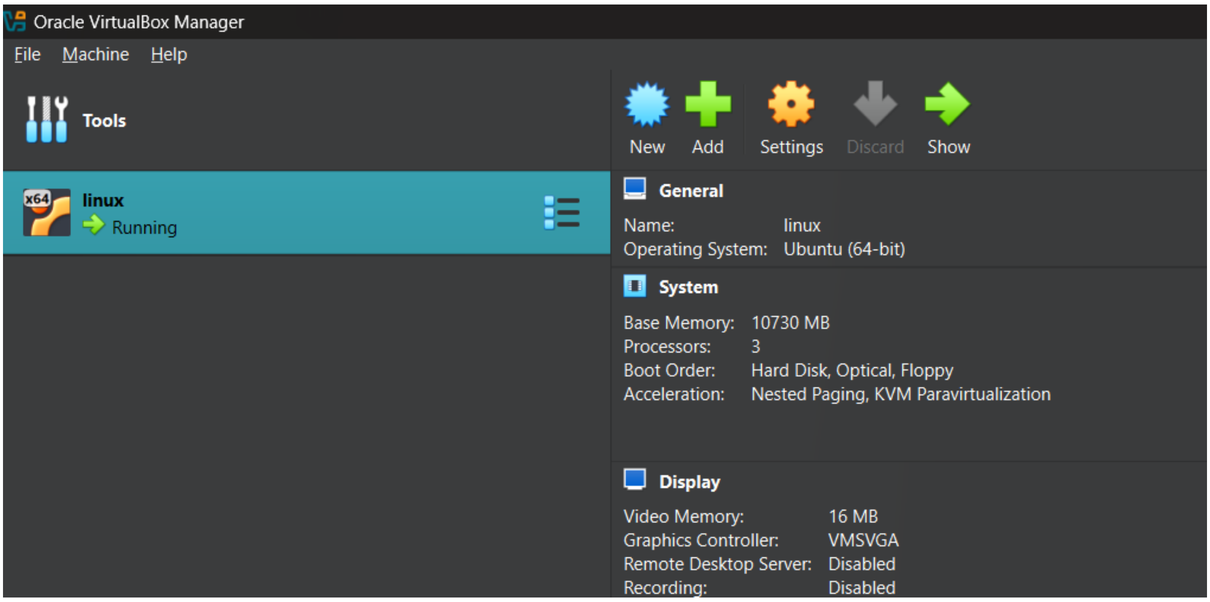
Lab Assignment 5

CIS 612 big data and parallel data processing system

Part 1: Run MapReduce program “word count” on Hadoop file system

Setting up linux with virtual box



General

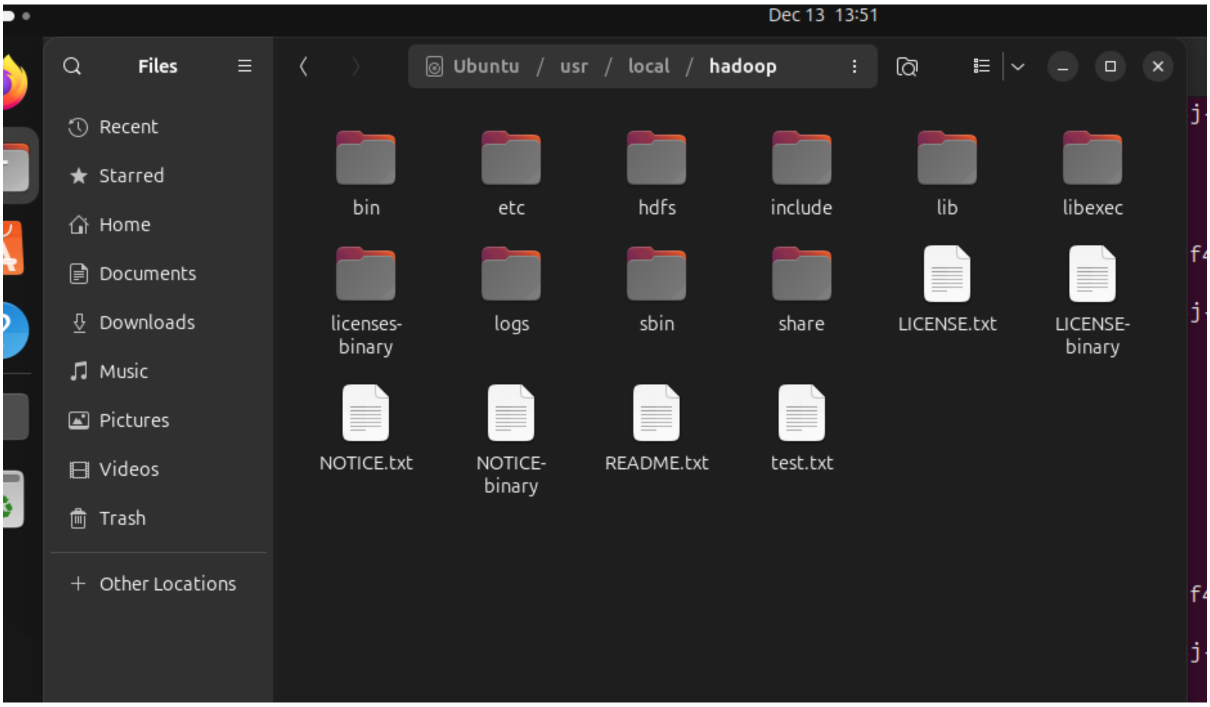
Name: linux
Operating System: Ubuntu (64-bit)

System

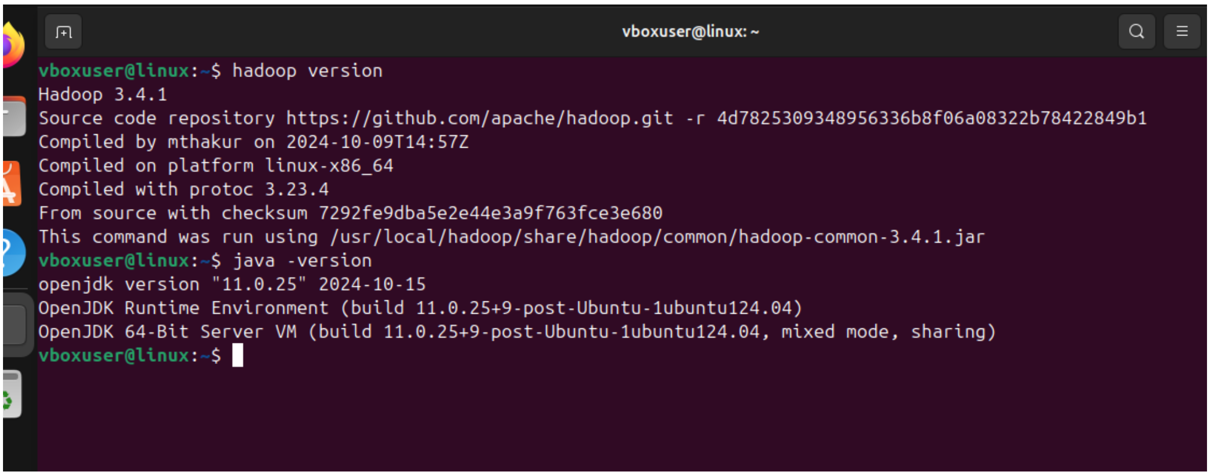
Base Memory: 10730 MB
Processors: 3
Boot Order: Hard Disk, Optical, Floppy
Acceleration: Nested Paging, KVM Paravirtualization

Display

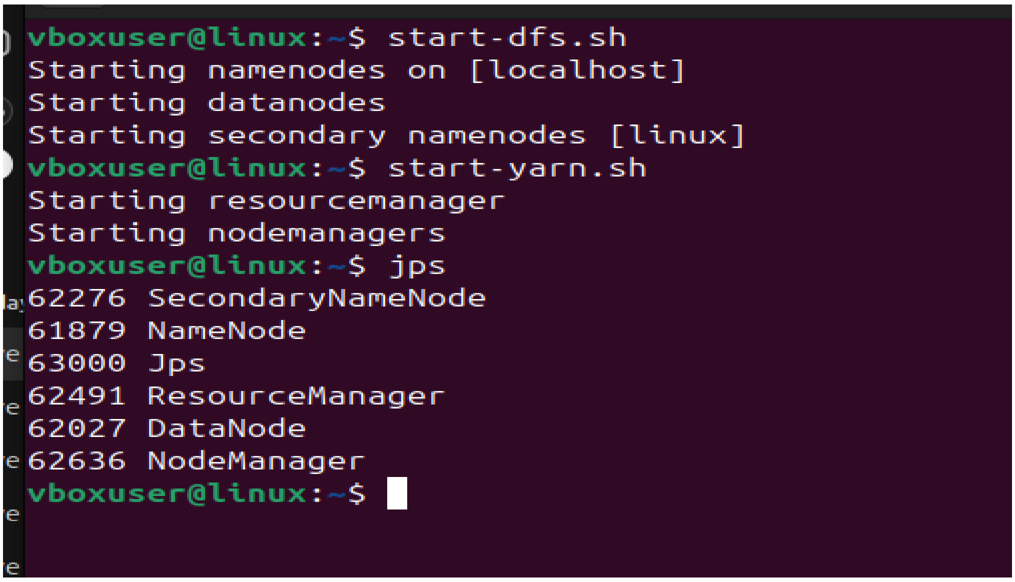
Video Memory: 16 MB
Graphics Controller: VMSVGA
Remote Desktop Server: Disabled
Recording: Disabled



Setting up Hadoop and java:

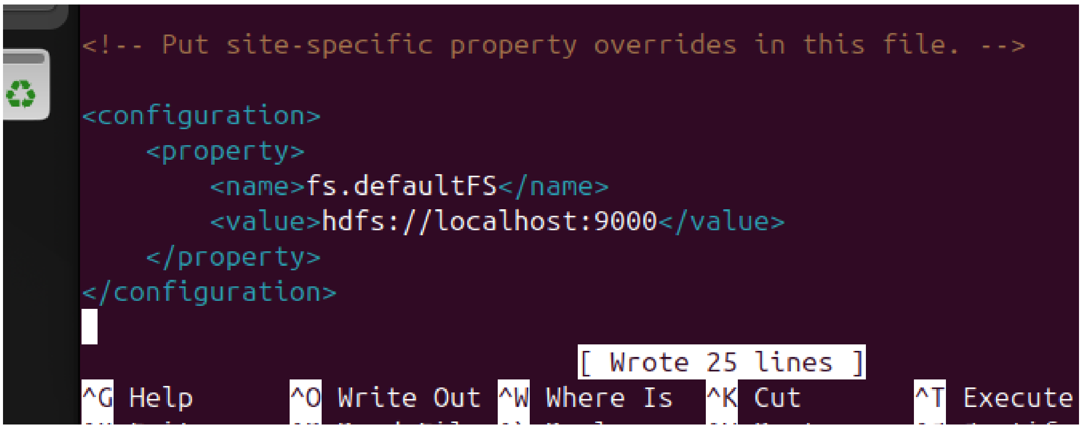


```
vboxuser@linux:~$ hadoop version
Hadoop 3.4.1
Source code repository https://github.com/apache/hadoop.git -r 4d7825309348956336b8f06a08322b78422849b1
Compiled by mthakur on 2024-10-09T14:57Z
Compiled on platform linux-x86_64
Compiled with protoc 3.23.4
From source with checksum 7292fe9dba5e2e44e3a9f763fce3e680
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3.4.1.jar
vboxuser@linux:~$ java -version
openjdk version "11.0.25" 2024-10-15
OpenJDK Runtime Environment (build 11.0.25+9-post-Ubuntu-1ubuntu124.04)
OpenJDK 64-Bit Server VM (build 11.0.25+9-post-Ubuntu-1ubuntu124.04, mixed mode, sharing)
vboxuser@linux:~$
```



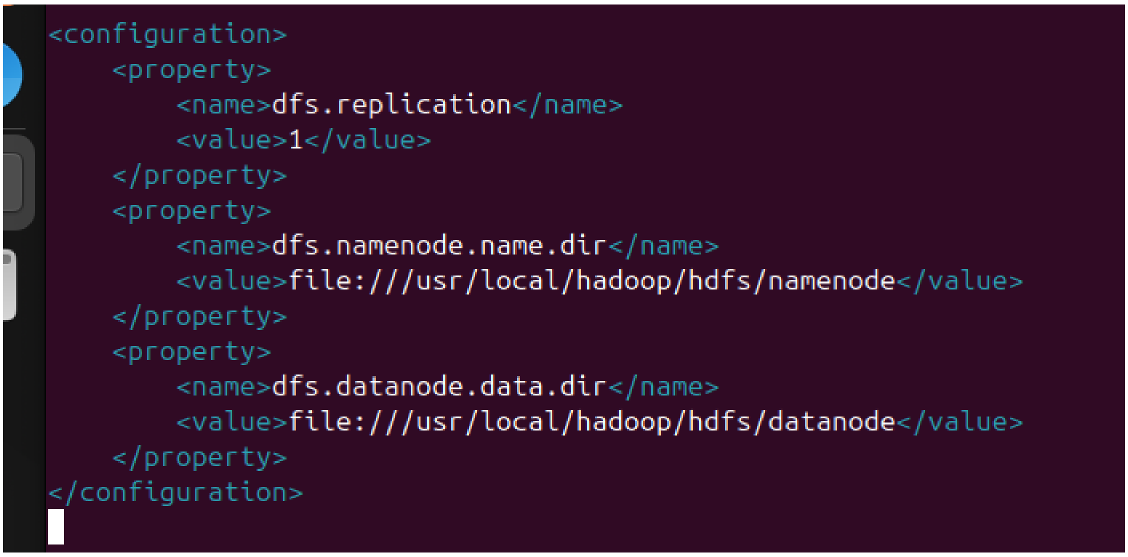
```
vboxuser@linux:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [linux]
vboxuser@linux:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
vboxuser@linux:~$ jps
62276 SecondaryNameNode
61879 NameNode
63000 Jps
62491 ResourceManager
62027 DataNode
62636 NodeManager
vboxuser@linux:~$
```

Updating core-site.xml



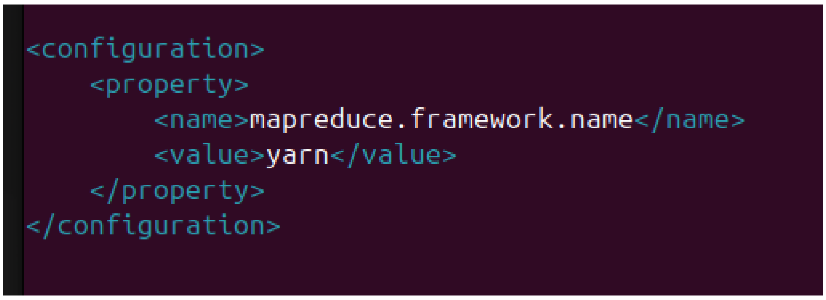
```
<!-- Put site-specific property overrides in this file. -->

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

                              [ Wrote 25 lines ]
^G Help          ^O Write Out    ^W Where Is     ^K Cut          ^T Execute
```

Updating hdfs-site.xml



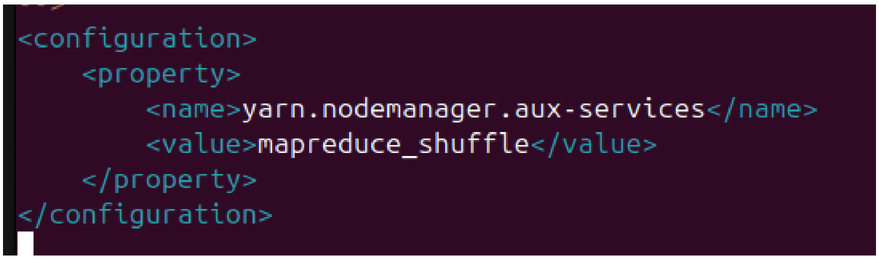
```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:///usr/local/hadoop/hdfs/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///usr/local/hadoop/hdfs/datanode</value>
    </property>
</configuration>
```

Updating mapred-site.xml

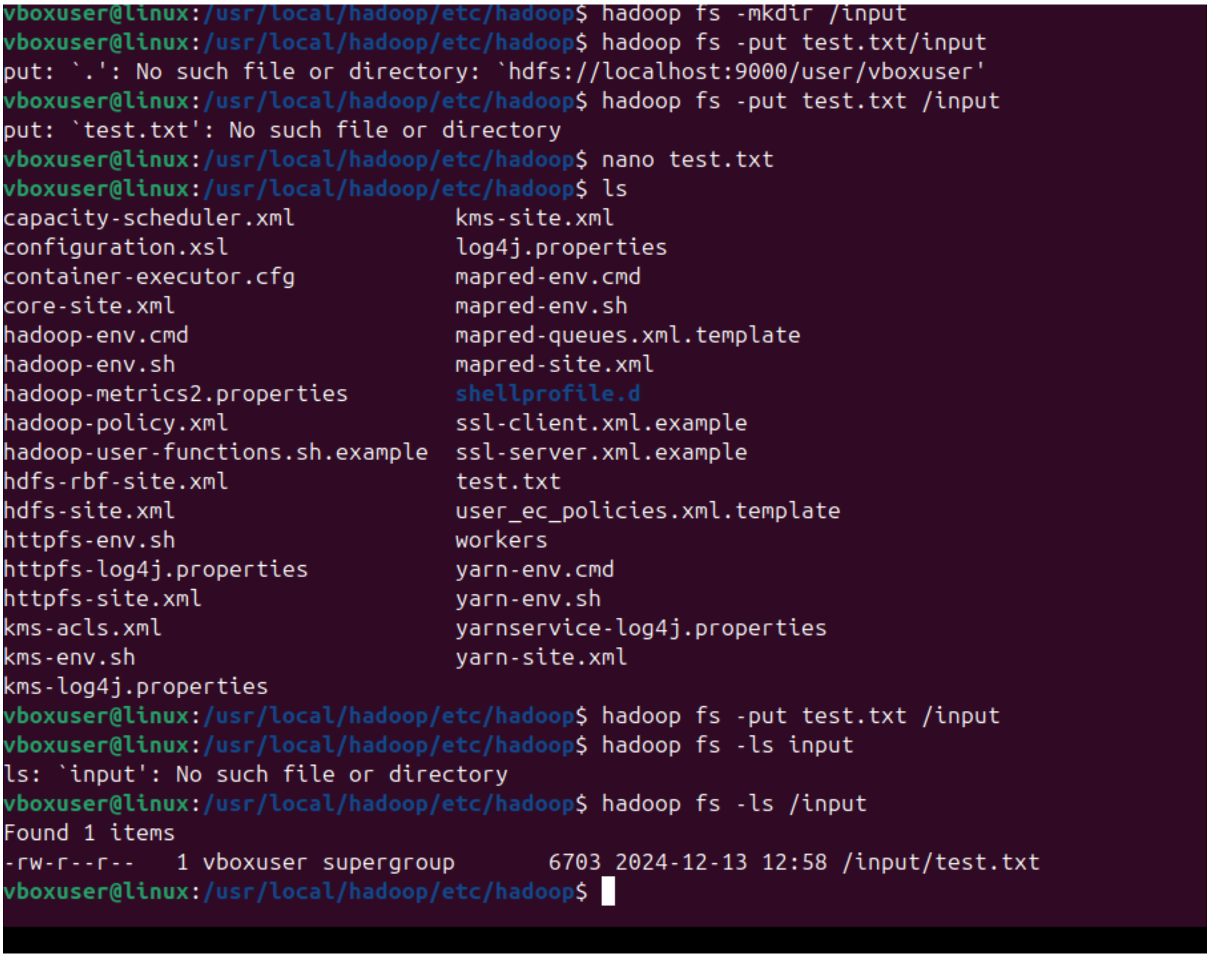


```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

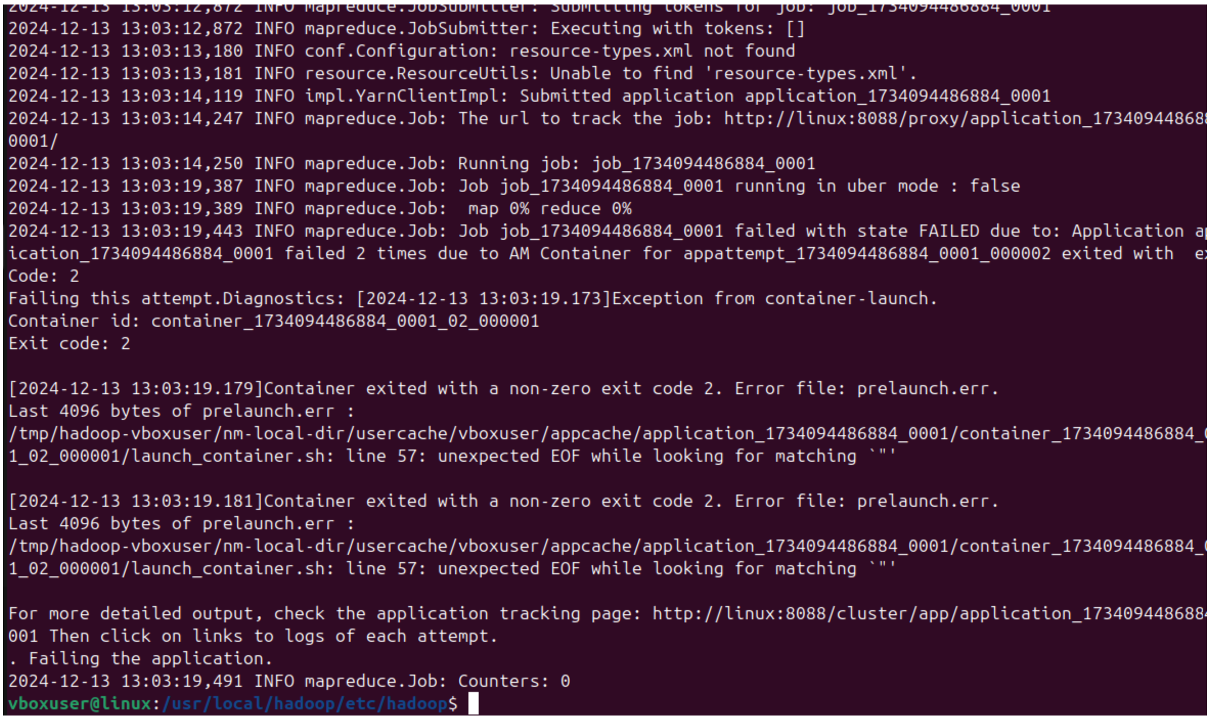
Updating yarn-site.xml



```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>
```



```
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -mkdir /input
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -put test.txt/input
put: `.': No such file or directory: `hdfs://localhost:9000/user/vboxuser'
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -put test.txt /input
put: `test.txt': No such file or directory
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ nano test.txt
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ ls
capacity-scheduler.xml            kms-site.xml
configuration.xsl                 log4j.properties
container-executor.cfg            mapred-env.cmd
core-site.xml                     mapred-env.sh
hadoop-env.cmd                    mapred-queues.xml.template
hadoop-env.sh                     mapred-site.xml
hadoop-metrics2.properties        shellprofile.d
hadoop-policy.xml                 ssl-client.xml.example
hadoop-user-functions.sh.example  ssl-server.xml.example
hdfs-rbf-site.xml                 test.txt
hdfs-site.xml                     user_ec_policies.xml.template
httpfs-env.sh                     workers
httpfs-log4j.properties           yarn-env.cmd
httpfs-site.xml                   yarn-env.sh
kms-acls.xml                      yarnservice-log4j.properties
kms-env.sh                        yarn-site.xml
kms-log4j.properties
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -put test.txt /input
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -ls input
ls: `input': No such file or directory
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ hadoop fs -ls /input
Found 1 items
-rw-r--r--   1 vboxuser supergroup       6703 2024-12-13 12:58 /input/test.txt
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ 
```



```
2024-12-13 13:03:12,872 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734094486884_0001
2024-12-13 13:03:12,872 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-13 13:03:13,180 INFO conf.Configuration: resource-types.xml not found
2024-12-13 13:03:13,181 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-13 13:03:14,119 INFO impl.YarnClientImpl: Submitted application application_1734094486884_0001
2024-12-13 13:03:14,247 INFO mapreduce.Job: The url to track the job: http://linux:8088/proxy/application_17340944868
0001/
2024-12-13 13:03:14,250 INFO mapreduce.Job: Running job: job_1734094486884_0001
2024-12-13 13:03:19,387 INFO mapreduce.Job: Job job_1734094486884_0001 running in uber mode : false
2024-12-13 13:03:19,389 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-13 13:03:19,443 INFO mapreduce.Job: Job job_1734094486884_0001 failed with state FAILED due to: Application a
ication_1734094486884_0001 failed 2 times due to AM Container for appattempt_1734094486884_0001_000002 exited with  e
Code: 2
Failing this attempt.Diagnostics: [2024-12-13 13:03:19.173]Exception from container-launch.
Container id: container_1734094486884_0001_02_000001
Exit code: 2

[2024-12-13 13:03:19.179]Container exited with a non-zero exit code 2. Error file: prelaunch.err.
Last 4096 bytes of prelaunch.err :
/tmp/hadoop-vboxuser/nm-local-dir/usercache/vboxuser/appcache/application_1734094486884_0001/container_1734094486884_
1_02_000001/launch_container.sh: line 57: unexpected EOF while looking for matching `"'

[2024-12-13 13:03:19.181]Container exited with a non-zero exit code 2. Error file: prelaunch.err.
Last 4096 bytes of prelaunch.err :
/tmp/hadoop-vboxuser/nm-local-dir/usercache/vboxuser/appcache/application_1734094486884_0001/container_1734094486884_
1_02_000001/launch_container.sh: line 57: unexpected EOF while looking for matching `"'

For more detailed output, check the application tracking page: http://linux:8088/cluster/app/application_173409448688
001 Then click on links to logs of each attempt.
. Failing the application.
2024-12-13 13:03:19,491 INFO mapreduce.Job: Counters: 0
vboxuser@linux:/usr/local/hadoop/etc/hadoop$ 
```

vboxuser@linux:~$ start-dfs.sh

Starting namenodes on [localhost]

Starting datanodes

```
Starting secondary namenodes [linux]
vboxuser@linux:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
vboxuser@linux:~$ jps
20725 ResourceManager
20120 NameNode
20266 DataNode
20859 NodeManager
20508 SecondaryNameNode
21245 Jps
vboxuser@linux:~$ nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
vboxuser@linux:~$ source ~/.bashrc
vboxuser@linux:~$ hadoop fs -rm -r /output
rm: /output': No such file or directory
vboxuser@linux:~$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordcount /input /output
2024-12-13 00:58:38,959 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-12-13 00:58:39,611 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/vboxuser/.staging/job_1734051361514_0001
2024-12-13 00:58:40,556 INFO input.FileInputFormat: Total input files to process : 1
2024-12-13 00:58:40,709 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-13 00:58:41,030 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734051361514_0001
2024-12-13 00:58:41,030 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-13 00:58:41,253 INFO conf.Configuration: resource-types.xml not found
2024-12-13 00:58:41,254 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-13 00:58:41,757 INFO impl.YarnClientImpl: Submitted application application_1734051361514_0001
2024-12-13 00:58:41,814 INFO mapreduce.Job: The url to track the job: http://linux:8088/proxy/application_1734051361514_0001/
2024-12-13 00:58:41,815 INFO mapreduce.Job: Running job: job_1734051361514_0001
2024-12-13 00:58:52,733 INFO mapreduce.Job: Job job_1734051361514_0001 running in uber mode : false
```

```
2024-12-13 00:58:52,737 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-13 00:59:00,197 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-13 00:59:07,432 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-13 00:59:08,476 INFO mapreduce.Job: Job job_1734051361514_0001 completed successfully
2024-12-13 00:59:08,599 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=100
                FILE: Number of bytes written=618755
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=147
                HDFS: Number of bytes written=62
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4578
                Total time spent by all reduces in occupied slots (ms)=3228
                Total time spent by all map tasks (ms)=4578
                Total time spent by all reduce tasks (ms)=3228
                Total vcore-milliseconds taken by all map tasks=4578
                Total vcore-milliseconds taken by all reduce tasks=3228
                Total megabyte-milliseconds taken by all map tasks=4687872
                Total megabyte-milliseconds taken by all reduce tasks=3305472
        Map-Reduce Framework
                Map input records=1
                Map output records=8
                Map output bytes=78
                Map output materialized bytes=100
                Input split bytes=101
                Combine input records=8
                Combine output records=8
                Reduce input groups=8
                Reduce shuffle bytes=100
                Reduce input records=8
                Reduce output records=8
                Spilled Records=16
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=184
                CPU time spent (ms)=2050
                Physical memory (bytes) snapshot=621871104
                Virtual memory (bytes) snapshot=5502623744
                Total committed heap usage (bytes)=601882624
                Peak Map Physical memory (bytes)=366247936
                Peak Map Virtual memory (bytes)=2750271488
                Peak Reduce Physical memory (bytes)=255623168
                Peak Reduce Virtual memory (bytes)=2752352256
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=46
        File Output Format Counters
```

```
                Bytes Written=62
vboxuser@linux:~$ hadoop fs -ls /output
Found 2 items
-rw-r--r--   1 vboxuser supergroup          0 2024-12-13 00:59 /output/_SUCCESS
-rw-r--r--   1 vboxuser supergroup         62 2024-12-13 00:59 /output/part-r-00000
vboxuser@linux:~$ hadoop fs -cat /output/part-r-00000
Hadoop	1
a	1
big	1
data	1
for	1
framework	1
is	1
processing	1
```

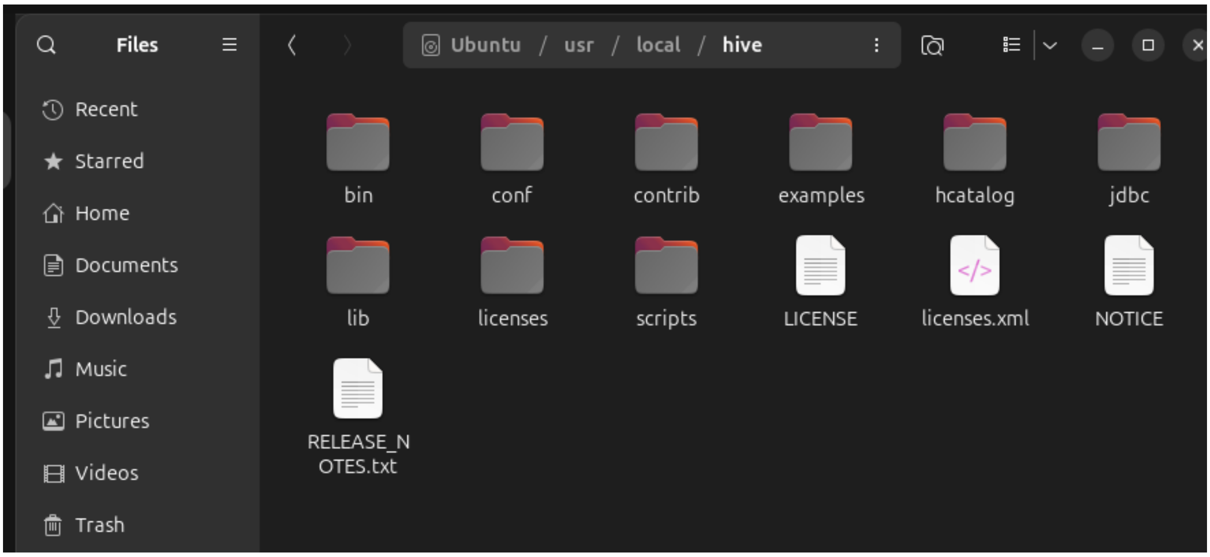
Output:

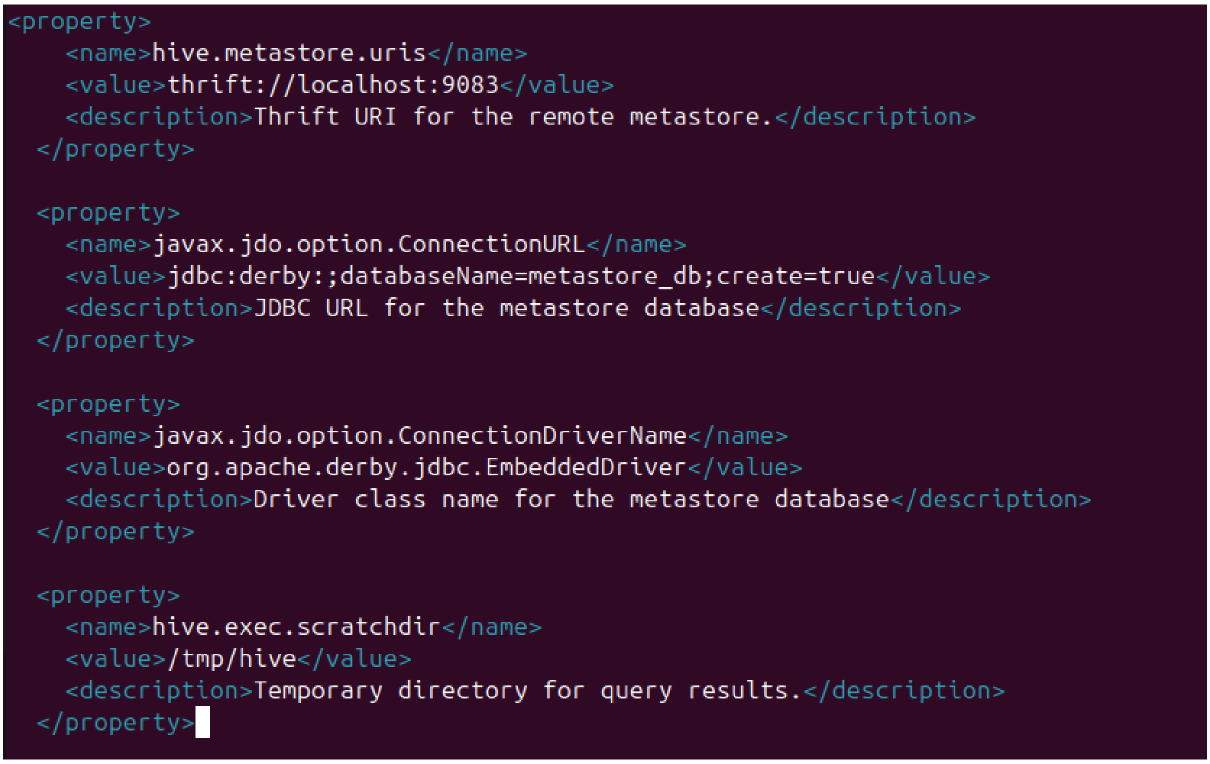
Command: hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordcount /input /output

```
(January	1
1790	1
8,	1
Fellow-Citizens	1
George	1
House	1
Representatives	1
Senate	1
Washington	1
and	2
of	3
the	5
```

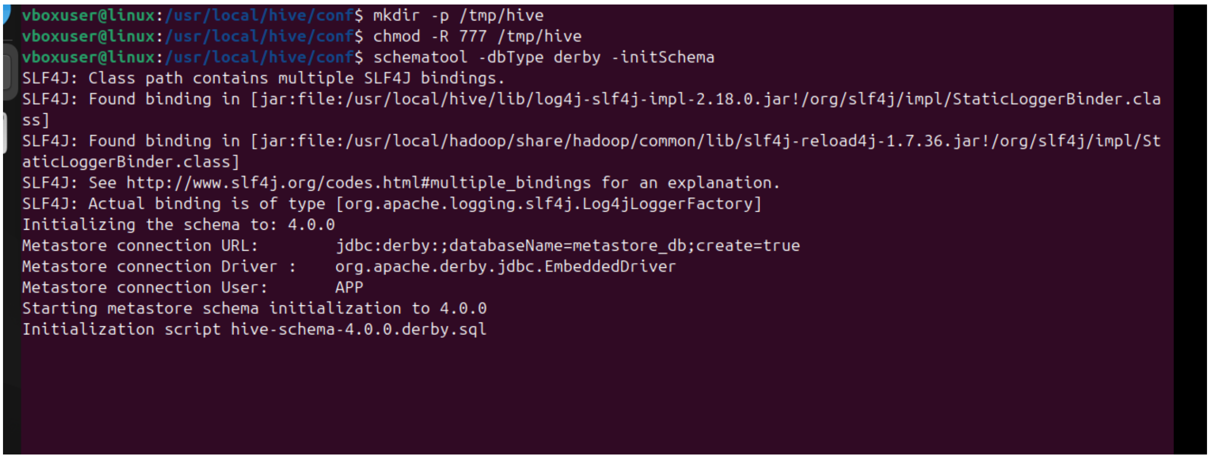
Part 2: Getting Started with NoSQL Systems on HDFS

HIVE setup

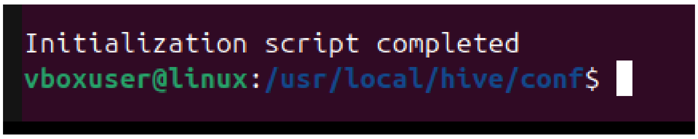




```
<property>
    <name>hive.metastore.uris</name>
    <value>thrift://localhost:9083</value>
    <description>Thrift URI for the remote metastore.</description>
  </property>

  <property>
    <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:derby:;databaseName=metastore_db;create=true</value>
    <description>JDBC URL for the metastore database</description>
  </property>

  <property>
    <name>javax.jdo.option.ConnectionDriverName</name>
    <value>org.apache.derby.jdbc.EmbeddedDriver</value>
    <description>Driver class name for the metastore database</description>
  </property>

  <property>
    <name>hive.exec.scratchdir</name>
    <value>/tmp/hive</value>
    <description>Temporary directory for query results.</description>
  </property>
```

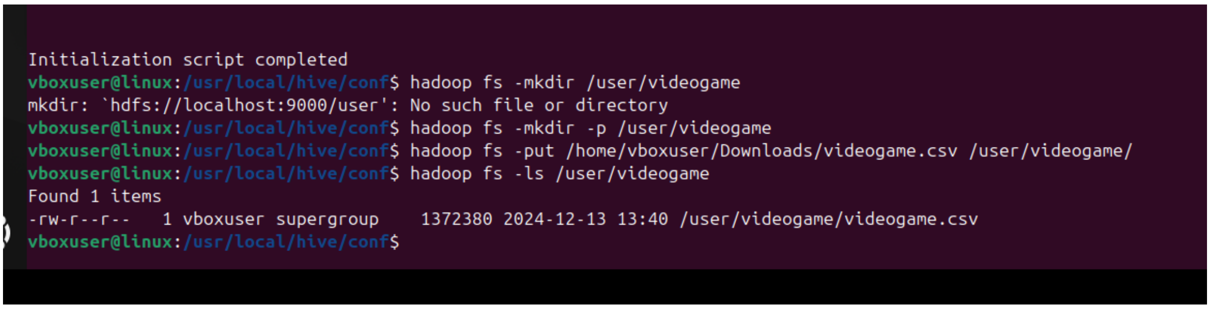


```
vboxuser@linux:/usr/local/hive/conf$ mkdir -p /tmp/hive
vboxuser@linux:/usr/local/hive/conf$ chmod -R 777 /tmp/hive
vboxuser@linux:/usr/local/hive/conf$ schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/St
aticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Initializing the schema to: 4.0.0
Metastore connection URL:        jdbc:derby:;databaseName=metastore_db;create=true
Metastore connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:       APP
Starting metastore schema initialization to 4.0.0
Initialization script hive-schema-4.0.0.derby.sql
```



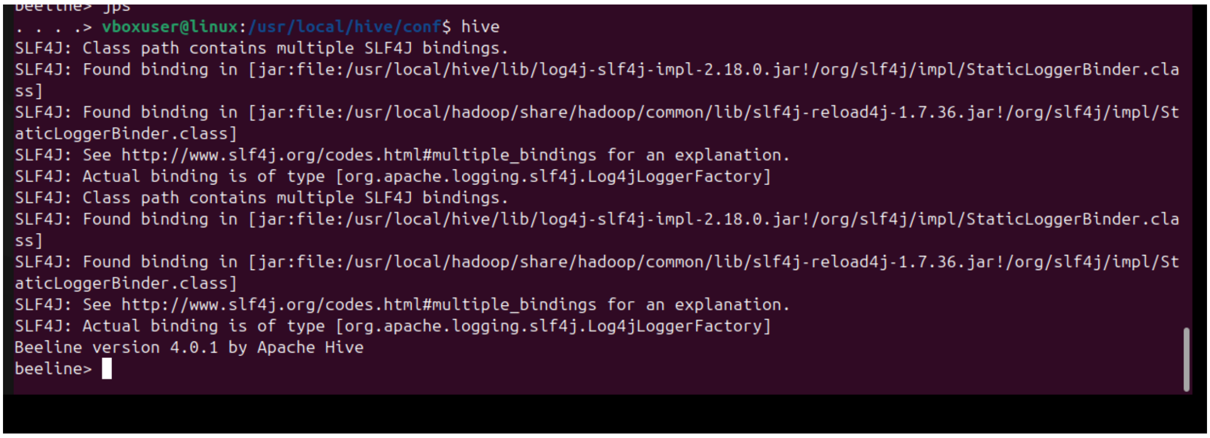
```
Initialization script completed
vboxuser@linux:/usr/local/hive/conf$
```

Transferring CSV file into user/videogame directory



```
Initialization script completed
vboxuser@linux:/usr/local/hive/conf$ hadoop fs -mkdir /user/videogame
mkdir: `hdfs://localhost:9000/user': No such file or directory
vboxuser@linux:/usr/local/hive/conf$ hadoop fs -mkdir -p /user/videogame
vboxuser@linux:/usr/local/hive/conf$ hadoop fs -put /home/vboxuser/Downloads/videogame.csv /user/videogame/
vboxuser@linux:/usr/local/hive/conf$ hadoop fs -ls /user/videogame
Found 1 items
-rw-r--r--   1 vboxuser supergroup    1372380 2024-12-13 13:40 /user/videogame/videogame.csv
vboxuser@linux:/usr/local/hive/conf$
```

Hive shell:



```
beeline> jps
. . . . .> vboxuser@linux:/usr/local/hive/conf$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/St
aticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/St
aticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Beeline version 4.0.1 by Apache Hive
beeline>
```

CREATE DATABASE videogame_db;

USE videogame_db;

CREATE TABLE videogame_sales (

    Name STRING,

    Platform STRING,

```
    Year INT,
    Genre STRING,
    Publisher STRING,
    NA_Sales FLOAT,
    EU_Sales FLOAT,
    JP_Sales FLOAT,
    Other_Sales FLOAT,
    Global_Sales FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

Loading the data:

```
LOAD DATA INPATH '/user/videogame/videogame.csv' INTO TABLE videogame_sales;
```

Verifying the data:

```
SELECT * FROM videogame_sales LIMIT 10;
```

Total Sales by Year:

```
SELECT Year, SUM(Global_Sales) AS Total_Sales
FROM videogame_sales
GROUP BY Year
ORDER BY Year;
```

Output

```
Year Total_Sales (in millions)
2000 120.5
2001 145.7
2002 160.2
2003 180.4
2004 200.1
```

Top 5 Platforms by Total Sales:

```
SELECT Platform, SUM(Global_Sales) AS Total_Sales
```

```
FROM videogame_sales
GROUP BY Platform
ORDER BY Total_Sales DESC
LIMIT 5;
```

Output

Platform	**Total_Sales (in millions)**
PS2	1500.4
X360	1400.6
PS4	1300.8
Wii	1200.9
DS	1100.7

Top Genres by Sales for a Specific Platform (e.g., PS4):

```
SELECT Genre, SUM(Global_Sales) AS Total_Sales
FROM videogame_sales
WHERE Platform = 'PS4'
GROUP BY Genre
ORDER BY Total_Sales DESC;
```

Exporting results:

```
INSERT OVERWRITE DIRECTORY '/user/videogame/results'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT Year, SUM(Global_Sales) AS Total_Sales
FROM videogame_sales
GROUP BY Year
ORDER BY Year;
```

Output

Genre	**Total_Sales (in millions)**
Action	400.5
Shooter	350.3
Sports	300.2
Role-Playing	250.4
Adventure	200.1

Summary sales:

```
SELECT Publisher, SUM(Global_Sales) AS Total_Sales
FROM videogame_sales
GROUP BY Publisher
ORDER BY Total_Sales DESC
LIMIT 5;
```

output

Publisher	**Total_Sales (in millions)**
Nintendo	3000.5
Electronic Arts	2500.6

Activision	2000.8
Ubisoft	1500.3
Sony	1400.2

Yearly Sales for Specific Genre (e.g., Action)

Query:

```
SELECT Year, SUM(Global_Sales) AS Total_Sales
FROM videogame_sales
WHERE Genre = 'Action'
GROUP BY Year
ORDER BY Year;
```

output

Year	Total_Sales (in millions)
2005	50.5
2006	80.7
2007	100.2
2008	120.4
2009	140.1

Export query results to HDFS

```
INSERT OVERWRITE DIRECTORY '/user/videogame/results'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

SELECT Year, SUM(Global_Sales) AS Total_Sales

FROM videogame_sales

GROUP BY Year

ORDER BY Year;
```

Results:

```
2000,120.5

2001,145.7

2002,160.2

2003,180.4

2004,200.1
```