Промежуточная аттестация 2

Формулировка задания

- Выполнить загрузку данных из БД;
- Написать запросы к базе данных и визуализировать результаты запросов;
- Провести статистический анализ данных.

Описание плана работы

Задача 1 Загрузка данных

Данные для задачи загрузить из дополнительных материалов или по ссылке: https://www.kaggle.com/datasets/atanaskanev/sqlite-sakila-sample-database?select=SQLite3+Sakila+Sample+Database+ERD.png

Выбрать СУБД PostgreSQL или MySQL. Загрузить файлы базы данных как таблицы в выбранную СУБД. При загрузке таблиц обратить внимание на отношения между таблицами: первичные ключи (Primary Key) и внешние ключи (Foreign Key).

Отношения между таблицами должны соответствовать Рис. 1.

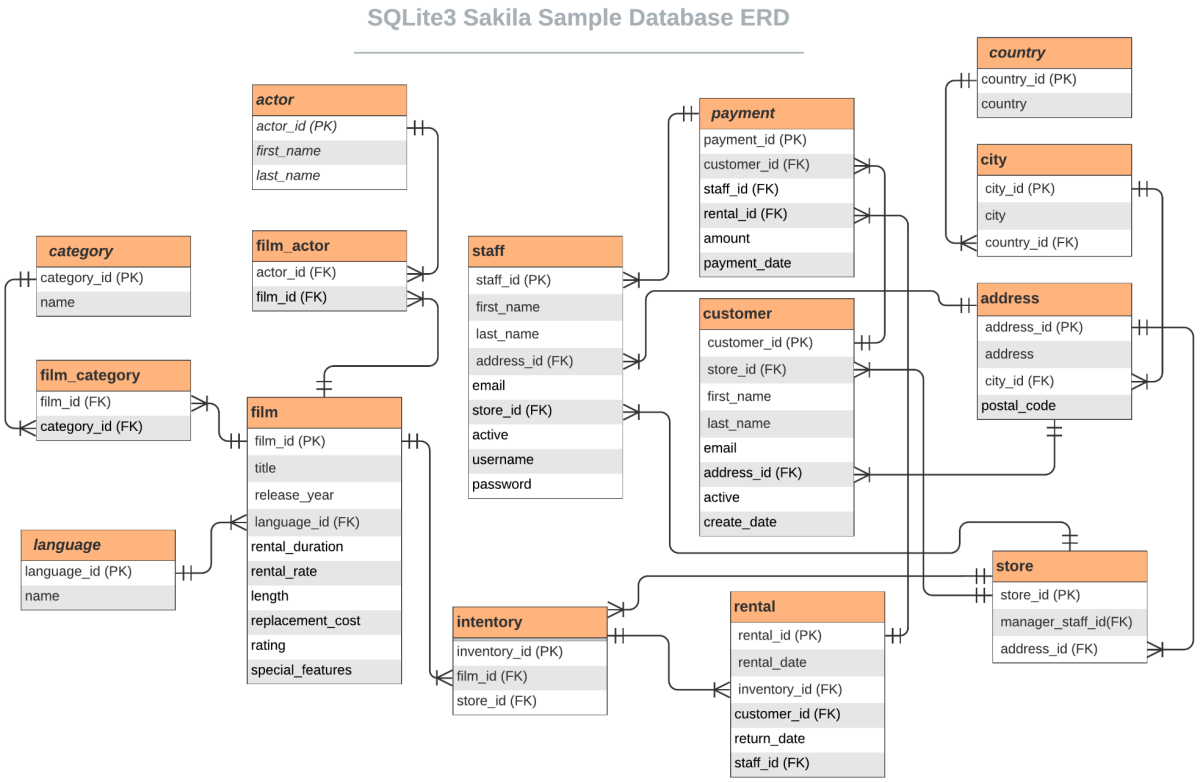

Рис. 1 Схема базы данных DVD-проката

После загрузки данных в таблицу проверить корректность типов данных и значений. Выгрузить dump базы данных:

Для PostgreSQL: https://www.dmosk.ru/miniinstruktions.php?mini=postgresql-dump

Для MySQL: https://www.dmosk.ru/miniinstruktions.php?mini=mysql-dump

Задача 2 Подключение к БД и выполнение запросов

Внимательно изучить схему данных и загруженную базу из Задания 1. Сформировать 1 или несколько классов по таблицам в базе данных.
Подключиться к базе данных, загруженной в задаче 1 из python. При подключении рекомендуется использовать менеджер контекста.

Выполнить запросы к данным для ответа на следующие вопросы и выгрузить в промежуточные таблицы:

1. Какова доля фильмов в каждой рейтинговой категории (G, PG, PG-13, R и т.д.) в нашем ассортименте?
2. Какие категории фильмов чаще всего арендуются клиентами?
3. Какова средняя продолжительность проката (rental duration) для каждой категории фильмов?
4. Каковы тенденции в ежемесячном доходе от проката (monthly rental revenue) и продажах (sales) за прошедший год?
5. Как соотносятся показатели продаж в разных магазинах?
6. Каковы средние затраты на замену (replacement_cost) фильмов в разных жанрах?
7. Какие актеры снимаются в самых разных жанрах фильмов?

Результат выполнения запросов сохранить в файлах .csv.

Построить визуализации результатов запросов с помощью библиотек matplotlib, seaborn или plotly.

Задача 3. Статистический анализ и визуализация данных

Подключиться к базе данных из Задания 1. Сформировать из 5-7 таблиц общую таблицу данных. Таблицу предварительно сохранить в файле .csv и загрузить в программу как DataFrame объект.

Провести разведочный анализ данных:

(а) для каждой числовой переменной вычислить:

• Долю пропусков

• Максимальное и минимальное значение

• Среднее значение

• Медиану

• Дисперсию

• Квантиль 0.1 и 0.9

• Квартиль 1 и 3

(b) для каждой категориальной переменной вычислить:

• Долю пропусков

• Количество уникальных значений

• Моду

Результат анализа записать в файл .csv

Входные данные для задач 1-3:

- Таблицы, выгруженные из Sakila Sample Database
- Структура данных: файл базы данных SQLite в формате *.db

Выходные данные для задач 1-3:

- Отчет со снимками экрана хода работы
- dump базы данных
- Код на python выполнения работы
- Промежуточные таблицы в формате csv для запросов 1-7
- Входные данные для статистического анализа в формате csv
- Результат статистического анализа в формате csv

Перечень необходимых инструментов

- Python
- venv
- psycopg2
- pymysql
- sqlalchemy
- PostgreSQL 14+
- MySQL Community 8+
- pandas
- matplotlib
- seaborn
- plotly
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданиям и решение задач. Решения должны быть представлены в формате .ipynb или .py.
3. Скрипты запросов к базе данных прописать в коде или вынести в отдельный файл.

Шкала оценивания

- **1.0 – отлично**

Решены все задачи. Учтены требования, код структурирован. Продемонстрирован навык работы с реляционными базами данных, библиотеками для анализа табличных данных и визуализации. Запросы 1-7 отрабатывают корректно. Визуализации данных информативны.

- **0.7–0.9 – хорошо**

Решены 2 или 3 задачи на высоком уровне. Код соответствует большинству требований, но может содержать небольшие ошибки или недочеты. Продемонстрирован навык работы с реляционными базами данных, библиотеками для анализа табличных данных и визуализации. Запросы 1-7 отрабатывают корректно с небольшими недочетами. Визуализации данных информативны, но есть небольшие замечания к реализации. Есть к чему стремиться в изучении темы.

- **0.5–0.6 – удовлетворительно**

Решены 1 или 2 задачи с множеством замечаний. Код соответствует большинству требований, но может содержать значительные ошибки или недочеты. Требуется дополнительное изучение тем модуля и исправление программы.

- **Менее 0.5 – задание не выполнено**

Задание выполнено на очень низком уровне или решено 0 задач. Требуется дополнительное изучение темы и решение заданий для практического изучения темы.