# Лабораторная работа по теме “Выбор хранилища данных”

## Задача:

Вы являетесь аналитиком данных в небольшой компании, которая решает перейти к современной системе управления данными. Вам поручено выбрать и обосновать выбор хранилища данных для компании. Ваша компания работает в сфере e-commerce и продает товары через онлайн-платформу.

## У вас есть следующие требования:

### Объем данных:

Ваша компания собирает миллионы записей о продажах, клиентах и инвентаре ежедневно.

### Аналитика:

Вам необходимо обеспечить возможность проведения сложных аналитических запросов к данным для выявления тенденций продаж, сегментации клиентов и определения эффективности маркетинговых кампаний.

### Масштабируемость:

Ваша компания растет, и вам нужна система, которая легко масштабируется с ростом объема данных и нагрузки.

### Реальное время:

Вам нужна возможность анализа данных в реальном времени для мониторинга заказов и складской информации.

## Задачи:

1. Выберите одно или несколько хранилищ данных (реляционная база данных, NoSQL, колоночное хранилище и т. д.), которые, по-вашему мнению, наилучшим образом соответствуют требованиям компании.

2. Обоснуйте свой выбор, предоставив аргументы, почему данное хранилище данных является оптимальным для конкретных требований.

3. Опишите структуру данных, которые будут храниться в выбранном хранилище (таблицы, колонки и связи).

4. Предложите схему интеграции данных из разных источников в хранилище.

5. Опишите, как бы вы настроили систему для обработки данных в реальном времени.

### п. 1.

Считаю, что для данной задачи достаточно хранилища Apache Cassandra

### п. 2.

#### Мощность, устойчивость, масштабируемость

Одним из важных свойств Cassandra является то, что ее базы данных распределены.

Базы данных Cassandra легко масштабируются, когда приложение находится под высокой нагрузкой, а дистрибутив также предотвращает потерю данных из-за аппаратного сбоя любого центра обработки данных.

Распределенная архитектура также обеспечивает техническую мощь; например, разработчик может настроить пропускную способность запросов на чтение или запись изолированно.

Cassandra позволяет легко увеличить объем данных, которыми она может управлять. Поскольку Cassandra основана на узлах, она масштабируется горизонтально (так называемое горизонтальное масштабирование), используя менее распространенное оборудование. Также есть возможность уменьшить масштаб, если хотите.

Эта линейная масштабируемость применима практически бесконечно. Эта способность стала одной из ключевых сильных сторон Кассандры.

#### Производительность

В Cassandra сами данные распределяются автоматически, что оказывает (положительное) влияние на производительность. Это достигается с помощью разделов. Каждый узел владеет определенным набором токенов, и Cassandra распределяет данные на основе диапазонов этих токенов по кластеру. Ключ раздела отвечает за распределение данных между узлами и важен для определения локальности данных. Когда данные вставляются в кластер, первым шагом является применение хэш-функции к ключу раздела. Выходные данные используются для определения того, какой узел (на основе диапазона токенов) получит данные.

#### Надежность и отказоустойчивость

Один фрагмент данных может быть реплицирован на несколько узлов (реплик), обеспечивая надежность и отказоустойчивость. Cassandra поддерживает понятие коэффициента репликации (RF), который описывает, сколько копий данных должно существовать в базе данных.

Использование нескольких реплик также дает преимущества в производительности. Потому что мы не ограничены одним экземпляром. У нас есть три узла (реплики), к которым можно получить доступ для предоставления данных для наших операций, между которыми мы можем балансировать нагрузку для достижения наилучшей производительности.

#### Вывод

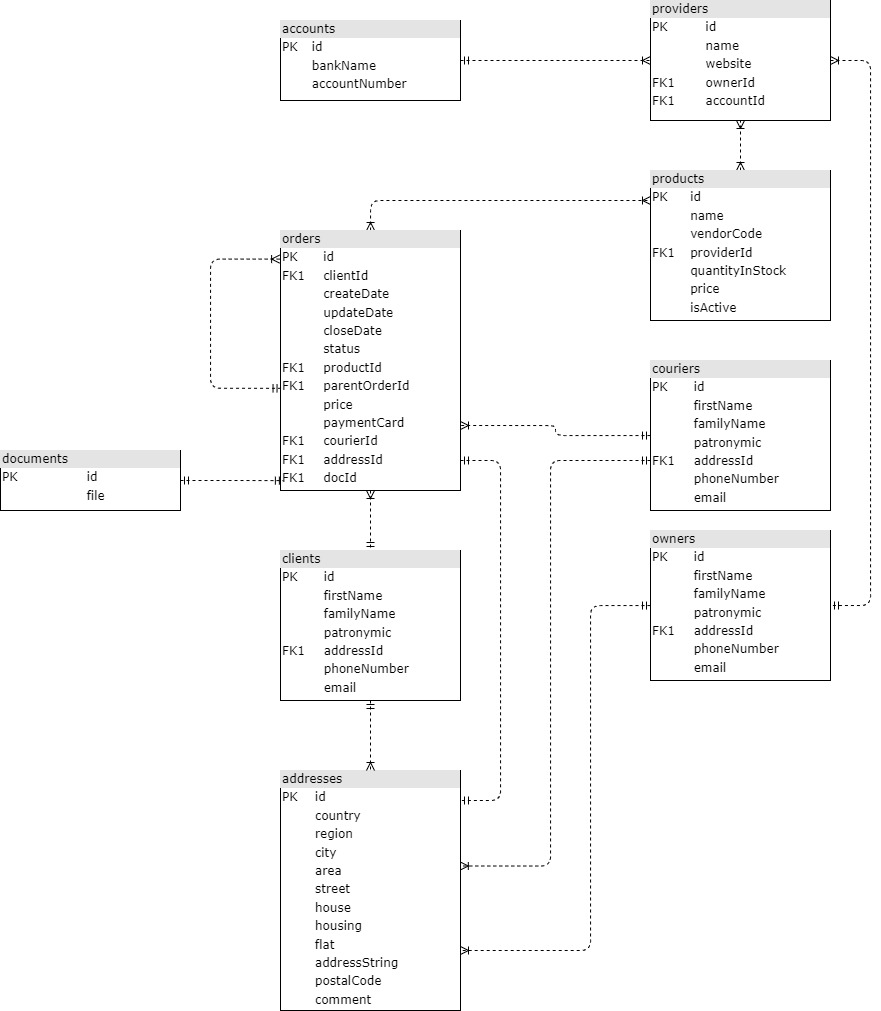
Учитывая все вышенаписанное, можно утверждать, что Cassandra полностью подойдет для поставленных задач, а именно:  
- Сможет хранить миллионы записей о продажах, клиентах и инвентаре ежедневно.

- Сможет обеспечить возможность проведения сложных аналитических запросов к данным для выявления тенденций продаж, сегментации клиентов и определения эффективности маркетинговых кампаний.

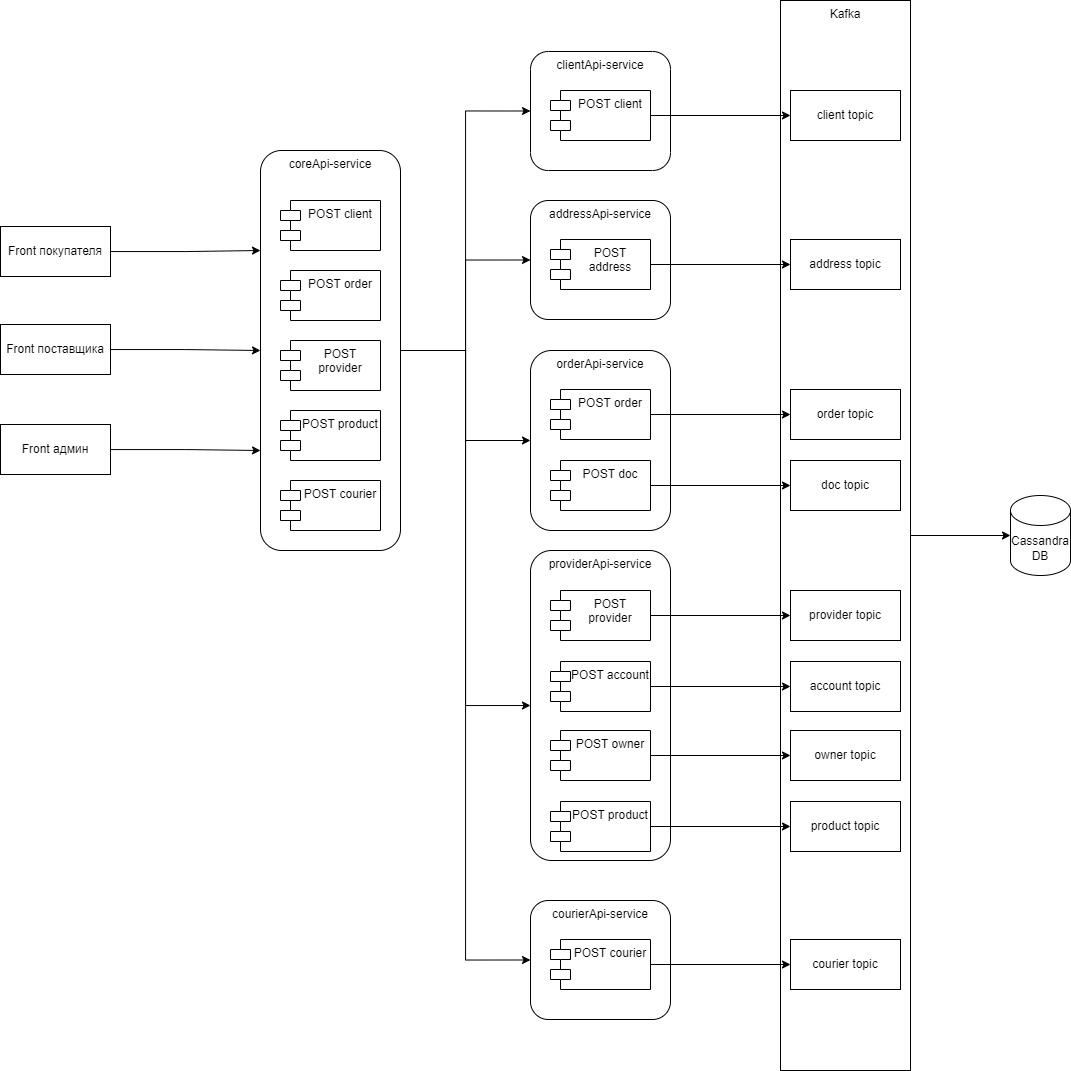
- Она легко масштабируется с ростом объема данных и нагрузки.

- Сможет предоставить данные для анализа в реальном времени для мониторинга заказов и складской информации.

### п. 3.



### п. 4.



### п. 5.

