Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

Лабораторная работа №3

Тема:

«Проектирование архитектуры хранилища больших данных» Вариант 10

Выполнил: Попов.А.Е., АДЭУ-221

Москва
2025

Энергетическая компания Smart Grid

Источники данных

В данной архитектуре учитываются три основных источника, типичных для Smart Grid-систем: данные от «умных» счетчиков энергопотребления, поступающие в виде временных рядов; технологические данные SCADA-систем, содержащие параметры работы оборудования, трансформаторных подстанций и сетевой инфраструктуры; а также внешние погодные API, предоставляющие сведения о температуре, ветре, осадках и других факторах, влияющих на нагрузку на энергосеть. Такой набор источников является ключевым для задач прогнозирования, контроля качества энергии и выявления аномалий.

Прием и транспортировка данных

Входящий поток данных обрабатывается через EMQX и Apache Kafka. EMQX используется как MQTT-брокер, принимающий телеметрию от smart meters, поскольку эти устройства передают данные с высокой частотой и небольшими пакетами. SCADA-системы также могут направлять агрегированные параметры в этот брокер. Apache Kafka выполняет роль распределенной шины данных, обеспечивая надежную доставку событий, их буферизацию и разделение потоков для разных компонентов обработки. Благодаря этому достигается устойчивость системы Smart Grid даже при резких изменениях интенсивности поступающих данных.

Оркестрация процессов

Работа всех пайплайнов координируется Apache Airflow, который управляет загрузкой данных, периодической обработкой и обновлением аналитических витрин. Airflow позволяет системно интегрировать потоковые данные smart meters и периодически обновляемые данные SCADA, а также автоматизировать прогнозные расчёты, необходимые для анализа нагрузки на сеть.

Обработка данных

Apache Spark является сердцем аналитической обработки. Он обеспечивает очистку, агрегацию и анализ временных рядов, что особенно важно для измерений энергопотребления. Spark может объединять данные погодных API с данными счетчиков, формируя расширенную модель для прогнозирования сети. Кроме того, Spark позволяет выявлять аномальные

значения, связанные с предположительными потерями электроэнергии или попытками хищения.

Хранение данных

Хранилище формируется на основе Delta Lake, что дает возможность хранить как сырые данные, поступающие в больших объемах от счетчиков, так и обработанные аналитические данные с транзакционной надежностью. Хранилище поддерживает эффективное ведение истории измерений, повторное воспроизведение данных, построение витрин для анализа нагрузки и качества энергоснабжения.

Аналитика и визуализация

Для визуализации результатов мониторинга энергосети используется Apache Superset, который позволяет строить дашборды, отображающие текущую нагрузку, историческую динамику, прогнозы, а также возможные аномалии в потреблении. Superset подключается к данным в Delta Lake и предоставляет аналитикам возможность оперативно анализировать показатели сети.

Мониторинг и управление

Обеспечение контроля над всей архитектурой выполняется средствами Prometheus и Grafana. Эти инструменты позволяют отслеживать состояние потоков данных, нагрузку на инфраструктуру, стабильность работы брокеров и пайплайнов Airflow. Для Smart Grid такая система мониторинга необходима для своевременного реагирования на нестабильность или ошибки в цепочке сбора данных.

Схема архитектуры

Ниже приводится схема, иллюстрирующая потоки данных и роль каждого компонента:

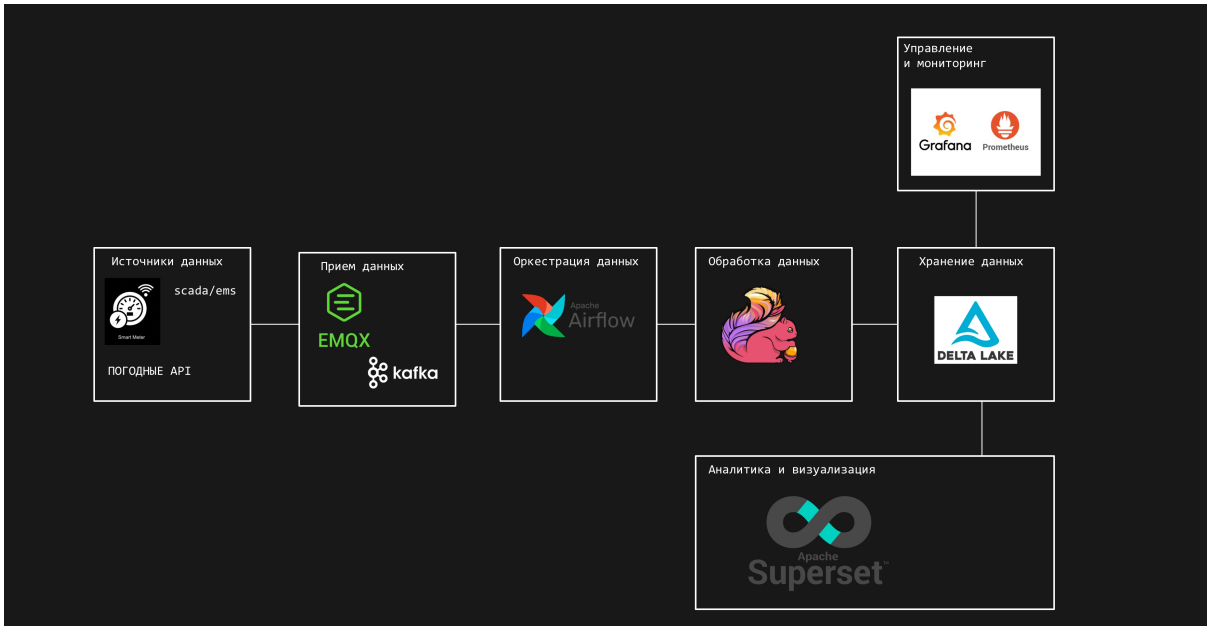

Рисунок 1 — Архитектура Smart Grid для сбора данных с умных счетчиков, SCADA-систем и погодных API

Вывод

Разработанная архитектура соответствует требованиям сценария Smart Grid и позволяет эффективно решать ключевые задачи энергетической компании: мониторинг энергопотребления, прогнозирование нагрузки, выявление потенциальных потерь и аномалий. Использование EMQX и Kafka обеспечивает надежный прием данных от больших массивов «умных» счетчиков, Spark предоставляет мощный механизм аналитической обработки временных рядов, а Delta Lake гарантирует надежное и масштабируемое хранение данных. Визуальный слой на базе Superset предоставляет специалистам удобный доступ к аналитике, а система мониторинга на основе Prometheus и Grafana обеспечивает устойчивую работу всей платформы. Такая архитектура обладает высокой гибкостью, масштабируемостью и позволяет развивать систему в направлении более продвинутой аналитики и машинного обучения.