

REVIEW 2

PDC J

COMPONENT

15BCE0466 Arjun Jayakrishnan

15BCE0437 Parth McPherson

15BCE0774 Rashi Agarwala

Project Topic:

Analysis of the video game dataset 'video_games_sales_as_at_22_dec_2016' which has been obtained from Kaggle.com.

The dataset describes video games sales of different games over the course of the last 20 years. It has more than 16500 entries.

It also describes critics and user ratings.

Visualisation:

1	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sal	Global_Sa	Critic_Sco	Critic_Coc	User_Scor	User_Cou	Developer	Rating
2	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76	51	8	322	Nintendo	E
3	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24						
4	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82	73	8.3	709	Nintendo	E
5	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80	73	8	192	Nintendo	E
6	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37						
7	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26						
8	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.14	6.5	2.88	29.8	89	65	8.5	431	Nintendo	E
9	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58	41	6.6	129	Nintendo	E
10	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.44	6.94	4.7	2.24	28.32	87	80	8.4	594	Nintendo	E
11	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31						
12	Nintendogs	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.67						
13	Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.9	23.21	91	64	8.6	464	Nintendo	E
14	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1						
15	Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.6	2.15	22.7	80	63	7.7	146	Nintendo	E
16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	15	4.89	0.24	1.69	21.81	61	45	6.3	106	Good Science Studio	E
17	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.01	8.49	2.53	1.77	21.79	80	33	7.4	52	Nintendo	E
18	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.02	9.09	0.98	3.96	21.04	97	50	8.2	3994	Rockstar North	M
19	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81	95	80	9	1588	Rockstar North	M
20	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61						
21	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.74	9.2	4.16	2.04	20.15	77	58	7.9	50	Nintendo	E
22	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.38	4.46	6.04	1.36	18.25						
23	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14						
24	Super Mario Bros. 3	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28						
25	Grand Theft Auto V	X360	2013	Action	Take-Two Interactive	9.66	5.14	0.06	1.41	16.27	97	58	8.1	3711	Rockstar North	M
26	Grand Theft Auto: Vice City	PS2	2002	Action	Take-Two Interactive	8.41	5.49	0.47	1.78	16.15	95	62	8.7	730	Rockstar North	M
27	Pokemon Ruby/Pokemon Sapphire	GBA	2002	Role-Playing	Nintendo	6.06	3.9	5.38	0.5	15.85						
28	Brain Age 2: More Training in Minutes a Day	DS	2005	Puzzle	Nintendo	3.43	5.35	5.32	1.18	15.29	77	37	7.1	19	Nintendo	E
29	Pokemon Black/Pokemon White	DS	2010	Role-Playing	Nintendo	5.51	3.17	5.65	0.8	15.14						
30	Gran Turismo 3: A-Spec	PS2	2001	Racing	Sony Computer Entertainment	6.85	5.09	1.87	1.16	14.98	95	54	8.4	314	Polyphony Digital	E
31	Call of Duty: Modern Warfare 3	X360	2011	Shooter	Activision	9.04	4.24	0.13	1.32	14.73	88	81	3.4	8713	Infinity Ward, Sledge M	E
32	Pokémon Yellow: Special Pikachu Edition	GB	1998	Role-Playing	Nintendo	5.89	5.04	3.12	0.59	14.64						
33	Call of Duty: Black Ops 3	PS4	2015	Shooter	Activision	6.03	5.86	0.36	2.38	14.63						

Project Plan:

We plan to analyse the data and report various findings that could help companies that make games to increase sales and know what consumers like.

To implement our analysis, Hortonworks Sandbox is used to instantiate the Hadoop Cluster. 'Hive 1.0' is then used to run SQL queries on the non-relational database.

We will also visualize the results using Hive and Python.

Example:

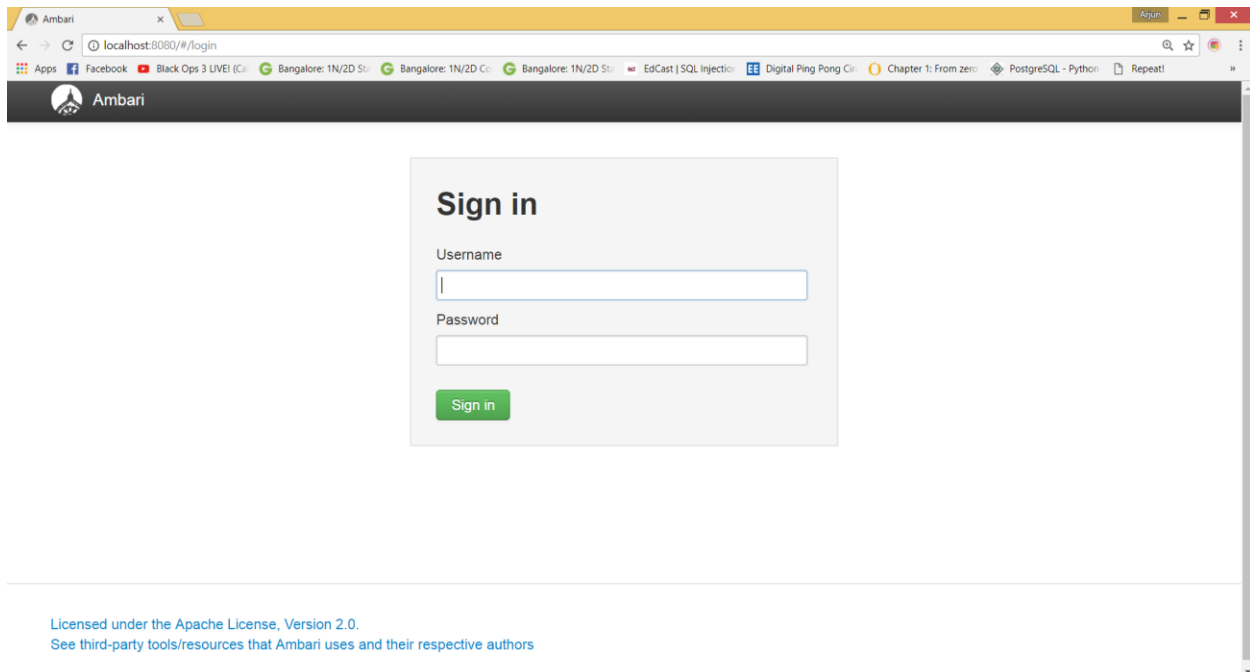
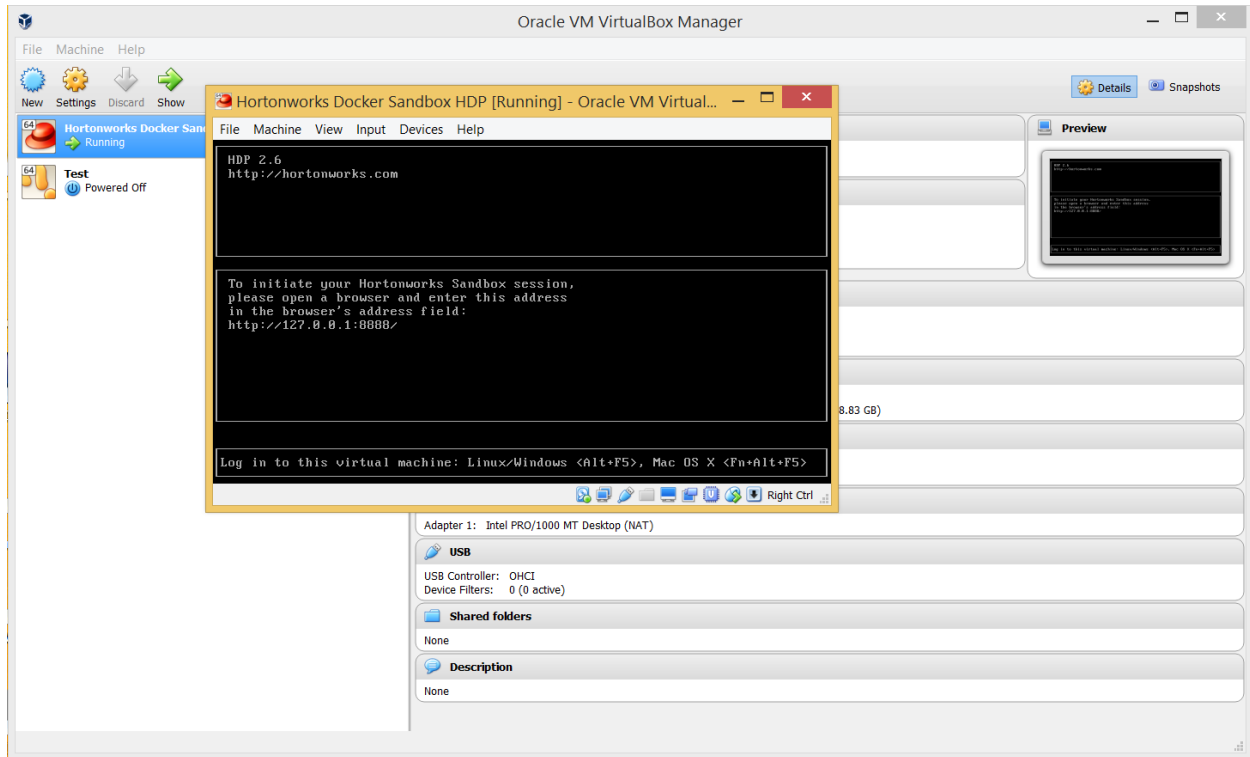
Let us take an example and find out the year when most games were released.

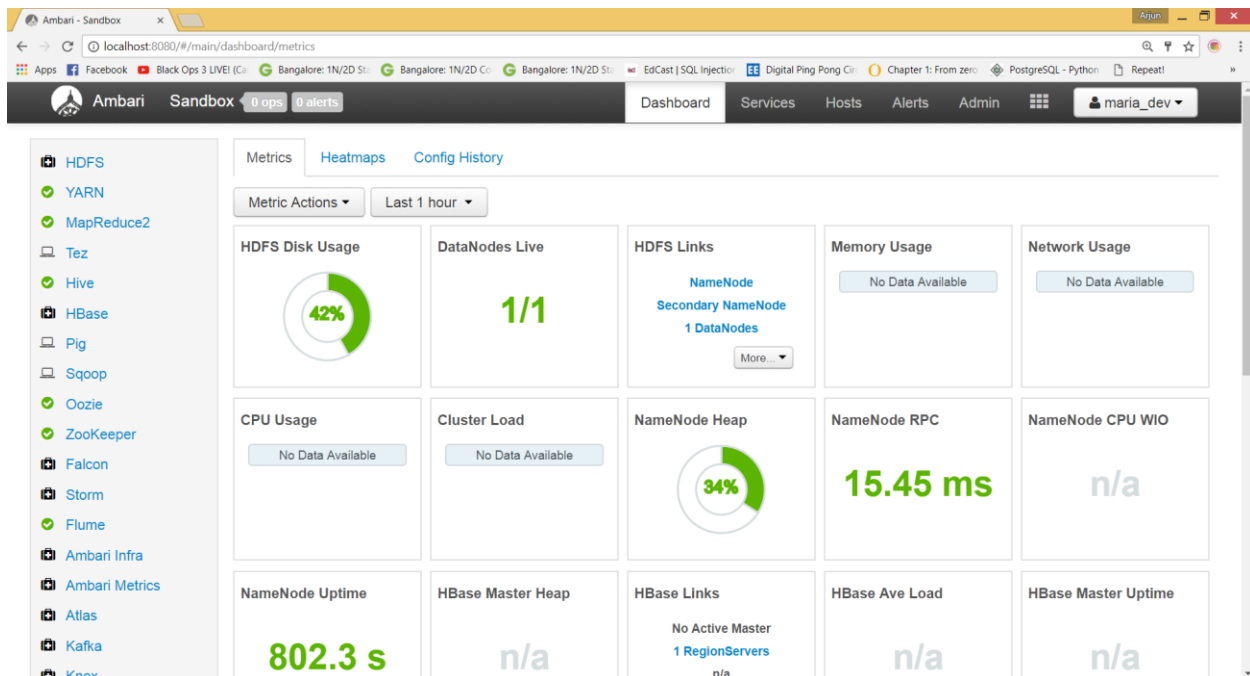
1. Open Sandbox in VirtualBox & open the browser to localhost:8080 to access Ambari.

```

Hortonworks Docker Sandbox HDP [Running] - Oracle VM Virtual...
File Machine View Input Devices Help
[ OK ] Started Authorization Manager.
Starting Network Manager Script Dispatcher Service...
[ OK ] Started Network Manager Script Dispatcher Service.
[ OK ] Started Network Manager Wait Online.
Starting LSB: Bring up/down networking...
[ OK ] Started LSB: Bring up/down networking.
[ OK ] Reached target Network.
Starting Dynamic System Tuning Daemon...
[ OK ] Started OpenSSH server daemon.
Starting OpenSSH server daemon...
Starting Postfix Mail Transport Agent...
Starting /etc/rc.d/rc.local Compatibility...
[ OK ] Reached target Network is Online.
Starting Crash recovery kernel arming...
Starting Docker Application Container Engine...
[ OK ] Started /etc/rc.d/rc.local Compatibility.
[FAILED] Failed to start Crash recovery kernel arming.
See 'systemctl status kdump.service' for details.
[ OK ] Started Postfix Mail Transport Agent.
[ OK ] Started Dynamic System Tuning Daemon.
Starting Network Manager Script Dispatcher Service...
[ OK ] Started Network Manager Script Dispatcher Service.
[ OK ] Started Docker Application Container Engine.
Starting SandboxContainer...

```





- Once the Hadoop Cluster is running, open Hive & upload Dataset. Columns have been renamed for convenience.

The screenshot shows the Hive 'Upload Table' interface. It displays the upload configuration and a preview of the data from a CSV file named 'Video_Games...c_2016.csv'.

Upload Configuration:

- Upload from: Local
- File type: CSV
- Database: default
- Stored as: ORC
- Table name: Video_Games_Sales_as_at_22_Dec_2016

Table Preview:

name_game	platform	year	genre	publisher	america_sales
Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales
Wii Sports	Wii	2006	Sports	Nintendo	41.36
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27
Tetris	GB	1989	Puzzle	Nintendo	23.2
New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28
Wii Play	Wii	2006	Misc	Nintendo	13.96
New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.44

3. Now run the sql query

Query:

```
Worksheet
1 SELECT year , count(year) as yearcount
2 FROM video_games_sales_as_at_22_dec_2016
3 GROUP BY year
4 ORDER BY yearcount;
```

Result:

1995	219
1996	263
1997	289
1999	338
2000	350
1998	379
2001	482
2016	502
2013	544
2014	581
2015	606
2012	653
2004	762
2003	775
2002	829
2005	939
2006	1006
2011	1136
2007	1197
2010	1255
2009	1426
2008	1427

Conclusion: Clearly, the most games were released in 2008.

We plan to find more such results by Review 3.