Департамент образования города Москвы

Государственное автономное образовательное учреждение высшего образования города Москва «Московский Городской Педагогический Университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ЛАБОРАТОРНАЯ РАБОТА №3.1 (Вар.9)

по дисциплине «Инструменты для хранения и обработки больших данных»

Тема: «Проектирование архитектуры хранилища больших данных»

Направление подготовки 38.03.05 – «бизнес-информатика»

Выполнила:

Студентка группы АДЭУ-221

Муханова А. И.

Преподаватель:

Босенко Т. М.

к.т.н., доц. департамента информатики, управления и технологий

Москва

2025

1. Цель работы

Сценарий: образовательная онлайн-платформа, предназначенная для персонализации траекторий обучения, анализа успеваемости студентов, рекомендации курсов и проведения A/B-тестирования новых функций.

Источники данных: логи взаимодействия с платформой (clickstream), результаты тестов, видеолекции, метаданные курсов и профили пользователей.

Типы данных:

Структурированные — профили пользователей, результаты тестов, оценки, расписания.

Полуструктурированные — логи взаимодействия (JSON), события кликов, события воспроизведения видео.

Неструктурированные — видеолекции, текстовые отзывы, загруженные файлы.

Объёмы и скорость:

- до 1–3 ТБ данных в год,

- потоковая генерация событий: 1 000–5 000 сообщений/сек,

- пакетные загрузки: результаты тестов, выгрузки LMS.

Скорость поступления: смешанный режим — потоковые события (до нескольких тысяч событий в секунду) и пакетные загрузки (результаты тестов, дампы).

Бизнес-цели: персонализация обучения в near-real-time, мониторинг успеваемости, A/B-тестирование интерфейсов и рекомендаций, отчётность для преподавателей и администраторов.

Требования к задержке/доступности:

Персональные рекомендации: задержка < 2 секунд для интерактивных сценариев.

Отчёты и модельное обучение: пакетная обработка с периодичностью от часов до дней.

Требования к безопасности: VK Cloud IAM: разграничение ролей, шифрование данных в S3-хранилище.

2. Выбор компонентов архитектуры

Модель: Data Lakehouse — объединение гибкости Data Lake и управляемости DW.

Компоненты архитектуры:

Слой	Реализация
Источники данных	Логи взаимодействия (веб/мобильные приложения), Результаты тестов (СУБД платформы), Видеолекции (файлы), Метаданные курсов и пользователей
Хранение данных	VK Cloud S3 (Data Lake, Delta Lake/Iceberg), ClickHouse (DWH Аналитика), PostgreSQL (OLTP, ML features via pgvector)

Обработка данных	Apache Flink (Потоковая обработка), Apache Spark (Пакетная обработка)
Аналитика	Apache Superset (BI дашборды), Grafana (Real-time мониторинг)
Обслуживание	Kubernetes VK Cloud (Управление ML)
Оркестрация и Мониторинг	Apache Airflow, Grafana
Безопасность и Управление	Apache Ranger

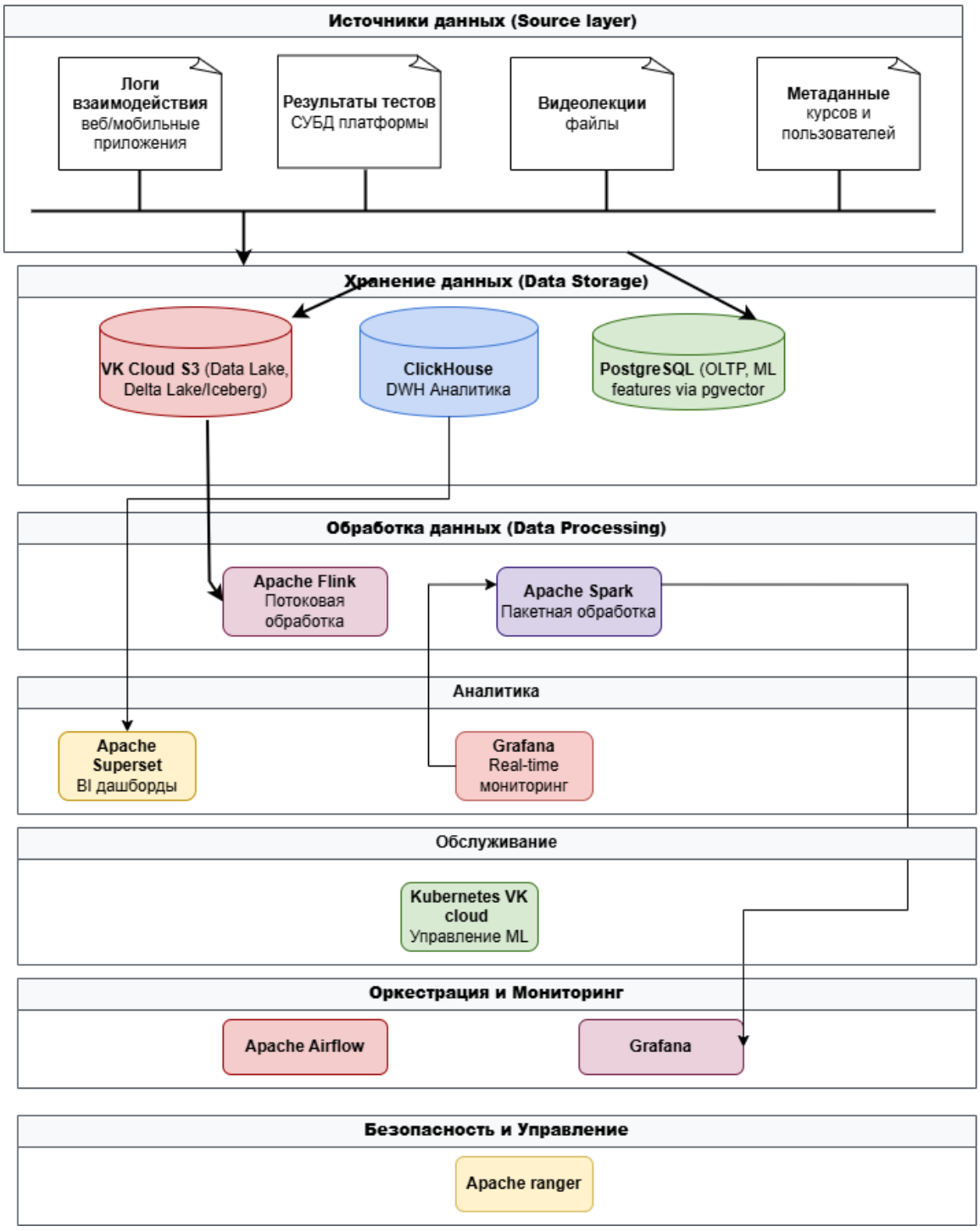
3. Обоснование выбора

Компонент	Почему выбран
VK Cloud S3 + Delta Lake/Iceberg	Российское облачное хранилище с поддержкой транзакционных форматов для надежного Data Lake
ClickHouse	Российская высокопроизводительная СУБД для аналитических запросов по успеваемости и активности
PostgreSQL + pgvector	Надежная OLTP-БД с векторными операциями для ML-рекомендаций и хранения фич
Apache Flink	Обработка потоковых данных логов взаимодействия в реальном времени для персонализации
Apache Spark	Пакетная обработка для ETL-пайплайнов, расчета

	метрик успеваемости, подготовки данных для ML
Apache Superset	Открытая BI-платформа для создания дашбордов для преподавателей и администраторов
Grafana	Мониторинг активности платформы в реальном времени и метрик производительности
Kubernetes VK Cloud	Российская облачная платформа для оркестрации ML-сервисов рекомендаций
Apache Airflow	Оркестрация всех данных пайплайнов и ML-процессов
Apache Ranger	Централизованное управление безопасностью и доступом к данным платформы

4. Диаграмма архитектуры

Архитектура хранилища больших данных для образовательной онлайн-платформы



5. Описание потоков данных

Пример 1 — Логи взаимодействия → Персонализация обучения:

1. Веб/мобильное приложение отправляет события взаимодействия (просмотр лекции, выполнение задания) → Apache Kafka
2. Apache Flink обрабатывает события в реальном времени → вычисляет активность студентов, предпочтения в обучении → результаты публикует в ClickHouse (для аналитики) и обновляет векторы в PostgreSQL (для ML-рекомендаций)
3. Apache Spark ежедневно обрабатывает накопленные логи из VK Cloud S3 → строит витрины успеваемости → обновляет ClickHouse

Пример 2 — Результаты тестов → Аналитика успеваемости:

1. СУБД платформы отправляет результаты тестов через Debezium (CDC) → VK Cloud S3 (Raw Zone)
2. Apache Spark обрабатывает данные → рассчитывает средний балл, прогресс по курсам, выявляет проблемные темы → записывает результаты в ClickHouse (для отчетов) и PostgreSQL (для адаптивного обучения)

Пример 3 — Видеолекции → Аналитика вовлеченности:

1. Видеофайлы загружаются через Apache NiFi → VK Cloud S3 (хранение)

2. Метрики просмотра (паузы, перемотки, завершение) отправляются → Apache Kafka → Apache Flink → агрегируются в ClickHouse для анализа вовлеченности

6. Отказоустойчивость и масштабируемость

- VK Cloud S3 — автоматическая репликация данных, неограниченное масштабирование хранилища
- ClickHouse — кластерная конфигурация с репликацией данных для отказоустойчивости аналитических запросов
- PostgreSQL — репликация и резервное копирование для обеспечения доступности ML-сервисов
- Apache Flink — checkpointing для восстановления состояния потоковой обработки
- Apache Spark — snapshot и восстановление состояний пакетной обработки
- Kubernetes VK Cloud — автоматическое масштабирование ML-сервисов в зависимости от нагрузки

7. Потенциальные проблемы и решения

1) Рост стоимости хранения образовательного контента:

- *Решение:* Внедрение политик жизненного цикла в VK Cloud S3 → автоматическое перемещение старых видео в холодное хранилище

2) Обеспечение конфиденциальности образовательных данных:

- *Решение:* Шифрование данных в хранилищах, управление доступом через Apache Ranger, минимизация хранения персональных данных, псевдонимизация в аналитических витринах

3) Интеграция потоковых и пакетных данных для единой аналитики:

- *Решение:* Использование формата Delta Lake/Iceberg в VK Cloud S3 как единого слоя для согласованного доступа к данным

8. Выводы

Предложенная архитектура обеспечивает эффективное управление образовательными данными и реализацию персонализированного подхода к обучению. Использование связки Apache Kafka + Apache Flink позволяет обрабатывать активность студентов в реальном времени для мгновенной адаптации учебного процесса. Data Lake на базе VK Cloud S3 с поддержкой Delta Lake гарантирует надежное хранение разнородных образовательных данных. ClickHouse обеспечивает высокоскоростную аналитику успеваемости, а PostgreSQL с pgvector — реализацию интеллектуальных систем рекомендаций. Архитектура масштабируема, отказоустойчива и соответствует требованиям современных образовательных платформ.