Name: Aneesh Partha CWID- A20376172

Table of Contents

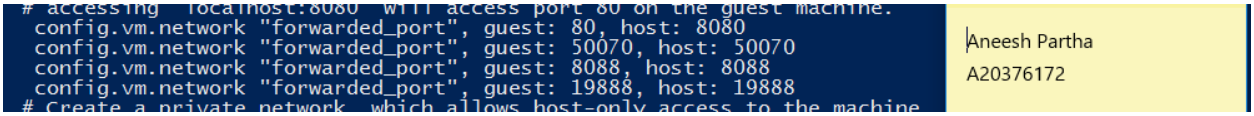
PART ONE – HADOOP INSTALLATION .....
PART TWO – ANALYSIS OF DATA SETS USING HADOOP .....
PART THREE – COMPARISON OF TIME TAKEN.....

Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

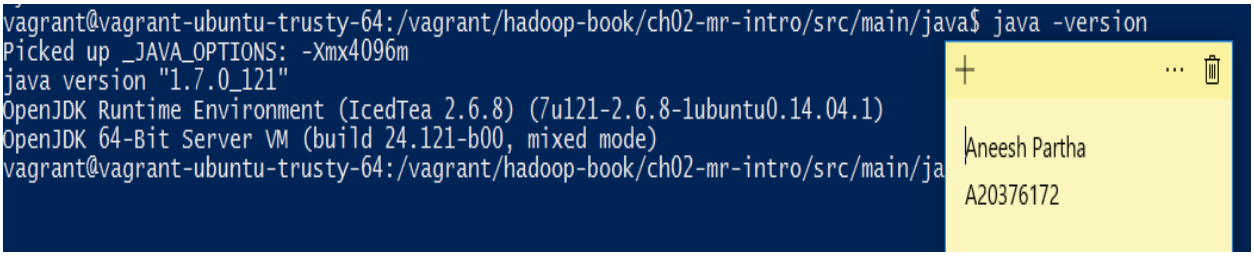
Name: Aneesh Partha **CWID- A20376172**

Hadoop Installation:

Enabling port:

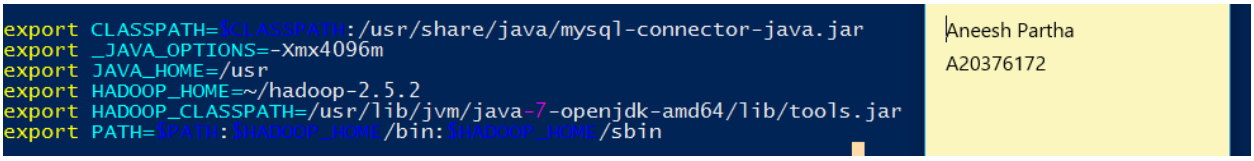


Check for Java availability:

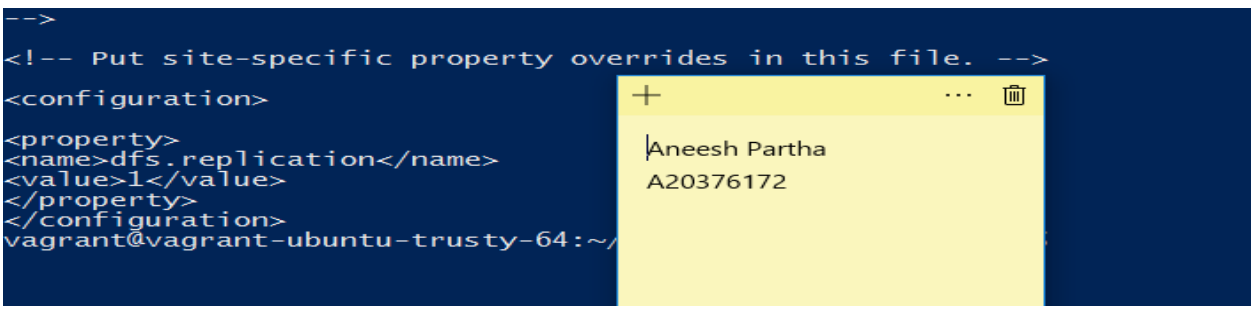


Check after Hadoop is installed in the system:

Below paths are set in .bashrc:



Contents of hdfs-site.xml:



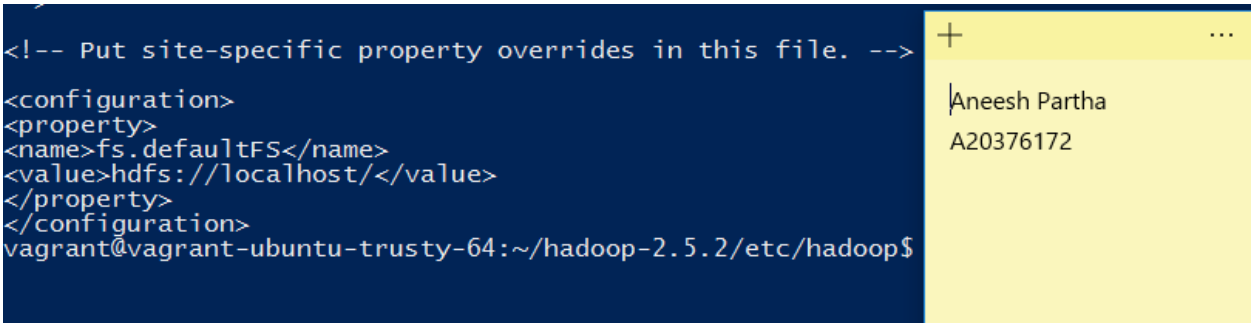

```
<!-- Put site-specific property overrides in this file. -->
<configuration>

<property>
<name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
vagrant@vagrant-ubuntu-trusty-64:~/
```

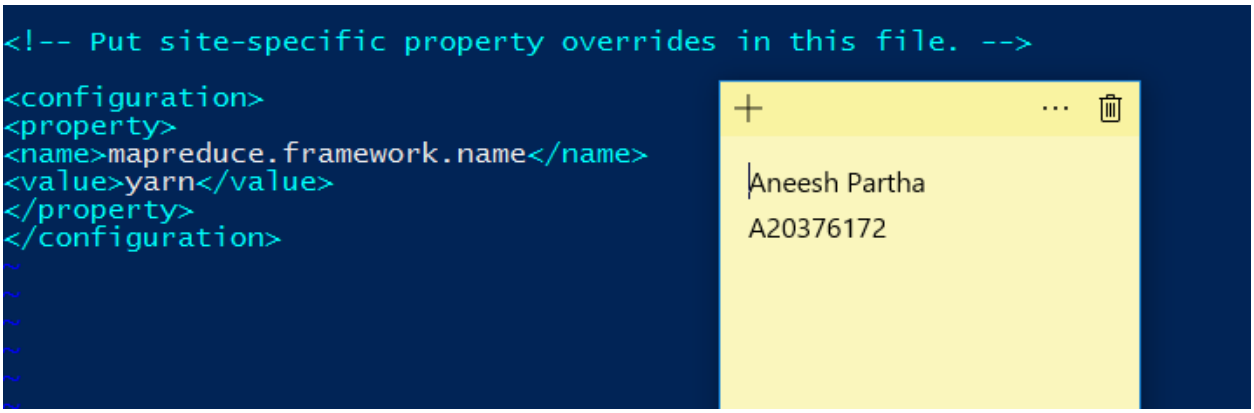
Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

Name: Aneesh Partha **CWID- A20376172**

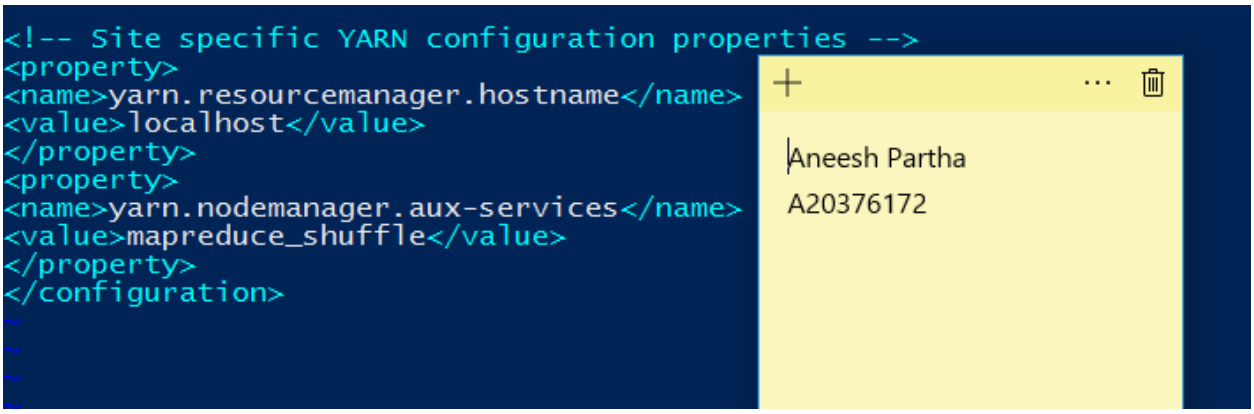
Contents of core-site.xml



Contents of Mapred-site.xml:



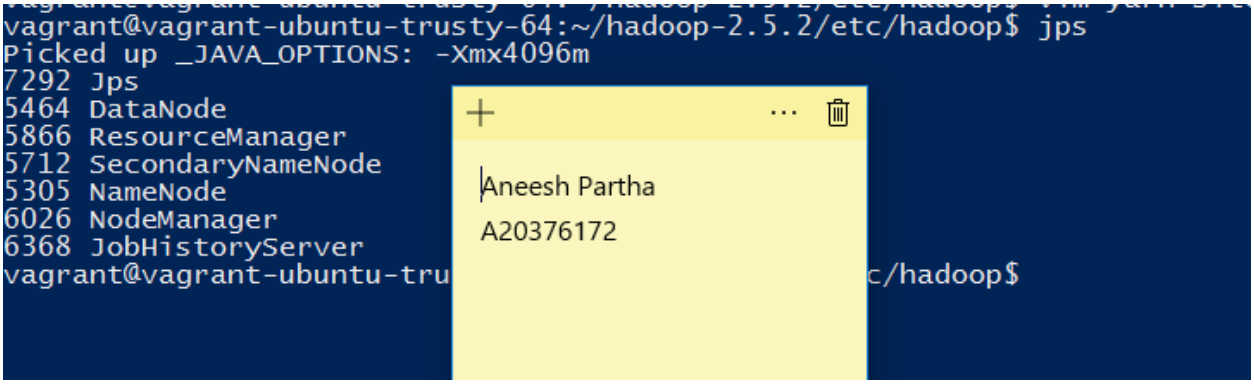
Contents of yarn-site.xml:



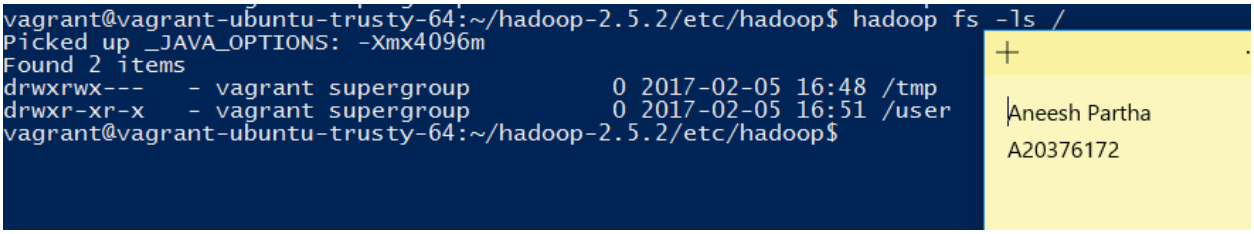
Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

Name: Aneesh Partha **CWID- A20376172**

Processes running in the system:



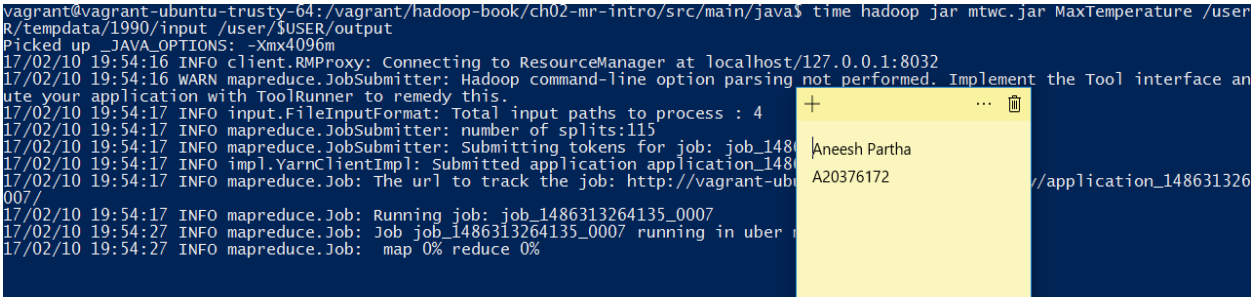
Hadoop file system:



Analysis of data sets using Hadoop:

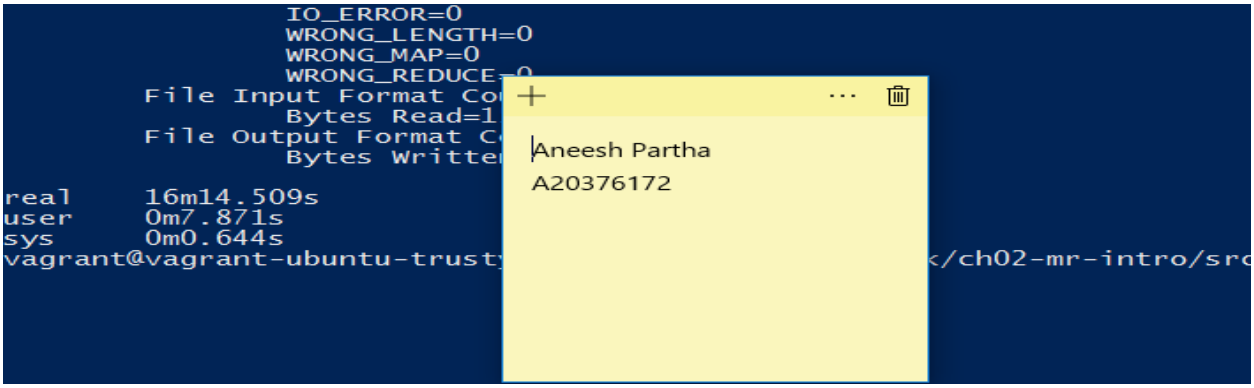
Without combiner:

1990,1991,1992,1993 data sets:



Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

Name: Aneesh Partha **CWID- A20376172**

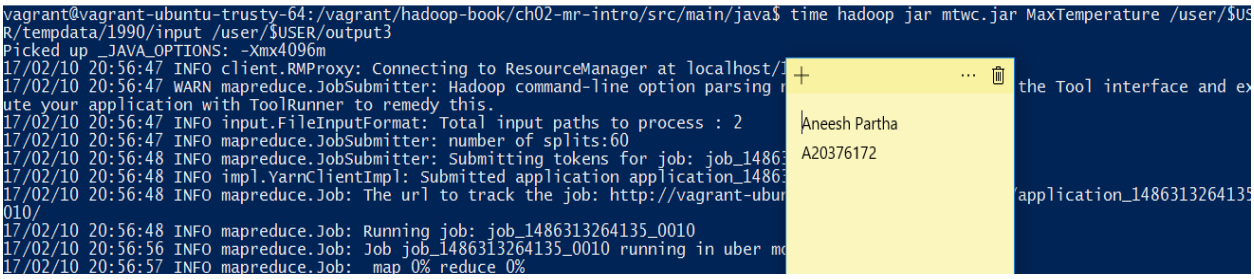


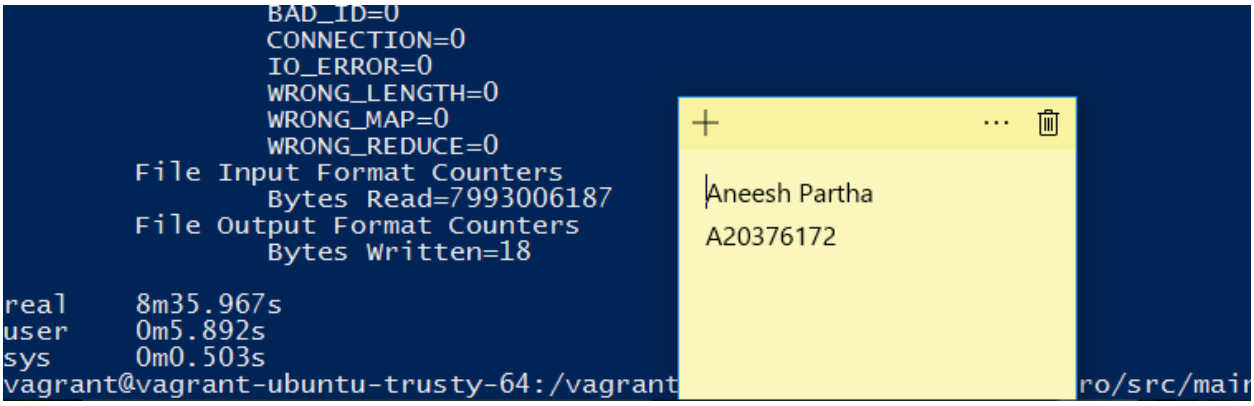
Retired Jobs

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 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

1990 and 1992 data sets:





Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

Name: Aneesh Partha **CWID- A20376172**

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 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

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
1992    605
vagrant@vagrant-ubuntu-trusty-64:/vagrant
```

Aneesh Partha
A20376172

```
USER/output3/part-r-00000
```

1990 data set

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/hadoop-book/ch02-mr-intro/src/main/java$ time hadoop jar mtwc.jar MaxTemperature /use
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/127.0.0.1:8032
17/02/10 21:18:10 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not perfor
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_
17/02/10 21:18:11 INFO impl.YarnClientImpl: Submitted application application_1486313264135_
17/02/10 21:18:11 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty
012/
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 : fals
17/02/10 21:18:18 INFO mapreduce.Job:  map 0% reduce 0%
```

Aneesh Partha
A20376172

```
        File Output F
                Bytes

real    1m31.243s
user    0m4.165s
sys     0m0.187s
vagrant@vagrant-ubunt
```

```
oop-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 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

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant$ hadoop fs -cat /user/$USER/output5/part-r-00000
Picked up _JAVA_OPTIONS: -Xmx4096m
1990    607
vagrant@vagrant-ubuntu-trusty-64:/vagran
```

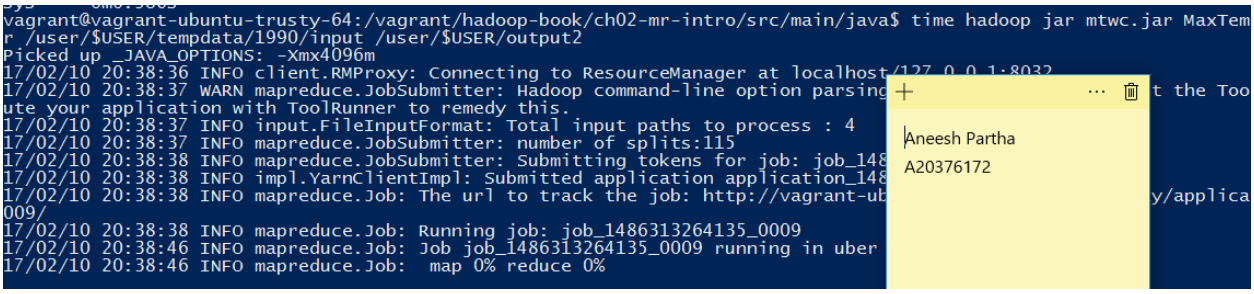
Aneesh Partha
A20376172

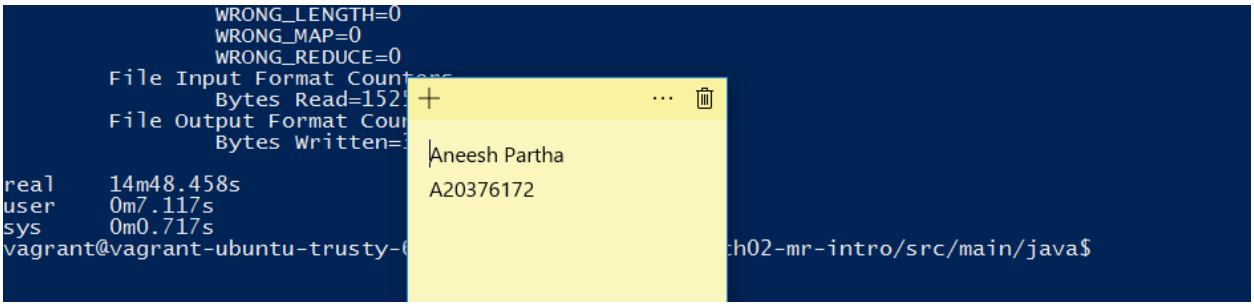
Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

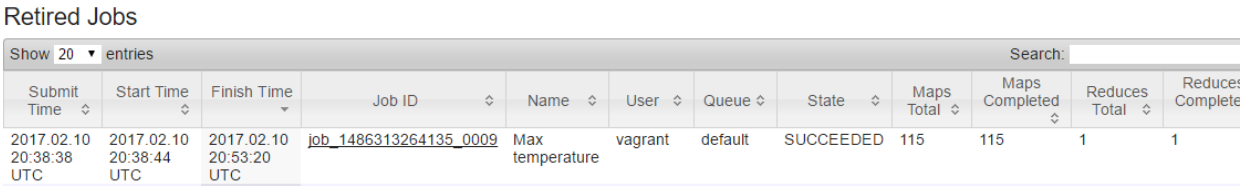
Name: Aneesh Partha **CWID- A20376172**

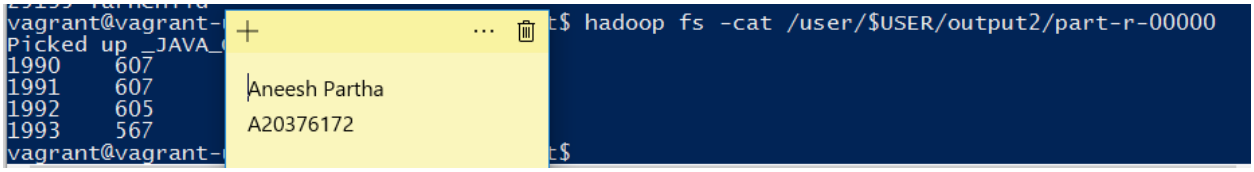
With combiner:

1990,1991,1992 and 1993 data sets

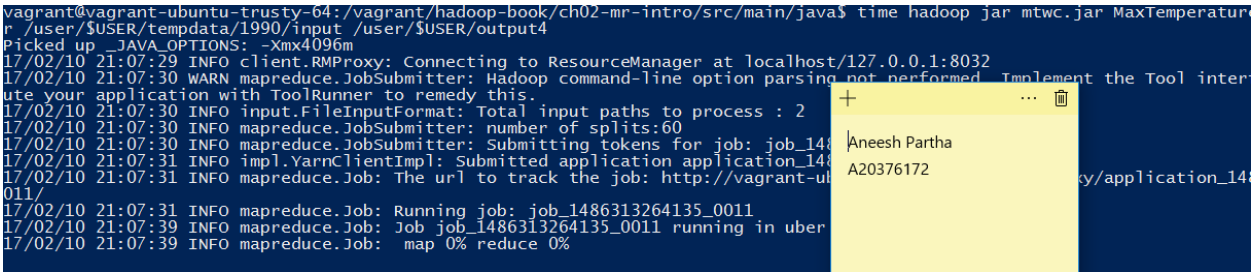






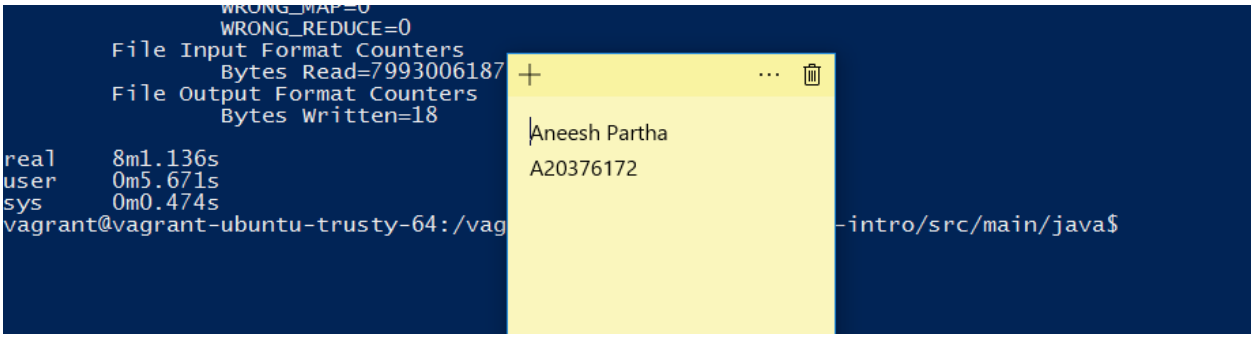


1990,1992 data sets



Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

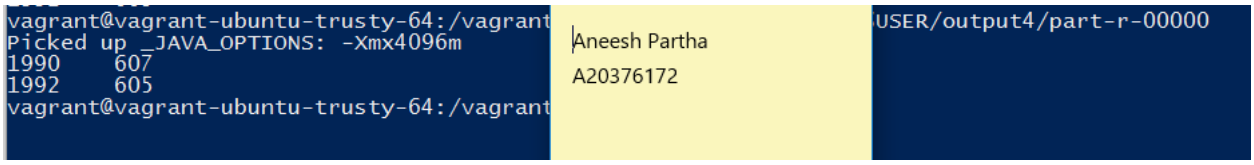
Name: Aneesh Partha **CWID- A20376172**



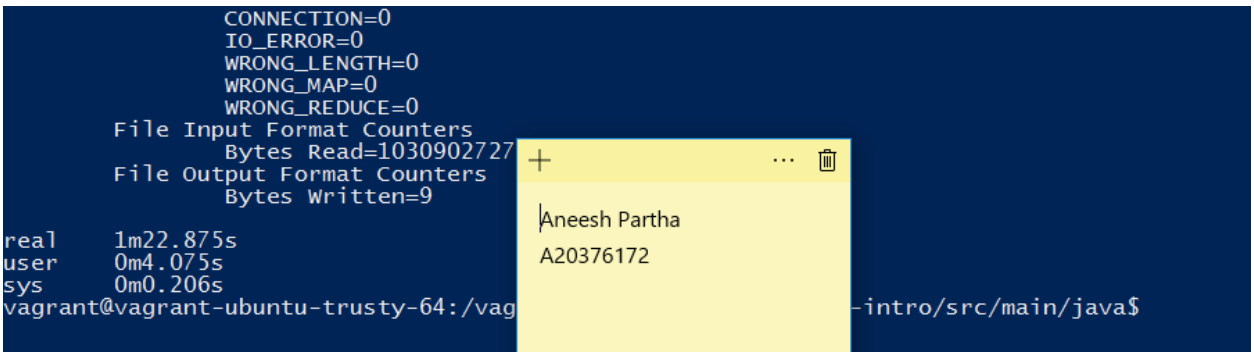
Retired Jobs

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 21:07:31 UTC	2017.02.10 21:07:37 UTC	2017.02.10 21:15:25 UTC	job_1486313264135_0011	Max temperature	vagrant	default	SUCCEEDED	60	60	1	1



1990 data set



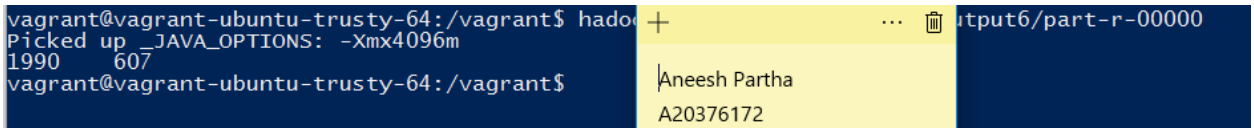
Retired Jobs

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	job_1486313264135_0013	Max temperature	vagrant	default	SUCCEEDED	8	8	1	1

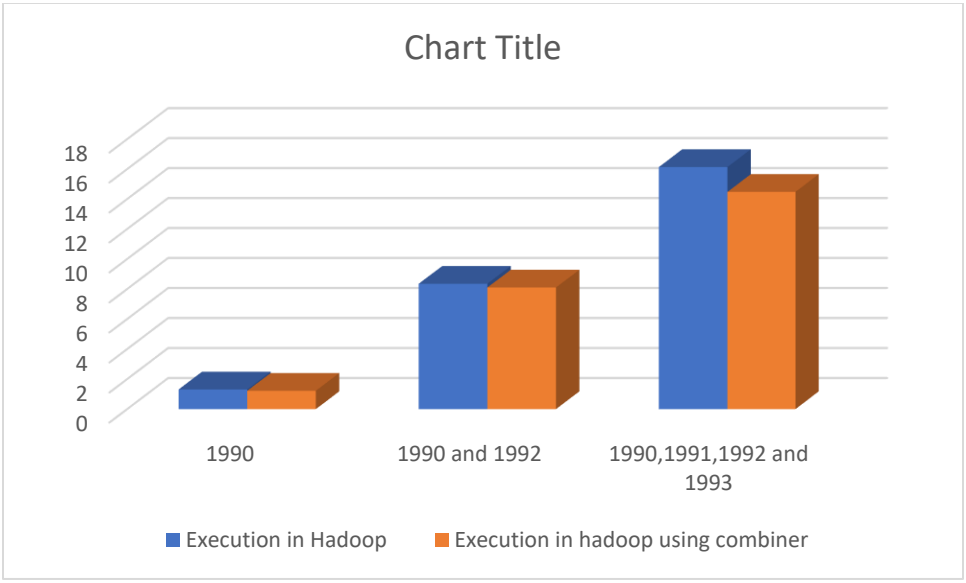
Illinois Institute of Technology	School of Applied Technology
ITMD – 521	Assignment - week03

Name: Aneesh Partha **CWID- 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.