**Вариант 3.2. Финансовая компания**

**Шаг 1: Определение требований**

* 1. **Объем данных**
     1. Ожидаемый объем 100 ТБ в год
     2. Ожидаемый рост: 30% ежегодно
  2. **Скорость получения данных:**

до 1000 транзакций в секунду.

* 1. **Типы данных:**
     1. Структурированные: транзакционные данные, клиентские профили, кредитные договоры, данные скоринга (кредитный рейтинг, платежеспособность), реестры счетов и вкладов. – 80%
     2. Полуструктурированные: заявки на кредит, данные внутренних аудиторских проверок (XML) – 15%
     3. Неструктурированные: сканы паспортов, договоров, справок о доходах, записи телефонных переговоров с клиентами. – 5%
  2. **Требования к обработке:** 
     1. Выявление мошенничества в реальном времени.
     2. Оценка кредитных рисков.
  3. **Доступность данных:** 
     1. Время отклика: <1 секунда.
     2. Доступность системы: 99.999%,
  4. **Безопасность данных:** 
     1. Сквозное шифрование.
     2. Соответствие 152-ФЗ и требованиям ЦБ РФ.

**Шаг 2: Выбор модели хранилища данных:**

Выбор схемы «Снежинка», т. к. она экономит место, что важно для данной компании, у которой по условию объем обрабатываемых и хранимых данных 100 ТБ в год, да еще и ожидается ежегодный рост этого значения на 30%. Также анализ может включать сложные иерархии и отношения.

**2.1. Источники данных**

⎯ Клиентские данные.

⎯ Счета.

⎯ Транзакции.

⎯ Кредитные договоры.

⎯ Логи операций.

⎯ Внутренние документы (сканы паспортов).

⎯ Шаблоны мошенничества.

⎯ История взаимодействий с клиентами, жалобы, заявки. – CRM**.**

⎯ Аудиозаписи разговоров, расшифровки (колл-центр).

**2.2. Слой сбора данных**

⎯ Apache Kafka – для сбора потоковых данных (транзакции, аудио) (высокая пропускная способность (десятки тысяч сообщений в секунду), горизонтальное масштабирование для роста объема данных, репликация, шифрование, открытое решение, рост нагрузки на 30% в год компенсируется добавлением партиций, гарантированная доставка за 10–50 мс).

⎯ Arenadata Streaming – для потоковой обработки, российская, с шифрованием, инструменты для ETL-преобразований в реальном времени и централизованным аудитом.

**2.3. Слой хранения данных**

⎯ PostgreSQL для структурированных и полуструктурированных данных. — надежная реляционная СУБД, для горячих транзакций, репликация, шифрование,

⎯ ClickHouse для аналитических запросов с высокой скоростью.

⎯ MinIO для неструктурированных, долгосрочное хранение логов, легко масштабируется.

⎯ Arenadata DB, долгосрочное хранение, масштабирование, для анализа, если не тянет ClickHouse

**2.4. Слой обработки данных**

⎯ Apache Flink – для маскирования персональных данных (есть и в Kafka, но необходимо определить, что за данные, например папорт или данные банковской карты, чтобы применить разные шаблоны маскирования, возможность напрямую интегрировать криптографические операции, фильтрация мусорных данных).

⎯ КриптоПро CSP – ЭЛ подпись логов (требование ЦБ РФ: Неизменяемость аудит-логов → подпись ЭП при записи).

⎯ Tinkoff VoiceKit – система распознавания и синтеза речи.

**2.5. Слой аналитики и машинного обучения**

⎯ Jupiter Notebook – для аналитик и ML. Стандарт для аналитиков и дата-сайентистов, поддерживает Python, ML-библиотеки (scikit-learn, TensorFlow, PyTorch).

⎯ Яндекс Даталенс – для визуализации и дашбордов. Бизнес-аналитика.

⎯ Grafana. Визуализация метрик из Prometheus (и других источников) в графиках/дашбордах. Мониторинг инфраструктуры / performance / SLA.

**2.6. Слой управления данными**

⎯ Arenadata Platform Security – для централизованного управления политиками безопасности (авторизация, маскировка данных, аудит).

**2.7. Слой оркестрации и мониторинга**

⎯ Apache Airflow – для оркестрации, объединения данных из разных СУБД.

⎯ Prometheus – для мониторинга и алертинга (система сбора метрик).

**Безопасность**

1. Шифрование: упомянуто на всех уровнях: при передаче с использованием сертифицированных ФСБ РФ алгоритмов через интеграцию с КриптоПро CSP, при хранении.
2. Выявление и учёт персональных данных (Arenadata Platform Security).
3. Контроль доступа к персональным данным.
4. Маскирование/Анонимизация/Токенизация персональных данных при загрузке, сохранении и доступе.
5. Неизменяемость Аудит-Логов.
6. Обработка персональных данных с согласия клиента.
7. Защита токенов.
8. Централизованное Управление Безопасностью и Аудит (Arenadata Platform Security).
9. Непрерывный Мониторинг и Реагирование.

**Отказоустойчивость:**

Tier IV ЦОД: N+2 Ресурсы: все компоненты (питание, охлаждение, ИБП, генераторы, сети) имеют минимум 2 независимых резервных линии, репликация, шардинг.

**Потянет рост данных**

Год 1: 130 ТБ

Год 2: 169 ТБ (+30%)

Год 3: 220 ТБ (+30%)

Рекомендации для поддержки роста:

1. Архивация в MinIO.
2. Шардинг PostgreSQL.
3. Автомасштабирование.

**Выводы:**

1. Объем данных (100 ТБ/год, рост 30% годовых): MinIO для архивов, шардинг PostgreSQL, автомасштабирование, использование ClickHouse/Arenadata DB для аналитики.
2. Скорость (1000 транз/сек): Kafka (десятки тысяч сообщ./сек) и Arenadata Streaming легко покрывают требуемую скорость приема. Гарантированная доставка в 10–50 мс Kafka удовлетворяет потребности потоковой обработки для мошенничества.
3. Мошенничество в РВ, оценка рисков: Apache Flink + Kafka идеальны для обработки транзакций на мошенничество. ClickHouse/Arenadata DB + Jupyter Notebook обеспечивают мощную базу для сложных аналитических моделей оценки рисков.
4. Доступность данных (<1с отклик, 99.999%): для оперативных запросов (PostgreSQL с индексами) и аналитики (ClickHouse). MinIO (S3 API) обеспечивает быстрый доступ к объектам, репликация, шардинг.
5. Безопасность данных (Шифрование, 152-ФЗ, ЦБ РФ): шифрование, управление персональными данными, аулит логов, ЭП.