**Business Objective:**

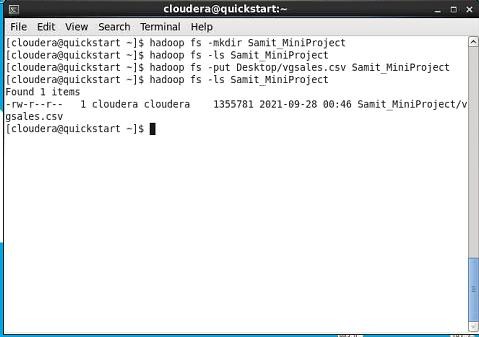
To analyze the psychological benefits and effects of playing Video Games on the data taken from the scrap of [vgchartz.com](https://www.vgchartz.com/)

**Business Questions**

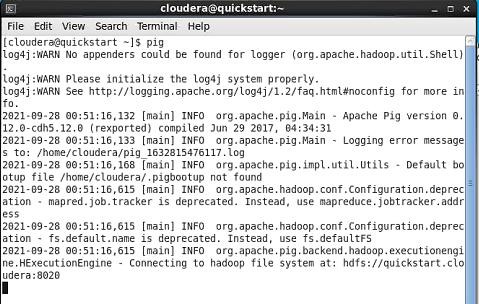
1. Which place did “Call of Duty: Ghosts” gain the maximum sales?
2. What genre of games in the most popular?
3. Out of the entire data given to you, how many percent games does Electronic Arts publish?
4. Which game reported the highest sales in Japan ever?
5. Which games in North America report a sale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

**Solutions:**

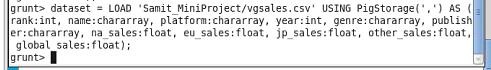
**Creating directory & putting vgsales.csv data into the directory;**



**Entering to grunt shell by the command; pig**



**Loading the Data,**

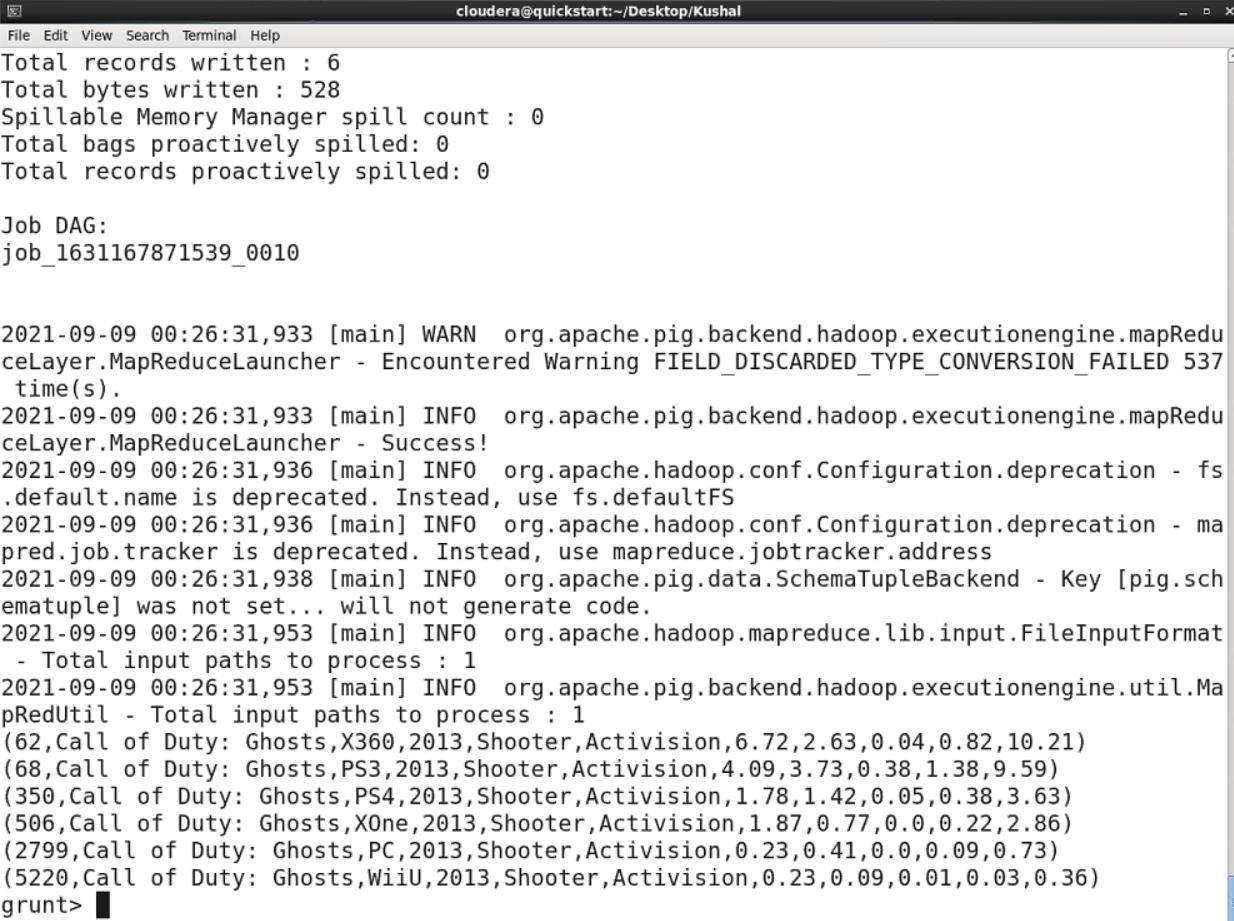
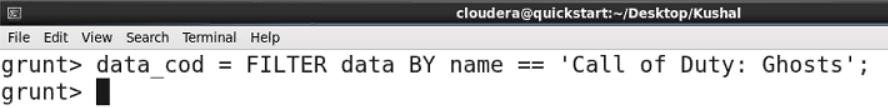


1. **Which place did “Call of Duty: Ghosts” gain the maximum sales?**

**Code:**

Grunt> data\_cod = FILTER data BY name == ‘Call of Duty: Ghosts’; Grunt> Dump data\_cod;

**Output:**



**Conclusion:**

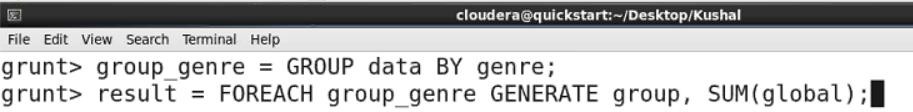
Used filter operator found the maximum sales from “Call of Duty: Ghosts”

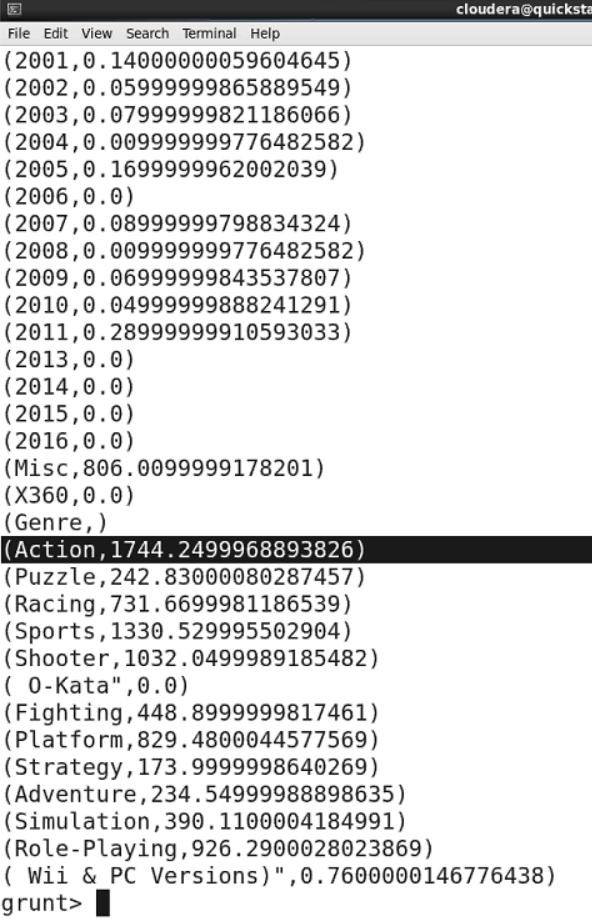
1. **What genre of games in the most popular?**

**Code:**

grunt> group\_genre = GROUP data BY genre;

grunt> result = FOREACH group\_genre GENERATE group, SUM(global); grunt> Dump result;

**Output:**



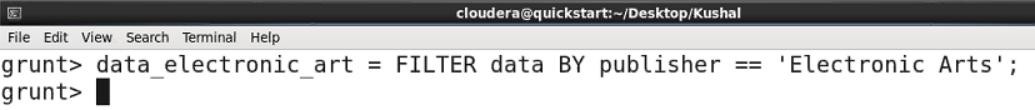
**Conclusion:**

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

1. **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:**



**Conclusion:**

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

1. **Which game reported the highest sales in Japan ever?**

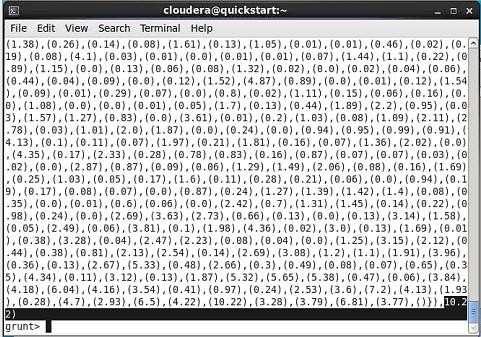
**Code:**

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:**



**Conclusion:**

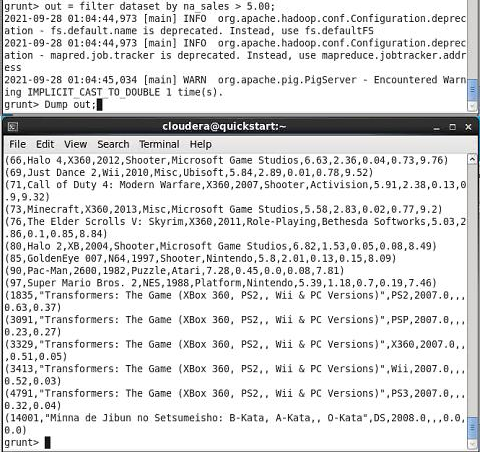
Firstly, we have grouped the data then using foreach grouped data, generated highest sales in Japan

1. **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:**



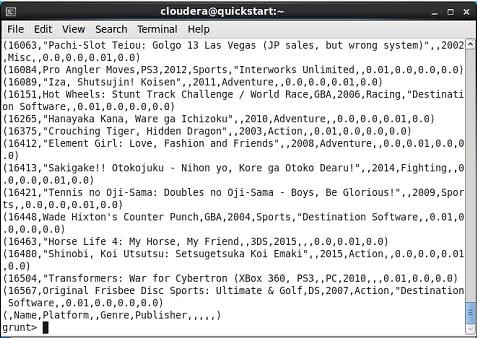
**Conclusion:**

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

1. **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:**



**Conclusion:**

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

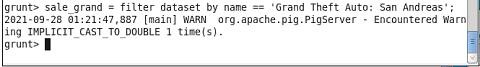
1. **How much sale did “Grand Theft Auto: San Andreas” report in the rest of the world?**

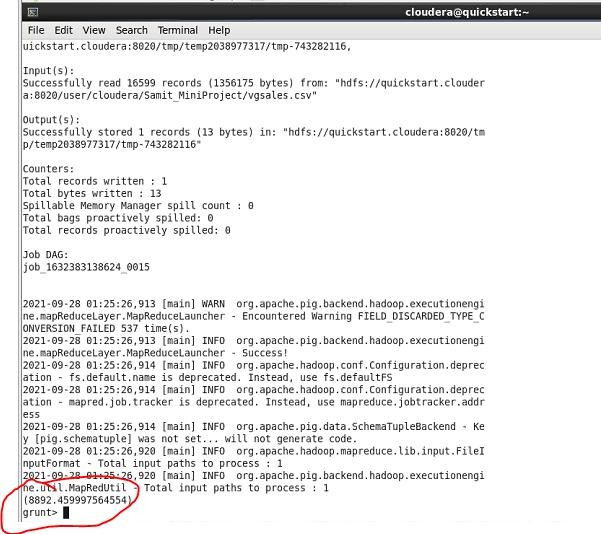
**Code:**

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:**



**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 genrated 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 reseacher who work large datasets frequently use scripting language. So, pig is better used in Hadoop.
* Pig causes lesser code efficiency.
* Pig requies less development time and effort.