Проект по SQL

- Цель исследования:

- Проанализировать базу данных.
- Получить информацию о книгах, издательствах, авторах, пользовательские обзоры книг.
- Помочь сформулировать ценностное предложение для нового продукта.

In [1]:
```
# импортируем библиотеки
import pandas as pd
from sqlalchemy import create_engine
# устанавливаем параметры
db_config = {'user': 'praktikum_student', # имя пользователя
'pwd': 'Sdf4$2;d-d30pp', # пароль
'host': 'rc1b-wcoijxj3yxfsf3fs.mdb.yandexcloud.net',
'port': 6432, # порт подключения
'db': 'data-analyst-final-project-db'} # название базы данных
connection_string = 'postgresql://{}:{}@{}:{}/{}'.format(db_config['user'],
db_config['pwd'],
db_config['host'],
db_config['port'],
db_config['db'])
# сохраняем коннектор
engine = create_engine(connection_string, connect_args={'sslmode':'require'})
```

Исследуем таблицы

Выведем первые строки

Таблица "books"

In [2]:
```
query="""
SELECT *
FROM books
LIMIT 5
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[2]:

	book_id	author_id	title	num_pages	publication_date	publisher_id
0	1	546	'Salem's Lot	594	2005-11-01	93
1	2	465	1 000 Places to See Before You Die	992	2003-05-22	336
2	3	407	13 Little Blue Envelopes (Little Blue Envelope...	322	2010-12-21	135
3	4	82	1491: New Revelations of the Americas Before C...	541	2006-10-10	309
4	5	125	1776	386	2006-07-04	268

- Содержит данные о книгах:
  - book_id — идентификатор книги;
  - author_id — идентификатор автора;
  - title — название книги;
  - num_pages — количество страниц;
  - publication_date — дата публикации книги;
  - publisher_id — идентификатор издателя.

Таблица "authors"

In [3]:
```
query="""
SELECT *
FROM authors
LIMIT 5
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[3]:

	author_id	author
0	1	A.S. Byatt
1	2	Aesop/Laura Harris/Laura Gibbs
2	3	Agatha Christie
3	4	Alan Brennert
4	5	Alan Moore/David Lloyd

- Содержит данные об авторах:
  - author_id — идентификатор автора;
  - author — имя автора.

Таблица "publishers"

In [4]:
```
query="""
SELECT *
FROM publishers
LIMIT 5
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[4]:

	publisher_id	publisher
0	1	Ace
1	2	Ace Book
2	3	Ace Books
3	4	Ace Hardcover
4	5	Addison Wesley Publishing Company

- Содержит данные об издательствах:
  - publisher_id — идентификатор издательства;
  - publisher — название издательства;

Таблица "ratings"

In [5]:
```
query="""
SELECT *
FROM ratings
LIMIT 5
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[5]:

	rating_id	book_id	username	rating
0	1	1	ryanfranco	4

	rating_id	book_id	username	rating
1	2	1	grantpatricia	2
2	3	1	brandtandrea	5
3	4	2	lorichen	3
4	5	2	mariokeller	2

- Содержит данные о пользовательских оценках книг:
  - rating_id — идентификатор оценки;
  - book_id — идентификатор книги;
  - username — имя пользователя, оставившего оценку;
  - rating — оценка книги.

Таблица "reviews"

In [6]:
```
query="""
SELECT *
FROM reviews
LIMIT 5
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[6]:

	review_id	book_id	username	text
0	1	1	brandtandrea	Mention society tell send professor analysis. ...
1	2	1	ryanfranco	Foot glass pretty audience hit themselves. Amo...
2	3	2	lorichen	Listen treat keep worry. Miss husband tax but ...
3	4	3	johnsonamanda	Finally month interesting blue could nature cu...
4	5	3	scotttamara	Nation purpose heavy give wait song will. List...

- Содержит данные о пользовательских обзорах:
  - review_id — идентификатор обзора;
  - book_id — идентификатор книги;
  - username — имя автора обзора;
  - text — текст обзора.

Задания:

задание:

Посчитаем, сколько книг вышло после 1 января 2000 года;

In [7]:
```
query="""
SELECT  count(book_id) as count_book
FROM books as b
Where publication_date >= '2000-01-01'
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[7]:

	count_book
0	821

Вывод:

- после 1 января 2000 года вышло 821 книга

задание

Для каждой книги посчитаем количество обзоров и среднюю оценку

In [8]:
```
query="""
SELECT b.book_id, title,COUNT(distinct(review_id)) as number_of_reviews ,AVG(rating) as average
    FROM books as b
    LEFT JOIN ratings as r ON b.book_id = r.book_id
    LEFT JOIN reviews as rev ON b.book_id = rev.book_id
    GROUP BY b.book_id
    ORDER BY number_of_reviews DESC
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[8]:

	book_id	title	number_of_reviews	average_grade
0	948	Twilight (Twilight #1)	7	3.662500
1	963	Water for Elephants	6	3.977273
2	734	The Glass Castle	6	4.206897
3	302	Harry Potter and the Prisoner of Azkaban (Harr...	6	4.414634
4	695	The Curious Incident of the Dog in the Night-Time	6	4.081081
...	...	...	...	...
995	83	Anne Rice's The Vampire Lestat: A Graphic Novel	0	3.666667
996	808	The Natural Way to Draw	0	3.000000
997	672	The Cat in the Hat and Other Dr. Seuss Favorites	0	5.000000
998	221	Essential Tales and Poems	0	4.000000
999	191	Disney's Beauty and the Beast (A Little Golden...	0	4.000000

1000 rows × 4 columns

- Вывод:
  - наибольшее количество обзоров 7
  - есть книги вообще без обзоров
  - наибольшая оценка 5

задание

Определим издательство, которое выпустило наибольшее число книг толще 50 страниц — так мы исключим из анализа брошюры

In [9]:
```
query="""
SELECT COUNT(p.publisher_id),p.publisher
    FROM books as b
    INNER JOIN publishers as p ON b.publisher_id = p.publisher_id
    WHERE b.num_pages>50
    GROUP BY p.publisher_id
    ORDER BY COUNT(p.publisher_id) DESC
    LIMIT 1
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[9]:

	count	publisher
0	42	Penguin Books

- Вывод:
  - Наибольшее число книг выпустило издательство "J.K. Rowling/Mary GrandPré" - 42 книги

задание

Определим автора с самой высокой средней оценкой книг — учитываем только книги с 50 и более оценками

In [10]:
```
query="""
SELECT a.author,AVG(r.rating) as average_rating_books
    FROM books as b
    JOIN authors as a ON b.author_id=a.author_id
    JOIN ratings as r ON b.book_id=r.book_id
    WHERE b.book_id IN (
            SELECT book_id
            FROM ratings as r
            GROUP BY book_id
            HAVING COUNT(rating)>=50
            ORDER BY AVG(rating) DESC)
GROUP BY a.author
ORDER BY average_rating_books DESC
LIMIT 1
"""
pd.io.sql.read_sql(query, con = engine)
```

Out[10]:

	author	average_rating_books
0	J.K. Rowling/Mary GrandPré	4.287097

- Вывод:
  - "J.K. Rowling/Mary GrandPré" автор с самой высокой средней оценкой книги(среди книг с 50 и более оценками) - 4.287097

задание

Посчитаем среднее количество обзоров от пользователей, которые поставили больше 50 оценок.

In [11]:
```
query="""
SELECT AVG(number_reviews)as average_number_reviews
    FROM (
        SELECT count(review_id) AS number_reviews
        FROM reviews
        WHERE username IN (
            SELECT username
            FROM ratings
            GROUP BY username
            HAVING count(rating_id)>=50)
GROUP BY username) AS RAT

"""
pd.io.sql.read_sql(query, con = engine)
```

Out[11]:

	average_number_reviews
0	24.222222

- Вывод:
  - среднее количество обзоров от пользователей, которые поставили больше 50 оценок - 24.222222