Big Data
Final Project

Dmitry Beresnev
d.beresnev@innopolis.university

Vsevolod Klyushev
v.klyushev@innopolis.university

Project description

GOAL –

create the recommendation system for Steam game platform.
In other words, we try to predict whether user would recommend some game for other users or not

1. Data

Datasets

48k

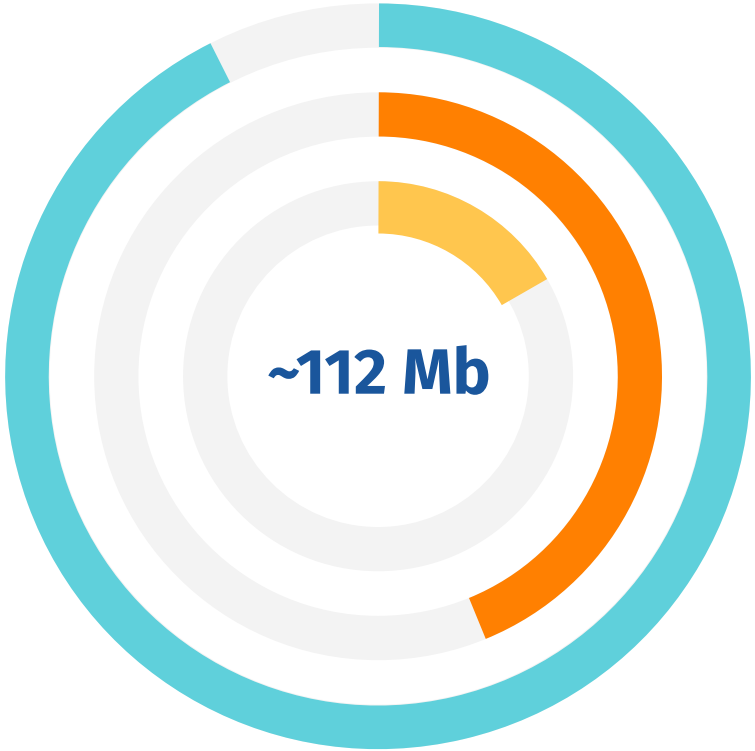
GAMES

Information about games from Steam gallery

1.2m

USERS

Anonymized information about Steam users



2.6m

RECOMMENDATIONS

Users' opinion about games

Game schema

app_id	title	date_release	win	mac	linux	rating
integer	varchar(256)	date	boolean	boolean	boolean	varchar(32)

Game schema (cont.)

positive_ratio	user_reviews	price_final	price_original	discount	steam_deck
integer	integer	real	real	real	boolean

User schema

user_id	products	reviews
integer	integer	integer

Recommendation schema

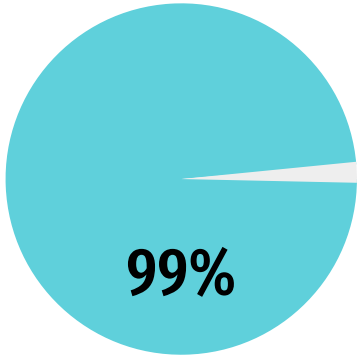
app_id	helpful	funny	date
integer	integer	integer	date

Recommendation schema (cont.)

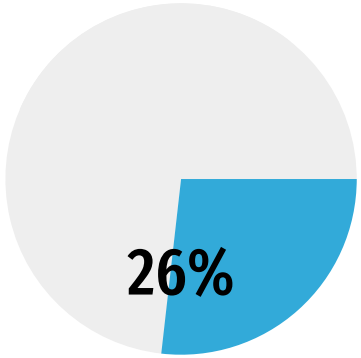
is_recommended	**hours**	**user_id**	**review_id**
boolean	real	integer	integer

2. Data analytics

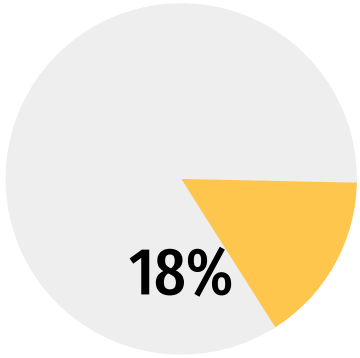
Games supporting platforms



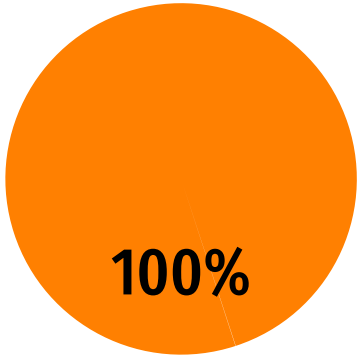
Windows



macOS



Linux

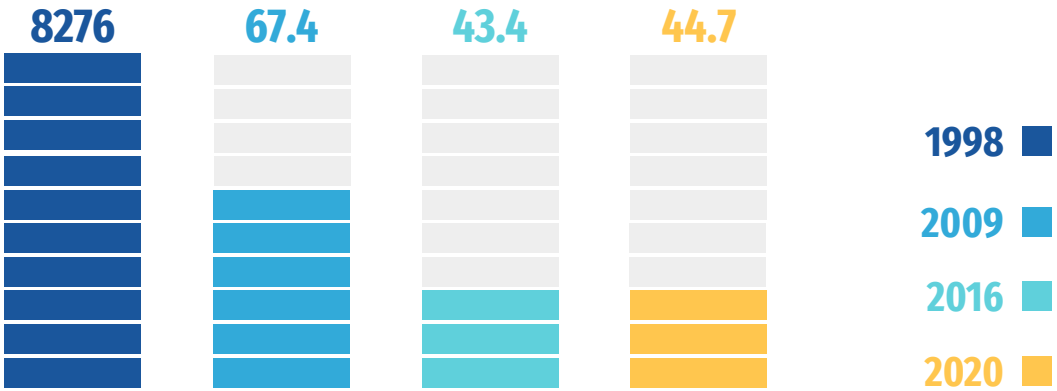


Steam Deck

Some years statistics

Number of games

Average reviews per game



Interesting fact!

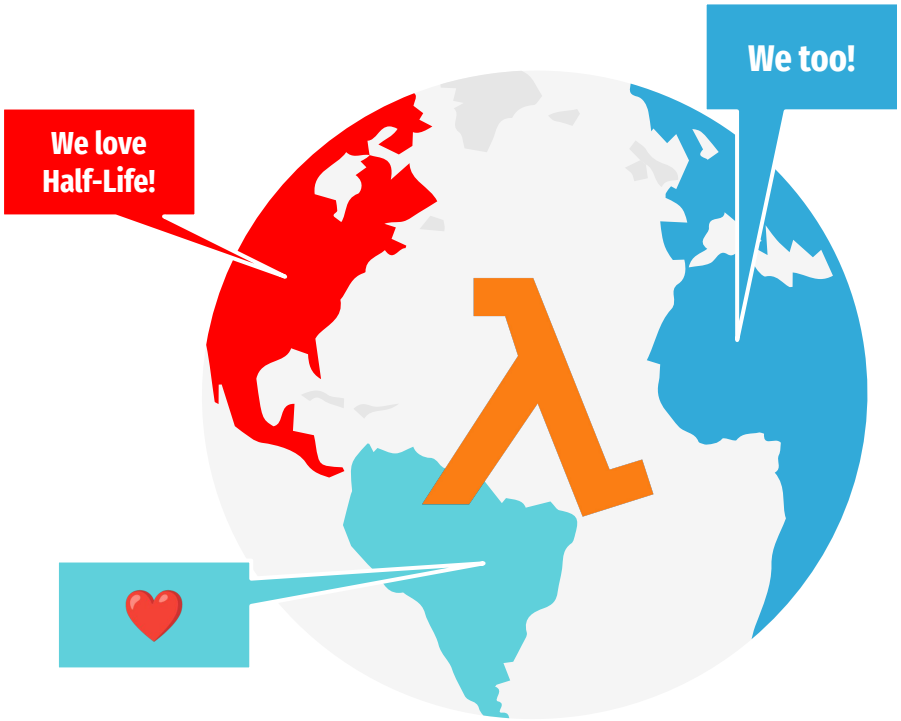
Half-Life 1998

is still extremely popular, despite its age:

SINGLE game of 1998 in our dataset

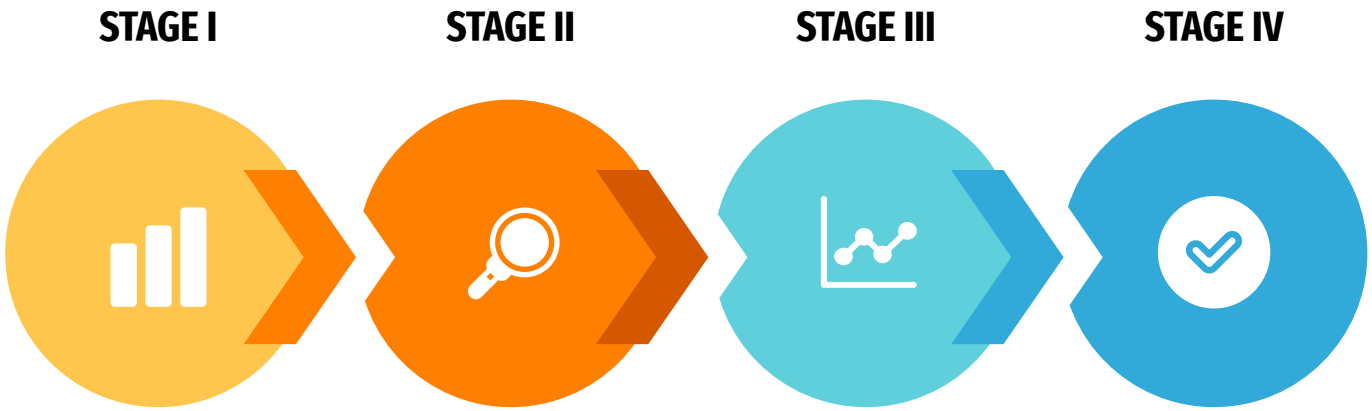
8276 reviews

NO bad reviews in our dataset



3. Work process

Progress


STAGE I
STAGE II
STAGE III
STAGE IV
Data preprocessing
Partitioning and data analysis
Building ML models
Streamlite dashboard

4. Faced challenges

Difficulties

01
Resources

Cluster limitations as runs on local machine

02
New stack

PySpark, Hive, Zeppelin, Avro

03
Old python

Strange pylint errors, no familiar API

04
Zeppelin

Problems with encoding, slow work

05
Data

Parsing and converting between stages

5. Performance

Model metrics

ALS

Random Forest

NDSG Normalized discounted cumulative gain

MAP Mean average precision

6.Streamlit Time

7. Conclusion

Afterwords

A distributed system is like a world...
It would work better being monolithic...

But seriously

It was quite an interesting experience working with distributed systems. We think it would be better if we were working with a real cluster. However, now we have an idea about the overflow of working with

BIG DATA

Thanks for attention!

If you have any questions - please, ask