Департамент образования города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

Инструменты для хранения и обработки больших данных

Лабораторная работа 3.1

Проектирование архитектуры хранилища больших данных

Выполнила: студентка группы АДЭУ-221

Ильина Алина Сергеевна

Проверил:

доцент департамента информатики, управления и технологий

Босенко Тимур Муртазович

Москва

2025

Вариант 8. Страховая компания: динамическое ценообразование на основе телематики (стиль вождения), автоматизированная оценка ущерба по фотографиям, выявление страхового мошенничества. Источники: данные телематики, фото/видео с ДТП, страховые полисы.

Задача: создать архитектуру хранилища больших данных для страховой компании.

Цель: обеспечить интеграцию и хранение телематических, мультимедийных и полисных данных, потоковую обработку данных для реального времени, автоматизацию оценки ущерба через анализ фото/видео, выявление страхового мошенничества с ML, обеспечение безопасности и соответствия 152-ФЗ, масштабируемость и высокую доступность системы.

1. Анализ требований:

1.1 Объем данных

- Ожидаемый объем: 20-40 ТБ в год.
- Рост: 40-60% ежегодно за счет увеличения количества телематических устройств и данных мультимедиа.

1.2 Скорость получения данных

- Телематические данные (стиль вождения): в реальном времени, до 2000 событий в секунду.
- Фото/видео с ДТП: загрузка в течение 5-15 минут после инцидента.
- Данные страховых полисов: ежедневные обновления.

1.3 Типы данных

- Структурированные: данные страховых полисов, заявки на возмещение (25%).
- Полуструктурированные: телематические данные (JSON, CSV), отчёты с устройств (50%).
- Неструктурированные: фото и видео с ДТП, аудиозаписи (25%).

1.4 Требования к обработке

- Анализ стиля вождения и динамическое ценообразование: в режиме реального времени.
- Автоматизированная оценка ущерба на основе фото/видео: еженедельно с дополнительной обработкой по инцидентам.
- Выявление мошенничества: еженедельно и при подаче заявок.

- Отчетность для руководства: ежедневно и по запросу.

1.5 Доступность данных

- Время отклика для аналитических запросов: <20 секунд.
- Доступность системы: 99.95% (допустимое время простоя ~4.4 часа в год).

1.6 Безопасность данных

- Шифрование данных в покое и при передаче по стандартам AES-256.
- Многофакторная аутентификация и разграничение доступа для сотрудников.
- Полный аудит и журналирование всех операций с данными.
- Соответствие требованиям 152-ФЗ "О персональных данных" и международным стандартам безопасности (ISO 27001).
- Формирование отчетов для актуариев (прогноз риска, убытков), операционных менеджеров (статус заявлений, качество фото), маркетинга (динамика тарифов, сегменты клиентов).

Для данной страховой компании лучше использовать **Hybrid Data Storage**, чтобы эффективно работать с большими объемами разнообразных данных и обеспечивать производительную аналитику в реальном времени.

2. Выбор компонентов архитектуры:

Источники данных

- Телематические устройства (стриминг style driving)
- Фото и видео с ДТП (неструктурированные)
- Страховые полисы (структурированные данные)

Слой сбора данных

- Apache Kafka: потоковый сбор телематических данных в реальном времени
- Airbyte/Fivetran: интеграция данных из полисов и внешних источников
- Logstash: сбор логов и событий

Слой хранения данных

- Облачное хранилище Amazon S3 с поддержкой Data Lakehouse

- Delta Lake или Apache Iceberg для ACID-транзакций и управления версиями данных
- PostgreSQL для метаданных и структурированных данных

Слой обработки данных

- Apache Spark (Databricks) для пакетной обработки данных
- Apache Flink для стриминговой обработки с низкой задержкой
- Apache Hive для SQL-запросов по большим данным

Слой аналитики, визуализации и машинного обучения

- Power BI / Tableau для построения бизнес-дашбордов
- Grafana для мониторинга и визуализации потоковых метрик
- Jupyter Notebooks для интерактивного анализа данных
- TensorFlow и PyTorch для моделей оценки ущерба и выявления мошенничества

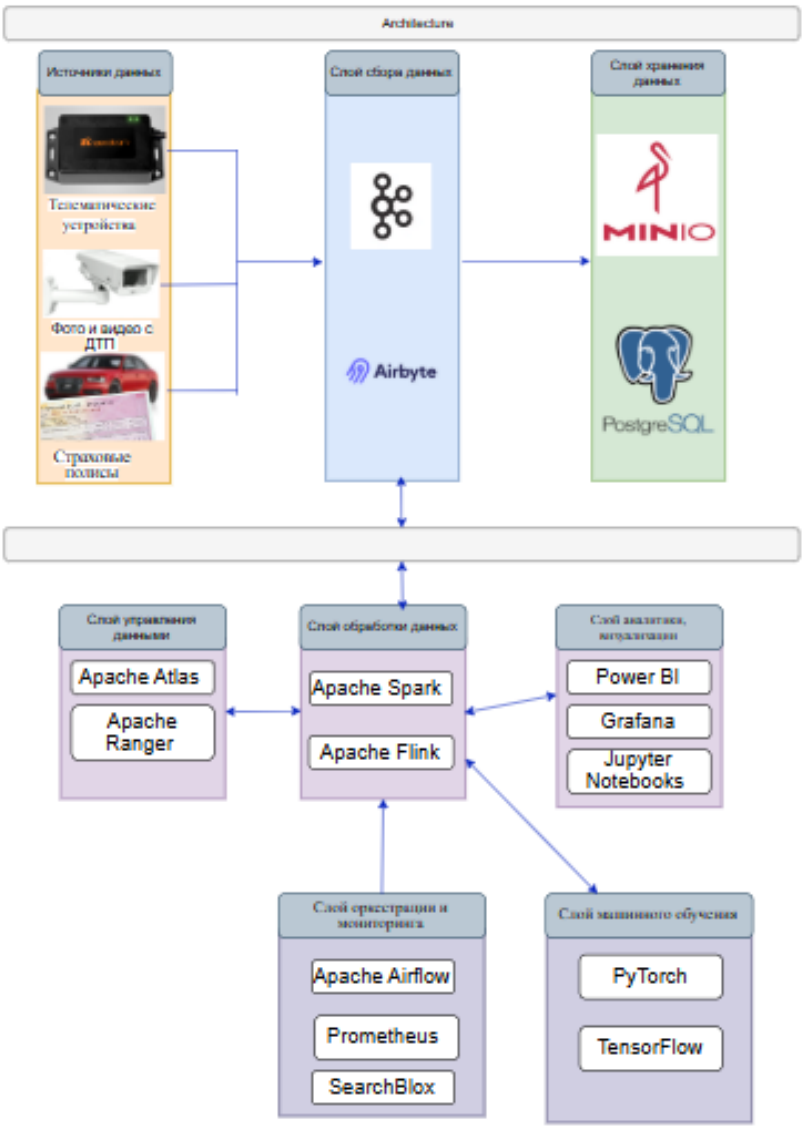
Слой оркестрации и мониторинга

- Apache Airflow для управления ETL и рабочих процессов
- Prometheus + Grafana для сбора и визуализации метрик
- ELK Stack для логирования и аудита

Слой управления данными

- Apache Atlas для каталогизации и управления метаданными
- Apache Ranger для контроля доступа и безопасности

3. **Схема архитектуры:**



Слой сбора данных:

- Apache Kafka: надежная потоковая платформа для сбора и передачи телематических и других данных в режиме реального времени; хорошо подходит для обработки событий и интеграции с обработчиками.
- Airbyte/Fivetran: удобные инструменты ETL/ELT для интеграции структурированных данных из страховых полисов и сторонних источников.

Слой хранения данных:

- **MinIO - бесплатное** объектное хранилище с открытым исходным кодом

- PostgreSQL: для хранения метаданных и структурированных данных, таких как страховые полисы.

Слой обработки данных:

- Apache Spark (Databricks): для пакетной обработки больших массивов данных, оптимизированный для аналитики.
- Apache Flink: для стриминговой обработки данных с низкой задержкой в реальном времени.

Слой аналитики, визуализации:

- Power BI / Tableau: для построения бизнес-дашбордов и визуализации ключевых метрик.
- Grafana: для мониторинга потоковых метрик и инфраструктуры.
- Jupyter Notebooks: для интерактивного анализа и разработки ML-моделей.

Слой машинного обучения:

- TensorFlow — лучше подходит для производства и масштабируемых решений, например, автоматической оценки ущерба по фото, благодаря высокой производительности и поддержке развёртывания.
- PyTorch — удобен для исследований и прототипирования моделей, например, для выявления мошенничества, за счёт гибкости и простоты разработки.

Слой оркестрации и мониторинга:

- Apache Airflow: для управления ETL-процессами и сложными рабочими процессами.
- Prometheus: для сбора и визуализации метрик системы и мониторинга.
- отечественные продукты для логирования SearchBlox

Слой управления данными:

- Apache Atlas: для каталогизации данных и управления метаданными.
- Apache Ranger: для реализации контроля доступа и безопасности данных.

4. Процесс обработки данных

- Данные собираются из телематических устройств, систем загрузки фото/видео ДТП и CRM через слой сбора данных.

- Сырые телематические данные и мультимедиа сохраняются в объектном хранилище (например, HDFS или S3) для долгосрочного хранения.
- Потоковые телематические данные обрабатываются в режиме реального времени с помощью Apache Flink или Kafka Streams для динамического ценообразования и обнаружения аномалий.
- Автоматизированная оценка ущерба запускается пакетными задачами Apache Spark, выполняющими обработку и анализ изображений и видео с использованием моделей компьютерного зрения.
- Выявление мошенничества осуществляется путем регулярных пакетных вычислений и моделей машинного обучения на Spark или Databricks по расписанию (еженедельно).
- Результаты аналитики и оценки ущерба сохраняются в базе данных с низкой задержкой доступа, например, Apache HBase или Cassandra.
- Аналитики используют Jupyter Notebooks для глубинного анализа данных и создание прототипов моделей, а дашборды и отчеты строятся в Superset или Power BI для мониторинга ключевых метрик.
- Модели машинного обучения обучаются на исторических данных телематики, мультимедиа и мошеннических паттернах, затем развертываются для онлайн-оценки риска, прогнозирования ущерба и выявления подозрительных действий.

5. Масштабирование и отказоустойчивость

- Использование распределённого хранилища данных (например, Apache HDFS, Amazon S3 или Azure Blob Storage) с возможностью горизонтального масштабирования для хранения больших объёмов телематических и мультимедийных данных.
- Репликация и резервное копирование данных в объектном хранилище и базе с быстрой доступностью (например, Apache HBase, Cassandra) для обеспечения высокой отказоустойчивости и сохранности данных.
- Применение Kubernetes для оркестрации микросервисов обработки данных, моделей машинного обучения и API сервисов с автоматическим масштабированием в зависимости от нагрузки.
- Горизонтальное масштабирование потоковой обработки данных на Apache Flink или Kafka Streams для стабильной обработки телематических событий в режиме реального времени.
- Внедрение системы мониторинга и алертинга (Prometheus, Grafana) для своевременного реагирования на сбои и поддержания SLA по доступности и времени отклика.
- Использование мульти-региональных развертываний и балансировщиков нагрузки для повышения отказоустойчивости и снижения времени простоя.

6. Безопасность

- Реализация шифрования данных в покое и при передаче с использованием стандартов AES-256, включая шифрование объектов в хранилищах (например, S3 SSE) и подключение TLS/SSL для передачи данных.
- Использование многофакторной аутентификации и централизованной системы управления идентификацией (например, LDAP, Active Directory) совместно с Kerberos для надежной аутентификации пользователей и сервисов.
- Применение решений для детального контроля доступа и аудита данных, таких как Apache Ranger или коммерческие аналоги, для тонкой настройки прав доступа к телематическим, мультимедийным и финансовым данным.
- Регулярное резервное копирование данных с использованием автоматизированных процедур и хранением бэкапов в гео-распределённых репозиториях.
- Разработка и тестирование плана аварийного восстановления (Disaster Recovery Plan), включая сценарии восстановления критичных компонентов и данных после сбоев или кибератак.
- Ведение полного журнала аудита всех изменений и операций с данными для обеспечения соответствия требованиям 152-ФЗ «О персональных данных» и внутренним политикам безопасности.