**Наименование проекта**

«Оптимизация процесса обновления витрин данных КХД Энергосбыт с использованием Apache Spark»

**Назначение проекта**

Основная цель проекта – сократить время и повысить надежность процесса обновления аналитических витрин данных в корпоративном хранилище данных (КХД) компании «Энергосбыт».

Проект направлен на решение следующих бизнес-задач:

* обеспечить бизнес-аналитиков и руководство компании актуальными данными для оперативного анализа потребления электроэнергии, расчетов с клиентами и эффективности сбытовых мероприятий;
* перенести ресурсоемкие операции по трансформации и агрегации больших объемов данных с традиционной СУБД хранилища на специализированный, высокопроизводительный кластер распределенных вычислений Apache Spark;
* создать масштабируемую платформу, которая позволит в будущем быстро разрабатывать и внедрять новые витрины данных под изменяющиеся требования бизнеса без значительных доработок основной архитектуры;
* разработать автоматизированный конвейер для регулярного и надежного обновления витрин, минимизируя ручное вмешательство и связанные с ним ошибки.

**Описание взаимодействия с потенциальным пользователем**

Потенциальные пользователи (акторы):

* Бизнес-аналитик: основной потребитель данных из витрин. Использует BI-инструменты для построения отчетов и дашбордов на основе обновленных данных;
* Дата-инженер: Ответственный за поддержку, мониторинг и развитие ETL-процесса. Взаимодействует с кодом Spark, настройками кластера и базой данных ClickHouse;
* Администратор КХД: Ответственный за стабильность работы источников данных и целевой базы данных, контролирует нагрузку и доступ к данным.

**Пользовательские истории**

Для Бизнес-аналитика:

* «Как бизнес-аналитик, я хочу получать обновленные данные по ежедневному потреблению электроэнергии в разрезе регионов до 9:00 утра, чтобы успевать готовить ежедневный отчет для руководства»;
* «Как бизнес-аналитик, я хочу, чтобы данные в отчетах загружались быстро, в течение нескольких секунд, чтобы я мог интерактивно исследовать информацию без задержек».

Для Дата-инженера:

* «Как дата-инженер, я хочу иметь автоматизированный процесс обновления витрин, который запускается по расписанию, чтобы не тратить время на ручной запуск скриптов»;
* «Как дата-инженер, я хочу получать уведомления в случае сбоя ETL-процесса, чтобы я мог оперативно устранить проблему»;
* «Как дата-инженер, я хочу, чтобы процесс добавления новых атрибутов в витрину был простым и не требовал полной перестройки всего конвейера данных, чтобы я мог быстро реагировать на запросы бизнеса».

**Пользовательские сценарии**

Сценарий №1. Получение ежедневного отчета по потреблению электроэнергии.

Основной сценарий:

1. аналитик открывает свой BI-инструмент;
2. аналитик открывает дашборд «Ежедневный отчет по потреблению»;
3. система BI отправляет SQL-запрос к витрине данных в ClickHouse;
4. ClickHouse быстро обрабатывает запрос и возвращает агрегированные данные;
5. дашборд обновляется, отображая актуальные данные за прошедший день (объемы потребления, суммы начислений, динамика по регионам);
6. аналитик выгружает отчет в формате PDF для отправки руководству.

Сценарий №2. Реакция на сбой при обновлении витрины данных.

Основной сценарий:

1. запланированный ETL-процесс на кластере Apache Spark завершается с ошибкой;
2. система мониторинга фиксирует сбой и отправляет уведомление  
   дата-инженеру;
3. дата-инженер подключается к интерфейсу управления кластером;
4. он находит лог выполнения упавшего задания Spark;
5. анализируя лог, он определяет причину сбоя;
6. дата-инженер устраняет проблему;
7. он вручную перезапускает ETL-процесс;
8. процесс успешно завершается, витрина в ClickHouse обновляется.

**Функциональные требования**

* система должна обеспечивать возможность подключения к различным источникам данных КХД «Энергосбыт», включая реляционные СУБД и файловые хранилища;
* система должна ежедневно извлекать новые и измененные данные по показаниям приборов учета, информации о клиентах, тарифах и платежах за прошедшие сутки;
* процесс извлечения данных должен быть распараллелен средствами Apache Spark для обеспечения высокой скорости чтения больших объемов;
* система должна выполнять очистку и валидацию данных;
* все бизнес-правила и логика трансформации должны быть реализованы в коде приложения Apache Spark;
* система должна загружать обработанные и агрегированные данные в целевую СУБД ClickHouse;
* должен поддерживаться режим полного обновления витрины для обеспечения консистентности данных;
* весь ETL-процесс должен запускаться автоматически по заданному расписанию;
* система должна логировать ключевые этапы выполнения процесса, а также любые ошибки и предупреждения.