

МИНОБРНАУКИ РОССИИ
ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И АНАЛИЗА ДАННЫХ
КАФЕДРА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СИСТЕМ

ОТЧЕТ ПО УЧЕБНОЙ ТЕХНОЛОГИЧЕСКОЙ (ПРОЕКТНО- ТЕХНОЛОГИЧЕСКОЙ) ПРАКТИКЕ

Студент
гр. БИС-22-01



Л.А. Богданович

Руководитель,
заведующий кафедрой,
канд. экон. наук, доцент



Е.В. Кийкова

Владивосток 2025

Содержание

Введение.....	3
1 Характеристика предприятия ООО «ФарПост»	4
1.1 Общая структура ООО «ФарПост».....	4
1.2 Организационная структура ООО «ФарПост»	5
1.3 Перечень основных направлений деятельности компании	7
2 Состояние информационной структуры ООО «ФарПост»	9
2.1 Информационный и аппаратно-технический комплекс предприятия.....	9
2.2 ИТ-архитектура отдела данных	10
2.3 Описание процессов отдела данных и уровень их автоматизации.....	12
2.4 Описание процесса разработки агрегации по сбору и обработке данных	13
3 Обоснование выбора технологии проектирования на основе анализа существующих разработок	15
3.1 Выбор инструментария для автоматизации выделения зависимостей.....	15
4 Техническое задание на автоматизацию бизнес процесса	17
4.1 Общие сведения об автоматизации.....	17
4.2 Цели и задачи проекта автоматизации	17
4.3 Требования к информационной системе	18
Заключение	21
Список использованных источников	22
Приложение А.....	23

Введение

В современном мире, цифровизация – обязательная часть успеха и развития предприятий. В таких условиях растет необходимость в молодых специалистах в сфере информационных технологий, прошедших практическую подготовку. Реализация IT-проектов требует не только глубокого понимания теоретических основ, но и навыков успешно применять эти знания и внедрять новые технологии в бизнес. Одну из ключевых ролей в развитии профессиональных навыков будущих специалистов играет практика. Она позволяет получить опыт работы на предприятии, научиться решать реальные задачи и ознакомиться со структурой информационной системы предприятия.

В данном отчете описаны цели и задачи практики в компании ООО «ФарПост». Целью учебной технологической практики является получение компетенций в области разработки информационных систем. Для выполнения практики были поставлены следующие задачи:

- 1) Изучение организационной структуры и деятельности предприятия.
- 2) Анализ существующей информационной структуры компании и проектов внутри нее.
- 3) Разработка проекта информационной системы с составлением технического задания на ее реализацию.
- 4) Приобретение практических навыков и компетенций в решении рабочих задач и закрепление теоретических основ языка программирования Python на основе задания по обновлению стратегий верификации данных.
- 5) Изучение баз данных предприятия и агрегационных функций для нее, во время выполнения задания по составлению диаграмм агрегаций, создание новых функций.

Данные задачи позволят легко и эффективно достичь цели получения компетенций в области разработки информационных систем.

1 Характеристика предприятия ООО «ФарПост»

1.1 Общая структура ООО «ФарПост»

ООО «ФарПост» – Одна из ведущих IT компаний на дальнем востоке. Основной вид деятельности компании – разработка и развитие программного обеспечения, публичных сайтов и приложений. Каждый из проектов является своей компанией, принадлежащей ООО «ФарПост». Большинство проектов разрабатываются в одном офисе, поэтому информационно сильно связаны. Компания занимает 1 место в отрасли по Приморскому краю [1].

ООО «ФарПост» существует с 1997 года. В компании созданы и поддерживаются такие проекты как:

- 1) Фарпост;
- 2) Дром;
- 3) VL.ru;
- 4) DVHUB.ru;
- 5) ЛовиКупон.

Бизнес-цели компании:

- 1) Оставаться номером один там, где мы номер один;
- 2) Стать номером один там, где мы не номер один;
- 3) Растить показатели ARPU и ARPPU.

Миссия компании ООО «ФарПост»: «Мы здесь для того, чтобы делать жизнь людей более комфортной, интересно и насыщенной за счет использования информационных технологий в решении задач».

1.2 Организационная структура ООО «ФарПост»

На рисунке 1 представлена организационная структура компании ООО «ФарПост»:

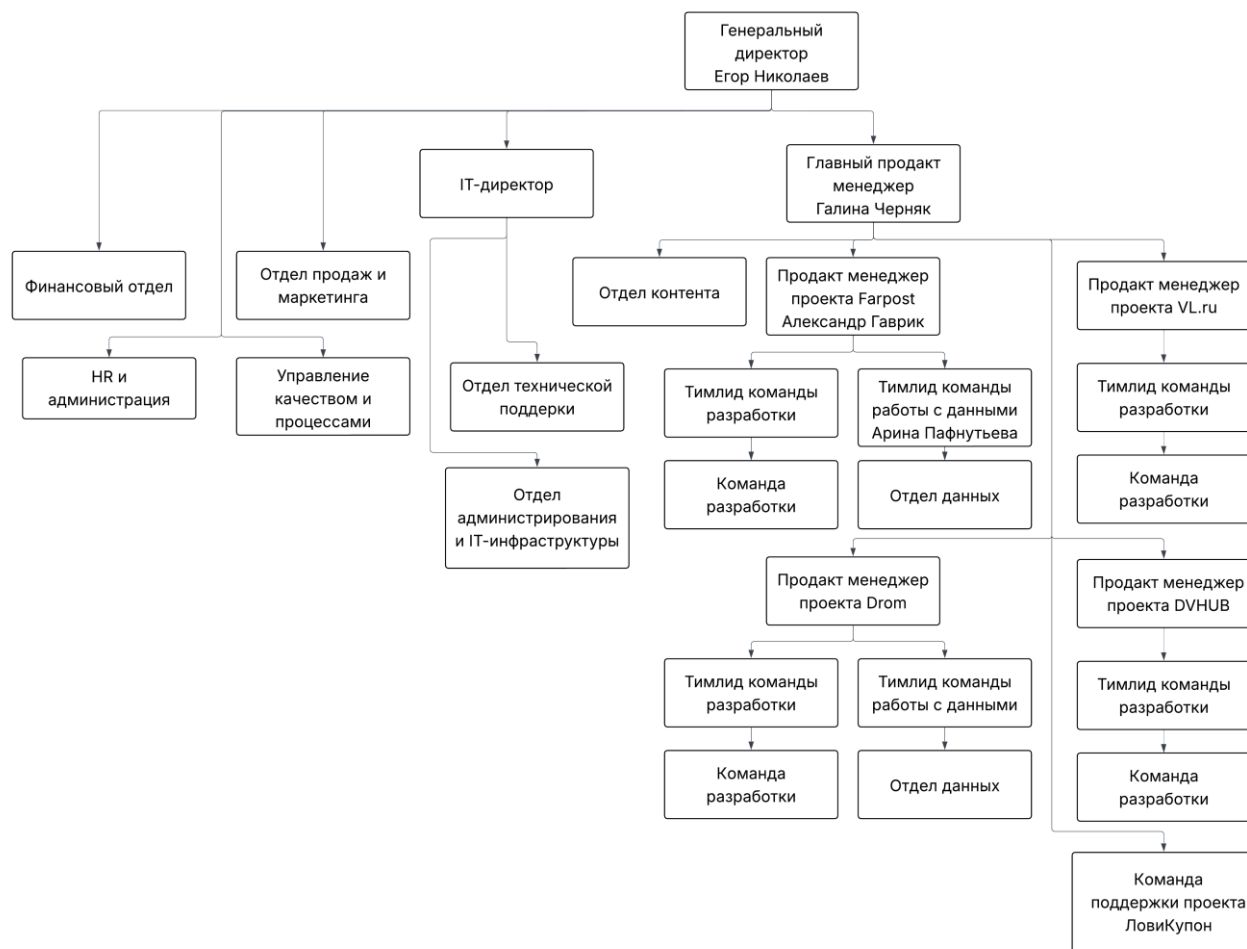


Рисунок 1 – Организационная структура компании ООО «ФарПост»

Организационная структура компании ООО «ФарПост» состоит из:

Топ-управления:

- 1) Генеральный директор. Отвечает за общее оправление компанией, а также ее стратегическое развитие.
- 2) ИТ-директор. Отвечает за техническое развитие компании, выбор технологий и их развитие.
- 3) Главный продуктовый менеджер. Отвечает за развитие проектов компании.

У каждого проекта есть свой продуктовый менеджер, который отвечает за развитие продукта и команда разработки. Для проектов Farpost и Drom, которые являются самыми большими, есть выделенные отделы данных. У проекта ЛовиКупон есть только отдел поддержки, так как проект маленький и на данный момент не развивается.

В каждой команде разработки есть:

1) Технический директор разработки или тимлид. Отвечает за техническое руководство отдела разработки.

2) Разработчики. Занимаются разработкой, поддержкой и развитием проектов.

3) Дизайнеры. Занимаются созданием пользовательского интерфейса и пользовательского опыта.

4) QA инженеры. Отвечают за тестирование разработанных функций и обеспечение качества продукта.

5) SEO-специалисты. Отвечают за продвижение сайтов в поисковых системах.

6) Аналитики. Занимаются подтверждением или опровержением гипотез, выдвинутых продуктовыми менеджерами.

В самых крупных проектах, как Farpost и Drom, есть свои отделы работы с данными. Их структура состоит из:

1) Технический директор разработки или тимлид. Отвечает за техническое руководство отдела разработки.

2) Инженеры данных. Занимаются разработкой агрегаций, баз данных и настройкой инструментов для управления ими. Также занимаются настройкой выгрузки данных с продуктовых баз данных, в базы данных отдела,

3) Разработчики данных, занимаются разработкой сайта Farlist и выгрузки данных в Metabase.

4) Аналитики. Занимаются подтверждением или опровержением гипотез, выдвинутых продуктовыми менеджерами на основе собранных в отделе данных инженерами данных.

Отдел технической поддержки:

1) Технический директор поддержки. Отвечает за организацию работы и эффективность технической поддержки пользователей.

2) Специалисты по поддержке. Занимаются решением технических вопросов и проблем пользователей сайтов.

Отдел продаж и маркетинга:

1) Менеджеры по продажам. Отвечают за продажи услуг внутри проектов.

2) Маркетологи. Занимаются продвижением сайтов, привлечением клиентов и управлением маркетинговыми кампаниями.

Отдел контента:

1) Руководитель отдела контента. Координирует работу сотрудников, занимается распределением задач.

2) Менеджеры контента. Отвечают за проработку контента и разделов.

3) Модераторы контента. Отвечают за юридическую чистоту контента, поскольку пользователи размещают на сайте свой собственный контент.

Отдел администрирования и IT-инфраструктуры:

1) Администраторы систем. Отвечают за управление и поддержку серверной и сетевой инфраструктуры компании.

2) Специалисты по безопасности информации. Отвечают за обеспечение безопасности данных и сетевых ресурсов компании.

Управление качеством и процессами:

1) Менеджер по управлению качеством. Отвечает за управление процессами и стандартами качества в разработке и поддержке сайтов.

2) SCRUM-мастер. Отвечает за развитие и модернизацию системы управления процессами в компании.

Финансовый отдел:

1) Финансовый директор. Отвечает за финансовое планирование, бюджетирование и управление финансами компании.

2) Бухгалтерия.

HR и администрация:

1) HR-менеджер. Отвечает за обучение сотрудников, проведение собеседований и помощь сотрудникам по любым организационным вопросам.

2) Кадровый специалист. Отвечает за найм и увольнение сотрудников.

3) Административный персонал. Занимается организационными вопросами, обеспечением офиса и делопроизводством.

Такая структура позволяет эффективно управлять процессами внутри компании и обеспечивать эффективную разработку проектов.

1.3 Перечень основных направлений деятельности компании

Основное направление деятельности компании – развитие программного обеспечения для IT-проектов. В основные проекты включают: Farpost, Drom, ЛовиКупон, VL.RU, DVHUB.RU. Кроме основных проектов, также существуют вспомогательные внутренние проекты различной направленности. В отделе данных к примеру, разрабатывается свой аналитический сервис для внутреннего пользования.

В течении практики из была изучена деятельность проекта Farpost. Среди основных направлений деятельности компании в рамках этого проекта можно выделить:

1) Поддержку и развитие сайта. Повышение надежности, скорости работы и удобства платформы.

2) Развитие вертикали «Авто и мото». Увеличение числа объявлений и трафика в сегменте.

3) Развитие вертикали «Недвижимость» и привлечение крупных клиентов и аудитории из других регионов.

4) Развитие вертикали «Услуги». Повышение доверия к исполнителям и рост числа заказов.

5) Развитие вертикали «Работа». Упрощение поиска вакансий и повышение конверсии.

6) Развитие вертикали «Товары» путем улучшения разделов, алгоритмов выдачи результатов и привлечения новых крупных продавцов.

Такие направления развития компании позволяют эффективно увеличивать бизнес показатели компании, и выполнять стратегические требования высшего руководства.

2 Состояние информационной структуры ООО «ФарПост»

2.1 Информационный и аппаратно-технический комплекс предприятия

Для разработки большого количества программного обеспечения, на предприятии используется большой список рабочих станций, как личных, так и корпоративных, выбор которых зависит от запросов и личных предпочтений сотрудников. Компания также полностью обеспечивает компьютерной периферией.

В офисе настроена локальная корпоративная сеть, с высокой степенью безопасности, построенной на разграничении доступов. Безопасность данных и ресурсов играет большую роль в компании, поэтому этой сфере выделено большое внимание, проводится постоянное обновление систем и также проводится регулярный мониторинг и проверка на угрозы. В этой сети есть доступ к серверам, базам данных, хранилищам данных и другой инфраструктуре компании. Управлением доступности и безопасностью информационной инфраструктуры занимаются администраторы этой сети. Получение любого доступа к любому ресурсу компании проходит через них. Каждая рабочая станция в офисе внесена в реестр и привязана к отдельному сотруднику, поэтому можно точно распределять доступы и выявлять нарушителей.

Для увеличения эффективности работы, ускорения обучения команды и проверки качества кода и программ внедрены технологии искусственного интеллекта. Компания выделила доступ к большому количеству инструментов, а также регулярно проводит тренинги и лекции по теме использования его в целях выполнения бизнес требований и повышения бизнес показателей компании.

Коммуникация выполняется в специализированном мессенджере Mattermost с высокой степенью защиты и шифрования, что обеспечивает высокую защиту переписок от утечек. Корпоративный мессенджер удобно настроен для общения как в личном формате, так и в формате групп, команд и каналов. Кроме того, мессенджер расположен на серверах компании, поэтому получить информацию из вне крайне тяжело. Эти факторы обеспечивают высокую эффективность и безопасность коммуникации в компании.

Кроме корпоративного мессенджера для коммуникации используются и другие приложения. В компании настроена корпоративная почта, которая используется в некоторых вопросах внутренней коммуникации. Для проведения видео встреч и конференций, а также любых встреч, где часть сотрудников находится в офисе, а часть сотрудников работают удаленно, используется приложение Zoom. Для него в каждой переговорной комнате установлены камеры, микрофоны и колонки. Это убирает любые преграды перед

коммуникацией между сотрудниками. Для каких-то интерактивных взаимодействий используется интерактивная доска Miro.

Для управления бизнес процессами компания использует сервис Jiro. Это сервис, позволяющий управлять бизнес процессы по методологии Agile. Все задачи распределены по отделам, в которых в свою очередь есть своя доска Kanban на которой висят все задачи по их статусу и сотруднику, а также проводятся спринты. Данный сервис и подход позволяет более эффективно выполнять работу и также выделять возможные проблемы в распределении и выполнении задач в команде.

Вся кодовая база находится в репозиториях GitLab, этот сервис позволяет настраивать контроль версий кода, проводить проверки кода, а также позволяет работать над кодом одновременно большому количеству разработчиков. Также в этом сервисе есть возможность настроить CI/CD – быструю сборку и развертывание кода на серверах. Это позволяет эффективно работать с кодом, а также быстро доставлять новое программное обеспечение. Также в этом сервисе обеспечена высокая степень безопасности, посредством шифрования данных. Кроме того, этот сервис находится на серверах компании и доступен только внутри корпоративной сети, что снижает риск получить данные кому-то из вне [2].

Все информационные ресурсы компании находятся на выделенных серверах в центре обработки информации. На серверах компании находятся все сайты, базы данных, внутренние инструменты, разработки. Также там находятся все сервисы, обеспечивающие безопасность в компании.

2.2 ИТ-архитектура отдела данных

У отдела данных есть своя информационная система с различными технологиями и инструментами для ее работы. В перечень различного программного обеспечения, которое использует отдел, входят:

1) Python – основной язык программирования, который используют в отделе. Язык применяется в широком спектре задач, от сбора и обработки данных, а также написания вспомогательных приложений, заканчивая анализом данных и получения выводов для последующего принятия бизнес решений на основе них.

2) Apache Airflow – открытая платформа для создания, выполнения, отслеживания и управления операциями по обработке данных. С помощью Python и SQL запросов, Airflow получает, проводит первичную обработку и загружает в удобные таблицы данные с помощью DAG'ов – ориентированных ациклических графов, выполняющих задачи в коде друг за другом с учетом зависимостей между ними [3].

3) Clickhouse – это высокопроизводительная, столбцовая система управления базами данных (СУБД) SQL для онлайн аналитической обработки [4]. С помощью этой СУБД хранятся практически все данные, которые используются в отделе.

4) React – это JavaScript-библиотека для создания пользовательских интерфейсов. Выбрана из-за своей гибкости, модульности и расширяемости. Используется для разработки сайта Farlist [5].

5) Mysql – система управления базами данных. Также используется для хранения данных, которые используются в отделе [6].

6) Draw.io и Lucidchart – программное обеспечение, используемое для рисования графиков и диаграмм. Используется для создания документации, в основном для графиков различных процессов и агрегаций, для визуальной наглядности и простоты понимания кода, на основе которого нарисованы графики.

7) Typescript – язык программирования, использующийся для созданий сайтов. Используется для разработки сайта Farlist.

8) Debezium – программное обеспечение категории CDC, то есть Capture Data Change. Также его можно охарактеризовать как набор коннекторов для различных СУБД, используется для чтения данных из баз данных в режиме реального времени, и выгрузки их в базы данных отдела. Также используется для записи логов [7].

9) Kafka – распределенная система обмена сообщениями между серверными приложениями в режиме реального времени. Используется в связке с Debezium [8].

10) Docker – это платформа с открытым исходным кодом для автоматизации разработки, доставки и развертывания приложений. Ее основная идея — создание стандартного и предсказуемого окружения, где приложения могут работать независимо от операционной системы или инфраструктуры. Используется для развертывания других сервисов и программного обеспечения в информационной системе отдела [9].

11) Kubernetes – это портативная расширяемая платформа с открытым исходным кодом для управления контейнеризованными рабочими нагрузками и сервисами, которая облегчает как декларативную настройку, так и автоматизацию [10]. Используется для развертывания программного обеспечения на всю локальную сеть отдела. С помощью Kubernetes, например, работает Apache Airflow, S3, множество worker'ов которые выполняют вычислительные задачи для других приложений. Kubernetes развернут на сервере отдела.

12) Visual studio – редактор кода для различных языков программирования. Используется частью отдела.

13) Pycharm – специализированная среда программирования для языка Python. Используется частью отдела.

14) S3 – хранилище данных без иерархии. Используется для хранения логов из различных источников, в основном подов Kubernetes, также используется для хранения изображений из баз данных.

Также в отделе есть список инструментов и сервисов собственной разработки. Чаще всего это различные веб приложения с интерфейсом REST API. Написаны они на Python с использованием библиотек Django или FastAPI.

Для стандартизации и поддержания чистоты кода в отделе используется несколько инструментов. Среди них Black, Ruff и basedpyright, программное обеспечение, помогающее следить и исправлять ошибки, связанные с чистотой кода. У отдела есть набор руководств по стилю кода для разных языков программирования.

Отдел данных эффективно использует современные технологии и программное обеспечение для решения задач по сбору, обработке и анализу данных, на основе чего менеджеры могут принять правильные бизнес решения. Широкий спектр применяемых технологий позволяет отделу эффективно выполнять свои цели.

2.3 Описание процессов отдела данных и уровень их автоматизации

Отдел данных в компании ООО «ФарПост» осуществляет несколько ключевых бизнес-процессов:

1) Разработка внутреннего инструмента для просмотра и аналитики данных собранных с различных проектов компании Farlist. Этот процесс включает в себя разработку сайта, в котором собраны различные отчеты, графики и диаграммы. Также разработчики добавляют разные параметры и фильтры, чтобы данные можно было более подробно рассмотреть.

2) Разработка, обновление и поддержка агрегаций для сбора данных. Этот процесс включает в себя написание скриптов для сбора, первичной обработки и загрузки данных во внутренние базы данных отдела.

3) Анализ данных. В этот процесс входит выставление гипотез и их проверка путем анализа данных собранных из различных источников и A/B тестирования. Это тестирование включает в себя изменения некоторых функций сайта, и предоставление новых функций ограниченной группе пользователей, чтобы посмотреть, как повлияли изменения на пользователей.

4) Улучшение качества данных. Включает в себя создание программного обеспечения, скриптов или добавление новых сервисов в существующую информационную систему. Также в этот процесс входит коммуникация с отделом разработки по теме предоставляемых ими данных.

5) Управление базами данных для нужд отдела. Включает в себя создание архитектуры баз данных, таблиц, связей между ними, используемых внутри отдела на этапах сбора и обработки данных, а также коммуникация с аналитиками, о том в каком виде должны быть предоставлены данные.

6) Написание парсеров данных. Нужно для получения данных из внешних источников. Эти парсеры создаются отдельными выделенными для этого процесса сотрудниками, и встраиваются в общую систему для дополнения и улучшения качества данных, а также для анализа конкурентов.

Каждый из процессов играет важную роль в работе отдела данных компании, обеспечивая качественную сбор и обработку данных, мониторинг информации, а также аналитику на основе которой бизнес может принимать правильные и эффективные решения для своего роста и успеха.

В ходе своей практики мне удалось ознакомиться с процессом разработки агрегаций для сбора и обработка данных.

2.4 Описание процесса разработки агрегации по сбору и обработке данных

Процесс разработки агрегации – один из самых важных в отделе, так как на основе данных полученных с агрегаций строятся все графики и диаграммы, а также проводится анализ данных, от выводов которого принимают бизнес решения. Подробная схема данного бизнес-процесса в нотации BPMN представлена в приложении А на рисунке А.1. Процесс разработки агрегации только частично автоматизирована, так как разработка каждой отдельной агрегации может сильно отличаться. Разработка агрегации по сбору и первичной обработки данных состоит из следующих этапов:

- 1) Получение запроса на анализ каких-либо данных и обсуждение технического задания.
- 2) Подготовка нужных для агрегации и анализа данных.
- 3) Разработка агрегации с последующей выгрузкой в Airflow.
- 4) Разработка и настройка отчетов в Metabase и Farlist.
- 5) Анализ данных и отправка результата.

В ходе практики был детально разобран этап с подготовкой нужных для агрегации и анализа данных. Он состоит из выделения нужных таблиц и данных, и последующее выявление зависимостей между таблицами и базами данных. Первая часть этапа уже автоматизирована с помощью существующих процессов и мероприятий для совещаний. Но вторая часть с выявлением зависимостей нет.

При выявлении зависимостей и связей между таблицами и базами данных может уходить много времени, нужно много делать действий вручную, каждый раз долго сидеть в документации, где могут быть неактуальные данные. Но для этого этапа не нужны никакие уникальные действия. Все таблицы и зависимости находятся в одной СУБД, реже в двух, и информация о том, что с чем связана также доступна, что показывает, что возможно автоматизировать данный этап процесса, а вместе с ним централизовать все данные о таблицах и базах данных, а также актуализировать их. Данный этап процесса создания агрегаций было принято взять на работу и автоматизировать, чтобы повысить эффективность и скорость процесса создания агрегаций.

3 Обоснование выбора технологии проектирования на основе анализа существующих разработок

3.1 Выбор инструментария для автоматизации выделения зависимостей

Для автоматизации выделения зависимостей и других данных о таблицах, для выделения любых связей между ними, необходимо средство для автоматизации сбора, анализа и отображения связей между элементами данных. Одна из подходящих платформ для этого – OpenMetadata. Это открытая платформа для управления метаданными таблиц и баз данных, поддерживающий широкий спектр функций для обозревания данных в информационных системах, в данном случае в нескольких базах данных и агрегациях, связанных с ними. Одна из полезных функций – анализ линейного следования, управление описанием данных и зависимостей, а также интеграция с различными источниками данных.

Для обоснования выбора платформы OpenMetadata, ее сравнили с похожими инструментами из той же области. Для сравнения был выбран DataHub, как похожий инструмент, который также, как и OpenMetadata является opensource [11]. В таблице 1 приведено сравнение этих платформ:

Таблица 1 – Сравнение платформ

Критерий	OpenMetadata	DataHub
Активность разработки	2800 звезд на Github и 180 участников.	8300 звезд на Github и 416 участников.
Количество источников данных.	Более 50 коннекторов.	Более 50 коннекторов.
Авторизация	Jaas авторизация и OIDC авторизация для Google, Azure и Okta	Поддерживаемые поставщики SSO: Auth0, Azure, Google, Okta, OneLogin, Amazon Cognito SSO, Key Cloak, LDAP Базовая авторизация и настраиваемый OIDC
Способы сохранения и загрузки резервных копий	Рекомендуется делать SQL дампы сервисной базы данных. Для некоторых данных нужны дополнительные действия.	Свой инструмент в консоли для сохранения и загрузки резервных копий всех важных данных.

Критерий	OpenMetadata	DataHub
История данных	История схемы данных и API для временной линии со всеми изменениями.	Хранит всю историю изменений в базе данных для всех объектов, для отображения изменений.
Документация	Возможность создать документацию с сохранением версий.	Возможность создать очень детальную документацию с поддержкой тегов и версий.
Система оповещений	SaaS решение позволяет отправлять оповещения в Slack и MS Teams. Возможно создать свои решения.	Настраиваемые коннекторы, REST API для любой автоматизации.
Расширяемость	Настраиваемые коннекторы, расширяемая модель метаданных, OpenAPI REST API.	Настраиваемые коннекторы, REST API для любой автоматизации.
Виды метаданных	Теги, владельцы, термины глоссария, продукты данных, домены.	Теги, уровни, владельцы, термины глоссария, предложения
Языки программирования	Python и Java	Python и Java
Поиск и исследуемость	Исследуемость – точка входа для Datahub, Поиск может быть широко настроен, но иногда он работает странно.	Хорошо работающая функция поиска. Отдельная функция изучения.

Хоть OpenMetadata и более молодой инструмент, но он имеет более проработанные инструменты для расширения платформы, и не уступает по остальным функциям Datahub. OpenMetadata отлично подходит отделу по своему функционалу, а также отвечает требованиям конкурентоспособности.

4 Техническое задание на автоматизацию бизнес процесса

4.1 Общие сведения об автоматизации

Предлагаемая автоматизация представляется собой внедрение платформы и инструментов для ее работы, которая собирает информацию о всех таблицах, базах данных и агрегациях, связях между ними и документацией, для того чтобы можно было быстро, автоматизировано и централизованно найти нужную информацию о связях и зависимостях. Данная платформа изначально не имеет никаких интеграций и функций для связи с нынешними системами в отделе. Кроме того, для работы платформы она должна быть правильно настроена и запущена в существующей ИТ-архитектуре отдела. Интерфейс реализован в виде веб-сайта, через который пользователь может смотреть данные и обновлять их, для которого возможно понадобится расширение функционала.

Требуемая функциональность:

1) Автоматический сбор технических и бизнес-метаданных из различных источников. Из Clickhouse: сбор структуры таблиц (столбцы, типы данных), схем, статистики объема данных, индексов и настроек партиционирования. Из MySQL: Извлечение информации о таблицах, первичных/внешних ключах, триггерах, процедурах и представлениях. Из Apache Airflow: Сбор метаданных о DAG-файлах, задачах (операторах), зависимостях между задачами, логах выполнения и параметрах запуска. Из Агрегации: Анализ предварительно подготовленных агрегатов (например, материализованных представлений или кубов) для отслеживания их источников и логики расчета.

2) Обеспечение целостности и актуальности метаданных. Проверка согласованности связей между таблицами (первичные/внешние ключи) и ETL-процессами. Регулярное сравнение метаданных в OpenMetadata с реальным состоянием источников.

3) Поддержка управления данных и наблюдаемости. Создание бизнес-гlossариев для классификации данных и мониторинг качества данных.

4) Полная документация по всем источникам и таблицам с данными.

Данная функциональность позволит максимально полно автоматизировать данный этап разработки агрегаций.

4.2 Цели и задачи проекта автоматизации

Цели проекта внедрения OpenMetadata в существующую информационную систему:

1) Повышение эффективности управления данными. В эту цель входит создание единого информационного пространства для хранения и управления метаданными, обеспечивающего доступ к актуальной информации о структуре данных, их источниках и зависимостях.

2) Поддержка управления и наблюдаемости, через обеспечение прозрачности данных с помощью визуализации lineage-зависимостей между таблицами, ETL-процессами и агрегациями.

3) Улучшение взаимодействия между командами при помощи централизации документации и бизнес-гlossариев для упрощения поиска и понимания данных аналитиками, разработчиками и менеджерами.

Задачи проекта внедрения OpenMetadata в существующую ИТ-архитектуру отдела:

1) Развертывание пустой платформы в ИТ-архитектуре отдела.

2) Интеграция с источниками данных, с помощью настройка автоматического сбора метаданных из ClickHouse. Также необходима интеграция с MySQL и Airflow.

3) Обеспечение актуальности метаданных через настройку регулярного сканирования источников и реализация событийных триггеров для мгновенного обновления при изменении схем или логики ETL-процессов.

4) Реализация функционала управления и мониторинга через разработку бизнес-гlossариев и тегов для классификации данных.

5) Создание интерфейса для анализа и поиска с фильтрацией таблиц по доменам и визуализации lineage-графов.

6) Обеспечение возможности мониторинга и наблюдения за данными с помощью интеграции с системами оповещения и внедрения метрик качества данных.

7) Написание документации.

Цели и задачи интеграции OpenMetadata направлены на автоматизацию сбора зависимостей и связей между таблицами и базами данных, а также на внедрение инструмента в существующую информационную систему, позволяющую повысить эффективность управления данными, улучшить их мониторинг и наблюдаемость.

4.3 Требования к информационной системе

Объектом автоматизации выступает выявление зависимостей и связей между таблицами и базами данных. Данная автоматизация ускорит процесс создания агрегаций, а также позволит отделу более эффективно следить за данными.

Данная автоматизация должна соответствовать следующим требованиям.

Функциональные требования:

1) Сбор структуры таблиц, типов данных, ключей, партиционирования и статистики объема из ClickHouse и MySQL.

2) Интеграция с Apache Airflow для извлечения мета данных DAG-файлов, задач, зависимостей и логов выполнения.

3) Обработка метаданных агрегаций.

4) Регулярное сканирование источников и мгновенное обновление при изменении схем или DAG-файлов.

5) Механизм валидации соответствия метаданных реальному состоянию источников с уведомлениями о несоответствиях.

6) Визуализация lineage-зависимостей между таблицами, ETL-процессами и агрегациями.

7) Контроль качества данных с помощью проверок.

8) Создание бизнес-гlossариев и тегов для классификации данных по доменам (например, «финансы», «маркетинг»).

9) Структурированное описание источников, таблиц, столбцов и их бизнес-значений.

Нефункциональные требования:

1) Аутентификация через LDAP/SSO и шифрование данных.

2) Ролевая модель доступа с разграничением прав на редактирование и просмотр метаданных.

3) Время поиска и анализа метаданных – менее 2 секунд.

4) Поддержка масштабирования для работы с сотнями источников и тысячами таблиц.

5) Резервное копирование метаданных и автоматическое восстановление после сбоев.

6) Интеграция с системами мониторинга для отслеживания состояния системы.

7) Производительность при параллельной обработке большого объема метаданных, система должна эффективно масштабироваться под нагрузкой.

8) Высокая доступность системы – резервирование компонентов и отказоустойчивость.

9) Логирование всех действий пользователей и системных событий с возможностью аудита.

10) Устойчивость к сетевым разрывам и ошибкам API источников – с последующим повтором попыток и уведомлениями.

11) Разделение ролей и прав доступа для безопасности.

Требования к интерфейсу:

1) Поддержка поиска, по ключевым словам, фильтрации таблиц по доменам и визуализации lineage-графов.

2) Возможность редактирования описаний, добавления тегов и просмотра истории изменений метаданных.

3) Экспорт метаданных в форматах, совместимых с Tableau, Power BI и другими аналитическими платформами.

Требования к документированию

- 1) Архитектурная схема интеграции OpenMetadata с источниками данных.
- 2) Руководство по настройке коннекторов и автоматизации сбора метаданных.
- 3) Пошаговые инструкции по поиску данных, добавлению тегов, анализу lineage и мониторингу качества данных.
- 4) Руководство по использованию сервиса.

Для того, чтобы определить сроки выполнения автоматизации, нужен доступ к некоторому программному обеспечению отдела, а также доступ на просмотр информации в базах данных. Поэтому на данный момент сложно определить сроки, но учитывая сложность интеграции такой большой системы, а также большого количества источников, таблиц и баз данных, можно сказать, что это большой проект на много месяцев.

По результатам выполнения интеграции OpenMetadata для автоматизации выявления зависимостей в базах данных, в ИТ-архитектуре отдела должна появиться платформа, с информацией о всех источниках данных, используемых во всех системах отдела, а также автоматизация сбора этих источников, определение зависимостей и подробной документацией про все источники и данные. Данная автоматизация ускорит процесс поиска зависимостей между таблицами и базами данных, позволит увеличить эффективность и скорость разработки агрегаций, а также позволит улучшить мониторинг и контроль над данными.

Заключение

По итогам прохождения практики в компании ООО «ФарПост», были получены знания о корпоративной ИТ-архитектуре отдела, а также информационных системах внутри нее, и структуре организации, а также навыки решения реальных ИТ-задач. Этот опыт помог понять функционирование больших корпоративных информационных систем, что важно для становления квалифицированным специалистом.

Также во время прохождения практики были выполнены все поставленные задачи и изучены основные бизнес процессы в отделе. Благодаря этим задачам были отточены навыки программирования, а также получены новые практические навыки. Также были получены навыки с взаимодействием с корпоративным программным обеспечением, что дало понимание, как и для чего оно работает.

Кроме того, во время изучения бизнес процесса создания агрегаций, и выполнения задач, связанных с ним, было определено, что можно автоматизировать в этом процессе. Предложенная мной автоматизация позволит более эффективно выполнять задачи, а также иметь больше управления и информации по всем источникам данных.

Таким образом, в ходе прохождения практики был получен опыт, способствующий развитию профессиональных навыков, понимания информационных систем, бизнес процессов и их автоматизации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1) ООО «ФарПост» // rusprofile URL: <https://www.rusprofile.ru/id/1430032> (дата обращения: 03.07.2025).
- 2) GitLab: что это такое и как им пользоваться // Skillbox Media URL: <https://skillbox.ru/media/code/gitlab-cto-eto-takoe-i-kak-im-polzovatsya/> (дата обращения: 05.07.2025).
- 3) Apache Airflow // SkillFactory media URL: <https://blog.skillfactory.ru/glossary/apache-airflow/> (дата обращения: 05.07.2025).
- 4) Что такое ClickHouse? // Clickhouse URL: <https://clickhouse.com/docs/ru/intro> (дата обращения: 05.07.2025).
- 5) Что такое React и как его освоить? // Яндекс Образование URL: <https://education.yandex.ru/journal/cto-takoe-react-i-kak-ego-osvoit> (дата обращения: 05.07.2025).
- 6) Что такое MySQL // RuWeb гибкий хостинг URL: <https://ruweb.net/articles/cto-takoe-mysql> (дата обращения: 05.07.2025).
- 7) Знакомство с Debezium — CDC для Apache Kafka // Хабр URL: <https://habr.com/ru/companies/flant/articles/523510/> (дата обращения: 05.07.2025).
- 8) Apache Kafka // SkillFactory media URL: <https://blog.skillfactory.ru/glossary/kafka-apache/> (дата обращения: 05.07.2025).
- 9) Docker: что это такое и как работать с контейнерами // Академия Selectel URL: <https://selectel.ru/blog/what-is-docker/> (дата обращения: 05.07.2025).
- 10) Что такое Kubernetes // Kubernetes URL: <https://kubernetes.io/ru/docs/concepts/overview/what-is-kubernetes/> (дата обращения: 05.07.2025).
- 11) Data Observability is Key: A Hands-on Comparison of Open Source Data Catalog Tools // inovex URL: <https://www.inovex.de/de/blog/data-observability-is-key-a-hands-on-comparison-of-open-source-data-catalog-tools/> (дата обращения: 07.07.2025).

Приложение А

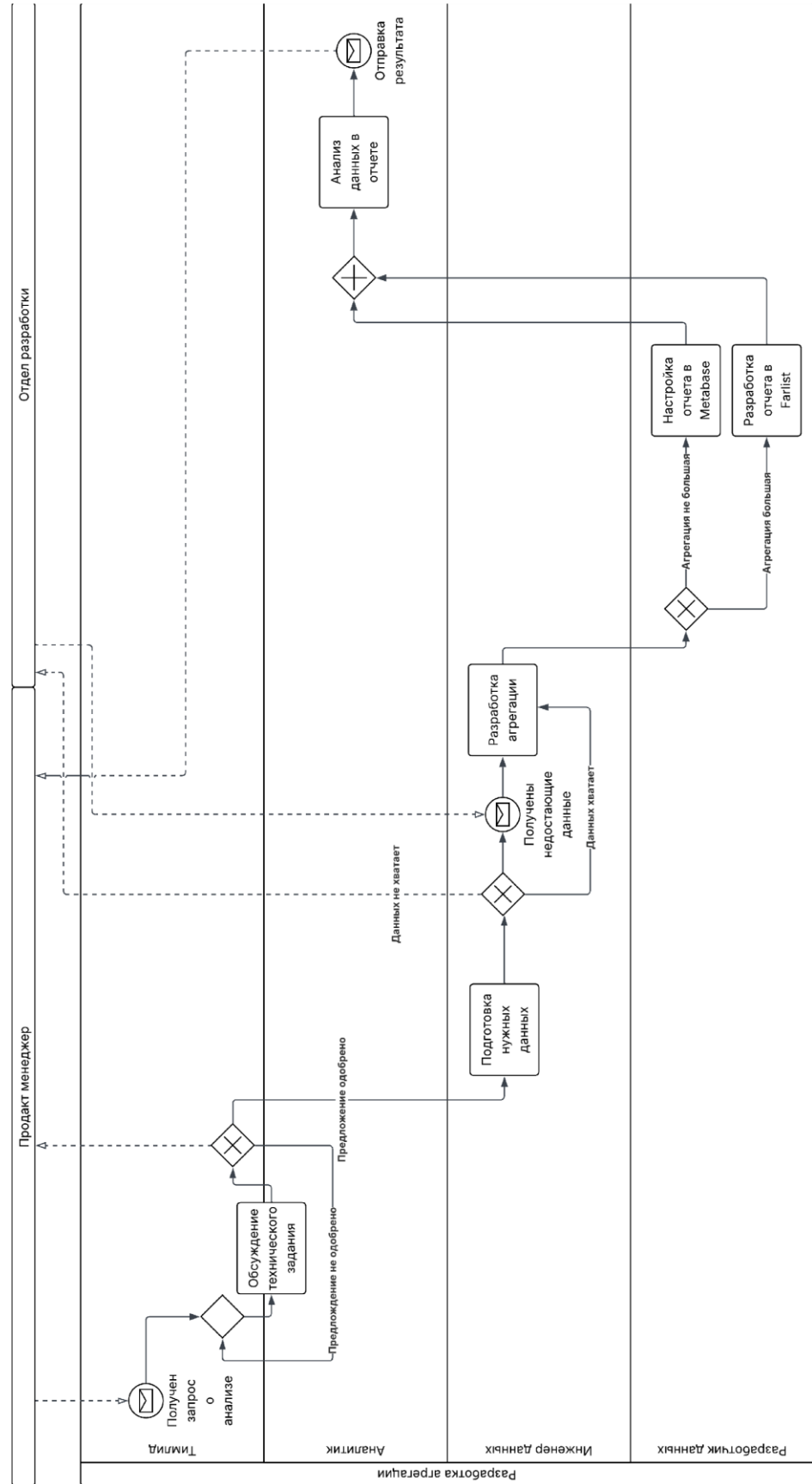


Рисунок А.1 – Визуализация бизнес-процесса Разработки агрегации