

# **Интеллектуальный анализ работы хранилища данных на основании обработки логов "Ростелеком"**

**команда DA44**

**25 сентября 2022**

**Skolkovo Hack 2022**

# Описание данных

Данные представляют схематичные логи запросов, выгруженные из базы Greenplum 5

rn	loguser	q
724484	etl_2048	from tbl_77830,from tbl_77829
884851	etl_1151	from tbl_336309
825173	etl_1151	into tbl_28425,JOIN tbl_28425,from tbl_97328
415920	etl_2048	from tbl_79634
469923	etl_1151	from tbl_19348

**rn** - это порядковый номер строки, он уникальный.

**loguser** - тип пользователя, выполняющий запрос. Где etl - группа разработчиков, которые загружают данные с помощью etl процессов, dev - обычные аналитики.

**q** - запрос query, упрощенный запрос, который пользователь отправлен в базу. Здесь через запятую склеены номера таблиц и оператор.

from, join означают, что данные извлекались,

into - запись данных

# Задача

узнать, какие объекты в базе являются бесполезными.

**Бесполезный объект** - объект, который продолжает наполняться данными, но никто из разработчиков к ним не обращается, то есть запрос `into` присутствует, но при этом нет ни одного селекта от обычного аналитика с именем, которое начинается на `dev`.

## Решение нашей команды:

- проведен исследовательский анализ проблемы в jupyter notebook;
- разработана метрика "бесполезности": отношение числа записей в таблицу к числу обращений от разработчиков;
- реализовано веб-приложение Dash, в котором выведены:
  - топ-10 таблиц по нашей метрике;
  - топ-10 пользователей по кол-ву запросов в разрезе типа: etl и dev;
  - топ-10 пользователей по кол-ву операций в запросе в разрезе типа: etl и dev.

Используемые OpenSource Технологии:  
Jupyter notebook, Python, Dash.

# Определение метрики "бесполезности" таблицы

## ТОП-10 "бесполезных таблиц"

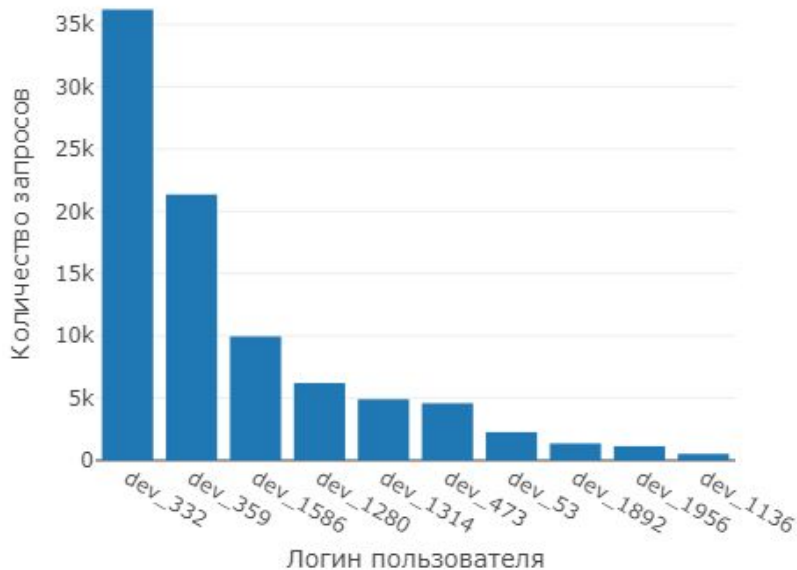
### Метрика **useless**

определена как  
отношение количества  
запросов into к  
количеству запросов  
from

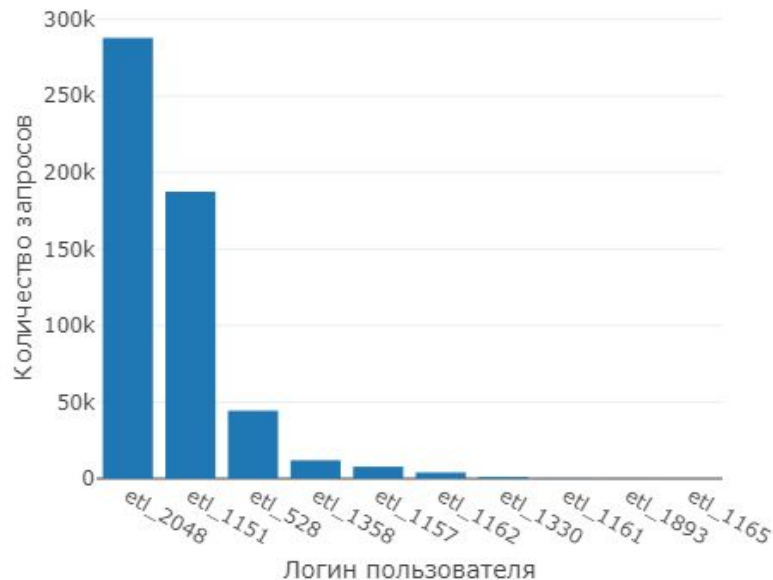
Имя таблицы	Число Froms	Число Intos	useless rate
tbl_27350	1	1536	1536
tbl_211513	9	9593	1066
tbl_29029	1	468	468
tbl_29481	1	466	466
tbl_29497	1	465	465
tbl_29495	1	465	465
tbl_27483	1	455	455
tbl_27346	1	455	455
tbl_29040	1	453	453
tbl_27761	1	447	447

# Распределение пользователей по кол-ву запросов

## ТОП-10 dev пользователей



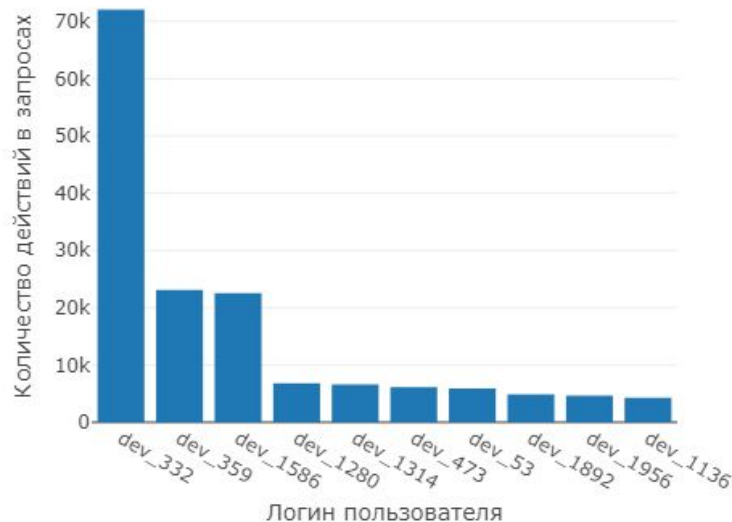
## ТОП-10 etl пользователей



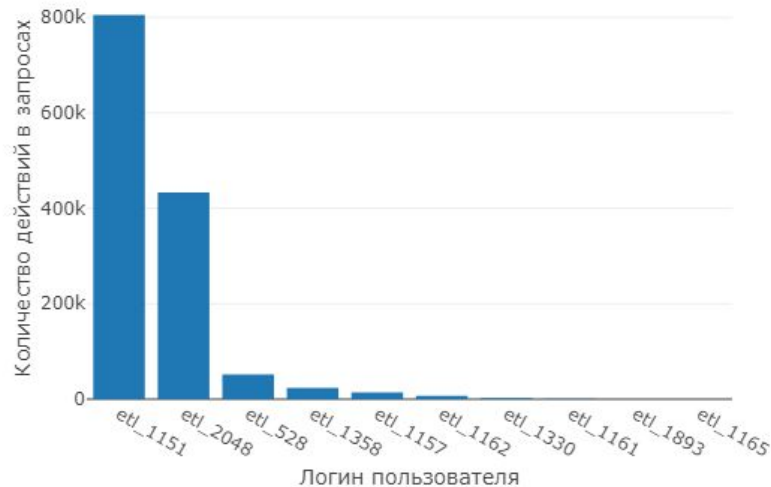
По графикам видно, что наибольшее количество запросов создают пользователи:  
dev\_332, dev\_359, dev\_1586, etl\_2048, etl\_1151, etl\_528.

# Распределение пользователей по кол-ву действий в запросах

## ТОП-10 dev пользователей



## ТОП-10 etl пользователей



По графикам видно, что запросы с наибольшим количеством действий создают те же пользователи, только меняются местами etl\_2048 и etl\_1151:

dev\_332, dev\_359, dev\_1586, etl\_1151, etl\_2048, etl\_528.

## Рекомендации

По результатам ранжирования по метрике “бесполезности” нужно обратить внимание на таблицы: tbl\_27350, tbl\_211513, tbl\_29029. Они очень часто записываются, но редко запрашиваются.

Также следует изучить действия пользователей dev\_332, dev\_359, dev\_1586, etl\_2048, etl\_1151, etl\_528, совершающих наибольшее количество запросов.



# Как использовать

Необходимое ПО:

- Python 3.9+
- библиотеки Python:
  - dash
  - regex
  - plotly
  - pandas

Путь к csv-файлу с логами необходимо указать в файле config.py. Для запуска веб-приложения необходимо открыть терминал и выполнить команду:

- для Windows и Mac: `python app.py`
- для Linux: `python3 app.py`
- для запуска на виртуальной машине: необходимо добавить host в конце app.py - `app.run_server(debug=True, host='0.0.0.0')`

# Спасибо за внимание

Команда DA44



Github: den-dw, Morjella, Mike-solk, taisiiap, AlexeyK12