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»

Выполнил: Дулис К.С., АДЭУ-221

Москва
2025

Анализ требований: для лабораторной работы по варианту "Промышленный холдинг (Индустрия 4.0)" требования сводятся к интеграции IoT-датчиков, SCADA и MES для предиктивного обслуживания оборудования, мониторинга линий и оптимизации качества продукции. Системы анализируют данные о вибрации, температуре и процессах в реальном времени, прогнозируя отказы, минимизируя простои и повышая OEE через ML-модели и цифровые двойники. Необходима бесперебойная интеграция по OPC UA/MQTT с высокой доступностью, безопасностью и масштабируемостью для холдинга, где данные из станков и систем обеспечивают traceability и снижение брака. В лабораторной достаточно смоделировать потоки в Python/SQL или UML для демонстрации KPI вроде наработки на отказ и затрат ТОиР.

Выбор компонентов архитектуры и их описание:

Архитектура включает несколько уровней. Источники данных: IoT-датчики, SCADA, MES и внешние API, которые генерируют телеметрию оборудования и параметры качества продукции. Прием данных реализуется через Kafka (стриминг событий), Logstash (ETL/нормализация логов и телеметрии) и gRPC-сервисы для двустороннего обмена с промышленными системами. Хранение обеспечивает связка PostgreSQL (операционные и справочные данные), Hadoop и HBase (масштабируемое распределённое хранилище исторических временных рядов). Обработка данных выполняется в Apache Spark, Flink и Hive: Spark для батч- и ML-обработки, Flink для стриминга в реальном времени, Hive как SQL-слой по данным в кластере. Управление и мониторинг данных обеспечивают Apache Atlas (каталог и линия данных) и Apache Ranger (политики безопасности и доступов). Для аналитики и визуализации используются Jupyter (исследовательский анализ, прототипы моделей), Superset (дашборды для бизнеса) и TensorFlow (модели предиктивного обслуживания). Prometheus выступает в роли системы оркестрации/мониторинга метрик инфраструктуры и сервисов.

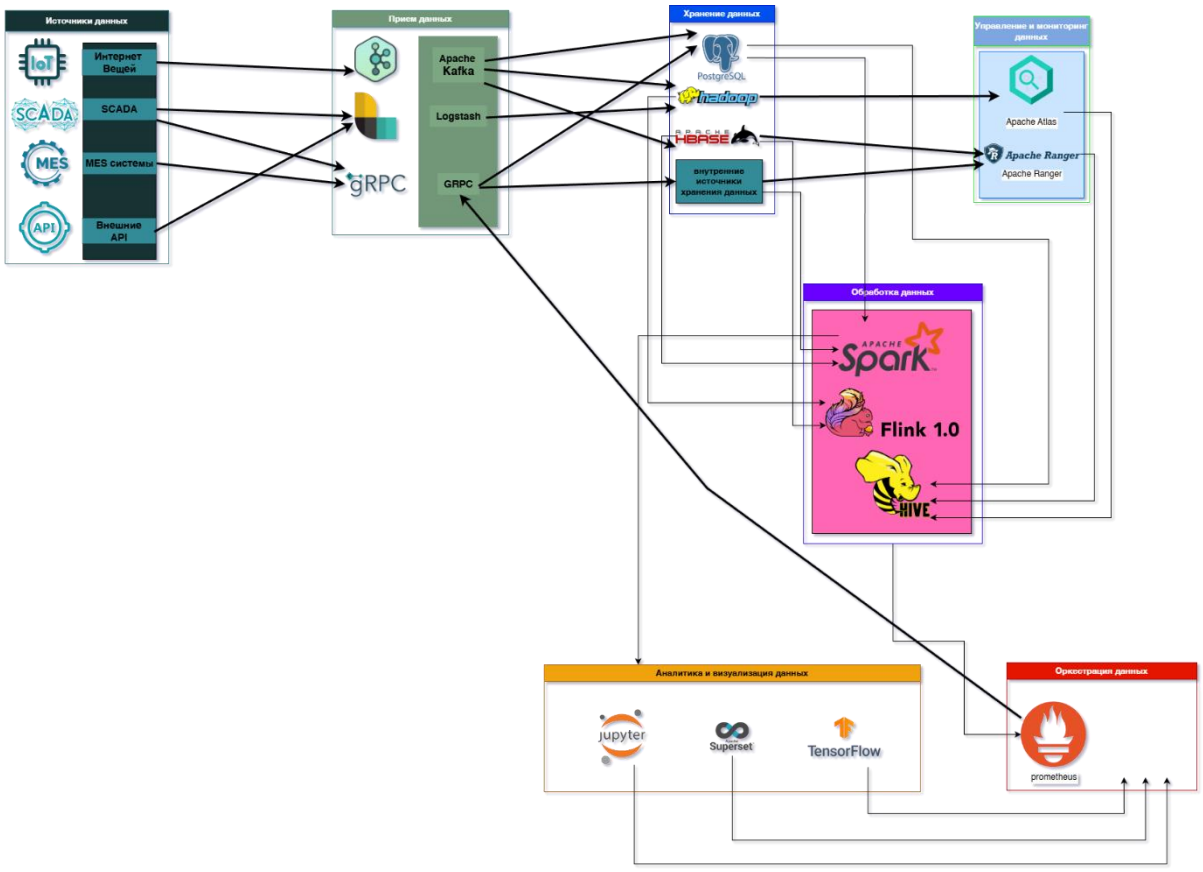
Обоснование выбора:

Этот стек оптимален, потому что сочетает зрелые open-source-технологии, хорошо зарекомендовавшие себя в промышленных сценариях Индустрии 4.0. Kafka, Hadoop/HBase, Spark и Flink масштабируются горизонтально и способны обрабатывать телеметрию с тысяч датчиков и линий в реальном времени, что критично для предиктивного обслуживания и мониторинга. Использование Atlas и Ranger обеспечивает управляемость и соответствие

требованиям безопасности холдинга, а связка Jupyter–TensorFlow–Superset закрывает полный цикл от экспериментальной аналитики до продуктовых дашбордов и моделей.

Описание потоков данных:

Потоки данных выглядят так: телеметрия и события с IoT/SCADA/MES поступают в шину Kafka или в Logstash по протоколам и коннекторам, а также через gRPC-сервисы. Logstash нормализует, фильтрует и обогащает события, после чего они записываются в хранилища: “тёплые” данные и агрегаты идут в PostgreSQL, “холодные” исторические массивы и сырые логи – в Hadoop/HBase. Spark и Flink читают потоки из Kafka и данные из Hadoop/HBase/PostgreSQL, считают показатели состояния оборудования, прогнозы отказов и ключевые KPI качества/производительности. Результаты агрегаций и предсказаний сохраняются обратно в БД и подаются в Superset и Jupyter, а инфраструктурные и бизнес-метрики снимаются Prometheus для дальнейшего мониторинга.



Вопросы производительности и масштабируемости:

Масштабируемость обеспечивается горизонтальным масштабированием брокеров Kafka, узлов Hadoop/HBase и кластеров Spark/Flink: при росте числа станков или частоты измерений добавляются новые ноды. Хранилище Hadoop позволяет линейно наращивать объём исторических данных без потери производительности за счёт распределения нагрузки, а Spark/Flink поддерживают параллельную обработку на многих исполнительных узлах. Prometheus фиксирует нагрузку на кластеры и сервисы, что позволяет автоматически масштабировать контейнеры/воркеры и предотвращать деградацию производительности.

Анализ потенциальных проблем и их решений:

Потенциальные проблемы: перегрузка Kafka и Logstash при пиковых потоках событий, рост задержек при обработке в Spark/Flink, “разрастание” кластера Hadoop, а также сложности с безопасностью и качеством данных. Для их снижения настраиваются политики ретенции и партиционирования тем Kafka, масштабирование Logstash по нескольким инстансам и back-pressure в стриминговых джобах. В Hadoop/HBase применяются компрессия и архивирование старых данных, а Atlas и Ranger помогают управлять доступом и отслеживать происхождение данных, что уменьшает риск утечек и некорректных расчётов.

Выводы:

В итоге архитектура строит полный конвейер данных от IoT/SCADA/MES до аналитики и предиктивных моделей, опираясь на масштабируемый стриминговый и batch-стек. Такое решение позволяет промышленному холдингу в духе Индустрии 4.0 обеспечивать непрерывный мониторинг линий, предиктивное обслуживание оборудования и улучшение качества продукции при контролируемой стоимости владения и высокой гибкости развития.