Лабораторная работа 3-1. Проектирование архитектуры хранилища больших данных

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

Алгоритм выполнения работы

1. Анализ требований:
   - Определить источники данных.
   - Выявить типы данных (структурированные, полуструктурированные, неструктурированные).
   - Оценить объемы данных и скорость их поступления.
   - Определить требования к аналитике и отчетности.
2. Выбор компонентов архитектуры:
   - Выбрать систему распределенного хранения (например, Hadoop HDFS, Amazon S3).
   - Определить систему обработки данных (например, Apache Spark, Flink).
   - Выбрать систему управления метаданными (например, Apache Atlas).
   - Определить инструменты для ETL процессов (например, Apache NiFi, Talend).
   - Выбрать инструменты для визуализации и аналитики (например, Tableau, Power BI).
3. Проектирование архитектуры:
   - Разработать схему потоков данных.
   - Определить компоненты для обеспечения безопасности и управления доступом.
   - Спроектировать систему мониторинга и логирования.
   - Разработать стратегию масштабирования.
4. Создание диаграммы архитектуры:
   - Использовать draw.io для создания визуального представления архитектуры.
   - Включить все основные компоненты и связи между ними.
5. Описание компонентов:
   - Для каждого компонента архитектуры предоставить краткое описание его роли и функций.
6. Обоснование выбора:
   - Объяснить причины выбора конкретных технологий и компонентов.
7. Рассмотрение вопросов производительности и масштабируемости:
   - Описать, как архитектура обеспечивает высокую производительность и масштабируемость.
8. Анализ потенциальных проблем и их решений:
   - Выявить возможные узкие места в архитектуре.
   - Предложить способы их устранения или минимизации влияния

9. Подготовка отчета:

- Составить подробный отчет, включающий все вышеперечисленные пункты.
- Приложить диаграмму архитектуры.

ПРИМЕР РЕШЕНИЯ ЗАДАНИЯ

Задача: создать архитектуру хранилища больших данных для компании, занимающейся анализом потребительского поведения.

Цель: обеспечить надежное хранение, эффективную обработку и анализ больших объемов данных, получаемых из различных источников, таких как веб-сайты, мобильные приложения, социальные сети и системы CRM.

Шаг 1. Определение требований

- **Объем данных:** Ожидаемый объем данных, которые будут храниться.
- **Скорость получения данных:** как часто данные будут поступать в хранилище.
- **Типы данных:** какие типы данных будут храниться (структурированные, неструктурированные, полуструктурированные).
- **Требования к обработке:** как данные будут использоваться (аналитика, машинное обучение, отчетность).
- **Доступность данных:** требования к доступности и времени отклика.
- **Безопасность данных:** требования к защите данных от несанкционированного доступа.

1. Требования к данным для средней компании в России

1.1 Объем данных

- Ожидаемый объем: 50-100 ТБ в год.
- Рост: 30-50% ежегодно.

1.2 Скорость получения данных

- Веб-сайты и мобильные приложения: в режиме реального времени, до 1000 событий в секунду.
- Социальные сети: обновление каждые 15 минут.
- CRM системы: ежедневные обновления.

1.3 Типы данных

- Структурированные: транзакционные данные, данные CRM (20%).
- Полуструктурированные: логи веб-сайтов и приложений, данные JSON/XML (50%).
- Неструктурированные: текстовые отзывы, посты в социальных сетях (30%).

1.4 Требования к обработке

- Анализ потребительских трендов: еженедельно.
- Сегментация клиентов: ежемесячно.
- Прогнозирование спроса: ежеквартально.

- Персонализация рекомендаций: в режиме реального времени.
- Отчетность для руководства: ежедневно/еженедельно.

1.5 Доступность данных

- Время отклика для аналитических запросов: <30 секунд.
- Доступность системы: 99.9% (допустимое время простоя ~8.8 часов в год).

1.6 Безопасность данных

- Шифрование данных в состоянии покоя и при передаче.
- Многофакторная аутентификация для доступа к данным.
- Аудит всех действий с данными.
- Соответствие требованиям 152-ФЗ "О персональных данных".

Шаг 2. Выбор модели хранилища данных

- **Data Lake:** хранение необработанных данных в едином репозитории.
- **Data Warehouse:** хранение структурированных данных, оптимизированных для аналитики.
- **Hybrid Data Storage:** сочетание Data Lake и Data Warehouse.

2. Архитектура хранилища больших данных

2.1 Компоненты архитектуры

Источники данных

- Веб-сайты и мобильные приложения.
- Социальные сети.
- CRM системы.
- Внешние API (например, данные о погоде, экономические показатели).

Слой сбора данных

- Apache Kafka для потоковых данных.
- Logstash для сбора логов.
- Пользовательские коннекторы для CRM и внешних API.

Слой хранения данных

- HDFS (Hadoop Distributed File System) для хранения сырых данных.
- Apache HBase для быстрого доступа к большим объемам данных.
- PostgreSQL для структурированных данных и метаданных.

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.

Слой аналитики и машинного обучения

- Jupyter Notebooks для интерактивной аналитики.
- TensorFlow и PyTorch для моделей машинного обучения.
- Apache Superset для визуализации и дашбордов.

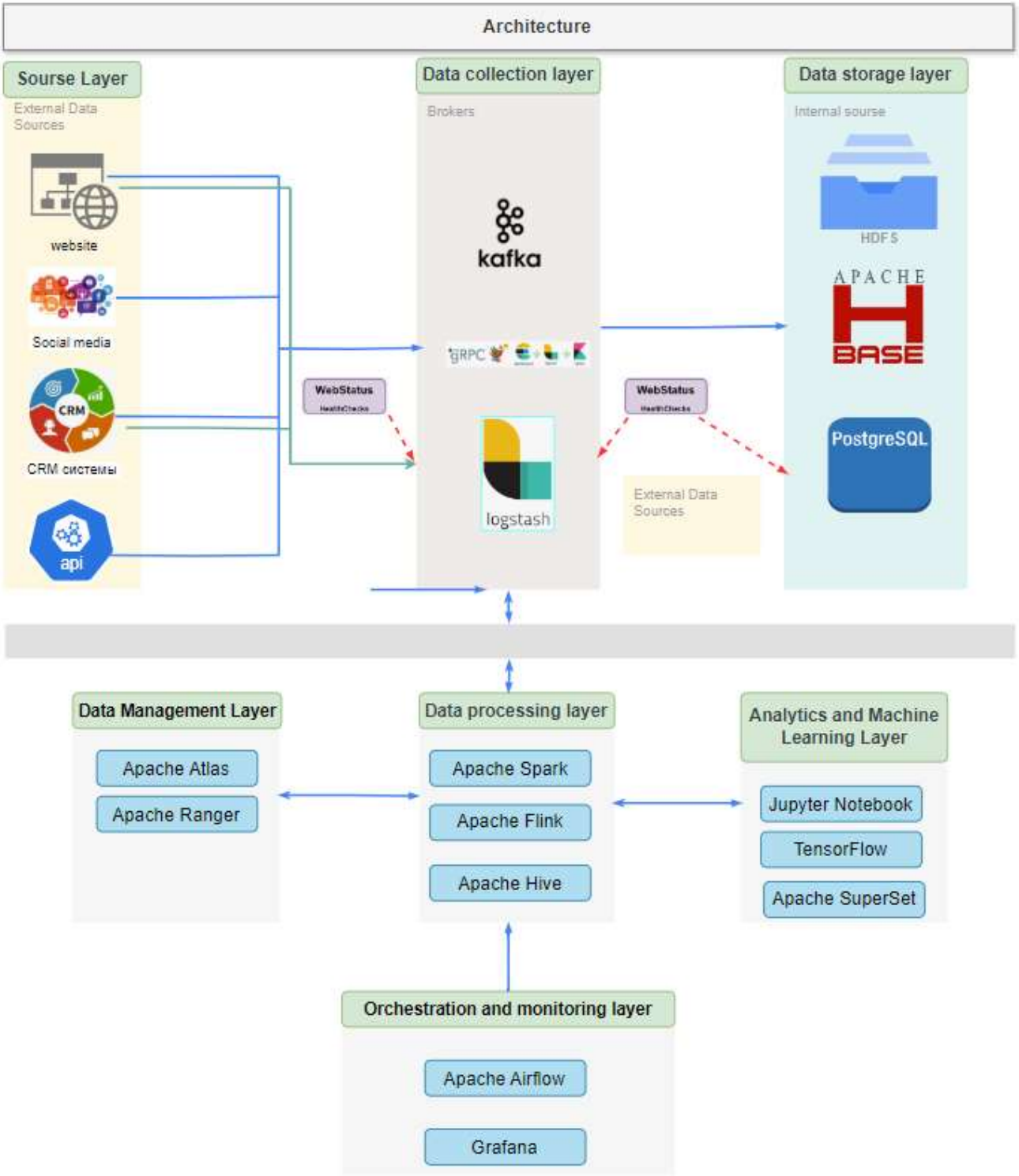
Слой управления данными

- Apache Atlas для управления метаданными.
- Apache Ranger для контроля доступа и аудита.

Слой оркестрации и мониторинга

- Apache Airflow для оркестрации рабочих процессов.
- Prometheus и Grafana для мониторинга и алертинга.

3. Схема архитектуры



4. Процесс обработки данных

- Данные собираются из различных источников через слой сбора данных.
- Сырые данные сохраняются в HDFS для долгосрочного хранения.
- Потоковые данные обрабатываются в реальном времени с помощью Flink для быстрой аналитики.
- Пакетные задачи, такие как сегментация клиентов, выполняются с помощью Spark по расписанию.
- Результаты анализа сохраняются в HBase для быстрого доступа.
- Аналитики используют Jupyter Notebooks и Superset для исследования данных и создания отчетов.
- Модели машинного обучения обучаются на исторических данных и развертываются для прогнозирования и рекомендаций.

5. Масштабирование и отказоустойчивость

- Использование кластерной архитектуры Hadoop для горизонтального масштабирования.
- Репликация данных в HDFS и HBase для обеспечения отказоустойчивости.
- Использование Kubernetes для оркестрации и автоматического масштабирования микросервисов.

6. Безопасность

- Реализация шифрования данных с помощью HDFS Transparent Encryption.
- Использование Kerberos для аутентификации.
- Применение Apache Ranger для детального контроля доступа к данным.
- Регулярное резервное копирование и план аварийного восстановления.

Основные этапы выполнения работы включают

1. Анализ требований.
2. Выбор компонентов архитектуры.
3. Проектирование архитектуры.
4. Создание диаграммы архитектуры.
5. Описание компонентов.
6. Обоснование выбора.
7. Рассмотрение вопросов производительности и масштабируемости.
8. Анализ потенциальных проблем и их решений.
9. Подготовка отчета.

Варианты заданий

Разработать архитектуру хранилища больших данных для компании, основываясь на предоставленных требованиях. Описать компоненты архитектуры, обосновать выбор технологий и предложить схему потока данных. Для создания диаграммы архитектуры студенты могут использовать **draw.io**, которое доступно как онлайн-инструмент и как приложение.

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.

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

Вариант 3. Телекоммуникационная компания
- Объем данных: 1 ПБ в год, рост 40% ежегодно.
- Скорость получения: до 10000 событий в секунду.
- Типы данных: 40% структурированные, 50% полуструктурированные, 10% неструктурированные.
- Требования к обработке: анализ качества связи, прогнозирование нагрузки на сеть.
- Доступность: 99.99%, время отклика <10 секунд.
- Безопасность: шифрование, соответствие 152-ФЗ и отраслевым стандартам.

Вариант 4. Средняя логистическая компания
- Объем данных: 50 ТБ в год, рост 25% ежегодно.
- Скорость получения: до 500 событий в секунду.
- Типы данных: 70% структурированные, 20% полуструктурированные, 10% неструктурированные.
- Требования к обработке: оптимизация маршрутов в реальном времени, прогнозирование спроса на услуги.
- Доступность: 99.9%, время отклика <30 секунд.
- Безопасность: базовое шифрование, соответствие 152-ФЗ.

Вариант 5. Крупная социальная сеть
- Объем данных: 2 ПБ в год, рост 60% ежегодно.
- Скорость получения: до 50000 событий в секунду.
- Типы данных: 20% структурированные, 30% полуструктурированные, 50% неструктурированные.
- Требования к обработке: анализ социальных связей, рекомендательные системы в реальном времени.
- Доступность: 99.99%, время отклика <3 секунды.
- Безопасность: продвинутое шифрование, строгое соответствие 152-ФЗ.

Вариант 6. Средняя медицинская организация
- Объем данных: 30 ТБ в год, рост 20% ежегодно.
- Скорость получения: до 100 событий в секунду.
- Типы данных: 50% структурированные, 30% полуструктурированные, 20% неструктурированные.
- Требования к обработке: анализ медицинских карт, прогнозирование заболеваний.
- Доступность: 99.99%, время отклика <20 секунд.
- Безопасность: сквозное шифрование, строгое соответствие 152-ФЗ и требованиям по защите медицинских данных.

Вариант 7. Крупный производственный холдинг
- Объем данных: 200 ТБ в год, рост 35% ежегодно.
- Скорость получения: до 2000 событий в секунду.
- Типы данных: 65% структурированные, 25% полуструктурированные, 10% неструктурированные.
- Требования к обработке: мониторинг производственных процессов в реальном времени, предиктивное обслуживание оборудования.
- Доступность: 99.9%, время отклика <1 минуты.
- Безопасность: шифрование, соответствие 152-ФЗ и отраслевым стандартам безопасности.

Вариант 8. Средняя страховая компания
- Объем данных: 80 ТБ в год, рост 30% ежегодно.
- Скорость получения: до 500 транзакций в секунду.
- Типы данных: 75% структурированные, 20% полуструктурированные, 5% неструктурированные.
- Требования к обработке: оценка страховых рисков, выявление мошенничества.
- Доступность: 99.95%, время отклика <15 секунд.
- Безопасность: шифрование, строгое соответствие 152-ФЗ и требованиям регулятора.

Вариант 9. Крупный образовательный портал
- Объем данных: 150 ТБ в год, рост 45% ежегодно.
- Скорость получения: до 3000 событий в секунду.
- Типы данных: 30% структурированные, 40% полуструктурированные, 30% неструктурированные.
- Требования к обработке: персонализация обучения, анализ успеваемости в реальном времени.
- Доступность: 99.9%, время отклика <10 секунд.
- Безопасность: шифрование, соответствие 152-ФЗ и требованиям к защите данных несовершеннолетних.

Вариант 10. Средняя энергетическая компания
- Объем данных: 120 ТБ в год, рост 25% ежегодно.
- Скорость получения: до 1500 событий в секунду.
- Типы данных: 70% структурированные, 25% полуструктурированные, 5% неструктурированные.
- Требования к обработке: мониторинг энергопотребления в реальном времени, прогнозирование нагрузки на сеть.
- Доступность: 99.99%, время отклика <5 секунд.
- Безопасность: продвинутое шифрование, соответствие 152-ФЗ и требованиям безопасности критической инфраструктуры.

Вариант 11. Крупная сеть отелей
- Объем данных: 80 ТБ в год, рост 35% ежегодно.
- Скорость получения: до 2000 событий в секунду.
- Типы данных: 55% структурированные, 35% полуструктурированные, 10% неструктурированные.
- Требования к обработке: персонализация обслуживания, прогнозирование загрузки.
- Доступность: 99.95%, время отклика <8 секунд.
- Безопасность: шифрование, соответствие 152-ФЗ и международным стандартам гостиничного бизнеса.

Вариант 12. Средняя компания в сфере кибербезопасности
- Объем данных: 200 ТБ в год, рост 60% ежегодно.
- Скорость получения: до 10000 событий в секунду.
- Типы данных: 30% структурированные, 50% полуструктурированные, 20% неструктурированные.
- Требования к обработке: анализ угроз в реальном времени, выявление аномалий.
- Доступность: 99.999%, время отклика <1 секунды.
- Безопасность: многоуровневое шифрование, строгое соответствие 152-ФЗ и международным стандартам безопасности.

Вариант 13. Крупный агропромышленный холдинг
- Объем данных: 50 ТБ в год, рост 25% ежегодно.
- Скорость получения: до 500 событий в секунду.
- Типы данных: 60% структурированные, 30% полуструктурированные, 10% неструктурированные.
- Требования к обработке: мониторинг урожайности, прогнозирование погодных условий.
- Доступность: 99.9%, время отклика <30 секунд.
- Безопасность: базовое шифрование, соответствие 152-ФЗ и отраслевым стандартам.

Вариант 14. Средняя биотехнологическая компания
- Объем данных: 300 ТБ в год, рост 50% ежегодно.
- Скорость получения: до 1000 событий в секунду.
- Типы данных: 40% структурированные, 40% полуструктурированные, 20% неструктурированные.
- Требования к обработке: анализ геномных данных, моделирование белковых структур.
- Доступность: 99.99%, время отклика <1 минуты.
- Безопасность: продвинутое шифрование, строгое соответствие 152-ФЗ и международным стандартам обработки биомедицинских данных.

Вариант 15. Крупная автомобильная компания
- Объем данных: 500 ТБ в год, рост 40% ежегодно.
- Скорость получения: до 5000 событий в секунду.
- Типы данных: 50% структурированные, 40% полуструктурированные, 10% неструктурированные
- Требования к обработке: анализ данных с подключенных автомобилей, оптимизация производственных процессов
- Доступность: 99.95%, время отклика <10 секунд
- Безопасность: шифрование, соответствие 152-ФЗ и международным автомобильным стандартам

Вариант 16. Средняя компания в сфере "умного города"
- Объем данных: 150 ТБ в год, рост 55% ежегодно
- Скорость получения: до 3000 событий в секунду
- Типы данных: 45% структурированные, 45% полуструктурированные, 10% неструктурированные
- Требования к обработке: управление городской инфраструктурой в реальном времени, анализ транспортных потоков
- Доступность: 99.99%, время отклика <5 секунд
- Безопасность: продвинутое шифрование, строгое соответствие 152-ФЗ и требованиям к критической инфраструктуре

Вариант 17. Крупная геологоразведочная компания
- Объем данных: 400 ТБ в год, рост 30% ежегодно.
- Скорость получения: до 1000 событий в секунду.
- Типы данных: 30% структурированные, 50% полуструктурированные, 20% неструктурированные.
- Требования к обработке: анализ сейсмических данных, 3D-моделирование месторождений.
- Доступность: 99.9%, время отклика <2 минуты.
- Безопасность: шифрование, соответствие 152-ФЗ и отраслевым стандартам.

Вариант 18. Средняя компания в сфере интернета вещей (IoT)
- Объем данных: 250 ТБ в год, рост 70% ежегодно.
- Скорость получения: до 20000 событий в секунду.
- Типы данных: 35% структурированные, 55% полуструктурированные, 10% неструктурированные.
- Требования к обработке: анализ данных с IoT-устройств в реальном времени, предиктивное обслуживание.
- Доступность: 99.999%, время отклика <2 секунды.
- Безопасность: многоуровневое шифрование, строгое соответствие 152-ФЗ и стандартам IoT-безопасности.

Вариант 19. Крупная компания в сфере виртуальной реальности (VR)
- Объем данных: 700 ТБ в год, рост 80% ежегодно.
- Скорость получения: до 8000 событий в секунду.
- Типы данных: 20% структурированные, 40% полуструктурированные, 40% неструктурированные.
- Требования к обработке: рендеринг VR-контента в реальном времени, анализ пользовательского опыта.
- Доступность: 99.99%, время отклика <10 миллисекунд.
- Безопасность: продвинутое шифрование, соответствие 152-ФЗ и стандартам обработки биометрических данных.

Вариант 20. Средняя компания в сфере экологического мониторинга
- Объем данных: 100 ТБ в год, рост 40% ежегодно.
- Скорость получения: до 2000 событий в секунду.
- Типы данных: 50% структурированные, 40% полуструктурированные, 10% неструктурированные.
- Требования к обработке: анализ загрязнения воздуха и воды в реальном времени, прогнозирование экологических рисков.
- Доступность: 99.95%, время отклика <15 секунд.
- Безопасность: шифрование, соответствие 152-ФЗ и экологическим стандартам.

Вариант 21. Крупная сеть фитнес-центров
- Объем данных: 40 ТБ в год, рост 30% ежегодно.
- Скорость получения: до 1500 событий в секунду.
- Типы данных: 65% структурированные, 25% полуструктурированные, 10% неструктурированные.
- Требования к обработке: анализ тренировок клиентов, персонализация фитнес-программ.
- Доступность: 99.9%, время отклика <20 секунд.
- Безопасность: шифрование, строгое соответствие 152-ФЗ и стандартам обработки медицинских данных.

Вариант 22. Средняя компания в сфере квантовых вычислений
- Объем данных: 180 ТБ в год, рост 65% ежегодно.
- Скорость получения: до 500 событий в секунду.
- Типы данных: 70% структурированные, 25% полуструктурированные, 5% неструктурированные.
- Требования к обработке: симуляция квантовых схем, оптимизация квантовых алгоритмов.
- Доступность: 99.99%, время отклика <1 минуты.
- Безопасность: продвинутое шифрование, соответствие 152-ФЗ и международным стандартам квантовой криптографии.

Вариант 23. Крупная сеть общественного питания
- Объем данных: 60 ТБ в год, рост 25% ежегодно.
- Скорость получения: до 3000 событий в секунду.
- Типы данных: 70% структурированные, 20% полуструктурированные, 10% неструктурированные.
- Требования к обработке: анализ предпочтений клиентов, оптимизация цепочки поставок в реальном времени.
- Доступность: 99.95%, время отклика <5 секунд.
- Безопасность: шифрование, соответствие 152-ФЗ и стандартам пищевой промышленности.

Вариант 24. Средняя компания в сфере блокчейн-технологий
- Объем данных: 350 ТБ в год, рост 75% ежегодно.
- Скорость получения: до 10000 транзакций в секунду.
- Типы данных: 80% структурированные, 15% полуструктурированные, 5% неструктурированные.
- Требования к обработке: верификация транзакций в реальном времени, анализ блокчейн-сети.
- Доступность: 99.999%, время отклика <1 секунды.
- Безопасность: криптографическая защита, соответствие 152-ФЗ и международным стандартам криптовалютных бирж.

Вариант 25. Крупная компания в сфере дополненной реальности (AR)
- Объем данных: 450 ТБ в год, рост 70% ежегодно.
- Скорость получения: до 15000 событий в секунду.
- Типы данных: 25% структурированные, 45% полуструктурированные, 30% неструктурированные.
- Требования к обработке: обработка AR-контента в реальном времени, анализ пользовательского взаимодействия.
- Доступность: 99.99%, время отклика <20 миллисекунд.
- Безопасность: многоуровневое шифрование, соответствие 152-ФЗ и стандартам обработки геолокационных данных.