Semestrálna práca

BI-BIG

Autor: Šimon Minárik

Obsah

1. Úvod 3
2. Dáta 3
    1. Zdroj dát 3
    2. Datasety 3
        1. vgsales.csv 3
        2. platform.csv 5
        3. publisher.csv 6
3. Spark 7
    1. Spustenie 7
    2. Načítanie dát 7
    3. Spracovanie dát 8
        1. Agregácia jedného datasetu 8
        2. Agregácia dvoch datasetov 8
            1. Uloženie datasetu na lokálny disk 9
        3. Agregácia agregovaného datasetu 9
4. ElasticSearch + Kibana 10
    1. Spustenie 10
    2. Dotazovanie do indexu 10
        1. Filtrovanie 10
        2. Triedenie 10
        3. Wildcard 11
    3. Vizualizácie 11
        1. Dashboard s pohl'admi na dáta 11
        2. Dashboard s grafmi (+ heatmap) 12
5. Záver 13

1. Úvod

Pre účely semestrálnej práce som sa rozhodol použiť dáta zaoberajúce sa predajom počítačových (a konzolových) hier. Dáta obsahujú informácie o samotných hrách, ako ich žáner, developer, vydavateľ, predaje.., základné informácie o ich vydavateľoch a taktiež obsahujú informácie o platformách, na ktoré boli hry vytvárané.
Účel semestrálnej práce bolo zamerať sa na spracovanie a vizualizáciu dát, pomocou technológií Spark (spracovanie), ElasticSearch a Kibana (vizualizácia). Dáta do indexu v ElasticSearch boli načítané pomocou LogStash.

2. Dáta

Dáta pokrývajú informácie ohľadom predaja hier na rôzne herné platformy, staré aj nové. Informácie o predajoch v iných kútoch sveta (Severná Amerika, Európa, Japonsko a iné) sú vhodné na analýzu možného úspechu novovytváranej hry - s pohľadom na úspech hier z jednotlivých žánrov v týchto častiach sveta.
Rovnako sa dá z dát vyčítať trend predaja hier na rôzne platformy, takisto v súľade s ich žánrom či lokalitou.

1. Zdroj dát

Dáta boli stiahnuté z https://data.world/mhoangvslev/steam-games-dataset, využité boli 3 z 4 dostupných datasetov - *vgsales.csv*, *platform.csv* a *publisher.csv*

2. Datasety

1. vgsales.csv

Dataset obsahujúci dáta o predaji jednotlivých hier.

Stĺpec	Popis obsahu stĺpca	Dátový typ	Integritné obmedzenie
Name	Názov hry	String	Bez obmedzenia
Platform	Názov hernej platformy	String	Bez obmedzenia
Year_of_Release	Rok vydania hry	Integer	V tvare YYYY
Genre	Žáner hry	String	Bez obmedzenia
Publisher	Názov spoločnosti, ktorá hru publikuje	String	Bez obmedzenia
NA_Sales	Počet predaných kusov v Severnej Amerike (v miliónoch)	Float	0.0+
EU_Sales	Počet predaných kusov v Európe (v miliónoch)	Float	0.0+
JP_Sales	Počet predaných kusov v Japonsku (v miliónoch)	Float	0.0+
Other_Sales	Počet predaných kusov v ostatných krajinách (v miliónoch)	Float	0.0+

Stĺpec	Popis obsahu stĺpca	Dátový typ	Integritné obmedzenie
Global_Sales	Počet predaných kusov celosvetovo (v miliónoch)	Float	súčet stĺpcov NA_Sales, EU_Sales, JP_Sales a Other_Sales
Critic_Score	Priemer hodnotenia hry od kritikov v celých číslach (0 - 100, kde 100 je najlepšie možné hodnotenie), získané od metacritic.com.. Kritik - recenzista na (hernom) spravodajskom portáli	Integer	0 - 100
Critic_Count	Počet kritikov, ktorý hru hodnotili	Integer	1+
User_Score	Priemer hodnotenia hry od hráčov v desatinných číslach (s jedným desatinným miestom, 0 - 10, kde 10 je najlepšie možné hodnotenie)	Float	0.0 - 10.0
User_Count	Počet hráčov, ktorí hru ohodnotili	Integer	1+
Developer	Názov spoločnosti, ktorá hru vyvinula	String	Bez obmedzenia
Rating	Parent advisory hodnotenia, ktoré sa odvýjajú od obsahu hry (obsahuje krvavé scény, vulgarizmy…) a od nich sa odvýja minimálny požadovaný vek na hranie hry.	String	jedna z možností [AO, E, E10+, EC, K-A, M, RP, T]

Ukážka dát:

Name	Platform	Year_o	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_	Global_S	Critic_S	Critic_	User_	User_C	Developer	Rating
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76	51	8	322	Nintendo	E
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24						
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82	73	8.3	709	Nintendo	E
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80	73	8	192	Nintendo	E
Pokemon Red/Pokem	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37						
Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26						
New Super Mario Bro	DS	2006	Platform	Nintendo	11.28	9.14	6.5	2.88	29.8	89	65	8.5	431	Nintendo	E
Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58	41	6.6	129	Nintendo	E
New Super Mario Bro	Wii	2009	Platform	Nintendo	14.44	6.94	4.7	2.24	28.32	87	80	8.4	594	Nintendo	E
Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31						
Nintendogs	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.67						
Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.9	23.21	91	64	8.6	464	Nintendo	E
Pokemon Gold/Poker	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1						
Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.6	2.15	22.7	80	63	7.7	146	Nintendo	E

2. platform.csv

Dataset obsahujúci informácie o herných platformách, na ktoré sú vyvíjané hry

Stĺpec	Popis obsahu stĺpca	Dátový typ	Integritné obmedzenie
Initiales	Názov, prípadne skratka názvu hernej plaformy	String	Bez obmedzenia
nom	Plný názov hernej platformy	String	Bez obmedzenia
Manufactureur	Spoločnosť, ktorá hernú platformu vyrába	String	Bez obmedzenia
nb_jeu	Kód platformy	Integer	Bez obmedzenia
sortie_eu	Dátum začiatku predaja platformy na európskom trhu	String	v tvare DD/MM/YYYY
sortie_usa	Dátum začiatku predaja platformy na americkom trhu	String	v tvare DD/MM/YYYY
sortir_jp	Dátum začiatku predaja platformy na japonskom trhu	String	v tvare DD/MM/YYYY
gen	Číslo generácie hernej platformy u jej výrobcu	Integer	1+

Ukážka dát:

Initiales	nom	manufactureur	nb_jeu	sortie_eu	sortie_usa	sortir_jp	gen
NES	Nintendo Entertainment System	Nintendo	715	27/10/1987	18/10/1985	15/07/1983	3
SNES	Super Nintendo Entertainment System	Nintendo	1757	06/06/1992	23/08/1991	21/11/1990	4
N64	Nintendo 64	Nintendo	388	01/03/1997	29/09/1996	23/06/1996	5
GC	Nintendo GameCube	Nintendo	657	03/05/2001	18/11/2001	14/09/2001	6
Wii	Wii	Nintendo	1541	08/12/2006	19/11/2006	02/12/2006	7
WiiU	Wii U	Nintendo	772	30/11/2012	18/11/2012	08/12/2012	8
GB	GameBoy	Nintendo	1055	28/09/1989	31/07/1989	21/04/1989	4
GBC	GameBoy Color	Nintendo	660	23/11/1998	18/11/1998	21/10/1998	5
GBA	GameBoy Advance	Nintendo	1504	22/06/2001	11/06/2001	21/03/2001	6
DS	Nintendo DS	Nintendo	1965	11/03/2005	21/11/2004	02/12/2004	7
3DS	Nindentendo 3DS	Nintendo	1330	25/03/2011	27/03/2011	26/02/2011	8

3. publisher.csv

Dataset obsahujúci informácie o spoločnostiach, ktoré hry publikujú

Stĺpec	Popis obsahu stĺpca	Dátový typ	Integritné obmedzenie
Publisher	Názov spoločnosti, ktorá publikuje hry	String	Bez obmedzenia
Headquarters	Lokalita sídla spoločnosti	String	v tvare *mesto*, *štát*, *krajina*, minimálne však *krajina*
Est.	Rok založenia spoločnosti	Integer	v tvare YYYY
Notable games published	Zoznam známych hier, ktoré spoločnosť publikovala	String	Bez obmedzenia
Notes	Rôzne poznámky k spoločnosti, ktoré neboli spomenuté v ostatných stĺpcoch	String	Bez obmedzenia
Active	Či spoločnosť stále funguje alebo nie.. 1 - je aktívna, 0 - nie je aktívna	Integer	0 alebo 1

Ukážka dát:

Publisher	Headquarters	Est.	Notable games published	Notes	Active
07th Expansion	Japan	2002	Higurashi When They Cry Umi		1
11 bit studios	Warsaw , Poland	2010	Frostpunk This War of Mine M	Also a video game develop	1
1C Company	Moscow, Russia	1991	Il-2 seriesMen of War series	Specializes in localization	1
2K Games	Novato, California, United States	2005		Also a video game develop	1
3D Realms	Garland, Texas, United States	1987	Duke Nukem series	Also a video game develop	1
The 3DO Company	Redwood City, California, United States	1991	Army Men	Also a video game and vid	0
505 Games	Milan, Italy	2006		Publishing division of Digit	1
5pb.	Tokyo, Japan	2005	Memories Off	Label of KID until 2006, an	1
7th Level	Dallas, Texas, United States	1993	Monty Python's Complete Was	Defunct in 1998.	0
A&F Software	Rochdale, UK	1981	Chuckie Egg	Merged with MC Lothlorie	0
Aackosoft	Zoeterwoude, Netherlands	1983	Various MSX games	Defunct in 1988.	0
Aardvark Software	UK	1983	Frak!	Defunct in 1989.	0
ABC Software	Buchs, St. Gallen, Switzerland	1991		Acquired by Electronic Art	0
Absolute Entertainment	Upper Saddle River, New Jersey, United States	1986	A Boy and His Blob: Trouble o	Defunct in 1995.	0
Access Software	Salt Lake City, Utah, United States	1982	Tex Murphy	Acquired by Microsoft in 1	0

3. Spark

1. Spustenie

Pre načítanie a spracovanie agregácií jednotlivých datasetov, je potrebné spustiť Zeppelin notebook, v ktorom sa vyhodnocujú Spark príkazy.
V prípade, že sa na lokálnom počítaci nenachádza docker image Sparku a Zeppelinu, stiahne sa a "zbuildí" sa pomocou príkazu:

docker image pull babubabu/spark-zeppelin-docker:v1 docker build -t babubabu/spark-zeppelin-docker .

Následne stačí Zeppelin spustiť príkazom:

docker run -it -v **/Users/mnrk/Documents/Skola/5.\ Semester/BI-BIG/Semestral\ Project/Datasets**:/work -v **/Users/mnrk/Documents/Skola/5.\ Semester/BI-BIG/Semestral\ Project/Spark/notebook**:/usr/zeppelin/zeppelin-0.8.1-bin-all/notebook/ -p 18080:18080 -p 8088:8080 -d babubabu/spark-zeppelin-docker:v1

Hrubo vyznačené sú absolútne cesty ku zložkám **Datasets** (obsahuje csv súbory) a **notebook** (obsahuje Zeppelin notebook)
Hash, ktorý sa v konzole vypíše sa skopíruje a pomocou príkazu docker attach <HASH> sa notebook spustí, prístupný bude na adrese localhost:8088

2. Načítanie dát

Datasety sa načítajú v Zeppelin notebooku pomocou Spark funkcie .read a načítané datasety budú uložené v premenných publishers, platforms a sales

```
val publishers = spark.read.format("csv")
  .option("sep", ",")
  .option("inferSchema", "true")
  .option("header", "true")
  .load("publisher.csv")

val platforms = spark.read.format("csv")
  .option("sep", ",")
  .option("inferSchema", "true")
  .option("header", "true")
  .load("platform.csv")

val sales = spark.read.format("csv")
  .option("sep", ",")
  .option("inferSchema", "true")
  .option("header", "true")
  .load("vgsales.csv")
```

3. Spracovanie dát

1. Agregácia jedného datasetu

Príkaz nižšie vytvorí nový dataset, ktorý agreguje 1 dataset, publishers, podľa roku založenia spoločnosti a k rokom pridáva počet firiem založených v tom roku

```
// Agregace 1 datasetu
val result1 = publishers.groupBy("`Est.`").count().sort($"count".desc)
result1.show()
```

```
+----+-----+
|Est.|count|
+----+-----+
|1982|   55|
|1983|   40|
|1987|   37|
|1981|   34|
|1984|   32|
|1988|   29|
|1990|   28|
```

2. Agregácia dvoch datasetov

Príkazy nižšie vytvoria nový dataset, ktorý agreguje 2 datsety, sales a platforms. Ukazuje úspešnosť predaja hier vydaných na platformách, ktoré vstúpili na európsky trh pred rokom 2000
Na dosiahnutie tohoto výsledku je potrebné najprv nájsť tie platformy (v datasete platforms), ktoré boli vyvinuté pred rokom 2000, pomocou filteru
Následne sa tieto platformy spoja s datasetom predajov (sales), pomocou funkcie join na skratke názvu platformy (kľúč pre inner join), agregujú sa podľa názvu platformy a roku vydania hier a súčtu celosvetového predaja jednotlivých hier.
Pre krajšie vyzerajúci dataset sa pomocou funkcie *as* zmení názov stĺpca sumy na “Platform_Global_Sales” a datset sa zoradí podľa stĺpca Year_of_Release, vzostupne.

```
//Agregace 2 datasetov
val tmp = platforms.filter(substring($"sortie_eu",7,10) < 2000).select("Initiales")
val result2 = sales.join(tmp, $"Platform" === $"Initiales", "inner")
                   .groupBy("Platform", "Year_of_Release")
                   .agg(round(sum("Global_Sales"), 2)
                   .as("Platform_Global_Sales"))
                   .sort("Year_of_Release")
result2.show()
```

```
+--------+---------------+--------------------+
|Platform|Year_of_Release|Platform_Global_Sales|
+--------+---------------+--------------------+
|    2600|           1980|               11.38|
|    2600|           1981|               35.77|
|    2600|           1982|               28.86|
|     NES|           1983|               10.96|
|    2600|           1983|                5.83|
|     NES|           1984|               50.09|
|    2600|           1984|                0.27|
|    2600|           1985|                0.45|
|     NES|           1985|               53.44|
|    2600|           1986|                0.66|
```

1. Uloženie datasetu na lokálny disk

Vzniknutý dataset som uložil lokálne na disk, pre neskoršiu prácu v Kibane, pomocou príkazu:

```
//Save dataframe
result2.coalesce(1).write.format("csv").option("sep", ",").option("header", "true").save
```

FINIS

3. Agregácia agregovaného datasetu

Príkazy nižšie vytvoria nový dataset, ktorý agreguje dáta z predošlej agregácie a z datasetu sales. Ukazuje, koľko “veľkých” hier (predaj hry je minimálne 10% z globálneho predaja hier na platforme za určitý rok) sa vydalo na jednotlivé platformy vydané na európskom trhu pred rokom 2000.
Na začiatok bolo potrebné predošlý dataset spojiť s datasetom sales, pomocou funkcie inner join na kľúčoch názvov platformy (stĺpec Platform) a roku vydania jednotlivých hier (stĺpec Year_of_Release). Funkcie drop boli použité na vyhodenie duplikátnych stĺpcov, ktoré vznikli po spojení 2 datasetov.
Následne sa z datasetu vyfiltrovali “veľké hry” a dáta sa agregovali na názve platformy a roku vydania hier + počet veľkých hier, ktoré v ten rok boli vydané (stĺpec Number_of_Big_Games)

```
//Agregace agregace a datasetu
val tmp2 = result2.join(sales, result2("Platform") === sales("Platform") && result2("Year_of_Release")
    === sales("Year_of_Release"), "inner")
                .drop(result2("Year_of_Release"))
                .drop(result2("Platform"))
val result3 = tmp2.filter($"Global_Sales" * 10 >= $"Platform_Global_Sales")
                .groupBy("Platform", "Year_of_Release")
                .count()
                .withColumnRenamed("count", "Number_of_Big_Games")
                .sort("Year_of_Release")
result3.show()
```

```
+--------+---------------+-------------------+
|Platform|Year_of_Release|Number_of_Big_Games|
+--------+---------------+-------------------+
|    2600|           1980|                  3|
|    2600|           1981|                  1|
|    2600|           1982|                  1|
|     NES|           1983|                  6|
|    2600|           1983|                  3|
|    2600|           1984|                  1|
|     NES|           1984|                  1|
|     NES|           1985|                  1|
|    2600|           1985|                  1|
|     NES|           1986|                  1|
```

4. ElasticSearch + Kibana

1. Spustenie

Na nahranie dát do indexu ElasticSearch som použil LogStash pipeline, ktorej konfiguračný súbor je možné vidieť v zložke *ElasticSearch+Kibana/logstash/pipeline*
Do indexu som si zvolil nahranie datasetu Sales (*vgsales.csv*) a novo vygenerovaného datasetu v Sparku (agregácia dvoch datasetov)
Po úspešnom vytvorení LogStash pipeline a pripravení všetkých konfiguračných súborov stačí spustiť docker-compose (v root zložke projektu):

```
docker-compose up -d
```

Kibana bude po chvíli dostupná na adrese localhost:5601
Pre prácu s dátami je potrebné index načítať v Management - názov indexu je “steam”, nepoužíva žiadny Time Filter

2. Dotazovanie do indexu

Do indexu pomocou ElasticSearch som sa dotazoval cez Dev Tools záložku v Kibane

1. Filtrovanie

```
GET steam/_search
{
  "query": {
    "match": {
      "Genre": "Sports"
    }
  }
}
```

Dotaz vypíše všetky hry žánru “Sports”

2. Triedenie

Dotaz vypíše platformy vydané na európskom trhu pred rokom 2000 s celosvetovým predajom hier vyšším ako 50 (miliónov) a zoradí ich zostupne

```
GET steam/_search
{
  "sort":
    {
      "Platform_Global_Sales": {
        "order": "desc"
      }
    },
  "query": {
    "range": {
      "Platform_Global_Sales": {
        "gte": 50
      }
    }
  }
}
```

3. Wildcard

Dotaz vypíše všetky hry, ktoré majú v názve slovo “the”, použitím wildcard vyhľadávania

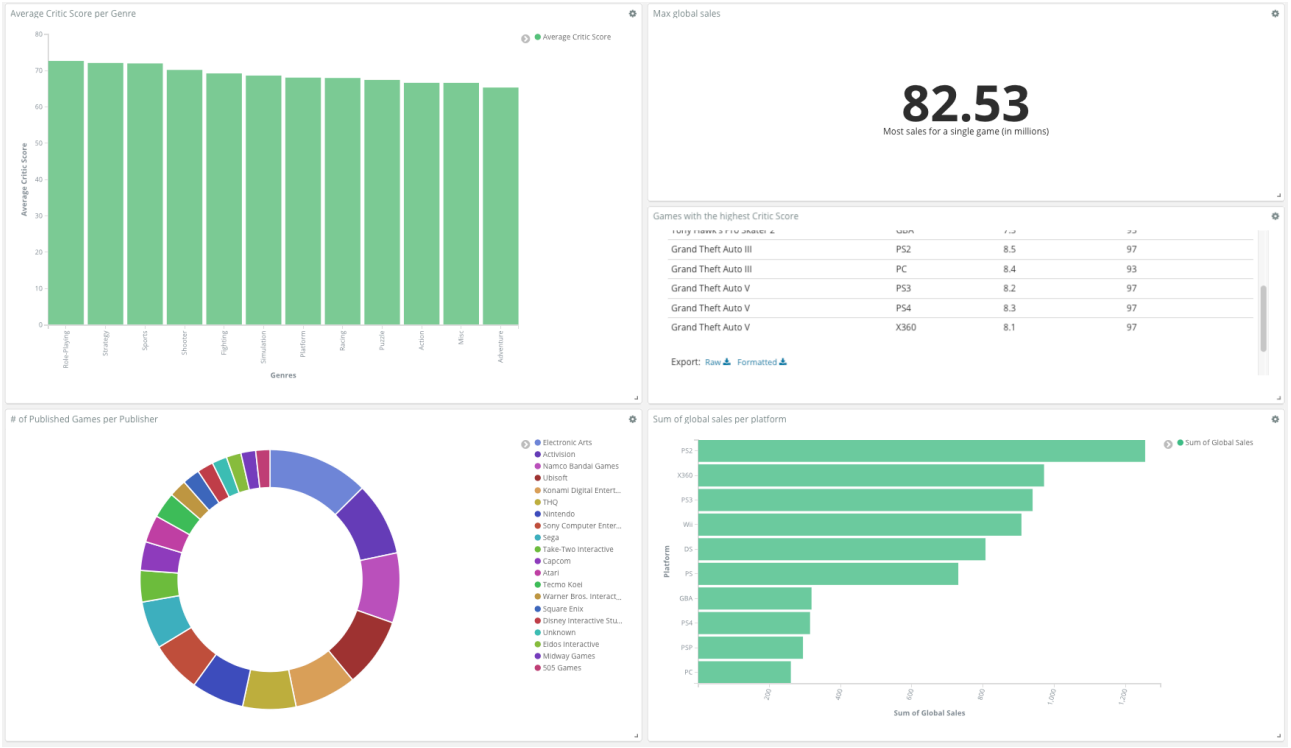
```
GET steam/_search
{
  "query": {
    "wildcard": {
      "Name": {
        "value": “*the*"
      }
    }
  }
}
```

3. Vizualizácie

Vizualizácie nad indexom boli vytvorené pomocou Kibany a následne pridané do Dashboardov, pre prehľadnejšiu navigáciu medzi jednotlivými grafmi.
Hotové vizualizácie a dashboardy spomínané nižšie je možné importovať cez Management - Saved Objects - Import - kibana.json

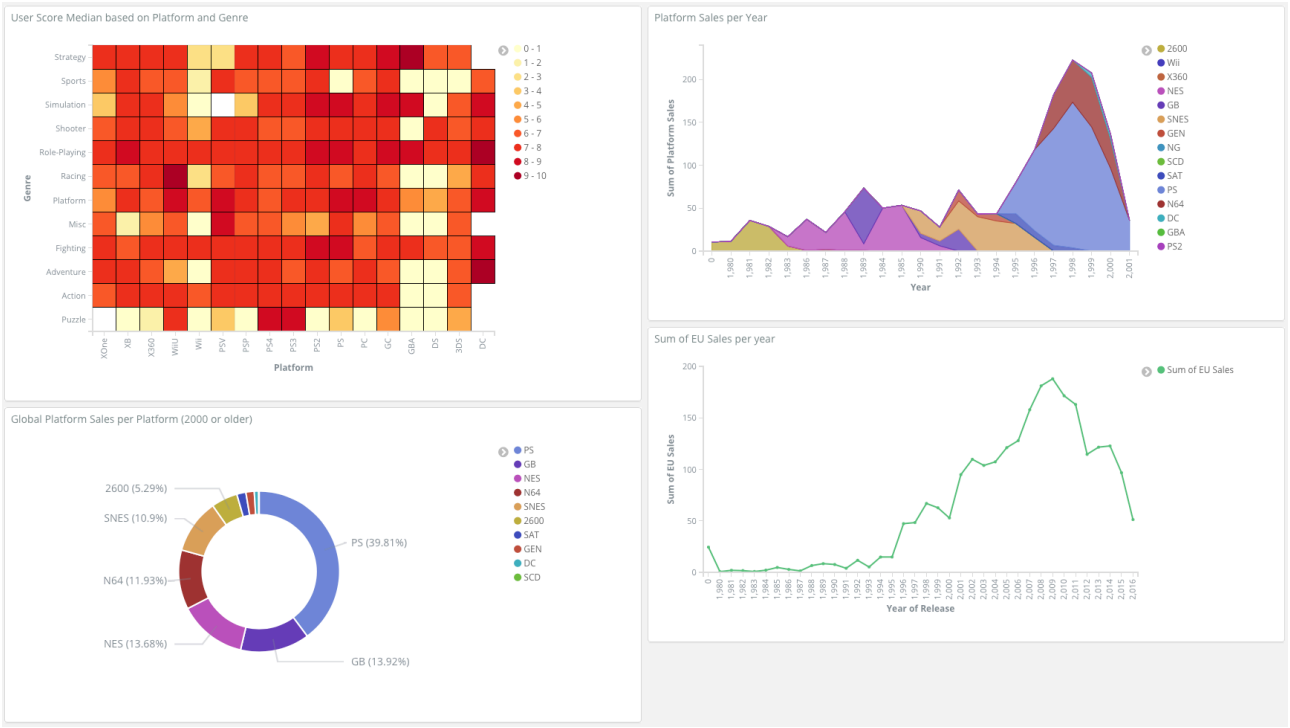
1. Dashboard s pohľadmi na dáta

- Vertikálny graf ukazuje priemerné hodnotenia hier od kritikov podľa žánrov hier
- Koláčový graf ukazuje percentuálne zastúpenie vydaných hier od jednotlivých publisherov
- Horizontálny graf ukazuje celkové predaje hier vytvorených na jednotlivé herné platformy
- Číslo predstavuje najvyšší predaj samostatnej hry (na ľubovoľnú hernú platformu)
- Tabuľka ukazuje názvy hier, platformy, na ktoré boli vyrobené a uživateľské hodnotenie, zoradené podľa hodnotenia kritikov zostupne



2. Dashboard s grafmi (+ heatmap)

- Heat mapa ukazuje medián užívateľského hodnotenia v závislosti na hernú platformu a žáner hry
- Koláčový graf ukazuje rozloženie celkového predaja hier pre herné platformi vydané na európskom trhu pred rokom 2000
- Area graf ukazuje rozloženie celkového predaja hier pre herné platformi podľa rokov vydania hier, pre jeden rok sú zobrazené maximálne 3 platformy (s najväčším predajom)
- Line graf ukazuje predaj hier v Európe v priebehu rokov



5. Záver

Vďaka semestrálnej práci som si mohol na “vlastnej koži” skúsiť prácu s big data technológiami ako ElasticSearch alebo Spark a prinútila ma zamýšľať sa nad dátami, ktoré mám pred sebou. Čo sa samotných dát týka, z Kibbana vizualizácií je vidieť, že nie sú kompletné a neodzrkadlujú 100% realitu, pretože napr. trend predaja hier by, podľa môjho neinformovaného laického pohľadu, určite stúpal, minimálne kvôli väčšej pestrosti herných platform a jednoduchšej možnosti nákupu online.