

## Table of Contents

PART ONE – HADOOP INSTALLATION .....	
PART TWO – ANALYSIS OF DATA SETS USING HADOOP .....	
PART THREE – COMPARISON OF TIME TAKEN.....	

Name: Aneesh Partha

CWID- A20376172

### Hadoop Installation:

Enabling port:

```
# accessing localhost:8080 will access port 80 on the guest machine.
config.vm.network "forwarded_port", guest: 80, host: 8080
config.vm.network "forwarded_port", guest: 50070, host: 50070
config.vm.network "forwarded_port", guest: 8088, host: 8088
config.vm.network "forwarded_port", guest: 19888, host: 19888
# Create a private network which allows host-only access to the machine
```

Aneesh Partha  
A20376172

Check for Java availability:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ java -version
Picked up _JAVA_OPTIONS: -Xmx4096m
java version "1.7.0_121"
OpenJDK Runtime Environment (IcedTea 2.6.8) (7u121-2.6.8-1ubuntu0.14.04.1)
OpenJDK 64-Bit Server VM (build 24.121-b00, mixed mode)
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$
```

Aneesh Partha  
A20376172

Check after Hadoop is installed in the system:

```
See http://www.oracle.com/technetwork/java/javase/documentation/index.html for more details.
vagrant@vagrant-ubuntu-trusty-64:/vagrant/github/apartha/ITMD521/week-03$ hadoop version
Picked up _JAVA_OPTIONS: -Xmx4096m
Hadoop 2.5.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r cc72e9b000545b86b75a61f483
Compiled by jenkins on 2014-11-14T23:45Z
Compiled with protoc 2.5.0
From source with checksum df7537a4faa4658983d397abf4514320
This command was run using /home/vagrant/hadoop-2.5.2/share/hadoop/common/hadoop-common-2.5.2.jar
vagrant@vagrant-ubuntu-trusty-64:/vagrant/github/apartha/ITMD521/week-03$
```

Aneesh Partha  
A20376172

Below paths are set in .bashrc:

```
export CLASSPATH=$CLASSPATH:/usr/share/java/mysql-connector-java.jar
export _JAVA_OPTIONS=-Xmx4096m
export JAVA_HOME=/usr
export HADOOP_HOME=~/.hadoop-2.5.2
export HADOOP_CLASSPATH=/usr/lib/jvm/java-7-openjdk-amd64/lib/tools.jar
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Aneesh Partha  
A20376172

Contents of hdfs-site.xml:

```
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
vagrant@vagrant-ubuntu-trusty-64:~/
```

Aneesh Partha  
A20376172

Contents of core-site.xml

```
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost/</value>
</property>
</configuration>
vagrant@vagrant-ubuntu-trusty-64:~/hadoop-2.5.2/etc/hadoop$
```

Contents of Mapred-site.xml:

```
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
~
~
~
```

Contents of yarn-site.xml:

```
<!-- Site specific YARN configuration properties -->
<property>
<name>yarn.resourcemanager.hostname</name>
<value>localhost</value>
</property>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
</configuration>
~
~
~
```

Processes running in the system:

```
vagrant@vagrant-ubuntu-trusty-64:~/hadoop-2.5.2/etc/hadoop$ jps
Picked up _JAVA_OPTIONS: -Xmx4096m
7292 Jps
5464 DataNode
5866 ResourceManager
5712 SecondaryNameNode
5305 NameNode
6026 NodeManager
6368 JobHistoryServer
vagrant@vagrant-ubuntu-trusty-64:~/hadoop-2.5.2/etc/hadoop$
```

Hadoop file system:

```
vagrant@vagrant-ubuntu-trusty-64:~/hadoop-2.5.2/etc/hadoop$ hadoop fs -ls /
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
drwxrwx--- - vagrant supergroup 0 2017-02-05 16:48 /tmp
drwxr-xr-x - vagrant supergroup 0 2017-02-05 16:51 /user
vagrant@vagrant-ubuntu-trusty-64:~/hadoop-2.5.2/etc/hadoop$
```

### Analysis of data sets using Hadoop:

Without combiner:

1990,1991,1992,1993 data sets:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperature /user
R/tempdata/1990/input /user/$USER/output
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 19:54:16 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 19:54:16 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and
ute your application with ToolRunner to remedy this.
17/02/10 19:54:17 INFO input.FileInputFormat: Total input paths to process : 4
17/02/10 19:54:17 INFO mapreduce.JobSubmitter: number of splits:115
17/02/10 19:54:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0007
17/02/10 19:54:17 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0007
17/02/10 19:54:17 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8032/application_1486313264135_0007
17/02/10 19:54:17 INFO mapreduce.Job: Running job: job_1486313264135_0007
17/02/10 19:54:27 INFO mapreduce.Job: Job job_1486313264135_0007 running in uber mode
17/02/10 19:54:27 INFO mapreduce.Job: map 0% reduce 0%
```

```

IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1
File Output Format Counters
  Bytes Written=1

real    16m14.509s
user    0m7.871s
sys     0m0.644s
vagrant@vagrant-ubuntu-trusty-64: /vagrant$

```

## Retired Jobs

Show 20 entries		Search:									
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.10 19:54:17 UTC	2017.02.10 19:54:25 UTC	2017.02.10 20:10:25 UTC	job_1486313264135_0007	Max temperature	vagrant	default	SUCCEEDED	115	115	1	1

```

vagrant@vagrant-ubuntu-trusty-64: /vagrant$ hadoop fs -cat /user/$USER/output/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990 607
1991 607
1992 605
1993 567

```

## 1990 and 1992 data sets:

```

vagrant@vagrant-ubuntu-trusty-64: /vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperature /user/$USER/tempdata/1990/input /user/$USER/output3
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 20:56:47 INFO client.RMPProxy: Connecting to ResourceManager at localhost/17/02/10 20:56:47 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing mis-
  ute your application with ToolRunner to remedy this.
17/02/10 20:56:47 INFO input.FileInputFormat: Total input paths to process : 2
17/02/10 20:56:47 INFO mapreduce.JobSubmitter: number of splits:60
17/02/10 20:56:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0010
17/02/10 20:56:48 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0010
17/02/10 20:56:48 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8021/
17/02/10 20:56:48 INFO mapreduce.Job: Running job: job_1486313264135_0010
17/02/10 20:56:56 INFO mapreduce.Job: Job job_1486313264135_0010 running in uber mode.
17/02/10 20:56:57 INFO mapreduce.Job: map 0% reduce 0%

```

```

BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=7993006187
File Output Format Counters
  Bytes Written=18

real    8m35.967s
user    0m5.892s
sys     0m0.503s
vagrant@vagrant-ubuntu-trusty-64: /vagrant$

```

## Retired Jobs

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.10 20:56:48 UTC	2017.02.10 20:56:54 UTC	2017.02.10 21:05:17 UTC	job_1486313264135_0010	Max temperature	vagrant	default	SUCCEEDED	60	60	1	1

```

deleted /user/vagrant/tempdata/1990/input
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ hbase fs -cat /user/$USER/output3/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990 607
1992 605
vagrant@vagrant-ubuntu-trusty-64:/vagrant$

```

## 1990 data set

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperature /user/$USER/output5
R/tempdata/1990/input /user/$USER/output5
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 21:18:09 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8022
17/02/10 21:18:10 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Your application may receive command-line options that no
ute your application with ToolRunner to remedy this.
17/02/10 21:18:10 INFO input.FileInputFormat: Total input paths to process : 1
17/02/10 21:18:10 INFO mapreduce.JobSubmitter: number of splits:8
17/02/10 21:18:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0012
17/02/10 21:18:11 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0012
17/02/10 21:18:11 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8021/job_1486313264135_0012/
17/02/10 21:18:11 INFO mapreduce.Job: Running job: job_1486313264135_0012
17/02/10 21:18:18 INFO mapreduce.Job: Job job_1486313264135_0012 running in uber mode : false
17/02/10 21:18:18 INFO mapreduce.Job: map 0% reduce 0%

```

```

File Output Format
Bytes
real    1m31.243s
user    0m4.165s
sys     0m0.187s
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$

```

## Retired Jobs

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.10 21:18:11 UTC	2017.02.10 21:18:17 UTC	2017.02.10 21:19:36 UTC	job_1486313264135_0012	Max temperature	vagrant	default	SUCCEEDED	8	8	1	1

```

deleted /user/vagrant/tempdata/1990/input
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ hbase fs -cat /user/$USER/output5/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990 607
vagrant@vagrant-ubuntu-trusty-64:/vagrant$

```

Name: Aneesh Partha

CWID- A20376172

With combiner:

1990,1991,1992 and 1993 data sets

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTem
r /user/$USER/tempdata/1990/input /user/$USER/output2
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 20:38:36 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 20:38:37 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool inter
ute your application with ToolRunner to remedy this.
17/02/10 20:38:37 INFO input.FileInputFormat: Total input paths to process : 4
17/02/10 20:38:37 INFO mapreduce.JobSubmitter: number of splits:115
17/02/10 20:38:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0009
17/02/10 20:38:38 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0009
17/02/10 20:38:38 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8042/jobdetails/application_1486313264135_0009/
17/02/10 20:38:38 INFO mapreduce.Job: Running job: job_1486313264135_0009
17/02/10 20:38:46 INFO mapreduce.Job: Job job_1486313264135_0009 running in uber mode.
17/02/10 20:38:46 INFO mapreduce.Job: map 0% reduce 0%

```

```

WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1521
File Output Format Counters
  Bytes Written=1521
real    14m48.458s
user    0m7.117s
sys     0m0.717s
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$

```

Retired Jobs

Show 20 entries		Search:									
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Complete
2017.02.10 20:38:38 UTC	2017.02.10 20:38:44 UTC	2017.02.10 20:53:20 UTC	job_1486313264135_0009	Max temperature	vagrant	default	SUCCEEDED	115	115	1	1

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop fs -cat /user/$USER/output2/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1991    607
1992    605
1993    567
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$

```

1990,1992 data sets

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperature
r /user/$USER/tempdata/1990/input /user/$USER/output4
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 21:07:29 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 21:07:30 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool inter
ute your application with ToolRunner to remedy this.
17/02/10 21:07:30 INFO input.FileInputFormat: Total input paths to process : 2
17/02/10 21:07:30 INFO mapreduce.JobSubmitter: number of splits:60
17/02/10 21:07:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0011
17/02/10 21:07:31 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0011
17/02/10 21:07:31 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8042/jobdetails/application_1486313264135_0011/
17/02/10 21:07:31 INFO mapreduce.Job: Running job: job_1486313264135_0011
17/02/10 21:07:39 INFO mapreduce.Job: Job job_1486313264135_0011 running in uber mode.
17/02/10 21:07:39 INFO mapreduce.Job: map 0% reduce 0%

```

```

WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=7993006187
File Output Format Counters
  Bytes Written=18

real    8m1.136s
user    0m5.671s
sys     0m0.474s
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ ./mr-intro/src/main/java$

```

## Retired Jobs

Show 20 entries		Search:									
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.10 21:07:31 UTC	2017.02.10 21:07:37 UTC	2017.02.10 21:15:25 UTC	<a href="#">job_1486313264135_0011</a>	Max temperature	vagrant	default	SUCCEEDED	60	60	1	1

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant$ ./mr-intro/src/main/java$
Picked up _JAVA_OPTIONS: -Xmx4096m
1990 607
1992 605
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ ./mr-intro/src/main/java$

```

## 1990 data set

```

vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperatureWith
r /user/$USER/tempdata/1990/input /user/$USER/output6
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 21:22:10 INFO client.RMProxy: Connecting to ResourceManager at localhost:8020
17/02/10 21:22:10 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing at the Tool interface
ute your application with ToolRunner to remedy this.
17/02/10 21:22:10 INFO input.FileInputFormat: Total input paths to process : 1
17/02/10 21:22:10 INFO mapreduce.JobSubmitter: number of splits:8
17/02/10 21:22:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486313264135_0013
17/02/10 21:22:11 INFO impl.YarnClientImpl: Submitted application application_1486313264135_0013
17/02/10 21:22:11 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8020/jobdetails/application_1486313264135_0013/
17/02/10 21:22:11 INFO mapreduce.Job: Running job: job_1486313264135_0013

```

```

CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1030902727
File Output Format Counters
  Bytes Written=9

real    1m22.875s
user    0m4.075s
sys     0m0.206s
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ ./mr-intro/src/main/java$

```

## Retired Jobs

Show 20 entries		Search:									
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.10 21:22:11 UTC	2017.02.10 21:22:17 UTC	2017.02.10 21:23:27 UTC	<a href="#">job_1486313264135_0013</a>	Max temperature	vagrant	default	SUCCEEDED	8	8	1	1



Name: Aneesh Partha

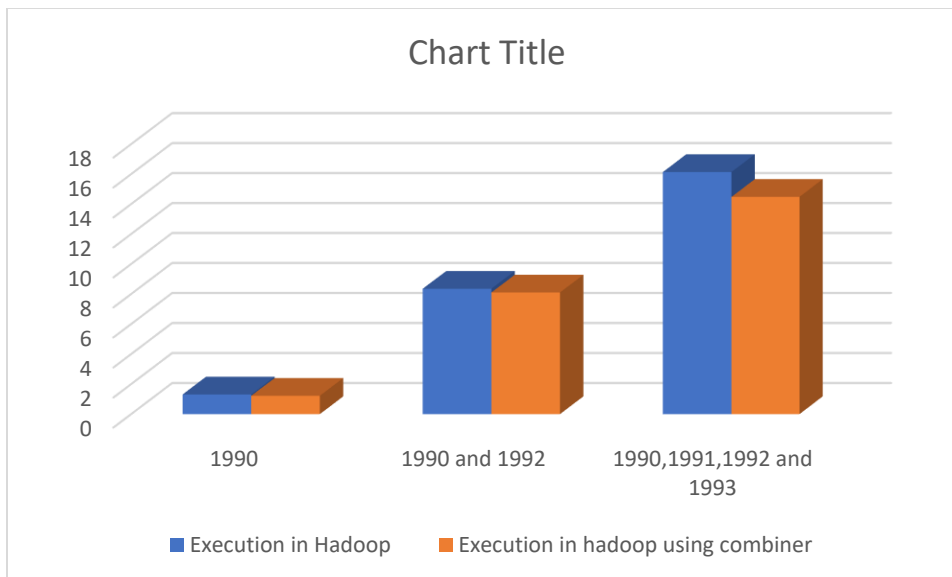
CWID- A20376172

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ hadoop + ... output6/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990 607
vagrant@vagrant-ubuntu-trusty-64:/vagrant$
```

Aneesh Partha  
A20376172

**Comparison of time taken:**

Dataset	Execution in Hadoop	Execution in hadoop using combiner
1990	1.31	1.22
1990 and 1992	8.35	8.1
1990,1991,1992 and 1993	16.14	14.48



As we can see from the above graph the time taken for execution using combiner is less when compared to time taken for execution without combiner. This is due to the fact that a partial reduce operation is completed in mapper phase. For smaller data sets the difference is not visible that much. But as the data sets are bigger the combiner plays an important role. As you can see the time different for 1990 small data set is negligible but the time difference for the larger data sets is more.