**Технический проект: Система обновления аналитических витрин данных «СОВД-Spark»**

Настоящий документ описывает и обосновывает архитектурные, технологические и проектные решения для создания распределенной системы обработки данных (далее – ПО) для оптимизации процесса обновления аналитических витрин.

Система предназначена для IT-инфраструктуры компании  
«ЭнергосбыТ Плюс». Она состоит из нескольких ключевых взаимодействующих компонентов:

1. Вычислительное ядро: кластер Apache Spark, выполняющий ресурсоемкие ETL-операции (извлечение, преобразование, агрегацию и загрузку) больших объемов данных;
2. Управляющий компонент: сервер Apache Airflow, отвечающий за оркестрацию, планирование по расписанию, мониторинг и управление жизненным циклом ETL-процессов.
3. Хранилище данных: включает в себя источники данных, промежуточное хранилище HDFS и целевую аналитическую СУБД ClickHouse для хранения итоговых витрин.

**Функциональная структура продукта**

Функциональная структура системы отражает её распределенное устройство и логику взаимодействия компонентов. Продукт включает основные блоки, реализующие классический ETL-конвейер:

* извлечение данных: чтение новых и измененных данных из различных источников КХД;
* преобразование данных: очистка, обогащение, применение бизнес-логики и агрегация данных в памяти кластера;
* загрузка данных: запись обработанных и агрегированных данных (витрин) в целевую СУБД;
* оркестрация и мониторинг: управление последовательностью, зависимостями и расписанием выполнения задач, а также отслеживание их статуса.

Такое разделение обеспечивает ясную логику, независимость компонентов и возможность масштабирования. Для описания структуры применяется диаграмма IDEF0, демонстрирующая функциональную иерархию системы.

IDEF0 и декомпозированная IDEF0 представлены на рисунках 1 и 2 соответственно.

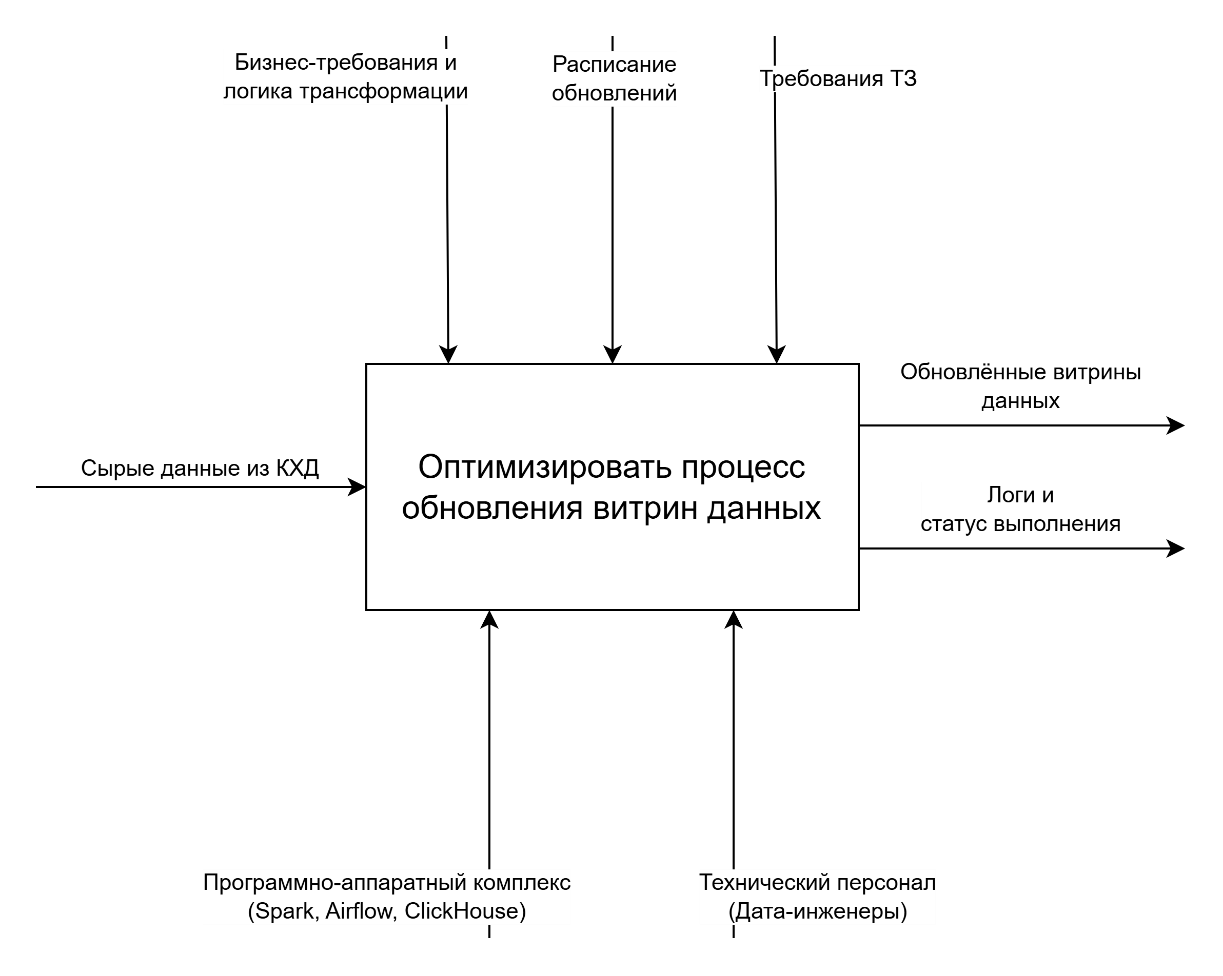


Рисунок 1 – IDEF0-диаграмма

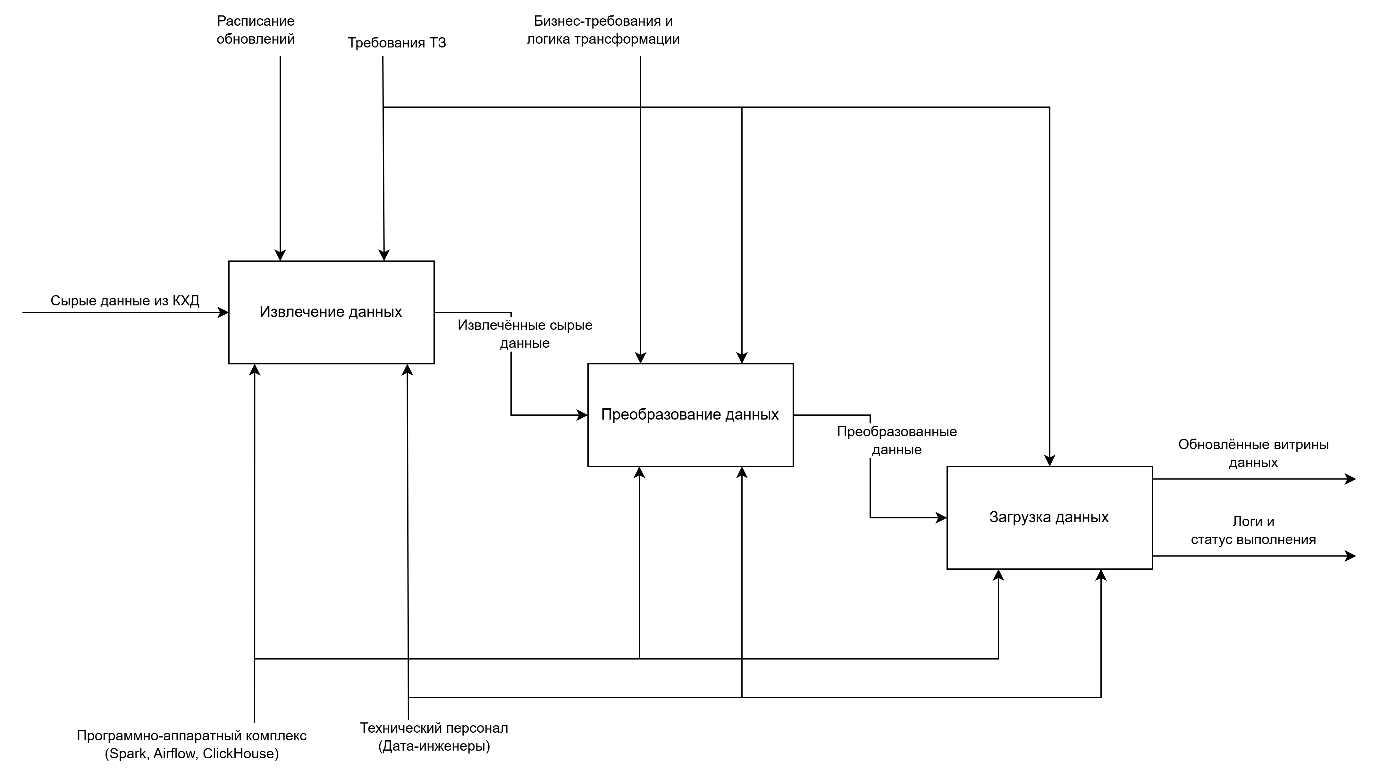


Рисунок 2 – Декомпозированная IDEF0-диаграмма

**Спецификация процессов**

Поведение системы описывается через основные сценарии использования и взаимодействие её компонентов.

Основные сценарии:

* запуск ETL-процесса по расписанию, успешное выполнение всех этапов и обновление данных в витрине;
* автоматическое обнаружение ошибки, прерывание процесса, отправка уведомления ответственному персоналу и последующее ручное вмешательство для устранения проблемы и перезапуска.

Для описания последовательности действий в рамках этих сценариев применяются диаграмма BPMN, которая показывает логику выполнения, точки принятия решений и зоны ответственности каждого компонента системы и персонала.

BPMN-диаграмма представлена на рисунке 3.

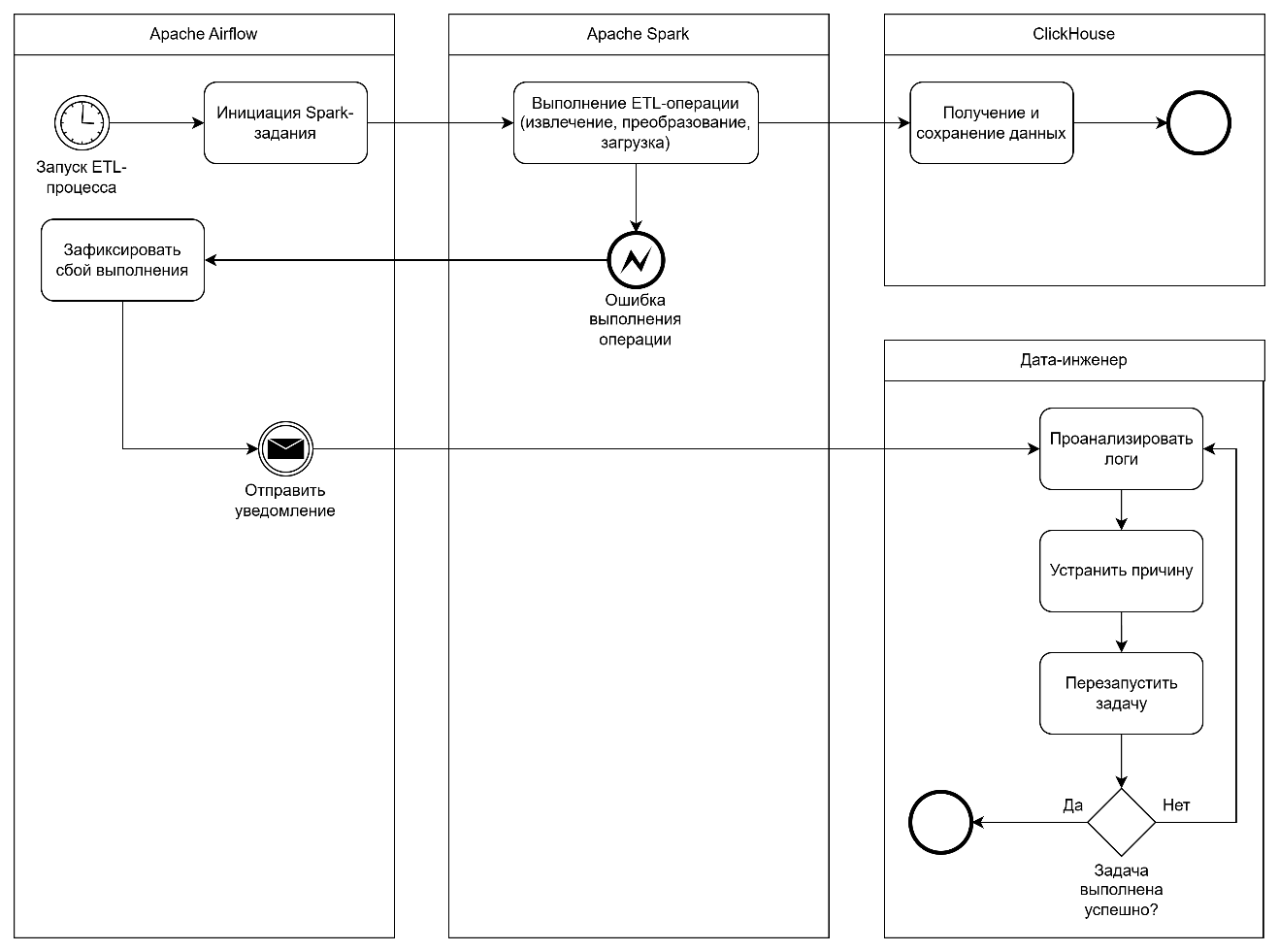


Рисунок 3 – BPMN-диаграмма

**Структура хранимой информации**

Данные в системе организованы вокруг аналитической модели "Звезда", которая является отраслевым стандартом для построения хранилищ и витрин данных. Она включает:

* таблицу фактов – хранит измеряемые числовые показатели (объем потребления, сумма начисления);
* таблицы-измерения – хранят описательные атрибуты (контекст), в разрезе которых анализируются факты.

Связи – одна запись в таблице-измерении соответствует множеству записей в таблице фактов (связь "один-ко-многим"). Такой подход обеспечивает высокую производительность аналитических запросов.

Для описания логической структуры базы данных используется  
ER-диаграмма, представленная на рисунке 4.

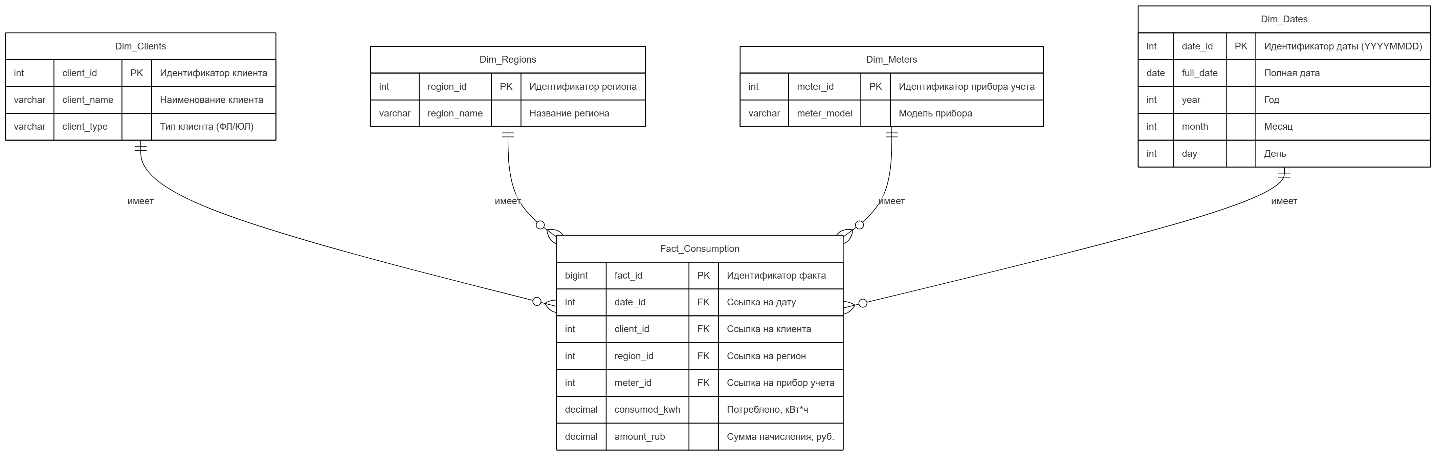


Рисунок 4 – ER диаграмма

**Архитектурно-структурное решение**

Продукт использует многоуровневую распределенную архитектуру, компоненты которой развернуты на различных физических или виртуальных узлах. Она состоит из следующих ключевых уровней:

1. Уровень управления: сервер Apache Airflow, который не выполняет вычислений, а лишь управляет запуском и мониторингом ETL-задач. Взаимодействие с ним осуществляется через веб-интерфейс.
2. Уровень обработки: кластер Apache Spark, который получает команды от Airflow и выполняет все ресурсоемкие операции по обработке данных в распределенном режиме.
3. Уровень хранения: включает серверы-источники КХД, файловую систему HDFS для временных данных и сервер СУБД ClickHouse для итоговых витрин.

Для визуализации физического расположения компонентов и их сетевого взаимодействия применяется диаграмма развертывания, представленная на рисунке 5.

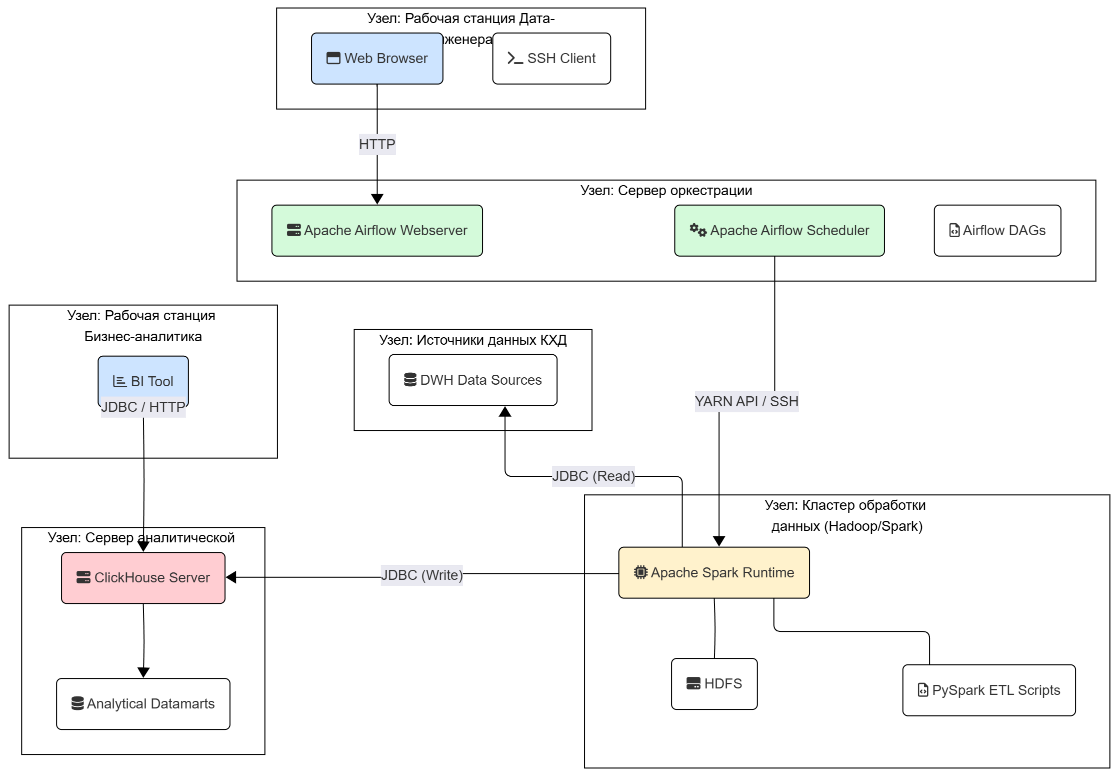


Рисунок 5 – Диаграмма развёртывания

**Интерфейсы**

Основным интерфейсом взаимодействия технического персонала  
(Дата-инженера) с системой является стандартный веб-интерфейс Apache Airflow. Он не является частью разрабатываемого ПО, но используется как основной инструмент управления.

Интерфейс Airflow предоставляет следующие возможности:

* просмотр списка всех ETL-процессов (DAGs) и их статусов;
* визуализация графа зависимостей задач внутри процесса;
* мониторинг истории запусков и статуса каждого конкретного выполнения;
* доступ к детальным логам для диагностики ошибок;
* ручной запуск, остановка и перезапуск процессов.

Конечные пользователи, бизнес-аналитики, не взаимодействуют с системой напрямую. Они получают доступ к результатам её работы через сторонние BI-инструменты, которые подключаются к СУБД ClickHouse.