1. **Основные бизнес-отчеты**
2. Отчет о статусе проекта внедрения системы хранилища данных и работы с большими данными

В отчете должны быть отражены:

- общий обзор состояния разных стадий проекта (например, закупка оборудования – выполнено на 70%, найм специалистов – выполнено на 50%, закупка и установка ПО – выполнено на 40%)

- основные вехи проекта и % их выполнения (или прогнозный срок полного выполнения на текущую дату) и сведения о графике реализации процедур внедрения

- бюджетная составляющая: расход, отношение к плановому

- описаны возникшие проблемы, методы, которыми их решили; или текущие проблемы, которые еще не решены и предложены способы устранения

1. Конкурентный и маркетинговый анализ

В отчете должны быть отражены:

- наименование и анализ конкурентных онлайн-кинотеатров (наименование, сайт, выручка, репутационный статус, количество пользователей)

- анализ веб-сайтов конкурентов: чем они привлекают пользователей, полезные «фичи»

- SWOT-анализ своего онлайн-кинотеатра с учетом сравнения с конкурентами

- анализ социальный сетей на предмет упоминаний в блогах, постах, рекламных записях (как наши, так и конкурентные)

- обзор и прогноз рынка онлайн-кинотеатров

- customer analysis: что нужно потребителю от онлайн-кинотеатра, какие устройства используют для просмотра, готовы ли (и сколько) платить за подписку

1. Отчеты по бюджету

В отчете должны быть отражены:

- различные категории бюджета (сегментирование на цели продвижения, покупки материальных и нематериальных активов)

- расходы за прошедший квартал или месяц по каждой категории бюджета

- области, для которых необходимо сократить или увеличить расходы

- бюджетный прогноз на следующий год или квартал

- отношение план/факт или текущий период к прошлому периоду

1. **Основные имеющиеся данные и источники из поступления**
2. Информация о фильмах, включая длительность, жанр, режиссера и т.д. Поступает от владельца фильма в описании к нему (например, киностудия при выпуске фильма заранее уже сообщает жанр, список актеров и многие другие показатели)
3. Информация о пользователях, включая регистрационные данные и некоторую личную информацию. Поступает от пользователя при регистрации на странице онлай-кинотеатра
4. Информация об оценках пользователей, поставленных разными людьми разным фильмах, а также словесные отзывы. Поступает от пользователей, которые на странице фильма пишут текст и ставят оценки
5. Финансовая информация – смотрит ли пользователь фильм по подписке, покупает разово, бывают скидки, акции, бесплатные фильмы. Поступает со страницы заказа (покупки) фильма
6. Информация по поведению пользователей: сколько раз он просмотрел фильм после покупки, досмотрел ли до конца, как часто заходит в кинотеатр, какие жанры предпочитает и т.д. Поступает из разных источников, но правильнее, наверное, сказать, что она не поступает, а собирается и обрабатывается, постоянно обновляется для каждого просмотренного фильма
7. **Основные сущности в хранилище данных и процесс заливки данных**

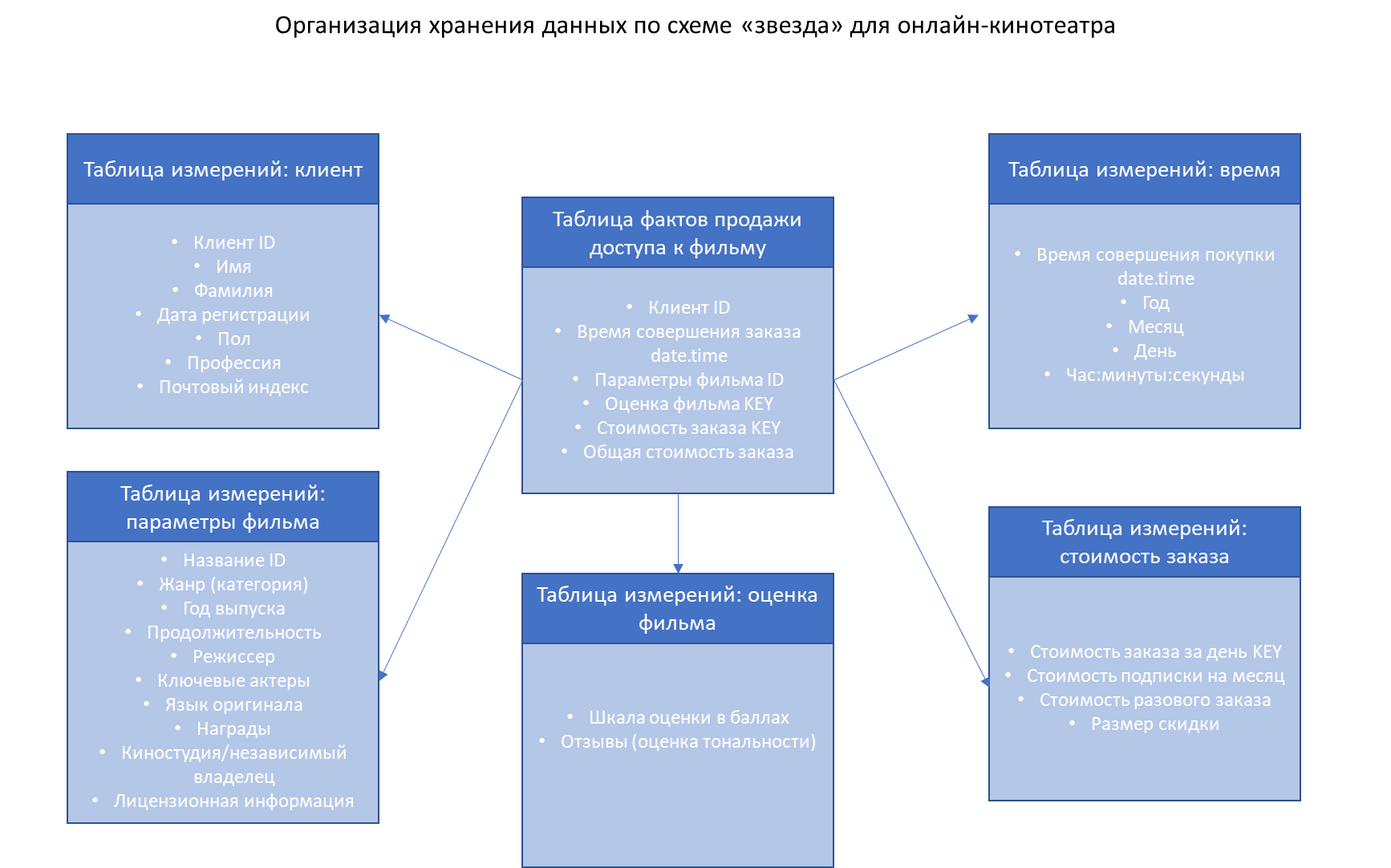


Схема «звезда» для онлайн-кинотеатра содержит колонку фактов (факт продажи доступа к фильму) и несколько колонок-ключей. Таблица фактов содержит основные сведения о том, кто, когда, почем купил какой фильм и с какой оценкой. При запросе к таблице фактов она с свою очередь обращается к таблицам измерений, причем разные параметры из разных таблиц измерений можно группировать для составления более точного запроса.

При регистрации пользователь указывает о себе сведения, позволяющие его однозначно идентифицировать среди всех остальных пользователей. Информация о пользователях хранится в своей таблице измерений.

При добавлении фильма хранилище онлайн-кинотеатра ему присваиваются параметры, позволяющие однозначно найти данный фильм, такие как название, год выпуска, режиссер и т.д. Для повышения пользовательского интереса к фильму добавлены сведения о наградах и номинациях (например, пользователь будет искать «номинанты премии Оскар-2020»).

Заказ в кинотеатре можно делать как в форме разовой покупки по тарифу для определенных фильмом (например, фильмы до 1990 года – по 100 руб, фильмы после 1990 г – по 200 руб) либо по подписке на месяц. Периодически бывают скидки либо на какие-то определенные категории фильмом, либо для определенного пользователя (бонус на день рождения). Тарифы, размеры скидки хранятся в таблице измерений по стоимости заказа.

Для удобства пользователя и удобства работы (или создания) рекомендательной системы каждому фильму, вновь внесенному в хранилище данных, присваивается рейтинг «0», после чего пользователи меняют его, просматривая фильм и оставляя отзывы. Отзывы могут быть как в числовом формате, так и в словесном (в таком случае рекомендация будет выглядеть как «75% положительных отзывов»). Данные сведения хранятся в таблице измерений по оценке фильма.

1. **Проверка на качество данных**
2. В справочнике «Клиент» каждому уникальному e-mail соответствует одна дата регистрации в системе
3. В справочнике «Клиент» указаны только реально существующие почтовые индексы, можно оставить поле пустым, но нельзя ввести неверные данные
4. В справочнике «Параметры фильма» не может быть двух фильмом с одинаковым названием одного режиссера и года выпуска.
5. В справочнике «Параметры фильма» один фильм не может принадлежать ко всем жанрам, в то время как должен принадлежать хотя бы к одному
6. В справочнике «Оценка фильма» оценка не может быть равна 0 и превышать 10
7. В справочнике «Оценка фильма» словесный отзыв не может состоять из одного слова

(отзыв «Супер!» - это не отзыв, это надо 10 баллов в баллах ставить, а не текстовое поле захламлять 😊)

1. В справочнике «Время» дата покупки не может быть раньше даты регистрации клиента и позже текущей даты и времени
2. В таблице «Оценка фильма» дата первого отзыва не должны быть раньше даты выхода фильма и даты включения фильма в хранилище данных онлайн-театра
3. В таблице «Стоимость заказа» стоимость заказа не может быть отрицательной
4. В таблице «Стоимость заказа» стоимость подписки на месяц должна отсчитываться от даты оплаты подписки, а не от начала календарного месяца
5. **Описание проекта по Crisp-DM**

Для улучшения работы онлайн-кинотеатра полезной будет построение рекомендательной системы, предсказывающей, какие фильмы могут понравиться конкретному пользователю, и показывающей их ему в рекомендациях. Для осуществления внедрения модели необходимо реализовать следующие шаги по Crisp-DM:

1. Business understanding

На этом шаге планируется выявить основные требования, недостатки текущего ведения дел и разобраться, чем именно поможет внедрение модели. В том числе получить ответы на следующие вопросы:

● Определить критерий, помогающий понять, какие фильмы нравятся пользователю

● Определить, как заказчик видит использование полученной модели, сформулировать минимально необходимое качество

● Оценить ожидаемый эффект от такой модели и сравнить его с ожидаемыми трудозатратами

1. Data Understanding

На этом шаге следует проверить, все ли необходимые данные доступны, их необходимо собрать, изучить и проверить их качество. Для выполнения этого этапа требуется:

- определить, какие данные по пользователям, по фильмам и по оценкам собираются, нет ли пропущенных важных данных, в данный момент не сохраняющиеся

- оценить способ хранения и тип данных. Если данные собираются в разные файлы или разные базы данных, собрать их вместе. Определить какие данные есть – категориальные, номинальные, числовые, …

- оценить взаимосвязи в данных, например – есть ли связь между полом пользователя и предпочитаемым жанром

- оценить полноту данных, при недостатке информации – сообщить заказчику

- оценить качество данных, например, нет ли дублирования пользователей, фильмов или оценок

1. Data Preparation

На этом этапе проводится подготовка данных к построению модели, а именно:

- очистка данных (выбрать значимые и незначимые (скорее всего) признаки), заполнить пропуски. Значимость признака определяется задачей. В случае построения рекомендательной системы пол и возраст пользователя, скорее всего, будут влиять на предпочтения в фильмах, а место проживания – скорее всего, можно на первом этапе учета признаков не учитывать. Для фильмов – будут важны режиссер, актеры, награды, а вот длительность фильма может не оказывать существенного влияния на предсказание

- могут быть введены дополнительные признаки, полученные математически из первичных признаков. Например, общая оценка фильма может складываться из цифровых оценок и словесных отзывов, просуммированных или взятых с каким-то коэффициентами

- определить принцип деления на тестовую и обучающую выборку (рандомно, сравнить старые фильмы с новыми, сравнить пользователей старше и младше определенного возраста, …)

В конце этапа подготовки должен получиться дата-фрейм, пригодный для анализа

1. Modeling

На данном этапе строится и обучается модель. Для этого:

- выбирается алгоритм(-мы) для построения рекомендации

- выбирается метрика(-ки) для оценки качества модели

- проводится оценка качества построенной модели согласно выбранной метрике

1. Evaluation

На этом этапе проводится оценка прогнозирования, понятность и полезность предсказываемых результатов, поиск неадекватной реакции модели на внешние обстоятельства. Требуется оценить:

- в каком виде лучше представить результат бизнесу (увеличение вероятности покупки - если советовать пользователю то, что ему понравится, скорее всего, он будет покупать больше; увеличение лояльности пользователей к данному онлайн-кинотеатру, что может выражаться в оттоке пользователей из другого кинотеатра в наш)

- найдены ли какие-то интересные зависимости. Например, может быть обнаружено, что пользователи, которые любят советские комедии, будут неравнодушны к американским ситкомам

- возможно, не хватает каких-то данных. Например, если к описанию фильма добавить имя оператора, это повысить точность предсказаний

- возможно, будет выявлено, что модель дает разную точность в разные промежутки времени. Например, перед вручение крупной кинопремии ряд пользователей может регистрироваться только для того, чтобы посмотреть нашумевшие фестивальные фильмы. Знание таких зависимостей позволить обогатить и улучшить модель

1. Deployment

На данном этапе происходит внедрение модели в работу заказчика. Для этого требуется:

- представить заказчику результат модели в понятном виде

- определить с заказчиком, какая финансовая польза получится от внедрения конкретной разработанной модели

- выбрать метод поддержки модели, как часто нужно обращаться за обновлением (раз в какой-то промежуток времени, при изменении интерфейса, при добавлении нового жанра фильмов, …)

- определить, кто и как отслеживает работу модели (аутсорс, сотрудник на стороне заказчика) и собирает с нее новые бизнес-показатели

- после согласования, «встроить» работу модели в действующий продукт (сайт онлайн-кинотеатра)

1. **Требуемые роли в команде**

За ходом проверок следят преимущественно ETL-специалист и дата-инженер.

За проверками по оценкам фильма дополнительно может следить аналитик, например, путем отслеживания количества и распределения оценок еженедельно. Если в какой-то день оценкок было значительно меньше, чем обычно, или все пользователи ставили «1» всем фильмам – возможно, был какой-то сбой.

По стадиям Crisp-DM:

- Business understanding. Для успешного выполнения данного этапа потребуется Аналитик, Владелец продукта, Бизнес-заказчик

- Data Understanding. Для выполнения требуется Аналитик, ETL-специалист, возможно – специалист по базам данным

- Data Preparation. Для выполнения нужны дата-сайентист и дата-инженер. Владелец продукта может дать совет по значимости показателя для конкретного бизнеса

- Modeling. Для выполнения этапа нужен дата-сайентист

- Evaluation. Для выполнения этапа нужны аналитик и владелец продукта

- Deployment. Для выполнения этапа требуются владелец продукта, аналитик BI, дата-сайентист, разработчик. Конечная презентация проводится перед бизнесом-заказчиком

За продвижением по проекту по стадиям и соблюдением заложенного графика по времени и по бюджету следит менеджер.