

Теоретическая часть

Итоговая работа по курсу
“Big Data с нуля, группа abd-21”

Теплова Людмила, 08.2022

1. Основные бизнес-отчеты

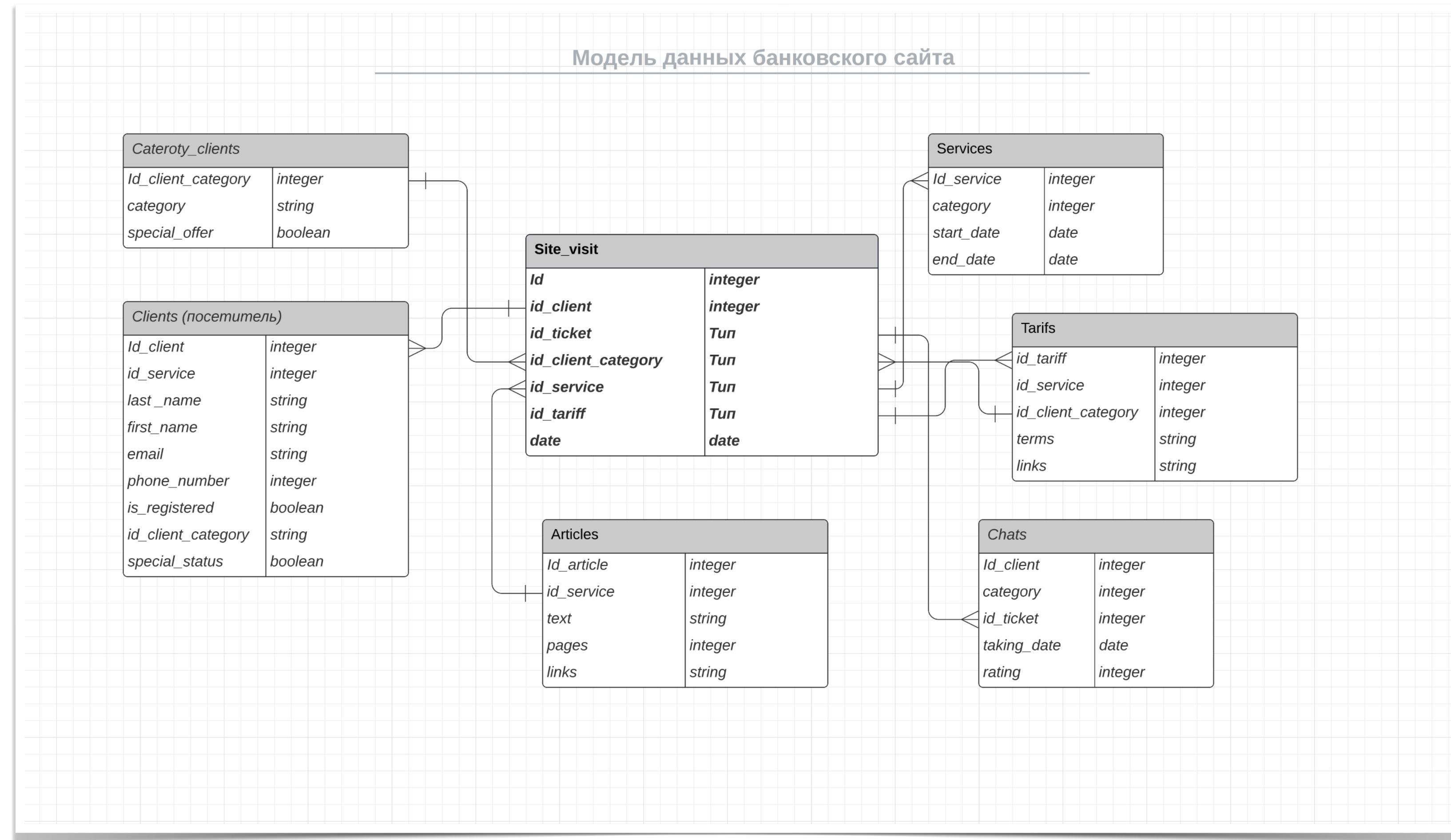
- Анализ эффективности маркетинговой политики через изменение (рост/спад) количества клиентов за последние 3 месяца
- Зависимость величины дохода от региона, зависимость количества клиентов от региона
- Регионы с наименьшим и наибольшим количеством клиентов
- Топ наиболее эффективных источников привлечения клиентов



2. Данные и источники их поступления

- Внутренние источники:
 - Корпоративные информационные системы (бд, хранилища)
 - Данные оборудования
 - Данные соцопросов
 - Внутренняя экспертиза
- Внешние источники:
 - Веб-аналитика
 - Порталы открытых данных
 - Веб-скрейпинг
 - Провайдеры платных услуг

3. Сущности в хранилище данных



Site_visit - таблица фактов, связана со справочниками через идентификаторы

4. Проверки на качество данных, используемые при заливке

- В таблице Clients проверка корректности введенного email (по шаблону *@*.*)
- В таблице Clients проверка корректности введенного телефонного номера (по шаблону 7XXXXXXXXX)
- В таблице Clients проверка заполненности поля с фактом регистрации
- В таблице Clients проверка заполненности поля с Фамилией, если есть Имя
- В таблице Visits проверка корректности даты и преобразование к нужному виду
- В таблице Client_Categories проверять на пустые значение поле special_offer
- В таблице с услугами (Services) проверять корректность даты начала действия услуги и окончания действия услуги
- В таблице с чатами проверять корректность даты
- В таблице с чатами проверять корректность введеной оценки (по шкале 1-5, должна присутствовать цифра от 1 до 5)
- В таблице Clients должны быть заполнены все строки с категорией, отсутствие пустых значений

5. Data-проект для улучшения показателей бизнеса

Построить модель, предсказывающую подключение посетителям сайта премиум подписки на получение дополнительных услуг с бонусами и скидками, на использование партнерских сервисов

Данные

- 1.Информация по пользователям, их статусе, тарифе
- 2.Информация по статьям для привлечения к подключению премиум подписки
- 3.Информация по использованию партнерских сервисов, включенных в премиум подписку



Шаги CrispDM

Business Understanding

Что планируем делать:

- Вывести метрику конверсии прочтения предложения о подключении премиум подписки в подключенную подписку
- Определить типичные значения, и взаимосвязь с использование клиентом сторонних партнерских сервисов

Каких результатов ожидаем:

- Определить как заказчик видит использование полученной модели, сформулировать минимально необходимое качество
- Оценить ожидаемый эффект от такой модели и сравнить его с ожидаемыми трудозатратами

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется Аналитик и Владелец продукта

Шаги CrispDM

Data Understanding

Что планируем делать:

- Собрать имеющиеся данные (внутренние данные (внутри компании - информация по пользователям, по подключениям услуг и подпискам, внешние данные (метрики) из открытых источников (web-аналитика))
- Описать данные (посмотреть основные статистики, оценить объем данных и достаточность ресурсов для обработки)
- Исследовать данные (объединить данные, увидеть скрытую информацию, построить визуализации, посмотреть зависимости "пути" знакомство с преимуществами премиум подписки - список партнерских сервисов - подключение премиум подписки)
- Определить качество данных (посмотреть на выбросы, посмотреть на типы данных, заполненность данных - пропуски, ошибки)

Каких результатов ожидаем:

- Заключение пригодность данных для анализа
- Инсайты
- Определить способы заполнения (обогащения данных)

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется Аналитик

Шаги CrispDM

Data preparation

Что планируем делать:

- Отобрать данные (например, взять только зарегистрированных пользователей)
- Заполнить пропуски, исправить ошибки, кодировки в данных и провести контроль качества данных (возможно частичное возвращение на предыдущий шаг)
- Обогатить данные и добавить новые признаки (метрики, полученные на первом шаге)
- Провести контроль качества данных (возможно частичное возвращение на предыдущий шаг Data Understanding)
- Перевести строковые данные с признаками в вектора
- Разделить данные на тестовую и тренировочную выборки

Каких результатов ожидаем:

- Получить dataframe для обучения модели
- Получить документацию по проведенным работам с данными

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется data engineer или data scientist

Шаги CrispDM

Modeling

Что планируем делать:

- Выбрать модель для обучения, построить план тестирования (определить количество итераций обучения, оптимизация работы)
- Построение модели
- Оценка результатов (тех анализ качества модели, готова ли модель к внедрению, достигаются ли критерии качества)

Каких результатов ожидаем:

- Получить обученную модель предсказаний с хорошим результатом
- Описать причины предпочтения выбора данной модели (исследования результатов разных применения разных результатов)
- Получить ответ на вопрос готова ли модель к внедрению

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется data scientist

Шаги CrispDM

Evaluation

Что планируем делать:

- Оценка результатов с точки зрения бизнес-целей (выявление инсайтов, полезных бизнесу, поиск вопросов), резолюция
- Контроль качества (выявление мест для оптимизации и получения более эффективных результатов, обсуждение ошибок, анализ несработавших гипотез, неожиданные результаты)

Каких результатов ожидаем:

- Принять решение - разрешение для перехода к следующему шагу (Deployment) или наоборот откат на предыдущий шаг с корректировками запросов

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется Аналитик

Шаги CrispDM

Deployment

Что планируем делать:

- Составить план внедрения модели
- Определить принципы мониторинга (какие показатели будут отслеживаться, какие признаки устаревания модели)
- Составить план для устранения сбоев при внедрении (быстрый откат)
- План поддержки актуальности модели
- Разработать стратегию технического обслуживания модели
- Подключить модель к рабочей системе
- Оценка результатов на “production” системе

Каких результатов ожидаем:

- Получить работающую систему с новым подходом, позволяющую разрабатывать и оценивать подходы для увеличения конверсии корзины в покупку
- Получить заключение об эффективности работы новой системы (оптимизации работы интернет-магазина)

Какие роли специалистов необходимы:

- Для успешного выполнения данного этапа потребуется Разработчик и Data scientist

6. Роли для реализации Data проекта

- аналитик
- владелец продукта
- data engineer
- data scientist
- разработчик

