# 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 |

1. - Отчет о дневном/месячном/годовом просмотрах в интернет-кинотеатре в динамике к прошлым показателям

- Отчет по выручке

- Динамику притока/оттока клиентов

1. - БД фильмов и их характеристик (свои данные)

- БД клиентов и их данных (свои данные)

- google, yandex аналитика (соответствующие сайты)

1. Основные сущности: - пользователь, фильм, дата.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| пользователь |  |  |  | дата/время |
| фио |  |  |  | дата |
| возраст |  | fact table |  | время |
| дата регистрации |  | просмотр |  |  |
| эл. почта |  | пользователь |  |  |
| логин |  | фильм |  |  |
|  |  | дата/время |  |  |
|  |  |  |  |  |
| фильм |  |  |  |  |
| название |  |  |  |  |
| год |  |  |  |  |
| жанр |  |  |  |  |
| актеры |  |  |  |  |
|  |  |  |  |  |

1. - валидный адрес эл. почты

- возраст - целое число > 0

- жанр - входит в справочник жанров

- актер - входит в справочник актеров

- название фильма - строка не более 200 символов

- дата имеет формат “yyyy-mm-dd”

- время в формате “hh:mm:ss”

- фио - строка, не пустая.

- логин - строка, не пустая, состоящая из латинских букв

- логин - не повторяет другие логины в БД

1. 6. Повысить продолжительность “жизни” клиента на сервисе. Исходим из того, что в кинотеатре подписка.

**Business Understanding**

**План-результат:**

1. Создать рекомендательную систему.
2. Растить метрику соотношения просмотренных фильмов из рекомендательной выдачи к найденным пользователем самостоятельно.
3. Определить как заказчик видит использование полученной модели, сформулировать минимально необходимое качество
4. Оценить ожидаемый эффект от такой модели и сравнить его с ожидаемыми трудозатратами

**Специалисты:** Аналитик, Владелец продукта/менеджер продукта

**Data understanding**

**План-результат:**

1. Оценить и собрать имеющиеся данные

* Информация о пользователе (своя)
* Информация о фильме (своя)
* Информация интернет статистики (сайты статистики)
* Информация о промо-акциях компании (своя)
* Информация об акциях иных кинотеатров

1. Исследовать данные (разведочный анализ)
2. Описать данные (количество, значения, связи)
3. Изучите качество данных (достоверность, отсутствующие данные, ошибки в данных, плохие метаданные)

**Специалисты:** Аналитик, Аналитик данных

**Data preparation**

**План-результат:**

1. При необходимости, парсинг данных с внешних сайтов, обогащение БД
2. Выбрать данные (разделите на train/test, выбрать признаки)
3. Очистить данные (заполните пропуски, исправьте ошибки)
4. Расширить данные (новые признаки и т.п.)
5. Сохранить данные (подготовить data frame, pipeline для обучения модели и сохранить его)

**Специалисты:** DS, DE

**Modeling**

**План-результат:**

1. Выбрать метод моделирования
2. Построить модель
3. Оценить модель
4. Описать результат

**Специалисты:** DS

**Evaluation**

**План-результат:**

1. Оценить результаты
2. Просмотреть процесс
3. Определить следующие шаги

**Специалисты:** DS, Аналитик

**Deployment**

**План-результат:**

1. Внедрение
2. Мониторинг и техническое обслуживание
3. Проведение итогового обзора проекта

**Специалисты:** DS, Разработчик, Аналитик