Теоретическая часть

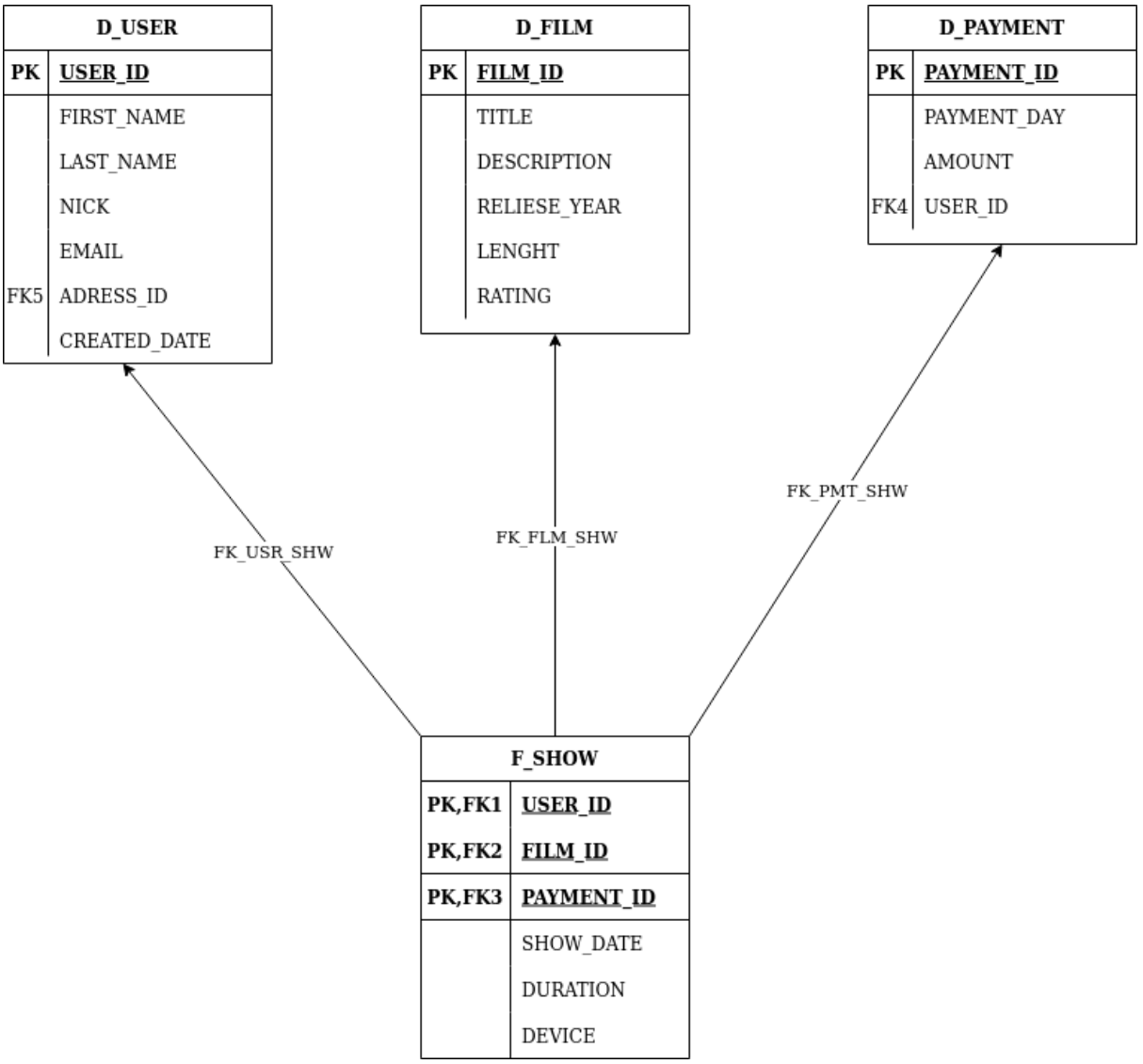
Вы - главный по данным в среднем по объему просмотров интернет-кинотеатре. Ваша задача разработать стратегию внедрения хранилища данных и работы с большими данными в этой компании. Задания:

1. Описать основные бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу
    - Отчёт по аудитории интернет-кинотеатра (
        - пол,
        - возраст,
        - местонахождение,
        - предпочтения,
        - с каких устройств в основном подключаются,
        - платежи
        - и т.п)
    - Отчёт по продажам (
        - контрольные цифры по продажам, показывающие достигнуты ли поставленные цели,
        - прибыль и издержки,
        - прогноз продаж на предстоящий период времени (месяц, квартал, год),
        - количество лидов и коэффициент конверсии за определённый период)
    - Отчет по интернет-маркетингу (
        - обзор используемой стратегии интернет-маркетинга,
        - основные цели маркетинга и информация об их достижении,
        - обзор тенденций: коэффициент конверсии, оплаченный и органический трафик, средняя цена за конверсию,
        - обзор показателей трафика по маркетинговым каналам,
        - обзор SEO-метрик, включая изменения поисковых позиций по ключевым словам,
        - обзор каналов в соцсетях, включая показатели вовлеченности и количество лидов с каждого канала)
2. Описать основные имеющиеся данные и источники их поступления

    - Информация по пользователям, истории их просмотров, покупок, поиска, используемых устройств (Источник данных: сайт интернет-кинотеатра, CRM, 1С)
    - Информация по фильмам, контенту (Источник данных: сайт интернет-кинотеатра)
    - Информация по каналам продвижения интернет-кинотеатра (Источники данных: порталы Я.Метрика, Google Analytics, рекламные площадки и т.п)

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

Для примера в качестве основной сущности для таблицы фактов возьмём “показ фильма” (SHOW), для таблиц измерений “пользователь” (USER), “фильм” (FILM), “платёж” (PAYMENT).



Показ фильма (F_SHOW)		
Аттрибут	Тип данных	Описание
USER_ID	FOGERNER KEY	идентификатор пользователя
FILM_ID	FOGERNER KEY	идентификатор фильма
PAYMENT_ID	FOGERNER KEY	идентификатор платежа

SHOW_DATE	DATE	дата показа
DURATION	INT	длительность просмотра в минутах
DEVICE	VARCHAR(25)	устройство для просмотра
…	…	…

Пользователь (D_USER)		
Аттрибут	Тип данных	Описание
USER_ID	SERIAL PRIMARY KEY	идентификатор пользователя
FIRST_NAME	VARCHAR(30)	имя пользователя
LAST_NAME	VARCHAR(30)	фамилия пользователя
NICK	VARCHAR(30)	псевдоним
EMAIL	VARCHAR(30)	эл.почта
ADDRESS_ID	FOGERNER KEY	идентификатор адреса пользователя
CREATED_DATE	DATE	дата создания аккаунта
…	…	…

Фильм (D_FILM)		
Аттрибут	Тип данных	Описание
FILM_ID	SERIAL PRIMARY KEY	идентификатор фильма
TITLE	VARCHAR(255)	название фильма
DESCRIPTION	TEXT	описание фильма
RELEASE_YEAR	INT	год выпуска фильма
LENGHT	INT	длительность фильма в минутах
RATING	FLOAT	рейтинг фильма по 10-бальной шкале
…	…	…

Платёж (D_PAYMENT)		
Аттрибут	Тип данных	Описание
PAYMENT_ID	SERIAL PRIMARY KEY	идентификатор платежа
PAYMENT_DATE	DATE	дата платежа
AMOUNT	FLOAT	сумма платежа
USER_ID	FOGERNER KEY	идентификатор пользователя совершившего платёж
…	…	…

4. Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке
    01. В справочнике ‘Пользователь’ у каждого человека должен быть заполнен email-адрес. Это должен быть валидный адрес, уникальный в рамках системы
    02. В справочнике ‘Пользователь’ у каждого пользователя должен быть указан ник. Значение должно быть уникальным в рамках системы
    03. В справочнике ‘Фильм’ поле LENGHT (длительность фильма) должно быть указана в минутах, в виде целочисленного положительного значения
    04. В справочнике ‘Фильм’ поле RATING (рейтинг фильма) должно содержать положительное значение от 0 до 10, и иметь не более 2-х знаков после запятой
    05. В справочнике ‘Фильм’ поле RELEASE_YEAR (год выхода фильма) должно содержать четырёхзначное целочисленное положительное значение
    06. В справочнике ‘Платёж’ поле AMOUNT (сумма платежа) должно содержать числовое значение, и иметь не более 2-х знаков после запятой
    07. В справочнике ‘Платёж’ обязательно должен быть указан идентификатор пользователя, совершившего оплату
    08. В справочнике ‘Платёж’ у каждого платежа должна быть заполнена дата совершения платежа. Это должна быть валидная дата
    09. В основной таблице ‘Показ фильма’ поле DEVICE (устройство для просмотра) должно быть указано одно из 3-х значений: “ПК”, “планшет”, “мобильный телефон”.
    10. В основной таблице ‘Показ фильма’ поле DURATION (длительность просмотра) должно быть указана в минутах, в виде целочисленного положительного значения

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

Для примера, необходимо создать модель, предсказывающую оформит ли пользователь ежемесячную подписку на сайте интернет-кинотеатра.

Business understanding

На этом шаге планируется выявить основную боль бизнеса. В том числе получить ответы на следующие вопросы:

- Вывести метрику конверсии просмотра условий интернет-подписки в покупку. Определить типичные значения, проверить наличие сезонности
- Определить как заказчик видит использование полученной модели, сформулировать минимально необходимое качество
- Оценить ожидаемый эффект от такой модели и сравнить его с ожидаемыми трудозатратами

Data understanding

На этом шаге собираем и изучаем имеющиеся у нас данные.

В результате чего:

- Собираем все имеющиеся данные воедино
- Описываем данные (объём данных, типы значений и т.п.)
- Исследуем данные (что также может помочь сформулировать гипотезы и оформить задачи по преобразованию данных, выполняемые во время подготовки данных)
- Проверяем качество данных. Корректно и полно ли представлена информация о товарах, пользователях и их покупках и т.д.

Data preparation

На этом шаге производится подготовка данных к дальнейшему моделированию.

В результате чего:

- Производим отбор данных. Например, база данных по пользователям будет содержать конфиденциальную информацию о пользователях интернет-кинотеатра, поэтому важно отфильтровать такие атрибуты, как имя покупателя, его адрес, телефон и номера банковских карт.
- Очищаем данные.
- Строим новые данные. Например, интернет-кинотеатр может пожелать, чтобы для событий, записываемых в журналах, создавались отметки времени, определялись посетители и сеансы и отмечалась посещаемая страница и представляемый событием тип операций
- Сохраняем данные в датафрейм для обучения модели

Modeling

На этом шаге планируется создать и обучить модель.

В результате чего:

- Выбираем конкретный способ моделирования на основе имеющихся данных и целей проводимого анализа.
- Строим модель.
- Оцениваем модель

Evaluation

На этом шаге планируется оценить результаты проведенных исследований при помощи критериев успешности бизнеса, установленных в самом начале проекта.

В результате чего:

- Оцениваем результаты
- Просматриваем процесс. Изучаем успехи и слабые места только что завершенного процесса
- Определяем следующие шаги. Продолжить внедрение или вернуться назад и уточнить/заменить модели

Deployment

На этом шаге планируется внедрение полученных результатов для внесения усовершенствований в организацию - интернет-магазин.

Для этого проводим:

- Планируем внедрение и составляем план для каждой модели, а также определяем возможные проблемные моменты
- Планируем мониторинг и техобслуживание
- Составляем итоговый отчёт по проекту

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

Этап Crisp-DM	**Специалист**
Business understanding	Аналитик и Владелец продукта
Data understanding	Аналитик

Data preparation	DS, DE
Modeling	DS
Evaluation	Аналитик
Deployment	DS, Разработчик

Также для сбора данных и подготовки выгрузок может понадобиться ETL-специалист. При разработке и настройке баз данных по сбору качественных данных, соответствующих проверкам п.4, понадобятся специалист-разработчик БД и разработчик.