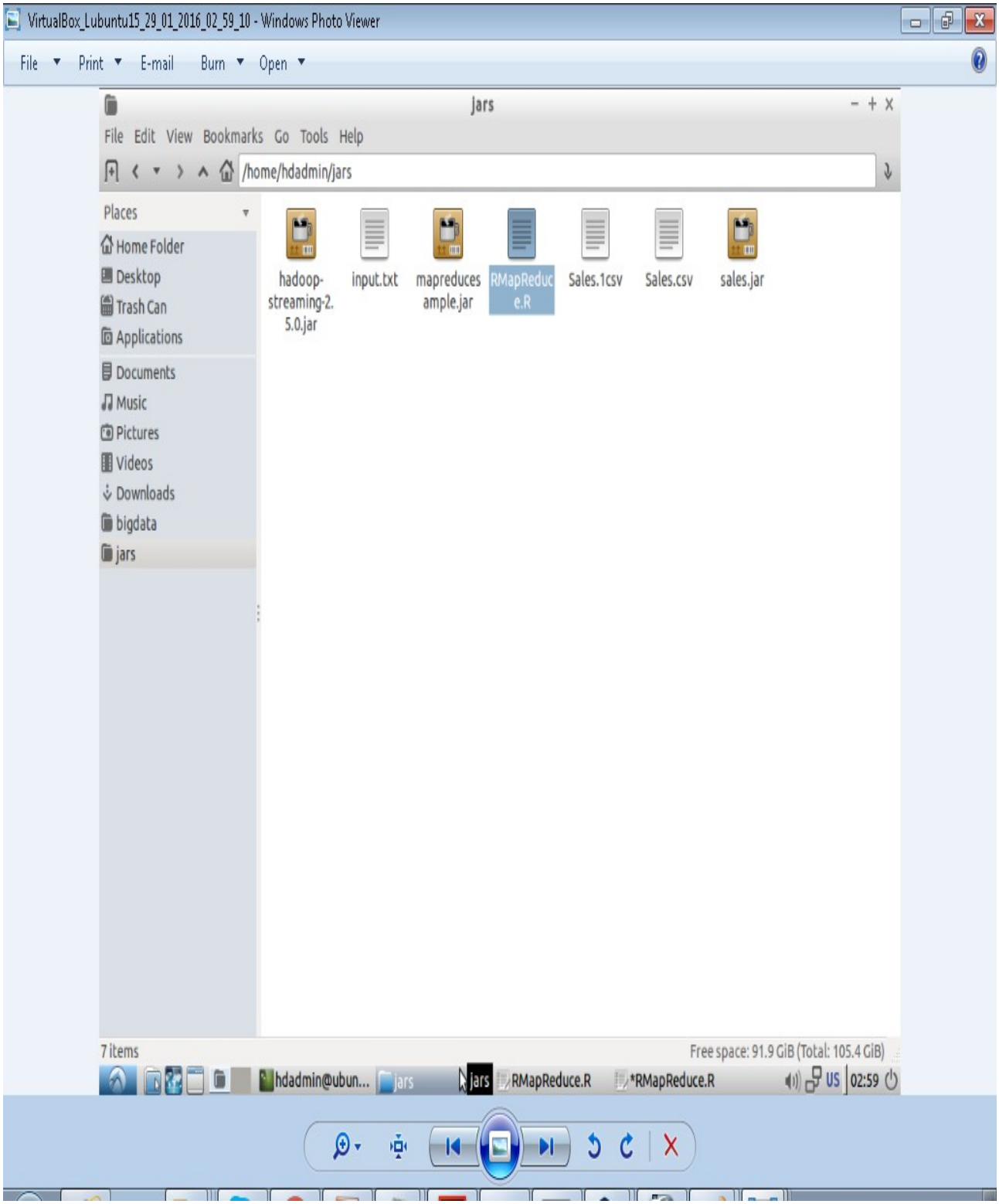
Install rmr2, rhdfs:



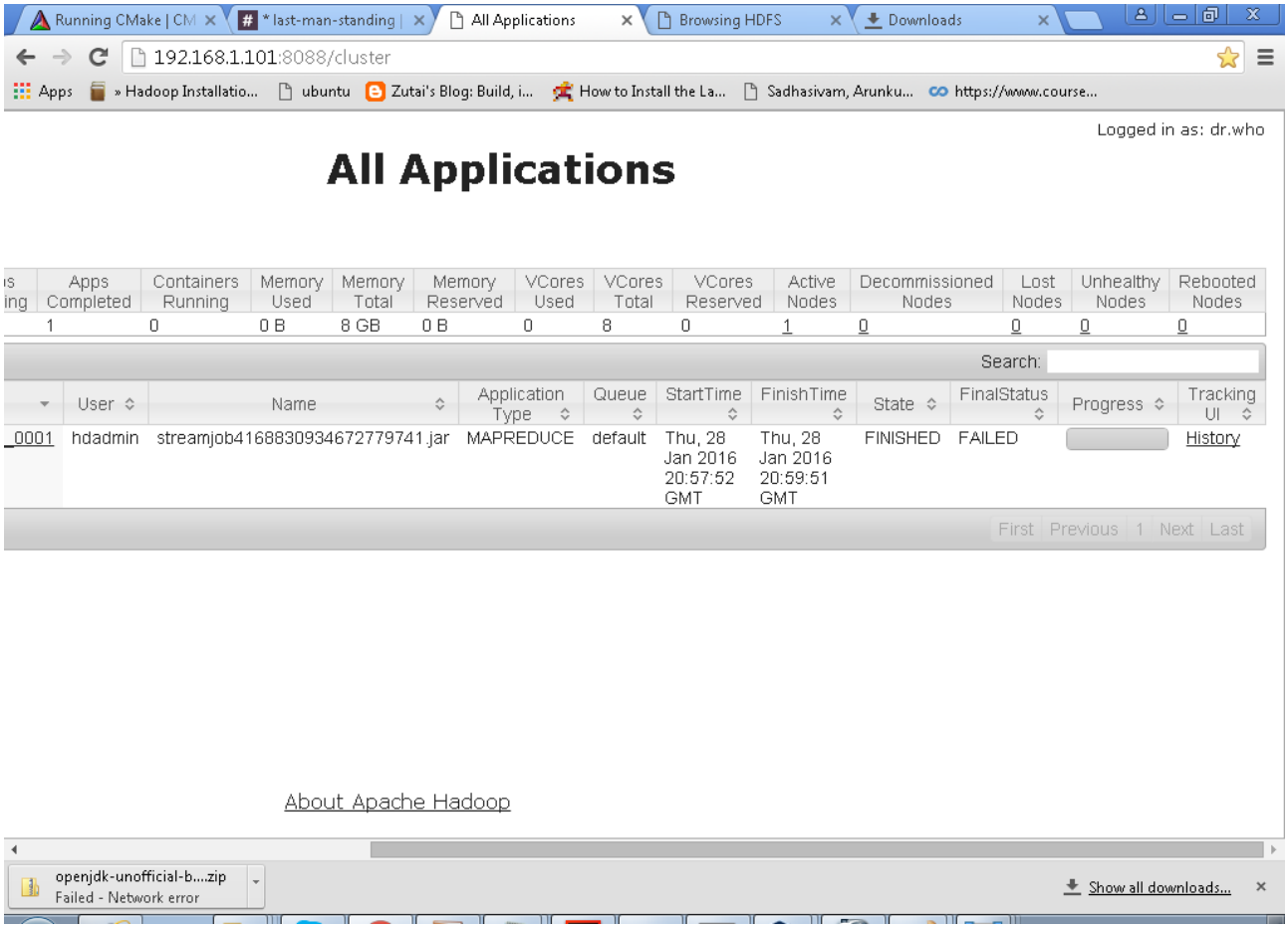
rmr2- r packages
rhdfs – to connect to hdfs.

```
hdadmin@ubntu:~/jars$ hadoop jar hadoop-streaming-2.5.0.jar
 -input /input/sales.csv -output /outputR – mapper RmapReduce.R
 -reducer Rmapreduce.R -file Rmapreduce.R
```



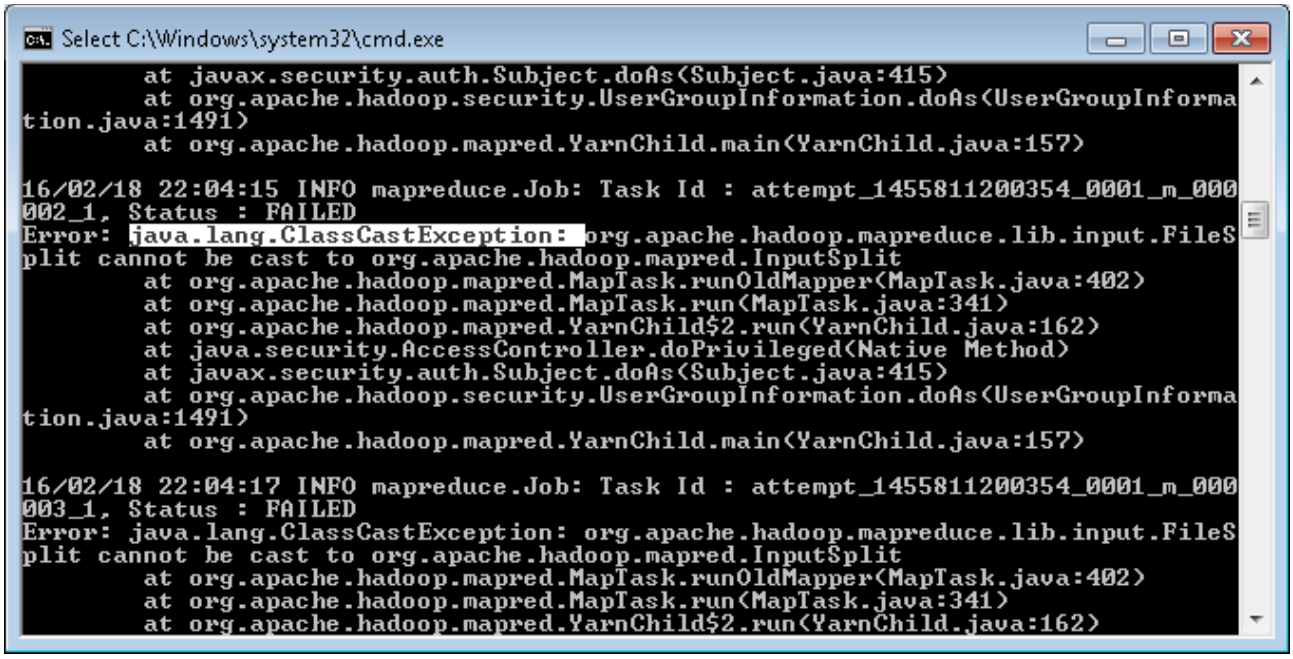

VirtualBox_Lubuntu15_29_01_2016_02_59_10 - Windows Photo Viewer
File
Print
E-mail
Burn
Open
jars
File Edit View Bookmarks Go Tools Help
/home/hdadmin/jars
Places
Home Folder
Desktop
Trash Can
Applications
Documents
Music
Pictures
Videos
Downloads
bigdata
jars
hadoop-streaming-2.5.0.jar
input.txt
mapreduces ample.jar
RMapReduce.R
Sales.1csv
Sales.csv
sales.jar
7 items
Free space: 91.9 GiB (Total: 105.4 GiB)
hdadmin@ubun...
jars
RMapReduce.R
*RMapReduce.R
US
02:59

Sample output



HADOOP STREAMING - WINDOWS

C:\HADOOPOUTPUT>yarn jar %HADOOP_HOME%/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.2.0.jar pi 16 10000



ERROR:

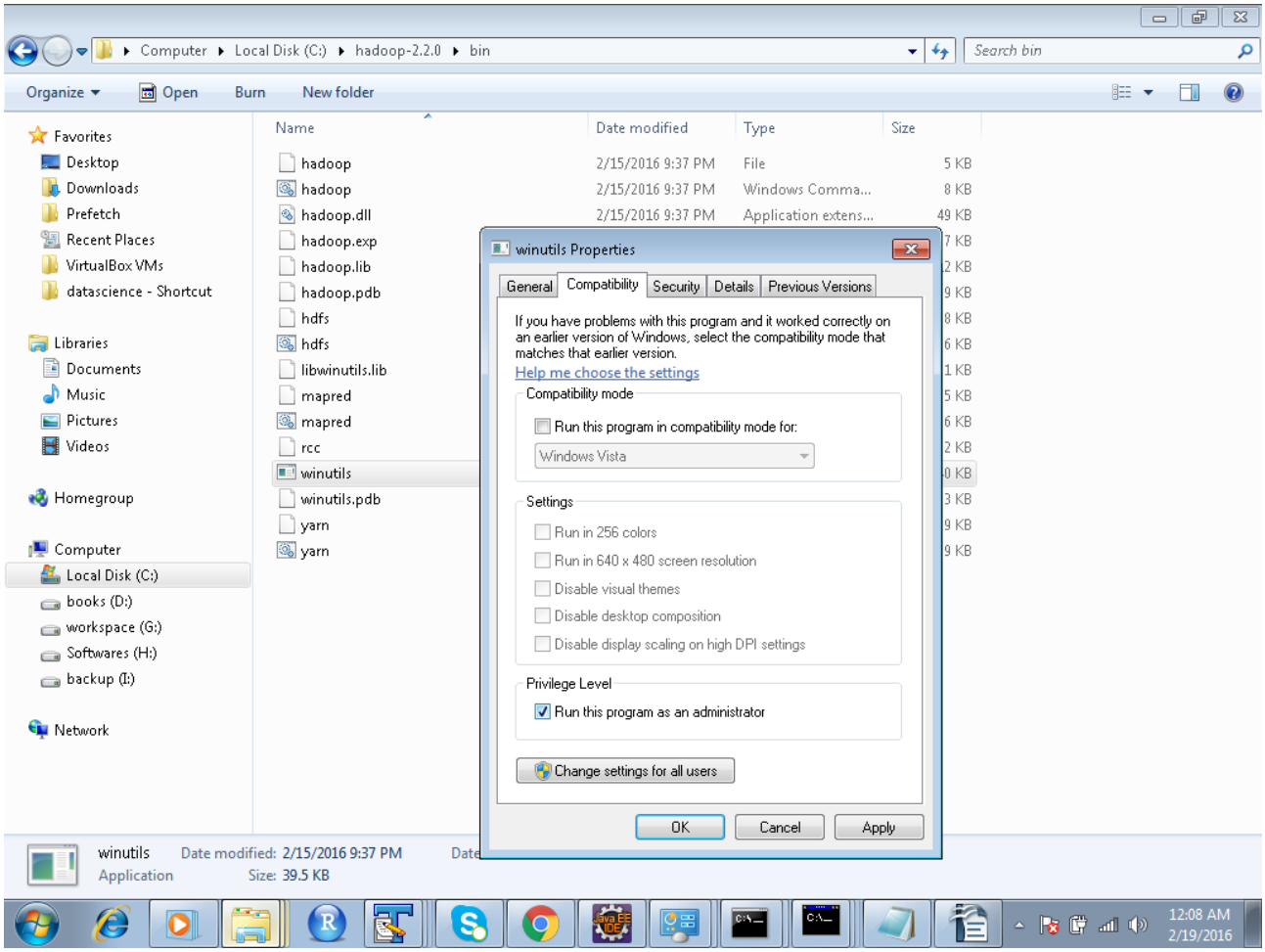
2016-02-14 12:13:27,999 INFO [IPC Server handler 0 on 49996] org.apache.hadoop.mapred.TaskAttemptListenerImpl: Status update from attempt_1455429464529_0001_m_000000_0
2016-02-14 12:13:27,999 INFO [IPC Server handler 0 on 49996] org.apache.hadoop.mapred.TaskAttemptListenerImpl: Progress of TaskAttempt attempt_1455429464529_0001_m_000000_0 is : 0.0
2016-02-14 12:13:28,012 FATAL [IPC Server handler 2 on 49996] org.apache.hadoop.mapred.TaskAttemptListenerImpl: Task: attempt_1455429464529_0001_m_000000_0 - exited : java.lang.ClassCastException: org.apache.hadoop.mapreduce.lib.input.FileSplit cannot be cast to org.apache.hadoop.mapred.InputSplit
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:402)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:415)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)

2016-02-14 12:13:28,025 INFO [AsyncDispatcher event handler]

org.apache.hadoop.mapreduce.v2.app.job.impl.TaskAttemptImpl: Diagnostics report from attempt_1455429464529_0001_m_000000_0: Error: java.lang.ClassCastException: org.apache.hadoop.mapreduce.lib.input.FileSplit cannot be cast to org.apache.hadoop.mapred.InputSplit

```
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:402)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)
```

2016-02-14 12:13:28,026 INFO [AsyncDispatcher event handler] org.apache.hadoop.mapreduce.v2.app.job.impl.TaskAttemptImpl: attempt_1455429464529_0001_m_000000_0 TaskAttempt Transitioned from RUNNING to FAIL_CONTAINER_CLEANUP
2016-02-14 12:13:28,026 INFO [ContainerLauncher #1] org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl: Processing the event EventType: CONTAINER_REMOTE_CLEANUP for container container_1455429464529_0001_01_000002 taskAttempt attempt_1455429464529_0001_m_000000_0
2016-02-14 12:13:28,026 INFO [ContainerLauncher #1] org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl: KILLING attempt_1455429464529_0001_m_000000_0



seems to be bug:

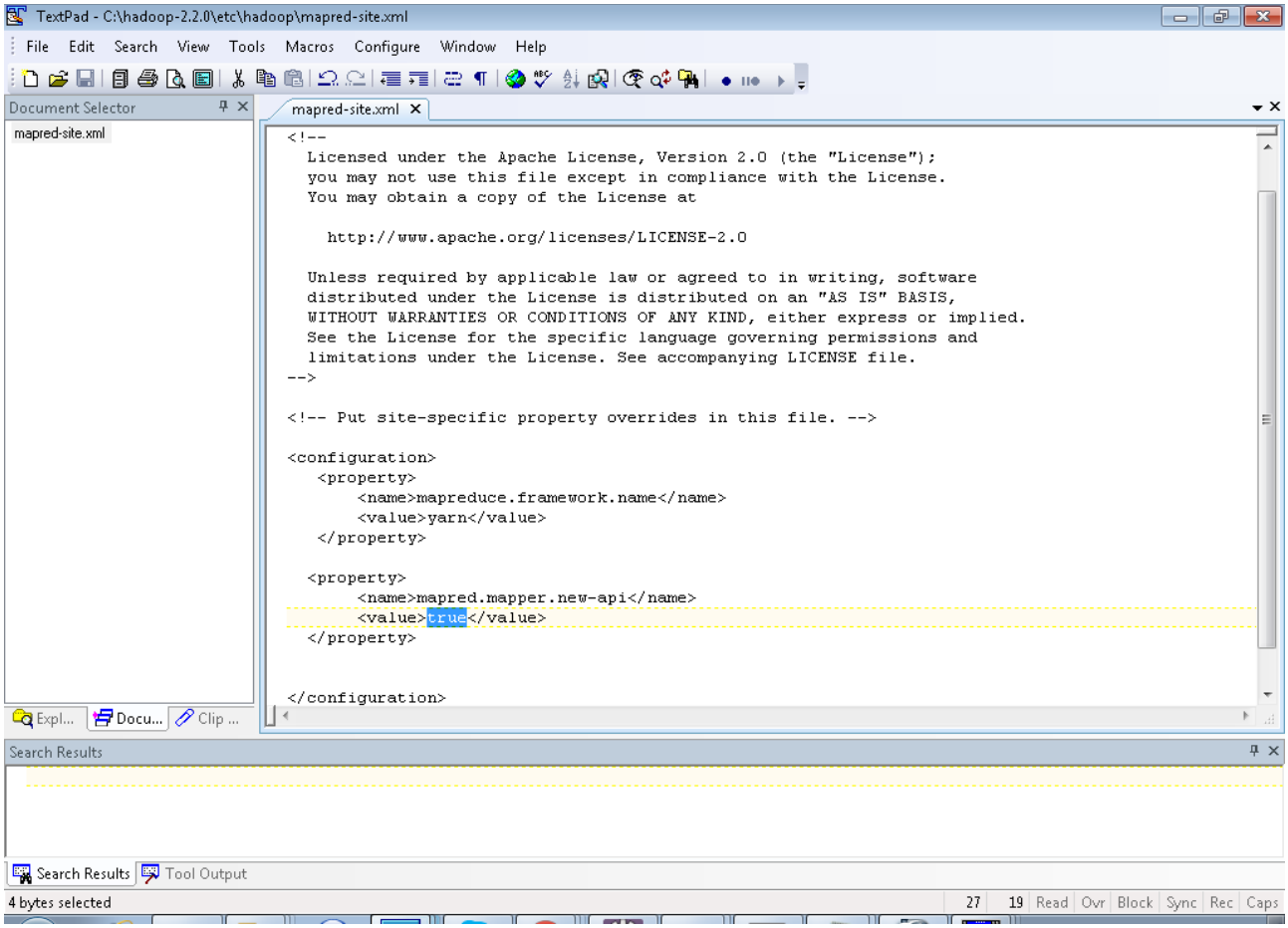
https://issues.apache.org/jira/browse/HADOOP-9110

The reason is that the Map task treated the Map as an old Map for some reason, but it was actually a new one with the new API org.apache.hadoop.mapreduce.lib.input.FileSplit. To solve this problem, you can set the parameter "**mapred.mapper.new-api" to true.**

C:\Windows\system32\cmd.exe

```
088/proxy/application_1455821614868_0001/
16/02/19 00:25:53 INFO mapreduce.Job: Running job: job_1455821614868_0001
16/02/19 00:26:02 INFO mapreduce.Job: Job job_1455821614868_0001 running in uber
 mode : false
16/02/19 00:26:02 INFO mapreduce.Job:  map 0% reduce 0%
16/02/19 00:26:23 INFO mapreduce.Job:  map 31% reduce 0%
16/02/19 00:26:24 INFO mapreduce.Job:  map 38% reduce 0%
16/02/19 00:26:42 INFO mapreduce.Job:  map 56% reduce 0%
16/02/19 00:26:44 INFO mapreduce.Job:  map 63% reduce 0%
16/02/19 00:26:48 INFO mapreduce.Job:  map 69% reduce 21%
16/02/19 00:26:52 INFO mapreduce.Job:  map 69% reduce 23%
16/02/19 00:26:57 INFO mapreduce.Job:  map 75% reduce 23%
16/02/19 00:26:59 INFO mapreduce.Job:  map 81% reduce 25%
16/02/19 00:27:02 INFO mapreduce.Job:  map 88% reduce 25%
16/02/19 00:27:03 INFO mapreduce.Job:  map 100% reduce 27%
16/02/19 00:27:03 INFO mapreduce.Job: Task Id : attempt_1455821614868_0001_r_000
000_0, Status : FAILED
Error: java.lang.NullPointerException
        at org.apache.hadoop.mapred.Task.getFsStatistics(Task.java:347)
        at org.apache.hadoop.mapred.ReduceTask$OldTrackingRecordWriter.<init>(Re
duceTask.java:496)
        at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:432
)
        at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:408)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInforma
tion.java:1491)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)

16/02/19 00:27:04 INFO mapreduce.Job:  map 100% reduce 0%
16/02/19 00:27:10 INFO mapreduce.Job: Task Id : attempt_1455821614868_0001_r_000
000_1, Status : FAILED
Error: java.lang.NullPointerException
        at org.apache.hadoop.mapred.Task.getFsStatistics(Task.java:347)
        at org.apache.hadoop.mapred.ReduceTask$OldTrackingRecordWriter.<init>(Re
duceTask.java:496)
        at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:432
)
        at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:408)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInforma
tion.java:1491)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)

16/02/19 00:27:18 INFO mapreduce.Job: Task Id : attempt_1455821614868_0001_r_000
000_2, Status : FAILED
Error: java.lang.NullPointerException
        at org.apache.hadoop.mapred.Task.getFsStatistics(Task.java:347)
        at org.apache.hadoop.mapred.ReduceTask$OldTrackingRecordWriter.<init>(Re
duceTask.java:496)
        at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:432
)
        at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:408)
```

Worked after changing %hadoop_home%/etc/hadoop/mapred-site.xml



ERROR 2:

2016-02-19 00:27:23,553 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.ShuffleSchedulerImpl: Arun-PC:13562 freed by fetcher#1 in 31ms
2016-02-19 00:27:23,554 INFO [EventFetcher for fetching Map Completion Events] org.apache.hadoop.mapreduce.task.reduce.EventFetcher: EventFetcher is interrupted.. Returning
2016-02-19 00:27:23,560 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: finalMerge called with 16 in-memory map-outputs and 0 on-disk map-outputs
2016-02-19 00:27:23,575 INFO [main] org.apache.hadoop.mapred.Merger: Merging 16 sorted segments
2016-02-19 00:27:23,575 INFO [main] org.apache.hadoop.mapred.Merger: Down to the last merge-pass, with 16 segments left of total size: 1986 bytes
2016-02-19 00:27:23,594 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merged 16 segments, 2146 bytes to disk to satisfy reduce memory limit
2016-02-19 00:27:23,596 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 1 files, 2120 bytes from disk
2016-02-19 00:27:23,596 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2016-02-19 00:27:23,597 INFO [main] org.apache.hadoop.mapred.Merger: Merging 1 sorted segments
2016-02-19 00:27:23,600 INFO [main] org.apache.hadoop.mapred.Merger: Down to the last merge-pass,

with 1 segments left of total size: 2106 bytes
2016-02-19 00:27:23,609 WARN [main] org.apache.hadoop.mapred.YarnChild: Exception running child : java.lang.NullPointerException

```
        at org.apache.hadoop.mapred.Task.getFsStatistics(Task.java:347)
        at org.apache.hadoop.mapred.ReduceTask$OldTrackingRecordWriter.<init>(ReduceTask.java:496)
        at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:432)
        at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:408)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)
```

2016-02-19 00:27:23,612 INFO [main] org.apache.hadoop.mapred.Task: Runnning cleanup for the task
2016-02-19 00:27:23,612 WARN [main] org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter: Output Path is null in abortTask()

```
C:\HADOOPOUTPUT>hdfs dfs -cat /input/wordcount.txt
hi
hi how are you
hadoop
hi how is hadoop
C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.WordCount /input/wordcount.txt /output
```

C:\Windows\system32\cmd.exe - yarn jar mapreduce.jar test.WordCount /input/wordcount.txt /o...

```
C:\HADOOPOUTPUT>hdfs dfs -cat /input/wordcount.txt
hi
hi how are you
hadoop
hi how is hadoop
C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.WordCount /input/wordcount.txt /outp
ut
16/02/19 00:58:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/02/19 00:58:08 INFO input.FileInputFormat: Total input paths to process : 1
16/02/19 00:58:08 INFO mapreduce.JobSubmitter: number of splits:1
16/02/19 00:58:08 INFO Configuration.deprecation: user.name is deprecated. Inste
ad, use mapreduce.job.user.name
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.jar is deprecated. Inst
ead, use mapreduce.job.jar
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.output.value.class is d
eprecated. Instead, use mapreduce.job.output.value.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.mapoutput.value.class i
s deprecated. Instead, use mapreduce.map.output.value.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapreduce.combine.class is dep
recated. Instead, use mapreduce.job.combine.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapreduce.map.class is depreca
ted. Instead, use mapreduce.job.map.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.job.name is deprecated.
 Instead, use mapreduce.job.name
16/02/19 00:58:08 INFO Configuration.deprecation: mapreduce.reduce.class is depr
ecated. Instead, use mapreduce.job.reduce.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapreduce.inputformat.class is
 deprecated. Instead, use mapreduce.job.inputformat.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.input.dir is deprecated
. Instead, use mapreduce.input.fileinputformat.inputdir
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.output.dir is deprecate
d. Instead, use mapreduce.output.fileoutputformat.outputdir
16/02/19 00:58:08 INFO Configuration.deprecation: mapreduce.outputformat.class i
s deprecated. Instead, use mapreduce.job.outputformat.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.output.key.class is dep
recated. Instead, use mapreduce.job.output.key.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.mapoutput.key.class is
deprecated. Instead, use mapreduce.map.output.key.class
16/02/19 00:58:08 INFO Configuration.deprecation: mapred.working.dir is deprecat
ed. Instead, use mapreduce.job.working.dir
16/02/19 00:58:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
55821614868_0004
```

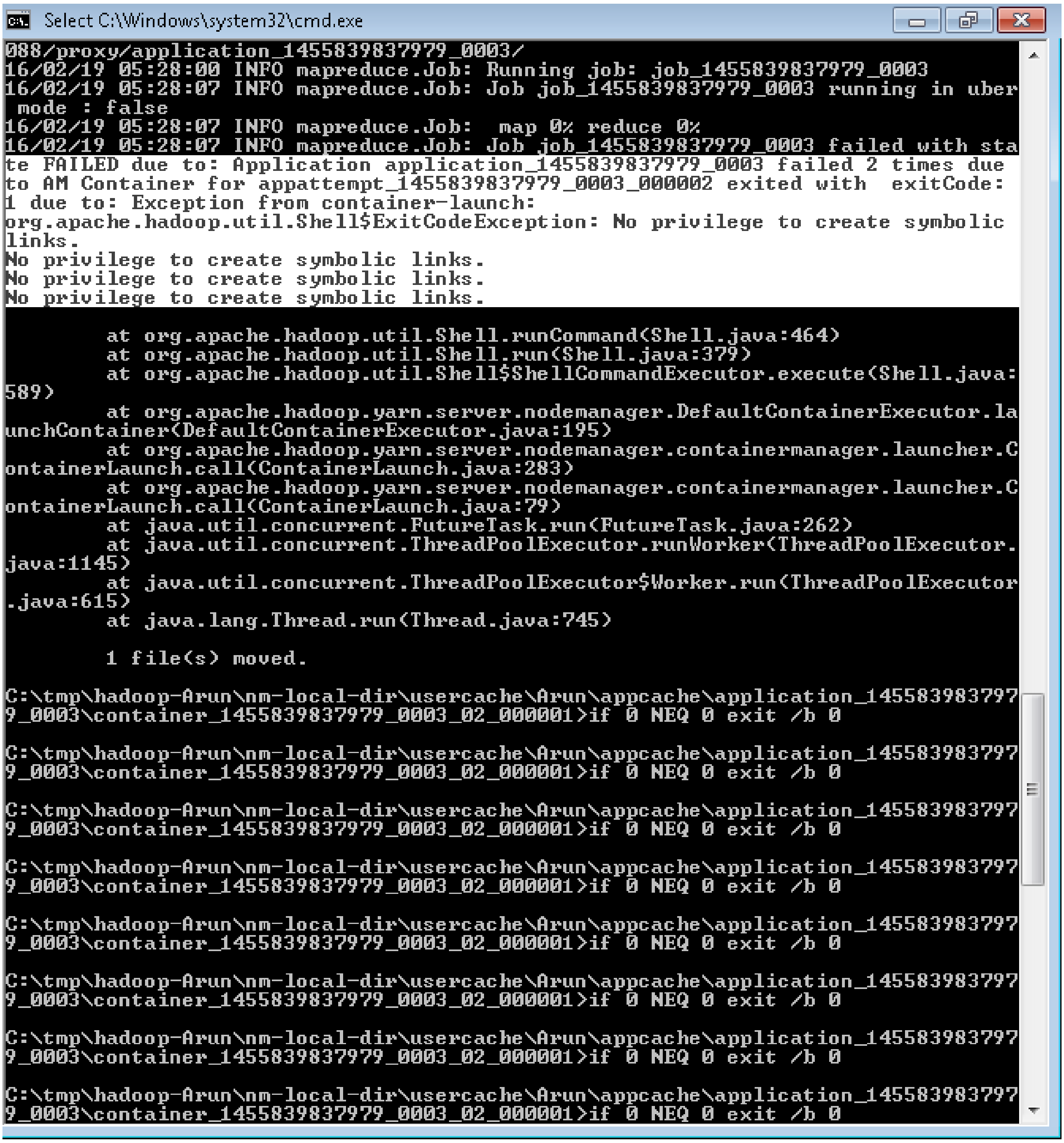
ERROR 3:

2016-02-19 00:27:23,552 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.Fetcher: fetcher#1 about to shuffle output of map attempt_1455821614868_0001_m_000013_0 decomp: 131 len: 135 to MEMORY
2016-02-19 00:27:23,552 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput: Read 131 bytes from map-output for attempt_1455821614868_0001_m_000013_0
2016-02-19 00:27:23,552 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 131, inMemoryMapOutputs.size() -> 15, commitMemory -> 1874, usedMemory ->2005
2016-02-19 00:27:23,552 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.Fetcher: fetcher#1 about to shuffle output of map attempt_1455821614868_0001_m_000015_0 decomp: 141 len: 145 to MEMORY
2016-02-19 00:27:23,553 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.InMemoryMapOutput: Read 141 bytes from map-output for attempt_1455821614868_0001_m_000015_0
2016-02-19 00:27:23,553 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 141, inMemoryMapOutputs.size() -> 16, commitMemory -> 2005, usedMemory ->2146
2016-02-19 00:27:23,553 INFO [fetcher#1] org.apache.hadoop.mapreduce.task.reduce.ShuffleSchedulerImpl: Arun-PC:13562 freed by fetcher#1 in 31ms
2016-02-19 00:27:23,554 INFO [EventFetcher for fetching Map Completion Events] org.apache.hadoop.mapreduce.task.reduce.EventFetcher: EventFetcher is interrupted.. Returning
2016-02-19 00:27:23,560 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: finalMerge called with 16 in-memory map-outputs and 0 on-disk map-outputs
2016-02-19 00:27:23,575 INFO [main] org.apache.hadoop.mapred.Merger: Merging 16 sorted segments
2016-02-19 00:27:23,575 INFO [main] org.apache.hadoop.mapred.Merger: Down to the last merge-pass, with 16 segments left of total size: 1986 bytes
2016-02-19 00:27:23,594 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merged 16 segments, 2146 bytes to disk to satisfy reduce memory limit
2016-02-19 00:27:23,596 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 1 files, 2120 bytes from disk
2016-02-19 00:27:23,596 INFO [main] org.apache.hadoop.mapreduce.task.reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce

As Merge using old api hence adding the config to change it.

Select C:\Windows\system32\cmd.exe

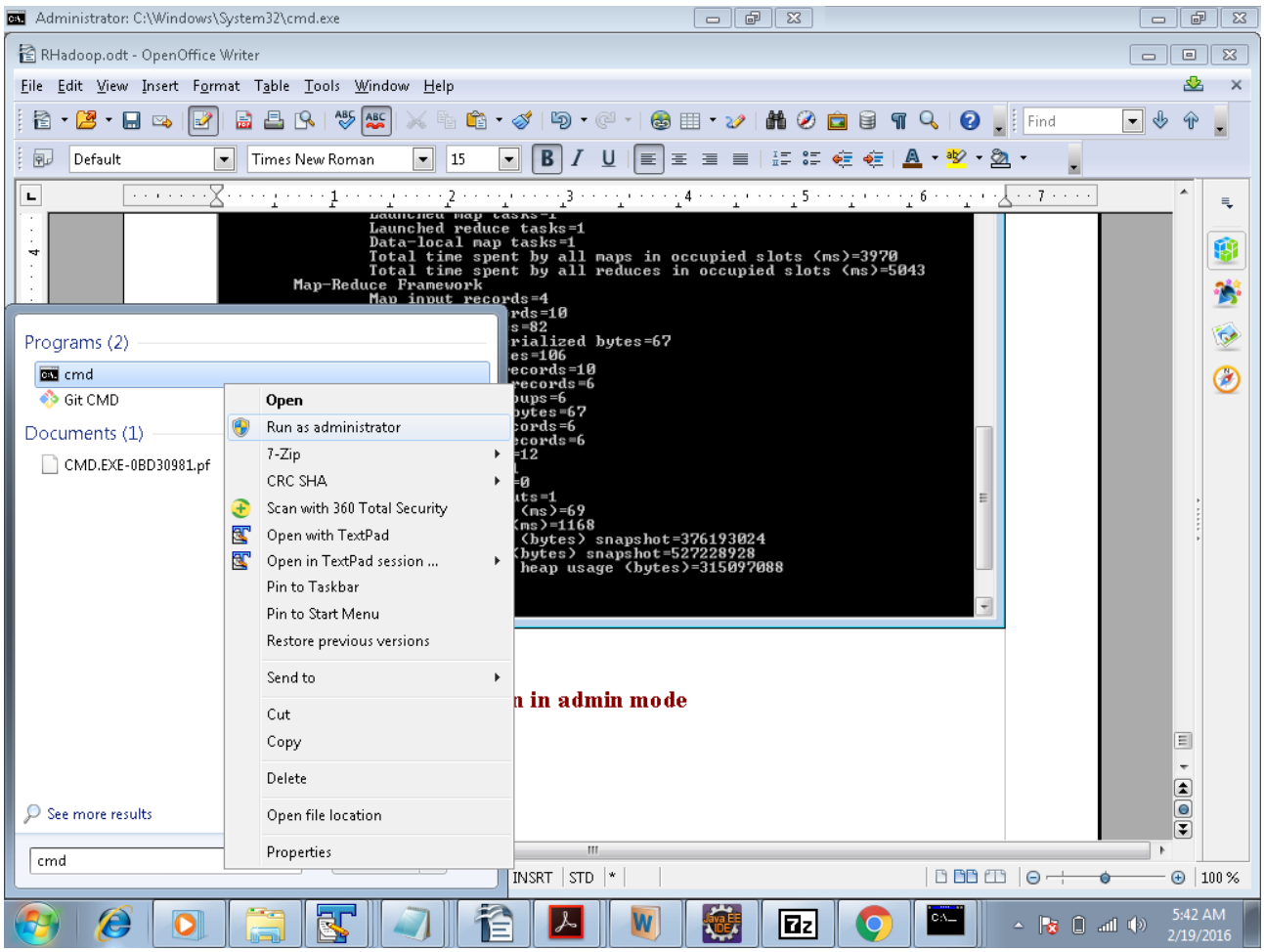
```
        [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] <path> ...]
        [-cp [-f] [-p] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] <path> ...]
        [-expunge]
        [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getmerge [-nl] <src> <localdst>]
        [-help [cmd ...]]
        [-ls [-d] [-h] [-R] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touchz <path> ...]
        [-usage [cmd ...]]

Generic options supported are
-conf <configuration file>     specify an application configuration file
-D <property=value>            use value for given property
-fs <local|namenode:port>      specify a namenode
-jt <local|jobtracker:port>    specify a job tracker
-files <comma separated list of files>    specify comma separated files to be co
pied to the map reduce cluster
-libjars <comma separated list of jars>    specify comma separated jar files to
include in the classpath.
-archives <comma separated list of archives>    specify comma separated archives
 to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]


C:\Users\Arun>hdfs dfs -chmod -R /input/wordcount.txt
-chmod: Not enough arguments: expected 2 but got 1
Usage: hadoop fs [generic options] -chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH
...

C:\Users\Arun>hdfs dfs -chmod -R 777
-chmod: Not enough arguments: expected 2 but got 1
Usage: hadoop fs [generic options] -chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH
...

C:\Users\Arun>hdfs dfs -chmod -R 777 /input/wordcount.txt

C:\Users\Arun>start-al_
```

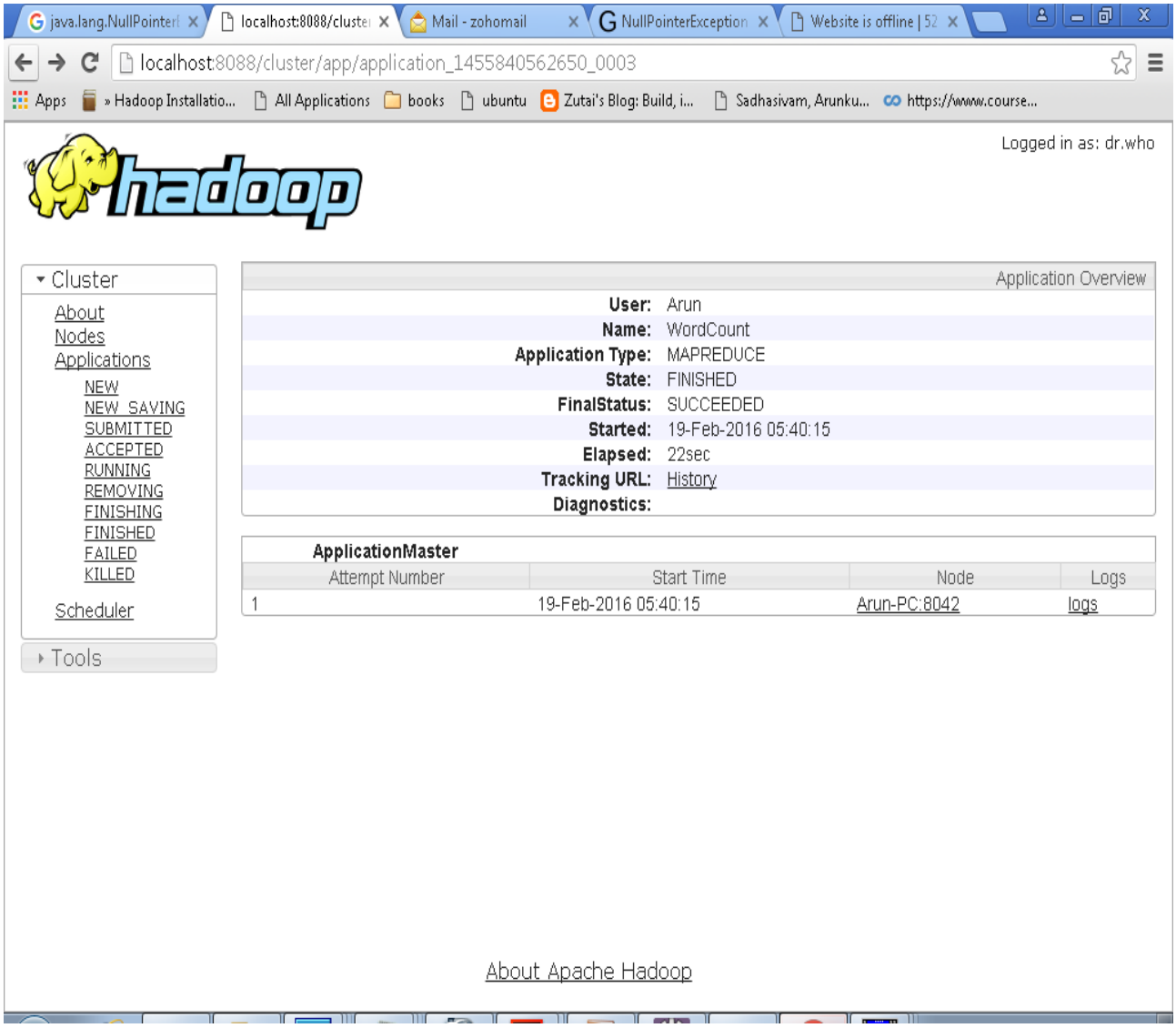


```
088/proxy/application_1455839837979_0003/
16/02/19 05:28:00 INFO mapreduce.Job: Running job: job_1455839837979_0003
16/02/19 05:28:07 INFO mapreduce.Job: Job job_1455839837979_0003 running in uber
 mode : false
16/02/19 05:28:07 INFO mapreduce.Job:  map 0% reduce 0%
16/02/19 05:28:07 INFO mapreduce.Job: Job job_1455839837979_0003 failed with sta
te FAILED due to: Application application_1455839837979_0003 failed 2 times due
to AM Container for appattempt_1455839837979_0003_000002 exited with  exitCode:
1 due to: Exception from container-launch:
org.apache.hadoop.util.Shell$ExitCodeException: No privilege to create symbolic
links.
No privilege to create symbolic links.
No privilege to create symbolic links.
No privilege to create symbolic links.

        at org.apache.hadoop.util.Shell.runCommand(Shell.java:464)
        at org.apache.hadoop.util.Shell.run(Shell.java:379)
        at org.apache.hadoop.util.Shell$ShellCommandExecutor.execute(Shell.java:
589)
        at org.apache.hadoop.yarn.server.nodemanager.DefaultContainerExecutor.la
unchContainer(DefaultContainerExecutor.java:195)
        at org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.C
ontainerLaunch.call(ContainerLaunch.java:283)
        at org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.C
ontainerLaunch.call(ContainerLaunch.java:79)
        at java.util.concurrent.FutureTask.run(FutureTask.java:262)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1145)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:615)
        at java.lang.Thread.run(Thread.java:745)

        1 file(s) moved.

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0

C:\tmp\hadoop-Arun\nm-local-dir\usercache\Arun\appcache\application_145583983797
9_0003\container_1455839837979_0003_02_000001>if 0 NEQ 0 exit /b 0
```

```
Administrator: C:\Windows\System32\cmd.exe
ed. Instead, use mapreduce.job.working.dir
16/02/19 05:40:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
55840562650_0003
16/02/19 05:40:15 INFO impl.YarnClientImpl: Submitted application application_14
55840562650_0003 to ResourceManager at /0.0.0.0:8032
16/02/19 05:40:15 INFO mapreduce.Job: The url to track the job: http://Arun-PC:8
088/proxy/application_1455840562650_0003/
16/02/19 05:40:15 INFO mapreduce.Job: Running job: job_1455840562650_0003
16/02/19 05:40:23 INFO mapreduce.Job: Job job_1455840562650_0003 running in uber
 mode : false
16/02/19 05:40:23 INFO mapreduce.Job:  map 0% reduce 0%
16/02/19 05:40:29 INFO mapreduce.Job:  map 100% reduce 0%
16/02/19 05:40:37 INFO mapreduce.Job:  map 100% reduce 100%
16/02/19 05:40:38 INFO mapreduce.Job: Job job_1455840562650_0003 completed succe
ssfully
16/02/19 05:40:38 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=67
                FILE: Number of bytes written=161456
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=150
                HDFS: Number of bytes written=37
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3970
                Total time spent by all reduces in occupied slots (ms)=5043
        Map-Reduce Framework
                Map input records=4
                Map output records=10
                Map output bytes=82
                Map output materialized bytes=67
                Input split bytes=106
                Combine input records=10
                Combine output records=6
                Reduce input groups=6
                Reduce shuffle bytes=67
                Reduce input records=6
                Reduce output records=6
                Spilled Records=12
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=69
                CPU time spent (ms)=1168
                Physical memory (bytes) snapshot=376193024
                Virtual memory (bytes) snapshot=527228928
                Total committed heap usage (bytes)=315097088
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
```

job run sucessfully once i run in admin mode


Administrator: C:\Windows\System32\cmd.exe
RHadoop.odt - OpenOffice Writer
File Edit View Insert Format Table Tools Window Help
Default
Times New Roman
15
Find
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=3970
Total time spent by all reduces in occupied slots (ms)=5043
Map-Reduce Framework
Map input records=4
Programs (2)
cmd
Git CMD
Documents (1)
CMD.EXE-0BD30981.pf
See more results
cmd
Open
Run as administrator
7-Zip
CRC SHA
Scan with 360 Total Security
Open with TextPad
Open in TextPad session ...
Pin to Taskbar
Pin to Start Menu
Restore previous versions
Send to
Cut
Copy
Delete
Open file location
Properties
in admin mode
INSRT
STD
100 %
5:42 AM
2/19/2016


localhost:8088/cluster/app/application_1455840562650_0003
Logged in as: dr.who
Cluster
About
Nodes
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
REMOVING
FINISHING
FINISHED
FAILED
KILLED
Scheduler
Tools
Application Overview
User: Arun
Name: WordCount
Application Type: MAPREDUCE
State: FINISHED
FinalStatus: SUCCEEDED
Started: 19-Feb-2016 05:40:15
Elapsed: 22sec
Tracking URL: History
Diagnostics:
ApplicationMaster
Attempt Number
Start Time
Node
Logs
1
19-Feb-2016 05:40:15
Arun-PC:8042
logs
About Apache Hadoop

R HADOOP STREAMING

CONFIGURE R:

C:\HADOOPOUTPUT>path
PATH=C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;C:\Windows\System32\WindowsPowerShell\v1.0\;C:\Program Files\Intel\WiFi\bin\;C:\Program Files\Common Files\Intel\WirelessCommon\;C:\Program Files (x86)\Skype\Phone\;C:\apache-maven-3.3.9\bin;C:\protoc;C:\Program Files\Microsoft SDKs\Windows\v7.1\bin;C:\Program Files\Git\bin;C:\zlib128;C:\zlib128\lib;C:\zlib128\include;C:\Program Files (x86)\CMake 2.6\bin;C:\hadoop-2.2.0\bin;C:\hadoop-2.2.0\sbin;C:\Java\jdk1.7.0_79\bin;C:\Anaconda2;C:\Anaconda2\Library\bin;C:\Anaconda2\Scripts

```
Caused by: java.lang.RuntimeException: configuration exception
        at org.apache.hadoop.streaming.PipeMapRed.configure(PipeMapRed.java:222)

        at org.apache.hadoop.streaming.PipeMapper.configure(PipeMapper.java:66)
        ... 22 more
Caused by: java.io.IOException: Cannot run program "c:/HADOOPOUTPUT/MapReduce.R"
: CreateProcess error=193, %1 is not a valid Win32 application
        at java.lang.ProcessBuilder.start(ProcessBuilder.java:1047)
        at org.apache.hadoop.streaming.PipeMapRed.configure(PipeMapRed.java:209)

        ... 23 more
Caused by: java.io.IOException: CreateProcess error=193, %1 is not a valid Win32
 application
        at java.lang.ProcessImpl.create(Native Method)
        at java.lang.ProcessImpl.<init>(ProcessImpl.java:385)
        at java.lang.ProcessImpl.start(ProcessImpl.java:136)
        at java.lang.ProcessBuilder.start(ProcessBuilder.java:1028)
```

After configuring R

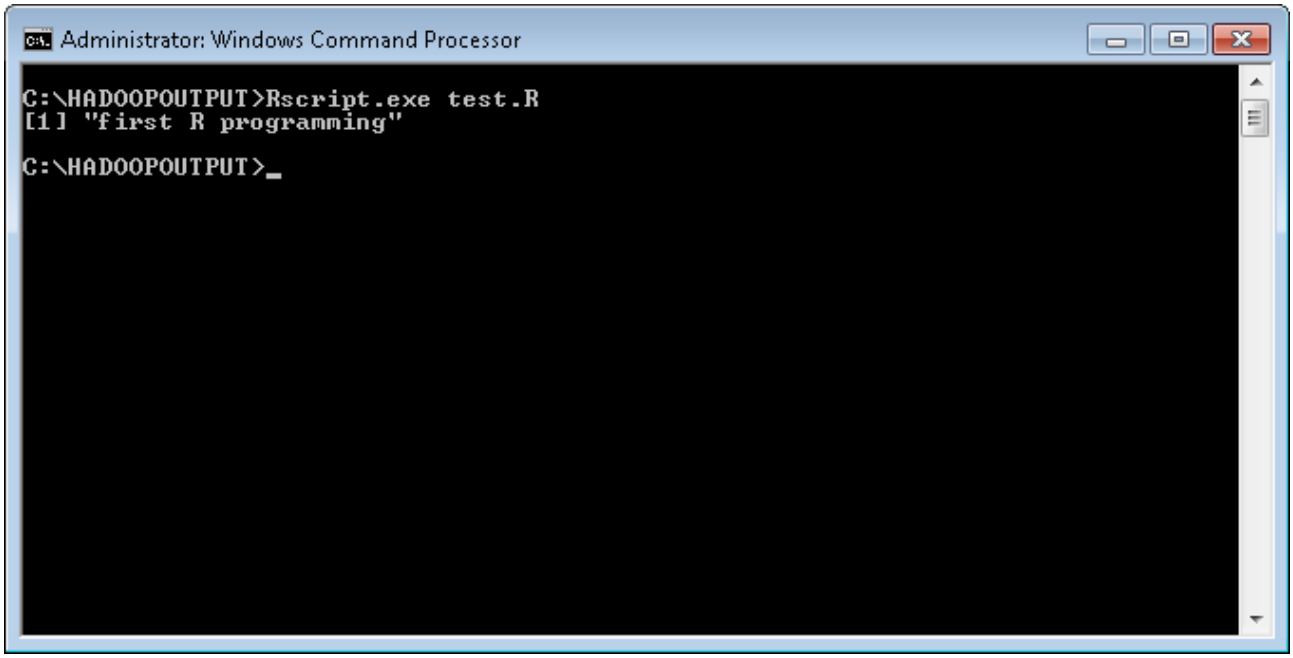
C:\Windows\system32>PATH
PATH=C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;C:\Windows\System32\WindowsPowerShell\v1.0\;C:\Program Files\Intel\WiFi\bin\;C:\Program Files\Common Files\Intel\WirelessCommon\;C:\Program Files (x86)\Skype\Phone\;C:\apache-maven-3.3.9\bin;C:\protoc;C:\Program Files\Microsoft SDKs\Windows\v7.1\bin;C:\Program Files\Git\bin;C:\zlib128;C:\zlib128\lib;C:\zlib128\include;C:\Program Files (x86)\CMake 2.6\bin;C:\hadoop-2.2.0\bin;C:\hadoop-2.2.0\sbin;C:\Java\jdk1.7.0_79\bin;C:\Anaconda2;C:\Anaconda2\Library\bin;C:\Anaconda2\Scripts;C:\Program Files\\R-3.2.3\bin

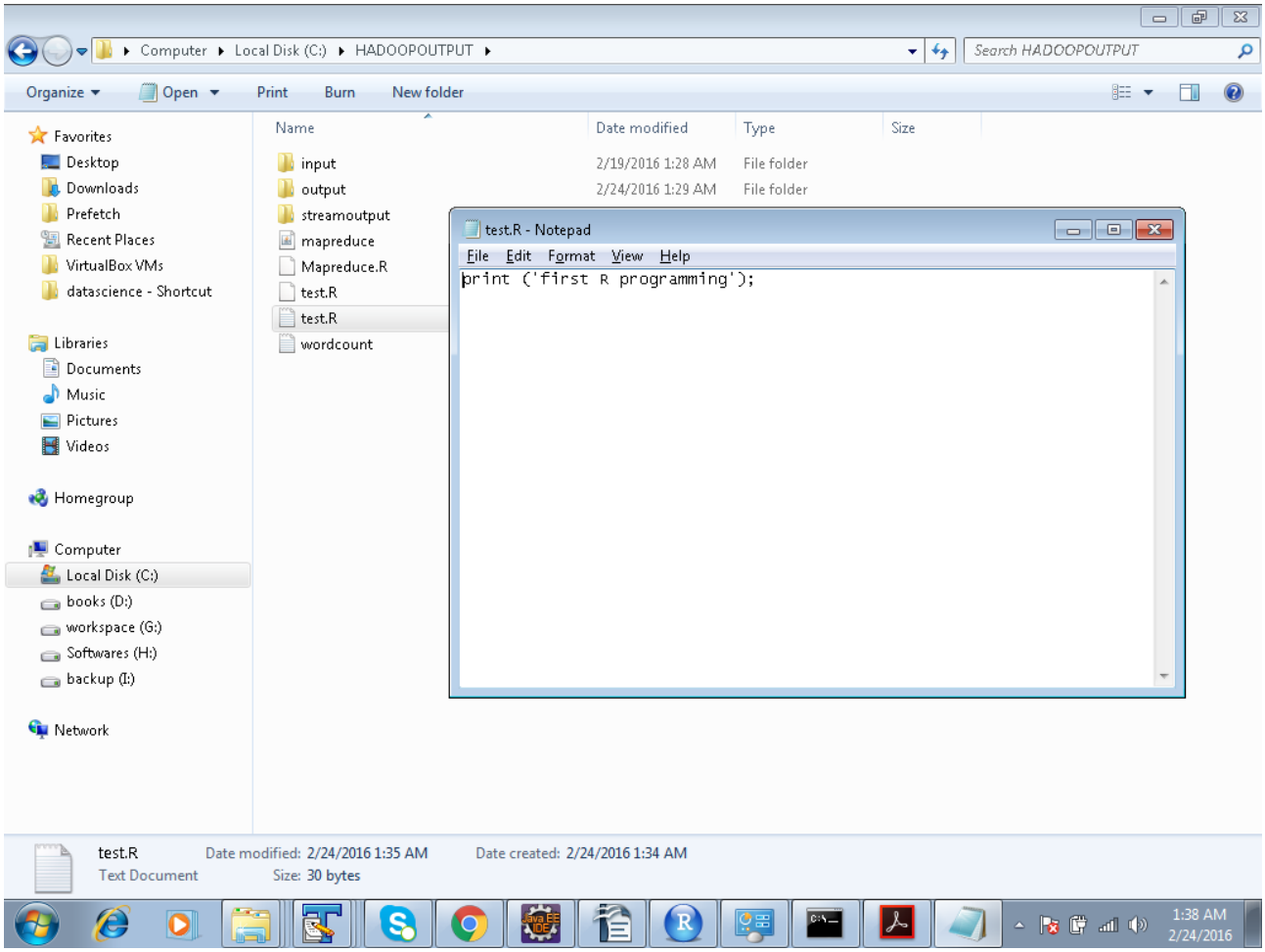
NOTE: **R api bin path is needed not R studio .**

It works Fine!!!

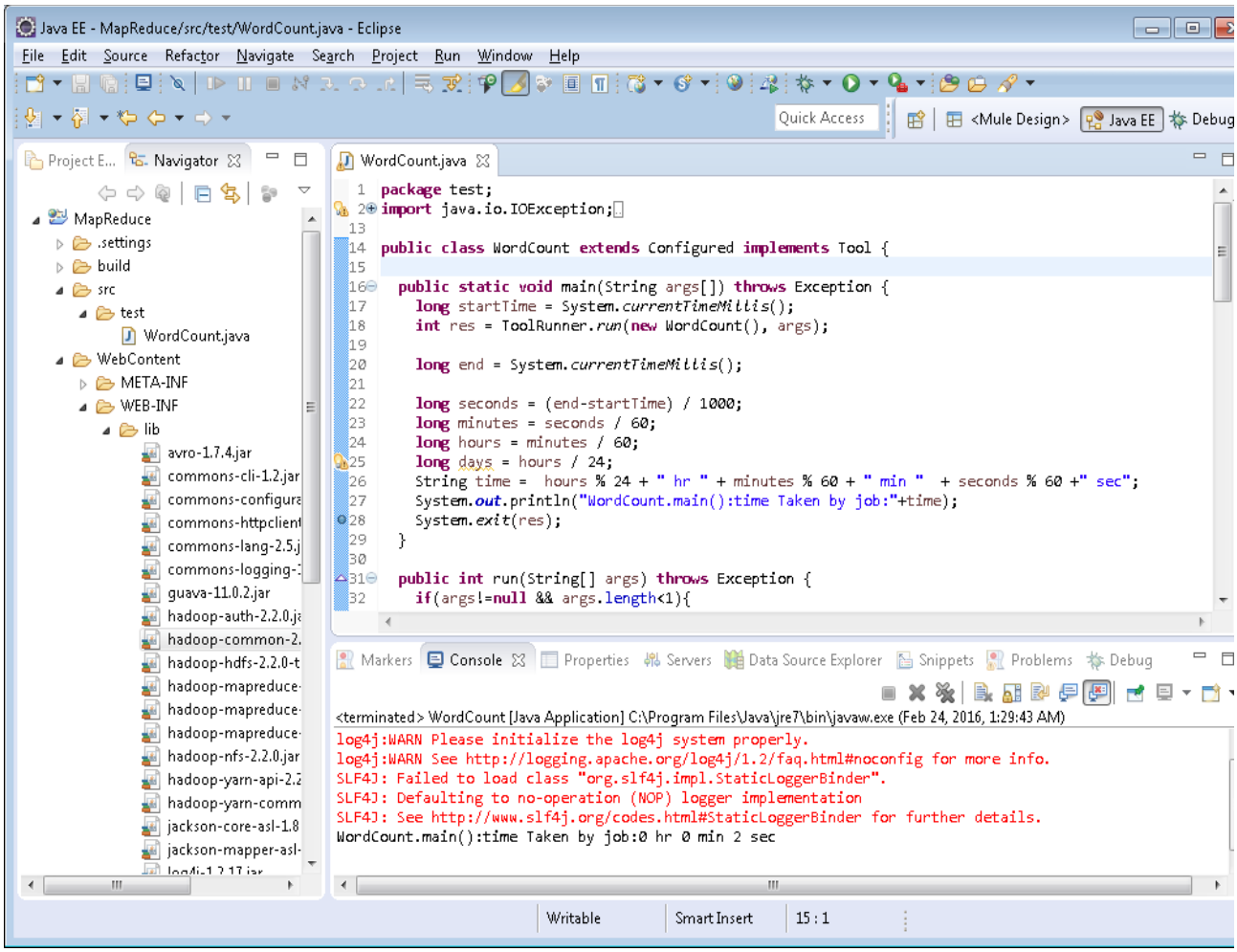
```
C:\HADOOPOUTPUT>yarn jar %HADOOP_HOME%/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar -input /input/wordcount.txt -output /Routput -mapper Mapreduce.R -reducer Mapreduce.R
```

To check R is working correctly





Job Run in standalone java class:



it takes 2 sec

Run same program in Hadoop

ubuntu:

hdadmin@ubuntu:/jars$
hadoop jar mapreduce.jar test.WordCount /input/wordcount.txt /output

windows:
c:\HADOOPOUTPUT>

yarn jar mapreduce.jar test.WordCount /input/wordcount.txt /output

Note:
in ubuntu both above command hadoop jar or yarn jar works fine.

Select Administrator: Apache Hadoop Distribution

```
ed. Instead, use mapreduce.job.working.dir
16/02/24 02:12:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
56259978769_0002
16/02/24 02:12:13 INFO impl.YarnClientImpl: Submitted application application_14
56259978769_0002 to ResourceManager at /0.0.0.0:8032
16/02/24 02:12:13 INFO mapreduce.Job: The url to track the job: http://Arun-PC:8
088/proxy/application_1456259978769_0002/
16/02/24 02:12:13 INFO mapreduce.Job: Running job: job_1456259978769_0002
16/02/24 02:12:21 INFO mapreduce.Job: Job job_1456259978769_0002 running in uber
 mode : false
16/02/24 02:12:21 INFO mapreduce.Job:  map 0% reduce 0%
16/02/24 02:12:27 INFO mapreduce.Job:  map 100% reduce 0%
16/02/24 02:12:35 INFO mapreduce.Job:  map 100% reduce 100%
16/02/24 02:12:36 INFO mapreduce.Job: Job job_1456259978769_0002 completed succe
ssfully
16/02/24 02:12:36 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=97
                FILE: Number of bytes written=161248
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=176
                HDFS: Number of bytes written=59
                HDFS: Number of read operations=6
```

All Applications | Create Your First Had | Email Hosting | Hoste | hadoopClient/HADO | All Applications

localhost:8088/cluster

Apps | » Hadoop Installatio... | All Applications | books | ubuntu | Zutai's Blog: Build, i... | Sadhasivam, Arunku... | https://www.course...

hadoop

All Applications

Cluster
- About
- Nodes
- Applications
  - NEW
  - NEW_SAVING
  - SUBMITTED
  - ACCEPTED
  - RUNNING
  - REMOVING
  - FINISHING
  - FINISHED
  - FAILED
  - KILLED
- Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes	Decommissioned Nodes
1	0	0	1	0	0 B	8 GB	0 B	1	0

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalSt
application_1456259978769_0002	Arun	WordCount	MAPREDUCE	default	Tue, 23 Feb 2016 20:42:12 GMT	Tue, 23 Feb 2016 20:42:35 GMT	FINISHED	SUCCEE

Showing 1 to 1 of 1 entries

About Apache Hadoop

Time Taken by Hadoop : 23 secs

SAME INPUT RUN IN HADOOP STREAMING

Select Administrator: Apache Hadoop Distribution

```
16/02/24 02:26:13 INFO Configuration.deprecation: mapred.working.dir is deprecat
ed. Instead, use mapreduce.job.working.dir
16/02/24 02:26:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
56259978769_0003
16/02/24 02:26:14 INFO impl.YarnClientImpl: Submitted application application_14
56259978769_0003 to ResourceManager at /0.0.0.0:8032
16/02/24 02:26:14 INFO mapreduce.Job: The url to track the job: http://Arun-PC:8
088/proxy/application_1456259978769_0003/
16/02/24 02:26:14 INFO mapreduce.Job: Running job: job_1456259978769_0003
16/02/24 02:26:21 INFO mapreduce.Job: Job job_1456259978769_0003 running in uber
 mode : false
16/02/24 02:26:21 INFO mapreduce.Job:  map 0% reduce 0%
16/02/24 02:26:30 INFO mapreduce.Job:  map 50% reduce 0%
16/02/24 02:26:31 INFO mapreduce.Job:  map 100% reduce 0%
16/02/24 02:26:37 INFO mapreduce.Job:  map 100% reduce 100%
16/02/24 02:26:38 INFO mapreduce.Job: Job job_1456259978769_0003 completed succe
ssfully
16/02/24 02:26:38 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=123
                FILE: Number of bytes written=240284
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=291
```

All Applications | Create Your First Had | Email Hosting | Hoste | hadoopClient/HADO | All Applications

localhost:8088/cluster

Apps | » Hadoop Installatio... | All Applications | books | ubuntu | Zutai's Blog: Build, i... | Sadhasivam, Arunku... | https://www.course...

hadoop

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes	Decommissioned Nodes
2	0	0	2	0	0 B	8 GB	0 B	1	0

Show 20 entries

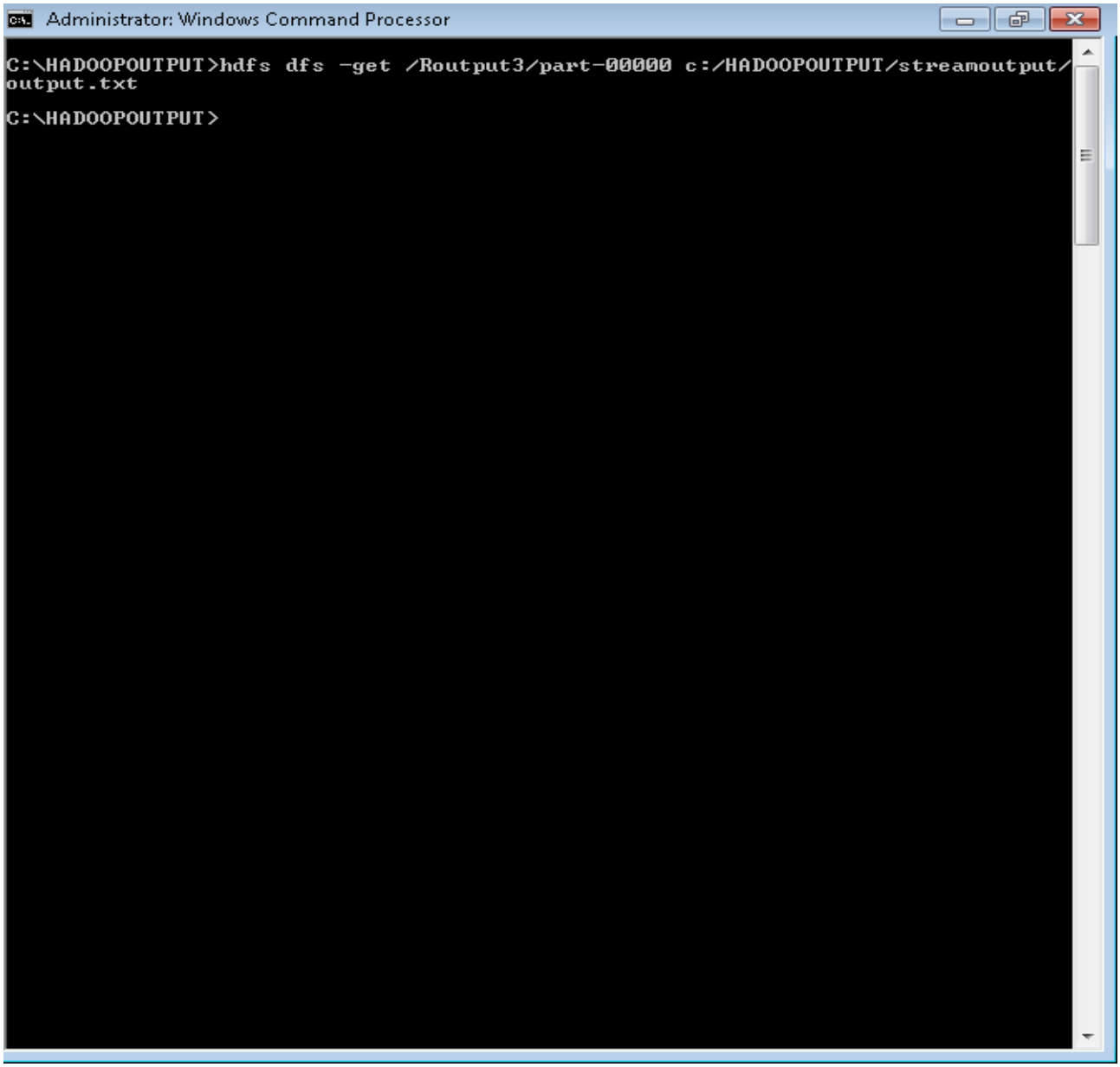
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State
application_1456259978769_0003	Arun	streamjob1368969199761739062.jar	MAPREDUCE	default	Tue, 23 Feb 2016 20:56:14 GMT	Tue, 23 Feb 2016 20:56:37 GMT	FINISHE
application_1456259978769_0002	Arun	WordCount	MAPREDUCE	default	Tue, 23 Feb 2016 20:42:12 GMT	Tue, 23 Feb 2016 20:42:35 GMT	FINISHE

Showing 1 to 2 of 2 entries

About Apache Hadoop

Time Taken by Hadoop : 23 secs

```
ubuntu:

hdadmin@ubuntu:/jars$
 hadoop jar hadoop-streaming-2.5.0.jar -input /input/wordcount.txt
-output /outputR -mapper RmapReduce.R -reducer RMapReduce.R -file
RMapReduce.R

windows:
c:\HADOOPOUTPUT>

yarn jar %HADOOP_HOME%/share/hadoop/tools/lib/hadoop-streaming-2
.2.0.jar -input /input/wordcount.txt -output /Routput

C:\HADOOPOUTPUT>hdfs dfs -cat /Routput/part-00000
hdfs dfs -get /Routput/part-00000 c:/HADOOPOUTPUT/streamoutput/
```



code:
https://github.com/arunsadhasivam/hadoopClient.git

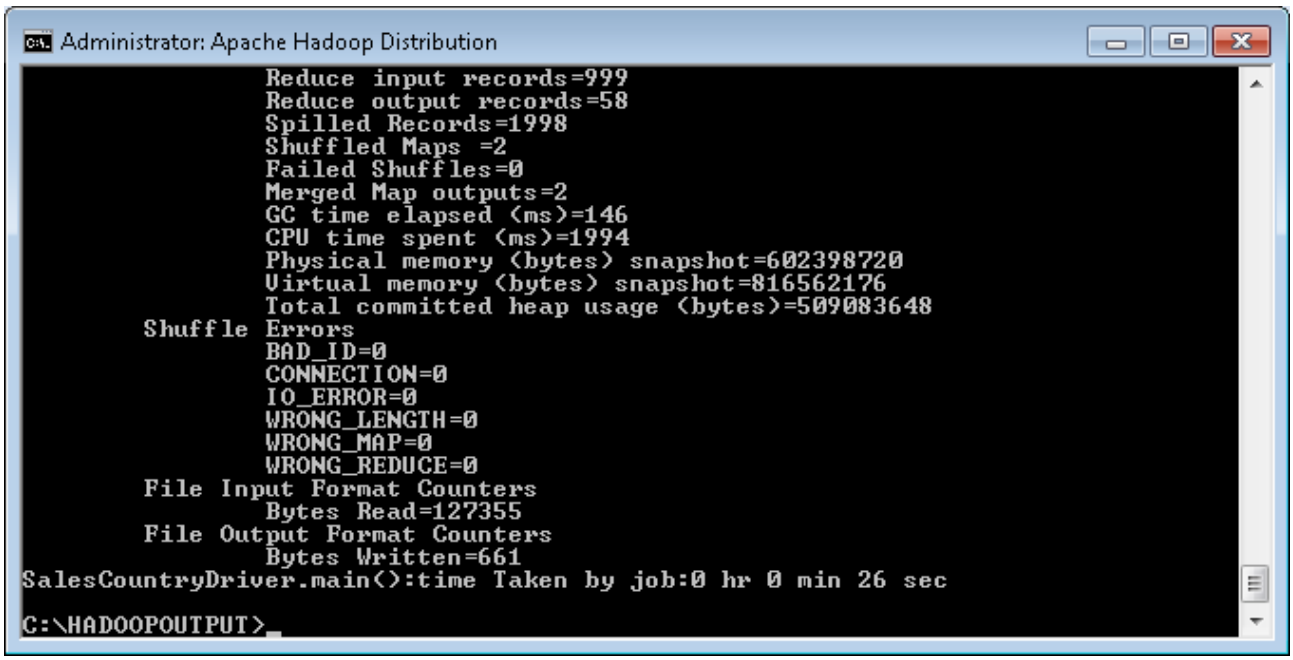
HADOOP PERFORMANCE COMPARISION ON LARGE DATASETS

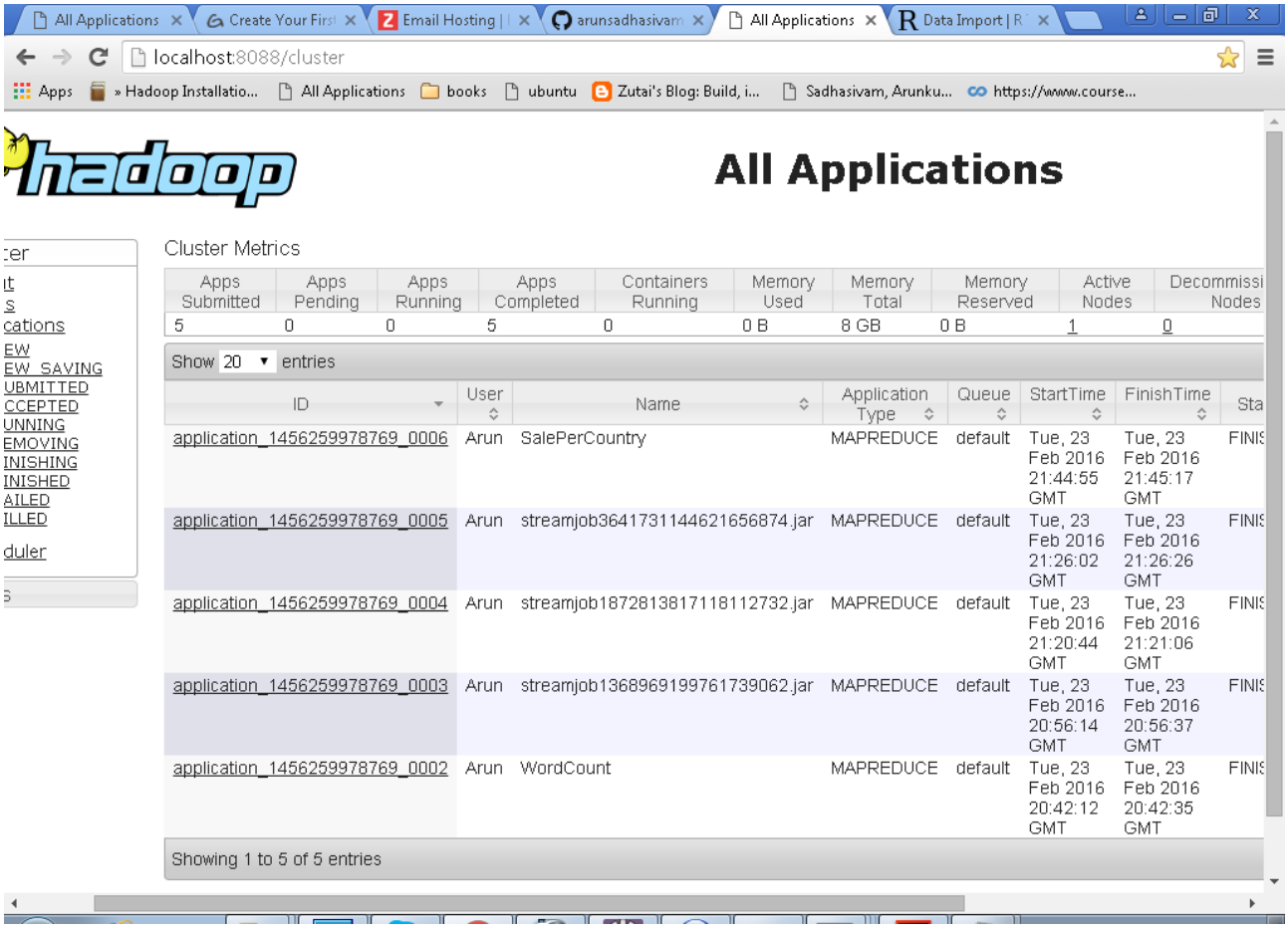
RUN EXCEL INPUT HAVING 1000 ROWS

In Normal java standalone it tooks 2 sec.

windows:
c:\HADOOPOUTPUT>

```
yarn jar mapreduce.jar test.SalesCountryDriver /input/sales.csv
/outputcsv
```





Time Taken by Hadoop : 22 secs

RUN EXCEL INPUT HAVING 3500 ROWS

In Normal java standalone it tooks 3 sec.

```
windows:
c:\HADOOPOUTPUT>

yarn jar mapreduce.jar test.SalesCountryDriver /input/sales.csv
/outputcsv
```

Standalone Job in java:

Java EE - MapReduce/src/test/SalesCountryDriver.java - Eclipse

File Edit Source Refactor Navigate Search Project Run Window Help

WordCount.java | SalesCountry...

```
package test;

import org.apache.hadoop.fs.Path;

public class SalesCountryDriver {
    public static void main(String[] args) {
        if(args!=null && args.length<1){
            args= new String[2];
            args[0]="c:/HADOOPOUTPUT/sales.csv";
            args[1]="c:/HADOOPOUTPUT/output";
        }
        long startTime = System.currentTimeMillis();
        JobClient my_client = new JobClient();
        // Create a configuration object for the job
        JobConf job_conf = new JobConf(SalesCountryDrive

        // Set a name of the Job
        job_conf.setJobName("SalePerCountry 3500 Records

        // Specify data type of output key and value
        job_conf.setOutputKeyClass(Text.class);
        job_conf.setOutputValueClass(IntWritable.class);

        // Specify names of Mapper and Reducer Class
        job_conf.setMapperClass(test.SalesMapper.class);
        job_conf.setReducerClass(test.SalesCountryReduce

        // Specify formats of the data
        job_conf.setInputFormat(TextIn
        job_conf.setOutputFormat(TextO

        // Set input and output directories using comman
        //arg[0] = name of input directory on HDFS, and

        FileInputFormat.setInputPaths(job_conf, new Path
```

test.SalesCountryReducer

Console

```
<terminated> SalesCountryDriver [Java Application] C:\Program Files\Java\jre7\bin\jav
g4j:WARN No appenders could be found for logger (org.apache.had
g4j:WARN Please initialize the log4j system properly.
g4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#nocon
F4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
F4J: Defaulting to no-operation (NOP) logger implementation
F4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for
lesCountryDriver.main():time Taken by job:0 hr 0 min 3 sec
```

Hadoop Mapreduce job:

Administrator: Windows Command Processor - yarn jar mapreduce.jar test.SalesCountryDriver /inp...

```
starting yarn daemons

C:\Windows\system32>hdfs dfs -copyFromLocal
-copyFromLocal: Not enough arguments: expected 1 but got 0
Usage: hadoop fs [generic options] -copyFromLocal [-f] [-p] <localsrc> ... <dst>


C:\Windows\system32>cd c:
C:\Windows\System32

C:\Windows\system32>c:

C:\Windows\system32>cd ..

C:\Windows>cd ..

C:\>cd HADOOPOUTPUT

C:\HADOOPOUTPUT>hdfs dfs -copyFromLocal sales.csv /input/sales3500.csv

C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.SalesCountryDriver /input/sales3500.
csv /outputcsv3500
16/02/24 22:57:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/02/24 22:57:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
```

Administrator: Windows Command Processor

```
                Reduce input records=3557
                Reduce output records=58
                Spilled Records=7114
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=129
                CPU time spent (ms)=3632
                Physical memory (bytes) snapshot=606785536
                Virtual memory (bytes) snapshot=829444096
                Total committed heap usage (bytes)=483917824
        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=466662
        File Output Format Counters
                Bytes Written=674
SalesCountryDriver.main():time Taken by job:0 hr 0 min 32 sec

C:\HADOOPOUTPUT>
```

About this learning sequer | All Applications | Library Genesis | awesome-public-datasets/

localhost:8088/cluster

Apps | » Hadoop Installatio... | All Applications | books | ubuntu | Zutai's Blog: Build, i... | Sadhasivam, Arunku... | https://www.course...

hadoop

All Applications

- Cluster
  - About
  - Nodes
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - REMOVING
    - FINISHING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes	Decommissioned Nodes
1	0	0	1	0	0 B	8 GB	0 B	1	0

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	Fi
application_1456334129495_0001	Arun	SalePerCountry 3500 Records	MAPREDUCE	default	Wed, 24 Feb 2016 17:27:33 GMT	Wed, 24 Feb 2016 17:28:01 GMT	FINISHED	SUC

Showing 1 to 1 of 1 entries

About Apache Hadoop

Time Taken by Hadoop : 28 secs

RUN EXCEL INPUT HAVING 10000 ROWS

In Normal java standalone it tooks 4 sec.

Java EE - MapReduce/src/test/SalesCountryDriver.java - Eclipse

```
package test;

import org.apache.hadoop.fs.Path;

public class SalesCountryDriver {
    public static void main(String[] arg
        if(args!=null && args.length<1){
            args= new String[2];
            args[0]="c:/HADOOPOUTPUT/sal
            args[1]="c:/HADOOPOUTPUT/out
        }
        long startTime = System.currentT
        JobClient my_client = new JobCli
        // Create a configuration object
        JobConf job_conf = new JobConf(S

        // Set a name of the Job
        job_conf.setJobName("SalePerCoun

        // Specify data type of output k
        job_conf.setOutputKeyClass(Text.
        job_conf.setOutputValueClass(Int

        // Specify names of Mapper and R
        job_conf.setMapperClass(test.Sal
        job_conf.setReducerClass(test.Sa

        // Specify formats of the data t
        job_conf.setInputFormat(TextInpu
        job_conf.setOutputFormat(TextOut

        // Set input and output director
        //arg[0] = name of input directo

        FileInputFormat.setInputPaths(jo
```

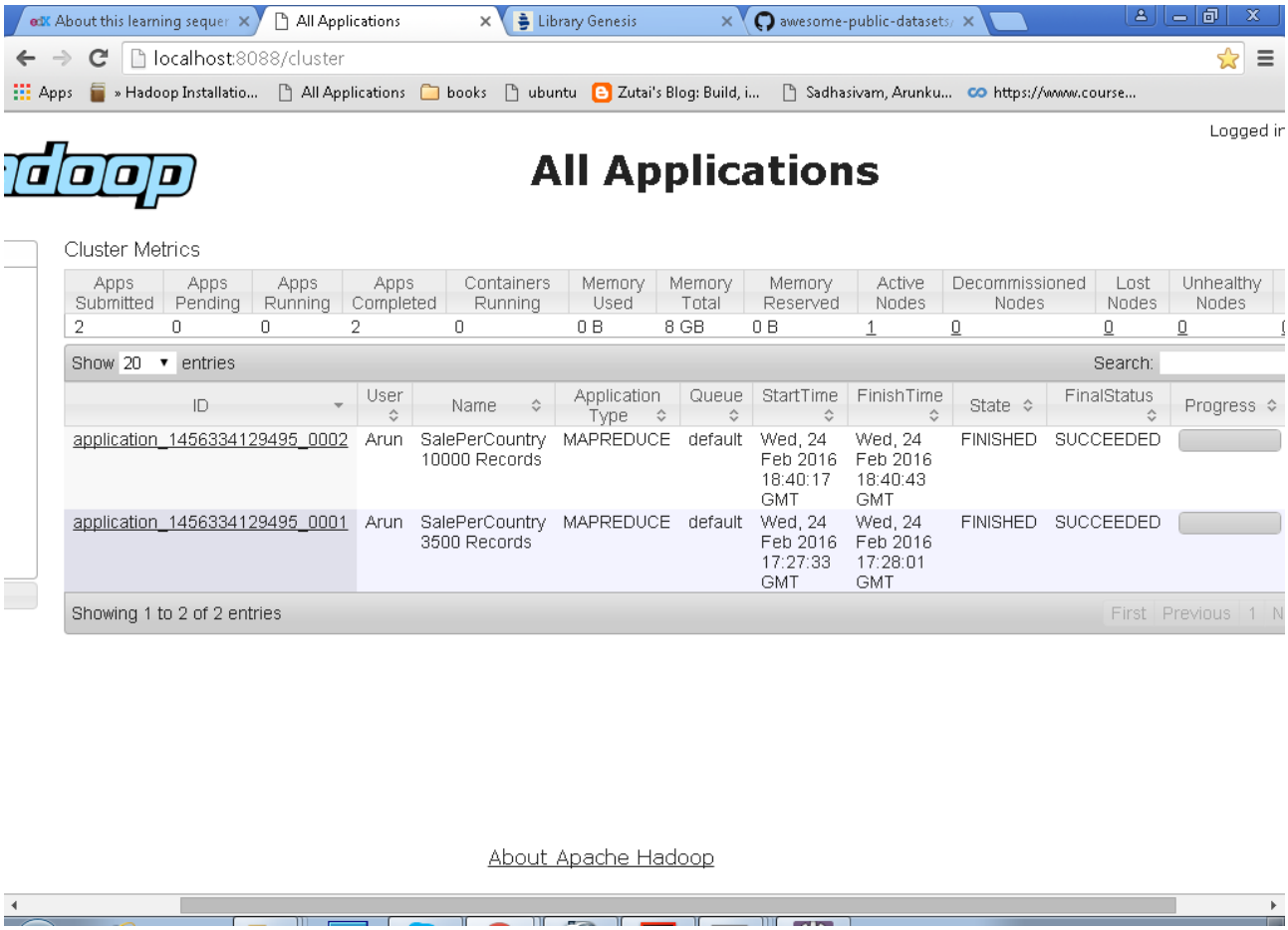
```
<terminated> SalesCountryDriver [Java Application] C:\Program F
enders could be found for logger (org.apache.ha
 initialize the log4j system properly.
tp://logging.apache.org/log4j/1.2/faq.html#noco
load class "org.slf4j.impl.StaticLoggerBinder".
 to no-operation (NOP) logger implementation
/www.slf4j.org/codes.html#StaticLoggerBinder fo
r.main():time Taken by job:0 hr 0 min 4 sec
```

Administrator: Windows Command Processor

```
                Reduce input records=3557
                Reduce output records=58
                Spilled Records=7114
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=56
                CPU time spent (ms)=3399
                Physical memory (bytes) snapshot=634642432
                Virtual memory (bytes) snapshot=841547776
                Total committed heap usage (bytes)=509083648
        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=466662
        File Output Format Counters
                Bytes Written=674
SalesCountryDriver.main():time Taken by job:0 hr 0 min 31 sec

C:\HADOOPOUTPUT>
```

Administrator: Windows Command Processor

```
16/02/25 00:27:20 INFO mapreduce.Job:  map 100% reduce 0%
16/02/25 00:27:29 INFO mapreduce.Job:  map 100% reduce 100%
16/02/25 00:27:30 INFO mapreduce.Job: Job job_1456334129495_0004 completed succe
ssfully
16/02/25 00:27:30 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=190553
                FILE: Number of bytes written=620075
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1400058
                HDFS: Number of bytes written=702
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=16015
                Total time spent by all reduces in occupied slots (ms)=5336
        Map-Reduce Framework
                Map input records=10669
                Map output records=10669
                Map output bytes=169209
                Map output materialized bytes=190559
                Input split bytes=188
                Combine input records=0
                Combine output records=0
                Reduce input groups=58
                Reduce shuffle bytes=190559
                Reduce input records=10669
                Reduce output records=58
                Spilled Records=21338
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=111
                CPU time spent (ms)=5364
                Physical memory (bytes) snapshot=617967616
                Virtual memory (bytes) snapshot=817680384
                Total committed heap usage (bytes)=509083648
        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=1399870
        File Output Format Counters
                Bytes Written=702
SalesCountryDriver.main():time Taken by job:0 hr 0 min 30 sec

C:\HADOOPOUTPUT>_
```



Time Taken by Hadoop : 26 secs

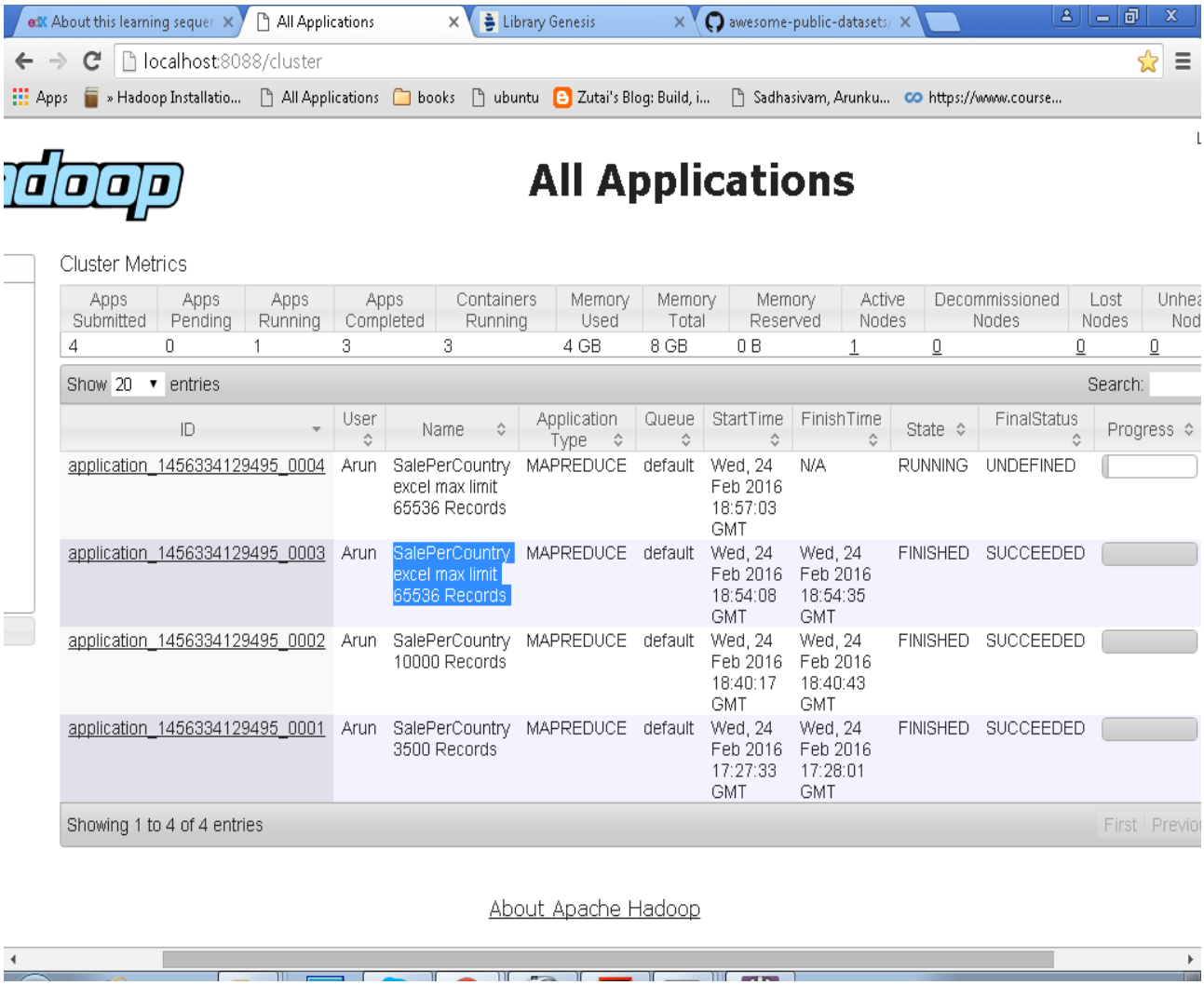
RUN EXCEL INPUT HAVING EXCEL WITH MAX LIMIT 65,536 ROWS.

In Normal java standalone it tooks 4 sec.

```
Administrator: Windows Command Processor - yarn jar mapreduce.jar test.SalesCountryDriver /inp...
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=466662
        File Output Format Counters
                Bytes Written=674
SalesCountryDriver.main():time Taken by job:0 hr 0 min 31 sec

C:\HADOOPOUTPUT>hdfs dfs -copyFromLocal sales.csv /input/sales65536.csv

C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.SalesCountryDriver /input/sales65536
.csv /outputcsv65000
16/02/25 00:24:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/02/25 00:24:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
16/02/25 00:24:07 WARN mapreduce.JobSubmitter: Hadoop command-line option parsin
g not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
16/02/25 00:24:08 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/25 00:24:08 INFO mapreduce.JobSubmitter: number of splits:2
16/02/25 00:24:08 INFO Configuration.deprecation: user.name is deprecated. Inste
```

Administrator: Windows Command Processor

```
16/02/25 00:24:27 INFO mapreduce.Job:  map 100% reduce 0%
16/02/25 00:24:35 INFO mapreduce.Job:  map 100% reduce 100%
16/02/25 00:24:36 INFO mapreduce.Job: Job job_1456334129495_0003 completed succe
ssfully
16/02/25 00:24:36 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=1170992
                FILE: Number of bytes written=2580953
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=8598369
                HDFS: Number of bytes written=741
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=16890
                Total time spent by all reduces in occupied slots (ms)=5720
        Map-Reduce Framework
                Map input records=65536
                Map output records=65536
                Map output bytes=1039914
                Map output materialized bytes=1170998
                Input split bytes=188
                Combine input records=0
                Combine output records=0
                Reduce input groups=58
                Reduce shuffle bytes=1170998
                Reduce input records=65536
                Reduce output records=58
                Spilled Records=131072
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=378
                CPU time spent (ms)=8033
                Physical memory (bytes) snapshot=642859008
                Virtual memory (bytes) snapshot=835715072
                Total committed heap usage (bytes)=525860864
        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=8598181
        File Output Format Counters
                Bytes Written=741
SalesCountryDriver.main():time Taken by job:0 hr 0 min 30 sec

C:\HADOOPOUTPUT>
```



Time Taken by Hadoop : 26-27 secs

NOTE:

1)when running mapreduce with 10 rows to 1000 rows hdfs takes much time than normal java standalone

2)hadoop streaming also took same time as normal java mapreduce program in hdfs.

3)for record of larger size only hadoop is useful.As you can see for 10 records 23 sec for 1000 records 22 secs.

Record size	Time Taken
5 Records	20 sec
3500 Records	22 sec
1000 Records	28 sec
10000 Records	26 sec
65,536 Records – Excel Max limit	26 to 27 sec run 3 times

Comparison chart with 6 Records vs 65,536 vs 1000 vs 3500

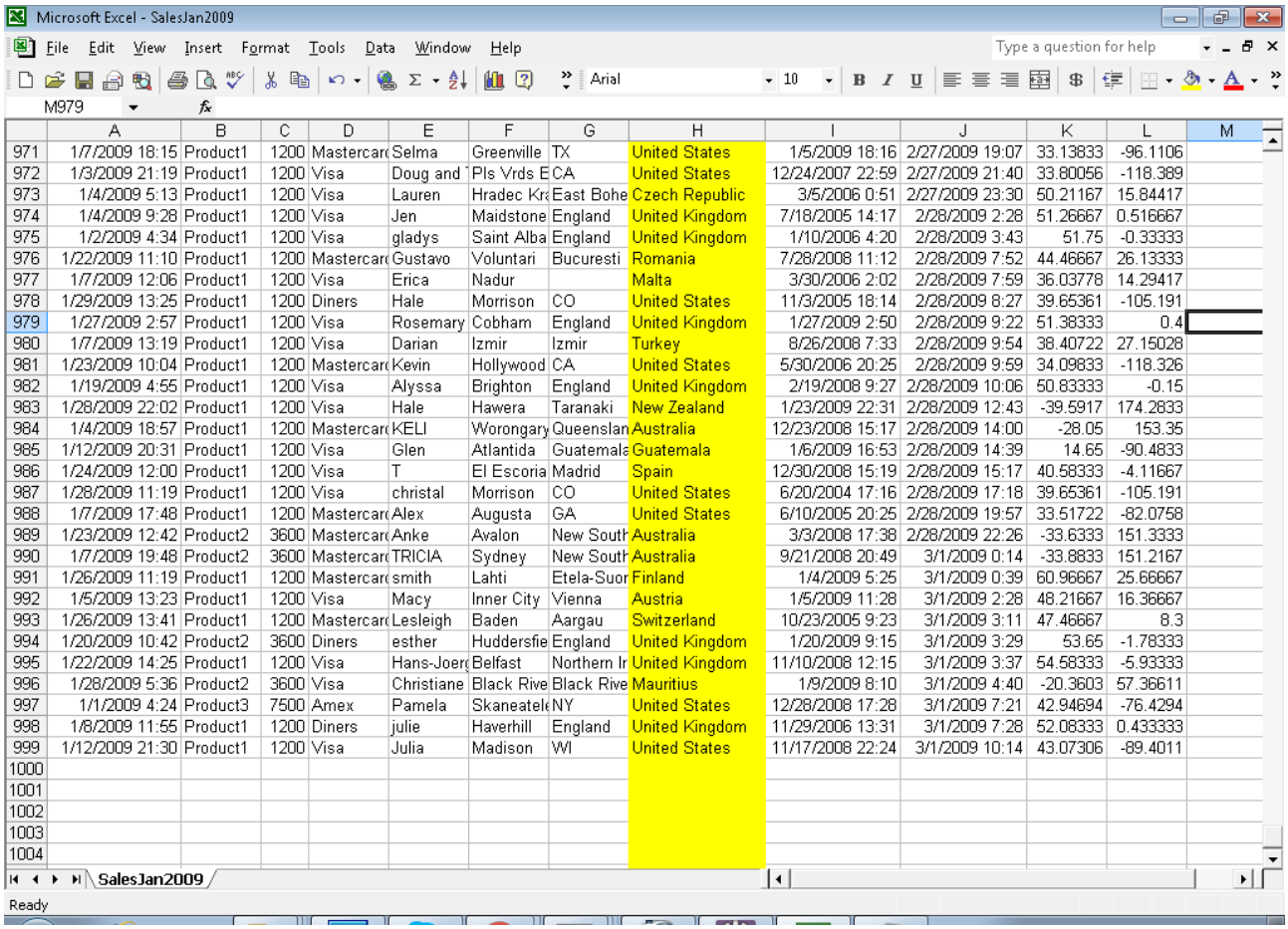
```
File System Counters
        FILE: Number of bytes read=97
        FILE: Number of bytes written=161282
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=176
        HDFS: Number of bytes written=59
        HDFS: Number of read operations=6
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=1    (2)         (2)          (2)
        Launched reduce tasks=1 (same)      (same)       (same)
        Data-local map tasks=1  (2)         (2)          (2)
        Total time spent by all maps in occupied slots (ms)=4110
        Total time spent by all reduces in occupied slots (ms)=4524
Map-Reduce Framework
        Map input records=5     (65536)     (10669)      (3557)
        Map output records=13   (65536)     (10669)      (3557)
        Map output bytes=119    (1039914)   (169209)     (56411)
        Map output materialized bytes=97
        Input split bytes=106   (188)       (188)        (186)
        Combine input records=13(0)         (0)          (0)
        Combine output records=8(0)         (0)          (0)
        Reduce input groups=8   (58)        (58)         (58)
        Reduce shuffle bytes=97
        Reduce input records=8  (65536)     (10669)      (3557)
        Reduce output records=8 (58)        (58)         (58)
        Spilled Records=16      (131072)    (21338)      (7114)
        Shuffled Maps =1        (2)         (2)          (2)
        Failed Shuffles=0
        Merged Map outputs=1    (2)         (2)           (2)
        GC time elapsed (ms)=41 (357)       (208)         (154)
        CPU time spent (ms)=1044
        Physical memory (bytes) snapshot=369164288
        Virtual memory (bytes) snapshot=520364032
        Total committed heap usage (bytes)=315097088
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=70
File Output Format Counters
        Bytes Written=59
```

```
t.main():time Taken by job:0 hr 0 min 24 sec
```

NOTE:

As you can see below original salesfor jan 2009 is 999 records. Since i copied the below 999 records multiple timeos to fill in 65,536 records. Hence it shows same

```
Reduce output records=8 (58)        (58)        (58)
Merged Map outputs=1    (2)         (2)         (2)
```



	A	B	C	D	E	F	G	H	I	J	K	L	M
971	1/7/2009 18:15	Product1	1200	Mastercard	Selma	Greenville	TX	United States	1/5/2009 18:16	2/27/2009 19:07	33.13833	-96.1106	
972	1/3/2009 21:19	Product1	1200	Visa	Doug and	Pls Vrds E	CA	United States	12/24/2007 22:59	2/27/2009 21:40	33.80056	-118.389	
973	1/4/2009 5:13	Product1	1200	Visa	Lauren	Hradec Kra	East Bohe	Czech Republic	3/5/2006 0:51	2/27/2009 23:30	50.21167	15.84417	
974	1/4/2009 9:28	Product1	1200	Visa	Jen	Maidstone	England	United Kingdom	7/18/2005 14:17	2/28/2009 2:28	51.26667	0.516667	
975	1/2/2009 4:34	Product1	1200	Visa	gladys	Saint Alba	England	United Kingdom	1/10/2006 4:20	2/28/2009 3:43	51.75	-0.33333	
976	1/22/2009 11:10	Product1	1200	Mastercard	Gustavo	Voluntari	Bucuresti	Romania	7/28/2008 11:12	2/28/2009 7:52	44.46667	26.13333	
977	1/7/2009 12:06	Product1	1200	Visa	Erica	Nadur		Malta	3/30/2006 2:02	2/28/2009 7:59	36.03778	14.29417	
978	1/29/2009 13:25	Product1	1200	Diners	Hale	Morrison	CO	United States	11/3/2005 18:14	2/28/2009 8:27	39.65361	-105.191	
979	1/27/2009 2:57	Product1	1200	Visa	Rosemary	Cobham	England	United Kingdom	1/27/2009 2:50	2/28/2009 9:22	51.38333	0.4	
980	1/7/2009 13:19	Product1	1200	Visa	Darian	Izmir	Izmir	Turkey	8/26/2008 7:33	2/28/2009 9:54	38.40722	27.15028	
981	1/23/2009 10:04	Product1	1200	Mastercard	Kevin	Hollywood	CA	United States	5/30/2006 20:25	2/28/2009 9:59	34.09833	-118.326	
982	1/19/2009 4:55	Product1	1200	Visa	Alyssa	Brighton	England	United Kingdom	2/19/2008 9:27	2/28/2009 10:06	50.83333	-0.15	
983	1/28/2009 22:02	Product1	1200	Visa	Hale	Hawera	Taranaki	New Zealand	1/23/2009 22:31	2/28/2009 12:43	-39.5917	174.2833	
984	1/4/2009 18:57	Product1	1200	Mastercard	KELI	Worongary	Queenslan	Australia	12/23/2008 15:17	2/28/2009 14:00	-28.05	153.35	
985	1/12/2009 20:31	Product1	1200	Visa	Glen	Atlantida	Guatemala	Guatemala	1/6/2009 16:53	2/28/2009 14:39	14.65	-90.4833	
986	1/24/2009 12:00	Product1	1200	Visa	T	El Escoria	Madrid	Spain	12/30/2008 15:19	2/28/2009 15:17	40.58333	-4.11667	
987	1/28/2009 11:19	Product1	1200	Visa	christal	Morrison	CO	United States	6/20/2004 17:16	2/28/2009 17:18	39.65361	-105.191	
988	1/7/2009 17:48	Product1	1200	Mastercard	Alex	Augusta	GA	United States	6/10/2005 20:25	2/28/2009 19:57	33.51722	-82.0758	
989	1/23/2009 12:42	Product2	3600	Mastercard	Anke	Avalon	New South	Australia	3/3/2008 17:38	2/28/2009 22:26	-33.6333	151.3333	
990	1/7/2009 19:48	Product2	3600	Mastercard	TRICIA	Sydney	New South	Australia	9/21/2008 20:49	3/1/2009 0:14	-33.8833	151.2167	
991	1/26/2009 11:19	Product1	1200	Mastercard	smith	Lahti	Etela-Suor	Finland	1/4/2009 5:25	3/1/2009 0:39	60.96667	25.66667	
992	1/5/2009 13:23	Product1	1200	Visa	Macy	Inner City	Vienna	Austria	1/5/2009 11:28	3/1/2009 2:28	48.21667	16.36667	
993	1/26/2009 13:41	Product1	1200	Mastercard	Lesleigh	Baden	Aargau	Switzerland	10/23/2005 9:23	3/1/2009 3:11	47.46667	8.3	
994	1/20/2009 10:42	Product2	3600	Diners	esther	Huddersfie	England	United Kingdom	1/20/2009 9:15	3/1/2009 3:29	53.65	-1.78333	
995	1/22/2009 14:25	Product1	1200	Visa	Hans-Joerg	Belfast	Northern Ir	United Kingdom	11/10/2008 12:15	3/1/2009 3:37	54.58333	-5.93333	
996	1/28/2009 5:36	Product2	3600	Visa	Christiane	Black Rive	Black Rive	Mauritius	1/9/2009 8:10	3/1/2009 4:40	-20.3603	57.36611	
997	1/1/2009 4:24	Product3	7500	Amex	Pamela	Skaneatele	NY	United States	12/28/2008 17:28	3/1/2009 7:21	42.94694	-76.4294	
998	1/8/2009 11:55	Product1	1200	Diners	julie	Haverhill	England	United Kingdom	11/29/2006 13:31	3/1/2009 7:28	52.08333	0.433333	
999	1/12/2009 21:30	Product1	1200	Visa	Julia	Madison	WI	United States	11/17/2008 22:24	3/1/2009 10:14	43.07306	-89.4011	
1000													
1001													
1002													
1003													
1004													

NOTE:

As you can see below unique total records is 56 hene it shows 58 records i.e 56 +(2)Merged Map = 58

Microsoft Excel - SalesJan2009

File Edit View Insert Format Tools Data Window Help

H2 United Kingdom

	A	B	C	D	E	F	G	H	I
1	Transaction_date	Product	Price	Payment_	Name	City	State	Country	Account_Crea
2	1/2/2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	1/2/2009 6
3	1/2/2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	1/2/2009
5	1/3/2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/2005 2
10	1/4/2009 13:17	Product1	1200	Mastercard	Renee Elis	Tel Aviv	Tel Aviv	Israel	1/4/2009 1
11	1/4/2009 14:11	Product1	1200	Visa	Aidan	Chatou	Ile-de-Fran	France	6/3/2008
13	1/5/2009 5:39	Product1	1200	Amex	Heidi	Eindhoven	Noord-Bra	Netherlands	1/5/2009
17	1/4/2009 1:05	Product1	1200	Diners	Leanne	Julianstow	Meath	Ireland	1/4/2009
18	1/5/2009 11:37	Product1	1200	Visa	Janet	Ottawa	Ontario	Canada	1/5/2009
19	1/6/2009 5:02	Product1	1200	Diners	barbara	Hyderabad	Andhra Pr	India	1/6/2009
24	1/5/2009 4:10	Product1	1200	Mastercard	Nicola	Roodepoor	Gauteng	South Africa	1/5/2009
26	1/2/2009 1:11	Product1	1200	Mastercard	Lena	Kuopio	Ita-Suome	Finland	12/31/2008
34	1/8/2009 1:59	Product1	1200	Mastercard	SYLVIA	Vesenaz	Geneve	Switzerland	11/28/2007 1
44	1/3/2009 13:24	Product1	1200	Visa	Mehmet F	Helsingor	Frederiksb	Denmark	1/3/2009 1
58	1/11/2009 14:17	Product1	1200	Visa	Stephanie	Brussels	Brussels (	Belgium	1/11/2009 1
62	1/5/2009 0:31	Product1	1200	Mastercard	Bonnie	Saltsjobad	Stockholm	Sweden	1/4/2009 2
69	1/11/2009 11:33	Product1	1200	Visa	Stefan	Stavanger	Rogaland	Norway	1/11/2009
77	1/12/2009 13:41	Product1	1200	Visa	Eric	Gasperich	Luxembou	Luxembourg	1/12/2009 1
79	1/13/2009 5:57	Product1	1200	Visa	Robin	Milan	Lombardy	Italy	10/15/2008
87	1/5/2009 8:58	Product2	3600	Mastercard	Marcia	Telgte	Nordrhein-	Germany	9/1/2008
92	1/6/2009 17:15	Product1	1200	Visa	Andrea	Bubuieci	Chisinau	Moldova	12/24/2008 1

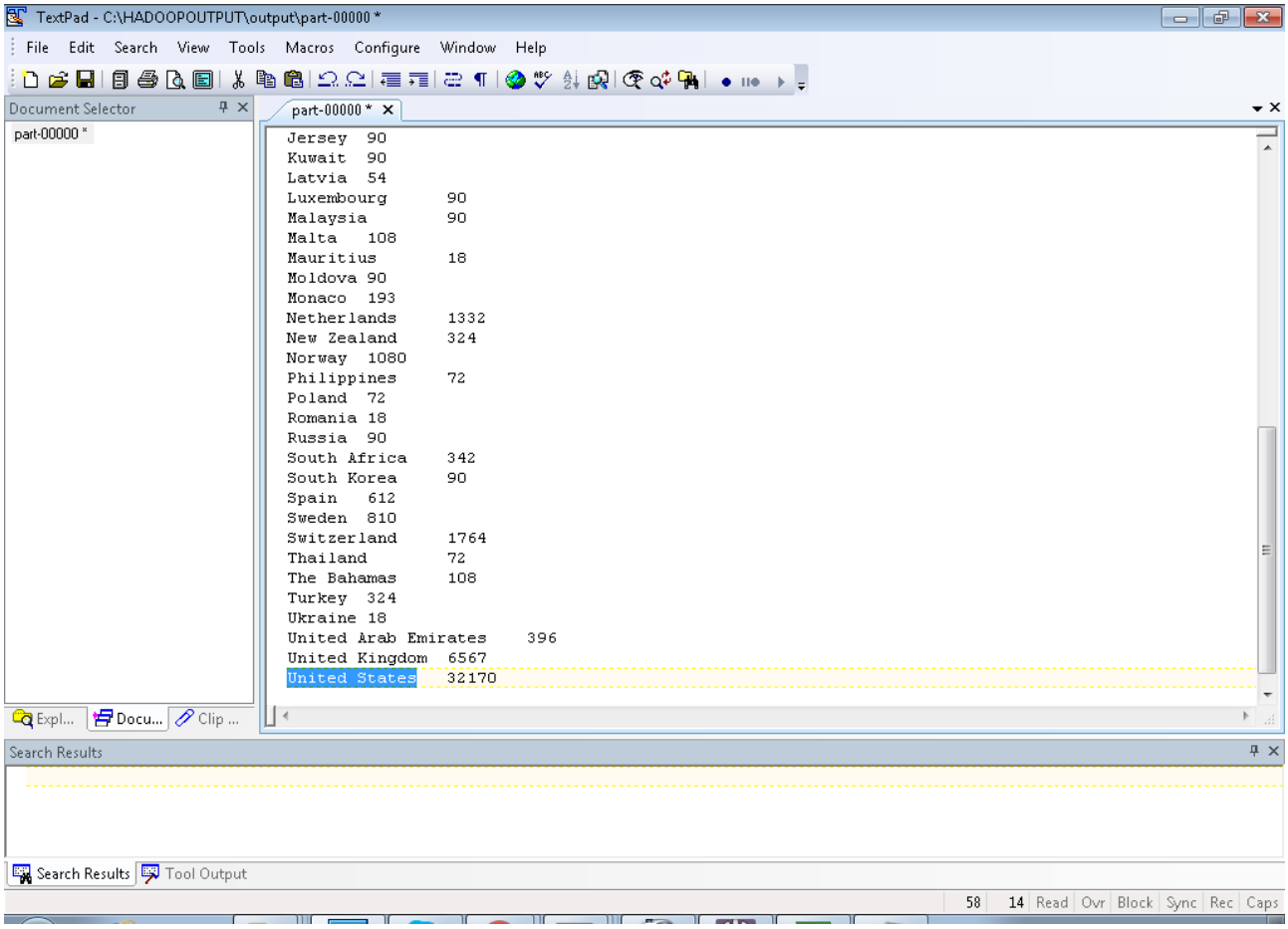
SalesJan2009

56 of 998 records found

NOTE:

As you can most of record size (3500,10000,65,536) when run in java standalone mode ,it takes 4 sec but in hdfs it takes 20-27 sec. Because hdfs proves good and improve performance only if size of record is very high.

Output



Code:

Mapper:

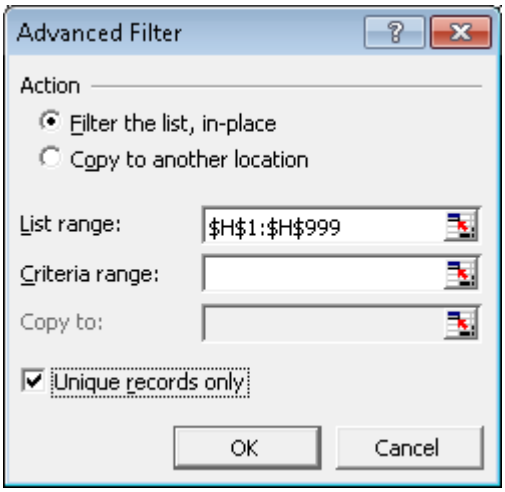
```
public class SalesMapper extends MapReduceBase implements Mapper<LongWritable, Text,
Text, IntWritable> {
	private final static IntWritable one = new IntWritable(1);

	public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
output, Reporter reporter) throws IOException {

		String valueString = value.toString();
		String[] SingleCountryData = valueString.split(",");
		output.collect(new Text(SingleCountryData[7]), one);
	}
}
```

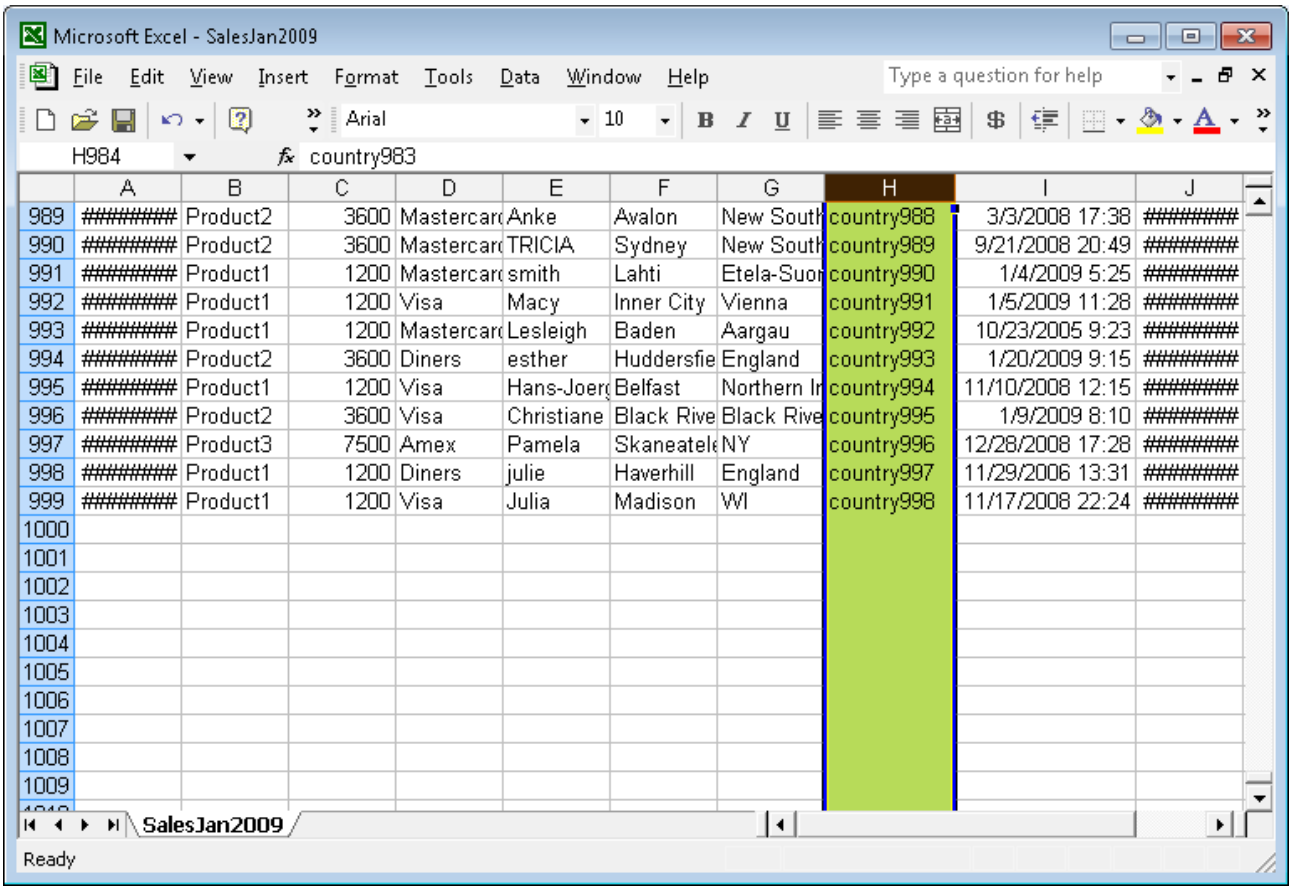
Reducer:

```
public class SalesCountryReducer extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable> {

	public void reduce(Text t_key, Iterator<IntWritable> values,
OutputCollector<Text,IntWritable> output, Reporter reporter) throws IOException {
		Text key = t_key;
		int frequencyForCountry = 0;
		while (values.hasNext()) {
			// replace type of value with the actual type of our value
			IntWritable value = (IntWritable) values.next();
			frequencyForCountry += value.get();

		}
		output.collect(key, new IntWritable(frequencyForCountry));
	}
}
```

NOTE:

Now make 999 records as unique





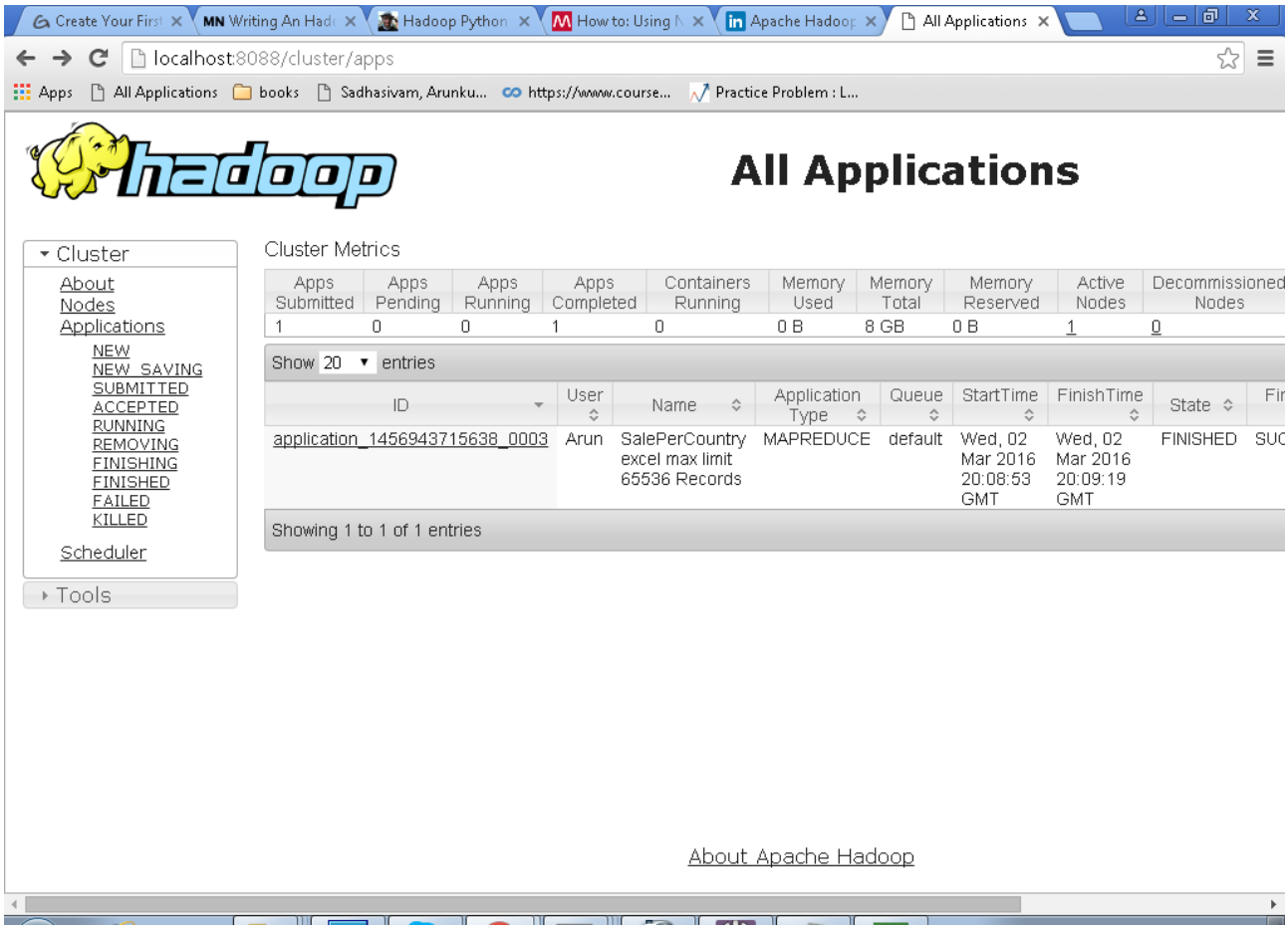
	A	B	C	D	E	F	G	H	I	J
989	#########	Product2	3600	Mastercar	Anke	Avalon	New South	country988	3/3/2008 17:38	#########
990	#########	Product2	3600	Mastercar	TRICIA	Sydney	New South	country989	9/21/2008 20:49	#########
991	#########	Product1	1200	Mastercar	smith	Lahti	Etela-Suor	country990	1/4/2009 5:25	#########
992	#########	Product1	1200	Visa	Macy	Inner City	Vienna	country991	1/5/2009 11:28	#########
993	#########	Product1	1200	Mastercar	Lesleigh	Baden	Aargau	country992	10/23/2005 9:23	#########
994	#########	Product2	3600	Diners	esther	Huddersfie	England	country993	1/20/2009 9:15	#########
995	#########	Product1	1200	Visa	Hans-Joerg	Belfast	Northern Ir	country994	11/10/2008 12:15	#########
996	#########	Product2	3600	Visa	Christiane	Black Rive	Black Rive	country995	1/9/2009 8:10	#########
997	#########	Product3	7500	Amex	Pamela	Skaneatele	NY	country996	12/28/2008 17:28	#########
998	#########	Product1	1200	Diners	julie	Haverhill	England	country997	11/29/2006 13:31	#########
999	#########	Product1	1200	Visa	Julia	Madison	WI	country998	11/17/2008 22:24	#########
1000										
1001										
1002										
1003										
1004										
1005										
1006										
1007										
1008										
1009										

Commands:

```
C:\HADOOPOUTPT>hdfs dfs -mkdir /input

C:\HADOOPOUTPUT>hdfs dfs -copyFromLocal SalesJan2009.csv
/input/salesunique.csv

C:\HADOOPOUTPUT>hdfs dfs -ls  /input/*
Found 1 items
-rw-r--r--   1 Arun supergroup     123637 2016-02-24 02:11
/input/sales.csv
Found 1 items
-rw-r--r--   1 Arun supergroup    1398907 2016-02-25 00:09
/input/sales10000.csv
Found 1 items
-rw-r--r--   1 Arun supergroup     466379 2016-02-24 22:53
/input/sales3500.csv
Found 1 items
-rw-r--r--   1 Arun supergroup    8594762 2016-02-25 00:22
/input/sales65536.csv
Found 1 items
-rw-r--r--   1 Arun supergroup     129745 2016-03-03 01:29
/input/salesunique.csv
Found 1 items
-rw-r--r--   1 Arun supergroup         70 2016-02-24 02:11
/input/wordcount.txt
```

RUN EXCEL INPUT HAVING EXCEL WITH UNIQUE 998 ROWS

```
C:\HADOOPOUTPT>yarn jar mapreduce.jar test.SalesCountryDriver
/input/salesunique.csv /outputUniquecsv
```



```
16/03/03 01:39:21 INFO mapreduce.Job: Job job_1456943715638_0003
completed successfully
16/03/03 01:39:21 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=16870
                FILE: Number of bytes written=272715
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=130599
                HDFS: Number of bytes written=12868
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=14589
                Total time spent by all reduces in occupied slots
(ms)=5780
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Map output bytes=14866
                Map output materialized bytes=16876
                Input split bytes=190
                Combine input records=0
                Combine output records=0
                Reduce input groups=999
                Reduce shuffle bytes=16876
                Reduce input records=999
                Reduce output records=999
                Spilled Records=1998
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=182
                CPU time spent (ms)=2214
                Physical memory (bytes) snapshot=598548480
                Virtual memory (bytes) snapshot=811290624
                Total committed heap usage (bytes)=509083648
        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=130409
        File Output Format Counters
                Bytes Written=12868
SalesCountryDriver.main():time Taken by job:0 hr 0 min 32 sec
```

Time Taken by Hadoop : 26 secs

RUN EXCEL INPUT HAVING EXCEL WITH UNIQUE 65536 ROWS

```
C:\HADOOPOUTPT>hdfs dfs -copyFromLocal SalesJan3500.csv
/input/salesunique65536.csv

C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.SalesCountryDriver
/input/salesunique65536.csv /outputUnique65536
```

Time Taken by Hadoop : 26 secs

```
16/03/03 01:51:24 INFO mapreduce.Job: Job job_1456943715638_0004
completed succe
ssfully
16/03/03 01:51:24 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=1234071
                FILE: Number of bytes written=2707138
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1479604
                HDFS: Number of bytes written=971921
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17106
                Total time spent by all reduces in occupied slots
(ms)=6297
        Map-Reduce Framework
                Map input records=65536
                Map output records=65536
                Map output bytes=1102993
                Map output materialized bytes=1234077
                Input split bytes=200
                Combine input records=0
                Combine output records=0
                Reduce input groups=65536
                Reduce shuffle bytes=1234077
                Reduce input records=65536
                Reduce output records=65536
                Spilled Records=131072
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=205
                CPU time spent (ms)=9153
                Physical memory (bytes) snapshot=646881280
                Virtual memory (bytes) snapshot=840310784
                Total committed heap usage (bytes)=532152320
        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=1479404
        File Output Format Counters
                Bytes Written=971921
SalesCountryDriver.main():time Taken by job:0 hr 0 min 31 sec
```

RUN EXCEL INPUT HAVING EXCEL WITH UNIQUE 3500 ROWS

```
C:\HADOOPOUTPUT>hdfs dfs -copyFromLocal SalesJan1000.csv
/input/salesunique3500.
csv

C:\HADOOPOUTPUT>yarn jar mapreduce.jar test.SalesCountryDriver
/input/salesuniqu
e3500.csv /outputUnique3500
```

Time Taken by Hadoop : 24 secs

```
16/03/03 01:59:52 INFO mapreduce.Job: Counters: 43
        File System Counters
                FILE: Number of bytes read=61905
                FILE: Number of bytes written=362800
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=180220
                HDFS: Number of bytes written=47895
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=15503
                Total time spent by all reduces in occupied slots
(ms)=5072
        Map-Reduce Framework
                Map input records=3501
```

```
                Map output records=3501
                Map output bytes=54897
                Map output materialized bytes=61911
                Input split bytes=198
                Combine input records=0
                Combine output records=0
                Reduce input groups=3501
                Reduce shuffle bytes=61911
                Reduce input records=3501
                Reduce output records=3501
                Spilled Records=7002
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=117
                CPU time spent (ms)=3696
                Physical memory (bytes) snapshot=650874880
                Virtual memory (bytes) snapshot=848662528
                Total committed heap usage (bytes)=520617984
        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=180022
        File Output Format Counters
                Bytes Written=47895
SalesCountryDriver.main():time Taken by job:0 hr 0 min 30 sec
```

Comparison chart with Unique Records -1000 vs 35000 vs 65,536

```
File System Counters
      FILE: Number of bytes read=16870
      FILE: Number of bytes written=272715
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=130599
      HDFS: Number of bytes written=12868
      HDFS: Number of read operations=9
      HDFS: Number of large read operations=0
      HDFS: Number of write operations=2
Job Counters
      Launched map tasks=2
      Launched reduce tasks=1
      Data-local map tasks=2
      Total time spent by all maps in occupied slots (ms)=14589
      Total time spent by all reduces in occupied slots (ms)=5780
Map-Reduce Framework
      Map input records=999                3501          65536
      Map output records=999               3501          65536
      Map output bytes=14866
      Map output materialized bytes=16876
      Input split bytes=190
      Combine input records=0
      Combine output records=0
      Reduce input groups=999              3501           65536
      Reduce shuffle bytes=16876           61911          1234077
      Reduce input records=999             3501           65536
      Reduce output records=999            3501           65536
      Spilled Records=1998                 7002           131072
      Shuffled Maps =2                     2               2
      Failed Shuffles=0
      Merged Map outputs=2
      GC time elapsed (ms)=182
      CPU time spent (ms)=2214
      Physical memory (bytes) snapshot=598548480
      Virtual memory (bytes) snapshot=811290624
      Total committed heap usage (bytes)=509083648
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=130409
File Output Format Counters
      Bytes Written=12868
```

NONUNIQUE VS UNIQUE RECORDS

NonUnique Records: 56 Records are unique
5 vs 65536 vs 10669 vs 3557 Records:

```
Launched map tasks=1            (2)         (2)          (2)
        Launched reduce tasks=1 (same)      (same)       (same)
        Data-local map tasks=1  (2)         (2)          (2)
        Total time spent by all maps in occupied slots (ms)=4110
        Total time spent by all reduces in occupied slots (ms)=4524
Map-Reduce Framework
        Map input records=5     (65536)     (10669)      (3557)
        Map output records=13   (65536)     (10669)      (3557)
        Map output bytes=119    (1039914)   (169209)     (56411)
        Map output materialized bytes=97
        Input split bytes=106   (188)       (188)        (186)
        Combine input records=13(0)         (0)          (0)
        Combine output records=8(0)         (0)          (0)
        Reduce input groups=8   (58)        (58)         (58)
        Reduce shuffle bytes=97
        Reduce input records=8  (65536)     (10669)      (3557)
        Reduce output records=8 (58)        (58)         (58)
        Spilled Records=16      (131072)    (21338)      (7114)
        Shuffled Maps =1        (2)         (2)          (2)
```

Unique Records:

```
Map-Reduce Framework
      Map input records=999                3501         65536
      Map output records=999               3501         65536
      Map output bytes=14866
      Map output materialized bytes=16876
      Input split bytes=190
      Combine input records=0
      Combine output records=0
      Reduce input groups=999              3501         65536
      Reduce shuffle bytes=16876           61911        1234077
      Reduce input records=999             3501         65536
      Reduce output records=999            3501         65536
      Spilled Records=1998                 7002         131072
      Shuffled Maps =2                     2            2
```

APACHE SPARK-IN MEMORY PROCESSING

- **Can Run 100 times faster than hadoop**
- **Support Real Time Stream Processing,Graph Analytics,Machine Learning and SQL.**
- **Pipe Oriented – create a pipe line e.g creat pipeline from spark streaming output to machine learning ML lib**

SPARK INSTALLATION

STEP 1:

- Download spark and add to PATH environment variable

STEP 2:

- install scala and add to PATH environment variable

STEP 3:

- open spark readme file and run spark source using mvn command given in the spark readme file according to spark version and hadoop version in use.

e.g like below

```
# Apache Hadoop 2.2.X

mvn -Pyarn -Phadoop-2.2 -Dhadoop.version=2.2.0 -DskipTests clean package

# Apache Hadoop 2.3.X

mvn -Pyarn -Phadoop-2.3 -Dhadoop.version=2.3.0 -DskipTests clean package
```

SPARK can be installed in 2 ways:

PROCEDURE 1:using pre-build hadoop with spark, scala – needs winutils to run hadoop on windows. Once installed hadoop , just download scala and spark and add to classpath will work.

PROCEDURE 2:separate spark,hadoop,sbt,scala installation- need to compile with SBT assembly

PROCEDURE 1:SPARK INSTALLATION

Install SBT:

since showing error in mvn with java , install SBT and install spark assembly.

TO RUN SPARK:

1)set SPARK_HOME environment variable.
If showing above error need to build spark.

Spark now comes packaged with a self-contained Maven installation to ease building and deployment of Spark from source located under the`build/` directory.

Mvn command:

```
mvn -Pyarn -Phadoop-2.2-Dhadoop.version=2.2.0 -DskipTests clean package
```

since i have Hadoop 2.2 on windows i prefer hadoop 2.2

```
mvn -X -Pyarn -Phadoop-2.2-Dhadoop.version=2.2.0 -DskipTests clean package
```

PROCEDURE 2:SPARK INSTALLATION

STEP 1:

on top of installed hadoop 2.2 or other which is installed Already install spark

```
C:\Windows\system32>path
PATH=C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;C:\Windows\System32
\WindowsPowerShell\v1.0\;C:\Program Files\Intel\WiFi\bin\;C:\Program Files\Commo
n Files\Intel\WirelessCommon\;C:\Program Files (x86)\Skype\Phone\;C:\apache-mave
n-3.3.9\bin;C:\protoc;C:\Program Files\Microsoft SDKs\Windows\v7.1\bin;C:\Progra
m Files\Git\bin;C:\Java\jdk1.7.0_79\bin;C:\Anaconda2;C:\Anaconda2\Library\bin;
C:\Anaconda2\Scripts;C:\Program Files\R\R-3.2.3\bin;C:\spark-1.6.0-bin-hadoop2.3\b
in;C:\scala-2.11.7\bin;C:\SBT-0.13\bin;C:\hadoop-2.2.0\bin;C:\hadoop-2.2.0\sbin
C:\Windows\system32>
```

Step 2:

set C:\spark-1.6.0-bin-hadoop2.3\sbin;C:\spark-1.6.0-bin-hadoop2.3\bin to class path

Spark can run hadoop built in

STEP 3:

install hadoop and run the hadoop once hadoop started , start spark-shell and run sample program.

Administrator: Windows Command Processor

```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Windows\system32>cd C:\spark-1.6.0-bin-hadoop2.3\sbin

C:\spark-1.6.0-bin-hadoop2.3\sbin>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\spark-1.6.0-bin-hadoop2.3\sbin>jps
5212 Jps
4748 sbt-launch.jar
4212 ResourceManager
5452 sbt-launch.jar
5712 SparkSubmit
4476 DataNode
5376 NameNode
1452 sbt-launch.jar
1652 NodeManager

C:\spark-1.6.0-bin-hadoop2.3\sbin>_
```

STEP 4:

```
Select C:\Windows\system32\cmd.exe - spark-shell
nto the
                                  working directory of each executor.
  --principal PRINCIPAL           Principal to be used to login to KDC, while runnin
g on
                                  secure HDFS.
  --keytab KEYTAB                 The full path to the file that contains the keytab
 for the
                                  principal specified above. This keytab will be cop
ied to
                                  the node running the Application Master via the Se
cure
                                  Distributed Cache, for renewing the login tickets
and the
                                  delegation tokens periodically.

C:\spark-1.6.0-bin-hadoop2.3\bin>spark-shell
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.li
b.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.propertie
s
To adjust logging level use sc.setLogLevel("INFO")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_79)
Type in expressions to have them evaluated.
Type :help for more information.
Spark context available as sc.
16/03/09 00:44:17 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is
 already registered. Ensure you dont have multiple JAR versions of the same plug
in in the classpath. The URL "file:/C:/spark-1.6.0-bin-hadoop2.3/bin/../lib/data
nucleus-rdbms-3.2.9.jar" is already registered, and you are trying to register a
n identical plugin located at URL "file:/C:/spark-1.6.0-bin-hadoop2.3/lib/datanu
cleus-rdbms-3.2.9.jar."
16/03/09 00:44:18 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is alr
eady registered. Ensure you dont have multiple JAR versions of the same plugin i
n the classpath. The URL "file:/C:/spark-1.6.0-bin-hadoop2.3/lib/datanucleus-api
-jdo-3.2.6.jar" is already registered, and you are trying to register an identic
al plugin located at URL "file:/C:/spark-1.6.0-bin-hadoop2.3/bin/../lib/datanucl
eus-api-jdo-3.2.6.jar."
16/03/09 00:44:18 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
istered. Ensure you dont have multiple JAR versions of the same plugin in the cl
asspath. The URL "file:/C:/spark-1.6.0-bin-hadoop2.3/bin/../lib/datanucleus-core
-3.2.10.jar" is already registered, and you are trying to register an identical
plugin located at URL "file:/C:/spark-1.6.0-bin-hadoop2.3/lib/datanucleus-core-3
.2.10.jar."
16/03/09 00:44:18 WARN Connection: BoneCP specified but not present in CLASSPATH
 (or one of dependencies)
```

STEP 5:

once spark shell started run sample program

Select Administrator: Windows Command Processor

```
C:\spark-1.6.0-bin-hadoop2.3\examples\src\main\java>spark-submit org/apache/spar
k/org/examples/JavaWordCount
Error: Cannot load main class from JAR file:/C:/spark-1.6.0-bin-hadoop2.3/exampl
es/src/main/java/org/apache/spark/org/examples/JavaWordCount
Run with --help for usage help or --verbose for debug output

C:\spark-1.6.0-bin-hadoop2.3\examples\src\main\java>cd ..

C:\spark-1.6.0-bin-hadoop2.3\examples\src\main>run-example SparkPi
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
16/03/09 01:12:23 INFO SparkContext: Running Spark version 1.6.0
16/03/09 01:12:25 INFO SecurityManager: Changing view acls to: Arun
16/03/09 01:12:25 INFO SecurityManager: Changing modify acls to: Arun
16/03/09 01:12:25 INFO SecurityManager: SecurityManager: authentication disabled
; ui acls disabled; users with view permissions: Set(Arun); users with modify pe
rmissions: Set(Arun)
16/03/09 01:12:26 INFO Utils: Successfully started service 'sparkDriver' on port
 63434.
16/03/09 01:12:27 INFO Slf4jLogger: Slf4jLogger started
16/03/09 01:12:28 INFO Remoting: Starting remoting
16/03/09 01:12:28 INFO Remoting: Remoting started; listening on addresses :[akka
.tcp://sparkDriverActorSystem@192.168.56.1:63447]
16/03/09 01:12:28 INFO Utils: Successfully started service 'sparkDriverActorSyst
em' on port 63447.
16/03/09 01:12:28 INFO SparkEnv: Registering MapOutputTracker
16/03/09 01:12:28 INFO SparkEnv: Registering BlockManagerMaster
16/03/09 01:12:28 INFO DiskBlockManager: Created local directory at C:\Users\Aru
n\AppData\Local\Temp\blockmgr-6bf528ac-71e3-4611-b96c-6cb98be0d769
16/03/09 01:12:28 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/03/09 01:12:28 INFO SparkEnv: Registering OutputCommitCoordinator
16/03/09 01:12:29 INFO Utils: Successfully started service 'SparkUI' on port 404
0.
16/03/09 01:12:29 INFO SparkUI: Started SparkUI at http://192.168.56.1:4040
16/03/09 01:12:29 INFO HttpFileServer: HTTP File server directory is C:\Users\Ar
un\AppData\Local\Temp\spark-42ffbf20-228a-47bb-a063-7dfe27815c5c\httpd-116064bc-
198a-4c9f-863f-293b09115d73
16/03/09 01:12:29 INFO HttpServer: Starting HTTP Server
16/03/09 01:12:29 INFO Utils: Successfully started service 'HTTP file server' on
 port 63450.
16/03/09 01:12:30 INFO SparkContext: Added JAR file:/C:/spark-1.6.0-bin-hadoop2.
3/bin/../lib/spark-examples-1.6.0-hadoop2.3.0.jar at http://192.168.56.1:63450/j
ars/spark-examples-1.6.0-hadoop2.3.0.jar with timestamp 1457466150856
16/03/09 01:12:31 INFO Executor: Starting executor ID driver on host localhost
16/03/09 01:12:31 INFO Utils: Successfully started service 'org.apache.spark.net
work.netty.NettyBlockTransferService' on port 63467.
16/03/09 01:12:31 INFO NettyBlockTransferService: Server created on 63467
16/03/09 01:12:31 INFO BlockManagerMaster: Trying to register BlockManager
16/03/09 01:12:31 INFO BlockManagerMasterEndpoint: Registering block manager loc
alhost:63467 with 511.5 MB RAM, BlockManagerId(driver, localhost, 63467)
16/03/09 01:12:31 INFO BlockManagerMaster: Registered BlockManager
16/03/09 01:12:32 INFO SparkContext: Starting job: reduce at SparkPi.scala:36
16/03/09 01:12:32 INFO DAGScheduler: Got job 0 (reduce at SparkPi.scala:36) with
 2 output partitions
16/03/09 01:12:32 INFO DAGScheduler: Final stage: ResultStage 0 (reduce at Spark
Pi.scala:36)
16/03/09 01:12:32 INFO DAGScheduler: Parents of final stage: List()
```