

Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
GRADUATION THESIS

Разработка системы онлайн-антифрода для сервиса новостей

Обучающийся / Student Гуммель Никита Константинович

Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования

Группа/Group М34041

Направление подготовки/ Subject area 09.03.02 Информационные системы и технологии

Образовательная программа / Educational program Программирование и интернет-технологии 2019

Язык реализации ОП / Language of the educational program Русский

Статус ОП / Status of educational program

Квалификация/ Degree level Бакалавр

Руководитель ВКР/ Thesis supervisor Койнов Руслан Васильевич, Университет ИТМО, факультет информационных технологий и программирования, преподаватель (квалификационная категория "преподаватель практики")

Консультант не из ИТМО / Third-party consultant Стрельцов Антон Алексеевич, ООО "Дзен.Платформа", Руководитель группы антифрод

Обучающийся/Student

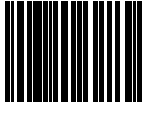
Документ подписан	
Гуммель Никита Константинович	
18.05.2023	

(эл. подпись/ signature)

Гуммель
Никита
Константинови
ч

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Койнов Руслан Васильевич	
18.05.2023	

(эл. подпись/ signature)

Койнов Руслан
Васильевич

(Фамилия И.О./ name
and surname)

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
SUMMARY OF A GRADUATION THESIS**

Обучающийся / Student Гуммель Никита Константинович
Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования
Группа/Group М34041
Направление подготовки/ Subject area 09.03.02 Информационные системы и технологии
Образовательная программа / Educational program Программирование и интернет-технологии 2019
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Бакалавр
Тема ВКР/ Thesis topic Разработка системы онлайн-антифрода для сервиса новостей
Руководитель ВКР/ Thesis supervisor Койнов Руслан Васильевич, Университет ИТМО, факультет информационных технологий и программирования, преподаватель (квалификационная категория "преподаватель практики")
Консультант не из ИТМО / Third-party consultant Стрельцов Антон Алексеевич, ООО "Дзен.Платформа", Руководитель группы антифрод

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
DESCRIPTION OF THE GRADUATION THESIS**

Цель исследования / Research goal

Проектирование и разработка системы онлайн-антифрода, поставляющей таблицы логов, проверенные на предмет фрода.

Задачи, решаемые в ВКР / Research tasks

- Определение требований к системе - Проектирование архитектуры системы - Реализация спроектированной архитектуры - Настройка мониторингов и алертов

Краткая характеристика полученных результатов / Short summary of results/findings

Составлены функциональные и нефункциональные требования к системе. Спроектирована архитектура, соответствующая требованиям. Спроектированная архитектура реализована, мониторинги и алерты настроены, система введена в эксплуатацию.

Обучающийся/Student

Документ подписан	
Гуммель Никита Константинович	
18.05.2023	

(эл. подпись/ signature)

Гуммель
Никита
Константинови
ч

(Фамилия И.О./ name)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Койнов Руслан Васильевич	
18.05.2023	

(эл. подпись/ signature)

and surname)

Койнов Руслан
Васильевич

(Фамилия И.О./ name
and surname)

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**ЗАДАНИЕ НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ /
OBJECTIVES FOR A GRADUATION THESIS**

Обучающийся / Student Гуммель Никита Константинович
Факультет/институт/кластер/ Faculty/Institute/Cluster факультет информационных технологий и программирования
Группа/Group М34041
Направление подготовки/ Subject area 09.03.02 Информационные системы и технологии
Образовательная программа / Educational program Программирование и интернет-технологии 2019
Язык реализации ОП / Language of the educational program Русский
Статус ОП / Status of educational program
Квалификация/ Degree level Бакалавр
Тема ВКР/ Thesis topic Разработка системы онлайн-антифрода для сервиса новостей
Руководитель ВКР/ Thesis supervisor Койнов Руслан Васильевич, Университет ИТМО, факультет информационных технологий и программирования, преподаватель (квалификационная категория "преподаватель практики")
Консультант не из ИТМО / Third-party consultant Стрельцов Антон Алексеевич, ООО "Дзен.Платформа", Руководитель группы антифрод

Основные вопросы, подлежащие разработке / Key issues to be analyzed

Целью выпускной квалификационной работы является проектирование и разработка системы онлайн-антифрода для сервиса новостей. Основной функциональностью разработанной системы является поставка таблиц логов, проверенных на фрод.

Дата выдачи задания / Assignment issued on: 15.12.2022

Срок представления готовой ВКР / Deadline for final edition of the thesis 15.05.2023

Характеристика темы ВКР / Description of thesis subject (topic)

Название организации-партнера / Name of partner organization: ООО "Дзен.Платформа"

Тема в области фундаментальных исследований / Subject of fundamental research: нет / not

Тема в области прикладных исследований / Subject of applied research: да / yes

СОГЛАСОВАНО / AGREED:

Руководитель ВКР/
Thesis supervisor

Документ подписан	
----------------------	--

Задание принял к
исполнению/ Objectives
assumed BY

Руководитель ОП/ Head
of educational program

	
Койнов Руслан Васильевич	
17.05.2023	


(эл. подпись)

Койнов Руслан
Васильевич

Документ подписан	
Гуммель Никита Константинович	
18.05.2023	

(эл. подпись)

Гуммель
Никита
Константинови
ч

Документ подписан	
Маятин Александр Владимирович	
19.05.2023	

(эл. подпись)

Маятин
Александр
Владимирович

Оглавление

Глоссарий	5
Введение	7
Глава 1. Описание предметной области	8
1.1 Исходный процесс	8
1.2 Формирование функциональных требований	11
1.3 Формирование нефункциональных требований	12
Глава 2. Проектирование	13
2.1 Системная архитектура	13
2.1 Программная архитектура	14
2.2 Архитектура данных	19
Глава 3. Реализация	25
3.1 Описание	25
3.2 News Xurma	25
3.3 Table Merger	30
3.4 Update Fat Thresholds	32
3.5 Alerts	33
3.6 Hitman	34
Заключение	36
Список используемой литературы	37

Глоссарий

Антифрод – система или процесс, направленный на борьбу с фродом, предсказание и мониторинг фрода.

Батч – набор таблиц.

Сервис Новости – новостной агрегатор, бывший Яндекс.Новости.

Фрод – мошенничество, незаконное использование информационных технологий в различных областях бизнеса.

Datalens - это сервис для бизнес-аналитики. Сервис позволяет подключаться к различным источникам данных, строить визуализации, собирать дашборды и делиться полученными результатами. Имеет удобный интерфейс работы с таблицами, расположенными в YТ.

Hitman – это универсальная платформа на базе Nirvana для создания продакшен-процессов по сбору и обработке данных. Hitman предоставляет интерфейс для настройки регулярных запусков этих процессов и мониторинга их выполнения.

Nirvana – облачная платформа для управления процессами, которые оформлены в виде ациклических графов.

Nirvana-граф – разовый процесс, имеющий графическое отображение. У каждого процесса есть своя конфигурация. Новые графы порождаются клонированием существующих. Граф может быть в одном из четырех состояний: черновик, согласованный, запущенный, завершенный. Все состояния, кроме черновика, являются неизменяемыми.

Nirvana-операция – блок исполнения в графе, который описывается конфигурацией, содержащей способ запуска произвольной программы (в т. ч. имя процессора) и другие опции. Сначала операцию нужно создать или найти, а уже потом – добавить в граф в редакторе и соединить связями с другими операциями или объектами данных.

SLA – это соглашение между заказчиком и исполнителем о том, какие, когда и как будут предоставляться услуги. В рамках данной выпускной квалификационной работы подразумевается соглашение о времени поставки таблиц онлайн-антифрода.

Telegram – мессенджер.

Workflow – набор Nirvana-графов, созданных клонированием Nirvana-графа из этого же набора.

YQL (Yandex Query Language) – универсальный декларативный язык запросов к системам хранения и обработки данных с основанным на SQL синтаксисом и поддержкой пользовательских функций на C++, Python и JavaScript. Может использоваться для работы с данными в YT.

YSON – разработанный в Яндексе формат данных, похожий на JSON.

YT – сервис распределенного хранения и обработки данных с поддержкой модели MapReduce, распределенной файловой системой и NoSQL key-value базой данных.

Введение

Многие крупные компании сталкиваются с такой проблемой, как фрод. В общем понимании фрод – это любое мошенническое действие, совершенное с целью обогащения. Однако есть множество непрямых мошеннических схем, поэтому мы будем использовать более широкое понятие, фрод – это любое действие, противоречащее правилам пользования платформой или совершенное ненастоящим пользователем, то есть роботом.

Для выявления фрода и борьбы с ним крупные компании создают свои группы антифрода или пользуются сторонними сервисами.

Распознавание фрода в событиях любой платформы является неотъемлемой частью продуктовой компании. Без четкой картины о взаимодействии пользователей с продуктом невозможны развитие, понимание узких мест, распознавание паттернов пользования продуктом.

Особенно проблема фрода актуальна для такого медиасервиса как Новости, в котором каждый день публикуются новостные СМИ и миллионы пользователей просматривают новостной контент.

Если не выявлять фрод и не бороться с ним, то модели ранжирования будут учиться на некорректных данных, не отражающих реальные действия пользователей, как следствие появляется возможность у злоумышленников сместить ранжирование в пользу каких-то конкретных СМИ или тематик новостей. Также большой поток фрода может смещать метрики во время проведения A/B тестов, поскольку один попавший в группу тестирования робот может совершать множество событий, вследствие чего продукту придется браковать тест или же результаты теста окажутся ошибочными.

Таким образом, целью выпускной квалификационной работы является разработка онлайн-антифрода для сервиса новостей.

Глава 1. Описание предметной области

1.1 Исходный процесс

Одно из основных направлений антифрода – нахождение фродовых событий и пользователей в логах.

Все события Новостей собираются воедино через logfeller и поставляются таблицами на хранилище YT. Все названия являются датами или временем, поэтому записываются в формате ISO8601. Описание поставляемых таблиц представлено на таблице 1.

Таблица 1 – Описание таблиц, поставляемых logfeller

Директория	Название таблицы	Описание	Изменяемы?	Срок жизни
stream/5min	Время поставки таблицы, всегда кратно пяти минутам	Данные таблицы хранят в себе события по мере их появления в logfeller, таким образом данные таблицы могут хранить в себе события за любой промежуток времени	Нет	2 суток
30min	Нижняя граница тридцатиминутного интервала	Таблицы хранят в себе данные за каждый тридцатиминутный интервал суток	В таблицы могут дописываться события, которые	3 суток
1d	Дата, за которую собраны события в данной таблице	Хранят в себе все события за какую-либо дату	logfeller получил позже времени поставки	Не удаляются

Выявление фрода в логах делится на два вида:

- Оффлайн-антифрод, который запускается раз в сутки. Данный процесс тяжелый, а его результатом работы является таблица с обнаруженными фродовыми пользователями за сутки;
- Онлайн-антифрод обрабатывает таблицы, поставляемые logfeller, и добавляет к каждой строке лога колонку rules, содержащую список идентификаторов антифрод-правил, которые сработали на данной строке лога. Соответственно, если в колонке rules пустой лист – событие не является фродом. Результатом онлайн антифрода являются таблицы с колонкой rules, поставляемые в том же формате, что и таблицы, которые предоставляет logfeller.

Проверенные через онлайн-антифрод таблицы используются для переобучения ранжирующих моделей, построения дашбордов продуктовой командой, а также для аналитических скриптов на свежих данных с возможностью фильтрации от фродовых событий. Дневные таблицы, полученные в результате работы онлайн-антифрода, вместе с оффлайн-антифродом используются для построения сессий, которые в дальнейшем используются командами ML и продуктовой аналитики.

До того, как Новости отделились от Яндекса, онлайн-антифрод совершался на стороне сервиса антифрода Яндекса, правила для обнаружения фрода задавала команда Новостей. После отказа от части сервисов Яндекса появилась необходимость создать свою систему онлайн-антифрода.

Задачу по разработке системы онлайн-антифрода можно разделить на три основные части:

- Определение требований к системе, таких как время и формат поставки таблиц;
- Разработка архитектуры системы с учетом требований;

- Реализация данной архитектуры с использованием технологий, определенных нефункциональными требованиями.

1.2 Формирование функциональных требований

Согласно поставленной задаче, можно выделить следующие функциональные требования к системе онлайн-антифрода сервиса новостей:

- Данные необходимо читать из stream/5min таблиц, поставляемых logfeller. Сделано это с целью уменьшения времени поставки таблиц, так как данные таблицы имеют наименьшую задержку поставки;
- Поставка таблиц должна осуществляться в том же формате, что и у таблиц logfeller'а. То есть stream/5min идентичны обрабатываемым таблицам, 30min и 1d таблицы должны содержать в себе данные, отфильтрованные по timestamp событий;
- Согласно SLA, задержка поставки таблиц онлайн-антифрода должна составлять не более 4 часов;
- Для системы должен быть настроен мониторинг времени поставки таблиц и времени работы отдельных компонентов;
- Должны быть настроены алерты:
 - Алерты об упавших процессах должны быть настроены через Hitman и отправляться через email и sms.
 - Алерты о задержке (или выявлении сигнала, что задержка возможна) поставки таблиц должны отправляться в канал алертов в Telegram и содержать в себе список задержавшихся таблиц и время поставки данных таблиц от logfeller.
- Поставляемые онлайн-антифродом таблицы должны иметь колонку rules, содержащую список идентификаторов правил, сработавших на событии. В остальной схеме таблицы должны быть идентичны тем таблицам, которые поставляет logfeller.

1.3 Формирование нефункциональных требований

- Необходимо использовать logfeller, из которого берутся исходные данные;
- Для написания скриптов обработки таблиц необходимо использовать YQL для экономии времени работы;
- Процессы должны уметь работать в асинхронном режиме, чтобы ускорять процесс поставки таблиц;
- Процессы не должны одновременно работать с чувствительными данными, чтобы не нарушить консистентность данных;
- Чтение данных из logfeller необходимо производить по батчам, это необходимо для асинхронной работы;
- Необходимо использовать инструменты Nirvana и Hitman;
- Хранить все табличные данные на YT;
- Должны быть настроены разграничения доступа, так как система работает с чувствительными данными.
- Графики мониторинга должны быть реализованы в Datalens.

Глава 2. Проектирование

2.1 Системная архитектура

В команде антифрода для построения регулярных процессов используют графы в Nirvana с запуском через сервис Hitman, соответственно все элементы системы реализованы в Nirvana.

Поскольку используются технологии Яндекса, то Nirvana, Hitman и YT находятся в одном серверном окружении.

Используемые компоненты системы:

- Сервис Hitman для регулярного запуска Nirvana-графов, мониторинга всего проекта и отдельных процессов в нем;
- Сервис Nirvana для составления графов исполнения;
- Для целей проекта использовались Nirvana-операции для запуска YQL-скриптов, Groovy-скриптов для работы с YSON, YT-операций;
- Все табличные данные находятся в хранилище данных YT;
- Запускаемые YQL-скрипты разбиваются на map-reduce операции на YT.

2.1 Программная архитектура

Одной из задач в разработке системы онлайн-антифрода для сервиса Новостей являлась разработка архитектуры всей системы. Исходя из выбранного инструментария и требований, было решено реализовать 4 Nirvana-графа, которые будут по cron запускаться сервисом Nitman. Графы из различных workflow должны работать независимо друг от друга.

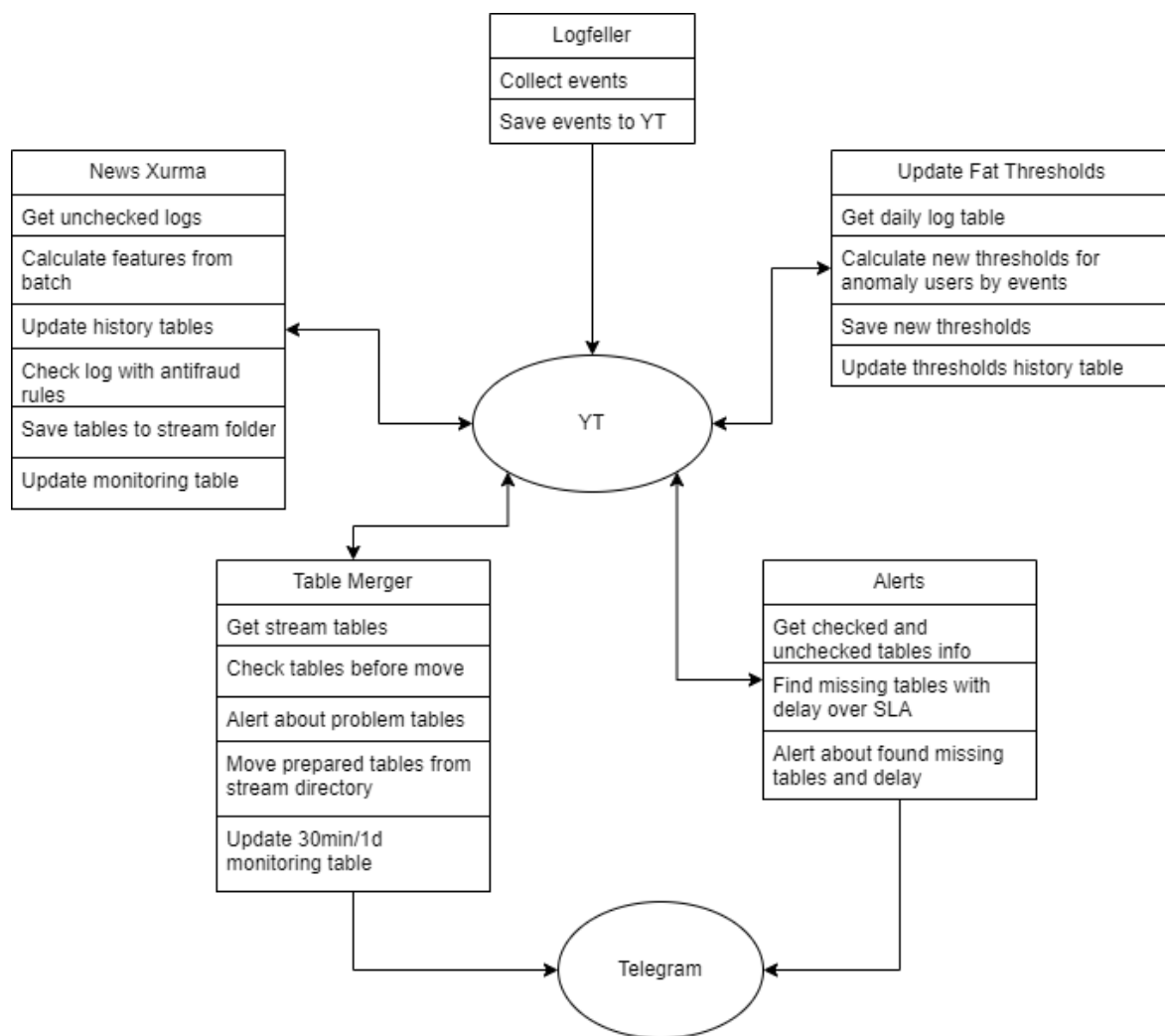


Рисунок 1 – Диаграмма классов для отображения функций графов и их взаимодействия

На рисунке 1 представлена диаграмма классов, на которой графы описаны в виде классов, а также компоненты, с которыми они взаимодействуют. Графы напрямую не взаимодействуют друг с другом, но пользуются общими таблицами.

Ниже приведена таблица с описанием каждого графа:

Таблица 2 – Описание основных графов архитектуры

Название графа	Функциональность	Результат работы
News Xurma	Основной граф системы, должен реализовывать подбор, обработку и применение антифрод-правил.	Потоковые таблицы в директории stream. Актуальная таблица мониторинга процесса.
Table Merger	Граф должен собирать таблицы 30min и 1d из stream директории, поставка осуществляется в формате logfeller'a.	Корректные таблицы в директории 30min и 1d. Актуальная таблица мониторинга процесса.
Alerts	Оповещение о задержках поставки таблиц.	Алерты в Telegram.
Update Fat Thresholds	Ежесуточное обновление порогов для правила пользователей с аномальными счетчиками событий.	Новые пороги. Актуальная таблица истории порогов.

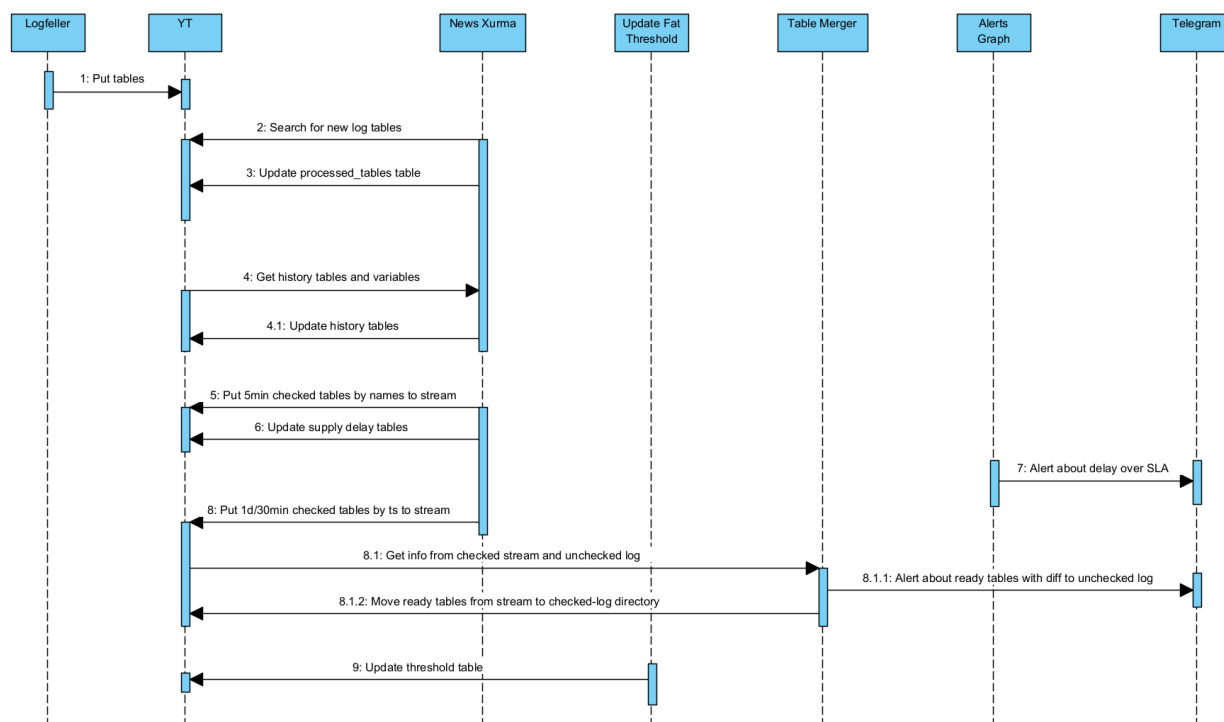


Рисунок 2 – Диаграмма последовательности взаимодействия компонентов системы

На рисунке 2 представлена диаграмма последовательности взаимодействия компонентов системы между собой.

1. Logfeller поставляет таблицы на YT в директорию zen-news-events-log (1d, 30min и stream/5min). stream/5min таблицы поставляются потоком, поток логов нарезается интервалами по 5 минут после чего поставляется как готовые таблицы. У logfeller возможны задержки, следствием чего является поставка нескольких таблиц одновременно.
2. Граф News Xurma проверяет, что прямо сейчас не обновляются таблицы истории, если обновляются – граф завершает свое выполнение. Если таблицы истории актуальны и не обновляются прямо сейчас – подбираются новые таблицы, которые еще не были обработаны.
3. Сохраняется информация о взятых в обработку таблицах.
4. Получение таблиц истории и переменных, которые затем будут использованы для применения правила онлайн-антифрода.

- 4.1 Обновление таблиц истории с добавлением в них информации из обрабатываемого батча, также происходит фильтрация событий старше суток.

Готовые таблицы со сформированной колонкой `rules`, содержащей список идентификаторов антифрод-правил, распределяются на три потока записи (пункты 5 и 8):

5. `stream/5min` таблицы записываются по названиям таблиц, так как схема данных в `zen-news-events-checked-log` должна быть идентична `zen-news-event-log`.
6. В таблицу мониторинга задержек поставок `5min` таблиц записывается информация о завершившейся поставке таблиц из батча.
7. Граф `Alerts` проверяет задержку поставок всех видов проверенных таблиц, и если задержка превышает `SLA`, то отправляет сообщение с алертом в Telegram в специализированный чат для алертов.
8. Если нужных для записи `1d` и `30min` таблиц нет в `stream`-директории, то создаются пустые таблицы с нужной схемой, далее происходит запись в данные таблицы. Это необходимо для `lock-free` записи таблиц несколькими `Nirvana`-графами.
 - 8.1 Граф `Table Merger` получает список `30min` и `1d` таблиц из `stream`-директории, которые не обновлялись более часа, данные таблицы считаются готовыми.
 - 8.2 Готовые таблицы, у которых разница с уже поставленными таблицами `zen-news-events-log` выше порога `0.0001`, считаются сигналом о проблемах в процессе или потенциальной задержке. Сообщение о проблеме отправляется в Telegram.
 - 8.3 Готовые `30min` и `1d` таблицы, с которыми не оказалось проблем из пункта 8.2, перемещаются из `stream` директории в директорию постоянного хранения. Для `30min` таблиц устанавливается время жизни 3 суток, через это время таблицы удаляются автоматически.

9. Раз в сутки запускается граф Update Fat Thresholds, который пересчитывает пороги для определения аномальных пользователей и записывает пороги в виде таблицы на YT.

2.2 Архитектура данных

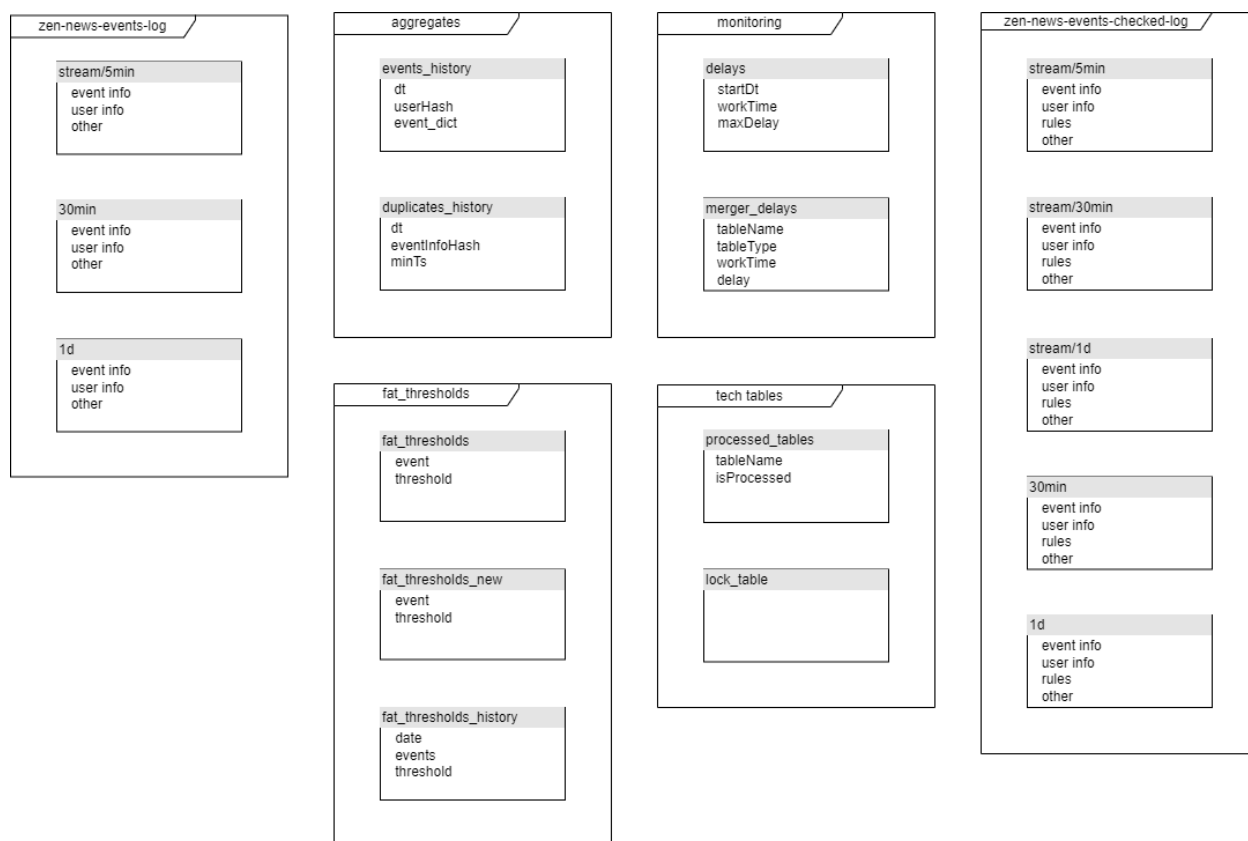


Рисунок 3 – Используемые таблицы

На рисунке 3 представлены таблицы, используемые в онлайн-антифродде:

1. Таблицы zen-news-events-log поставляются logfeller. Все таблицы имеют одинаковую схему, различие только в том, как составляются данные таблицы. Основные используемые таблицы – таблицы из директории stream/5min, из остальных берется только время поставки и размер (количество строк).
2. Агрегаты:

Данные таблицы обрабатываются только News Xurma и поддерживают окно наблюдения в сутки, фильтруя записи старше суток по dt и дописывая информацию из текущего обрабатываемого батча логов.

Таблица 3 – Таблицы истории агрегатов

Таблица	Схема		
	Ключ	Тип	Смысл
events_history	dt	Timestamp	Время запуска графа, который внес записи
	userHash	Uint64	Хэш информации о пользователе
	event_dict	Yson	Словарь событий с количеством
duplicates_history	dt	Timestamp	Время запуска графа, который внес записи
	eventInfoHash	Uint64	Хэш информации о событии и пользователе
	minTs	Timestamp	Минимальное время появления eventInfoHash в батче

Таблица events_history используется для антифрод-правил, для которых нужно знать, какие события за последние сутки совершал пользователь и сколько раз. Так же по данной таблице можно получить полный словарь действий пользователя за сутки, что позволяет проверить правило пользователей с аномальными счетчиками.

Таблица дубликатов нужна для применения правила дубликатов. Правило вычистки дубликатов является техническим и должно предотвращать спам событиями в логах от одного и того же пользователя или сгладить эффект от ошибок логирования. Если ts события не равен минимальному событию с таким же eventInfoHash за последние сутки, то событие является дубликатом.

3. Правило аномальных пользователей:

Таблица 4 – Описание таблиц для правила нахождения аномальных пользователей

Таблица	Схема			Смысл
	Ключ	Тип	Смысл	
fat_thresholds	event	String	Название события	Используемые пороги
	threshold	UInt64	Порог для правила аномальных пользователей	
fat_thresholds_new	event	String	Название события	Новые посчитанные пороги. Таблица удаляется после переноса данных из нее в fat_thresholds.
	threshold	UInt64	Порог для правила аномальных пользователей	
fat_thresholds_history	date	String	Название используемой для расчета дневной таблицы (дата рассматриваемых событий)	Таблица хранит рассчитанные пороги за каждую дату. Потенциальные всплески порогов могут свидетельствовать о фрод атаке.
	event	String	Название события	
	threshold	UInt64	Порог для правила аномальных пользователей	

Для правила нахождения аномальных по счетчикам событий пользователей граф Update Fat Thresholds запускается раз в сутки на дневной таблице лога и считает новые пороги. Пороги записываются в таблицу fat_thresholds_new, которая создается в момент записи. Так же новые посчитанные пороги дописываются в таблицу fat_thresholds_history, где для каждого дня хранятся рассчитанные пороги за день.

Запись происходит изначально в таблицу `fat_thresholds_new`, чтобы обновлять таблицу `fat_thresholds` только тогда, когда ни один процесс не будет читать из неё.

4. Технические таблицы:

Таблица 5 – Описание технических таблиц

Таблица	Схема			Смысл
	Ключ	Тип	Смысл	
<code>lock_table</code>				Используется для блокировки
<code>processed_tables</code>	<code>tableName</code>	String	Название таблицы из stream/5min	Знание о том, что какой-то граф еще не актуализировал таблицы истории исключает запуск нового графа
	<code>isProcessed</code>	Bool	False, если таблица в обработке	
			True, если история уже актуализирована	

Две данные таблицы гарантируют, что в критическом блоке графа `News Хирта`, где происходит подбор батча таблиц для обработки, будет только один граф, и что с таблицами истории работает только один граф.

5. Мониторинги:

Таблица 6 – Описание таблиц мониторинга

Таблица	Схема			Смысл
	Ключ	Тип	Смысл	
delays	startDt	Timestamp	Время запуска графа News Xurma	Граф News Xurma записывает в данную таблицу данные в ходе своей работы
	workTime	Uint64	Время в секундах между startDt и временем поставки таблиц 5min	
	maxDelay	Uint64	Максимальная задержка поставки таблицы 5min в секундах	
merger_delays	tableName	String	Название перемещаемой таблицы из stream-директории	Граф Table Merger для каждой таблицы, которую он поставляет в ходе запуска, записывает информацию о времени своей работы и о задержке поставки.
	tableType	String	“30min”/”1d”, тип таблицы, перемещаемой из stream-директории	
	worktime	Uint64	Время работы графа Table Merger в секундах	
	delay	Uint64	Задержка поставки таблицы в секундах	

Обе эти таблицы необходимы для мониторинга времени работы графов и задержки поставки таблиц.

- Таблицы zen-news-events-checked-log являются результатом работы всей системы. Все таблицы имеют одинаковую схему. Схема таблиц отличается от zen-news-events-log только наличием колонки rules,

хранящей список идентификаторов отработавших правил для каждой строки логов.

- stream/5min – директория таблиц пятиминуток, таблицы в ней хранятся двое суток, после чего автоматически удаляются;
- stream/30min – директория для временного накопления таблиц 30min. По мере готовности таблицы из stream/30min граф Table Merger забирает таблицу и производит запись в директорию 30min. Если в директории 30min уже есть таблица с таким названием – дозаписывает, иначе создает таблицу и производит запись;
- stream/1d – директория для временного накопления дневных таблиц. По мере готовности таблицы из stream/1d граф Table Merger забирает таблицу и производит запись в директорию 1d. Если в директории 1d уже есть таблица с таким названием – дозаписывает, иначе создает таблицу и производит запись;
- 30min – директория для хранения таблиц 30min, таблицы в ней хранятся трое суток, после чего автоматически удаляются;
- 1d – директория для хранения дневных таблиц.

Глава 3. Реализация

3.1 Описание

Согласно разработанной архитектуре системы онлайн-антифрода были реализованы 4 Nirvana-графа. Для каждого из них были созданы процессы в Hitman, настроены триггеры, клонирующие и запускающие графы с заранее заданными параметрами, а также мониторинги на падения процессов, частоту запуска и время выполнения с нотификацией через SMS и корпоративную почту.

3.2 News Xurma

News Xurma является основным графом всей архитектуры. В его функции входят:

- Определение нового батча необработанных таблиц;
- Предобработка данных в таблицах;
- Поддержание актуальности таблиц истории (events_history и duplicates_history);
- Обновление таблицы порогов fat_thresholds при наличии fat_thresholds_new для правила аномальных пользователей;
- Применение к каждой строке лога в батче правил онлайн-антифрода, результатом чего является список идентификаторов правил;
- Запись лога в соответствующие таблицы в директорию stream;
- Обновление таблицы мониторинга задержек поставок таблиц stream/5min.

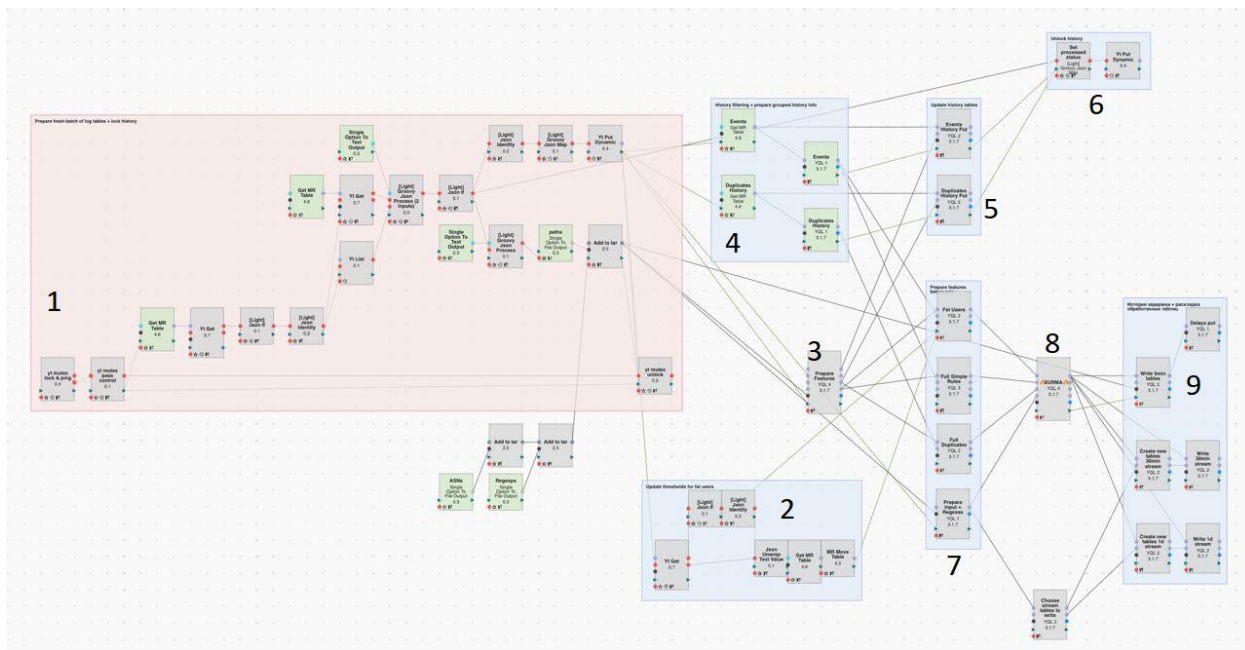


Рисунок 4 – Реализация графа News Xurta



Рисунок 5 – Очередность и связь между этапами графа News Xurta

На рисунке 4 представлена реализация архитектуры графа с подписанными номерами этапов. Для простоты понимания очередность и связь между этапами отображена в виде схемы на рисунке 5.

1. Подбор таблиц

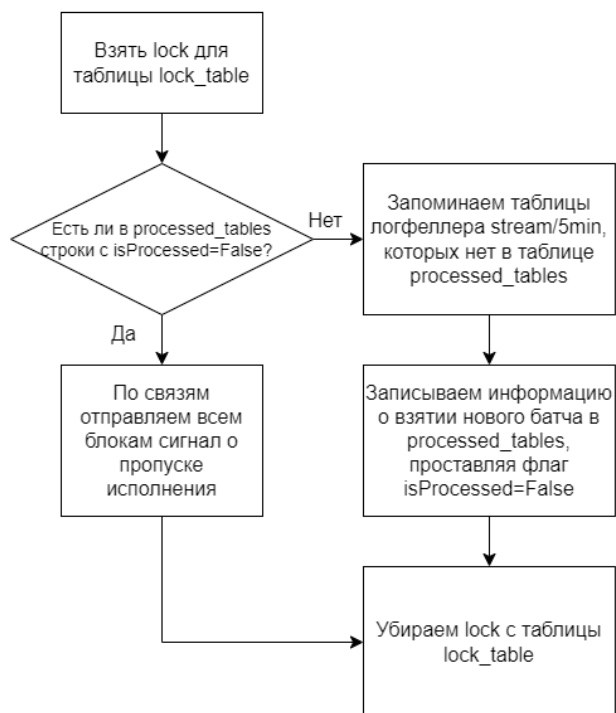


Рисунок 6 – Алгоритм работы первого этапа графа News Xurta

На рисунке 6 представлена блок-схема алгоритма работы первого этапа. Поскольку в данном блоке есть запись в техническую таблицу `processed_tables`, необходимо защитить данный этап блокировкой, чтобы предотвратить параллельную запись, а следовательно, и параллельную обработку, возможно пересекающихся, батчей.

На данном этапе определяется, будет ли исполняться граф или пропустится. А также определяется список таблиц, попавших в батч на обработку. Добавляя информацию о новом батче в таблицу `isProcessed`, мы гарантируем, что только один граф может работать с таблицами истории.

Таким образом, если таблицы истории используются каким-либо графом, то граф завершит своё исполнение, предварительно освободив таблицу `lock_table`.

2. Обновление таблицы порогов для правила определения пользователей с аномальными счетчиками событий.

На данном этапе проверяется наличие таблицы `fat_thresholds_new` и, если таблица `fat_thresholds` никаким процессом не используется, то перезаписывает данные из `fat_thresholds_new` в `fat_thresholds`, удаляя версию `fat_thresholds_new`. Сделано это для поддержания асинхронности работы с таблицей порогов.

3. На данном этапе происходит объединение всех таблиц батча в одну, по которой в последующем происходит агрегация по пользователям или событиям. Данные агрегаты в последующем будут использоваться в этапах 5 и 7.

4. Параллельно с выполнением этапа 3, происходит фильтрация таблиц истории агрегатов от событий, которым больше суток. Так же на данном этапе производится агрегация таблиц историй по пользователям и событиям, чтобы ускорить объединение истории с данными из свежего батча.

Результатом данного этапа является таблица уникальных пользователей с их словарем событий за сутки, а также таблица уникальных событий с их минимальным временем появления в логе за последние сутки. Данные таблицы будут использоваться в этапе 7.

5. После того, как этапы 3 и 4 завершились, в таблицах истории находятся актуальные данные, но неполные. Поэтому на данном этапе в историю дозаписываются данные из этапа 3.
6. После того, как все таблицы истории актуализированы. У таблиц из текущего батча в таблице `processed_tables` значение `isProcessed` меняется на `True`, таким образом происходит разблокировка работы для остальных графов.
7. На данном этапе производится подготовка списка найденных пользователей с аномальными счетчиками, объединение старой истории из этапа 4 со свежими данными из этапа 3. Также к таблице, содержащей

все события батча, добавляются ключи для дальнейшего join в 8 этапе, а именно хэш информации о пользователе и хэш информации о событии. Во время расчета ключей так же применяются правила, которые не требуют данных об истории.

8. Самый долгий этап исполнения графа, производится тяжелый join, а также непосредственно запись всех сработавших правил в колонку rules в виде списка идентификаторов правил. Результатом данного этапа является таблица, содержащая в себе все события из обрабатываемого лога, где для каждой строки дополнительно добавлены названия таблиц пятиминуток, из которых эти события были взяты, и колонка rules.
9. Финальная часть графа, в которой производится запись в stream-директорию. Все таблицы записываются без колонки, содержащей информацию о том, из какой таблицы пятиминутки было событие. Раскладка событий для таблиц stream/5min происходит по ранее сохраненным названиям таблиц пятиминуток. После чего обновляется таблица мониторинга delays. На рисунке 7 представлен график, который в автоматическом режиме строится по таблице мониторинга delays.

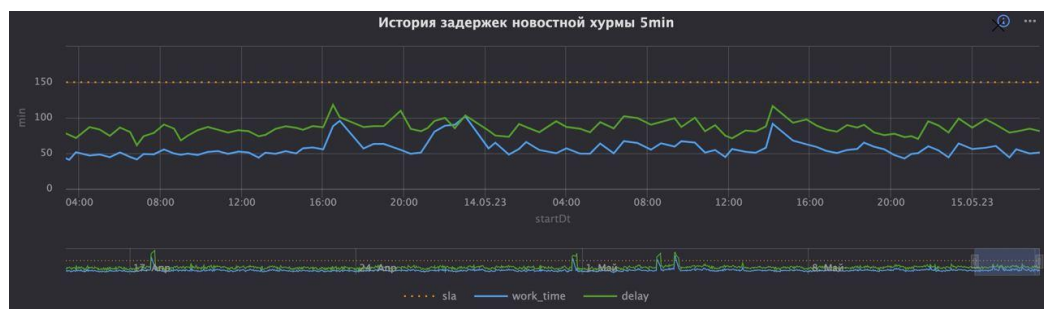


Рисунок 7 — Мониторинг поставок таблиц stream/5min

Если для таблиц stream/30min и stream/1d не было в stream-директории таблиц, в которую должна происходить запись, то создаются новые пустые таблицы с правильной схемой. После чего производится запись в данные таблицы, каждая строка попадает в ту таблицу, которая соответствует её timestamp события.

3.3 Table Merger

Вспомогательный граф в архитектуре, реализующий поставку таблиц, идентичную по формату тому, как это делает logfeller. Функции данного графа:

- Нахождение готовых таблиц в stream/30min или stream/1d, в зависимости от параметра графа. Данные таблицы перемещаются из stream-директории в место постоянного хранения.
- Информация о некорректных таблицах отправляется в чат алертов в Telegram.
- Обновление таблицы мониторинга задержки поставок 30min и 1d таблиц.

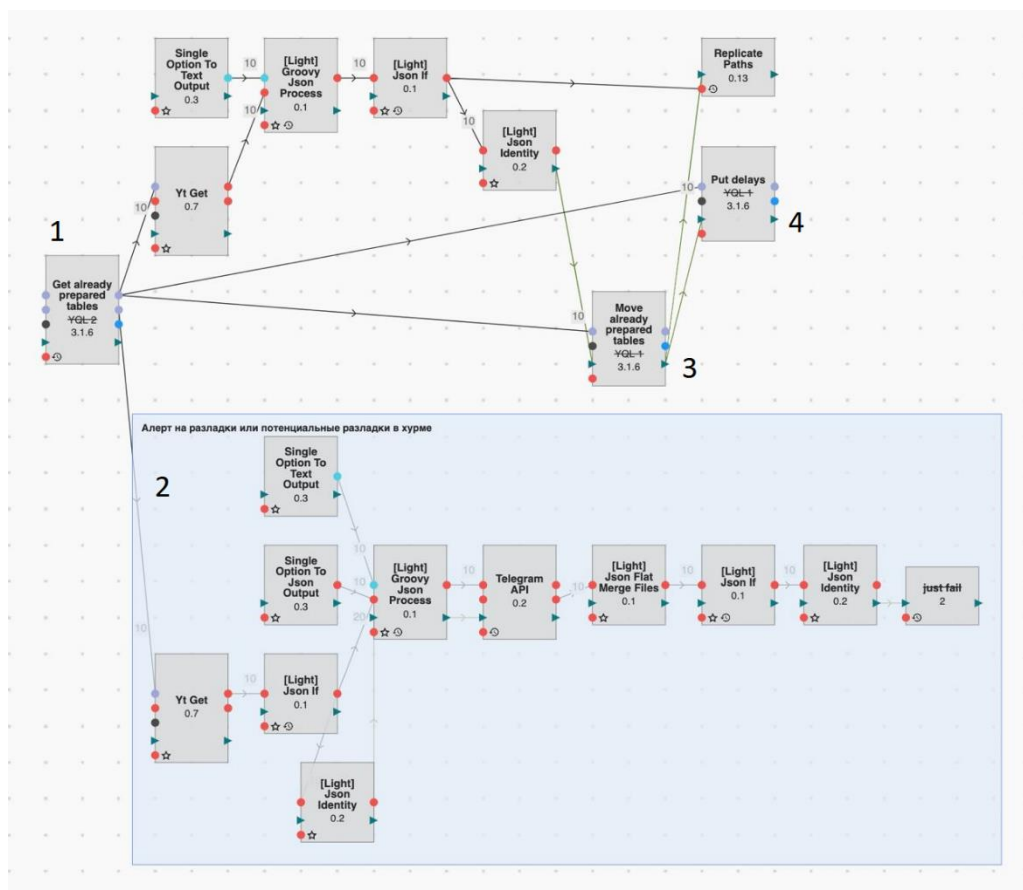


Рисунок 8 – Реализация графа Table Merger

На рисунке 8 представлен реализованный граф с пронумерованными этапами. В данном графе на 1 этапе берутся таблицы из соответствующей параметрам

запуска графа stream-директории (stream/30min или stream/1d), которые не обновлялись больше часа. Такие таблицы считаются готовыми к переносу. Однако готовые таблицы необходимо проверить на разладки по числу строк:

Пусть tableType - тип таблицы 30min/1d, tableName - название таблицы, тогда:

S - количество строк в таблице stream/tableType/tableName.

M - количество строк в таблице tableType/tableName, таблица существует, иначе 0.

U - количество строк в таблице tableType/tableName из непроверенного лога.

Разладкой называется ситуация, когда $\left| \frac{S+M}{U} - 1 \right| > 1^{-4}$.

Если на 1 этапе были найдены таблицы с разладками, то на 2 этапе формируется и отправляется алерт в Telegram. Пример алерта представлен на рисунке 9.

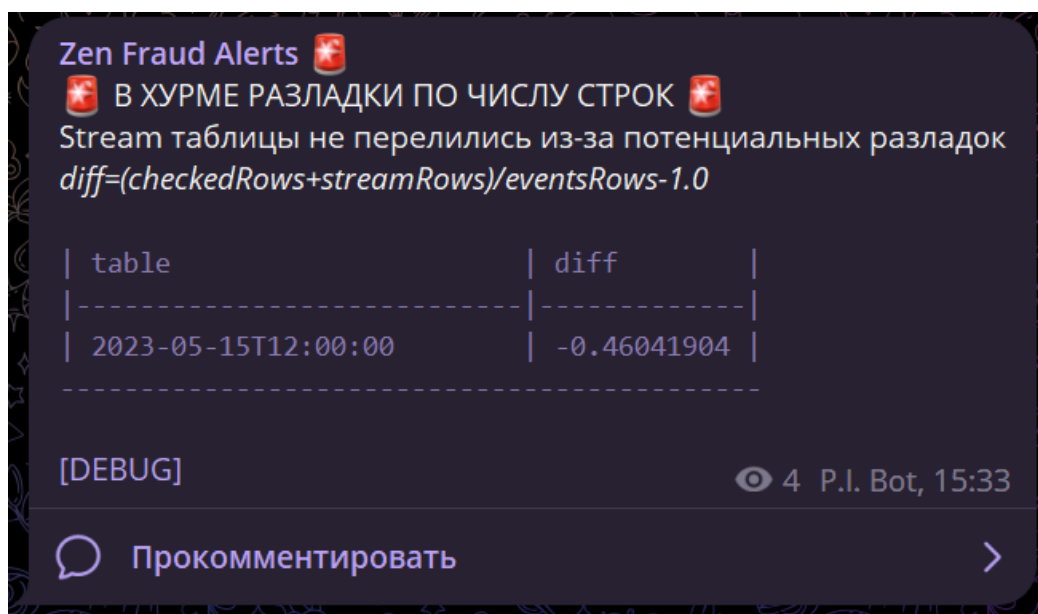


Рисунок 9 – Алерт о потенциальных разладках

На 3 этапе берутся готовые таблицы без разладок из 1 этапа, и данные из этих таблиц переносятся из stream-директории в место постоянного хранения. Если

же таблица с таким названием уже существует в месте постоянного хранения, то происходит дозапись.

На 4 этапе обновляется таблица мониторинга работы графа Table Merger merger_delays. На рисунке 10 представлен график, который в автоматическом режиме строится по таблице мониторинга merger_delays для дневных таблиц. late_events_delay определяется как максимальная задержка дозаписи в таблицу.



Рисунок 10 – Мониторинг поставки дневных таблиц

3.4 Update Fat Thresholds

Данный граф реализует обсчет порогов для определения аномальных по количеству событий пользователей.

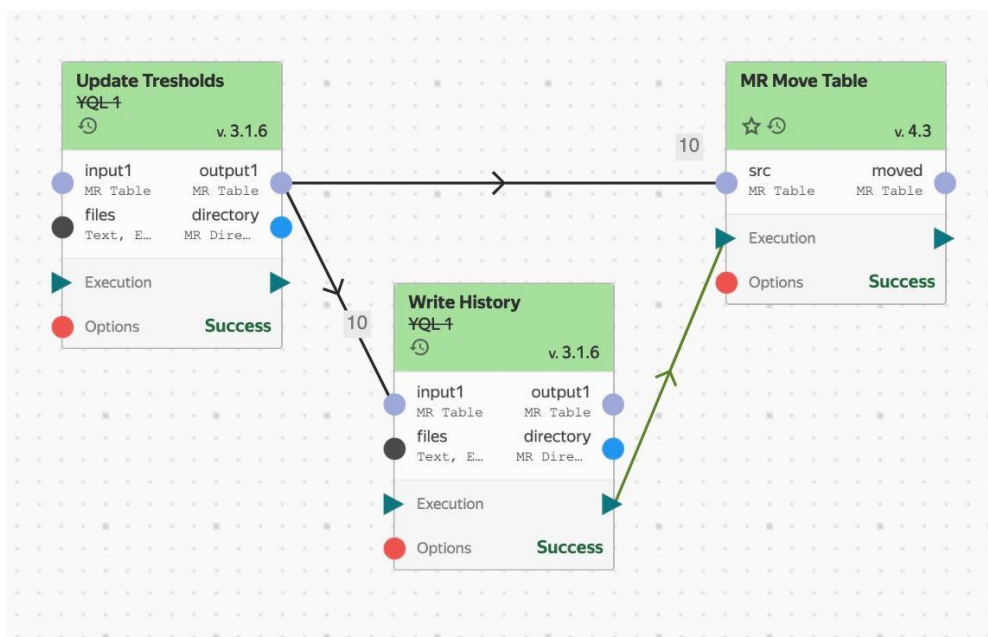


Рисунок 11 – Реализация графа Update Fat Thresholds

На рисунке 11 представлен реализованный граф, состоящий из кубиков расчета порогов, записи в историю таблиц, а также записи новых посчитанных порогов в таблицу `fat_thresholds_new`.



Рисунок 12 – История порогов для аномальных пользователей

График, который строится по таблице истории порогов `fat_thresholds_history`, изображен на рисунке 12. По данному графику можно косвенно отслеживать активность роботов на платформе.

3.5 Alerts

Данный граф является одним из ключевых сигналов о поломках системы. Он считает для каждой таблицы, появившейся за последние сутки в `zen-news-events-log`, для которой нет таблицы с таким же названием в `zen-news-events-checked-log`, сколько времени прошло с момента поставки таблицы до текущего момента, это время называется задержкой. Если задержка превышает 4 часа, то граф отправляет алерт в чат алертов в Telegram. Пример алерта представлен на рисунке 13.

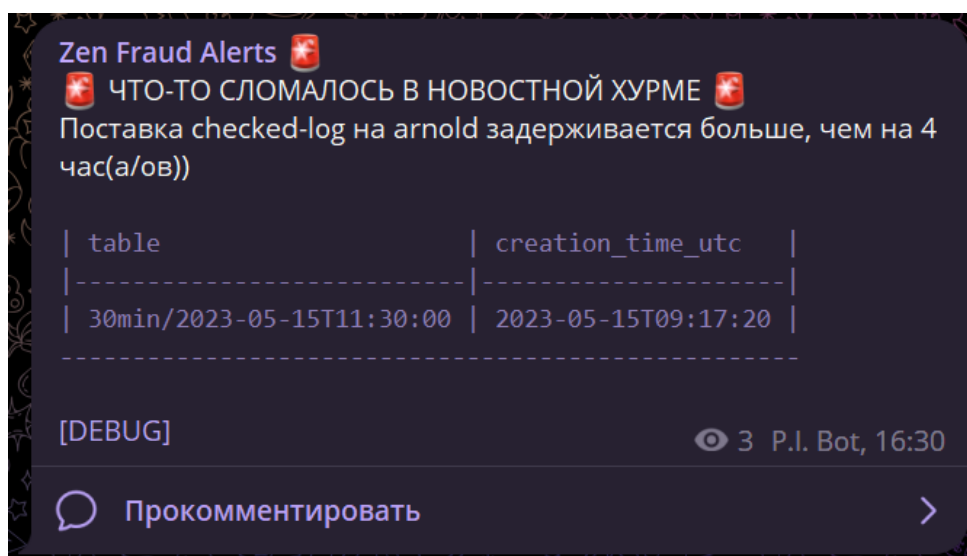


Рисунок 13 – Пример алерта о задержках поставки таблицы

3.6 Hitman

Для всей системы был создан проект в Hitman, содержащий отдельный процесс под каждый граф.

Для каждого процесса были настроены мониторинги:

- Число успешно завершенных запусков за час должно быть не меньше одного;
- Не должно быть графов, завершившихся с ошибкой;
- Верхняя граница времени работы каждого графа настраивалась для каждого процесса отдельно.

Если мониторинг заметил отклонение от одного из правил – он отправляет алерт в виде SMS, так же дублируя алерт на корпоративную почту. Пример алертов представлен на рисунке 14.

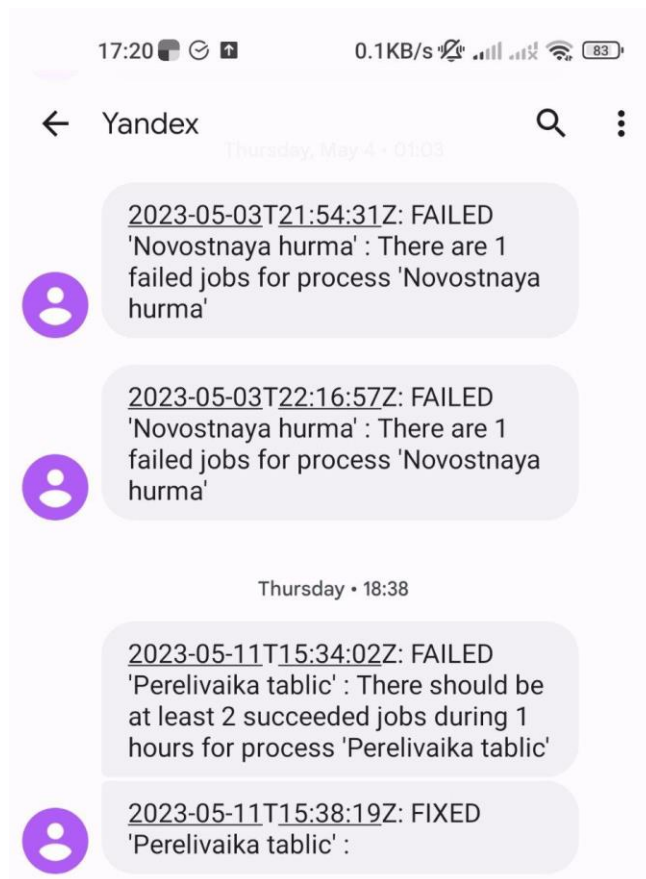


Рисунок 14 – SMS алерты от мониторинга Hitman

Заключение

В ходе выполнения выпускной квалификационной работы была разработана архитектура системы онлайн-антифрода для сервиса Новостей, а также реализована данная архитектура с использованием графов Nirvana, настроены мониторинги и построены графики для отслеживания состояния системы.

Таблицы, поставляемые онлайн-антифродом активно используются аналитиками Новостей и Дзена.

Благодаря настроенным мониторингам и алертам обо всех неполадках в системе команда антифрода узнает своевременно.

Все функциональные и нефункциональные требования исполнены в полном объеме.

Список используемой литературы

1. Yson [Электронный ресурс] – URL: <https://ydb.tech/ru/docs/yql/reference/udf/list/yson> (дата обращения 05.05.2023)
2. YQL [Электронный ресурс] – URL: <https://ydb.tech/ru/docs/yql/reference/> (дата обращения 05.05.2023)
3. YT [Электронный ресурс] – URL: <https://habr.com/ru/companies/yandex/articles/311104/> (дата обращения 05.05.2023)
4. Datalens [Электронный ресурс] – URL: <https://cloud.yandex.ru/docs/datalens/> (дата обращения 05.05.2023)
5. Дзен Новости [Электронный ресурс] – URL: <https://dzen.ru/news> (дата обращения 10.05.2023)
6. SLA (соглашение о уровне услуг) [Электронный ресурс] – URL: https://ru.wikipedia.org/wiki/Соглашение_об_уровне_услуг (дата обращения 10.05.2023)