

Analiza popularnosti video igara pomoću podataka o prenosima sa *Twitch*-a

OPIS PROJEKTA IZ PREDMETA SISTEMI BAZA PODATAKA

Opis skupa podataka

KORIŠTENE DATOTEKE

- <https://www.kaggle.com/nikdavis/steam-store-games>

Skup podataka o video igrama prikupljen sa platforme za distribuciju video igara *Steam*. Skup sadrži podatke o 27000 video igara. Korištena je datoteka sa osnovnim podacima o video igrama, koja sadrži 18 atributa.

- <https://clivecast.github.io/>

Skup podataka o prenosima sa *Twitch*-a. *Twitch* je platforma za prenošenje video materijala uživo, sa fokusom na video igre. Skup podataka je sastavljen tako što su u periodu od 1. do 28. februara 2015. godine svakih 5 minuta prikupljani podaci o svim aktivnim prenosima na Twitch-u. Podaci prikupljeni u jednom trenutku posmatranja upisani su u jednu datoteku, a trenutak posmatranja je sadržan u nazivu datoteke. Podaci uključuju trenutni broj gledalaca prenosa, podatke o autoru prenosa, kao i naziv video igre koja se igra na prenosu.

Zbog veličine skupa podataka za potrebe projekta je korišten njegov manji dio. Upotrebljeno je oko 18 miliona zapisa iz perioda od 1. do 12. februara 2015. godine.

Opis skupa podataka

SEMANTIKA KOLONA DATOTEKE O VIDEO IGRAMA

- **appid** – jedinstveni identifikator igre na Steam-u
- **name** – naziv igre
- **release_date** – datum objavljivanja igre
- **english** – da li je igra na engleskom jeziku
- **developer** – kompanije koje su radile na razvoju igre
- **publisher** – izdavač
- **platforms** – platforme na kojima je dostupna igra
- **required_age** – starosno ograničenje prema PEGI UK standardu
- **categories** – kategorije kojima pripada igra, na primjer da li je za jednog ili više igrača
- **genres** – žanrovi kojima pripada igra
- **steamspy_tags** – slično žanrovima, ali izglasano od strane igrača
- **achievements** – broj postignuća u igri
- **positive_ratings** – broj pozitivnih ocena
- **negative_ratings** – broj negativnih ocena
- **average_playtime** – prosečno vreme igranja
- **median_playtime** – medijana vremena igranja
- **owners** – procenjeni broj igrača
- **price** – cena u GBP

Opis skupa podataka

SEMANTIKA KOLONA DATOTEKA O PRENOSIMA

- **stream ID** – jedinstveni identifikator prenosa
- **current views** – broj gledalaca u trenutku prikupljanja podataka
- **stream created time** – kada je počeo prenos
- **broadcaster ID** – jedinstveni identifikator autora prenosa
- **broadcaster name** – ime autora prenosa
- **delay setting** – postavka kašnjenja autora
- **follower number** – broj pratilaca autora
- **partner status** – da li je autor partner na Twitch-u
- **broadcaster language** – jezik autora
- **language** – jezik prenosa
- **total views of broadcaster** – ukupan broj pregleda autora
- **broadcaster's created time** – kada se autor priključio Twitch-u
- **source resolution** – rezolucija videa

Opis skupa podataka

PROSTIRANJE PODATAKA

Prenos 1 - * 0 - * Video igra

Agregacije

1. Pronaći 10 igrica koje su imale najviše gledalaca u jednom trenutku (u okviru svih prenosa koji su u tom trenutku bili aktivni).
2. Za svakog autora odrediti koju je igru igrao najveći broj minuta.
3. Za svaki sat u danu pronaći koja je igra imala u prosjeku najviše prenosa.
4. Koje igre je prenosio autor koji je ostvario najveći rast broja pratilaca u posmatranom vremenskom periodu?
5. Koje žanrove je prenosio najveći broj autora?
6. Za svaku igru odrediti koliko različitih autora je igralo u bar jednom prenosu, i pronaći 3 autora koja su je igrala u najviše prenosa.
7. Pronaći 5 igara koje su prenošene najveći broj minuta.
8. Koliki je broj prenosa, i prosjek, minimum, i maksimum prosječnog broja gledalaca po prenosu za svaku igru?
9. Za svaki žanr odrediti prosječno i maksimalno trajanje prenosa, i prosječni broj gledalaca po prenosu.
10. Za sve autore koji su najveći broj minuta igrali igru The Evil Within, a pored toga su igrali bar još jednu igru, pronaći koje su još igre igrali, u koliko prenosa, koliko ukupno minuta, i koliko minuta u prosjeku po prenosu.

Inicijalna logička šema

KOLEKCIJA *GAMES*

- Dokument predstavlja zapis o jednoj video igri
- Odgovara jednom redu iz datoteke sa osnovnim podacima o video igrama

▼ (1) 10	{ 18 fields }
" _id	10
" name	Counter-Strike
" release_date	2000-11-01 00:00:00.000Z
" english	false
" developer	Valve
" publisher	Valve
> " platforms	[3 elements]
" required_age	0
> " categories	[4 elements]
> " genres	[1 element]
> " steamspy_tags	[3 elements]
" achievements	0
" positive_ratings	124534
" negative_ratings	3339
" average_playtime	17612
" median_playtime	317
" owners	10000000-20000000
" price	7.19

Inicijalna logička šema

KOLEKCIJA *STREAMS*

- Dokument odgovara jednom redu iz datoteka skupa podataka o prenosima
- Dodat je podatak o vremenu iz naziva datoteke
- Dodata je proširena referenca na video igru

▼ (1) ObjectId("60c1e3126faa4ef0359a7d22")	{ 8 fields }
_id	ObjectId("60c1e3126faa4ef0359a7d22")
timestamp	2015-02-01 00:00:00.000Z
stream_id	12932973168
current_views	34846
start_time	2015-02-01 02:48:25.000Z
▼ game	{ 7 fields }
_id	570
name	Dota 2
release_date	2013-07-09 00:00:00.000Z
publisher	Valve
> categories	[7 elements]
> genres	[3 elements]
price	0.0
language	en
▼ broadcaster	{ 7 fields }
id	29578325
name	beyondthesummit
follower_number	197236
partner	true
language	en
total_views	197742366
created	2012-04-07 04:16:39.000Z

Inicijalna logička šema

PRIMJERI AGREGACIJA

8. Koliki je broj prenosa, i prosjek, minimum, i maksimum prosječnog broja gledalaca po prenosu za svaku igru? – 176 sekundi

```
db.getCollection('streams').aggregate([
  {$group: {_id: {$game: "$game.name", stream_id: "$stream_id"}, average_views: {$avg: "$current_views"}}},
  {$group: {_id: "$_id.game",
    average_views: {$avg: "$average_views"},
    max_views: {$max: "$average_views"},
    min_views: {$min: "$average_views"},
    count: {$sum: 1}}},
  {$sort: {"average_views": -1}},
  {$project: {_id: 0,
    game: "$_id",
    max_views_per_stream: "$max_views",
    average_views_per_stream: "$average_views",
    min_views_per_stream: "$min_views",
    number_of_streams: "$count"}}
], {allowDiskUse: true})
```

Inicijalna logička šema

PRIMJERI AGREGACIJA

4. Koje igre je prenosio autor koji je ostvario najveći rast broja pratilaca u posmatranom vremenskom periodu? – 467 sekundi

```
db.getCollection('streams').aggregate([
  {$sort: {"timestamp": 1}},
  {$group: {_id: "$broadcaster.name",
    follower_counts: {$push: "$broadcaster.follower_number"},
    played_games: {$addToSet: "$game.name"}}},
  {$project: {_id: 0,
    broadcaster: "$_id",
    played_games: 1,
    start_follower_count: {$arrayElemAt: [ "$follower_counts", 0 ]},
    end_follower_count: {$arrayElemAt: [ "$follower_counts", -1 ]}},
  {$project: {broadcaster: 1, played_games: 1, follower_change: {$subtract: ["$end_follower_count", "$start_follower_count"]}},
  {$sort: {"follower_change": -1}},
  {$limit: 1}
], {allowDiskUse: true})
```

Inicijalna logička šema

PRIMJERI AGREGACIJA

1. Pronaći 10 igrica koje su imale najviše gledalaca u jednom trenutku (u okviru svih prenosa koji su u tom trenutku bili aktivni). – 154 sekunde

```
db.getCollection('streams').aggregate([
  {$group: {_id: {game: "$game.name", timestamp: "$timestamp"}, total_views: {$sum: "$current_views"}}},
  {$group: {_id: "$_id.game", max_views: {$max: "$total_views"}}},
  {$sort: {"max_views": -1}},
  {$limit: 10},
  {$project: {_id: 0, game: "$_id", max_views: 1}}
], {allowDiskUse: true})
```

Inicijalna logička šema

PRIMJERI AGREGACIJA

2. Za svakog autora odrediti koju je igru igrao najveći broj minuta. – 192 sekunde

```
db.getCollection('streams').aggregate([
  {$group: {_id: {game: "$game.name", stream_id: "$stream_id", broadcaster: "$broadcaster.name"},
    start: {$min: "$timestamp"}, end: {$max: "$timestamp"}}},
  {$project: {game: "$_id.game",
    broadcaster: "$_id.broadcaster",
    duration_in_minutes: {$divide: [{$subtract: ["$end", "$start"]}, 60000]}},
  {$group: {_id: {game: "$game", broadcaster: "$broadcaster"}, total_minutes: {$sum: "$duration_in_minutes"}}},
  {$project: {game: "$_id.game", broadcaster: "$_id.broadcaster", total_minutes: 1}},
  {$sort: {"total_minutes": -1}},
  {$group: {_id: "$broadcaster", games: {$push: {name: "$game", total_minutes: "$total_minutes"}}}},
  {$project: {_id: 0, broadcaster: "$_id", most_played_game: {$arrayElemAt: [ "$games", 0 ]}}}
], {allowDiskUse: true})
```

Logička šema prilagođena agregacijama

- Sastoji se iz 5 kolekcija: *games*, *streams*, *broadcasters*, *games-per-time*, i *games-broadcasters*
- Kolekcija *games* je ista kao u inicijalnoj šemi
- Ostale kolekcije su dobijene primjenom šablona baketiranja i proračunavanja na postojeće kolekcije
- Grupisanje pri baketiranju je vršeno po onim atributima koji su se često koristili za grupisanje u agregacijama

Logička šema prilagođena agregacijama

KOLEKCIJA *STREAMS*

- Dobijena primjenom šablona baketiranja i proračunavanja na kolekciju *streams* iz inicijalne šeme, grupisanjem dokumenata po id-ju prenosa i video igri
- Jedan dokument nove kolekcije se odnosi na sekciju jednog prenosa u toku koje je igrana jedna video igra
- Podaci specifični za vremenski trenutak grupisani su u listu
- Dodati su atributi dobijeni proračunavanjem koji su često korišteni u upitima

▼ (3) { 2 fields }	{ 12 fields }
▼ _id	{ 2 fields }
game	Wings of Vi
stream_id	12897646832
▼ details	[5 elements]
▼ [0]	{ 3 fields }
timestamp	2015-02-01 00:00:00.000Z
current_views	1981
▼ broadcaster_details	{ 2 fields }
follower_number	293656
total_views	46208589
> [1]	{ 3 fields }
> [2]	{ 3 fields }
> [3]	{ 3 fields }
> [4]	{ 3 fields }
start_time	2015-02-01 00:00:00.000Z
end_time	2015-02-01 00:20:00.000Z
max_views	1981
min_views	348
view_sum	3501
timestamp_count	5.0
▼ game	{ 7 fields }
_id	318530
name	Wings of Vi
release_date	2014-11-28 00:00:00.000Z
publisher	Grynssoft
> categories	[5 elements]
> genres	[3 elements]
price	10.99
stream_id	12897646832
▼ broadcaster	{ 3 fields }
_id	8330235
name	manvsgame
created	2009-09-17 19:42:52.000Z
duration_in_minutes	20.0

Logička šema prilagođena agregacijama

KOLEKCIJA *BROADCASTERS*

- Dobijena primjenom šablona baketiranja na kolekciju *streams* iz inicijalne šeme, grupisanjem po autoru
- Izbačeni su atributi koji se ne odnose direktno na autora
- Podaci specifični za vremenski trenutak grupisani su u listu

▼ (8) { 2 fields }	{ 4 fields }
> _id	{ 2 fields }
▼ details	[2 elements]
▼ [0]	{ 4 fields }
# follower_number	0
# total_views	0
timestamp	2015-02-01 07:45:00.000Z
partner	false
> [1]	{ 4 fields }
name	001jonas001
created	2014-12-21 14:32:43.000Z

Logička šema prilagođena agregacijama

KOLEKCIJA *GAMES-PER-TIME*

- Dobijena grupisanjem kolekcije *streams* iz inicijalne šeme po video igri i vremenskom trenutku
- Proračunati su podaci o gledanosti i broju prenosa za igru u posmatranom trenutku koji su bili potrebni za upite

```
▼ { 1 } { 2 fields }
  ▼ { 1 } _id
    game
    timestamp
  total_views
  total_streams
  ▼ { 1 } game
    name
    _id
    > { 1 } genres
    timestamp
```

```
{ 5 fields }
{ 2 fields }
#KILLALLZOMBIES
2015-02-07 09:45:00.000Z
0
1.0
{ 3 fields }
#KILLALLZOMBIES
303720
[ 2 elements ]
2015-02-07 09:45:00.000Z
```


Logička šema prilagođena agregacijama

KOLEKCIJA *GAMES-BROADCASTERS*

- Dobijena grupisanjem kolekcije *streams* iz nove šeme po video igri i autoru
- Proračunati su podaci o broju prenosa koje je autor imao za igru, i ukupnom trajanju svih tih prenosa
- Podignuti su indeksi nad atributima *number_of_streams* i *total_duration_in_minutes*

```
▼ (1) { 2 fields }
  ▼ _id
    game
    broadcaster
    number_of_streams
    total_duration_in_minutes
  ▼ game
    name
    _id
  > genres
  ▼ broadcaster
    name
    _id
```

```
{ 5 fields }
{ 2 fields }
Counter-Strike: Global Offensive
chrischunli
1.0
105.0
{ 3 fields }
Counter-Strike: Global Offensive
730
[ 2 elements ]
{ 2 fields }
chrischunli
10005214
```

Logička šema prilagođena agregacijama

PRIMJERI AGREGACIJA

8. Koliki je broj prenosa, i prosjek, minimum, i maksimum prosječnog broja gledalaca po prenosu za svaku igru? – 9.61 sekundi

```
db.getCollection('streams').aggregate([
  {$project: {game: "$game.name", average_views: {$divide: ["$view_sum", "$timestamp_count"]}}},
  {$group: {_id: "$game",
    average_views: {$avg: "$average_views"},
    max_views: {$max: "$average_views"},
    min_views: {$min: "$average_views"},
    count: {$sum: 1}}},
  {$sort: {"average_views": -1}},
  {$project: {_id: 0,
    game: "$_id",
    max_views_per_stream: "$max_views",
    average_views_per_stream: "$average_views",
    min_views_per_stream: "$min_views",
    number_of_streams: "$count"}}
], {allowDiskUse: true})
```

Logička šema prilagođena agregacijama

PRIMJERI AGREGACIJA

4. Koje igre je prenosio autor koji je ostvario najveći rast broja pratilaca u posmatranom vremenskom periodu? – 5.31 sekundi

```
db.getCollection('broadcasters').aggregate([
  {$project: {_id: 0,
    broadcaster: "$name",
    start_details: {$arrayElemAt: [ "$details", 0 ]},
    end_details: {$arrayElemAt: [ "$details", -1 ]}}},
  {$project: {broadcaster: 1, follower_change: {$subtract: ["$end_details.follower_number", "$start_details.follower_number"]}},
  {$sort: {"follower_change": -1}},
  {$limit: 1},
  {$lookup:
    {
      from: "games-broadcasters",
      localField: "broadcaster",
      foreignField: "_id.broadcaster",
      as: "games"
    }
  },
  {$project: {broadcaster: 1, follower_change: 1, games: "$games.game.name"}}
], {allowDiskUse: true})
```

Logička šema prilagođena agregacijama

PRIMJERI AGREGACIJA

1. Pronaći 10 igrica koje su imale najviše gledalaca u jednom trenutku (u okviru svih prenosa koji su u tom trenutku bili aktivni). – 2.53 sekundi

```
db.getCollection('games-per-time').aggregate([
  {$group: {_id: "$_id.game", max_views: {$max: "$total_views"}}},
  {$sort: {"max_views": -1}},
  {$limit: 10},
  {$project: {_id: 0, game: "$_id", max_views: 1}}
], {allowDiskUse: true})
```

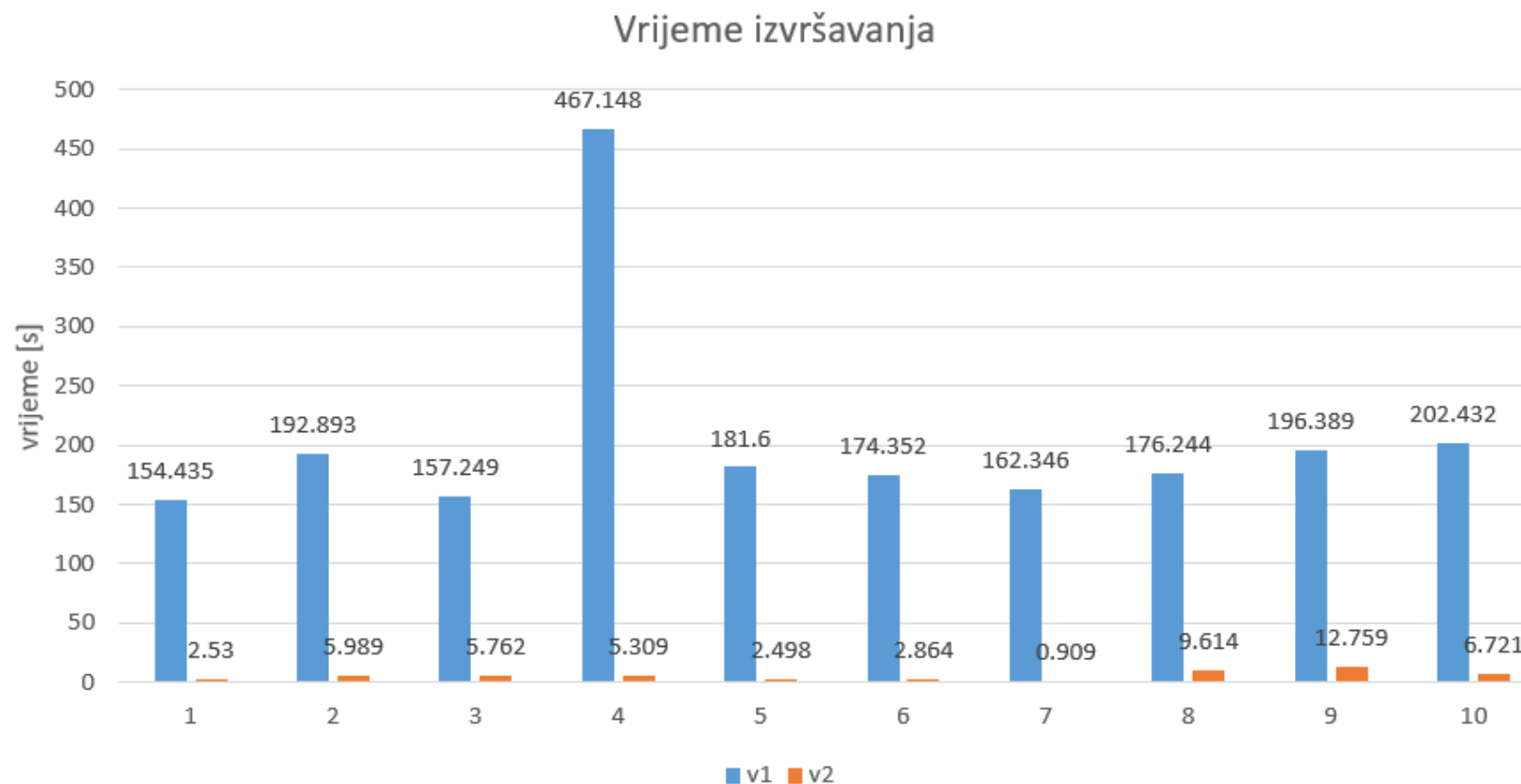
Logička šema prilagođena agregacijama

PRIMJERI AGREGACIJA

2. Za svakog autora odrediti koju je igru igrao najveći broj minuta. – 5.99 sekundi sa indeksom, oko 8 sekundi bez indeksa

```
db.getCollection('games-broadcasters').aggregate([
  {$sort: {"total_duration_in_minutes": -1}},
  {$group: {_id: "$broadcaster.name", games: {$push: {name: "$game.name", total_minutes: "$total_duration_in_minutes"}}}},
  {$project: {_id: 0, broadcaster: "$_id", most_played_game: {$arrayElemAt: [ "$games", 0 ]}}}
], {allowDiskUse: true})
```

Poređenje performansi



Poređenje performansi

