# 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:

|  |  |  |
| --- | --- | --- |
| Номер | Задание | Баллы |
| 1 | Описать основные бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу | 10 |
| 2 | Описать основные имеющиеся данные и источники их поступления | 20 |
| 3 | Описать основные сущности в хранилище данных (схема звезда) и процесс заливки данных | 20 |
| 4 | Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке | 10 |
| 5 | Придумать Data-проект, который должен улучшить показатели Вашего бизнеса и расписать его по Crisp-DM | 20 |
| 6 | Описать требуемые роли в команде по работе с данными на этапах 4 и 5 | 30 |
|  | Итого | 110 |

Итого - максимум 110. Для зачета необходимо набрать 90

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

В отчете можно проанализировать:

- количество пользователей в периоде,

- количество уникальных пользователей,

- количество новых пользователей,

- количество оцененных фильмов на 1 пользователя

- активность по возрастам, профессиям, полу.

* 1. **Отчет по фильмам**

Отчет позволит оценить:

- количество фильмов на сайте,

- рейтинг фильмов, жанров

- количество просмотров,

- количество новых фильмов за период,

1. Описать основные имеющиеся данные и источники их поступления

База данных состоит из 2-х справочников (пользователи, фильмы) и таблицы событий, в которой пользователь присваивает рейтинг просмотренному фильму.

Справочник пользователей формируется из анкет, заполняемых при регистрации на сайте.

Справочник фильмов формируется сотрудниками сайта. Обновление справочника происходит в момент заливки нового фильма на сайт.

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

**u.user** – справочник посетителей сайта.

- id

- возраст

- пол

- профессия

**u.item** – справочник фильмов

- id

- название фильма

- фдата выхода

- ссылка на фильм

- жанр

**u.data** – таблица событий

- id посетителя

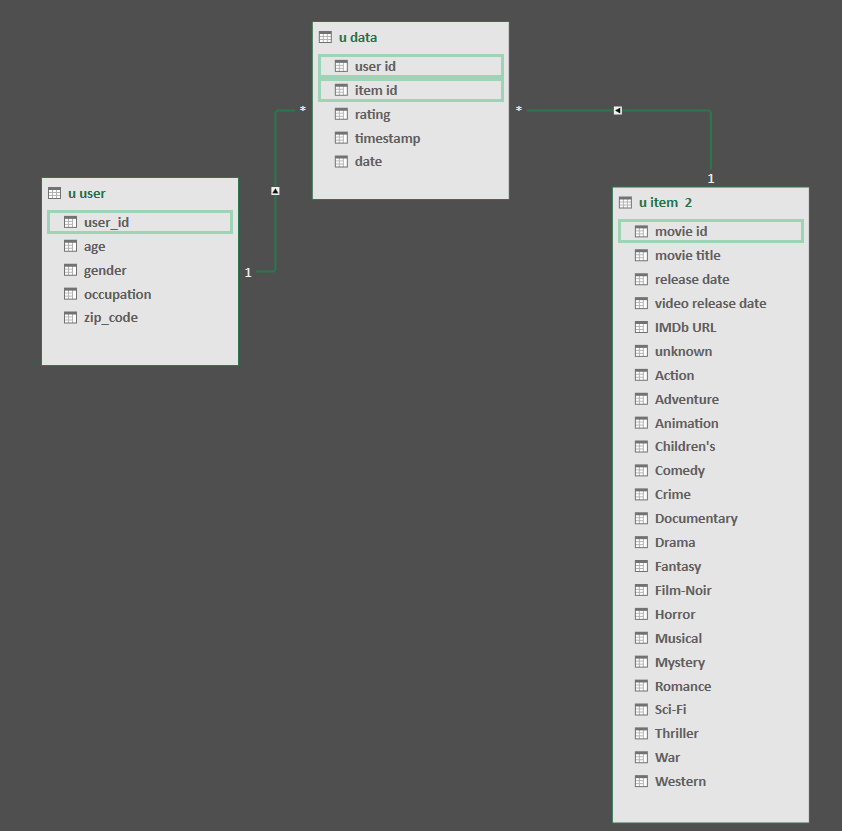
- id фильма

- оценка

- время оценки

Справочники содержат дополнительные признаки для расшифровки таблицы событий.

Связь через ключевое поле id\_user и id\_movie.



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

- в справочниках пользователей и фильмов id не должны повторяться и содержать только цифры (последовательность чисел).

- для каждого пользователя должен быть указан возраст – должно быть число, например от 18 до 100 лет

- для каждого пользователя нужно выбрать из списка пол – М или Ж.

- профессия должна быть выбрана из списка.

- обязательно должно быть заполнено поле «movie title»

- обязательно должно быть заполнено поле «release date» в виде даты ДД.ММ.ГГГГ

- обязательно должен быть указан жанр,

- отчеты (по посетителям сайта или фильмам) должны содержать не больше строк, чем сами справочники. Если строк больше, значит где-то есть дублирование строк и сводная информация некорректная,

- отчеты не должны содержать пустых строк – либо применен неверный join, либо справочники неактуальны,

- средний рейтинг не может быть больше 5.

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

**Разработать модель рекомендации фильмов.**

**Бизнес-анализ**

Определить цель, метрики, сформировать план действий, желаемые сроки и затраты.

Роли: Владелец, Аналитик.

**Анализ данных**

- Определение источников данных и описание содержащейся информации;

- Изучение данных на достаточность информации и ее качества.

Роли: Аналитик, специалист по БД.

**Подготовка данных**

- Объединение данных из разных источников в единый датафрейм;

- Очистка от выбросов и избыточной информации;

- Форматирование данных в согласованную структуру.

Роли: Аналитик

**Моделирование**

- Выбор модели;

- Проведение обучения модели;

- Оценка качества модели.

Роли: DS

**Оценка решения**

- Оценка результатов анализа и получившейся модели;

- Оценка процесса.

Роли: Владелец, Аналитик, специалист по БД, DS

**Внедрение**

- внедрение модели;

- поддержка модели;

- формирование отчетности.

Роли: Разработчик, Аналитик, DS

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

**Бизнес-анализ**

Роли: Владелец, Аналитик.

**Анализ данных**

Роли: Аналитик, специалист по БД.

**Подготовка данных**

Роли: Аналитик

**Моделирование**

Роли: DS

**Оценка решения**

Роли: Владелец, Аналитик, специалист по БД, DS

**Внедрение**

Роли: Разработчик, Аналитик, DS