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

Вариант - 3

Лабораторная работа 3.1

Тема: «Проектирование архитектуры хранилища больших данных»

Дисциплина «Инструменты для хранения и обработки больших данных»

Выполнила:
Арлинская Александра Викторовна
Проверил:
Босенко Тимур Муртазович

Курс обучения: 4

Форма обучения: очная

Москва

2025

1. Определение требований

1.1 Объем данных:

- Ожидаемый объем: 100–200 ТБ годовых, с ростом ~40% в год (учитывая детализацию CDR/IPDR и логи с базовых станций).

1.2 Скорость поступления данных:

- Базовые станции: пайплайн в реальном времени, до 10 000 сообщений/секунда.
- Биллинг: обновления каждые несколько минут/часов, порциями.
- Техподдержка: обновления событий, тикеты, дневные выгрузки.

1.3 Типы данных:

- Структурированные: CDR (Call Detail Records), IPDR, биллинг, логи обращений.
- Полуструктурированные: json/xml отчеты по оборудованию, статистика нагрузки.
- Неструктурированные: текстовые обращения в поддержку, отчеты об инцидентах.

1.4 Требования к обработке:

- Прогнозирование нагрузки на сеть: почасово/помесячно.
- Анализ качества связи: ежедневные отчеты, оперативные панели.
- Отток клиентов (churn prediction): ежемесячные и ежеквартальные прогнозы.
- Интерактивная аналитика и визуализация для бизнес-аналитиков.

1.5 Доступность:

- Реальное время для мониторинга и прогнозирования.
- SLA — 99.9%, отклик <30 секунд для критических запросов.

1.6 Безопасность:

- Шифрование данных при хранении и передаче.
- Многофакторная аутентификация, сегментация доступов.
- Аудит действий, соблюдение 152-ФЗ, GDPR.

2. Выбор модели хранилища

- Data Lake — для хранения сырых CDR/IPDR, биллинга, логов.
- Data Warehouse — для агрегированных и аналитических данных по клиентам и нагрузке.
- Hybrid Storage — сочетание lake (быстрая потоковая запись) и warehouse (аналитика, отчеты).

3. Проектирование архитектуры

1. Источники:

- Базовые станции (системы OSS/BSS, SNMP брокеры),
- Биллинг-CRM,
- Техподдержка (тикеты, чаты, обращения).

2. Слой сбора данных:

- Kafka — потоковая доставка;
- Confluent — обработка, доп.безопасность.

3. Слой хранения:

- Delta Lake
- Apache Iceberg

4. Слой обработки:

- Apache Spark — пакетная/стрим-аналитика, ML обработка;
- Apache Flink — обработка событий в реальном времени;
- ClickHouse — быстрая агрегация для больших данных.

5. Слой аналитики и ML:

Jupyter Notebook, TensorFlow, Apache Superset — аналитика, построение ML, визуализация результатов и динамики по регионам/кластерам.

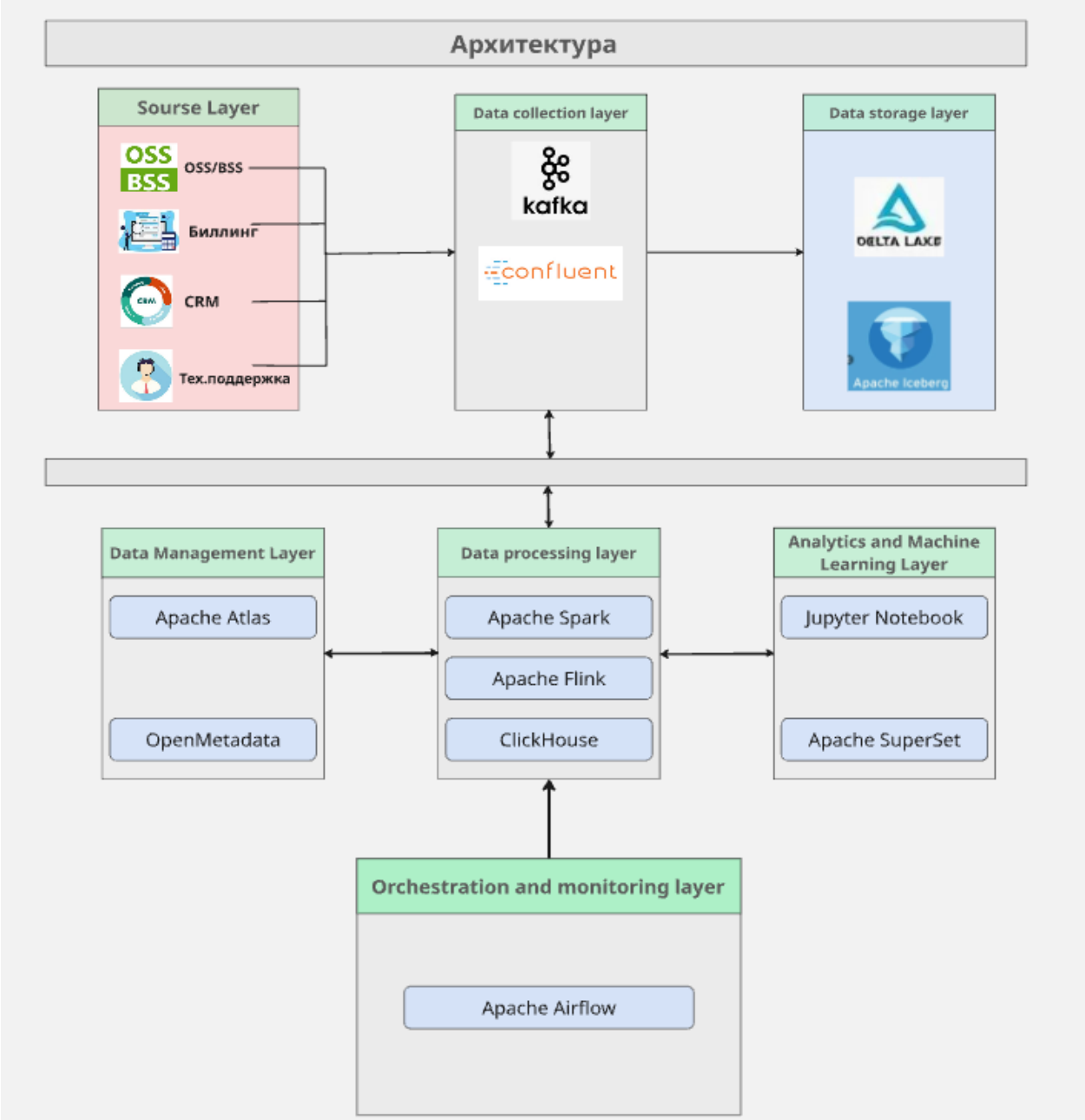
6. Слой управления и безопасности:

- Apache Atlas — управление метаданными;

7. Оркестрация и мониторинг:

- Apache Airflow — построение ETL пайплайнов, расписание задач;

4. Диаграмма



5. Описание компонентов выбранных инструментов и обоснование выбора

Apache Kafka служит надежной основой для загрузки потоковых данных в реальном времени, минимизируя задержки и обеспечивая предупреждения о высоких нагрузках. Инструменты Confluent добавляют дополнительный уровень безопасности и расширенную поддержку, что делает работу с данными еще более надежной.

Delta Lake поддерживает ACID-транзакции, обеспечивая быстрые обновления данных, таких как биллинг и записи звонков. Для долгосрочного хранения объемной истории данных подходит Apache Hadoop. Все сервисы являются бесплатными.

Потоки данных из Kafka могут быть записаны в таблицы, где они проходят очистку и подготовку к дальнейшей обработке. Apache Airflow предоставляет удобные инструменты для создания, планирования и мониторинга этапов обработки данных, обеспечивая отличную интеграцию с другими компонентами экосистемы.

Spark отлично справляется с обработкой CDR (Call Detail Record) и IPDR (Internet Protocol Detail Record), предоставляя подробные записи о телефонных вызовах и интернет-активности.

Flink же обеспечивает реальный мониторинг данных с базовых станций. ClickHouse подходит для получения сверхбыстрых ответов и агрегированных показателей по огромным массивам телекоммуникационных данных. Apache Atlas предоставляет мощные инструменты для управления данными биллинга, а OpenMetadata обеспечивает отличную интеграцию с BI-системами, способствуя командной работе и оптимизации процессов.

6. Анализ потенциальных проблем и их решений

Потоковая обработка с ultra-low latency (узкое место в Flink/Logstash):

Решение: масштабирование кластеров Flink, настройка параметров памяти и backpressure; мониторинг нагрузки для предотвращения узких мест.

Рост расходов на хранение больших объемов данных:

Решение: Использование Data Lifecycle Management, архивирование старых данных и tiered storage; внедрение сжатия и дельта-обновлений.

Сложность управления большим количеством потоков и коннекторов:

Решение: централизованный мониторинг через Grafana и Apache Atlas; автоматизация оркестрации через Airflow с alerting.