

ITMD 521 – WEEK-03-HADOOP CLUSTER INSTALL

1. Single node cluster installation

```
vagrant@precise64:~$ hadoop version
Hadoop 2.5.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r cc72e9b000545b86b75a61f4835eb86d57bfafc0
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@precise64:~$
```

2. Data set used in week 02 is inserted into HDFS

```
vagrant@precise64:~$ hadoop fs -ls /user/$USER/pradeep/input
Found 4 items
-rw-r--r-- 1 vagrant supergroup 1030874055 2017-02-12 00:37 /user/vagrant/pradeep/input/1990
-rw-r--r-- 1 vagrant supergroup 3760031646 2017-02-12 00:44 /user/vagrant/pradeep/input/1991
-rw-r--r-- 1 vagrant supergroup 6961894564 2017-02-12 00:51 /user/vagrant/pradeep/input/1992
-rw-r--r-- 1 vagrant supergroup 3504568450 2017-02-12 00:54 /user/vagrant/pradeep/input/1993
vagrant@precise64:~$
```

3. Running the MaxTemperature without combiner against 1990 dataset

```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperature /user/$USER/pradeep/input/1990 /user/$USER/pradeep/output
17/02/11 02:11:01 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 02:11:02 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy
17/02/11 02:11:02 INFO InputFileInputFormat: Total input paths to process : 1
17/02/11 02:11:02 INFO mapreduce.JobSubmitter: number of splits:8
17/02/11 02:11:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0001
17/02/11 02:11:03 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0001
17/02/11 02:11:03 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0001/
17/02/11 02:11:13 INFO mapreduce.Job: Job job_1486778797450_0001 running in uber mode : false
17/02/11 02:11:13 INFO mapreduce.Job: map 0% reduce 0%
17/02/11 02:11:35 INFO mapreduce.Job: map 6% reduce 0%
17/02/11 02:11:36 INFO mapreduce.Job: map 8% reduce 0%
17/02/11 02:11:37 INFO mapreduce.Job: map 9% reduce 0%
17/02/11 02:11:38 INFO mapreduce.Job: map 18% reduce 0%
17/02/11 02:11:39 INFO mapreduce.Job: map 29% reduce 0%
real    1m9.990s
user    0m4.996s
sys     0m0.372s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$
```

- Time taken for running the command is 1 minute 9 seconds

```
File Output Format Counters
Bytes Written=9
real    1m9.990s
user    0m4.996s
sys     0m0.372s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$
```

- MaxTemperature result without combiner is **1990 - 607**

```
-rw-r--r-- 1 vagrant supergroup 9 2017-02-11 02:12 /user/vagrant/pradeep/output/part-r-00000
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output/part-r-00000
1990 607
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$
```

- Job History displaying the time taken for running the MaxTemperature without combiner against 1990



JobHistory

▼ Application	Retired Jobs							
About Jobs	Show 20 ▼ entries							
Tools	Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
	2017.02.11 02:11:03 UTC	2017.02.11 02:11:11 UTC	2017.02.11 02:12:07 UTC	job_1486778797450_0001	Max temperature	vagrant	default	SUCCEEDED
	Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
	Showing 1 to 1 of 1 entries							

4. Running MaxTemperatureWithCombiner against 1990

```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/pradeep/input/1990 /user/$USER/pradeep/output1
17/02/11 02:21:19 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 02:21:20 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/11 02:21:20 INFO input.FileInputFormat: Total input paths to process : 1
17/02/11 02:21:20 INFO mapreduce.JobSubmitter: number of splits:8
17/02/11 02:21:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0002
17/02/11 02:21:20 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0002
17/02/11 02:21:21 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0002/
17/02/11 02:21:21 INFO mapreduce.Job: Running job: job_1486778797450_0002
17/02/11 02:21:28 INFO mapreduce.Job: Job job_1486778797450_0002 running in uber mode : false
17/02/11 02:21:28 INFO mapreduce.Job: map 0% reduce 0%
17/02/11 02:21:51 INFO mapreduce.Job: map 9% reduce 0%
17/02/11 02:21:52 INFO mapreduce.Job: map 12% reduce 0%
17/02/11 02:21:53 INFO mapreduce.Job: map 14% reduce 0%
```

- Time taken for running the command is 1 minute 3 seconds

```
Bytes Written=9
real    1m3.949s
user    0m4.952s
sys     0m0.352s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ |
```

- MaxTemperatureWithCombiner result with combiner is **1990 - 607**

```
File Output Format Counters
  Bytes Written=9
real    1m3.949s
user    0m4.952s
sys     0m0.352s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output1/part-r-00000
1990
607
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ |
```

- Job History displaying the time taken for running the MaxTemperatureWithCombiner against 1990



JobHistory

▼ Application	Retired Jobs							
About Jobs	Show 20 ▼ entries							
Tools	Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
	2017.02.11 02:21:20 UTC	2017.02.11 02:21:26 UTC	2017.02.11 02:22:18 UTC	job_1486778797450_0002	Max temperature	vagrant	default	SUCCEEDED

5. Running the MaxTemperature without combiner against 1990 and 1992

```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperature /user/$USER/pradeep/input/* /user/$USER/pradeep/output2
17/02/11 02:37:47 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 02:37:47 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/11 02:37:48 INFO input.FileInputFormat: Total input paths to process : 2
17/02/11 02:37:48 INFO mapreduce.JobSubmitter: number of splits:60
17/02/11 02:37:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0003
17/02/11 02:37:49 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0003
17/02/11 02:37:49 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0003/
17/02/11 02:37:49 INFO mapreduce.Job: Running job: job_1486778797450_0003
17/02/11 02:37:58 INFO mapreduce.Job: Job job_1486778797450_0003 running in uber mode : false
17/02/11 02:38:25 INFO mapreduce.Job: map 0% reduce 0%
17/02/11 02:38:28 INFO mapreduce.Job: map 1% reduce 0%
17/02/11 02:38:28 INFO mapreduce.Job: map 2% reduce 0%
```

- Time taken for running the command is 6 minute 55 seconds

```
File Output Format Counters
  Bytes Written=18
real    6m55.606s
user    0m6.800s
sys     0m0.920s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ |
```


- MaxTemperature result without combiner against 1990 and 1992

1990 - 607

1992 - 605

```
sys    0m0.920s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output2/part-r-00000
1990    607
1992    605
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ |
```

- Job History displaying the time taken for running the MaxTemperature Without Combiner against 1990 and 1992



JobHistory

▼ Application

About Jobs

Tools

Retired Jobs

Show 20 ▼ entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
2017-02-11 02:37:49 UTC	2017-02-11 02:37:56 UTC	2017-02-11 02:44:37 UTC	job_1486778797450_0003	Max temperature	vagrant	default	SUCCEEDED

6. Running MaxTemperatureWithCombiner from the compiled jar file against 1990 and 1992


```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/pradeep/input/* /user/$USER/pradeep/output3
17/02/11 02:47:27 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 02:47:28 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/11 02:47:28 INFO input.FileInputFormat: Total input paths to process : 2
17/02/11 02:47:28 INFO mapreduce.JobSubmitter: number of splits:60
17/02/11 02:47:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0004
17/02/11 02:47:29 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0004
17/02/11 02:47:29 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0004/
17/02/11 02:47:29 INFO mapreduce.Job: Running job: job_1486778797450_0004
17/02/11 02:47:37 INFO mapreduce.Job: Job job_1486778797450_0004 running in uber mode : false
17/02/11 02:47:37 INFO mapreduce.Job: map 0% reduce 0%
```

- Time taken for running the command is 6 minute 5 seconds

```
File Output Format Counters
  Bytes Written=18
real    6m5.265s
user    0m6.216s
sys     0m0.904s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output3/part-r-00000
1990    607
1992    605
```

- Result observed: 1990 – 607, 1992 – 605

- Job History displaying the time taken for running the MaxTemperatureWithCombiner against 1990 and 1991



JobHistory

Application

About
Jobs

Tools

Retired Jobs

Show 20 entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
2017-02-11 02:47:29 UTC	2017-02-11 02:47:36 UTC	2017-02-11 02:53:28 UTC	job_1486778797450_0004	Max temperature	vagrant	default	SUCCEEDED


7. Running the MaxTemperature without combiner against 1990, 1991, 1992,1993

```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperature /user/$USER/pradeep/input/* /user/$USER/pradeep/output4
17/02/11 03:10:22 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 03:10:23 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/11 03:10:23 INFO input.FileInputFormat: Total input paths to process : 4
17/02/11 03:10:24 INFO mapreduce.JobSubmitter: number of splits:115
17/02/11 03:10:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0005
17/02/11 03:10:24 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0005
17/02/11 03:10:25 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0005/
17/02/11 03:10:25 INFO mapreduce.Job: Running job: job_1486778797450_0005
17/02/11 03:10:34 INFO mapreduce.Job: Job job_1486778797450_0005 running in uber mode : false
17/02/11 03:10:34 INFO mapreduce.Job: map 0% reduce 0%
17/02/11 03:10:57 INFO mapreduce.Job: map 1% reduce 0%
17/02/11 03:11:05 INFO mapreduce.Job: map 2% reduce 0%
17/02/11 03:11:08 INFO mapreduce.Job: map 3% reduce 0%
17/02/11 03:11:13 INFO mapreduce.Job: map 4% reduce 0%
17/02/11 03:11:14 INFO mapreduce.Job: map 5% reduce 0%
17/02/11 03:11:34 INFO mapreduce.Job: map 6% reduce 0%
```

- Time taken for running the command is 12 minute 34 seconds

```
bytes written=30
real    12m34.634s
user    0m9.521s
sys     0m1.648s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output4/part-r-00000
1990      607
1991      607
1992      605
1993      567
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ |
```

- Result observed: 1990 – 607, 1991 - 607, 1992 – 605, 1993 – 567
- Job History displaying the time taken for running the MaxTemperature without combiner against 1990, 1991, 1992,1993



JobHistory

Application

About
Jobs

Tools

Retired Jobs

Show 20 entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
2017-02-11 03:10:24 UTC	2017-02-11 03:10:32 UTC	2017-02-11 03:22:51 UTC	job_1486778797450_0005	Max temperature	vagrant	default	SUCCEEDED

8. Running the MaxTemperatureWithCombiner against 1990, 1991, 1992,1993

```
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mt.jar MaxTemperatureWithCombiner /user/$USER/pradeep/input/* /user/$USER/pradeep/output5
17/02/11 03:25:30 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/11 03:25:30 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/11 03:25:31 INFO input.FileInputFormat: Total input paths to process : 4
17/02/11 03:25:32 INFO mapreduce.JobSubmitter: number of splits:115
17/02/11 03:25:32 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486778797450_0006
17/02/11 03:25:32 INFO impl.YarnClientImpl: Submitted application application_1486778797450_0006
17/02/11 03:25:33 INFO mapreduce.Job: The url to track the job: http://precise64:8088/proxy/application_1486778797450_0006/
17/02/11 03:25:33 INFO mapreduce.Job: Running job: job_1486778797450_0006
17/02/11 03:25:41 INFO mapreduce.Job: Job job_1486778797450_0006 running in uber mode : false
17/02/11 03:25:41 INFO mapreduce.Job: map 0% reduce 0%
```

- Time taken for running the command is 10 minute 43 seconds

```
real    10m43.787s
user    0m8.225s
sys     0m1.400s
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$ hadoop fs -cat /user/$USER/pradeep/output4/part-r-00000
1990    607
1991    607
1992    605
1993    567
vagrant@precise64:~/hadoop-book/ch02-mr-intro/src/main/java$
```

- Result observed: 1990 – 607, 1991 - 607, 1992 – 605, 1993 – 567
- Job History displaying the time taken for running the MaxTemperatureWithCombiner against 1990, 1991, 1992,1993



JobHistory

Application

About Jobs

Tools

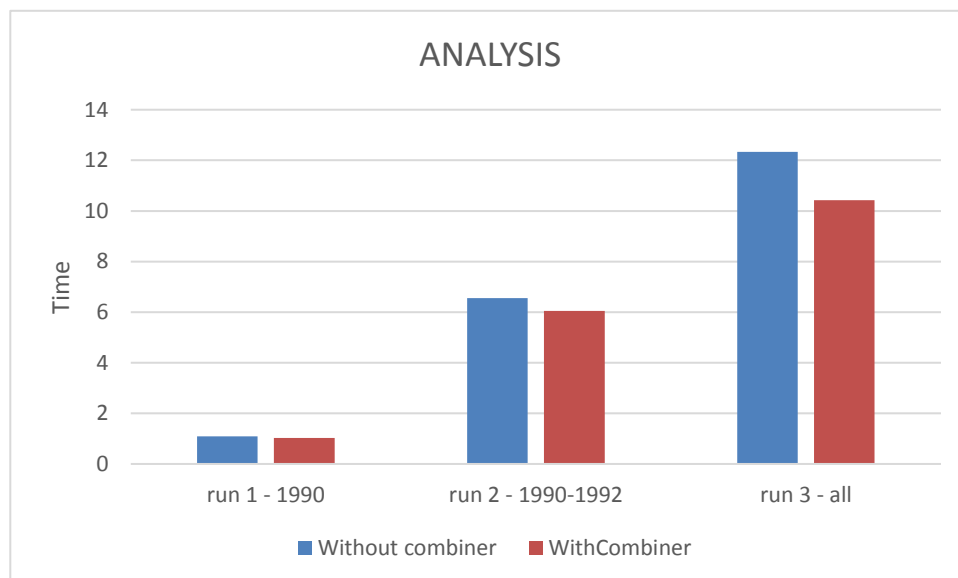
Retired Jobs

Show 20 entries

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State
2017.02.11 03:25:32 UTC	2017.02.11 03:25:39 UTC	2017.02.11 03:36:09 UTC	job_1486778797450_0006	Max temperature	vagrant	default	SUCCEEDED

9. Analysis of time difference

years	Without combiner	WithCombiner
run 1 - 1990	1.09	1.03
run 2 - 1990-1992	6.55	6.05
run 3 - all	12.34	10.43



- SYSTEM MEMORY: 4096 MB & SYSTEM SPEED: 2.53GHz

From running all three combination of the given dataset following is observed

- Year 1990 takes the lowest time because it is the smallest file and using combiner takes less time when compared to using without combiner.
- Year 1990 and 1991 takes 6 minutes and 55 seconds without combiner and 6 minutes 05 seconds using with combiner, by which using combiner makes the processing time very less compared to running without combiner
- Running against all the year is the largest file combination gives a bigger running difference on running with combiner (10.43) and without combiner (12.34).
- It is evident that when we start handling larger files we can save processing time by using Hadoop with combiner to save time.