ASSIGNMENT III

1) Against the dataset 1990.

The below are the screen shots of the screen shot which is obtained when we run the jar of 1990 dataset.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar M
axTemperature /user/$USER/tempdata/1990/1990 /user/$USER/output3
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/09 21:24:36 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/09 21:24:36 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/09 21:24:37 INFO input.FileInputFormat: Total input paths to process : 1
17/02/09 21:24:37 INFO mapreduce.JobSubmitter: number of splits:8
17/02/09 21:24:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486674570489_0001
17/02/09 21:24:38 INFO impl.YarnClientImpl: Submitted application application_1486674570489_0001
17/02/09 21:24:38 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486674570489_0001/
17/02/09 21:24:38 INFO mapreduce.Job: Running job: job_1486674570489_0001
17/02/09 21:24:48 INFO mapreduce.Job: Job job_1486674570489_0001 running in uber mode : false
17/02/09 21:24:48 INFO mapreduce.Job:  map 0% reduce 0%
17/02/09 21:25:25 INFO mapreduce.Job:  map 1% reduce 0%
17/02/09 21:25:28 INFO mapreduce.Job:  map 2% reduce 0%
17/02/09 21:25:31 INFO mapreduce.Job:  map 8% reduce 0%
17/02/09 21:25:32 INFO mapreduce.Job:  map 14% reduce 0%
17/02/09 21:25:35 INFO mapreduce.Job:  map 23% reduce 0%
17/02/09 21:25:36 INFO mapreduce.Job:  map 25% reduce 0%
17/02/09 21:25:38 INFO mapreduce.Job:  map 35% reduce 0%
17/02/09 21:25:39 INFO mapreduce.Job:  map 37% reduce 0%
17/02/09 21:25:41 INFO mapreduce.Job:  map 46% reduce 0%
17/02/09 21:25:42 INFO mapreduce.Job:  map 48% reduce 0%
17/02/09 21:25:44 INFO mapreduce.Job:  map 54% reduce 0%
17/02/09 21:25:47 INFO mapreduce.Job:  map 67% reduce 0%
17/02/09 21:25:48 INFO mapreduce.Job:  map 75% reduce 0%
17/02/09 21:26:08 INFO mapreduce.Job:  map 76% reduce 0%
17/02/09 21:26:10 INFO mapreduce.Job:  map 77% reduce 0%
17/02/09 21:26:11 INFO mapreduce.Job:  map 81% reduce 25%
17/02/09 21:26:13 INFO mapreduce.Job:  map 88% reduce 25%
17/02/09 21:26:14 INFO mapreduce.Job:  map 91% reduce 25%
17/02/09 21:26:15 INFO mapreduce.Job:  map 100% reduce 25%
17/02/09 21:26:18 INFO mapreduce.Job:  map 100% reduce 56%
17/02/09 21:26:21 INFO mapreduce.Job:  map 100% reduce 84%
17/02/09 21:26:22 INFO mapreduce.Job:  map 100% reduce 100%
17/02/09 21:26:23 INFO mapreduce.Job: Job job_1486674570489_0001 completed successfully
17/02/09 21:26:23 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=51062743
                FILE: Number of bytes written=102995989
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1030903631
                HDFS: Number of bytes written=9
                HDFS: Number of read operations=27
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

The job history obtained for running the MaxTemprature file. It is clearly seen that it takes 95 sec

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
2017.02.09 21:24:37 UTC	2017.02.09 21:24:46 UTC	2017.02.09 21:26:21 UTC	job_1486674570489_0001	Max temperature	vagrant	default	SUCCEEDED	8	8	1	1
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Tota	Maps Complete	Reduces Tota	Reduces Comp

On Running the class MaxTempratureWithCombiner jar file for the 1990 dataset.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main$ cd java
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/tempdata/1990/1990 /user/$USER/output13
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 08:49:27 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 08:49:28 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/10 08:49:28 INFO input.FileInputFormat: Total input paths to process : 1
17/02/10 08:49:28 INFO mapreduce.JobSubmitter: number of splits:8
17/02/10 08:49:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486699840359_0011
17/02/10 08:49:29 INFO impl.YarnClientImpl: Submitted application application_1486699840359_0011
17/02/10 08:49:29 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486699840359_0011/
17/02/10 08:49:29 INFO mapreduce.Job: Running job: job_1486699840359_0011
17/02/10 08:49:37 INFO mapreduce.Job: Job job_1486699840359_0011 running in uber mode : false
17/02/10 08:49:37 INFO mapreduce.Job:  map 0% reduce 0%
17/02/10 08:50:06 INFO mapreduce.Job:  map 2% reduce 0%
17/02/10 08:50:09 INFO mapreduce.Job:  map 15% reduce 0%
17/02/10 08:50:12 INFO mapreduce.Job:  map 32% reduce 0%
17/02/10 08:50:16 INFO mapreduce.Job:  map 50% reduce 0%
17/02/10 08:50:18 INFO mapreduce.Job:  map 54% reduce 0%
```

The time taken to run the dataset for the year 1990 is 61sec.

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 08:49:29 UTC	2017.02.10 08:49:35 UTC	2017.02.10 08:50:36 UTC	job_1486699840359_0011	Max temperature	vagrant	default	SUCCEEDED	8	8	1	1

The output for the year 1990 is show below in the screenshot.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user/$USER/output13
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup          0 2017-02-10 08:50 /user/vagrant/output13/_SUCCESS
-rw-r--r--   1 vagrant supergroup          9 2017-02-10 08:50 /user/vagrant/output13/part-r-00000
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/output13/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$
```

2) Against the dataset 1990 and 1992.

On Running the class MaxTemprature jar file for the 1990 and 1992 dataset.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user/$USER/tempdata/1990-92
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup 1030874055 2017-02-10 05:57 /user/vagrant/tempdata/1990-92/1990
-rw-r--r--   1 vagrant supergroup 6961894564 2017-02-10 05:56 /user/vagrant/tempdata/1990-92/1992
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar MaxTemperature /user/$USER/tempdata/1990-92/* /user/$USER/output9
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 06:15:08 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 06:15:08 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/10 06:15:08 INFO input.FileInputFormat: Total input paths to process : 2
17/02/10 06:15:09 INFO mapreduce.JobSubmitter: number of splits:60
17/02/10 06:15:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486699840359_0007
17/02/10 06:15:10 INFO impl.YarnClientImpl: Submitted application application_1486699840359_0007
17/02/10 06:15:10 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486699840359_0007/
17/02/10 06:15:10 INFO mapreduce.Job: Running job: job_1486699840359_0007
17/02/10 06:15:20 INFO mapreduce.Job: Job job_1486699840359_0007 running in uber mode : false
```

The time taken to run the dataset for the year 1990 and 1992 is 469sec.

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 08:29:07 UTC	2017.02.10 08:29:13 UTC	2017.02.10 08:36:06 UTC	job_1486699840359_0010	Max temperature	vagrant	default	SUCCEEDED	60	60	1	1

The output for the two years 1990 and 1992.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user
/$USER/output9
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup          0 2017-02-10 06:23 /user/vagrant/output9/_SUCCESS
-rw-r--r--   1 vagrant supergroup         18 2017-02-10 06:23 /user/vagrant/output9/part-r-00000
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /use
r/$USER/output9/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1992    605
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$
```

On Running the class MaxTempratureWithCombiner jar file for the 1990 and 1992 dataset.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/tempdata/1990-92/* /user/$USER/output12
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 08:29:06 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 08:29:06 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/10 08:29:06 INFO input.FileInputFormat: Total input paths to process : 2
17/02/10 08:29:07 INFO mapreduce.JobSubmitter: number of splits:60
17/02/10 08:29:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486699840359_0010
17/02/10 08:29:07 INFO impl.YarnClientImpl: Submitted application application_1486699840359_0010
17/02/10 08:29:07 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486699840359_0010/
17/02/10 08:29:07 INFO mapreduce.Job: Running job: job_1486699840359_0010
17/02/10 08:29:15 INFO mapreduce.Job: Job job_1486699840359_0010 running in uber mode : false
17/02/10 08:29:15 INFO mapreduce.Job:  map 0% reduce 0%
17/02/10 08:29:47 INFO mapreduce.Job:  map 1% reduce 0%
17/02/10 08:29:48 INFO mapreduce.Job:  map 2% reduce 0%
17/02/10 08:29:50 INFO mapreduce.Job:  map 3% reduce 0%
17/02/10 08:29:51 INFO mapreduce.Job:  map 4% reduce 0%
17/02/10 08:29:56 INFO mapreduce.Job:  map 5% reduce 0%
```

The time taken for running the dataset 1990 and 1992 is 353sec.

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 08:29:07 UTC	2017.02.10 08:29:13 UTC	2017.02.10 08:36:06 UTC	job_1486699840359_0010	Max temperature	vagrant	default	SUCCEEDED	60	60	1	1

Output is:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user/$USER/output12
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup          0 2017-02-10 08:36 /user/vagrant/output12/_SUCCESS
-rw-r--r--   1 vagrant supergroup         18 2017-02-10 08:36 /user/vagrant/output12/part-r-00000
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/output12/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1992    605
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$
```

3) Against the dataset for all the 4 years.

Run the jar MaxTemperature for all the four years.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar M
axTemperature /user/$USER/tempdata/19all/* /user/$USER/output10
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 07:46:41 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 07:46:42 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/10 07:46:42 INFO input.FileInputFormat: Total input paths to process : 4
17/02/10 07:46:42 INFO mapreduce.JobSubmitter: number of splits:115
17/02/10 07:46:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486699840359_0008
17/02/10 07:46:43 INFO impl.YarnClientImpl: Submitted application application_1486699840359_0008
17/02/10 07:46:43 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486699840359_0008/
17/02/10 07:46:43 INFO mapreduce.Job: Running job: job_1486699840359_0008
17/02/10 07:46:50 INFO mapreduce.Job: Job job_1486699840359_0008 running in uber mode : false
17/02/10 07:46:50 INFO mapreduce.Job:  map 0% reduce 0%
17/02/10 07:47:23 INFO mapreduce.Job:  map 1% reduce 0%
17/02/10 07:47:27 INFO mapreduce.Job:  map 2% reduce 0%
17/02/10 07:47:32 INFO mapreduce.Job:  map 3% reduce 0%
17/02/10 07:47:40 INFO mapreduce.Job:  map 5% reduce 0%
17/02/10 07:48:15 INFO mapreduce.Job:  map 6% reduce 0%
17/02/10 07:48:18 INFO mapreduce.Job:  map 7% reduce 0%
```

The time taken is 878sec.

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 07:46:43 UTC	2017.02.10 07:46:48 UTC	2017.02.10 08:01:26 UTC	job_1486699840359_0008	Max temperature	vagrant	default	SUCCEEDED	115	115	1	1

The output for the years 1990-1993.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user
/$USER/output10
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup          0 2017-02-10 08:01 /user/vagrant/output10/_SUCCESS
-rw-r--r--   1 vagrant supergroup         36 2017-02-10 08:01 /user/vagrant/output10/part-r-00000
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /use
r/$USER/output10/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1991    607
1992    605
1993    567
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$
```

On Running the class MaxTempratureWithCombiner jar file for the 1990 t0 1994.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/tempdata/19all/* /user/$USER/output11
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/10 08:12:01 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/10 08:12:01 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/10 08:12:02 INFO input.FileInputFormat: Total input paths to process : 4
17/02/10 08:12:02 INFO mapreduce.JobSubmitter: number of splits:115
17/02/10 08:12:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486699840359_0009
17/02/10 08:12:02 INFO impl.YarnClientImpl: Submitted application application_1486699840359_0009
17/02/10 08:12:02 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1486699840359_0009/
17/02/10 08:12:02 INFO mapreduce.Job: Running job: job_1486699840359_0009
17/02/10 08:12:10 INFO mapreduce.Job: Job job_1486699840359_0009 running in uber mode : false
17/02/10 08:12:10 INFO mapreduce.Job:  map 0% reduce 0%
17/02/10 08:12:43 INFO mapreduce.Job:  map 1% reduce 0%
17/02/10 08:12:49 INFO mapreduce.Job:  map 2% reduce 0%
17/02/10 08:12:55 INFO mapreduce.Job:  map 3% reduce 0%
17/02/10 08:12:59 INFO mapreduce.Job:  map 4% reduce 0%
17/02/10 08:13:00 INFO mapreduce.Job:  map 5% reduce 0%
```

The time taken is 805 sec.

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 08:12:02 UTC	2017.02.10 08:12:08 UTC	2017.02.10 08:25:33 UTC	job_1486699840359_0009	Max temperature	vagrant	default	SUCCEEDED	115	115	1	1

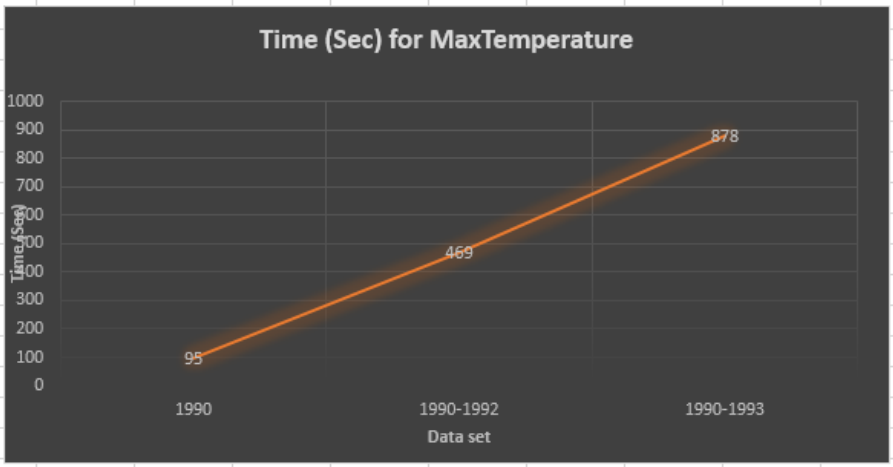
The output for all the four years.

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -ls /user/$USER/output11
Picked up _JAVA_OPTIONS: -Xmx4096m
Found 2 items
-rw-r--r--   1 vagrant supergroup          0 2017-02-10 08:25 /user/vagrant/output11/_SUCCESS
-rw-r--r--   1 vagrant supergroup         36 2017-02-10 08:25 /user/vagrant/output11/part-r-00000
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/output11/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1991    607
1992    605
1993    567
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/hadoop_book/hadoop-book/ch02-mr-intro/src/main/java$
```

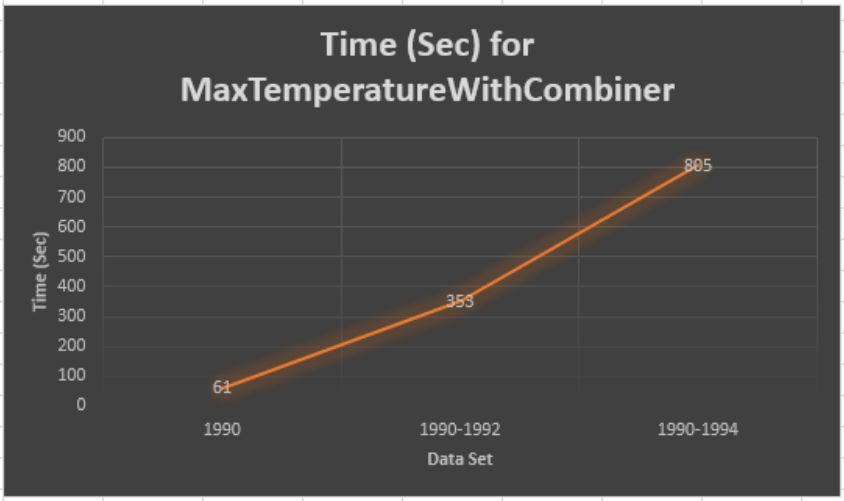
Analysis:

Line Graph

DataSet	Time (Sec)
1990	95
1990-1992	469
1990-1993	878



DataSet	Time(Sec)
1990	61
1990-1992	353
1990-1994	805

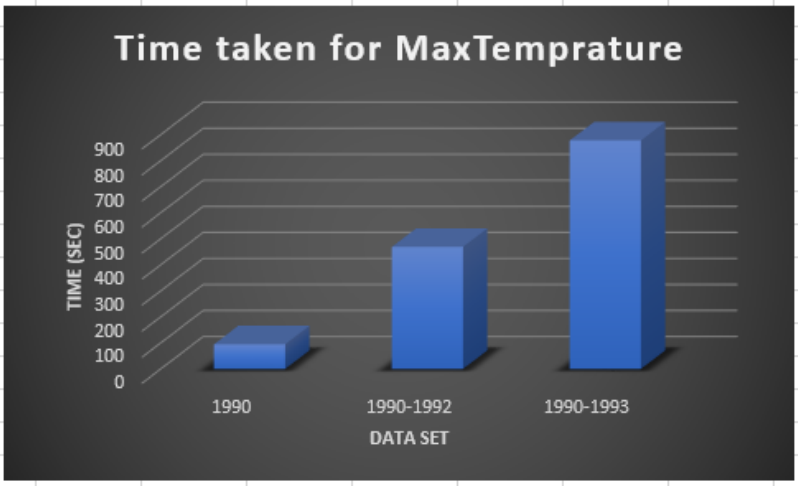


The line graphs above clearly show that in Hadoop the processing of data increases linearly.

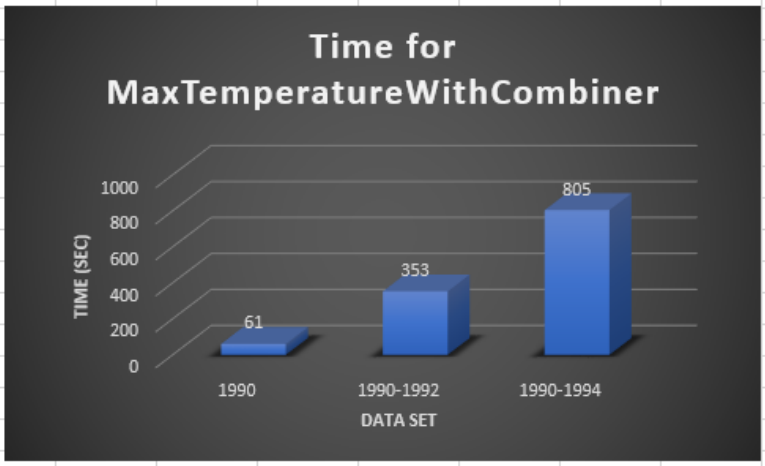
Bar Graph

The bar graph is used for data comparison.

DataSet	Time (Sec)
1990	95
1990-1992	469
1990-1993	878



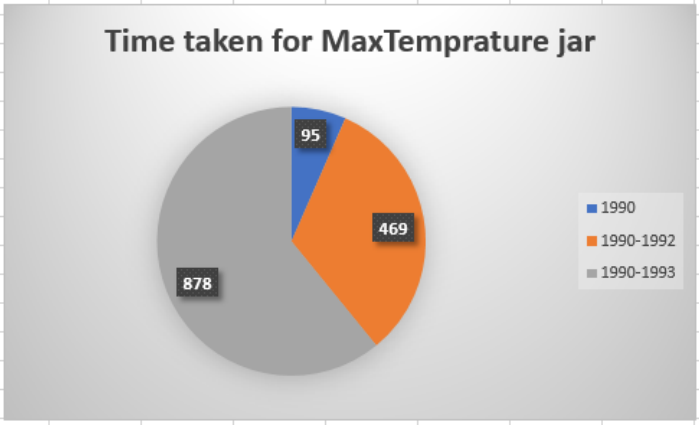
DataSet	Time(Sec)
1990	61
1990-1992	353
1990-1994	805



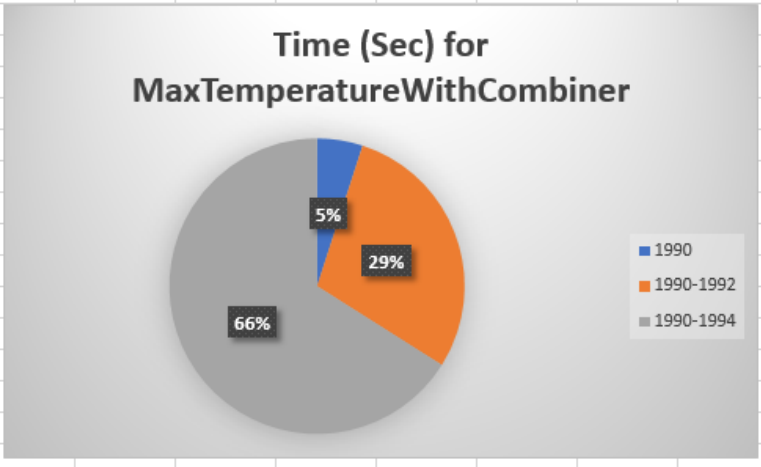
Pie Chart

The pie chart is used to know the percentage the data uses to the whole percentage.

DataSet	Time (Sec)
1990	95
1990-1992	469
1990-1993	878



DataSet	Time(Sec)
1990	61
1990-1992	353
1990-1994	805



- From the above graphs, it is very clear that the time taken to execute jobs in Hadoop is a linear curve to the one seen to the before exercise. In the previous exercise while using java and MySQL it was clear that the time taken increases exponentially. So, for large dataset Hadoop MapReduce works fine when compared to other databases.

- The main reason MapReduce works fine when compared to other database is due to the local cluster that it creates with data nodes and the job is split amongst these nodes. Thus, the tasks are executed parallelly. Since simultaneous work is being done the time taken reduces.
- The simultaneous running of job MapReduce manages by using two phase the Map and Reduce phase. In the Map phase records are read and processed simultaneously and the reduce phase reassembles them.
- It is clearly visible that the time take to execute MaxTemperatureWithCombiner is much lesser compared to MaxTemprature. The main reason here would be the optimized code.