Теоретическая часть

Вы - главный по данным в среднем по объему просмотров интернет-кинотеатре. Ваша задача разработать стратегию внедрения хранилища данных и работы с большими данными в этой компании.
Задания:

1. Описать основные бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу

- Анализ фильмов:
  - дата выхода
  - жанр
  - количество просмотров
  - рейтинг
  - себестоимость фильма

- Анализ пользователей:
  - возраст
  - пол
  - местонахождение
  - активность
  - история просмотров
  - предпочтения
  - транзакции

- Анализ продаж:
  - продажи за период
  - продажи по клиентам
  - продажи по фильмам и жанрам
  - прибыль
  - затраты

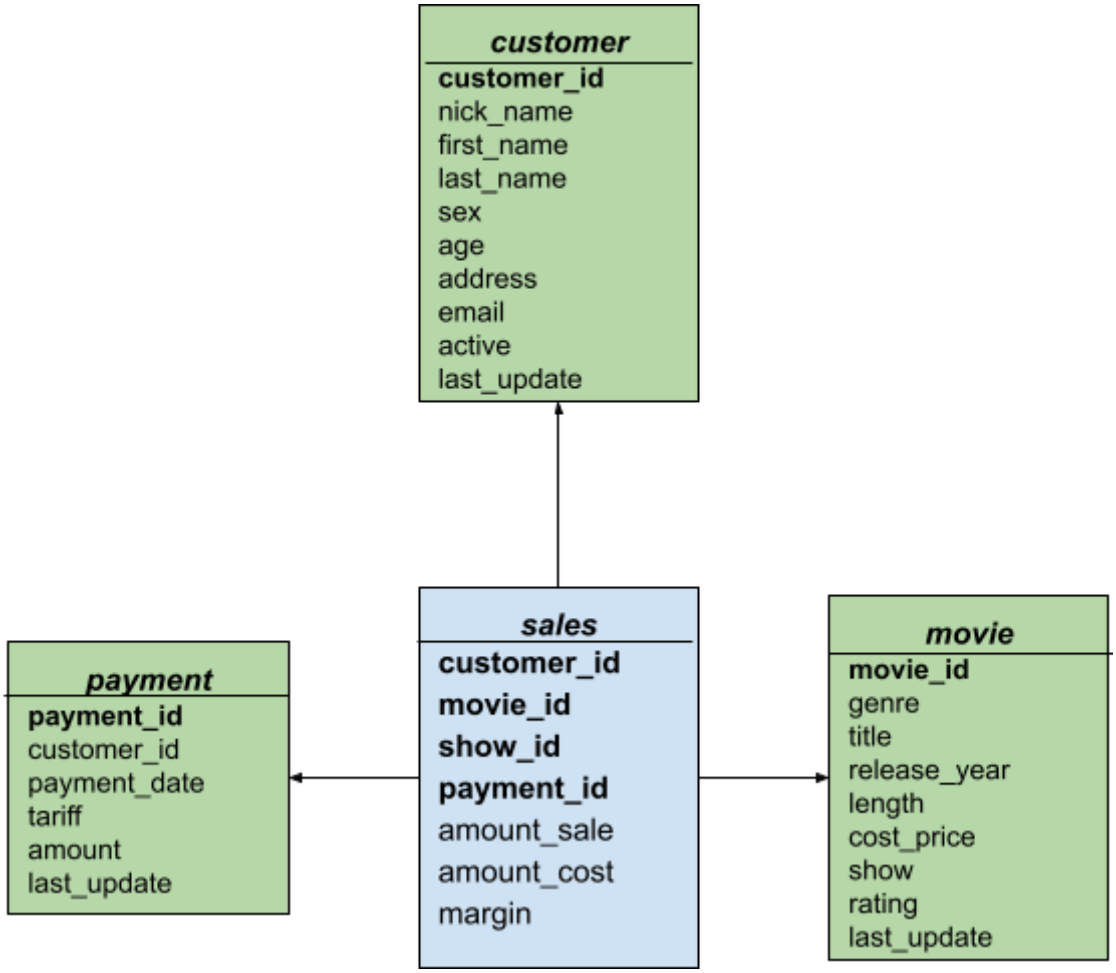
2. Описать основные имеющиеся данные и источники их поступления

- Фильмы: описательная информация по фильмам, их рейтинг, количество показов (сайт интернет - кинотеатра)
- Пользователи: данные по аудитории, история действий (сайт интернет - кинотеатра, CRM - система)
- Продажи: количество продаж; объем продажи в разрезе по фильмам, по покупателям, по периодам (1С)

3. Описать основные сущности в хранилище данных (схема звезда) и процесс заливки данных

- Таблица фактов:
  sales (продажи), которая состоит из следующих столбцов: customer_id (идентификатор пользователя) movie_id (идентификатор фильма); show_id (идентификатор показа), payment_id (идентификатор платежа); amount_sale (сумма в ценах продажи); amount_cost (сумма в себестоимости); margin (прибыль)
- Таблицы измерений:
  - ***customer*** (пользователь) содержит: customer_id (идентификатор пользователя); nick_name (ник); его имя и фамилию (first_name и last_name); sex (возраст); age (возраст); address (адрес); email (электронная почта); active (действующий или нет), last_update (дата обновления записи)

- ***movie*** (фильмы) содержит: movie_id (идентификатор фильма); genre (жанр); title (название фильма); release_year (дата выхода фильма); length (длина фильма), cost_prise (себестоимость фильма); show (количество показов); rating (рейтинг), last_update (дата обновления записи)
- ***payment*** (платежи) содержит: payment _id (идентификатор платежа); payment_date (дата платежа), tariff (тарифный план), amount (сумма платежа); customer_id (идентификатор пользователя), last_update (дата обновления записи)



Заливка данных будет осуществляться с помощью ETL-процесса. Сначала данные извлекаются из исходной системы (сайт, 1С, CRM), затем данные очищаются и преобразовываются, далее загружаются в хранилище данных.

4. Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке.

- Отсутствующие данные по полям таблиц
- Неверный тип данных по всем числовым полям
- Повторяющиеся данные, дубли по именам и названиям
- Невалидные данные по адресам и email
- Неверный диапазон данных в полях по полям с датами и возрастом
- Орфографические ошибки в написании названий жанров, имен и названий фильмов
- Нарушение разрядности в числовых полях
- Наличие посторонних данных в полях с определенным значение (например, пол, жанр и т.п.)
- Ошибочные данные (например возраст и длина фильма не входит в установленный диапазон)
- Данные, хранящиеся в неправильном поле (фамилия в имени и т.п.)

5. Придумать Data-проект, который должен улучшить показатели Вашего бизнеса и расписать его по Crisp-DM

.

Разработать модель, которая правильно подбирает фильмы и сериалы для каждого конкретного пользователя

Business understanding

- Определить кто принимает ключевые решения, кто финансирует, кто будет основным пользователем)
- Определить бизнес-цель (с помощью системы рекомендации привлечь зрителя и увеличить прибыль)
- Оценить наличие существующих решений. (Есть ли уже разработанные модели, почему они не отвечают намеченной цели)
- Оценить ресурсы для реализации. (Качество ПО, достаточно ли данных, наличие специалистов)
- Описать риски (нарушение сроков реализации, проблемы с финансированием, проблемы качества и количества данных)
- Предварительно оценить ROI
- Определить критерии оценки модели (минимальный и оптимальный)
- Согласовать ожидаемое качество с заказчиком
- Составить план проекта

Data understanding

- Проанализировать все имеющиеся данные, оценить достаточно ли их, при необходимости организовать сбор новых данных.
- Описать структуру источников данных (Таблицы, ключи, размер данных, количество строк и столбцов).
- Рассчитать ключевые статистики, собрать данные по событиям.
- Исследовать данные. Зафиксировать аномалии данных, распределение. Отметить потенциально полезные атрибуты.
- Оценить качество данных. Проверить пропущенные значения, форматы, релевантность дат. Проверить данные на ошибки, опечатки и единообразие ввода данных.

Data Preparation

- Отобрать данные для обучения модели: определить какие данные потенциально имеют отношение к проверяемой гипотезе, достаточно ли они качественны.
- Выполнить очистку данных: заполнить или удалить пустые значения, исправить опечатки, исключить дубли
- Выполнить генерацию признаков из визитов и контента.
- Выполнить конвертацию типов данных, нормализовать числовые данные.
- Произвести интеграцию (выполнить слияние информационных систем по зрителям, фильмам, транзакциям и пр.)
- Выполнить форматирование, которое требуется для инструментов моделирования.

Modeling

- Определить какие модели использовать
- Сравнить качество разных моделей и выбрать лучшую комбинацию
- Обучить модель
- Провести анализ качества модели
- Оценить готова ли модель к внедрения
- Оценить результат с технической точки зрения.

Evaluation

- Оценить результат с точки зрения бизнес-цели (анализ количественных характеристик)
- Пилотный запуск выбранной модели
- Проанализировать ход проекта, эффективность шагов и допущенные ошибки
- Определить следующие шаги (внедрять или улучшать модель)

Deployment

- Разработать план по внедрению полученной модели и подготовить технический проект внедрения функционала
- Настроить мониторинг и запланировать поддержку
- Составить отчет о результатах моделирования.

6. Описать требуемые роли в команде по работе с данными на этапах 4 и 5

На этапе Modeling необходимы компетенции DS-специалиста

На этапе Evaluation - Аналитик