Департамент образования города Москвы

Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»

Институт цифрового образования
Департамент информатики, управления и технологий

Лабораторная работа 3-1
по дисциплине «Инструменты для хранения и обработки больших данных»

Тема: «Проектирование архитектуры хранилища больших данных»

Направление подготовки 38.03.05 – бизнес-информатика
Профиль подготовки «Аналитика данных и эффективное управление»
(очная форма обучения)

Выполнила:
Студентка группы АДЭУ-211
Белик Мария Константиновна

Преподаватель:
Босенко Т.М.

Москва
2024

ВВЕДЕНИЕ

Цель работы: разработать архитектуру хранилища больших данных для заданного сценария использования.

Задачи:

1. разработать архитектуру хранилища больших данных для компании, основываясь на предоставленных требованиях;
2. описать компоненты архитектуры, обосновать выбор технологий и предложить схему потока данных.

Исходные данные (Вариант 1): Крупный онлайн-ритейлер

— Объем данных: 500 ТБ в год, рост 50% ежегодно.

— Скорость получения: до 5000 транзакций в секунду.

— Типы данных: 60% структурированные, 30% полуструктурированные, 10% неструктурированные.

— Требования к обработке: анализ поведения пользователей в реальном времени, прогнозирование спроса.

— Доступность: 99.99%, время отклика <5 секунд.

— Безопасность: шифрование, соответствие 152-ФЗ и PCI DSS.

ХОД РАБОТЫ

Шаг 1. Определение требований для крупного онлайн-ритейлера

1.1. Объем данных

— Ожидаемый объем: 500 ТБ в год.

— Ожидаемый рост: 50% ежегодно.

1.2. Скорость получения данных:

— Мобильные приложения и веб-сайты: до 5000 транзакций в секунду.

— Социальные сети: обновление каждые 15 минут.

— CRM системы: в режиме реального времени.

1.3. Типы данных:

— Структурированные: транзакционные данные (информация о продажах, покупках, возвратах, включая идентификаторы товаров, количество, цена, дата, время транзакции), данные о клиентах, складские данные, данные CRM (60%).

— Полуструктурированные: логи событий (данные о взаимодействиях пользователей с сайтом и мобильным приложением, включая время на странице, клики, навигацию), данные о продуктах (описание категорий, атрибутов товаров в формате JSON/XML), обратная связь от клиентов (отзывы и рейтинги, собранные в виде текстовых данных, сообщения, реакции на публикации) (30%).

— Неструктурированные: данные о товарах (изображения и видео товаров для размещения на сайте или в мобильном приложении), массовые рассылки (текстовые и мультимедийные данные с предложениями и акциями), посты в социальных сетях (10%).

1.4. Требования к обработке

— Прогнозирование спроса: в режиме реального времени.

— Анализ поведения пользователей: в режиме реального времени.
— Персонализация предложений: в режиме реального времени.
— Обработка транзакций: в режиме реального времени.
— Сегментация клиентов: ежемесячно.

1.5. Доступность данных

— Время отклика для аналитических запросов: <5 секунд.
— Доступность системы: 99,99% (допустимое время простоя ~ 8,5 часов в год)

1.6. Безопасность данных

— Шифрование
— Соответствие 152-ФЗ и PCI DSS.

Шаг 2. Выбор модели хранилища данных: компоненты архитектуры

Выбор модели Hybrid Data Storage, т.к. несмотря на то, что 60% данных компании структурированные, все-таки остальные 40% занимают полуструктурированные и неструктурированные данные, для которых подойдет озеро данных.

Выбор схемы «Снежинка», т.к. она экономит место, что важно для данной компании, у которой по условию объем обрабатываемых и хранимых данных 500 ТБ в год, да еще и ожидается ежегодный рост этого значения на 50%. Тем более, «Снежинка» подразумевает нормализованные таблица измерений, а Вы говорили, что к нормализации надо стремиться.

2.1. Источники данных

— Мобильное приложение и веб-сайт.
— Социальные сети (ВКонтакте, Telegram, Одноклассники, VC.ru, Tik Tok, Дзен).
— CRM системы.

— Платежные системы банков (обработка платежей на стороне банка).

— Внешние API (данные о погоде для специализации предложений и рассылок, курс валют, данные о поведении клиента на сторонних приложениях или веб-сайтах, данные о браузерных запросах пользователей для персонализации подборок товаров).

2.2. Слой сбора данных

— Apache Pulsar – для сбора потоковых данных.

— Fluentd – для сбора логов.

2.3. Слой хранения данных

— Apache Cassandra – для быстрого доступа к большим объемам данных, для хранения структурированных данных.

— HDFS (Hadoop Distributed File System) для хранения сырых данных.

— DBeaver для структурированных данных.

2.4. Слой обработки данных

— Apache Spark – для пакетной и потоковой обработки.

— Apache Storm – для обработки в реальном времени.

— Apache Hive – для SQL-подобных запросов к большим данным.

2.5. Слой аналитики и машинного обучения

— VSCode – для аналитики.

— PowerBI – для визуализации и дашбордов.

— Apache Druid – для аналитики в режиме реального времени.

2.6. Слой управления данными

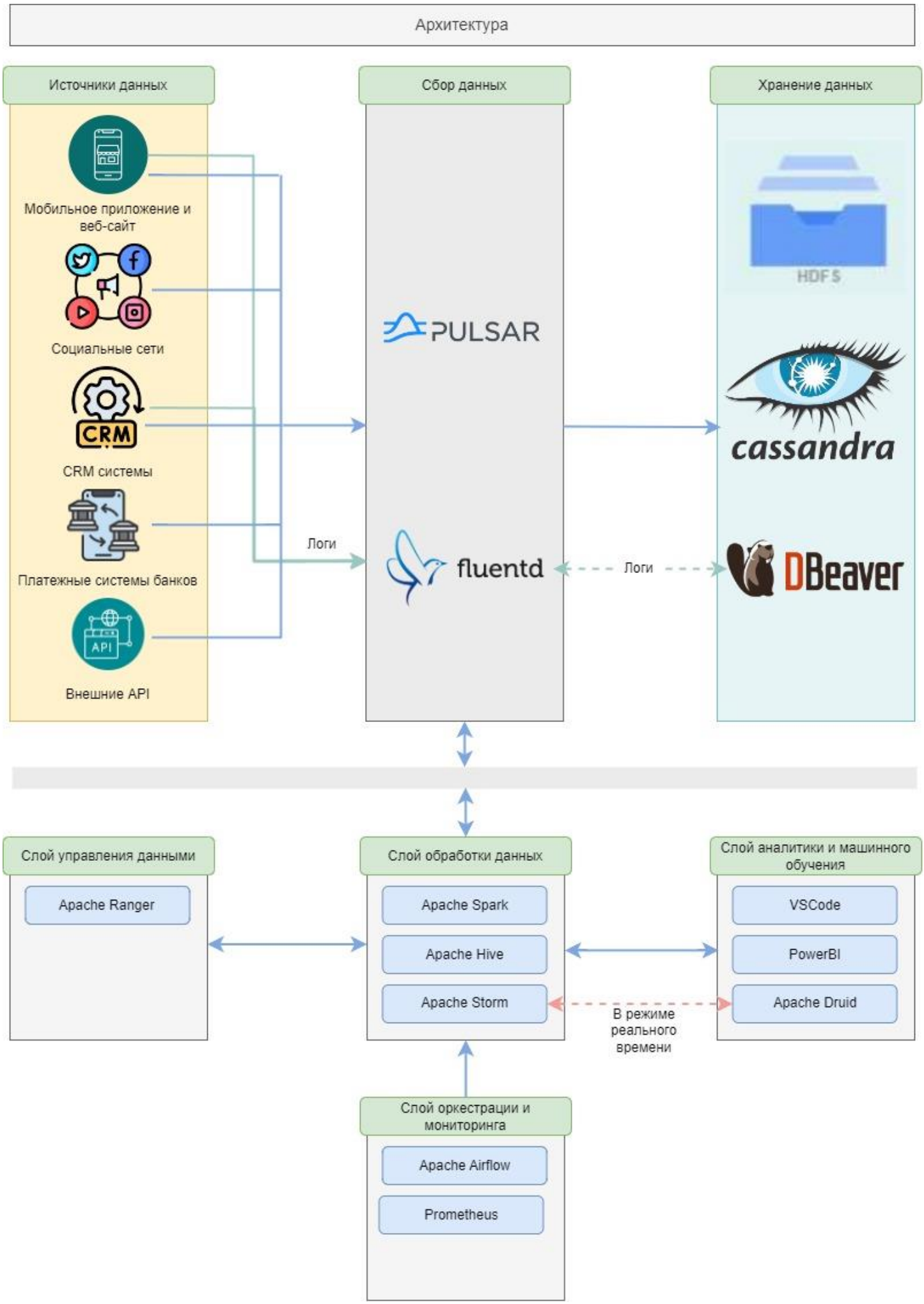
— Apache Ranger – для контроля доступа и аудита

2.7. Слой оркестрации и мониторинга

— Apache Airflow – для оркестрации.

— Prometheus – для мониторинга и алертинга.

Шаг 3. Схема архитектуры


Архитектура
Источники данных
Мобильное приложение и веб-сайт
Социальные сети
CRM системы
Платежные системы банков
Внешние API
Сбор данных
PULSAR
fluentd
Логи
Логи
Хранение данных
HDFS
cassandra
DBeaver
Слой управления данными
Apache Ranger
Слой обработки данных
Apache Spark
Apache Hive
Apache Storm
В режиме реального времени
Слой аналитики и машинного обучения
VSCode
PowerBI
Apache Druid
Слой оркестрации и мониторинга
Apache Airflow
Prometheus

Шаг 4. Процесс обработки данных

- Данные собираются из различных источников через слой сбора данных.
- Сырые данные сохраняются в HDFS для долгосрочного хранения.
- Потоковые данные обрабатываются с помощью Apache Storm для быстрой аналитики в режиме реального времени (анализ поведения потребителей, персонализация предложений, обработка транзакций).
- Пакетные задачи, такие как сегментация клиентов, выполняются с помощью Spark по расписанию.
- Результаты анализа сохраняются в Apache Cassandra для быстрого доступа.
- Аналитики используют VSCode для исследования данных, Apache Druid для работы в режиме реального времени, PowerBI для создания профессиональных дашбордов.
- Модели машинного обучения обучаются на исторических данных и развертываются для прогнозирования и рекомендаций.

Шаг 5. Масштабирование и отказоустойчивость

- Использование HDFS и Apache Cassandra для обеспечения отказоустойчивости.
- Использование Apache Airflow и Prometheus для оркестрации и масштабирования.

Шаг 6. Безопасность

- Применение Apache Ranger для детального контроля доступа к данным.

— Регулярное резервное копирование и план аварийного восстановления.

ВЫВОДЫ

В ходе выполнения лабораторной работы были выполнены все поставленные задачи и достигнута цель:

1. были описаны требования к архитектуре хранилища данных для компании, являющейся крупным онлайн-ритейлером;
2. были описаны компоненты архитектуры хранилища больших данных для компании, основываясь на предоставленных требованиях;
3. была построена схема архитектуры хранилища больших данных компании.