PART 2:

Getting Started with NoSql Systems on HDFS

Fathima Syeda

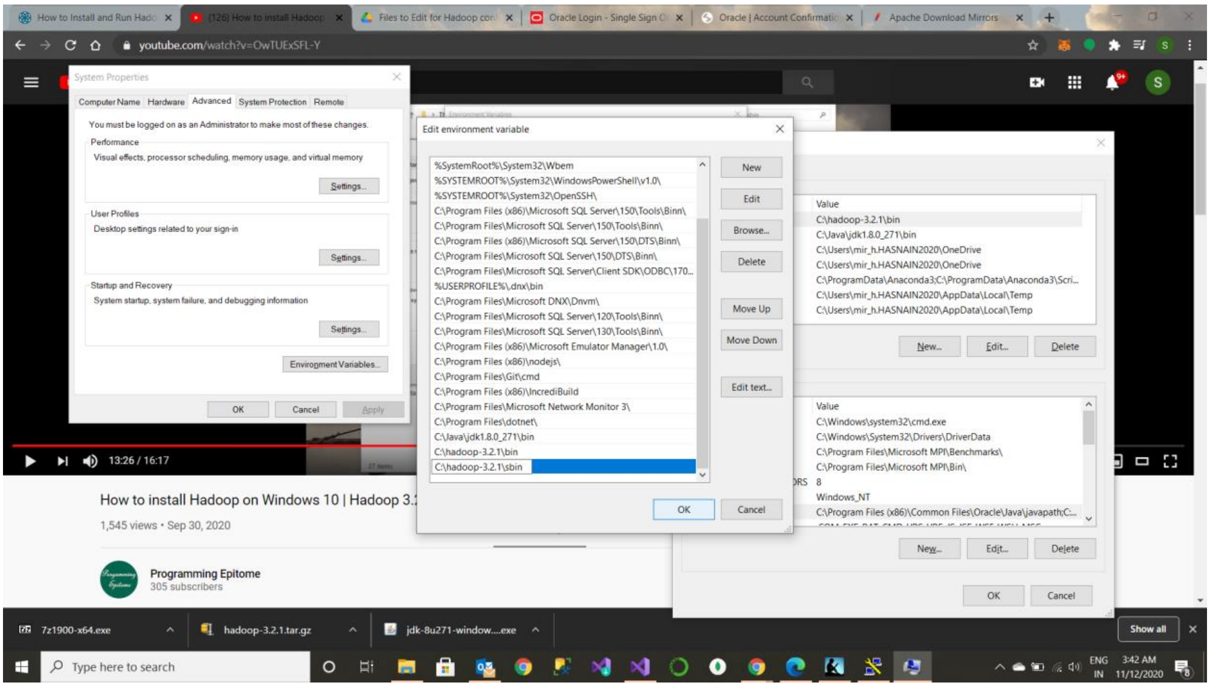
Platform Setup:

Hadoop 3.2.1 is installed in this lab, but in order to do that we must have JAVA SDK-8 installed on our system , so we install that first.

Hadoop would be setup as a single node system in pseudo-distributed mode

We download Hadoop from- https://hadoop.apache.org/releases.html

Once Hadoop is installed , we must set the path and environment variables for it as:



```
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files (x86)\Microsoft SQL Server\150\Tools\Binn\
C:\Program Files\Microsoft SQL Server\150\Tools\Binn\
C:\Program Files (x86)\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\170...
%USERPROFILE%\.dnx\bin
C:\Program Files\Microsoft DNX\Dnvm\
C:\Program Files\Microsoft SQL Server\120\Tools\Binn\
C:\Program Files\Microsoft SQL Server\130\Tools\Binn\
C:\Program Files (x86)\Microsoft Emulator Manager\1.0\
C:\Program Files (x86)\nodejs\
C:\Program Files\Git\cmd
C:\Program Files (x86)\IncrediBuild
C:\Program Files\Microsoft Network Monitor 3\
C:\Program Files\dotnet\
C:\Java\jdk1.8.0_271\bin
C:\hadoop-3.2.1\bin
C:\hadoop-3.2.1\sbin
```

```
Value
C:\hadoop-3.2.1\bin
C:\Java\jdk1.8.0_271\bin
C:\Users\mir_h.HASNAIN2020\OneDrive
C:\Users\mir_h.HASNAIN2020\OneDrive
C:\ProgramData\Anaconda3;C:\ProgramData\Anaconda3\Scri...
C:\Users\mir_h.HASNAIN2020\AppData\Local\Temp
C:\Users\mir_h.HASNAIN2020\AppData\Local\Temp
```

```
Value
C:\Windows\system32\cmd.exe
C:\Windows\System32\Drivers\DriverData
C:\Program Files\Microsoft MPI\Benchmarks\
C:\Program Files\Microsoft MPI\Bin\
8
Windows_NT
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:...
```

After that 5 configuration files in the C:\hadoop-3.2.1\etc\hadoop folder, viz

core-site.xml ,mapred-site.xml, yarn -site.xml , hdfs-site.xml and hadoop-env.cmd are edited.

a) File C:/Hadoop-3.2.1/etc/hadoop/core-site.xml, paste below xml paragraph and save this file.

```
<configuration>
 <property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>
 </property>
</configuration>
```

b) C:/Hadoop-3.2.1/etc/hadoop/mapred-site.xml, paste below xml paragraph and save this file.

```
<configuration>
 <property>
 <name>mapreduce.framework.name</name>
 <value>yarn</value>
 </property>
</configuration>
```

c) Create folder "data" under "C:\Hadoop-3.2.1"

1) Create folder "datanode" under "C:\Hadoop-3.2.1\data"

2) Create folder "namenode" under "C:\Hadoop-3.2.1\data" data

d) Edit file C:\Hadoop-3.2.1/etc/hadoop/hdfs-site.xml, paste below xml paragraph and save this file.

```
<configuration>
 <property>
<name>dfs.replication</name>
 <value>1</value>
 </property>
 <property>
 <name>dfs.namenode.name.dir</name>
 <value>C:\hadoop-3.2.1\data\namenode</value>
 </property>
 <property>
 <name>dfs.datanode.data.dir</name>
 <value>C:\hadoop-3.2.1\data\datanode</value>
 </property>
</configuration>
```

e) Edit file C:/Hadoop-3.2.1/etc/hadoop/yarn-site.xml, paste below xml paragraph and save this file.

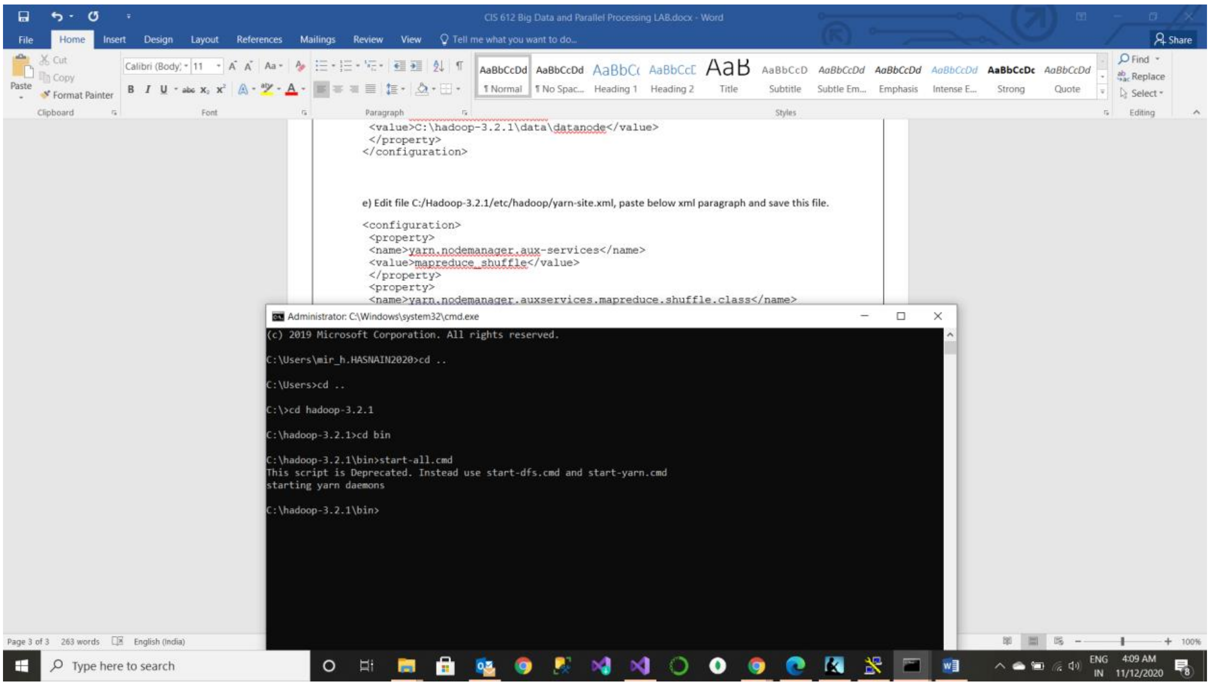
```
<configuration>
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>
 <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
 </property>
</configuration>
```

f) save the java path in the hadoop-env.cmd file as the path of the java sdk’s bin folder.

Once the configurations files have been edited and saved, the Hadoop configuration is completed successfully.

Running the Hadoop single node system

Go to the cmd prompt and go to the Hadoop-3.2.1/bin folder and type start-all.cmd to start all the nodes in the Hadoop system.



```
 <value>C:\hadoop-3.2.1\data\datanode</value>
 </property>
</configuration>
```

e) Edit file C:/Hadoop-3.2.1/etc/hadoop/yarn-site.xml, paste below xml paragraph and save this file.

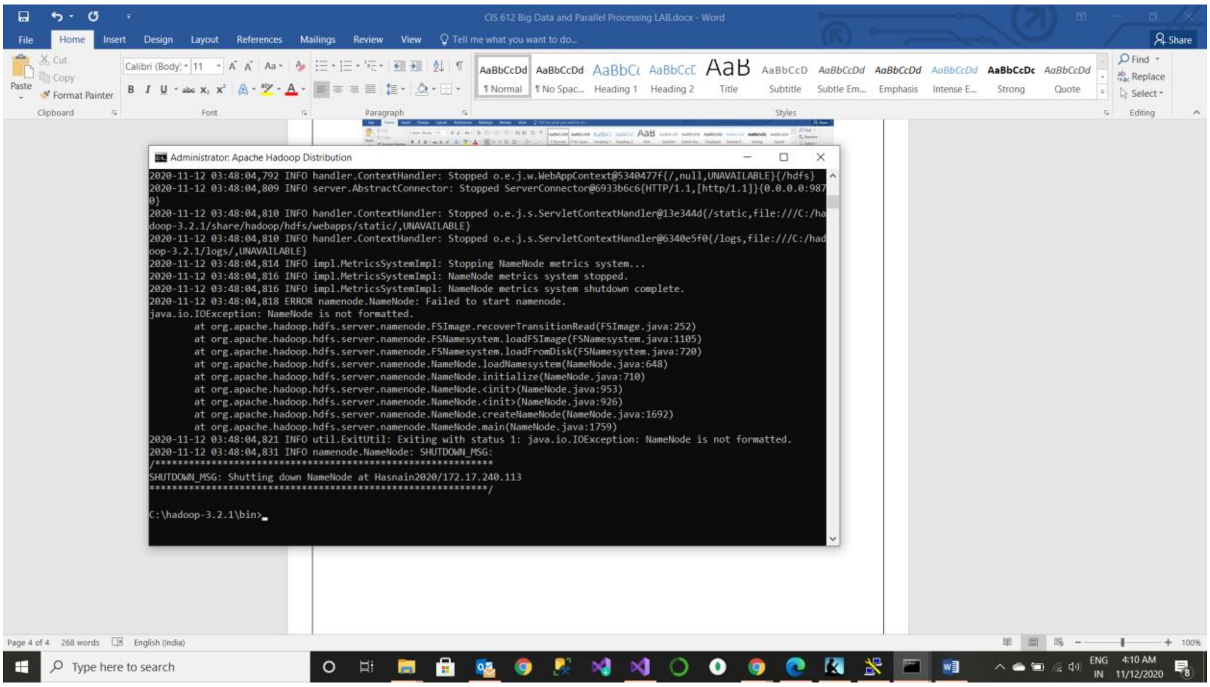
```
<configuration>
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>
 <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
```

Administrator: C:\Windows\system32\cmd.exe

```
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\mir_h.HASNAIN2020>cd ..

C:\Users>cd ..

C:\>cd hadoop-3.2.1

C:\hadoop-3.2.1>cd bin

C:\hadoop-3.2.1\bin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\hadoop-3.2.1\bin>
```

This would start the following

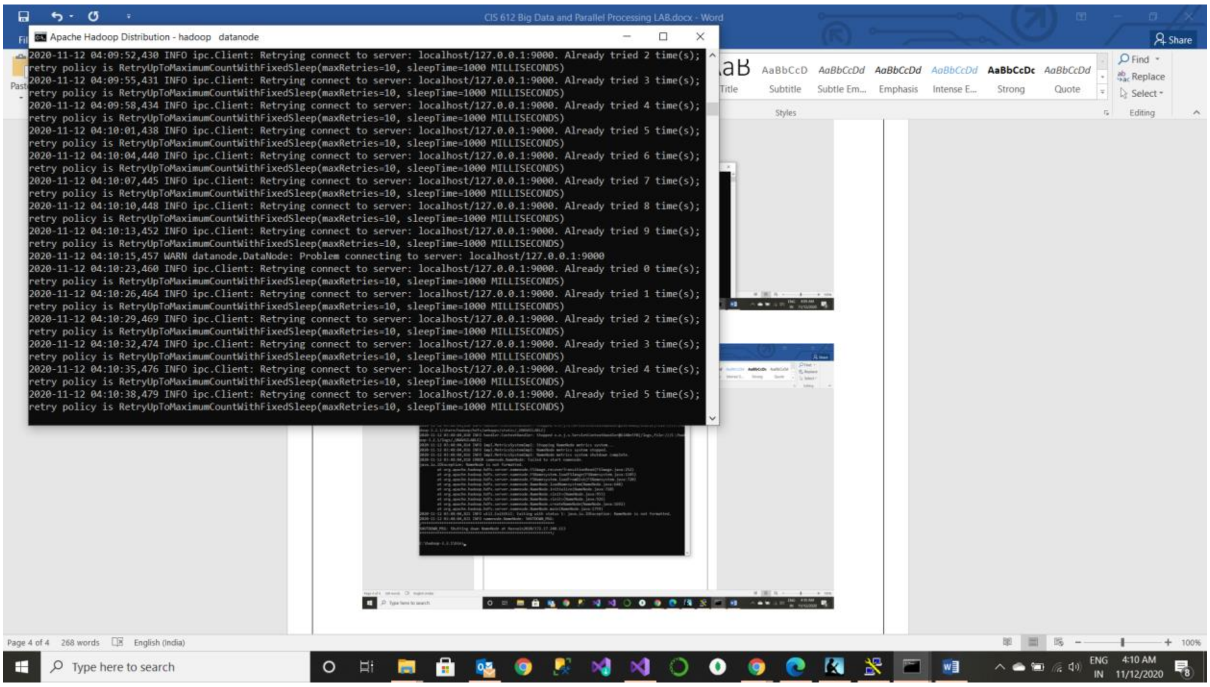
Name Node



Administrator: Apache Hadoop Distribution

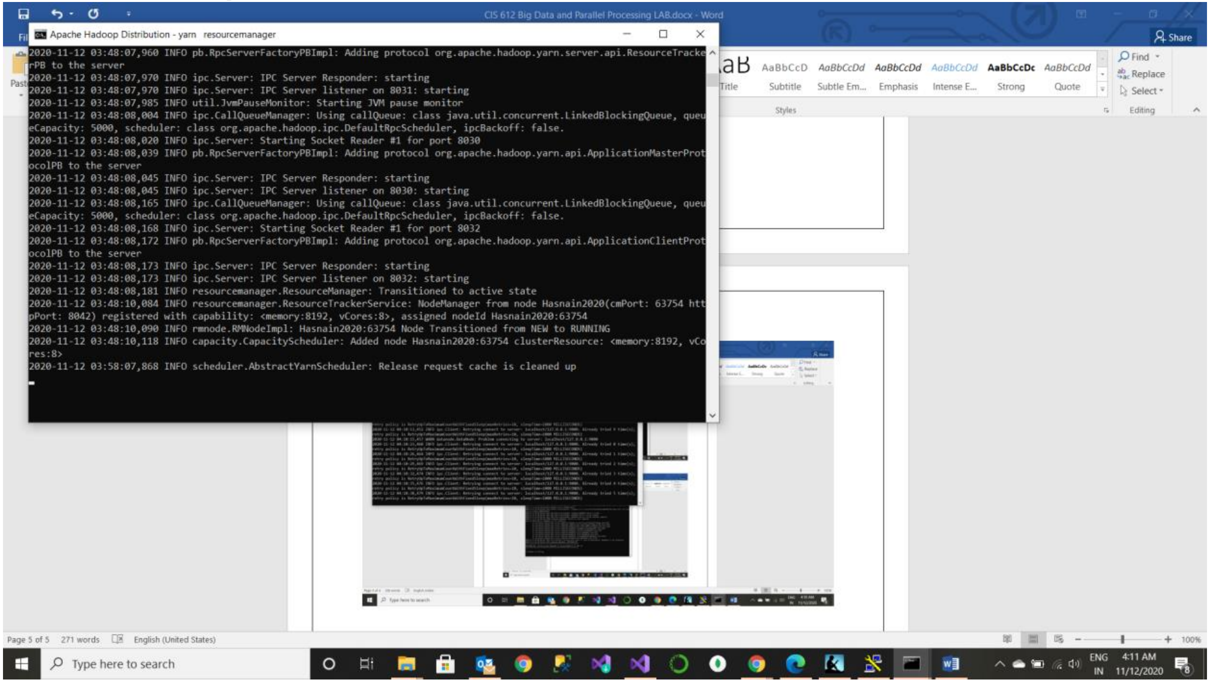
```
2020-11-12 03:48:04,792 INFO handler.ContextHandler: Stopped o.e.j.w.WebAppContext@5340477f{/,null,UNAVAILABLE}{/hdfs}
2020-11-12 03:48:04,809 INFO server.AbstractConnector: Stopped ServerConnector@6933b6c6{HTTP/1.1,[http/1.1]}{0.0.0.0:987
0}
2020-11-12 03:48:04,810 INFO handler.ContextHandler: Stopped o.e.j.s.ServletContextHandler@13e344d{/static,file:///C:/ha
doop-3.2.1/share/hadoop/hdfs/webapps/static/,UNAVAILABLE}
2020-11-12 03:48:04,810 INFO handler.ContextHandler: Stopped o.e.j.s.ServletContextHandler@6340e5f0{/logs,file:///C:/had
oop-3.2.1/logs/,UNAVAILABLE}
2020-11-12 03:48:04,814 INFO impl.MetricsSystemImpl: Stopping NameNode metrics system...
2020-11-12 03:48:04,816 INFO impl.MetricsSystemImpl: NameNode metrics system stopped.
2020-11-12 03:48:04,816 INFO impl.MetricsSystemImpl: NameNode metrics system shutdown complete.
2020-11-12 03:48:04,818 ERROR namenode.NameNode: Failed to start namenode.
java.io.IOException: NameNode is not formatted.
        at org.apache.hadoop.hdfs.server.namenode.FSImage.recoverTransitionRead(FSImage.java:252)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFSImage(FSNamesystem.java:1105)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.loadFromDisk(FSNamesystem.java:720)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.loadNamesystem(NameNode.java:648)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.initialize(NameNode.java:710)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.java:953)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.<init>(NameNode.java:926)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(NameNode.java:1692)
        at org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java:1759)
2020-11-12 03:48:04,821 INFO util.ExitUtil: Exiting with status 1: java.io.IOException: NameNode is not formatted.
2020-11-12 03:48:04,831 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at Hasnain2020/172.17.240.113
************************************************************/

C:\hadoop-3.2.1\bin>
```

Data Node


CIS 612 Big Data and Parallel Processing LAB.docx - Word
Apache Hadoop Distribution - hadoop datanode
Page 4 of 4 268 words English (India)
Type here to search
4:10 AM
11/12/2020

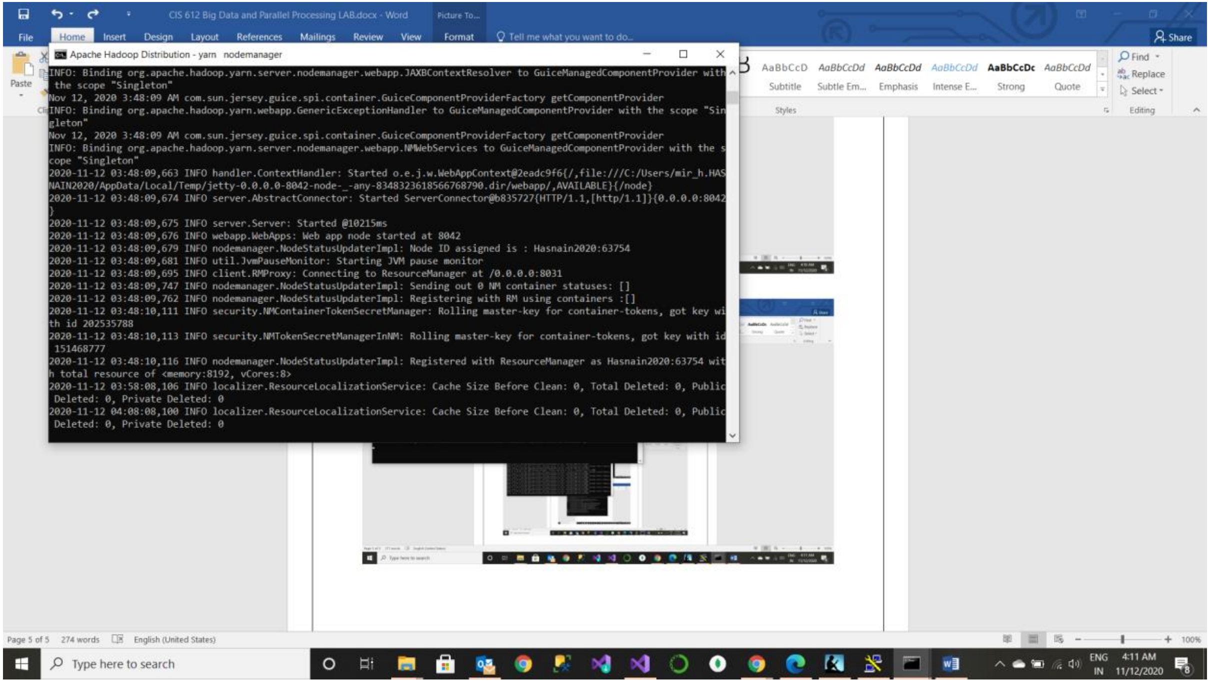
```
2020-11-12 04:09:52,430 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 2 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:09:55,431 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 3 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:09:58,434 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 4 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:01,438 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 5 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:04,440 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 6 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:07,445 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 7 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:10,448 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 8 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:13,452 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 9 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:15,457 WARN datanode.DataNode: Problem connecting to server: localhost/127.0.0.1:9000
2020-11-12 04:10:23,460 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 0 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:26,464 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 1 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:29,469 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 2 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:32,474 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 3 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:35,476 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 4 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2020-11-12 04:10:38,479 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 5 time(s);
retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
```

Resource Manager


CIS 612 Big Data and Parallel Processing LAB.docx - Word
Apache Hadoop Distribution - yarn resourcemanager
Page 5 of 5 271 words English (United States)
Type here to search
4:11 AM
11/12/2020

```
2020-11-12 03:48:07,960 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.server.api.ResourceTracke
rPB to the server
2020-11-12 03:48:07,970 INFO ipc.Server: IPC Server Responder: starting
2020-11-12 03:48:07,970 INFO ipc.Server: IPC Server listener on 8031: starting
2020-11-12 03:48:07,985 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2020-11-12 03:48:08,004 INFO ipc.CallQueueManager: Using callQueue: class java.util.concurrent.LinkedBlockingQueue, queu
eCapacity: 5000, scheduler: class org.apache.hadoop.ipc.DefaultRpcScheduler, ipcBackoff: false.
2020-11-12 03:48:08,020 INFO ipc.Server: Starting Socket Reader #1 for port 8030
2020-11-12 03:48:08,039 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.api.ApplicationMasterProt
ocolPB to the server
2020-11-12 03:48:08,045 INFO ipc.Server: IPC Server Responder: starting
2020-11-12 03:48:08,045 INFO ipc.Server: IPC Server listener on 8030: starting
2020-11-12 03:48:08,165 INFO ipc.CallQueueManager: Using callQueue: class java.util.concurrent.LinkedBlockingQueue, queu
eCapacity: 5000, scheduler: class org.apache.hadoop.ipc.DefaultRpcScheduler, ipcBackoff: false.
2020-11-12 03:48:08,168 INFO ipc.Server: Starting Socket Reader #1 for port 8032
2020-11-12 03:48:08,172 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.api.ApplicationClientProt
ocolPB to the server
2020-11-12 03:48:08,173 INFO ipc.Server: IPC Server Responder: starting
2020-11-12 03:48:08,173 INFO ipc.Server: IPC Server listener on 8032: starting
2020-11-12 03:48:08,181 INFO resourcemanager.ResourceManager: Transitioned to active state
2020-11-12 03:48:10,084 INFO resourcemanager.ResourceTrackerService: NodeManager from node Hasnain2020(cmPort: 63754 htt
pPort: 8042) registered with capability: <memory:8192, vCores:8>, assigned nodeId Hasnain2020:63754
2020-11-12 03:48:10,090 INFO rmnode.RMNodeImpl: Hasnain2020:63754 Node Transitioned from NEW to RUNNING
2020-11-12 03:48:10,118 INFO capacity.CapacityScheduler: Added node Hasnain2020:63754 clusterResource: <memory:8192, vCo
res:8>
2020-11-12 03:58:07,868 INFO scheduler.AbstractYarnScheduler: Release request cache is cleaned up
```

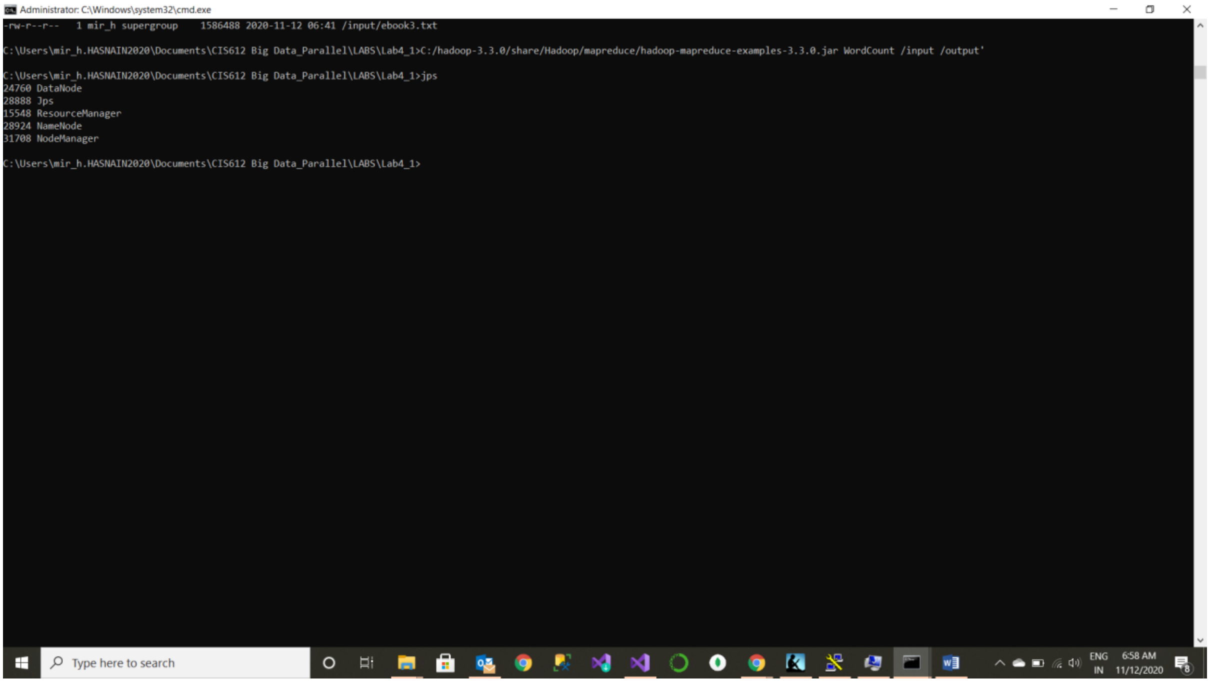
Node Manager



Apache Hadoop Distribution - yarn nodemanager

```
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver to GuiceManagedComponentProvider with
 the scope "Singleton"
Nov 12, 2020 3:48:09 AM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope "Sin
gleton"
Nov 12, 2020 3:48:09 AM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the s
cope "Singleton"
2020-11-12 03:48:09,663 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@2eadc9f6{/,file:///C:/Users/mir_h.HAS
NAIN2020/AppData/Local/Temp/jetty-0.0.0.0-8042-node-_-any-8348323618566768790.dir/webapp/,AVAILABLE}{/node}
2020-11-12 03:48:09,674 INFO server.AbstractConnector: Started ServerConnector@b835727{HTTP/1.1,[http/1.1]}{0.0.0.0:8042
}
2020-11-12 03:48:09,675 INFO server.Server: Started @10215ms
2020-11-12 03:48:09,676 INFO webapp.WebApps: Web app node started at 8042
2020-11-12 03:48:09,679 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : Hasnain2020:63754
2020-11-12 03:48:09,681 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2020-11-12 03:48:09,695 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031
2020-11-12 03:48:09,747 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
2020-11-12 03:48:09,762 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2020-11-12 03:48:10,111 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key wi
th id 202535788
2020-11-12 03:48:10,113 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id
 151468777
2020-11-12 03:48:10,116 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as Hasnain2020:63754 wit
h total resource of <memory:8192, vCores:8>
2020-11-12 03:58:08,106 INFO localizer.ResourceLocalizationService: Cache Size Before Clean: 0, Total Deleted: 0, Public
 Deleted: 0, Private Deleted: 0
2020-11-12 04:08:08,100 INFO localizer.ResourceLocalizationService: Cache Size Before Clean: 0, Total Deleted: 0, Public
 Deleted: 0, Private Deleted: 0
```

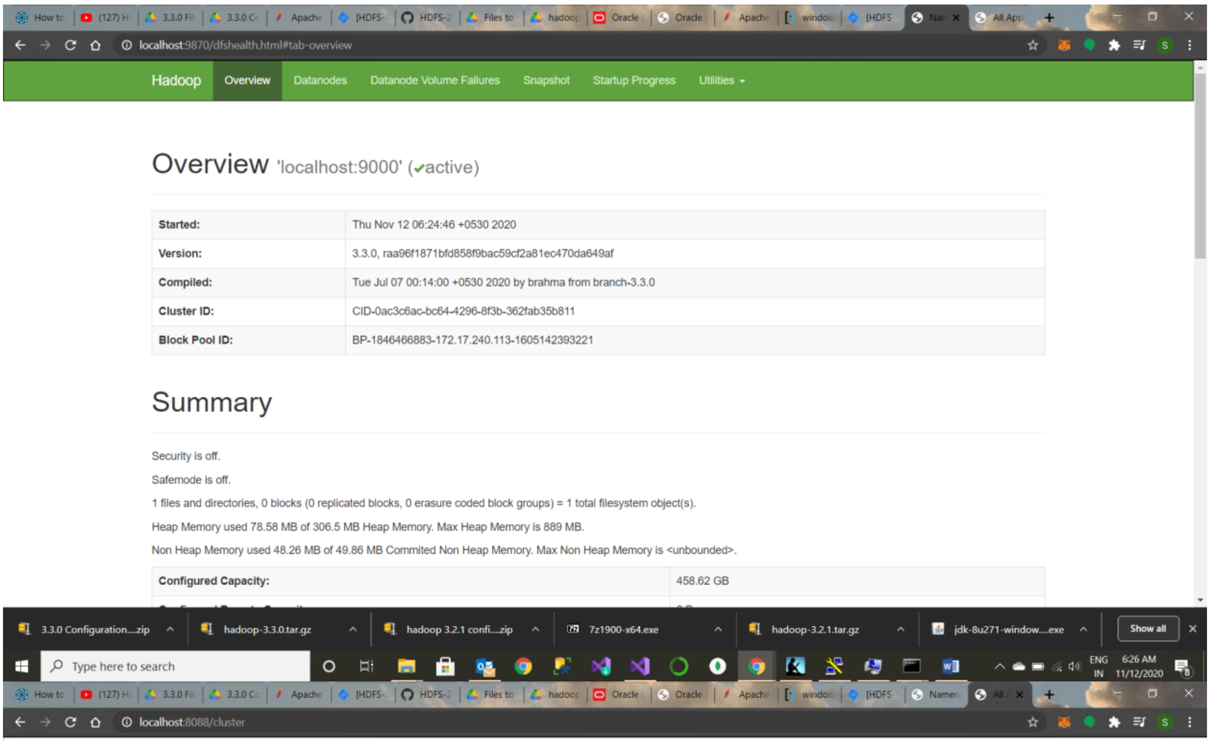
Checking the port of the daemons using **jps** command



Administrator: C:\Windows\system32\cmd.exe

```
-rw-r--r--   1 mir_h supergroup    1586488 2020-11-12 06:41 /input/ebook3.txt

C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parallel\LABS\Lab4_1>C:/hadoop-3.3.0/share/Hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar WordCount /input /output'

C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parallel\LABS\Lab4_1>jps
24760 DataNode
28888 Jps
15548 ResourceManager
28924 NameNode
31708 NodeManager

C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parallel\LABS\Lab4_1>
```

Open the browser and type localhost:9870 anf localhost:8086/clusters to see the Hadoop connection and the nodes running on it .



Overview 'localhost:9000' (✓active)

Started:	Thu Nov 12 06:24:46 +0530 2020
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 00:14:00 +0530 2020 by brahma from branch-3.3.0
Cluster ID:	CID-0ac3c6ac-bc64-4296-8f3b-362fab35b811
Block Pool ID:	BP-1846466883-172.17.240.113-1605142393221

Summary

Security is off.

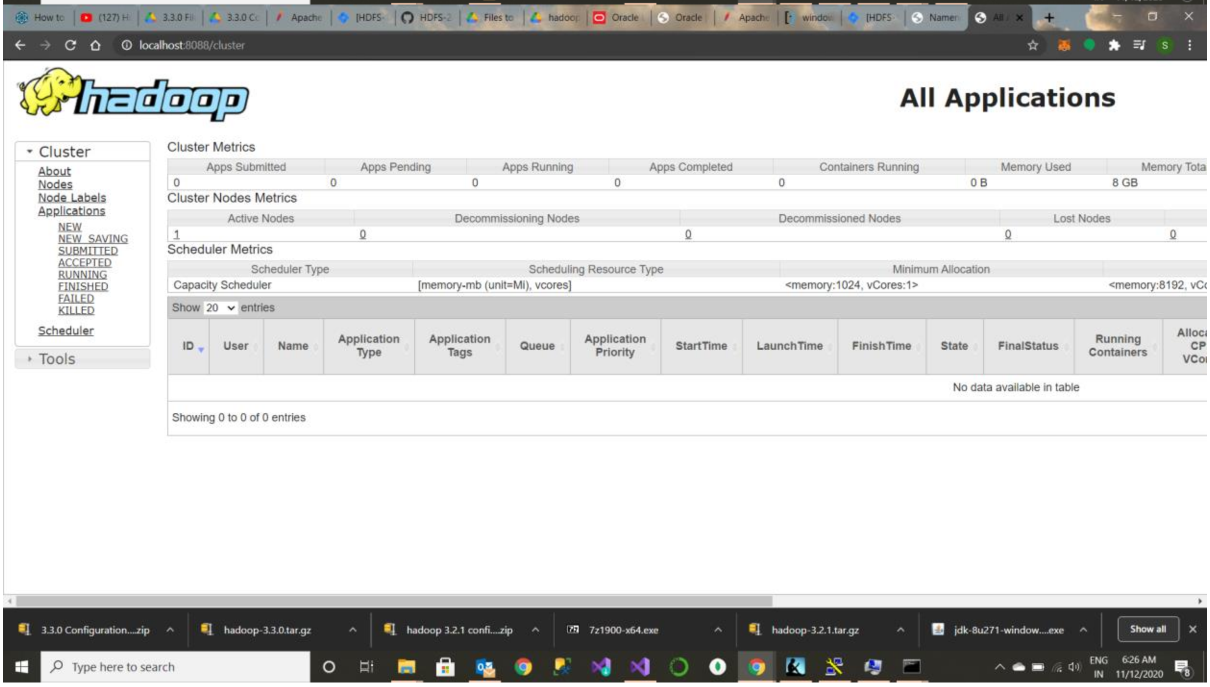
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 78.58 MB of 306.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 48.26 MB of 49.86 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	458.62 GB



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Tota
0	0	0	0	0	0 B	8 GB

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	
1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCo

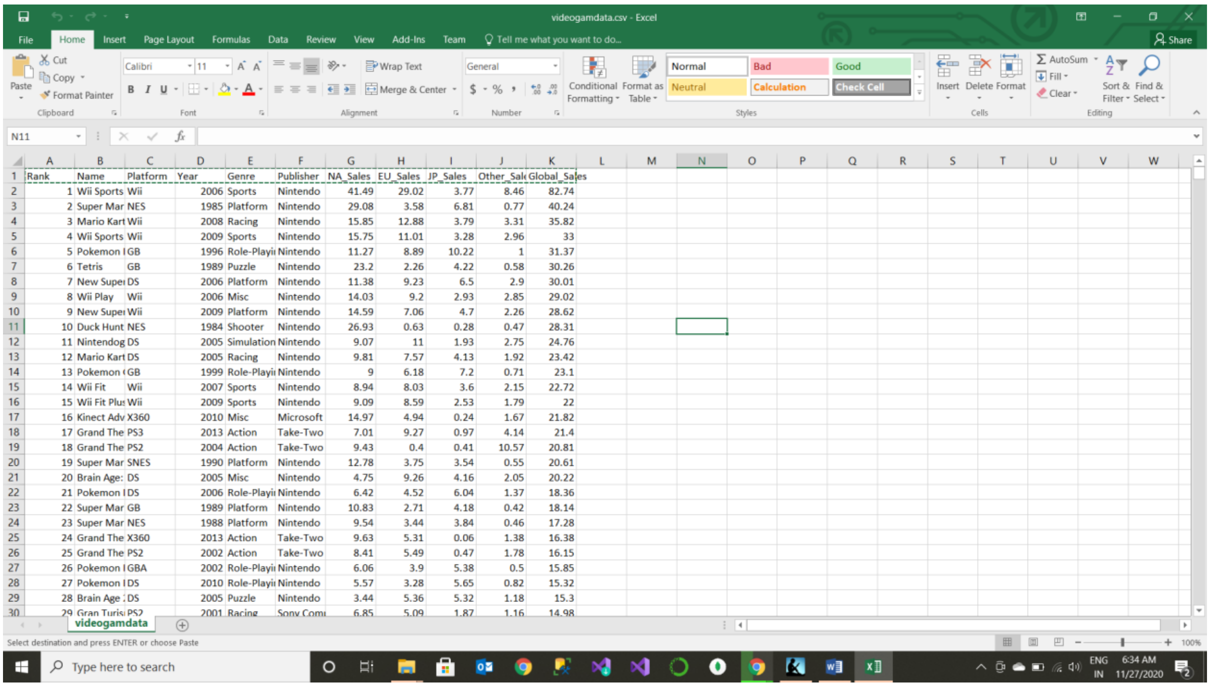
Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Alloca CP VCo
No data available in table													

Showing 0 to 0 of 0 entries

Part-1 Hive

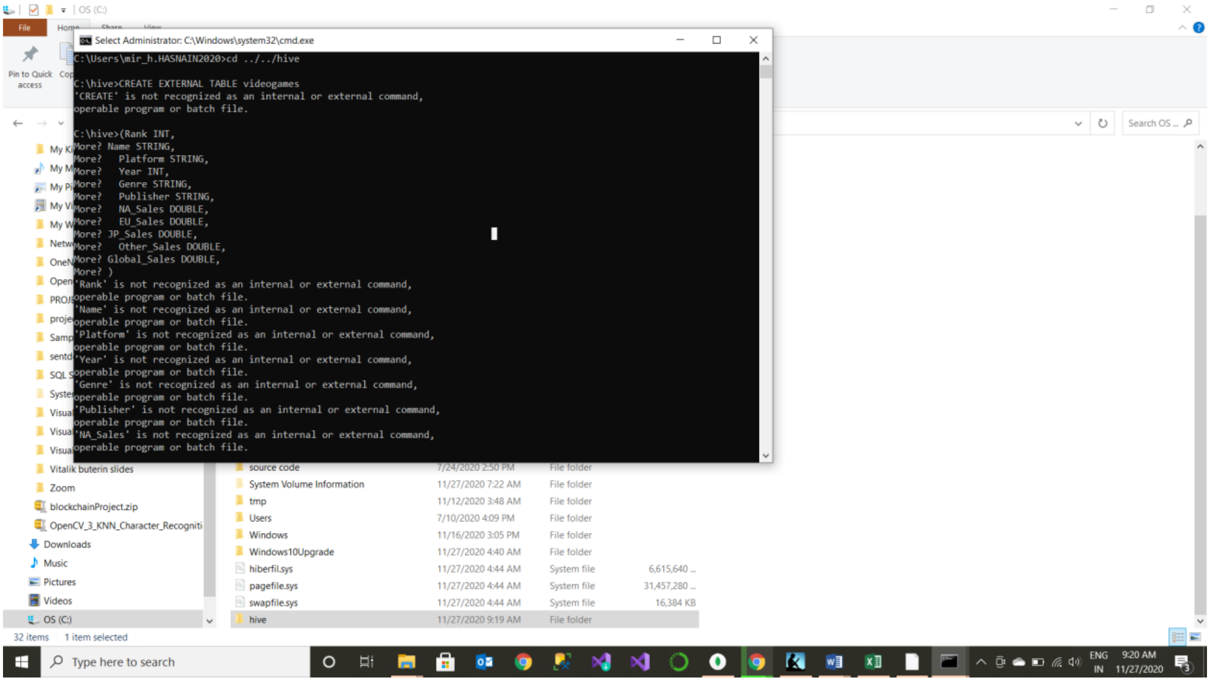
Our input data is the videogame sales data



Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sal	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mar	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon I	GB	1996	Role-Playi	Nintendo	11.27	8.89	10.22	1	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	New Super	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	Nintendog	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
12	Mario Kart	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	Pokemon (	GB	1999	Role-Playi	Nintendo	9	6.18	7.2	0.71	23.1
14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
16	Kinect Adv	X360	2010	Misc	Microsoft	14.97	4.94	0.24	1.67	21.82
17	Grand The	PS3	2013	Action	Take-Two	7.01	9.27	0.97	4.14	21.4
18	Grand The	PS2	2004	Action	Take-Two	9.43	0.4	0.41	10.57	20.81
19	Super Mar	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
20	Brain Age:	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22
21	Pokemon I	DS	2006	Role-Playi	Nintendo	6.42	4.52	6.04	1.37	18.36
22	Super Mar	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14
23	Super Mar	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28
24	Grand The	X360	2013	Action	Take-Two	9.63	5.31	0.06	1.38	16.38
25	Grand The	PS2	2002	Action	Take-Two	8.41	5.49	0.47	1.78	16.15
26	Pokemon I	GBA	2002	Role-Playi	Nintendo	6.06	3.9	5.38	0.5	15.85
27	Pokemon I	DS	2010	Role-Playi	Nintendo	5.57	3.28	5.65	0.82	15.32
28	Brain Age :	DS	2005	Puzzle	Nintendo	3.44	5.36	5.32	1.18	15.3
29	Gran Turis	PS2	2001	Racing	Sony Comp	6.85	5.09	1.87	1.16	14.98

1. Creation of Table

```
CREATE EXTERNAL TABLE videogames
(           Rank INT,
            Name STRING,
            Platform STRING,
            Year INT,
            Genre STRING,
            Publisher STRING,
            NA_Sales DOUBLE,
            EU_Sales DOUBLE,
            JP_Sales DOUBLE,
            Other_Sales DOUBLE,
            Global_Sales DOUBLE,
)
```



```
C:\Users\mir_h.HASNAIN2020>cd ../../hive

C:\hive>CREATE EXTERNAL TABLE videogames
'CREATE' is not recognized as an internal or external command,
operable program or batch file.

C:\hive>(Rank INT,
More? Name STRING,
More?   Platform STRING,
More?   Year INT,
More?   Genre STRING,
More?   Publisher STRING,
More?   NA_Sales DOUBLE,
More?   EU_Sales DOUBLE,
More? JP_Sales DOUBLE,
More?   Other_Sales DOUBLE,
More? Global_Sales DOUBLE,
More? )
'Rank' is not recognized as an internal or external command,
operable program or batch file.
'Name' is not recognized as an internal or external command,
operable program or batch file.
'Platform' is not recognized as an internal or external command,
operable program or batch file.
'Year' is not recognized as an internal or external command,
operable program or batch file.
'Genre' is not recognized as an internal or external command,
operable program or batch file.
'Publisher' is not recognized as an internal or external command,
operable program or batch file.
'NA_Sales' is not recognized as an internal or external command,
operable program or batch file.
```

2. Loading data into created table all states

Load data local inpath '/home/hduser/Desktop/videogames.csv' into table `videogames;`

3. Creation of partition table

create table Platform _part(Rank INT,

Name STRING,

Year INT,

Genre STRING,

Publisher STRING,

NA_Sales DOUBLE,

EU_Sales DOUBLE,

JP_Sales DOUBLE,

Other_Sales DOUBLE,

Global_Sales DOUBLE,

) PARTITIONED BY(Platform STRING);

4. Set the following property on the partition
   set hive.exec.dynamic.partition.mode=nonstrict

5. Loading data into partition table

INSERT OVERWRITE TABLE Platform _part PARTITION(Platform)

SELECT Name, Year, Genre, Publisher ,NA_Sales , EU_Sales , JP_Sales , Other_Sales , Global_Sales from videogames;

6. Creating Bucket

   In Hive, we have to enable buckets by using the **set.hive.enforce.bucketing=true;**

Create table sample_bucket { Rank INT,

Name STRING,

Year INT,

Platform STRING,

Publisher STRING,

NA_Sales DOUBLE,

EU_Sales DOUBLE,

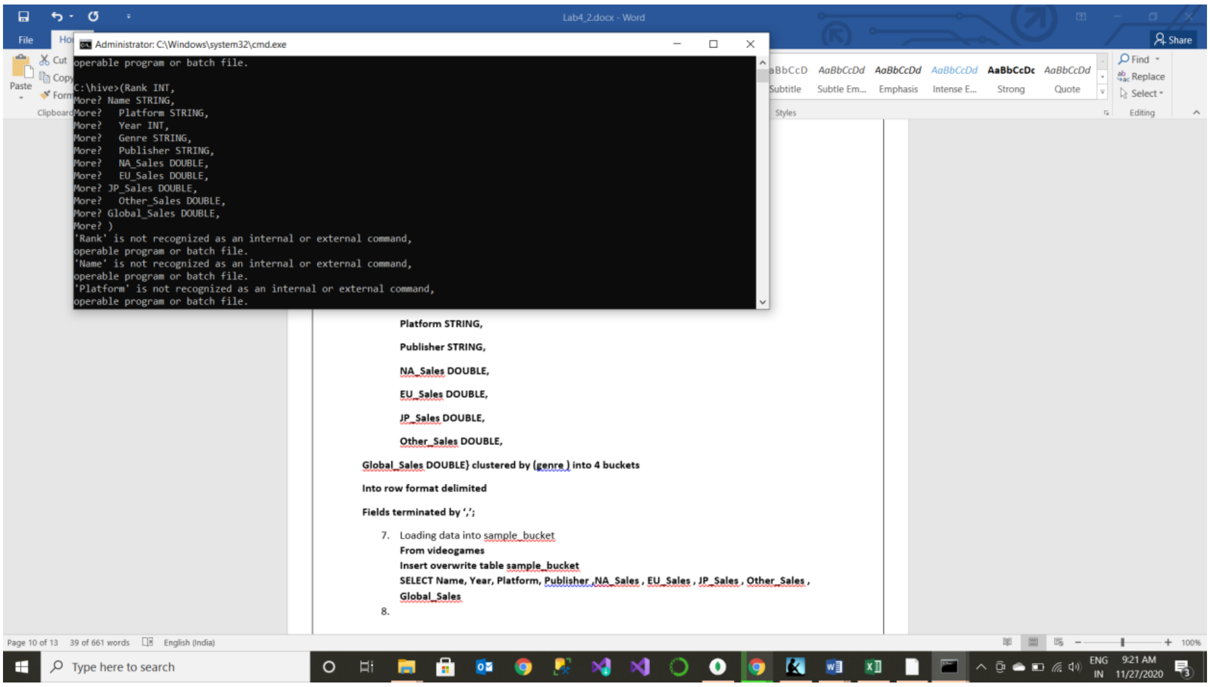
JP_Sales DOUBLE,

Other_Sales DOUBLE,

Global_Sales DOUBLE} clustered by (genre) into 4 buckets

Into row format delimited

Fields terminated by ‘,’;



```
operable program or batch file.

C:\hive>(Rank INT,
More? Name STRING,
More?   Platform STRING,
More?   Year INT,
More?   Genre STRING,
More?   Publisher STRING,
More?   NA_Sales DOUBLE,
More?   EU_Sales DOUBLE,
More? JP_Sales DOUBLE,
More?   Other_Sales DOUBLE,
More? Global_Sales DOUBLE,
More? )
'Rank' is not recognized as an internal or external command,
operable program or batch file.
'Name' is not recognized as an internal or external command,
operable program or batch file.
'Platform' is not recognized as an internal or external command,
operable program or batch file.
```

Platform STRING,

Publisher STRING,

NA_Sales DOUBLE,

EU_Sales DOUBLE,

JP_Sales DOUBLE,

Other_Sales DOUBLE,

Global_Sales DOUBLE} clustered by (genre) into 4 buckets

Into row format delimited

Fields terminated by ‘,’;

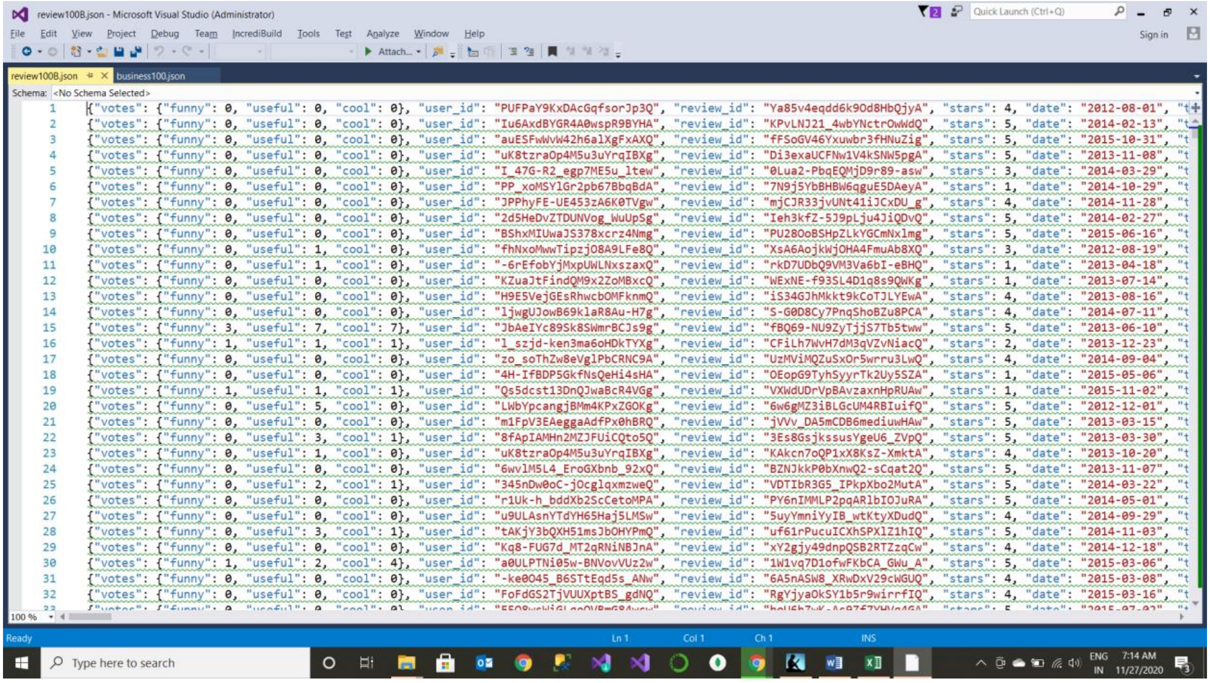
7. Loading data into sample_bucket
   From videogames
   Insert overwrite table sample_bucket
   SELECT Name, Year, Platform, Publisher ,NA_Sales , EU_Sales , JP_Sales , Other_Sales , Global_Sales
8.

7. Loading data into sample_bucket
   From videogames
   Insert overwrite table sample_bucket
   SELECT Name, Year, Platform, Publisher ,NA_Sales , EU_Sales , JP_Sales , Other_Sales , Global_Sales
8.

Part-2 MongoDb on HDFS

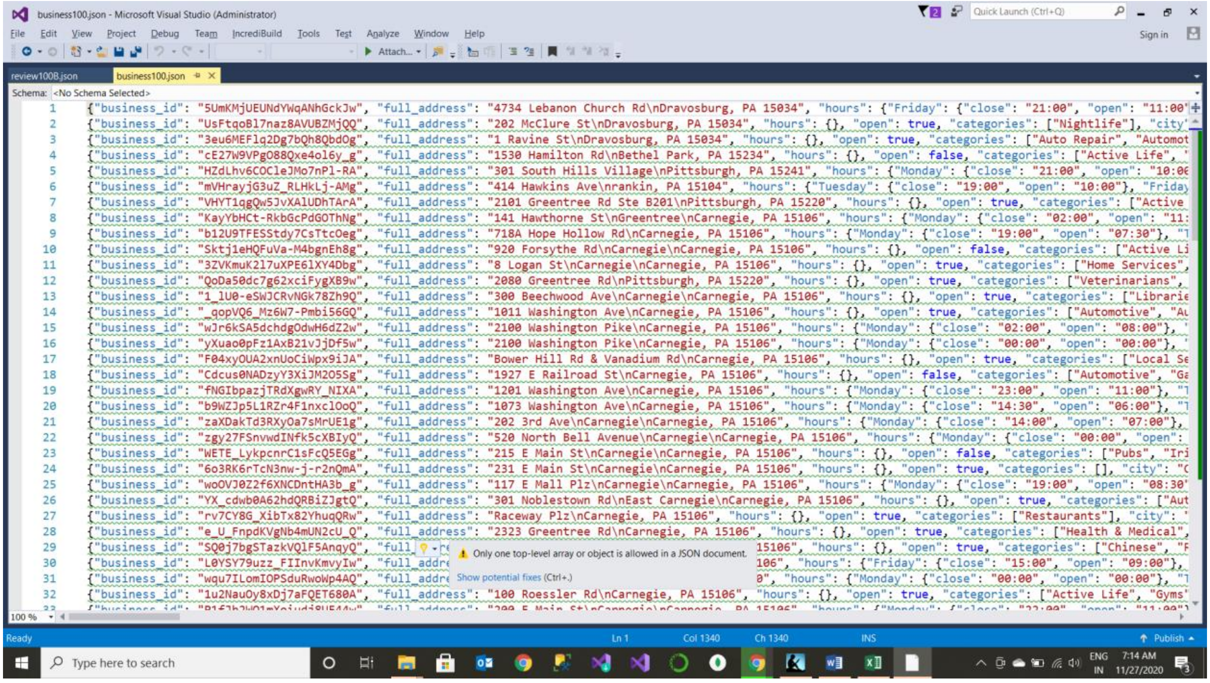
Download sample data

Using the business100.json and review100.json files



review100B.json - Microsoft Visual Studio (Administrator)
File Edit View Project Debug Team IncrediBuild Tools Test Analyze Window Help
Attach...
Quick Launch (Ctrl+Q)
Sign in
review100B.json
business100.json
Schema: <No Schema Selected>
100 %
Ready
Ln 1 Col 1 Ch 1 INS
Type here to search
ENG IN 7:14 AM 11/27/2020

```
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "PUFPaY9KxDAcGqfsorJp3Q", "review_id": "Ya85v4eqdd6k9Od8HbQjyA", "stars": 4, "date": "2012-08-01", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "Iu6AxdBYGR4A0wspR9BYHA", "review_id": "KPvLNJ21_4wbYNctrOwWdQ", "stars": 5, "date": "2014-02-13", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "auESFwWvW42h6alXgFxAXQ", "review_id": "fFSoGV46Yxuwbr3fHNuZig", "stars": 5, "date": "2015-10-31", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "uK8tzraOp4M5u3uYrqIBXg", "review_id": "Di3exaUCFNw1V4kSNW5pgA", "stars": 5, "date": "2013-11-08", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "I_47G-R2_egp7ME5u_ltew", "review_id": "0Lua2-PbqEQMjD9r89-asw", "stars": 3, "date": "2014-03-29", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "PP_xoMSYlGr2pb67BbqBdA", "review_id": "7N9j5YbBHBW6qguE5DAeyA", "stars": 1, "date": "2014-10-29", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "JPPhyFE-UE453zA6K0TVgw", "review_id": "mjCJR33jvUNt41iJCxDU_g", "stars": 4, "date": "2014-11-28", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "2d5HeDvZTDUNVog_WuUpSg", "review_id": "Ieh3kfZ-5J9pLju4JiQDvQ", "stars": 5, "date": "2014-02-27", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "BShxMIUwaJS378xcrz4Nmg", "review_id": "PU28OoBSHpZLkYGCmNxlmg", "stars": 5, "date": "2015-06-16", "t
{"votes": {"funny": 0, "useful": 1, "cool": 0}, "user_id": "fhNxoMwwTipzjO8A9LFe8Q", "review_id": "XsA6AojkWjOHA4FmuAb8XQ", "stars": 3, "date": "2012-08-19", "t
{"votes": {"funny": 0, "useful": 1, "cool": 0}, "user_id": "-6rEfobYjMxpUWLNxszaxQ", "review_id": "rkD7UDbQ9VM3Va6bI-eBHQ", "stars": 1, "date": "2013-04-18", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "KZuaJtFindQM9x2ZoMBxcQ", "review_id": "WExNE-f93SL4D1q8s9QWKg", "stars": 1, "date": "2013-07-14", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "H9E5VejGEsRhwcbOMFknmQ", "review_id": "iS34GJhMkkt9kCoTJLYEwA", "stars": 4, "date": "2013-08-16", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "ljwgUJowB69klaR8Au-H7g", "review_id": "S-G0D8Cy7PnqShoBZu8PCA", "stars": 4, "date": "2014-07-11", "t
{"votes": {"funny": 3, "useful": 7, "cool": 7}, "user_id": "JbAeIYc89Sk8SWmrBCJs9g", "review_id": "fBQ69-NU9ZyTjjS7Tb5tww", "stars": 5, "date": "2013-06-10", "t
{"votes": {"funny": 1, "useful": 1, "cool": 1}, "user_id": "l_szjd-ken3ma6oHDkTYXg", "review_id": "CFiLh7WvH7dM3qVZvNiacQ", "stars": 2, "date": "2013-12-23", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "zo_soThZw8eVglPbCRNC9A", "review_id": "UzMViMQZuSxOr5wrru3LwQ", "stars": 4, "date": "2014-09-04", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "4H-IfBDP5GkfNsQeHi4sHA", "review_id": "OEopG9TyhSyyrTk2Uy5SZA", "stars": 1, "date": "2015-05-06", "t
{"votes": {"funny": 1, "useful": 1, "cool": 1}, "user_id": "Qs5dcst13DnQJwaBcR4VGg", "review_id": "VXWdUDrVpBAvzaxnHpRUAw", "stars": 1, "date": "2015-11-02", "t
{"votes": {"funny": 0, "useful": 5, "cool": 0}, "user_id": "LWbYpcangjBMm4KPxZGOKg", "review_id": "6w6gMZ3iBLGcUM4RBIuifQ", "stars": 5, "date": "2012-12-01", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "m1FpV3EAeggaAdfPx0hBRQ", "review_id": "jVVv_DA5mCDB6mediuwHAw", "stars": 5, "date": "2013-03-15", "t
{"votes": {"funny": 0, "useful": 3, "cool": 1}, "user_id": "8fApIAMHn2MZJFUiCQto5Q", "review_id": "3Es8GsjkssusYgeU6_ZVpQ", "stars": 5, "date": "2013-03-30", "t
{"votes": {"funny": 0, "useful": 1, "cool": 0}, "user_id": "uK8tzraOp4M5u3uYrqIBXg", "review_id": "KAkcn7oQP1xX8KsZ-XmktA", "stars": 4, "date": "2013-10-20", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "6wvlM5L4_EroGXbnb_92xQ", "review_id": "BZNJkkP0bXnwQ2-sCqat2Q", "stars": 5, "date": "2013-11-07", "t
{"votes": {"funny": 0, "useful": 2, "cool": 1}, "user_id": "345nDw0oC-jOcglqxmzweQ", "review_id": "VDTIbR3G5_IPkpXbo2MutA", "stars": 5, "date": "2014-03-22", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "r1Uk-h_bddXb2ScCetoMPA", "review_id": "PY6nIMMLP2pqARlbIOJuRA", "stars": 5, "date": "2014-05-01", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "u9ULAsnYTdYH65Haj5LMSw", "review_id": "5uyYmniYyIB_wtKtyXDudQ", "stars": 4, "date": "2014-09-29", "t
{"votes": {"funny": 0, "useful": 3, "cool": 1}, "user_id": "tAKjY3bQXH51msJbOHYPmQ", "review_id": "uf61rPucuICXhSPXlZ1hIQ", "stars": 5, "date": "2014-11-03", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "Kq8-FUG7d_MT2qRNiNBJnA", "review_id": "xY2gjy49dnpQSB2RTZzqCw", "stars": 4, "date": "2014-12-18", "t
{"votes": {"funny": 1, "useful": 2, "cool": 4}, "user_id": "a0ULPTNi05w-BNVovVUz2w", "review_id": "1W1vq7D1ofwFKbCA_GWu_A", "stars": 5, "date": "2015-03-06", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "-ke0O45_B6STtEqd5s_ANw", "review_id": "6A5nASW8_XRwDxV29cWGUQ", "stars": 4, "date": "2015-03-08", "t
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "FoFdGS2TjVUUXptBS_gdNQ", "review_id": "RgYjyaOkSY1b5r9wirrfIQ", "stars": 4, "date": "2015-03-16", "t
```


business100.json - Microsoft Visual Studio (Administrator)
File Edit View Project Debug Team IncrediBuild Tools Test Analyze Window Help
Attach...
Quick Launch (Ctrl+Q)
Sign in
review100B.json
business100.json
Schema: <No Schema Selected>
100 %
Ready
Ln 1
Col 1340
Ch 1340
INS
Publish
Type here to search
ENG
IN
7:14 AM
11/27/2020

```
{"business_id": "5UmKMjUEUNdYWqANhGckJw", "full_address": "4734 Lebanon Church Rd\nDravosburg, PA 15034", "hours": {"Friday": {"close": "21:00", "open": "11:00"
{"business_id": "UsFtqoBl7naz8AVUBZMjQQ", "full_address": "202 McClure St\nDravosburg, PA 15034", "hours": {}, "open": true, "categories": ["Nightlife"], "city"
{"business_id": "3eu6MEFlq2Dg7bQh8QbdOg", "full_address": "1 Ravine St\nDravosburg, PA 15034", "hours": {}, "open": true, "categories": ["Auto Repair", "Automot
{"business_id": "cE27W9VPgO88Qxe4ol6y_g", "full_address": "1530 Hamilton Rd\nBethel Park, PA 15234", "hours": {}, "open": false, "categories": ["Active Life", "
{"business_id": "HZdLhv6COCleJMo7nPl-RA", "full_address": "301 South Hills Village\nPittsburgh, PA 15241", "hours": {"Monday": {"close": "21:00", "open": "10:00
{"business_id": "mVHrayjG3uZ_RLHkLj-AMg", "full_address": "414 Hawkins Ave\nrankin, PA 15104", "hours": {"Tuesday": {"close": "19:00", "open": "10:00"}, "Friday
{"business_id": "VHYT1qgQw5JvXAlUDhTArA", "full_address": "2101 Greentree Rd Ste B201\nPittsburgh, PA 15220", "hours": {}, "open": true, "categories": ["Active
{"business_id": "KayYbHCt-RkbGcPdGOThNg", "full_address": "141 Hawthorne St\nGreentree\nCarnegie, PA 15106", "hours": {"Monday": {"close": "02:00", "open": "11:
{"business_id": "b12U9TFESStdy7CsTtcOeg", "full_address": "718A Hope Hollow Rd\nCarnegie, PA 15106", "hours": {"Monday": {"close": "19:00", "open": "07:30"}, "T
{"business_id": "Sktj1eHQFuVa-M4bgnEh8g", "full_address": "920 Forsythe Rd\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Active Li
{"business_id": "3ZVKmuK2l7uXPE6lXY4Dbg", "full_address": "8 Logan St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Home Services",
{"business_id": "QoDa50dc7g62xciFygXB9w", "full_address": "2080 Greentree Rd\nPittsburgh, PA 15220", "hours": {}, "open": true, "categories": ["Veterinarians",
{"business_id": "1_lU0-eSWJCRvNGk78Zh9Q", "full_address": "300 Beechwood Ave\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Librarie
{"business_id": "_qopVQ6_Mz6W7-Pmbi56GQ", "full_address": "1011 Washington Ave\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Automotive", "Au
{"business_id": "wJr6kSA5dchdgOdwH6dZ2w", "full_address": "2100 Washington Pike\nCarnegie, PA 15106", "hours": {"Monday": {"close": "02:00", "open": "08:00"},
{"business_id": "yXuao0pFz1AxB21vJjDf5w", "full_address": "2100 Washington Pike\nCarnegie, PA 15106", "hours": {"Monday": {"close": "00:00", "open": "00:00"},
{"business_id": "F04xyOUA2xnUoCiWpx9iJA", "full_address": "Bower Hill Rd & Vanadium Rd\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Local Se
{"business_id": "Cdcus0NADzyY3XiJM2O5Sg", "full_address": "1927 E Railroad St\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Automotive", "Ga
{"business_id": "fNGIbpazjTRdXgwRY_NIXA", "full_address": "1201 Washington Ave\nCarnegie, PA 15106", "hours": {"Monday": {"close": "23:00", "open": "11:00"}, "T
{"business_id": "b9WZJp5L1RZr4F1nxclOoQ", "full_address": "1073 Washington Ave\nCarnegie, PA 15106", "hours": {"Monday": {"close": "14:30", "open": "06:00"}, "T
{"business_id": "zaXDakTd3RXyOa7sMrUE1g", "full_address": "202 3rd Ave\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "14:00", "open": "07:00"},
{"business_id": "zgy27FSnvwdINfk5cXBIyQ", "full_address": "520 North Bell Avenue\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "00:00", "open":
{"business_id": "WETE_LykpcnrC1sFcQ5EGg", "full_address": "215 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Pubs", "Iri
{"business_id": "6o3RK6rTcN3nw-j-r2nQmA", "full_address": "231 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": [], "city": "C
{"business_id": "woOVJ0Z2f6XNCDntHA3b_g", "full_address": "117 E Mall Plz\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "19:00", "open": "08:30"
{"business_id": "YX_cdwb0A62hdQRBiZJgtQ", "full_address": "301 Noblestown Rd\nEast Carnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Aut
{"business_id": "rv7CY8G_XibTx82YhuqQRw", "full_address": "Raceway Plz\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Restaurants"], "city": "
{"business_id": "e_U_FnpdKVgNb4mUN2cU_Q", "full_address": "2323 Greentree Rd\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Health & Medical",
{"business_id": "SQ0j7bgSTazkVQlF5AnqyQ", "full_re 15106", "hours": {}, "open": true, "categories": ["Chinese", "F
{"business_id": "L0YSY79uzz_FIInvKmvyIw", "full_addre 106", "hours": {"Friday": {"close": "15:00", "open": "09:00"},
{"business_id": "wqu7ILomIOPSduRwoWp4AQ", "full_addre 0", "hours": {"Monday": {"close": "00:00", "open": "00:00"}, "T
{"business_id": "1u2NauOy8xDj7aFQET680A", "full_address": "100 Roessler Rd\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Active Life", "Gyms"
```

Only one top-level array or object is allowed in a JSON document.

Show potential fixes (Ctrl+.)

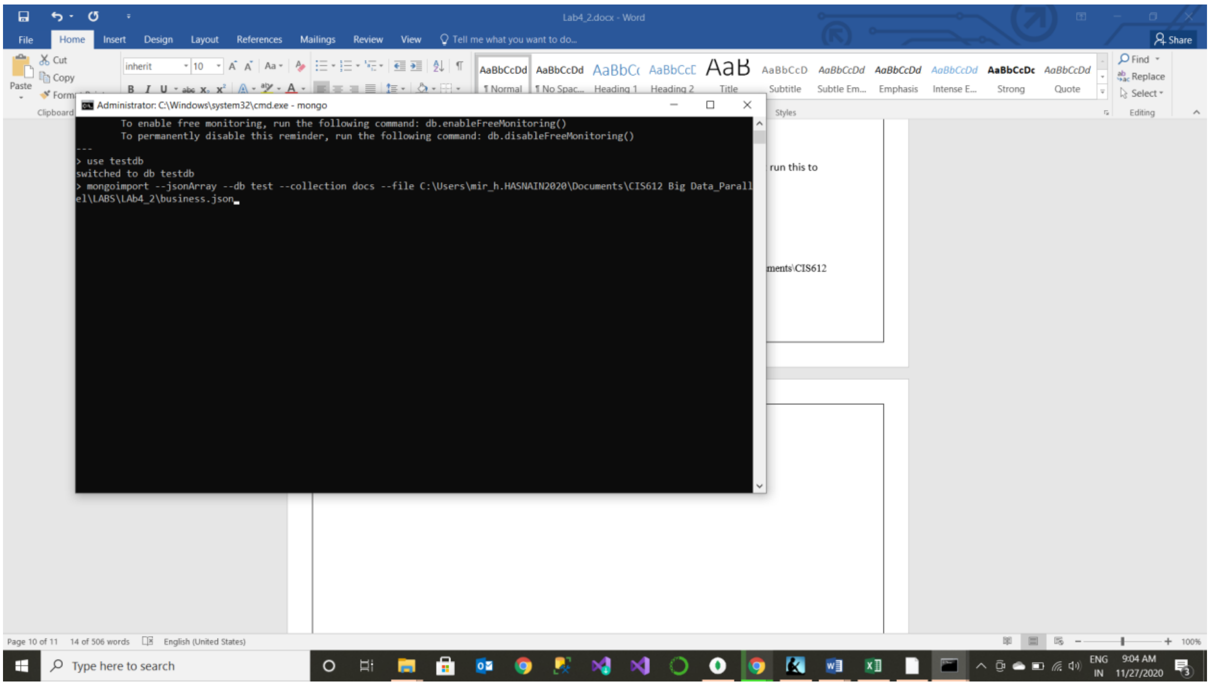
- Install MongoDB on the VM

- Start MongoDB – a default configuration file is installed by yum so you can just run this to start on localhost and the default port 27017

mongod -f /etc/mongod.conf

- Install MongoDB Hadoop Connector
- Inserting into the mongodb collection using

mongoimport --jsonArray --db test --collection docs --file C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parallel\LABS\LAb4_2\business.json
mongoimport --jsonArray --db test --collection docs --file C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parallel\LABS\LAb4_2\review.json



```
        To enable free monitoring, run the following command: db.enableFreeMonitoring()
        To permanently disable this reminder, run the following command: db.disableFreeMonitoring()
---
> use testdb
switched to db testdb
> mongoimport --jsonArray --db test --collection docs --file C:\Users\mir_h.HASNAIN2020\Documents\CIS612 Big Data_Parall
el\LABS\LAb4_2\business.json
```

Displaying the data in the collection using mongodb via Hadoop input format

# set up parameters for reading from MongoDB via Hadoop input format

config = {"mongo.input.uri": "mongodb://localhost:27017/YelpBusiness.business"}

inputFormatClassName = "com.mongodb.hadoop.MongoInputFormat"

keyClassName = "org.apache.hadoop.io.Text"

valueClassName = "org.apache.hadoop.io.MapWritable"

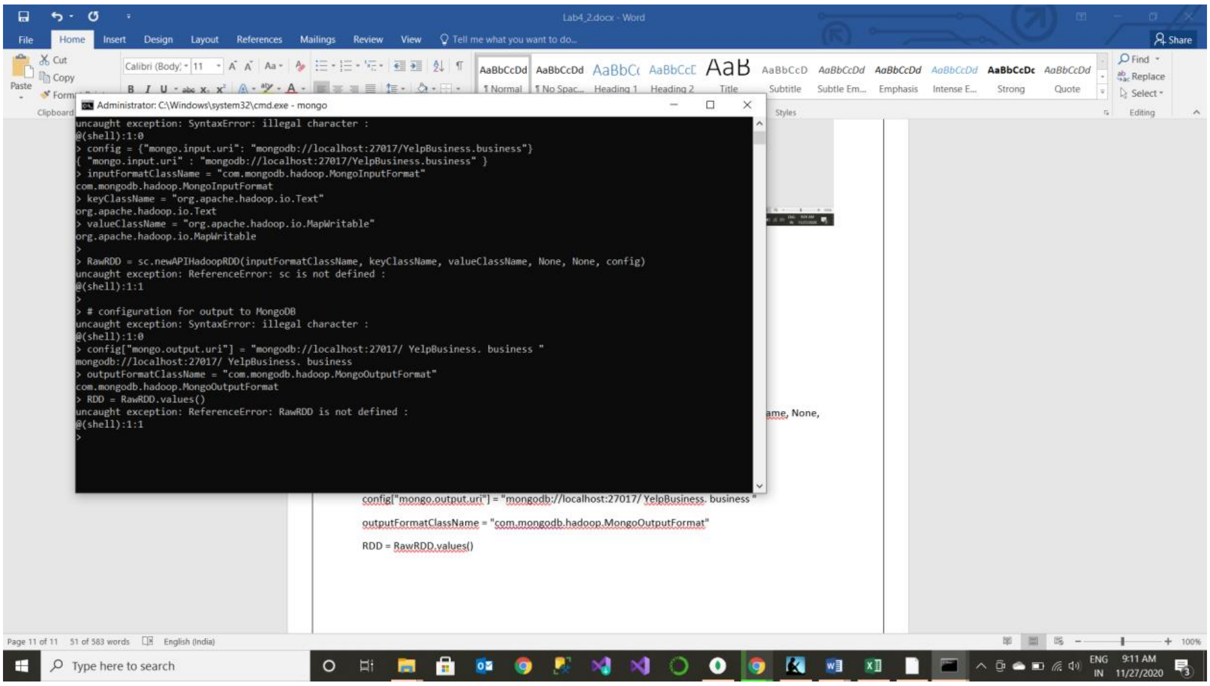
RawRDD = sc.newAPIHadoopRDD(inputFormatClassName, keyClassName, valueClassName, None, None, config)

# configuration for output to MongoDB

config["mongo.output.uri"] = "mongodb://localhost:27017/ YelpBusiness. business "

outputFormatClassName = "com.mongodb.hadoop.MongoOutputFormat"

RDD = RawRDD.values()



```
uncaught exception: SyntaxError: illegal character :
@(shell):1:0
> config = {"mongo.input.uri": "mongodb://localhost:27017/YelpBusiness.business"}
{ "mongo.input.uri" : "mongodb://localhost:27017/YelpBusiness.business" }
> inputFormatClassName = "com.mongodb.hadoop.MongoInputFormat"
com.mongodb.hadoop.MongoInputFormat
> keyClassName = "org.apache.hadoop.io.Text"
org.apache.hadoop.io.Text
> valueClassName = "org.apache.hadoop.io.MapWritable"
org.apache.hadoop.io.MapWritable
>
> RawRDD = sc.newAPIHadoopRDD(inputFormatClassName, keyClassName, valueClassName, None, None, config)
uncaught exception: ReferenceError: sc is not defined :
@(shell):1:1
>
> # configuration for output to MongoDB
uncaught exception: SyntaxError: illegal character :
@(shell):1:0
> config["mongo.output.uri"] = "mongodb://localhost:27017/ YelpBusiness. business "
mongodb://localhost:27017/ YelpBusiness. business
> outputFormatClassName = "com.mongodb.hadoop.MongoOutputFormat"
com.mongodb.hadoop.MongoOutputFormat
> RDD = RawRDD.values()
uncaught exception: ReferenceError: RawRDD is not defined :
@(shell):1:1
>
```

ame, None,

config["mongo.output.uri"] = "mongodb://localhost:27017/ YelpBusiness. business "

outputFormatClassName = "com.mongodb.hadoop.MongoOutputFormat"

RDD = RawRDD.values()