Оглавление

[Титульный слайд Даниил 1](#_Toc121376942)

[Предметная область Цуй 1](#_Toc121376943)

[Цель и постановка задач Даниил 1](#_Toc121376944)

[Аналогичные исследования Даниил 2](#_Toc121376945)

[Инфраструктура сервиса Цуй 2](#_Toc121376946)

[Датасет и Ограничения источника данных Даниил 2](#_Toc121376947)

[Реализация модуля загрузки Даниил 2](#_Toc121376948)

[Реализация БД Даниил 2](#_Toc121376949)

[Горизонтальное расширение Арина 3](#_Toc121376950)

# 1 Титульный слайд Даниил

Добрый день.  
Наша команда состоит из 4 программистов. Мы занимались исследованием игрового рынка на основе данных с площадки Steam.

# 2 Предметная область Цуй

Мы выбрали именно эту тему поскольку игры близки каждому из нас. Мы видим широкое будущее игровой индустрии и хотим туда попасть. Более того, некоторые члены нашей команды уже несколько лет разрабатывают игры.

Игры сейчас выходят десятками каждый день и нам стало интересно проследить за тенденциями игровой индустрии как с точки зрения игроков, так и с точки зрения заказчиков.

# 3 Цель и постановка задач Даниил

Наши цели выходят за рамки этого курса, так как мы планируем заниматься этим проектом еще и следующий семестр. Поэтому на слайде показаны цели конечного продукта, а постановка задач – только на текущий семестр.

Одна из конечных целей – сделать удобный и независимый от площадок инструмент для рекомендации игр пользователям.

Другая – для бизнес-аналитиков игровой индустрии. Мы хотим сделать инструмент для анализа трендов популярности игр в зависимости от их жанров, выявить корреляции и на основе машинного обучения создать инструмент-предсказатель успешности игры по ее параметрам.

Соответственно, в этот семестр мы вынесли большую часть аналитических задач, касающихся только игр.

# 4 Аналогичные исследования Даниил

Поверхностное исследование на малой выборке игр.

Статистика от Гугла. Ключевая оценка – прибыльность. Только ее и исследовали.

Схожая работа по тематике. Однако там проводились исследования а - игроков и б - среди друзей.  
У нас это впереди и масштабнее.

# 5 Инфраструктура сервиса Цуй

На данный момент реализована такая инфраструктура сервиса.

Майнер с помощью стимАПИ выкачивает данные, проверяет, чистит и кладет их в БД.  
Оттуда анализаторы подтягивают информацию и анализируют.

Также, был реализован сервер, который выдает результаты анализа в виде веб-интерфейса.

Сервер имеет возможность обновлять гистограммы и графики по данным из базы в реальном времени.

# 6 Датасет и Ограничения источника данных Даниил

Датасет содержит 1,5 ГБ json файлов. В них хранится информация о 150к приложений, из которых большую часть составляют игры. Остальное – это дополнения, музыка, сторонние разработки.

Для заполнения данных об одной игре требовалось сделать минимум 2 запроса. Скорость скачивания данных – 1 запрос в 2 секунды. У стима есть хромающая документация и хромающие данные, так как некоторым записям уже почти 20 лет. Не идеальная, но и плохой ее не назвать. Однако, из-за нее код много раз переписывался и потерял изначальную красоту.

Данные можно получать в формате json/xml. В нашем случае это был джсон.   
Внимательные зрители могут заметить несоответствие между 100к запросов в день и 6 сутками выкачки…

# 7 Реализация модуля загрузки Даниил

А получилось так потому что я сначала слишком сильно распараллелил майнер и стим ограничил мне скорость выкачки до 1 запроса в 2 секунды.

# 8 Реализация БД Даниил

БД была сделана в Postgresql. По образу и подобию моделей хранения данных в стиме.

Её можно мысленно разделить на 2 части: левую и правую. Левая будет содержать информацию о пользователях для дальнейшего анализа, правая – атрибуты игр. Складировали в БД все для полноты анализа.

Возможность хоста БД (как на своей машине, так и на внешних серверах) подключений за пределами локалхоста рассматривалась, но реализовано это не было, так как не было острой необходимости. Передача данных производилась с помощью бекапов.

# 9 Горизонтальное расширение Арина

Производительность системы можно поднять путем размножения майнеров, выведения БД на отдельный сервер и размножения анализаторов. В таком случае, узким местом станет база данных, но к ней можно применить партицирование или шардирование. Однако это имеет смысл только по отношению к части бд, связанной с игроками, так как игр всего 150к и растут они медленно, а игроков в сотни раз больше и рост быстрее.

# 10 Реализация анализатора Арина

# 11 Реализация анализатора Арина

# 12 Визуализация результатов Арина

# 13 14 15 16 17 18 результаты Арина

# Переход к веб интерфейсу, остальные результаты

# 19 Корреляции Женя

# 20 Кластеризации Женя

# 21 Характеристики системы и времени Женя

# 22 Формат работы и временные затраты команды Даниил