Департамент образования города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ОТЧЕТ

по дисциплине «Инструменты для хранения и обработки больших данных»

Направление подготовки 38.03.05 – бизнес-информатика

Профиль подготовки «Аналитика данных и эффективное управление»

(очная форма обучения)

Выполнил:

Студент группы АДЭУ-211

Буц Игорь Денисович

Руководитель:

Босенко Тимур Муртазович

Москва  
2024

**Задача**: создать архитектуру хранилища больших данных для компании, предоставляющей финансовые услуги.

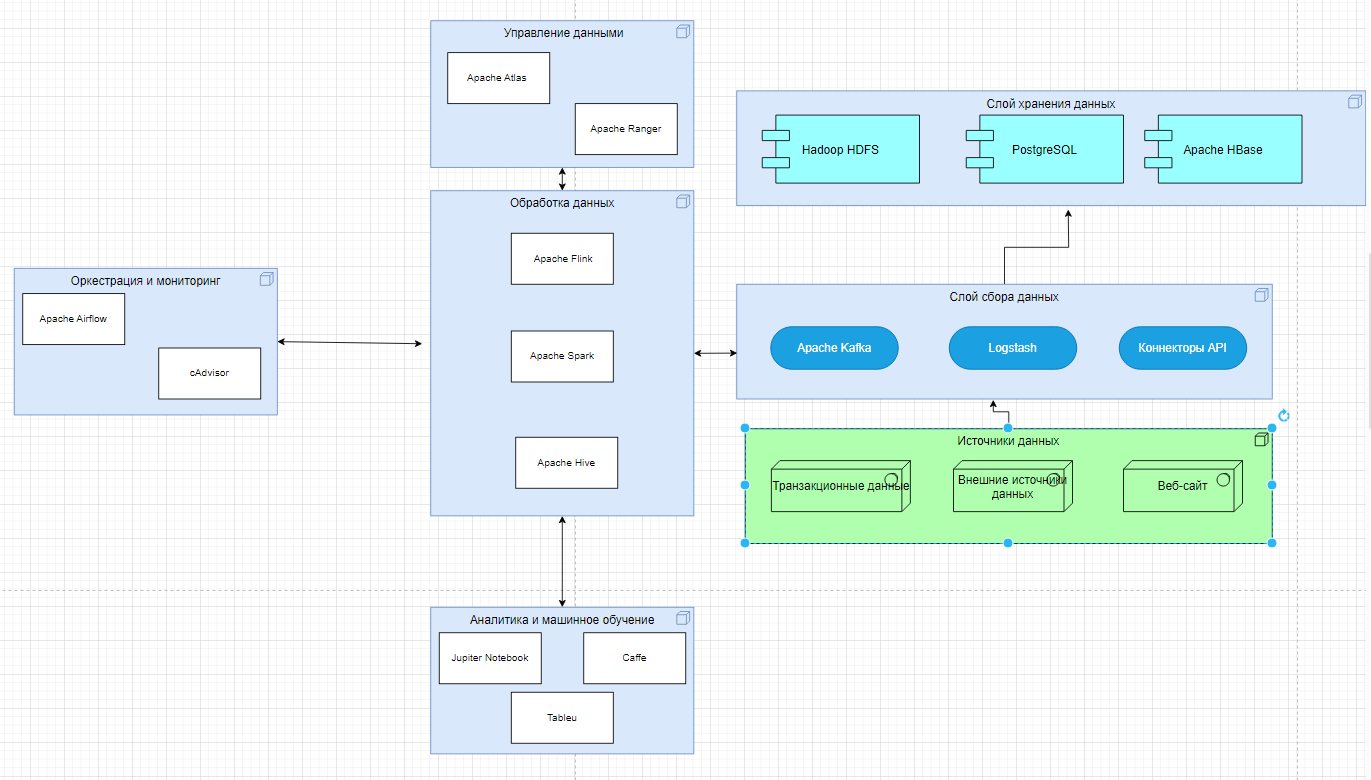
**Цель**: обеспечить надежное хранение, эффективную обработку и анализ больших объемов данных, получаемых из различных источников, таких как вебсайты, рыночные данные, сторонние финансовые службы.

Вариант 2

1. **Объем данных**: 100 ТБ в год, рост 30% ежегодно.
2. **Скорость получения данных**: до 1000 транзакций в секунду.
3. **Типы данных**: 80% структурированные, 15% полуструктурированные, 5% неструктурированные.
4. **Требования к обработке**: выявление мошенничества в реальном времени, оценка кредитных рисков.
5. **Доступность данных**: 99.999%, время отклика
6. **Безопасность данных**: сквозное шифрование, строгое соответствие 152-ФЗ и требованиям ЦБ РФ
7. **Компонент архитектуры**

2.1 Источники данных

* **Транзакционные системы**
* Рыночные данные, кредитные бюро, регулирующие органы и сторонние финансовые службы.
* Вебсайт
  1. Слой сбора данных
* Apache Kafka для потоковых данных.
* Logstash для сбора логов.
* Пользовательские коннекторы для CRM и внешних API.
  1. Слой хранения данных
* Hadoop HDFS
* PostgreSQL для структурированных данных и метаданных.
* Apache HBase для быстрого доступа к большим объемам данных.
  1. Слой обработки данных
* Apache Spark для пакетной и потоковой обработки.
* Apache Flink для обработки в реальном времени.
* Apache Hive для SQL-подобных запросов к большим данным.
  1. Слой аналитики и машинного обучения
* Jupyter Notebooks для интерактивной аналитики
* Caffe – фреймворк для моделей машинного обучения
* Tableu – для визуализации данных
  1. Слой управления данными
* Apache Atlas для управления метаданными.
* Apache Ranger для контроля доступа и аудита
  1. Слой оркестрации и мониторинга
* Apache Airflow - для оркестрации рабочих процессов.
* cAdvisor – для мониторинга

1. **Схема** **архитектуры**
2. **Процесс обработки данных**
   1. Данные собираются из различных источников через слой сбора данных. −
   2. Сырые данные сохраняются в HDFS для долгосрочного хранения.
   3. Потоковые данные обрабатываются в реальном времени с помощью Apache Kafka для быстрой аналитики.
   4. Пакетные задачи, такие как сегментация клиентов, выполняются с помощью Spark по расписанию. − Результаты анализа сохраняются в HBase для быстрого доступа.
   5. Аналитики используют Jupyter Notebooks и Tableu для исследования данных и создания отчетов.
   6. Модели машинного обучения обучаются на исторических данных и развертываются для прогнозирования и рекомендаций в Caffe.
3. **Масштабирование и отказоустойчивость** 
   1. Использование кластерной архитектуры Hadoop для горизонтального масштабирования.
   2. Репликация данных в HDFS и HBase для обеспечения отказоустойчивости.
   3. Использование Apache Airflow для оркестрации и автоматического масштабирования микросервисов.
4. **Безопасность** 
   1. Реализация шифрования данных с помощью HDFS Transparent Encryption.
   2. Использование Accenture для аутентификации.
   3. Применение Apache Ranger для детального контроля доступа к данным.
   4. Регулярное резервное копирование и план аварийного восстановления.