Урок 5. Интерактивный анализ данных в Apache Zeppelin

- 1. Установите Apache Zeppelin и настройте интеграцию с Apache Spark или Apache Hive
- 2. Скачайте датасет Video Game Sales (https://www.kaggle.com/gregorut/videogamesales)
- 3. Выведите самую продаваемую игру за всё время
- 4. Какая платформа самая популярная в каждом регионе (NA, EU, JP)?
- 5. Какой жанр популярен больше всего в каждом регионе (NA, EU, JP)?
- 6. Выведите самый популярный жанр на каждый год

```
%spark.sql

create table if not exists vgsales
    using csv
    options (
        path "/user/hduser/videogamesales/vgsales.csv",
        header true,
        inferSchema true
    );

select * from vgsales limit 10;
```

=	<u>l.111</u>	¢	<u>~</u>	<u> #</u>	<u>*</u>	•	settings ▼

Rank	Name	Platform ~	Year ~	Genre ≡
1	Wii Sports	Wii	2006	Sports
2	Super Mario Bros.	NES	1985	Platform
3	Mario Kart Wii	Wii	2008	Racing
4	Wii Sports Resort	Wii	2009	Sports
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playir
6	Tetris	GB	1989	Puzzle
7	New Super Mario Bros.	DS	2006	Platform
4				

Самая продаваемая игра за всё время

```
%spark.sql
select *
from vgsales
order by Global_Sales desc
limit 1;
```

 \sim

<u>lılıl</u>

Rank ~	Name ~	Platform ~	Year ~	Genre ≡
1	Wii Sports	Wii	2006	Sports

settings ▼

4

Самая популярная платформа в каждом регионе

```
%spark.sql
with x as (
    select
        Platform,
        sum(NA_Sales) NA_Sales_Sum,
        sum(EU_Sales) EU_Sales_Sum,
        sum(JP_Sales) JP_Sales_Sum
    from vgsales
    group by Platform
    select 'NA' as region, Platform
    from x
    order by NA_Sales_Sum desc
    limit 1
union
    select 'EU' as region, Platform
    order by EU_Sales_Sum desc
    limit 1
union
(
    select 'JP' as region, Platform
    order by JP_Sales_Sum desc
    limit 1
)
```

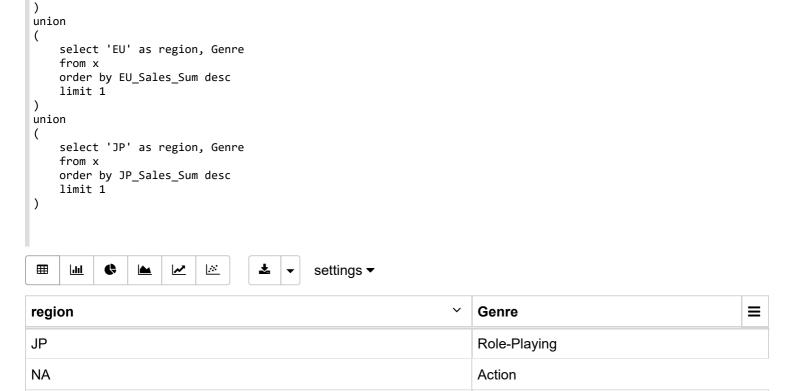


region	Platform =
JP	DS
EU	PS3
NA	X360
4	

Самый популярный жанр в каждом регионе

```
with x as (
    select
        Genre,
        sum(NA_Sales) NA_Sales_Sum,
        sum(EU_Sales) EU_Sales_Sum,
        sum(JP_Sales) JP_Sales_Sum
    from vgsales
    group by Genre
)
(
    select 'NA' as region, Genre
    from x
    order by NA_Sales_Sum desc
    limit 1
```

%spark.sql



Action

Самый популярный жанр на каждый год

EU

```
%spark.sql
with x as (
    select
    Year,
    Genre,
    sum(Global_Sales) Global_Sales,
    row_number() over(partition by Year order by sum(Global_Sales) desc) as row_number
    from vgsales
    where Year != 'N/A'
    group by Year, Genre
    order by Year, Global_Sales desc
)
select Year, Genre
from x
where row_number == 1
```



Year	Genre ≡
1980	Shooter
1981	Action
1982	Puzzle
1983	Platform
1984	Shooter
1985	Platform
1986	Action
1987	Fighting

1988	Platform
1989	Puzzle
1990	Platform
1991	Platform
1992	Fighting
1993	Platform
1994	Platform
1995	Platform
1996	Role-Playing
1997	Racing
1998	Sports
1999	Role-Playing
2000	Sports
2001	Action
2002	Action
2003	Action
2004	Action
2005	Action
2006	Sports
2007	Action
2008	Action
2009	Action
2010	Action
2011	Action
2012	Action
2013	Action
2014	Action
2015	Action
2016	Action
2017	Role-Playing
2020	Simulation
4	