

Business Objective:

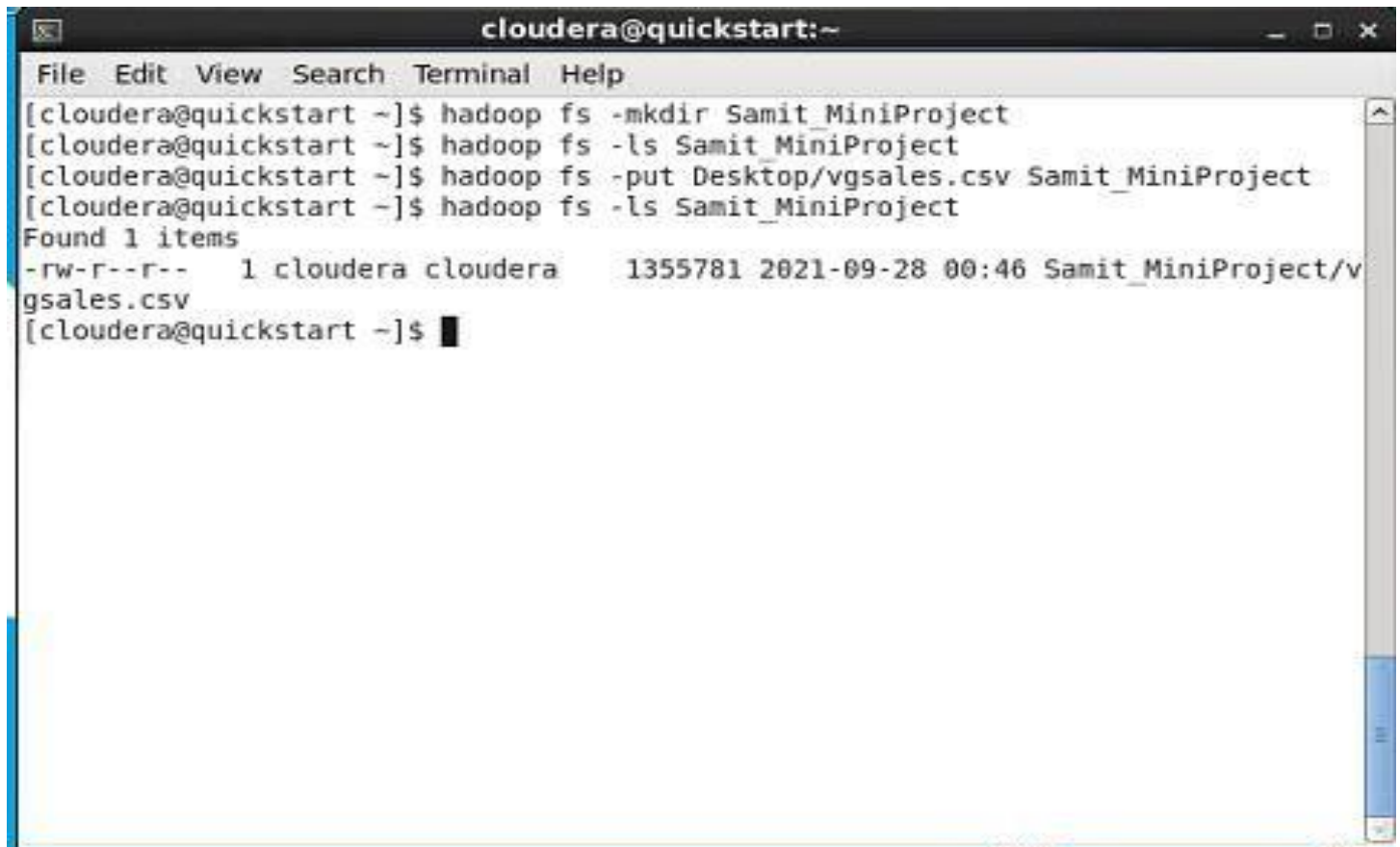
To analyze the psychological benefits and effects of playing Video Games on the data taken from the scrap of vgchartz.com

Business Questions

- Which place did “Call of Duty: Ghosts” gain the maximum sales?
- What genre of games in the most popular?
- Out of the entire data given to you, how many percent games does Electronic Arts publish?
- Which game reported the highest sales in Japan ever?
- Which games in North America report a sale of over 5 million USD?
- Which game reported the highest sale globally and in what year?
- How much sale did “Grand Theft Auto: San Andreas” report in the rest of the world?

Solutions:

Creating directory & putting vgsales.csv data into the directory;



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hadoop fs -mkdir Samit_MiniProject  
[cloudera@quickstart ~]$ hadoop fs -ls Samit_MiniProject  
[cloudera@quickstart ~]$ hadoop fs -put Desktop/vgsales.csv Samit_MiniProject  
[cloudera@quickstart ~]$ hadoop fs -ls Samit_MiniProject  
Found 1 items  
-rw-r--r--  1 cloudera cloudera    1355781 2021-09-28 00:46 Samit_MiniProject/vgsales.csv  
[cloudera@quickstart ~]$
```

Entering to grunt shell by the command; pig

```

cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2021-09-28 00:51:16,132 [main] INFO org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.12.0 (rexported) compiled Jun 29 2017, 04:34:31
2021-09-28 00:51:16,133 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1632815476117.log
2021-09-28 00:51:16,168 [main] INFO org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2021-09-28 00:51:16,615 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-28 00:51:16,615 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-28 00:51:16,615 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.clo
udera:8020

```

Loading the Data,

```

grunt> dataset = LOAD 'Samit_MiniProject/vgsales.csv' USING PigStorage(',') AS (
rank:int, name:chararray, platform:chararray, year:int, genre:chararray, publish
er:chararray, na_sales:float, eu_sales:float, jp_sales:float, other_sales:float,
global_sales:float);
grunt>

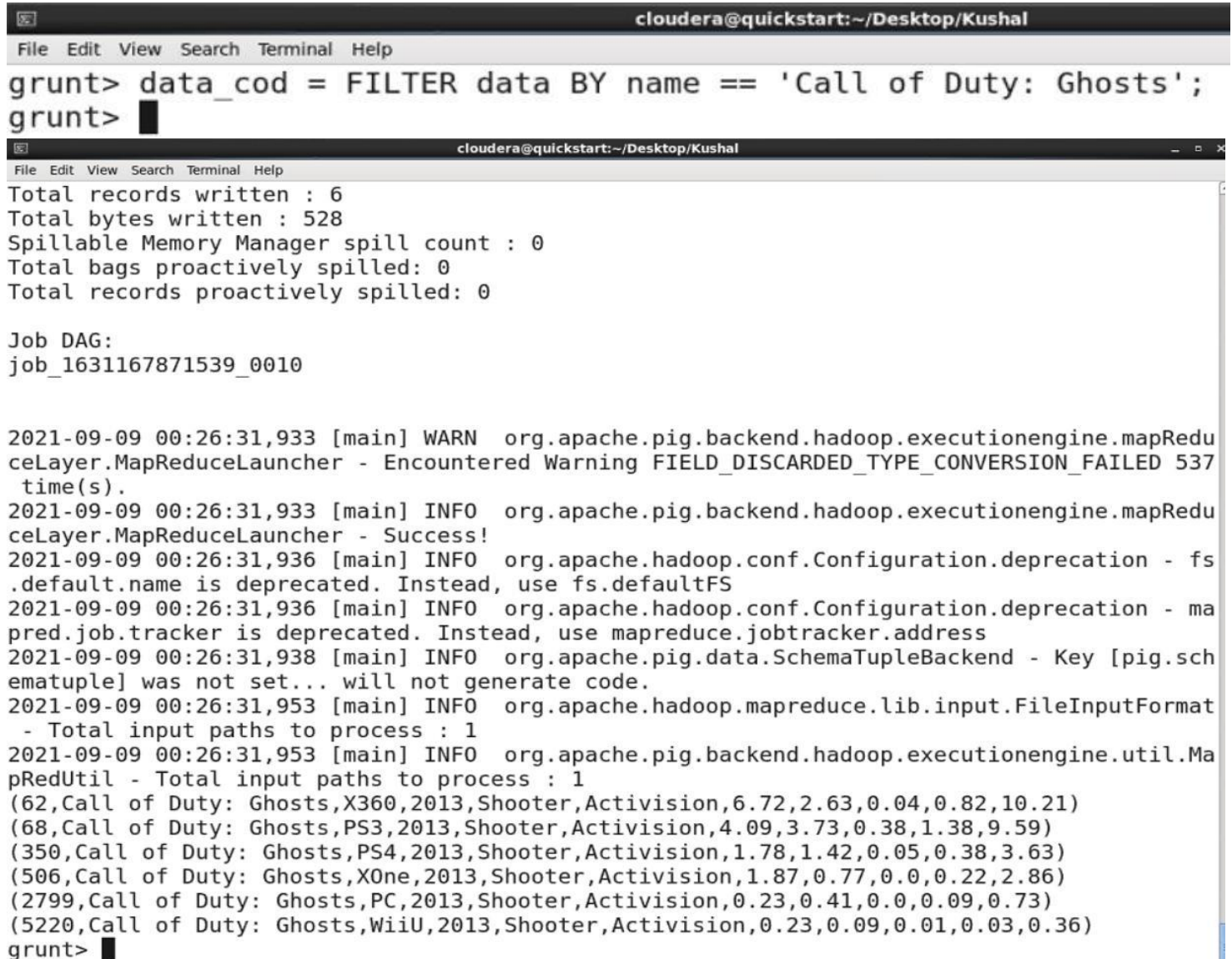
```

a. Which place did “Call of Duty: Ghosts” gain the maximum sales?

Code:

```
Grunt> data_cod = FILTER data BY name == 'Call of Duty: Ghosts';
Grunt> Dump data_cod;
```

Output:



```
cloudera@quickstart:~/Desktop/Kushal
File Edit View Search Terminal Help
grunt> data_cod = FILTER data BY name == 'Call of Duty: Ghosts';
grunt>

cloudera@quickstart:~/Desktop/Kushal
File Edit View Search Terminal Help
Total records written : 6
Total bytes written : 528
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1631167871539_0010

2021-09-09 00:26:31,933 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapRedu
ceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 537
time(s).
2021-09-09 00:26:31,933 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapRedu
ceLayer.MapReduceLauncher - Success!
2021-09-09 00:26:31,936 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs
.default.name is deprecated. Instead, use fs.defaultFS
2021-09-09 00:26:31,936 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - ma
pred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-09 00:26:31,938 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.sch
ematuple] was not set... will not generate code.
2021-09-09 00:26:31,953 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat
- Total input paths to process : 1
2021-09-09 00:26:31,953 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Ma
pRedUtil - Total input paths to process : 1
(62,Call of Duty: Ghosts,X360,2013,Shooter,Activision,6.72,2.63,0.04,0.82,10.21)
(68,Call of Duty: Ghosts,PS3,2013,Shooter,Activision,4.09,3.73,0.38,1.38,9.59)
(350,Call of Duty: Ghosts,PS4,2013,Shooter,Activision,1.78,1.42,0.05,0.38,3.63)
(506,Call of Duty: Ghosts,XOne,2013,Shooter,Activision,1.87,0.77,0.0,0.22,2.86)
(2799,Call of Duty: Ghosts,PC,2013,Shooter,Activision,0.23,0.41,0.0,0.09,0.73)
(5220,Call of Duty: Ghosts,WiiU,2013,Shooter,Activision,0.23,0.09,0.01,0.03,0.36)
grunt>
```

Conclusion:

Used filter operator found the maximum sales from “Call of Duty: Ghosts”

b. What genre of games in the most popular?

Code:

```
grunt> group_genre = GROUP data BY genre;
grunt> result = FOREACH group_genre GENERATE group, SUM(global);
grunt> Dump result;
```

Output:



```
cloudera@quickstart:~/Desktop/Kushal
File Edit View Search Terminal Help

grunt> group_genre = GROUP data BY genre;
grunt> result = FOREACH group_genre GENERATE group, SUM(global);
```

```
cloudera@quickstart
File Edit View Search Terminal Help

(2001,0.14000000059604645)
(2002,0.059999999865889549)
(2003,0.079999999821186066)
(2004,0.009999999776482582)
(2005,0.1699999962002039)
(2006,0.0)
(2007,0.08999999798834324)
(2008,0.009999999776482582)
(2009,0.06999999843537807)
(2010,0.04999999888241291)
(2011,0.28999999910593033)
(2013,0.0)
(2014,0.0)
(2015,0.0)
(2016,0.0)
(Misc,806.0099999178201)
(X360,0.0)
(Genre,)
(Action,1744.2499968893826)
(Puzzle,242.83000080287457)
(Racing,731.6699981186539)
(Sports,1330.529995502904)
(Shooter,1032.0499989185482)
( O-Kata",0.0)
(Fighting,448.8999999817461)
(Platform,829.4800044577569)
(Strategy,173.9999998640269)
(Adventure,234.54999988898635)
(Simulation,390.1100004184991)
(Role-Playing,926.2900028023869)
( Wii & PC Versions)",0.7600000146776438)
grunt>
```

Conclusion:

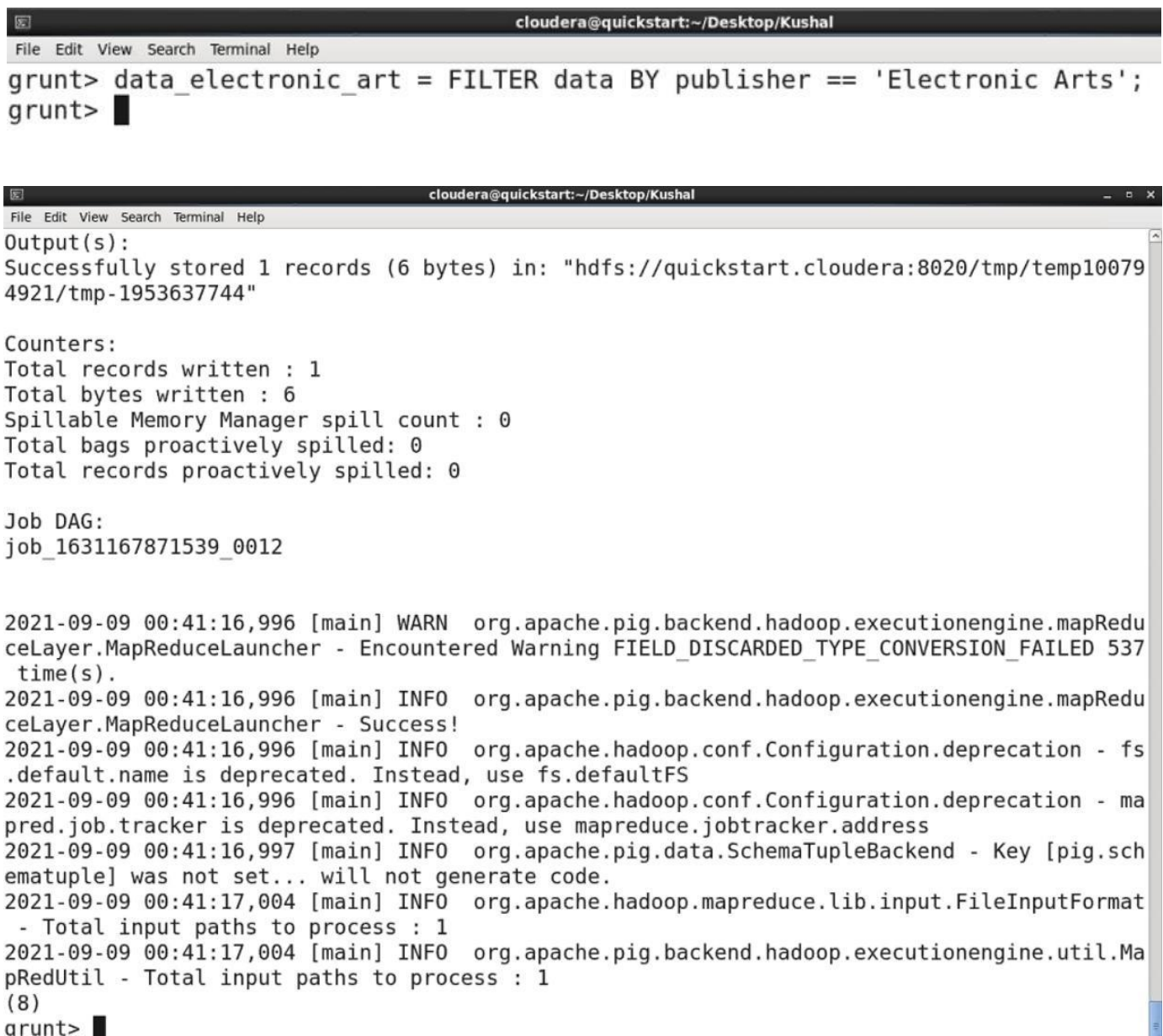
Grouped data by genre then using foreach operator generated genre of the games.

c. Out of the entire data given to you, how many percent games does Electronic Arts publish?

Code:

```
Grunt> data_electronic_art = FILTER data BY publisher == 'Electronic Arts';
Grunt> Dump data_electronic_art;
```

Output:



```
cloudera@quickstart:~/Desktop/Kushal
File Edit View Search Terminal Help
grunt> data_electronic_art = FILTER data BY publisher == 'Electronic Arts';
grunt>
```

```
cloudera@quickstart:~/Desktop/Kushal
File Edit View Search Terminal Help
Output(s):
Successfully stored 1 records (6 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp100794921/tmp-1953637744"

Counters:
Total records written : 1
Total bytes written : 6
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1631167871539_0012

2021-09-09 00:41:16,996 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 537 time(s).
2021-09-09 00:41:16,996 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-09-09 00:41:16,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-09 00:41:16,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-09 00:41:16,997 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schemaTuple] was not set... will not generate code.
2021-09-09 00:41:17,004 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-09-09 00:41:17,004 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(8)
arunt>
```

Conclusion:

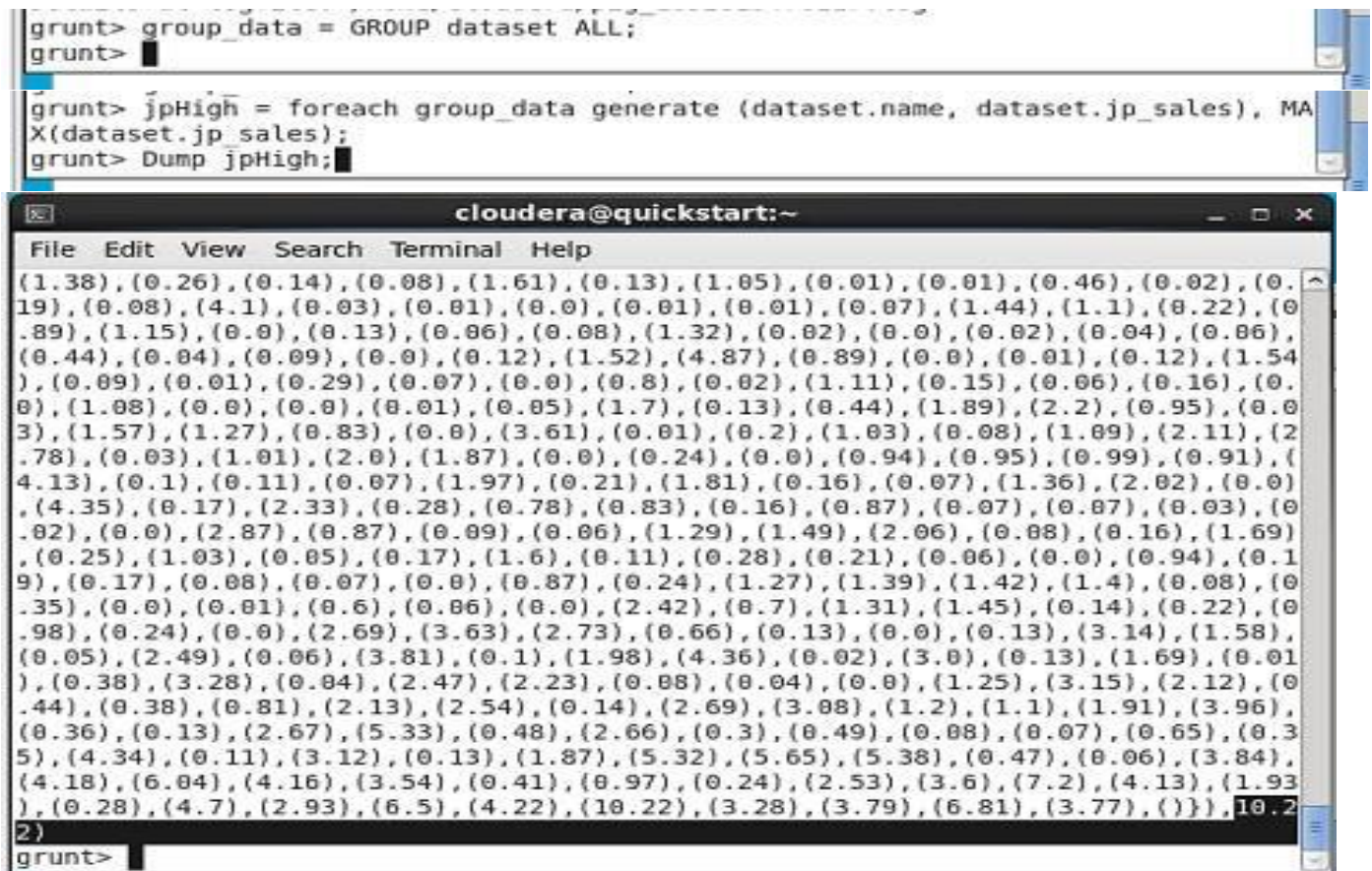
Firstly, used filter operator and filtered data by publisher by electronic arts and dumped it and got the result.

d. Which game reported the highest sales in Japan ever?

Code:

```
grunt> group_data = GROUP dataset ALL;
grunt> jpHigh = foreach group_data generate (dataset.name, dataset.jp_sales),
MAX(dataset.jp_sales);
grunt> Dump jpHigh;
```

Output:



```
grunt> group_data = GROUP dataset ALL;
grunt>
grunt> jpHigh = foreach group_data generate (dataset.name, dataset.jp_sales), MA
X(dataset.jp_sales);
grunt> Dump jpHigh;
```

```
cloudera@quickstart:~
File Edit View Search Terminal Help
(1.38), (0.26), (0.14), (0.08), (1.61), (0.13), (1.05), (0.01), (0.01), (0.46), (0.02), (0.
19), (0.08), (4.1), (0.03), (0.01), (0.0), (0.01), (0.01), (0.07), (1.44), (1.1), (0.22), (0
.89), (1.15), (0.0), (0.13), (0.06), (0.08), (1.32), (0.02), (0.0), (0.02), (0.04), (0.06),
(0.44), (0.04), (0.09), (0.0), (0.12), (1.52), (4.87), (0.89), (0.0), (0.01), (0.12), (1.54
), (0.09), (0.01), (0.29), (0.07), (0.0), (0.8), (0.02), (1.11), (0.15), (0.06), (0.16), (0.
0), (1.08), (0.0), (0.0), (0.01), (0.05), (1.7), (0.13), (0.44), (1.89), (2.2), (0.95), (0.0
3), (1.57), (1.27), (0.83), (0.0), (3.61), (0.01), (0.2), (1.03), (0.08), (1.09), (2.11), (2
.78), (0.03), (1.01), (2.0), (1.87), (0.0), (0.24), (0.0), (0.94), (0.95), (0.99), (0.91), (
4.13), (0.1), (0.11), (0.07), (1.97), (0.21), (1.81), (0.16), (0.07), (1.36), (2.02), (0.0)
, (4.35), (0.17), (2.33), (0.28), (0.78), (0.83), (0.16), (0.87), (0.07), (0.07), (0.03), (0
.02), (0.0), (2.87), (0.87), (0.09), (0.06), (1.29), (1.49), (2.06), (0.08), (0.16), (1.69)
, (0.25), (1.03), (0.05), (0.17), (1.6), (0.11), (0.28), (0.21), (0.06), (0.0), (0.94), (0.1
9), (0.17), (0.08), (0.07), (0.0), (0.87), (0.24), (1.27), (1.39), (1.42), (1.4), (0.08), (0
.35), (0.0), (0.01), (0.6), (0.06), (0.0), (2.42), (0.7), (1.31), (1.45), (0.14), (0.22), (0
.98), (0.24), (0.0), (2.69), (3.63), (2.73), (0.66), (0.13), (0.0), (0.13), (3.14), (1.58),
(0.05), (2.49), (0.06), (3.81), (0.1), (1.98), (4.36), (0.02), (3.0), (0.13), (1.69), (0.01
), (0.38), (3.28), (0.04), (2.47), (2.23), (0.08), (0.04), (0.0), (1.25), (3.15), (2.12), (0
.44), (0.38), (0.81), (2.13), (2.54), (0.14), (2.69), (3.08), (1.2), (1.1), (1.91), (3.96),
(0.36), (0.13), (2.67), (5.33), (0.48), (2.66), (0.3), (0.49), (0.08), (0.07), (0.65), (0.3
5), (4.34), (0.11), (3.12), (0.13), (1.87), (5.32), (5.65), (5.38), (0.47), (0.06), (3.84),
(4.18), (6.04), (4.16), (3.54), (0.41), (0.97), (0.24), (2.53), (3.6), (7.2), (4.13), (1.93
), (0.28), (4.7), (2.93), (6.5), (4.22), (10.22), (3.28), (3.79), (6.81), (3.77), ()), 10.2
2)
grunt>
```

Conclusion:

Firstly, we have grouped the data then using foreach grouped data, generated highest sales in Japan

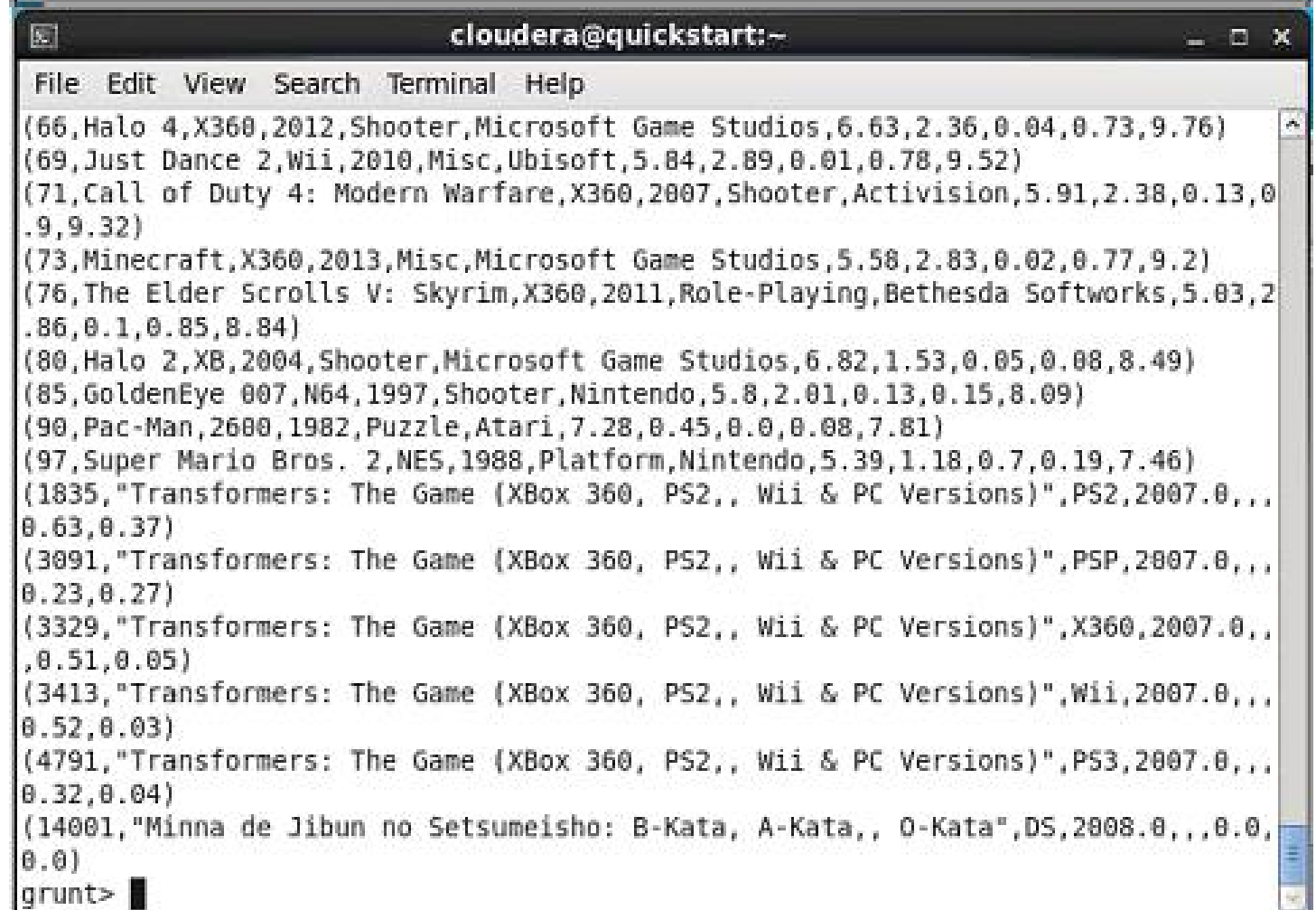
e. Which games in North America report a sale of over 5 million USD?

Code:

```
grunt> out = filter dataset by na_sales > 5.00;
grunt> Dump out;
```

Output:

```
grunt> out = filter dataset by na_sales > 5.00;
2021-09-28 01:04:44,973 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-28 01:04:44,973 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-28 01:04:45,034 [main] WARN org.apache.pig.PigServer - Encountered Warn
ing IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> Dump out;
```



```
cloudera@quickstart:~
File Edit View Search Terminal Help
(66,Halo 4,X360,2012,Shooter,Microsoft Game Studios,6.63,2.36,0.04,0.73,9.76)
(69,Just Dance 2,Wii,2010,Misc,Ubisoft,5.84,2.89,0.01,0.78,9.52)
(71,Call of Duty 4: Modern Warfare,X360,2007,Shooter,Activision,5.91,2.38,0.13,0
.9,9.32)
(73,Minecraft,X360,2013,Misc,Microsoft Game Studios,5.58,2.83,0.02,0.77,9.2)
(76,The Elder Scrolls V: Skyrim,X360,2011,Role-Playing,Bethesda Softworks,5.83,2
.86,0.1,0.85,8.84)
(80,Halo 2,XB,2004,Shooter,Microsoft Game Studios,6.82,1.53,0.05,0.08,8.49)
(85,GoldenEye 007,N64,1997,Shooter,Nintendo,5.8,2.01,0.13,0.15,8.09)
(90,Pac-Man,2600,1982,Puzzle,Atari,7.28,0.45,0.0,0.08,7.81)
(97,Super Mario Bros. 2,NES,1988,Platform,Nintendo,5.39,1.18,0.7,0.19,7.46)
(1835,"Transformers: The Game (XBox 360, PS2,, Wii & PC Versions)",PS2,2007.0,,
0.63,0.37)
(3091,"Transformers: The Game (XBox 360, PS2,, Wii & PC Versions)",PSP,2007.0,,
0.23,0.27)
(3329,"Transformers: The Game (XBox 360, PS2,, Wii & PC Versions)",X360,2007.0,,
0.51,0.05)
(3413,"Transformers: The Game (XBox 360, PS2,, Wii & PC Versions)",Wii,2007.0,,
0.52,0.03)
(4791,"Transformers: The Game (XBox 360, PS2,, Wii & PC Versions)",PS3,2007.0,,
0.32,0.04)
(14001,"Minna de Jibun no Setsumeisho: B-Kata, A-Kata,, O-Kata",DS,2008.0,,0.0,
0.0)
grunt>
```

Conclusion:

By using filter operator we have filtered the data which has sale over 5 million USD in NorthAmerica.

f. Which game reported the highest sale globally and in what year?

Code:

```
grunt> order_data = order dataset by global_sales desc;
grunt> Dump order_data;
```

Output:



```
grunt> order_data = order dataset by global_sales desc;
2021-09-28 01:08:18,833 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> Dump order_data;
```

```
cloudera@quickstart:~
File Edit View Search Terminal Help
(16063,"Pachi-Slot Teiou: Golgo 13 Las Vegas (JP sales, but wrong system)",,2002
,Misc,,0.0,0.0,0.01,0.0)
(16084,Pro Angler Moves,PS3,2012,Sports,"Interworks Unlimited,,0.01,0.0,0.0,0.0)
(16089,"Iza, Shutsujin! Koisen",,2011,Adventure,,0.0,0.0,0.01,0.0)
(16151,Hot Wheels: Stunt Track Challenge / World Race,GBA,2006,Racing,"Destinati
on Software,,0.01,0.0,0.0,0.0)
(16265,"Hanayaka Kana, Ware ga Ichizoku",,2010,Adventure,,0.0,0.0,0.01,0.0)
(16375,"Crouching Tiger, Hidden Dragon",,2003,Action,,0.01,0.0,0.0,0.0)
(16412,"Element Girl: Love, Fashion and Friends",,2008,Adventure,,0.0,0.01,0.0,0
.0)
(16413,"Sakigake!! Otokojuku - Nihon yo, Kore ga Otoko Dearu!",,2014,Fighting,,0
.0,0.0,0.01,0.0)
(16421,"Tennis no Oji-Sama: Doubles no Oji-Sama - Boys, Be Glorious!",,2009,Spor
ts,,0.0,0.0,0.01,0.0)
(16448,Wade Hixton's Counter Punch,GBA,2004,Sports,"Destination Software,,0.01,0
.0,0.0,0.0)
(16463,"Horse Life 4: My Horse, My Friend,,3DS,2015,,0.0,0.01,0.0)
(16480,"Shinobi, Koi Utsutsu: Setsugetsuka Koi Emaki",,2015,Action,,0.0,0.0,0.01
,0.0)
(16504,"Transformers: War for Cybertron (XBox 360, PS3,,PC,2010,,,0.01,0.0,0.0)
(16567,Original Frisbee Disc Sports: Ultimate & Golf,DS,2007,Action,"Destination
Software,,0.01,0.0,0.0,0.0)
(,Name,Platform,,Genre,Publisher,,,,)
grunt>
```

Conclusion:

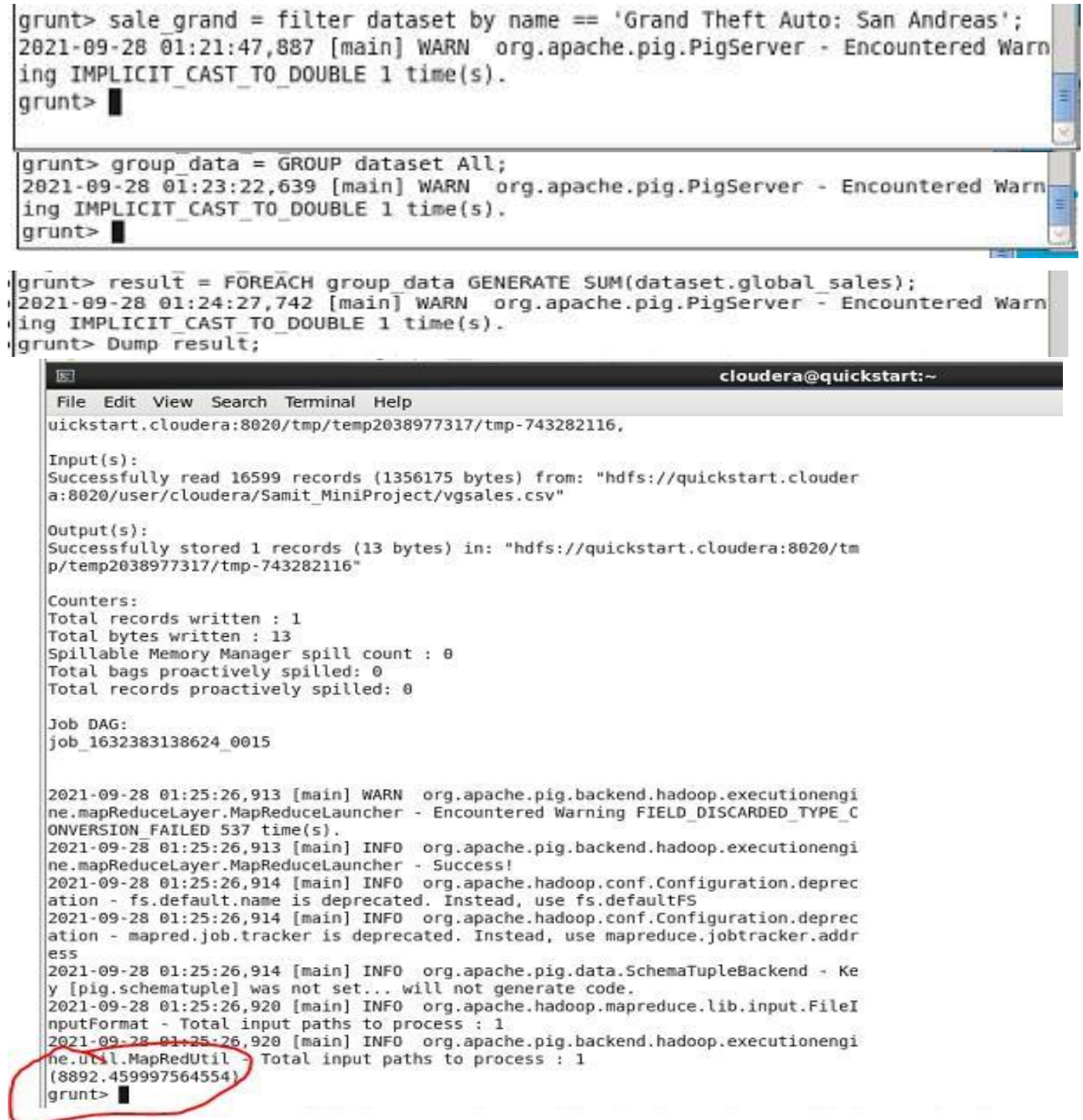
By using order and desc operator I have displayed highest sale globally.

g. How much sale did “Grand Theft Auto: San Andreas” report in the rest of the world?

Code:

```
grunt> sale_grand = filter dataset by name == 'Grand Theft Auto: San
Andreas';
grunt> group_data = GROUP dataset All;
grunt> result = FOREACH group_data GENERATE SUM(dataset.global_sales);
grunt> Dump result;
```

Output:



```
grunt> sale_grand = filter dataset by name == 'Grand Theft Auto: San
Andreas';
2021-09-28 01:21:47,887 [main] WARN org.apache.pig.PigServer - Encountered Warn
ing IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt>

grunt> group_data = GROUP dataset All;
2021-09-28 01:23:22,639 [main] WARN org.apache.pig.PigServer - Encountered Warn
ing IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt>

grunt> result = FOREACH group_data GENERATE SUM(dataset.global_sales);
2021-09-28 01:24:27,742 [main] WARN org.apache.pig.PigServer - Encountered Warn
ing IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> Dump result;
```

```
cloudera@quickstart:~
File Edit View Search Terminal Help
quickstart.cloudera:8020/tmp/temp2038977317/tmp-743282116,

Input(s):
Successfully read 16599 records (1356175 bytes) from: "hdfs://quickstart.clouder
a:8020/user/cloudera/Samit_MiniProject/vgsales.csv"

Output(s):
Successfully stored 1 records (13 bytes) in: "hdfs://quickstart.cloudera:8020/tm
p/temp2038977317/tmp-743282116"

Counters:
Total records written : 1
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1632383138624_0015

2021-09-28 01:25:26,913 [main] WARN org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION FAILED 537 time(s).
2021-09-28 01:25:26,913 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-28 01:25:26,914 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-28 01:25:26,914 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-28 01:25:26,914 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-28 01:25:26,920 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-28 01:25:26,920 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(8892.459997564554)
grunt>
```

Conclusion:

Firstly, we have filtered the data by name “Grand Theft Auto: San Andreas” then grouped all dataset then using foreach operator I have calculated the sale generated by “Grand Theft Auto:San Andreas”.

About Pig

- Pig is a high level platform/ tool which is used to process the large set of data.
- Pig uses it's own language called Pig Latin.
- Pig Latin is a data flow language, not declarative unlike SQL. Hence, we can easily follow the commands.
- Pig is very fast that is nearly about 36% faster than hive for join operations on datasets.
- Pig is 46% faster than hive for arithmetic operations
- Pig is 10% faster than hive for filtering about 10% of the data.
- Pig is very easy to write and read.
- It provides data operations like ordering, filters and joins and we can perform very easily.
- It needs less development time.
- Pig helped a lot which helps us to write complex data transformations without knowing java.
- As it is data flow language, it is very easy to work with. Just we need the commands.
- That's why it is very helpful in my project to work with Pig.
- Pig helps you save memory for storage in your local or data server storage.
- Pig latin is parallel language.
- Data researcher who work large datasets frequently use scripting language. So, pig is better used in Hadoop.
- Pig causes lesser code efficiency.
- Pig requires less development time and effort.