Bases de Dados 2

GoodBooks Dataset

Relatório Final

Grupo 4

Índice

Introdução global

No âmbito da disciplina de Bases de Dados 2, os estudantes foram desafiados a trabalhar em grupos para aplicar os conhecimentos adquiridos em sala de aula em projetos práticos. Até ao momento desta disciplina, a nossa experiência estava limitada à modelação de dados e a uma abordagem teórica da tecnologia MySQL. Agora, em Bases de Dados 2, tivemos a oportunidade de explorar não apenas essa tecnologia previamente estudada, mas também outras duas que nos eram desconhecidas. Pela primeira vez, pudemos trabalhar de forma prática com conjuntos de dados reais e de forma autónoma, ampliando assim as nossas competências e conhecimentos no campo das bases de dados.

O objetivo principal deste projeto foi aprofundar a nossa compreensão das diversas abordagens de armazenamento e manipulação de dados, assim como as vantagens e desafios de cada tecnologia selecionada. Para atingir esse objetivo, realizamos a migração do conjunto de dados Goodbooks para três tecnologias distintas: MySQL, Apache Cassandra e NEO4J. Cada uma dessas etapas representou uma oportunidade única para explorar a tecnologia em questão e aplicar os conceitos aprendidos.

No primeiro estágio, migramos o conjunto de dados Goodbooks para o MySQL, onde exploramos a modelação de dados relacional e aplicamos consultas SQL selecionadas. Na segunda fase, realizamos a migração para o Cassandra, uma tecnologia de bases de dados distribuída, onde exploramos os recursos disponíveis para lidar com grandes volumes de dados. Por fim, na terceira fase, realizamos a transição para o NEO4J, utilizando a modelação de grafos para analisar as relações entre os dados.

Cada fase, juntamente com a sua introdução e conclusão respetivas, contém todos os elementos necessários para uma interpretação correta do conjunto de dados, da modelação dos dados e do código envolvido. Desta forma, o leitor deste relatório terá uma compreensão precisa de todos os elementos e poderá até utilizá-lo como referência para criar os seus próprios bancos de dados.

Para uma melhor interpretação, aconselhamos a uma leitura seguida deste documento.

MySQL

Introdução

No âmbito da unidade curricular bases de dados 2, os alunos foram desafiados a trabalhar em grupos para aplicar os conhecimentos adquiridos em sala de aula em projetos práticos. Nesta primeira fase do relatório, apresentamos a análise de um conjunto de dados usando o sistema de gerenciamento de banco de dados MySQL, realizada pelo nosso grupo.

O conjunto de dados selecionado para este trabalho é o Goodbooks, que contém informações sobre livros, autores, classificações e avaliações de utilizadores. Nesta primeira parte, o nosso objetivo foi demonstrar como os recursos do MySQL podem ser usados para explorar, analisar e visualizar dados de um conjunto de dados real.

Para atingir esse objetivo, descrevemos a estrutura do conjunto de dados e os modelos criados para armazená-lo no banco de dados MySQL. Além disso, utilizamos a ferramenta DBeaver, recomendada pelos docentes da UC, para administrar a base de dados e executar consultas SQL para extrair informações valiosas do conjunto de dados.

Informação sobre o Dataset

Descrição do dataset:

O conjunto de dados selecionado para análise contém informações detalhadas sobre diversos livros disponíveis na plataforma GoodReads, conhecida como GoodBooks. Essa plataforma oferece uma vasta gama de livros, com informações como sinopse, autor, editora, data de publicação e outros detalhes relevantes para os leitores.

Uma característica importante da plataforma GoodReads é a possibilidade de os utilizadores atribuírem uma nota de 1 a 5 estrelas aos livros que leram, permitindo que outros leitores saibam o que esperar da obra antes de iniciá-la. Além disso, a plataforma permite que os utilizadores adicionem livros à sua lista "ver mais tarde", tornando mais fácil para eles lembrarem-se dos títulos que desejam ler no futuro.

Com base nesses dados, é possível realizar uma análise aprofundada sobre os livros mais populares, os autores mais bem-sucedidos e outros insights interessantes sobre o mundo da literatura.

O dataset é composto por 5 ficheiros csv (5 entidades):

- Books(lojas): Contém diversas informações sobre cada livro no conjunto de dados.
  Atributos:
  - Book_id: Id que caracteriza cada livro no conjunto de dados GoodBooks.
  - Goodreads_book_id: Id do livro atribuído pela GoodReads.
  - Best_book_id: Id do “melhor livro” de um determinado trabalho literário.
  - Work_id: Id de cada trabalho literário.
  - Books_count: Contagem de livros diferentes associados a um determinado “Work_Id”.
  - Isbn: International Standart Book Number – Um número de 10 dígitos que identifica cada livro (cada obra), número esse dado pela agência internacional.
  - Isbn13: International Standart Book Number – Semelhante ao Isbn, porém, com 13 dígitos. Os três dígitos a mais identificam a língua ou o país, algo que não é distinguível no ibsn.
  - Authors: Autor/es de cada livro. Quando há mais que um, os autores estão separados por vírgulas.
  - Original_publication_year: Ano de publicação original do livro (primeira vez em que foi laçado).
  - Original_tittle: Título original do livro.
  - Tittle: Nome do livro.
  - Language_code: Código do idioma em que o livro foi originalmente escrito (Ex: Pt).
  - Average Rating: Representa a classificação média do livro atribuída pelos utilizadores do Goodreads. (Os utilizadores atribuem uma avaliação de 1 a 5).
  - Ratings_count: Quantidade total de avaliações que foram dadas pelos utilizadores para um determinado livro.

- Work_ratings_count: Quantidade total de avaliações que foram dadas pelos utilizadores para uma determinada obra, independentemente da brochura, capa ou idioma.
- Work_text_reviews_count: Número de comentários para uma determinada obra na plataforma Goodreads.
- Ratings_1: Número de avaliações que um livro recebeu com classificação de 1 estrela.
- Ratings_2: Número de avaliações que um livro recebeu com classificação de 2 estrelas.
- Ratings_3: Número de avaliações que um livro recebeu com classificação de 3 estrelas.
- Ratings_4: Número de avaliações que um livro recebeu com classificação de 4 estrelas.
- Ratings_5: Número de avaliações que um livro recebeu com classificação de 5 estrelas.
- Image_url: URL da imagem da capa do livro.
- Small_image_url: URL da imagem da capa do livro em formato pequeno.

➢ Tags: Contém informações sobre as tags (etiquetas) existentes na plataforma Goodreads.

Atributos:

- Tag_id: Identificador único para cada tag.
- Tag_name: Nome da tag, ou seja, a palavra ou frase que os utilizadores utilizaram para rotular o livro.

➢ Book_tags: Contém informações sobre as tags atribuídas pelos utilizadores do GoodReads a cada livro.

Atributos:

- Book_id: Identificador único para cada livro.
- Tag_id: O ID da tag atribuída ao livro.
- Count: Número de utilizadores que atribuíram a tag ao livro.

➢ Ratings: Contém informações sobre as avaliações dos utilizadores para livros específicos. Cada entrada na tabela apresenta uma avaliação de um utilizador para um livro. Uma avaliação pode variar entre 1 (mais baixa) e 5 estrelas (avaliação mais alta).

Atributos:

- User_id: Id que caracteriza cada utilizador.
- Book_id: Id que caracteriza cada livro no conjunto de dados GoodBooks.
- Rating: Avaliação dada ao livro (entre 1 a 5).

- To_read: Contém informações sobre a lista de livros que cada utilizador do Goodreads pretende ler. Cada linha da tabela representa um livro adicionado à lista “to read” por um utilizador específico.

  Atributos:
  - User_id: Id do utilizador que adicionou o livro à lista “to read”.
  - Book_id: Id do livro que foi adicionado à lista “to read”.

Importação dos dados e criação das tabelas

Para esta primeira fase, foi necessário realizar a criação e carga de várias tabelas na base de dados para armazenar o conjunto de dados Goodbooks. Através do código apresentado ao lado, foram criadas tabelas como "books", "ratings", "tags" e outras, que fornecem uma estrutura para organizar os dados relacionados a livros, avaliações, etiquetas e outros elementos relevantes.

Estas tabelas são essenciais para a correta interpretação e utilização dos dados contidos no conjunto Goodbooks. Através delas, é possível estabelecer relações entre os livros, as avaliações dos utilizadores, as etiquetas associadas e outras informações importantes.

Ao executar o código fornecido, os dados contidos nos ficheiros CSV são carregados nas tabelas correspondentes, permitindo que a base de dados seja populada com as informações relevantes para a análise e manipulação posterior.

Este processo de criação e carga das tabelas é fundamental para estabelecer uma base sólida para as fases subsequentes do projeto, onde serão exploradas diferentes tecnologias de base de dados para migrar e analisar esses dados.

É importante ressaltar que o código apresentado é apenas uma parte do processo de preparação do ambiente de trabalho e não abrange aspectos específicos do funcionamento interno das consultas ou dos recursos técnicos envolvidos.

Antes desta etapa, foi necessário preparar o ambiente de trabalho, incluindo a configuração do DBeaver e a criação dos containers para as bases de dados. Recomendamos uma instalação prévia dos requisitos necessários, uma vez que este relatório não abrange esses detalhes.

```sql
USE goodbooks;

DROP TABLE IF EXISTS book_tags;
CREATE TABLE book_tags (
    book_id INT,
    tag_id INT,
    count INT,
    FOREIGN KEY (book_id) REFERENCES books (book_id),
    FOREIGN KEY (tag_id) REFERENCES tags (tag_id),
    PRIMARY KEY (book_id, tag_id, count)
);

DROP TABLE IF EXISTS books;
CREATE TABLE books (
    book_id INT,
    goodreads_book_id INT,
    best_book_id INT,
    work_id INT,
    books_count INT,
    isbn VARCHAR(20),
    isbn13 VARCHAR(20),
    authors VARCHAR(200),
    original_publication_year INT,
    original_title VARCHAR(200),
    title VARCHAR(200),
    language_code VARCHAR(10),
    average_rating DECIMAL(3,2),
    ratings_count INT,
    work_ratings_count INT,
    work_text_reviews_count INT,
    ratings_1 INT,
    ratings_2 INT,
    ratings_3 INT,
    ratings_4 INT,
    ratings_5 INT,
    image_url VARCHAR(500),
    small_image_url VARCHAR(500),
    PRIMARY KEY (book_id)
);

DROP TABLE IF EXISTS ratings;
CREATE TABLE ratings (
    user_id INT NOT NULL,
    book_id INT NOT NULL,
    rating INT NOT NULL,
    FOREIGN KEY (book_id) REFERENCES books (book_id),
    PRIMARY KEY (user_id, book_id, rating)
);

DROP TABLE IF EXISTS tags;
CREATE TABLE tags (
    tag_id INT NOT NULL,
    tag_name VARCHAR(50) NOT NULL,
    PRIMARY KEY (tag_id)
);

DROP TABLE IF EXISTS to_read;
CREATE TABLE to_read (
    user_id INT NOT NULL,
    book_id INT NOT NULL,
    FOREIGN KEY (book_id) REFERENCES books (book_id),
    PRIMARY KEY (user_id, book_id)
);


LOAD DATA INFILE '/var/lib/mysql/csv/book_tags.csv'
INTO TABLE book_tags
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/books.csv'
INTO TABLE books
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/ratings.csv'
INTO TABLE ratings
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/tags.csv'
INTO TABLE tags
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/to_read.csv'
INTO TABLE to_read
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

Figura 1- Código referente a importação de tabelas SQL

Modelos de dados

Modelo conceptual de dados:

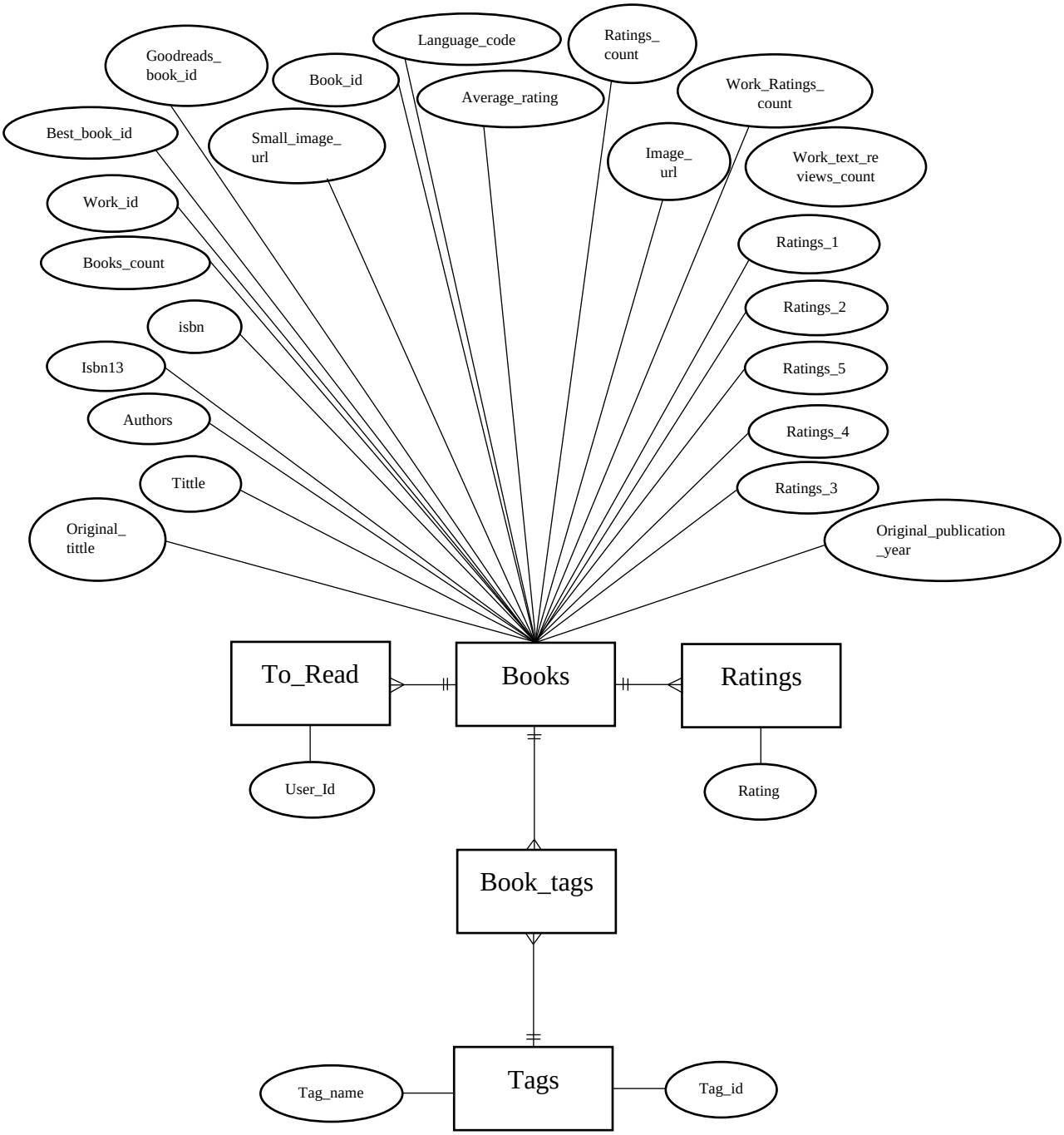


Diagrama 1- Modelo conceptual de dados para MySQL

Diagrama Entidade-Relacionamento (ER)

books (book_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title, language_code, average_rating, ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url)

tags (tag_id, tag_name)

book_tags (book_id, tag_id, count)

ratings (user_id, book_id, rating)

to_read (user_id, book_id)

Dicionário de Dados

Tabela	Books		
Descrição	Este conjunto de dados inclui um conjunto enorme de dados sobre cada livro.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
book_id	Código identificador do livro	int	PK
goodreads_book_id	código do livro atribuido pela Good Reads	int	
best_book_id	código do melhor livro de um determinado trabalho literário	int	
work_id	código de cada trabalho literário	int	
books_count	Número de livros diferentes associados a um determinado trabalho literário	int	
Isbn	Número de 10 dígitos que identifica cada livro	varchar	
isbn13	Número de 13 dígitos que identifica cada livro. Funciona de forma semelhante ao Isbn mas identifica também a língua ou o país.	varchar	
Authors	Autor/es de cada livro	varchar	
original_publication_year	Ano de publicação original do livro	int	
original_tittle	Título original do livro	varchar	
Tittle	Nome do livro	varchar	
language_code	Código do idioma em que o livro foi originalmente escrito	varchar	
average Rating	Classificação média do livro atribuída pelos utilizadores do Goodreads	decimal	
ratings_count	Número total de avaliações para um determinado livro	int	
work_ratings_count	Número total de avaliações para uma determinada obra	int	
work_text_reviews_count	Número de comentários para uma determinada obra	int	
ratings_1	Quantidade de avaliações com 1 estrela	int	
ratings_2	Quantidade de avaliações com 2 estrela	int	
ratings_3	Quantidade de avaliações com 3 estrela	int	
ratings_4	Quantidade de avaliações com 4 estrela	int	
ratings_5	Quantidade de avaliações com 5 estrela	int	
image_url	URL da imagem da capa do livro	varchar	
small_image_url	URL da imagem da capa do livro em formato pequeno	varchar	

Tabela	Tags		
Descrição	Este conjunto de dados inclui dados sobre as tags (etiquetas) existentes na plataforma Goodreads.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
tag_id	código identificador de cada tag	int	PK
tag_name	nome da tag respetiva	varchar	

Tabela	Book_tags		
Descrição	Este conjunto de dados inclui dados sobre as tags atribuídas pelos utilizadores do GoodReads a cada livro.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
book_id	Código identificador do livro	int	PK
tag_id	Código identificador da tag	int	PK
Count	Número de utilizadores que atribuíram a tag ao livro	int	

Tabela	Ratings		
Descrição	Este conjunto de dados inclui dados sobre as avaliações dos utilizadores para livros específicos.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
user_id	código do utilizador	int	PK
book_id	código do livro	int	PK, FK
rating	avaliação dada ao livro	int	PK

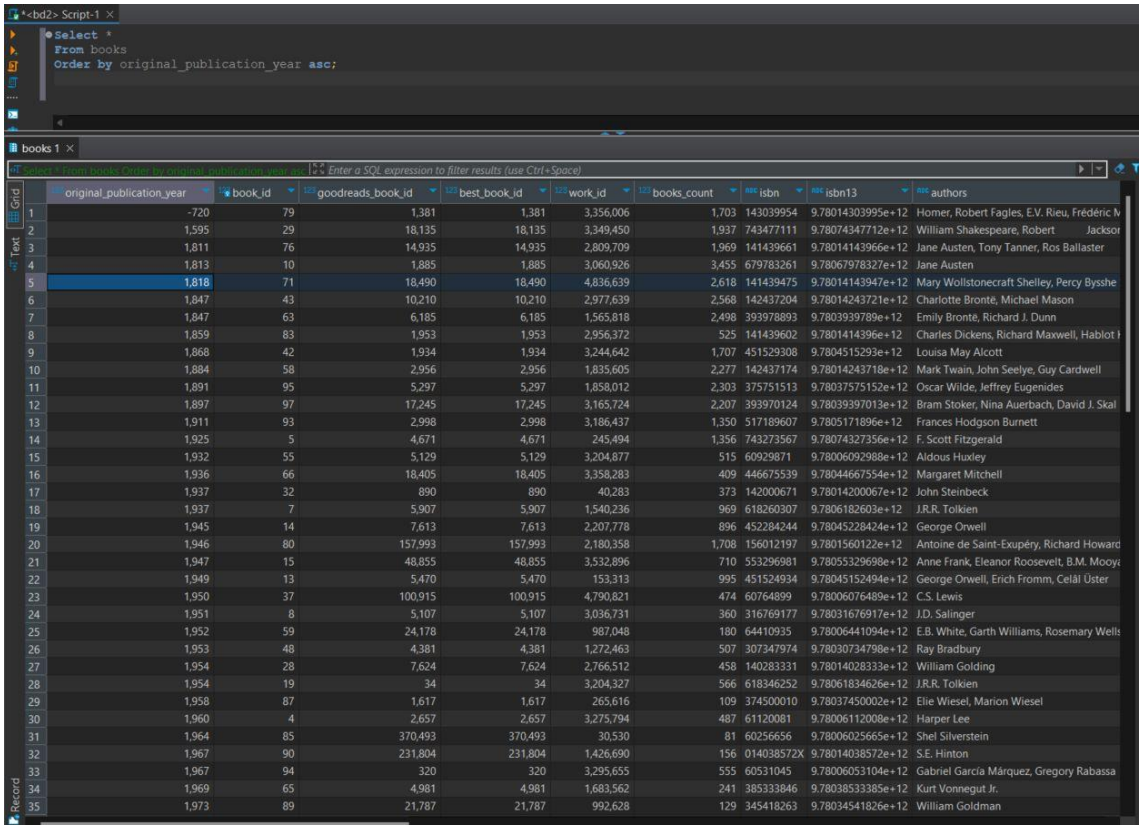
Tabela	To_read		
Descrição	Este conjunto de dados inclui dados sobre a lista de livros que cada utilizador do Goodreads adicionou à lista “ver mais tarde”.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
user_id	código do utilizador que adicionou o livro à lista “to read”	int	PK
book_id	código do livro que foi adicionado à lista “to read”	int	PK, FK

Questões analíticas

1-Listar os livros por ordem crescente pelo respetivo ano de publicação.

```
Select *
From books
Order by original_publication_year asc;
```

Resultado da query:



	original_publication_year	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors
1	-720	79	1,381	1,381	3,356,006	1,703	143039954	9.78014303995e+12	Homer, Robert Fagles, E.V. Rieu, Frédéric M
2	1,595	29	18,135	18,135	3,349,450	1,937	743477111	9.78074347712e+12	William Shakespeare, Robert Jackso
3	1,811	76	14,935	14,935	2,809,709	1,969	141439661	9.78014143966e+12	Jane Austen, Tony Tanner, Ros Ballaster
4	1,813	10	1,885	1,885	3,060,926	3,455	679783261	9.78067978327e+12	Jane Austen
5	1,818	71	18,490	18,490	4,836,639	2,618	141439475	9.78014143947e+12	Mary Wollstonecraft Shelley, Percy Bysshe
6	1,847	43	10,210	10,210	2,977,639	2,568	142437204	9.78014243721e+12	Charlotte Brontë, Michael Mason
7	1,847	63	6,185	6,185	1,565,818	2,498	393978893	9.7803939789e+12	Emily Brontë, Richard J. Dunn
8	1,859	83	1,953	1,953	2,956,372	525	141439602	9.7801414396e+12	Charles Dickens, Richard Maxwell, Hablot
9	1,868	42	1,934	1,934	3,244,642	1,707	451529308	9.7804515293e+12	Louisa May Alcott
10	1,884	58	2,956	2,956	1,835,605	2,277	142437174	9.78014243718e+12	Mark Twain, John Seelye, Guy Cardwell
11	1,891	95	5,297	5,297	1,858,012	2,303	375751513	9.78037575152e+12	Oscar Wilde, Jeffrey Eugenides
12	1,897	97	17,245	17,245	3,165,724	2,207	393970124	9.78039397013e+12	Bram Stoker, Nina Auerbach, David J. Skal
13	1,911	93	2,998	2,998	3,186,437	1,350	517189607	9.7805171896e+12	Frances Hodgson Burnett
14	1,925	5	4,671	4,671	245,494	1,356	743273567	9.78074327356e+12	F. Scott Fitzgerald
15	1,932	55	5,129	5,129	3,204,877	515	60929871	9.78006092988e+12	Aldous Huxley
16	1,936	66	18,405	18,405	3,358,283	409	446675539	9.78044667554e+12	Margaret Mitchell
17	1,937	32	890	890	40,283	373	142000671	9.78014200067e+12	John Steinbeck
18	1,937	7	5,907	5,907	1,540,236	969	618260307	9.7806182603e+12	J.R.R. Tolkien
19	1,945	14	7,613	7,613	2,207,778	896	452284244	9.78045228424e+12	George Orwell
20	1,946	80	157,993	157,993	2,180,358	1,708	156012197	9.7801560122e+12	Antoine de Saint-Exupéry, Richard Howard
21	1,947	15	48,855	48,855	3,532,896	710	553296981	9.78055329698e+12	Anne Frank, Eleanor Roosevelt, B.M. Mooy
22	1,949	13	5,470	5,470	153,313	995	451524934	9.78045152494e+12	George Orwell, Erich Fromm, Celâl Üster
23	1,950	37	100,915	100,915	4,790,821	474	60764899	9.78006076489e+12	C.S. Lewis
24	1,951	8	5,107	5,107	3,036,731	360	316769177	9.78031676917e+12	J.D. Salinger
25	1,952	59	24,178	24,178	987,048	180	64410935	9.78006441094e+12	E.B. White, Garth Williams, Rosemary Wells
26	1,953	48	4,381	4,381	1,272,463	507	307347974	9.78030734798e+12	Ray Bradbury
27	1,954	28	7,624	7,624	2,766,512	458	140283331	9.78014028333e+12	William Golding
28	1,954	19	34	34	3,204,327	566	618346252	9.78061834626e+12	J.R.R. Tolkien
29	1,958	87	1,617	1,617	265,616	109	374500010	9.78037450002e+12	Elie Wiesel, Marion Wiesel
30	1,960	4	2,657	2,657	3,275,794	487	61120081	9.78006112008e+12	Harper Lee
31	1,964	85	370,493	370,493	30,530	81	60256656	9.78006025665e+12	Shel Silverstein
32	1,967	90	231,804	231,804	1,426,690	156	014038572X	9.78014038572e+12	S.E. Hinton
33	1,967	94	320	320	3,295,655	555	60531045	9.78006053104e+12	Gabriel García Márquez, Gregory Rabassa
34	1,969	65	4,981	4,981	1,683,562	241	385333846	9.78038533385e+12	Kurt Vonnegut Jr.
35	1,973	89	21,787	21,787	992,628	129	345418263	9.78034541826e+12	William Goldman

Figura 2- Resultado da pergunta 1

2-Para cada livro com classificação inferior a quatro, listar o título que lhe corresponde.

```
select title
from books
where book_id in (select book_id
                  from ratings
                  where rating<4);
```

Resultado da query:

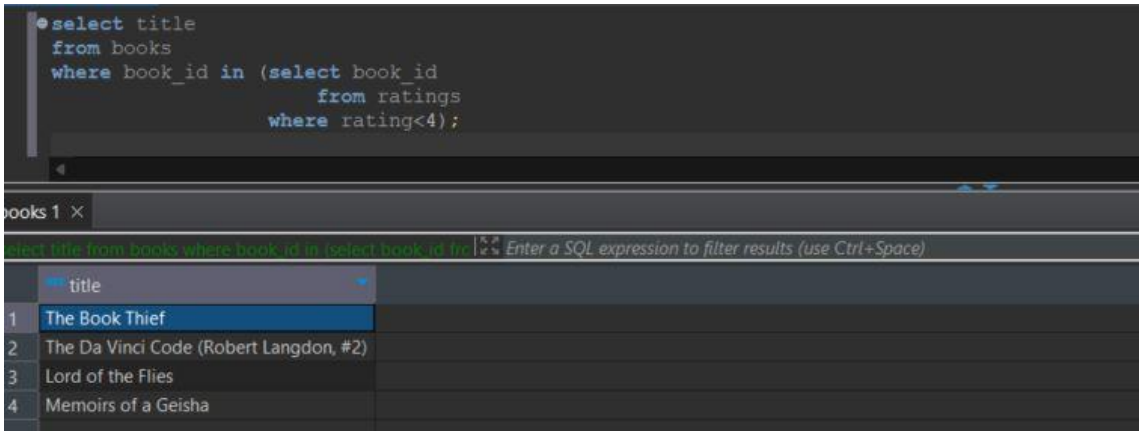


Figura 3- Resultado da 2ª questão analítica

3-Quantos livros existem sobre o utilizador user_id=8?

```
select count(*)
from ratings
where user_id = 8;
```

Resultado da query:

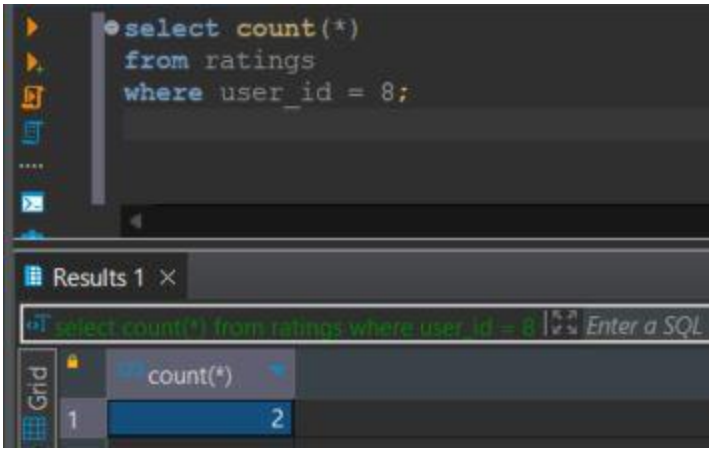
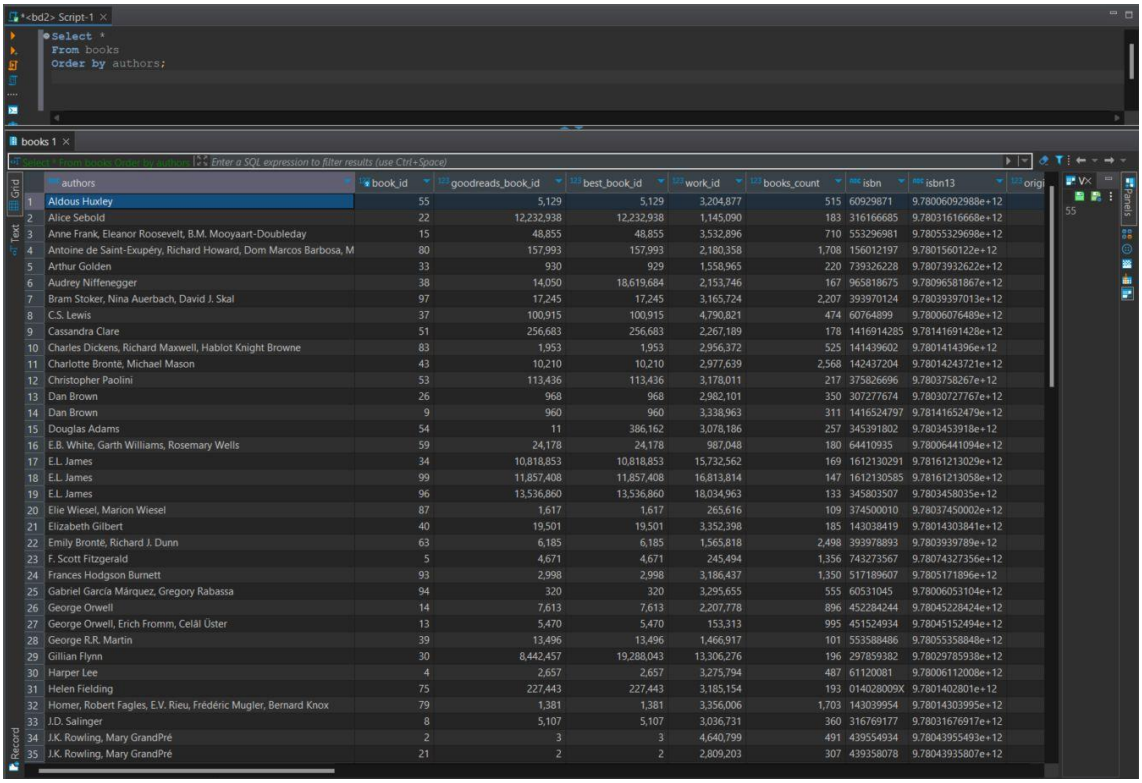


Figura 4- Resultado da 3ª questão analítica

4-Listar os vários livros por ordem do respetivo autor.

```
Select *
From books
Order by authors;
```

Resultado da query:



	authors	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13
1	Aldous Huxley	55	5,129	5,129	3,204,877	515	60929871	9.78006092988e+12
2	Alice Sebold	22	12,232,938	12,232,938	1,145,090	183	316166685	9.78031616668e+12
3	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-Doubleday	15	48,855	48,855	3,532,896	710	553296981	9.78055329698e+12
4	Antoine de Saint-Exupéry, Richard Howard, Dom Marcos Barbosa, M	80	157,993	157,993	2,180,358	1,708	156012197	9.78015601222e+12
5	Arthur Golden	33	930	929	1,558,965	220	739326228	9.78073932622e+12
6	Audrey Niffenegger	38	14,050	18,619,684	2,153,746	167	965818675	9.78096581867e+12
7	Bram Stoker, Nina Auerbach, David J. Skal	97	17,245	17,245	3,165,724	2,207	393970124	9.78039397013e+12
8	C.S. Lewis	37	100,915	100,915	4,790,821	474	60764899	9.78006076489e+12
9	Cassandra Clare	51	256,683	256,683	2,267,189	178	1416914285	9.78141691428e+12
10	Charles Dickens, Richard Maxwell, Hablot Knight Browne	83	1,953	1,953	2,956,372	525	141439602	9.7801414396e+12
11	Charlotte Brontë, Michael Mason	43	10,210	10,210	2,977,639	2,568	142437204	9.78014243721e+12
12	Christopher Paolini	53	113,436	113,436	3,178,011	217	375826696	9.7803758267e+12
13	Dan Brown	26	968	968	2,982,101	350	307277674	9.78030727767e+12
14	Dan Brown	9	960	960	3,338,963	311	1416524797	9.78141652479e+12
15	Douglas Adams	54	11	386,162	3,078,186	257	345391802	9.7803453918e+12
16	E.B. White, Garth Williams, Rosemary Wells	59	24,178	24,178	987,048	180	64410935	9.78006441094e+12
17	E.L. James	34	10,818,853	10,818,853	15,732,562	169	1612130291	9.78161213029e+12
18	E.L. James	99	11,857,408	11,857,408	16,813,814	147	1612130585	9.78161213058e+12
19	E.L. James	96	13,536,860	13,536,860	18,034,963	133	345803507	9.7803458035e+12
20	Elie Wiesel, Marion Wiesel	87	1,617	1,617	265,616	109	374500010	9.78037450002e+12
21	Elizabeth Gilbert	40	19,501	19,501	3,352,398	185	143038419	9.78014303841e+12
22	Emily Brontë, Richard J. Dunn	63	6,185	6,185	1,565,818	2,498	393978893	9.7803939789e+12
23	F. Scott Fitzgerald	5	4,671	4,671	245,494	1,356	743273567	9.78074327356e+12
24	Frances Hodgson Burnett	93	2,998	2,998	3,186,437	1,350	517189607	9.7805171896e+12
25	Gabriel García Márquez, Gregory Rabassa	94	320	320	3,295,655	555	60531045	9.78006053104e+12
26	George Orwell	14	7,613	7,613	2,207,778	896	452284244	9.78045228424e+12
27	George Orwell, Erich Fromm, Celâl Üster	13	5,470	5,470	153,313	995	451524934	9.78045152494e+12
28	George R.R. Martin	39	13,496	13,496	1,466,917	101	553588486	9.78055358848e+12
29	Gillian Flynn	30	8,442,457	19,288,043	13,306,276	196	297859382	9.78029785938e+12
30	Harper Lee	4	2,657	2,657	3,275,794	487	61120081	9.78006112008e+12
31	Helen Fielding	75	227,443	227,443	3,185,154	193	014028009X	9.7801402801e+12
32	Homer, Robert Fagles, E.V. Rieu, Frédéric Mugler, Bernard Knox	79	1,381	1,381	3,356,006	1,703	143039954	9.78014303995e+12
33	J.D. Salinger	8	5,107	5,107	3,036,731	360	316769177	9.78031676917e+12
34	J.K. Rowling, Mary GrandPré	2	3	3	4,640,799	491	439554934	9.78043955493e+12
35	J.K. Rowling, Mary GrandPré	21	2	2	2,809,203	307	439358078	9.78043935807e+12

Figura 5- Resultado da 4ª questão analítica

Pergunta 5- Book_id, obra original e avaliação média de todos os livros com classificação superior a 4.5 estrelas.

```
Select book_id, original_tittle, average_rating

From books

Where average_rating > 4.5;
```

Resultado da query:

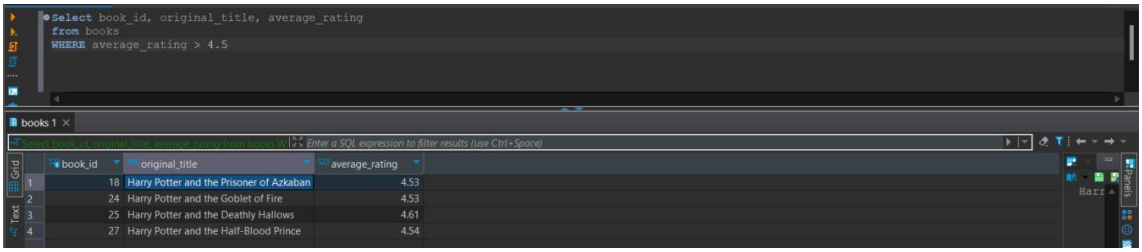


Figura 6- Resultado da 5ª questão analítica

Pergunta 6- Mostrar autores e o livro original que possuam mais de 4 milhões de avaliações.

```
Select authors, original_tittle

From books

Where ratings_count > 4000000
```

Resultado da query:

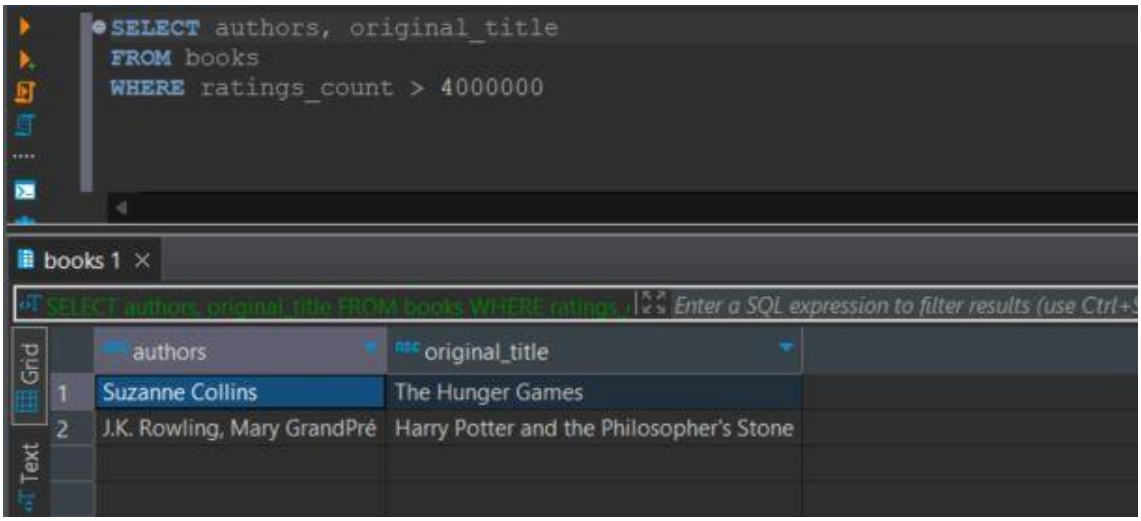


Figura 7- Resultado da 6ª questão analítica

Conclusão

Nesta primeira fase do relatório, tivemos a oportunidade de aplicar os conhecimentos adquiridos até ao momento, desenvolvendo habilidades em análise de dados, criação de bases de dados e consultas SQL. Ao trabalhar em grupo, analisamos cuidadosamente o conjunto de dados selecionado, extraindo informações valiosas e consolidando a nossa compreensão dos conceitos fundamentais de análise de dados e engenharia de software. Além disso, fortalecemos as nossas habilidades em colaboração e comunicação ao partilhar ideias e superar desafios juntos.

Durante o processo, tivemos a oportunidade de explorar as funcionalidades do sistema de gestão de bases de dados MySQL. Isso permitiu-nos extrair informações relevantes do conjunto de dados Goodbooks e obter insights sobre a plataforma.

Através da criação dos modelos de dados, como o modelo conceptual de dados e o diagrama Entidade-Relacionamento (ER), estabelecemos a estrutura necessária para armazenar os dados de forma organizada e estabelecemos as relações entre as diferentes entidades. Com base nesses modelos, criamos tabelas no MySQL e importamos os dados relevantes do conjunto de dados Goodbooks.

Ao realizar consultas SQL, pudemos responder a várias questões analíticas propostas. Essas consultas proporcionaram nos uma compreensão mais profunda dos dados e a capacidade de extrair informações úteis para análise posterior. Essas questões analíticas serão respondidas mais à frente nas tecnologias que se seguem.

Em suma, esta primeira fase do projeto proporcionou-nos uma experiência valiosa na aplicação prática dos conceitos aprendidos em sala de aula, reforçando as nossas capacidades em análise de dados, gestão de bases de dados e trabalho em equipe.

Agora, iremos migrar para o Cassandra, um modelo NoSQL altamente escalável.

Apache Cassandra

Introdução

Nesta segunda fase do projeto, iremos abordar a migração do sistema de gestão de bases de dados MySQL para o Apache Cassandra. O nosso objetivo é realizar a transição do GoodBooks, originalmente projetado para o MySQL, para um ambiente baseado em Cassandra.

O Apache Cassandra é uma base de dados distribuída e altamente escalável, projetada para lidar com grandes volumes de dados em ambientes distribuídos, sem pontos únicos de falha. Ao contrário do MySQL, que adota uma abordagem relacional, o Cassandra segue um modelo NoSQL, com uma estrutura de colunas distribuídas.

Durante esta fase, iremos analisar o esquema de dados do GoodBooks e adaptá-lo para um formato adequado para o Cassandra. Isso envolverá a reestruturação do modelo de dados para tirar partido dos recursos e da arquitetura do Cassandra.

Ao migrar para o Cassandra, iremos explorar as vantagens desta tecnologia, tais como a escalabilidade horizontal, a alta disponibilidade e o desempenho otimizado para grandes volumes de dados. Iremos compreender como lidar com os desafios e as considerações específicas associadas ao uso do Cassandra em comparação com o MySQL.

Através da migração do GoodBooks para o Cassandra, iremos adquirir conhecimentos valiosos sobre a implementação e administração de bases de dados NoSQL distribuídas. Desta forma, estaremos preparados para enfrentar desafios em projetos futuros que envolvam sistemas de gestão de bases de dados distribuídos e altamente escaláveis.

Com esta introdução, estamos prontos para iniciar a migração do GoodBooks para o Cassandra e explorar todo o potencial desta tecnologia no âmbito do nosso projeto.

Fase de transformação

Para utilizar o sistema NOSQL Cassandra, é necessário passar por um processo de transformação, pois é preciso consolidar todas as informações em uma única tabela, ao invés de mantê-las dispersas em 5 tabelas interconectadas. A seguir, será apresentada uma explicação sobre como realizar essa transformação.

1. Ainda no servidor MYSQL, devemos ter as tabelas criadas e carregadas.

```sql
DROP DATABASE IF EXISTS goodbooks;

CREATE DATABASE IF NOT EXISTS goodbooks;

USE goodbooks;

DROP TABLE IF EXISTS books;
CREATE TABLE IF NOT EXISTS books (
    book_id INT PRIMARY KEY,
    goodreads_book_id INT,
    best_book_id INT,
    work_id INT,
    books_count INT,
    isbn VARCHAR(20),
    isbn13 VARCHAR(20),
    authors VARCHAR(200),
    original_publication_year INT,
    original_title VARCHAR(200),
    title VARCHAR(200),
    language_code VARCHAR(10),
    average_rating DECIMAL(3,2),
    ratings_count INT,
    work_ratings_count INT,
    work_text_reviews_count INT,
    ratings_1 INT,
    ratings_2 INT,
    ratings_3 INT,
    ratings_4 INT,
    ratings_5 INT,
    image_url VARCHAR(500),
    small_image_url VARCHAR(500)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

DROP TABLE IF EXISTS tags;
CREATE TABLE IF NOT EXISTS tags (
    tag_id INT PRIMARY KEY,
    tag_name VARCHAR(50) NOT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

DROP TABLE IF EXISTS book_tags;
CREATE TABLE IF NOT EXISTS book_tags (
    book_id INT,
    tag_id INT,
    count INT,
    PRIMARY KEY (book_id, tag_id, count),
    FOREIGN KEY (book_id) REFERENCES books (book_id),
    FOREIGN KEY (tag_id) REFERENCES tags (tag_id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

DROP TABLE IF EXISTS ratings;
CREATE TABLE IF NOT EXISTS ratings (
    user_id INT NOT NULL,
    book_id INT NOT NULL,
    rating INT NOT NULL,
    PRIMARY KEY (user_id, book_id, rating),
    FOREIGN KEY (book_id) REFERENCES books (book_id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

DROP TABLE IF EXISTS to_read;
CREATE TABLE IF NOT EXISTS to_read (
    user_id INT NOT NULL,
    book_id INT NOT NULL,
    PRIMARY KEY (user_id, book_id),
    FOREIGN KEY (book_id) REFERENCES books (book_id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

Fig 8- Código referente à criação de tabelas

```sql
LOAD DATA INFILE '/var/lib/mysql/csv/book_tags.csv'
INTO TABLE book_tags
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/books.csv'
INTO TABLE books
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/ratings.csv'
INTO TABLE ratings
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/tags.csv'
INTO TABLE tags
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA INFILE '/var/lib/mysql/csv/to_read.csv'
INTO TABLE to_read
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

Figura 9- Código referente ao carregamento das tabelas

2. Após a criação e carregamento das tabelas, já podemos criar a tabela com todos os dados.

```sql
DROP TABLE IF EXISTS books_all;
CREATE TABLE books_all AS
SELECT
  books.book_id,
  books.goodreads_book_id,
  books.best_book_id,
  books.work_id,
  books.books_count,
  books.isbn,
  books.isbn13,
  books.authors,
  books.original_publication_year,
  books.original_title,
  books.title,
  books.language_code,
  books.average_rating,
  books.ratings_count,
  books.work_ratings_count,
  books.work_text_reviews_count,
  books.ratings_1,
  books.ratings_2,
  books.ratings_3,
  books.ratings_4,
  books.ratings_5,
  books.image_url,
  books.small_image_url,
  COALESCE(tags.tag_id, 00) AS tag_id,
  COALESCE(tags.tag_name, '00') AS tag_name,
  COALESCE(book_tags.count, 00) AS book_tag_count,
  COALESCE(ratings.user_id, 00) AS ratings_user_id,
  COALESCE(ratings.rating, 00) AS user_rating,
  COALESCE(to_read.user_id, 00) AS to_read_user_id
FROM books
LEFT JOIN book_tags ON books.book_id = book_tags.book_id
LEFT JOIN tags ON book_tags.tag_id = tags.tag_id
LEFT JOIN ratings ON books.book_id = ratings.book_id
LEFT JOIN to_read ON books.book_id = to_read.book_id;


select * from books_all;
```

Figura 10- Código utilizado para a criação da "tabela gigante".

O código acima cria a tabela com todos os dados e dá-lhe o nome "books_all".

O comando COALESCE foi necessário (no nosso caso) para trocar valores que poderiam ficar a NULL por 0. Assim, evitamos problemas futuros de erros referentes a chaves estrangeiras.

O select final foi necessário para visualizar a tabela completa e, utilizando as ferramentas do DBeaver, exportá-la em formato CSV para utilizá-la como o nosso Dataset Cassandra.

Neste ponto já temos o novo Dataset organizado numa só tabela, que será caracterizada mais à frente.

3. Já no servidor Cassandra, criamos o keyspace e a respetiva tabela.

```
CREATE KEYSPACE IF NOT EXISTS goodbooks_keyspace WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};

DROP TABLE IF EXISTS goodbooks_keyspace.books_all;
CREATE TABLE goodbooks_keyspace.books_all (
    book_id INT,
    goodreads_book_id INT,
    best_book_id INT,
    work_id INT,
    books_count INT,
    isbn TEXT,
    isbn13 TEXT,
    authors TEXT,
    original_publication_year INT,
    original_title TEXT,
    title TEXT,
    language_code TEXT,
    average_rating DECIMAL,
    ratings_count INT,
    work_ratings_count INT,
    work_text_reviews_count INT,
    ratings_1 INT,
    ratings_2 INT,
    ratings_3 INT,
    ratings_4 INT,
    ratings_5 INT,
    image_url TEXT,
    small_image_url TEXT,
    tag_id INT,
    tag_name TEXT,
    book_tag_count INT,
    ratings_user_id INT,
    user_rating INT,
    to_read_user_id INT,
    PRIMARY KEY (book_id, tag_id, ratings_user_id, to_read_user_id)
);
```

Figura 11- Código utilizado para a criação da tabela gigante na Cassandra.

Um keyspace é uma espécie de database para Cassandra.

Após criarmos o keyspace, criamos a tabela, de forma semelhante à criação de tabelas em SQL.

4. Abrir o CQLS no terminal e colocar o comando COPY.

```
pires — com.docker.cli ◂ docker exec -it bd2_cassandra cqlsh — 108×34
Last login: Fri Apr 14 15:14:02 on ttys000
pires@Pedros-MacBook-Pro ~ % docker exec -it bd2_cassandra cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.0.0 | Cassandra 4.0.3 | CQL spec 3.4.5 | Native protocol v5]
Use HELP for help.
cqlsh> COPY goodbooks_keyspace.books_all (book_id, goodreads_book_id, best_book_id, work_id, books_count, is
bn, isbn13, authors, original_publication_year, original_title, title, language_code, average_rating, rating
s_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5,
 image_url, small_image_url, tag_id, tag_name, book_tag_count, ratings_user_id, user_rating, to_read_user_id
)
   ... FROM '/var/lib/cassandra/csv/books_all.csv'
   ... WITH DELIMITER = ',' AND HEADER = true;
Using 3 child processes

Starting copy of goodbooks_keyspace.books_all with columns [book_id, goodreads_book_id, best_book_id, work_i
d, books_count, isbn, isbn13, authors, original_publication_year, original_title, title, language_code, aver
age_rating, ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3, rat
ings_4, ratings_5, image_url, small_image_url, tag_id, tag_name, book_tag_count, ratings_user_id, user_ratin
g, to_read_user_id].
```

Figura 12- Conexão ao Test Cluster seguida do comando COPY utilizado para preencher a tabela.

Informação sobre o Dataset

O conjunto de dados selecionado para análise contém informações detalhadas sobre diversos livros disponíveis na plataforma GoodReads, conhecida como GoodBooks. Essa plataforma oferece uma vasta gama de livros, com informações como sinopse, autor, editora, data de publicação e outros detalhes relevantes para os leitores.

Uma característica importante da plataforma GoodReads é a possibilidade de os utilizadores atribuírem uma nota de 1 a 5 estrelas aos livros que leram, permitindo que outros leitores saibam o que esperar da obra antes de iniciá-la. Além disso, a plataforma permite que os utilizadores adicionem livros à sua lista "ver mais tarde", tornando mais fácil para eles lembrarem-se dos títulos que desejam ler no futuro.

Com base nesses dados, é possível realizar uma análise aprofundada sobre os livros mais populares, os autores mais bem-sucedidos e outros insights interessantes sobre o mundo da literatura.

O dataset passou a ser composto apenas por 1 ficheiro CSV:

- Books_all: Contém diversas informações sobre cada livro no conjunto de dados, sobre as suas tags(etiquetas) atribuídas, sobre as avaliações dos utilizadores e sobre os utilizadores que colocaram determinado livro na sua fila “ler mais tarde”.
  Atributos:
  - Book_id: Id que caracteriza cada livro no conjunto de dados GoodBooks.
  - Goodreads_book_id: Id do livro atribuído pela GoodReads.
  - Best_book_id: Id do “melhor livro” de um determinado trabalho literário.
  - Work_id: Id de cada trabalho literário.
  - Books_count: Contagem de livros diferentes associados a um determinado “Work_Id”.
  - Isbn: International Standart Book Number – Um número de 10 dígitos que identifica cada livro (cada obra), número esse dado pela agência internacional.
  - Isbn13: International Standart Book Number – Semelhante ao Isbn, porém, com 13 dígitos. Os três dígitos a mais identificam a língua ou o país, algo que não é distinguível no ibsn.
  - Authors: Autor/es de cada livro. Quando há mais que um, os autores estão separados por vírgulas.
  - Original_publication_year: Ano de publicação original do livro (primeira vez em que foi laçado).
  - Original_tittle: Título original do livro.
  - Tittle: Nome do livro.
  - Language_code: Código do idioma em que o livro foi originalmente escrito (Ex: Pt).
  - Average Rating: Representa a classificação média do livro atribuída pelos utilizadores do Goodreads. (Os utilizadores atribuem uma avaliação de 1 a 5).
  - Ratings_count: Quantidade total de avaliações que foram dadas pelos utilizadores para um determinado livro.

- Work_ratings_count: Quantidade total de avaliações que foram dadas pelos utilizadores para uma determinada obra, independentemente da brochura, capa ou idioma.
- Work_text_reviews_count: Número de comentários para uma determinada obra na plataforma Goodreads.
- Ratings_1: Número de avaliações que um livro recebeu com classificação de 1 estrela.
- Ratings_2: Número de avaliações que um livro recebeu com classificação de 2 estrelas.
- Ratings_3: Número de avaliações que um livro recebeu com classificação de 3 estrelas.
- Ratings_4: Número de avaliações que um livro recebeu com classificação de 4 estrelas.
- Ratings_5: Número de avaliações que um livro recebeu com classificação de 5 estrelas.
- Image_url: URL da imagem da capa do livro.
- Small_image_url: URL da imagem da capa do livro em formato pequeno.
- Tag_id: Identificador único para cada tag.
- Tag_name: Nome da tag, ou seja, a palavra ou frase que os utilizadores utilizaram para rotular o livro.
- Book_tags_Count: Número de utilizadores que atribuíram a tag ao livro.
- Ratings_user_id: Utilizador que deu a avaliação (user_rating).
- User_rating: Avaliação dada por determinado utilizador (Ratings_user_id).
- To_read_user_id: Utilizador que colocou determinado livro na lista “Ler mais tarde”.

Dicionário de Dados

Tabela	books_all		
Descrição	Este conjunto de dados inclui um conjunto enorme de dados sobre cada livro, assim como as avaliações, utilizadores que o colocaram em "ler mais tarde" e as suas respetivas etiquetas (tags).		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
book_id	código do livro	int	PK
goodreads_book_id	código do livro atribuído pela GoodReads	int	
best_book_id	código do melhor livro de um determinado trabalho literário	int	
work_id	código de cada trabalho literário	int	
books_count	Número de livros diferentes associados a um determinado trabalho literário	int	
isbn	Número de 10 dígitos que identifica cada livro	text	
isbn13	Número de 13 dígitos que identifica cada livro. Funciona de forma semelhante ao Isbn mas identifica também a língua ou o país.	text	
authors	Autor/es de cada livro	text	
original_publication_year	Ano de publicação original do livro	int	
original_tittle	Título original do livro	text	
tittle	Nome do livro	text	
language_code	Código do idioma em que o livro foi originalmente escrito	text	
average_rating	Classificação média do livro atribuída pelos utilizadores do Goodreads	decimal	
ratings_count	Número total de avaliações para um determinado livro	int	
work_ratings_count	Número total de avaliações para uma determinada obra	int	
work_text_reviews_count	Número de comentários para uma determinada obra	int	
ratings_1	Quantidade de avaliações com 1 estrela	int	
ratings_2	Quantidade de avaliações com 2 estrela	int	
ratings_3	Quantidade de avaliações com 3 estrela	int	
ratings_4	Quantidade de avaliações com 4 estrela	int	
ratings_5	Quantidade de avaliações com 5 estrela	int	
image_url	URL da imagem da capa do livro	text	
small_image_url	URL da imagem da capa do livro em formato pequeno	text	
tag_id	Identificador único para cada tag		PK
tag_name	Nome da tag, ou seja, a palavra ou frase que os utilizadores utilizaram para rotular o livro.	text	
book_tags_count	Número de utilizadores que atribuíram a tag ao livro.	int	
ratings_book_id	Utilizador que deu a avaliação (user_rating).	int	PK
user_rating	Avaliação dada por determinado utilizador (ratings_user_id).	int	
to_read_user_id	Utilizador que colocou determinado livro na lista "Ler mais tarde".	int	PK

Modelos de dados

Modelo conceptual de dados:

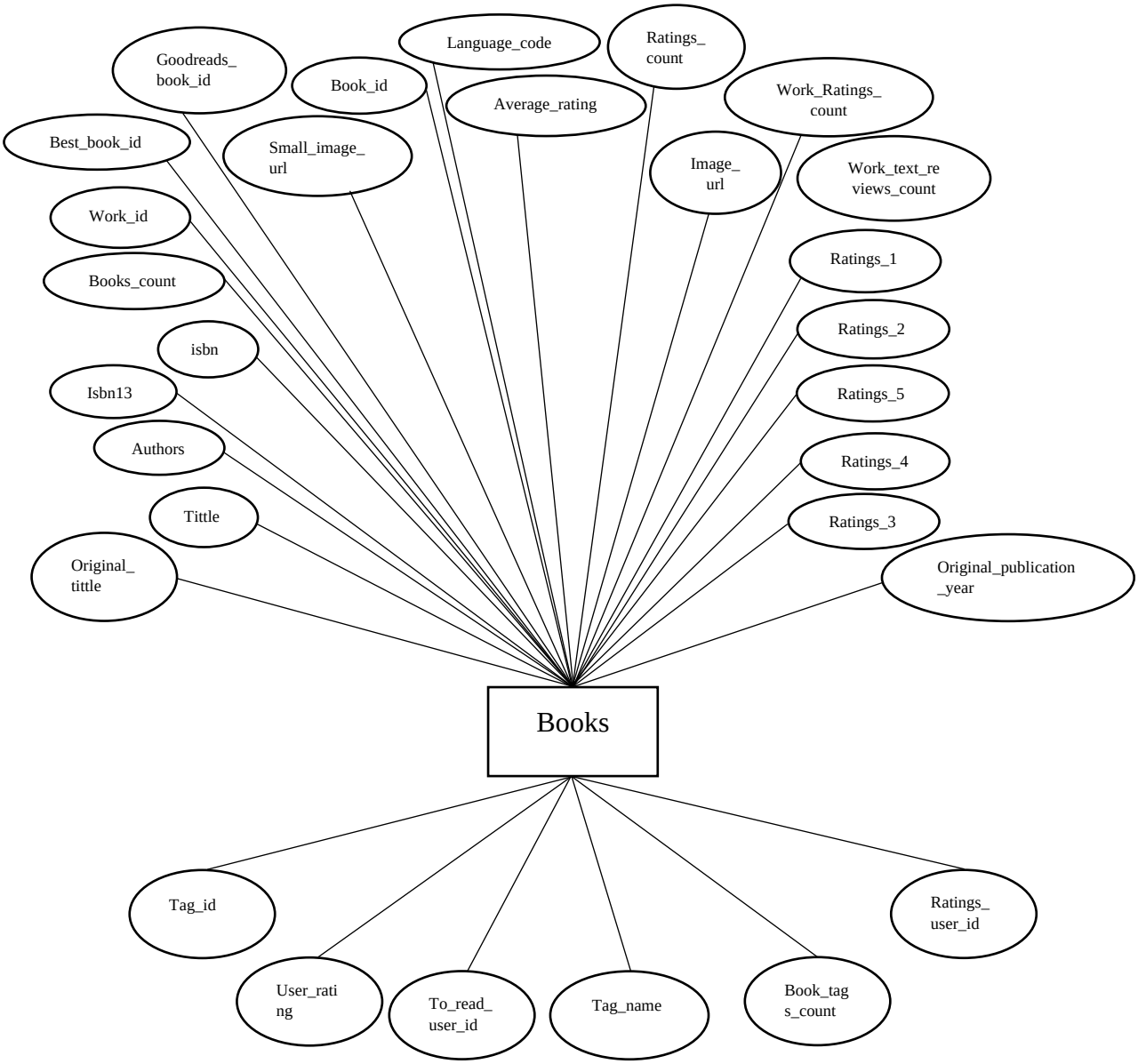


Diagrama 2- Modelo conceptual Cassandra

Diagrama Entidade-Relacionamento (ER)

books (book_id, tag_id, ratings_user_id, to_read_user_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title, language_code, average_rating, ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url, tag_name, book_tags_count, user_rating)

Questões analíticas

Ao utilizar o Cassandra, as consultas SQL tradicionais não funcionam da mesma forma que em SQL. O grupo, baseando-se em exemplos de CQL (Cassandra Query Language), conseguiu adaptar as consultas SQL, que antes eram simples, para um método inovador e até então desconhecido.

Ao trabalhar com Cassandra, é necessário criar tabelas específicas para cada consulta, preenchendo-as com os valores da tabela unificada "books_all". Assim, é possível executar uma consulta adaptada que o Cassandra possa interpretar.

Por meio de um processo de tentativa e erro, bem como extensa pesquisa online, aprendemos conceitos essenciais que nos permitiram obter as respostas corretas. O mais importante foram os conceitos de clustering e chave de partição, que nos ajudaram a contornar as limitações do Cassandra (por exemplo: no Cassandra, só é possível utilizar "GROUP BY" em colunas de chave de partição e/ou colunas de clustering).

Embora ainda não dominemos completamente o assunto, compreendemos como utilizar esses conceitos para formular adequadamente as perguntas que precisam ser respondidas.

Foi dessa maneira que conseguimos obter os resultados desejados.

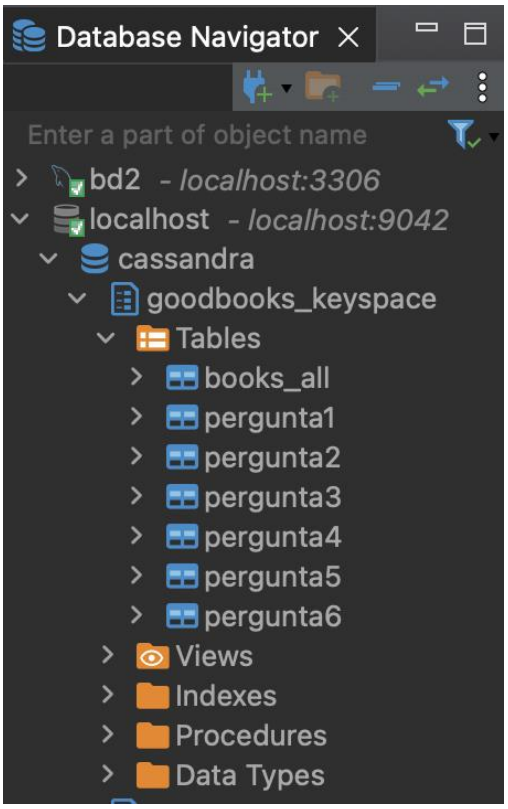


Figura 13- Tabelas criadas para as respostas às questões analíticas.

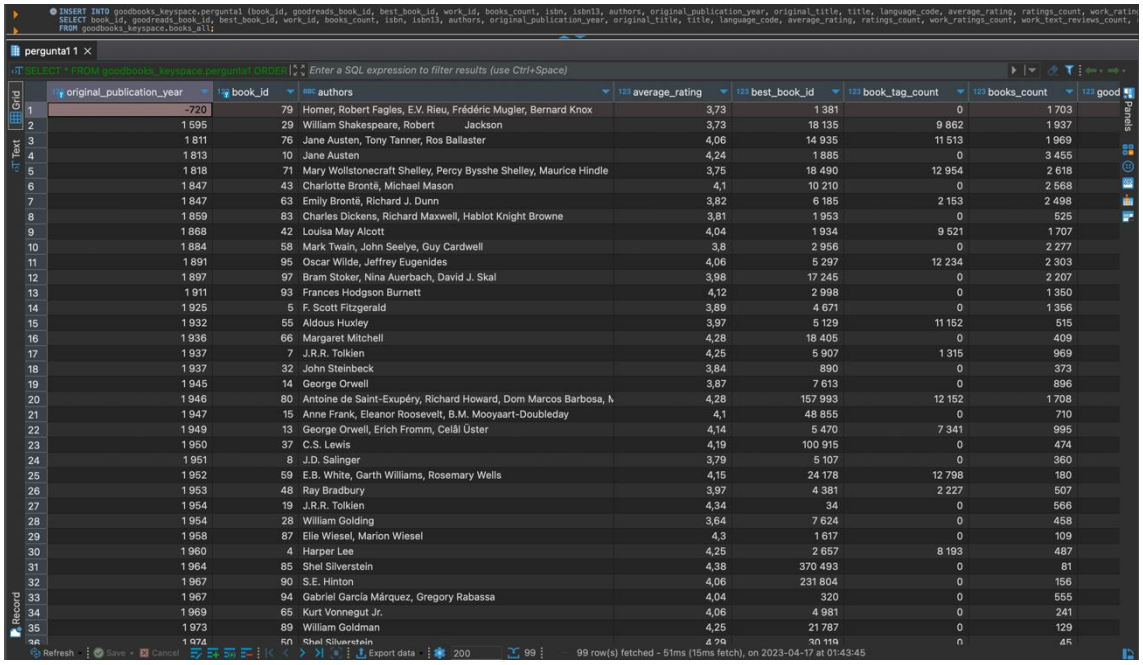
1-Listar os livros por ordem crescente pelo respetivo ano de publicação.

```
CREATE TABLE goodbooks_keyspace.pergunta1 (
        book_id INT,
        goodreads_book_id INT,
        best_book_id INT,
        work_id INT,
        books_count INT,
        isbn TEXT,
        isbn13 TEXT,
        authors TEXT,
        original_publication_year INT,
        original_title TEXT,
        title TEXT,
        language_code TEXT,
        average_rating DECIMAL,
        ratings_count INT,
        work_ratings_count INT,
        work_text_reviews_count INT,
        ratings_1 INT,
        ratings_2 INT,
        ratings_3 INT,
        ratings_4 INT,
        ratings_5 INT,
        image_url TEXT,
        small_image_url TEXT,
        tag_id INT,
        tag_name TEXT,
        book_tag_count INT,
        ratings_user_id INT,
        user_rating INT,
        to_read_user_id INT,
        PRIMARY KEY ((original_publication_year), book_id)
) WITH CLUSTERING ORDER BY (book_id ASC);

INSERT INTO goodbooks_keyspace.pergunta1 (book_id, goodreads_book_id, best_book_id,
work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title,
language_code, average_rating, ratings_count, work_ratings_count, work_text_reviews_count,
ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url, tag_id,
tag_name, book_tag_count, ratings_user_id, user_rating, to_read_user_id)
SELECT book_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13,
authors, original_publication_year, original_title, title, language_code, average_rating,
ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3,
ratings_4, ratings_5, image_url, small_image_url, tag_id, tag_name, book_tag_count,
ratings_user_id, user_rating, to_read_user_id
FROM goodbooks_keyspace.books_all;
```

```
SELECT *
FROM goodbooks_keyspace.pergunta1
ORDER BY original_publication_year ASC, book_id ASC;
```

Resultado da Query:



	original_publication_year	book_id	authors	average_rating	best_book_id	book_tag_count	books_count	good
1	-720	79	Homer, Robert Fagles, E.V. Rieu, Frédéric Mugler, Bernard Knox	3,73	1 381	0	1 703	
2	1 595	29	William Shakespeare, Robert Jackson	3,73	18 135	9 862	1 937	
3	1 811	76	Jane Austen, Tony Tanner, Ros Ballaster	4,06	14 935	11 513	1 969	
4	1 813	10	Jane Austen	4,24	1 885	0	3 455	
5	1 818	71	Mary Wollstonecraft Shelley, Percy Bysshe Shelley, Maurice Hindle	3,75	18 490	12 954	2 618	
6	1 847	43	Charlotte Brontë, Michael Mason	4,1	10 210	0	2 568	
7	1 847	63	Emily Brontë, Richard J. Dunn	3,82	6 185	2 153	2 498	
8	1 859	83	Charles Dickens, Richard Maxwell, Hablot Knight Browne	3,81	1 953	0	525	
9	1 868	42	Louisa May Alcott	4,04	1 934	9 521	1 707	
10	1 884	58	Mark Twain, John Seelye, Guy Cardwell	3,8	2 956	0	2 277	
11	1 891	95	Oscar Wilde, Jeffrey Eugenides	4,06	5 297	12 234	2 303	
12	1 897	97	Bram Stoker, Nina Auerbach, David J. Skal	3,98	17 245	0	2 207	
13	1 911	93	Frances Hodgson Burnett	4,12	2 998	0	1 350	
14	1 925	5	F. Scott Fitzgerald	3,89	4 671	0	1 356	
15	1 932	55	Aldous Huxley	3,97	5 129	11 152	515	
16	1 936	66	Margaret Mitchell	4,28	18 405	0	409	
17	1 937	7	J.R.R. Tolkien	4,25	5 907	1 315	969	
18	1 937	32	John Steinbeck	3,84	890	0	373	
19	1 945	14	George Orwell	3,87	7 613	0	896	
20	1 946	80	Antoine de Saint-Exupéry, Richard Howard, Dom Marcos Barbosa, M	4,28	157 993	12 152	1 708	
21	1 947	15	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-Doubleday	4,1	48 855	0	710	
22	1 949	13	George Orwell, Erich Fromm, Celâl Üster	4,14	5 470	7 341	995	
23	1 950	37	C.S. Lewis	4,19	100 915	0	474	
24	1 951	8	J.D. Salinger	3,79	5 107	0	360	
25	1 952	59	E.B. White, Garth Williams, Rosemary Wells	4,15	24 178	12 798	180	
26	1 953	48	Ray Bradbury	3,97	4 381	2 227	507	
27	1 954	19	J.R.R. Tolkien	4,34	34	0	566	
28	1 954	28	William Golding	3,64	7 624	0	458	
29	1 958	87	Elie Wiesel, Marion Wiesel	4,3	1 617	0	109	
30	1 960	4	Harper Lee	4,25	2 657	8 193	487	
31	1 964	85	Shel Silverstein	4,38	370 493	0	81	
32	1 967	90	S.E. Hinton	4,06	231 804	0	156	
33	1 967	94	Gabriel García Márquez, Gregory Rabassa	4,04	320	0	555	
34	1 969	65	Kurt Vonnegut Jr.	4,06	4 981	0	241	
35	1 973	89	William Goldman	4,25	21 787	0	129	
36	1 974	50	Shel Silverstein	4,29	30 119	0	45	

Figura 14- Resultado da 1ª questão analítica

2-Para cada livro com classificação inferior a quatro, listar o título que lhe corresponde.

```
CREATE TABLE goodbooks_keyspace.pergunta2 (
        book_id INT,
        title TEXT,
        ratings_user_id INT,
        user_rating INT,
PRIMARY KEY (book_id, user_rating, ratings_user_id)
);

INSERT INTO goodbooks_keyspace.pergunta2 (book_id, title, ratings_user_id, user_rating)
SELECT book_id, title, ratings_user_id, user_rating
FROM goodbooks_keyspace.books_all

WHERE user_rating > 0 AND user_rating < 4;
SELECT DISTINCT book_id, title
FROM goodbooks_keyspace.pergunta2;
```

Resultado da Query:

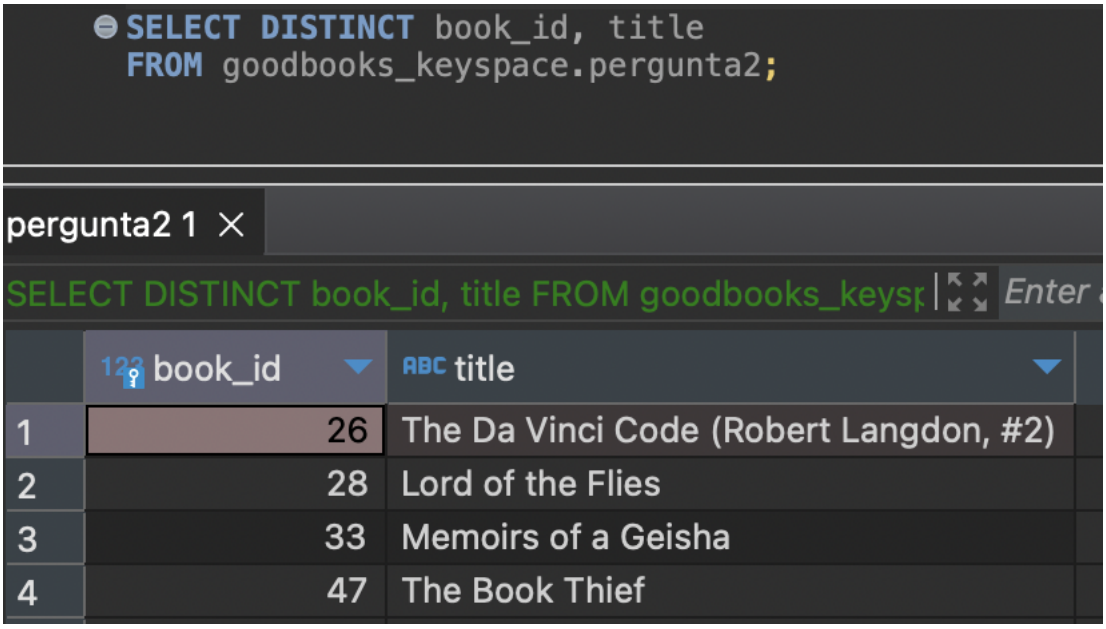


Figura 15- Resultado da 2ª questão analítica

3-Quantos livros existem sobre o utilizador user_id=8?

```
DROP TABLE IF EXISTS goodbooks_keyspace.pergunta3;
CREATE TABLE goodbooks_keyspace.pergunta3 (
        book_id INT,
        ratings_user_id INT,
        PRIMARY KEY (ratings_user_id, book_id)
);

INSERT INTO goodbooks_keyspace.pergunta3 (book_id, ratings_user_id)
SELECT book_id, ratings_user_id
FROM goodbooks_keyspace.books_all;

SELECT COUNT(*)
FROM goodbooks_keyspace.pergunta3
WHERE ratings_user_id = 8;
```

Resultado da Query:

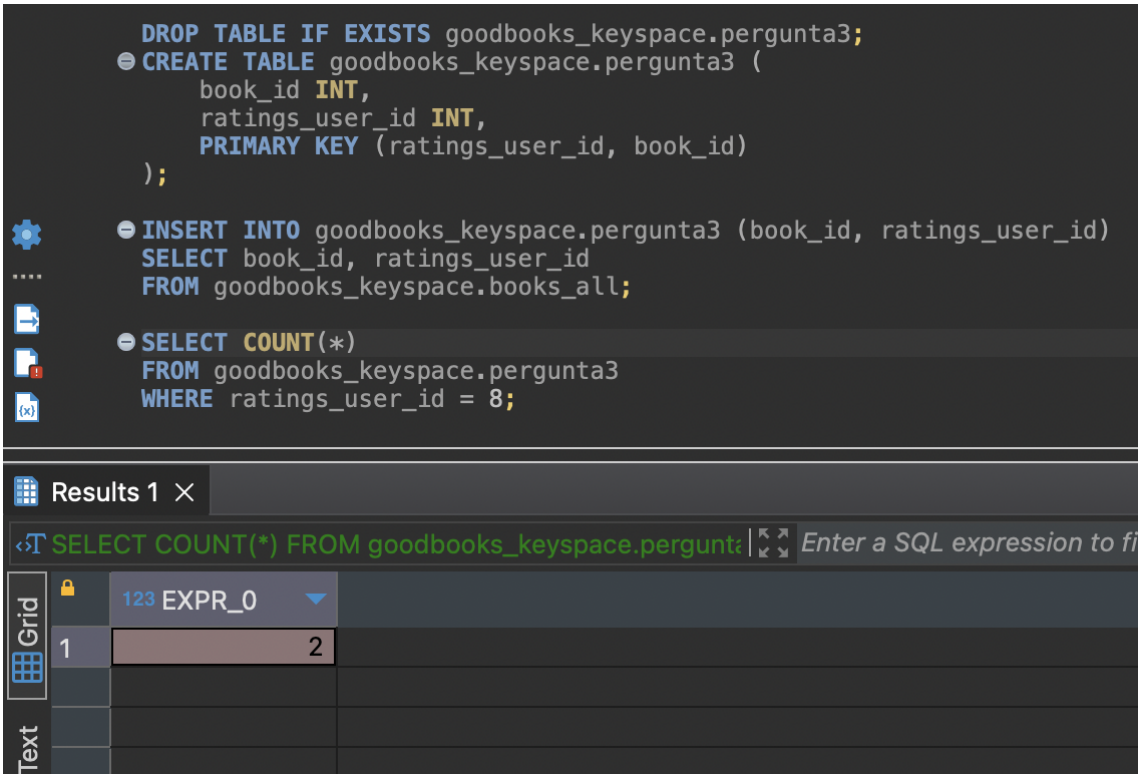


Figura 16- Resultado da 3ª questão analítica.

4-Listar os vários livros por ordem do respetivo autor.

```
CREATE TABLE goodbooks_keyspace.pergunta4 (
        book_id INT,
        goodreads_book_id INT,
        best_book_id INT,
        work_id INT,
        books_count INT,
        isbn TEXT,
        isbn13 TEXT,
        authors TEXT,
        original_publication_year INT,
        original_title TEXT,
        title TEXT,
        language_code TEXT,
        average_rating DECIMAL,
        ratings_count INT,
        work_ratings_count INT,
        work_text_reviews_count INT,
        ratings_1 INT,
        ratings_2 INT,
        ratings_3 INT,
        ratings_4 INT,
        ratings_5 INT,
        image_url TEXT,
        small_image_url TEXT,
        tag_id INT,
        tag_name TEXT,
        book_tag_count INT,
        ratings_user_id INT,
        user_rating INT,
        to_read_user_id INT,
        PRIMARY KEY ((authors), book_id)
) WITH CLUSTERING ORDER BY (book_id ASC);

INSERT INTO goodbooks_keyspace.pergunta4 (book_id, goodreads_book_id, best_book_id,
work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title,
language_code, average_rating, ratings_count, work_ratings_count, work_text_reviews_count,
ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url, tag_id,
tag_name, book_tag_count, ratings_user_id, user_rating, to_read_user_id)
SELECT book_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13,
authors, original_publication_year, original_title, title, language_code, average_rating,
ratings_count, work_ratings_count, work_text_reviews_count, ratings_1, ratings_2, ratings_3,
ratings_4, ratings_5, image_url, small_image_url, tag_id, tag_name, book_tag_count,
ratings_user_id, user_rating, to_read_user_id
FROM goodbooks_keyspace.books_all;
```

```
SELECT *
FROM goodbooks_keyspace.pergunta4
ORDER BY authors ASC, book_id ASC;
```

Resultado da Query:

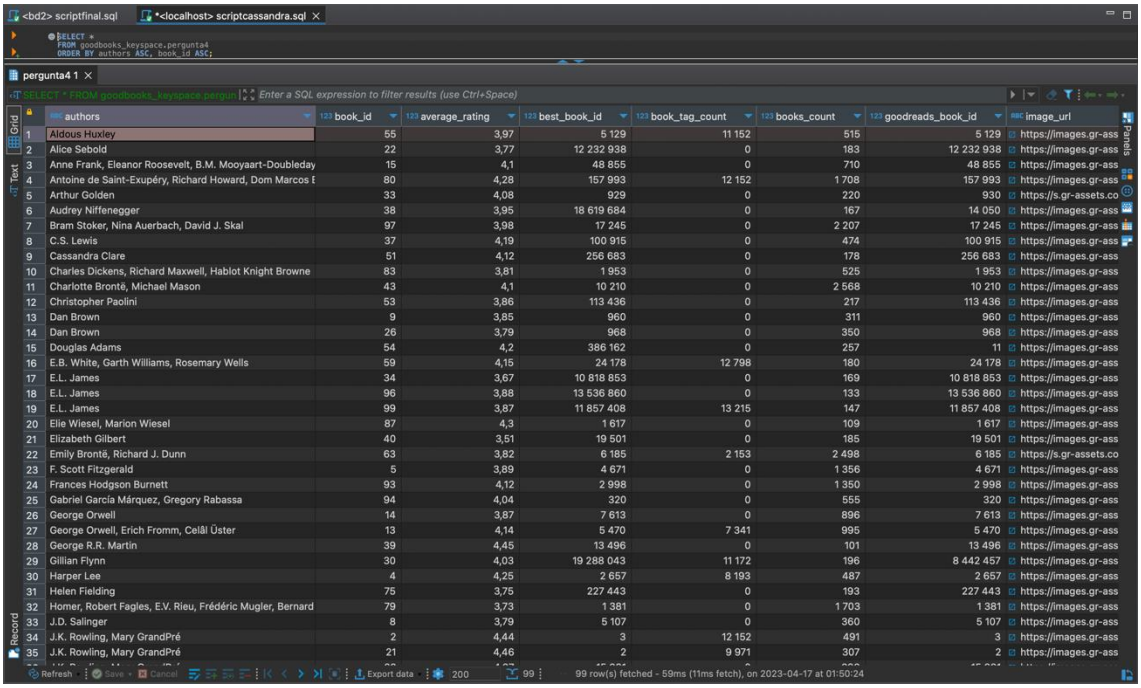


Figura 17 – Resultado da 4ª questão analítica.

5- Book_id, obra original e avaliação média de todos os livros com classificação superior a 4.5 estrelas.

```
CREATE TABLE goodbooks_keyspace.books_by_average_rating (
        book_id INT,
        original_title TEXT,
        average_rating DECIMAL,
        PRIMARY KEY (average_rating, book_id)
) WITH CLUSTERING ORDER BY (book_id ASC);

INSERT INTO goodbooks_keyspace.books_by_average_rating (book_id, original_title, average_rating)
SELECT book_id, original_title, average_rating
FROM goodbooks_keyspace.books_all;

SELECT book_id, original_title, average_rating
FROM goodbooks_keyspace.books_by_average_rating
WHERE average_rating > 4.5;
```

Resultado da Query:

```
DROP TABLE IF EXISTS goodbooks_keyspace.pergunta5;
CREATE TABLE goodbooks_keyspace.pergunta5 (
    book_id INT,
    original_title TEXT,
    average_rating DECIMAL,
    PRIMARY KEY (average_rating, book_id)
) WITH CLUSTERING ORDER BY (book_id ASC);

INSERT INTO goodbooks_keyspace.pergunta5 (book_id, original_title, average_rating)
SELECT book_id, original_title, average_rating
FROM goodbooks_keyspace.books_all;

SELECT book_id, original_title, average_rating
FROM goodbooks_keyspace.pergunta5
WHERE average_rating > 4.5;
```

unta5 1 ×

CT book_id, original_title, average_rating FROM goc | Enter a SQL expression to filter

123 book_id	ABC original_title	123 average_rating
18	Harry Potter and the Prisoner of Azkaban	4,53
24	Harry Potter and the Goblet of Fire	4,53
27	Harry Potter and the Half-Blood Prince	4,54
25	Harry Potter and the Deathly Hallows	4,61

Figura 18 – Resultado da 5ª questão analítica.

6- Mostrar autor e o livro original que contenha mais de 4 milhões de avaliações.

```
CREATE TABLE goodbooks_keyspace.pergunta6 (
        book_id INT,
        authors TEXT,
        original_title TEXT,
        ratings_count INT,
        PRIMARY KEY (ratings_count, book_id)
) WITH CLUSTERING ORDER BY (book_id ASC);

INSERT INTO goodbooks_keyspace.pergunta6 (book_id, authors, original_title, ratings_count)
SELECT book_id, authors, original_title, ratings_count
FROM goodbooks_keyspace.books_all;

SELECT authors, original_title
FROM goodbooks_keyspace.pergunta6
WHERE ratings_count > 4000000;
```

Resultado da Query:

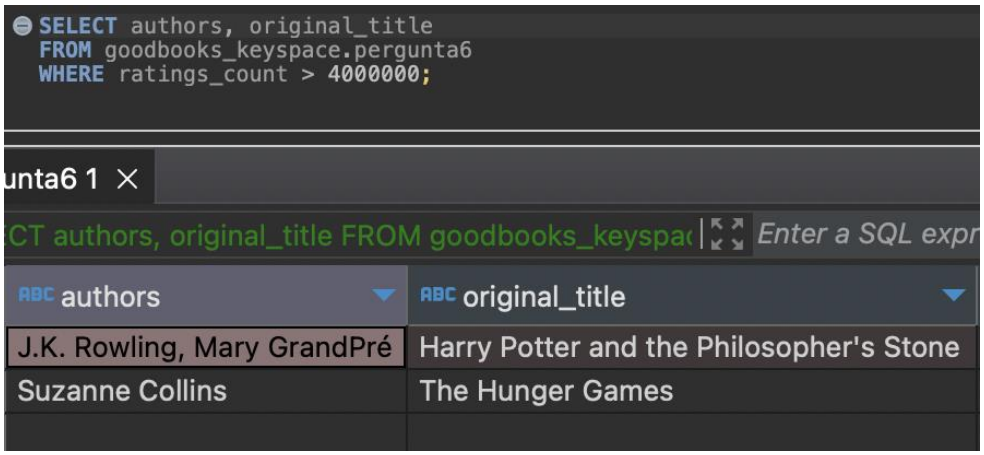


Figura 19 – Resultado da 6ª questão analítica.

Conclusão

Durante o processo de migração do conjunto de dados Goodbooks do banco de dados MySQL para o Cassandra, deparamo-nos com desafios que exigiram a nossa atenção e esforço, porém, conseguimos superá-los com sucesso. Ao longo desta transição, notamos diferenças significativas em relação ao MySQL utilizado na fase anterior, tais como a estrutura de dados e as funcionalidades disponíveis no Cassandra.

Essas diferenças levaram-nos a adotar uma abordagem cuidadosa e aprofundada na migração, buscando compreender e adaptar o modelo de dados para o Cassandra. Tivemos que considerar aspetos como a distribuição dos dados, a modelagem adequada para aproveitar a natureza distribuída do Cassandra e as técnicas de armazenamento e recuperação eficientes.

Apesar dos desafios enfrentados, acreditamos que a migração para o Cassandra foi uma experiência extremamente valiosa. Esta transição permitiu-nos explorar uma nova tecnologia de banco de dados distribuída e aprofundar o nosso conhecimento sobre as suas características e funcionalidades.

No final desta fase, concluímos que o processo de migração do conjunto de dados Goodbooks para o banco de dados Cassandra foi bem-sucedido. Além de cumprirmos o objetivo proposto, respondendo a todas as questões analíticas e obtendo as respostas esperadas, esta experiência contribuiu significativamente para a nossa aprendizagem e evolução profissional na área de bases de dados. Adquirimos conhecimentos práticos na utilização do Cassandra, compreendendo as suas particularidades e tornando-nos capazes de aproveitar os seus recursos para lidar com dados distribuídos de forma eficiente.

Consideramos esta transição como um marco importante no nosso projeto, pois proporcionou-nos a oportunidade de explorar diferentes tecnologias de banco de dados e expandir o nosso conjunto de habilidades, sendo esta a primeira tecnologia NoSQL com a qual tivemos o prazer de trabalhar. A migração do MySQL para o Cassandra fortaleceu a nossa compreensão das particularidades e desafios, sendo também importante para nós, pois o fizemos sem qualquer conhecimento prévio sobre o Cassandra. O grupo também aprendeu a utilizar várias fontes para obter informação e a aprender uma tecnologia de base de dados de forma mais autónoma, o que nos fez sentir uma maior aproximação ao que será o nosso mundo real no mercado de trabalho.

NEO4J

Introdução

Nesta terceira fase do projeto, realizamos a transição do banco de dados MySQL para o NEO4J, utilizando o conjunto de dados Goodbooks para análise. O objetivo principal desta etapa é demonstrar a aplicação dos recursos oferecidos pelo NEO4J em um modelo de dados previamente baseado em SQL.

Durante o processo de migração, exploramos a modelagem de grafos utilizando a linguagem Cypher e importamos os dados a partir de arquivos CSV. Adaptamos o modelo de dados para tirar proveito das funcionalidades específicas do NEO4J.

O NEO4J é um banco de dados orientado a grafos projetado para lidar com dados altamente conectados, representados por nós e relacionamentos. Ao migrar do MySQL para o NEO4J, podemos aproveitar as vantagens dessa abordagem não relacional, permitindo a representação eficiente de dados complexos e seus relacionamentos.

Durante a transição do Goodbooks para o NEO4J, repensamos e redesenhamos o modelo de dados, explorando as funcionalidades exclusivas oferecidas pelo NEO4J. Utilizando a linguagem Cypher, executamos consultas e operações de manipulação de dados para explorar a capacidade do NEO4J em lidar com dados altamente conectados de forma eficiente.

Ao concluir esta fase do projeto, adquirimos habilidades práticas na modelagem de grafos com o NEO4J e na importação de dados por meio de arquivos CSV. Estamos preparados para aproveitar as vantagens do NEO4J como uma alternativa ao modelo relacional do MySQL, permitindo a representação e manipulação de dados em formato de grafo.

Com esta introdução, estamos prontos para prosseguir com a terceira e última fase do projeto, explorando as possibilidades do NEO4J na transição do Goodbooks do MySQL para um banco de dados orientado a grafos.

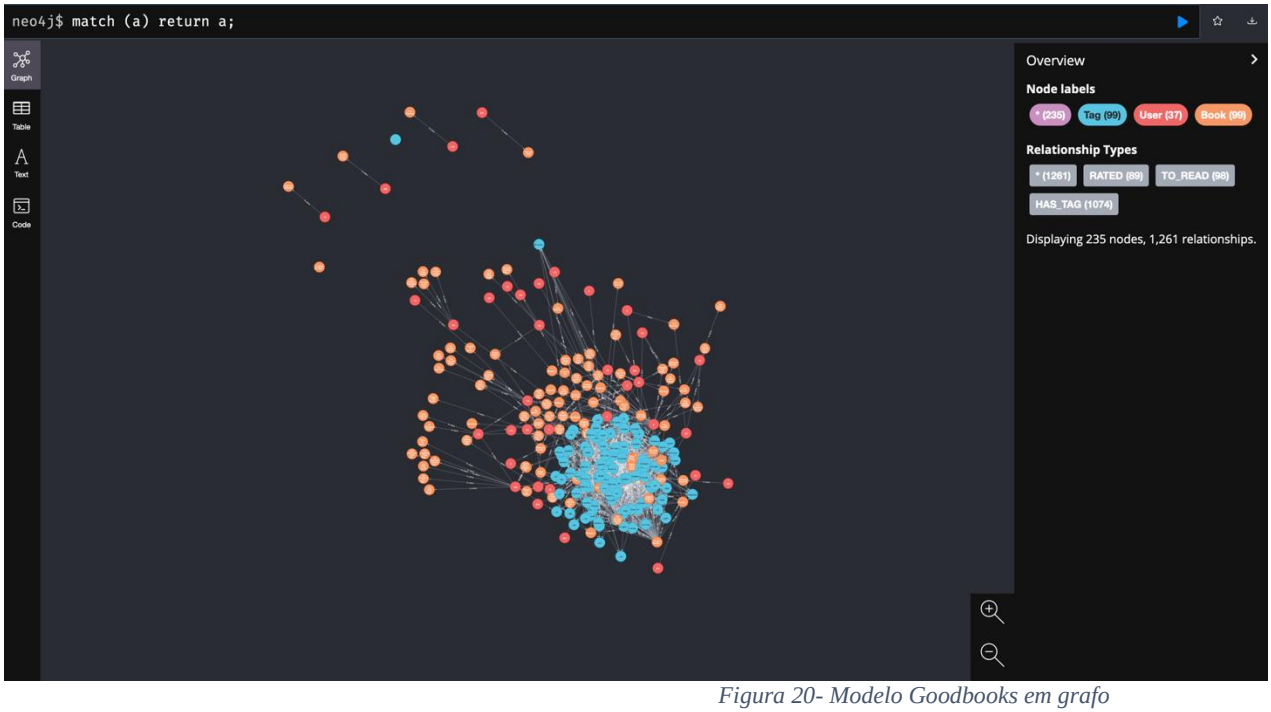


Figura 20- Modelo Goodbooks em grafo

Informação sobre o Dataset

O Dataset utilizado para NEO4J é exatamente o mesmo que foi necessário para a fase 1 (MySQL). Assim, sendo composto pelos mesmos 5 ficheiros csv, torna-se desnecessário voltar a repetir informação neste relatório.

Para aceder à informação do Dataset (5 ficheiros CSV) relativos ao GoodBooks,

Clique aqui

Importação das tabelas

Ao migrar do SQL para o Neo4j, adaptamos o processo de importação dos dados para funcionar com grafos. Utilizamos arquivos CSV para criar nós representando livros, etiquetas e utilizadores, estabelecendo conexões entre eles. O código a seguir cria as entidades "Book", "User" e "Tag", e estabelece os relacionamentos necessários, como "HAS_TAG" (possui etiqueta), "RATED" e "TO_READ ". Esse processo de importação e criação dos nós e relacionamentos permite uma estrutura mais adequada para representar os dados no Neo4j.

```
LOAD CSV WITH HEADERS FROM 'file:///books.csv' AS row
CREATE (b:Book {
  book_id: toInteger(row.book_id),
  goodreads_book_id: toInteger(row.goodreads_book_id),
  best_book_id: toInteger(row.best_book_id),
  work_id: toInteger(row.work_id),
  books_count: toInteger(row.books_count),
  isbn: row.isbn,
  isbn13: row.isbn13,
  authors: row.authors,
  original_publication_year: toInteger(row.original_publication_year),
  original_title: row.original_title,
  title: row.title,
  language_code: row.language_code,
  average_rating: toFloat(row.average_rating),
  ratings_count: toInteger(row.ratings_count),
  work_ratings_count: toInteger(row.work_ratings_count),
  work_text_reviews_count: toInteger(row.work_text_reviews_count),
  ratings_1: toInteger(row.ratings_1),
  ratings_2: toInteger(row.ratings_2),
  ratings_3: toInteger(row.ratings_3),
  ratings_4: toInteger(row.ratings_4),
  ratings_5: toInteger(row.ratings_5),
  image_url: row.image_url,
  small_image_url: row.small_image_url
});

LOAD CSV WITH HEADERS FROM 'file:///tags.csv' AS row
CREATE (t:Tag {tag_id: toInteger(row.tag_id), tag_name: row.tag_name});

LOAD CSV WITH HEADERS FROM 'file:///ratings.csv' AS row
MERGE (u:User {user_id: toInteger(row.user_id)});

LOAD CSV WITH HEADERS FROM 'file:///book_tags.csv' AS row
MATCH (b:Book {book_id: toInteger(row.book_id)}), (t:Tag {tag_id: toInteger(row.tag_id)})
CREATE (b)-[:HAS_TAG {count: toInteger(row.count)}]->(t);

LOAD CSV WITH HEADERS FROM 'file:///ratings.csv' AS row
MATCH (u:User {user_id: toInteger(row.user_id)}), (b:Book {book_id: toInteger(row.book_id)})
MERGE (u)-[r:RATED]->(b)
ON CREATE SET r.rating = toInteger(row.rating);

LOAD CSV WITH HEADERS FROM 'file:///to_read.csv' AS row
MERGE (u:User {user_id: toInteger(row.user_id)})
MERGE (b:Book {book_id: toInteger(row.book_id)})
MERGE (u)-[:TO_READ]->(b);
```

Figura 21- Código referente à importação e criação de entidades e relacionamentos

Modelos de dados

Modelo de grafos:

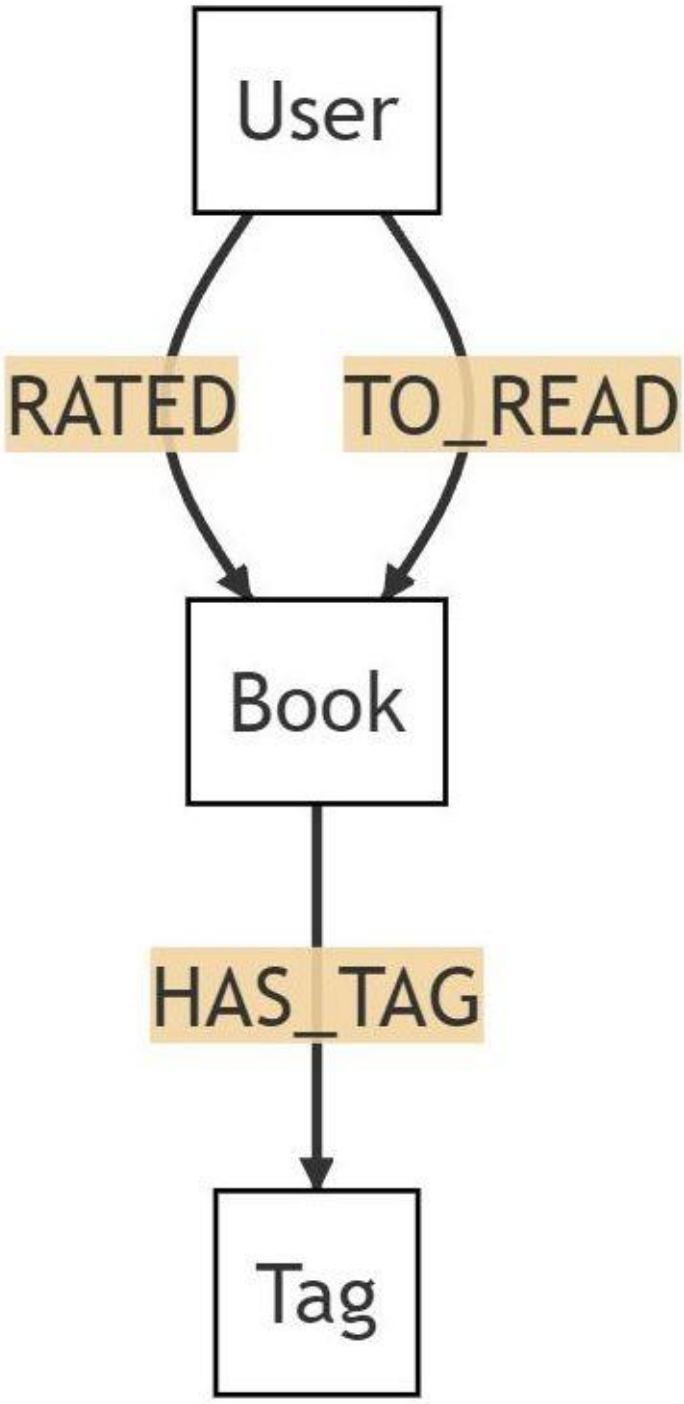


Diagrama 3- Modelo de Grafo

Dicionário de Dados

- Entidades

Nodo	Book		
Descrição	Cada nó contém informações detalhadas sobre um livro específico		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
book_id	código do livro	Integer	PK
goodreads_book_id	código do livro atribuido pela Good Reads	Integer	
best_book_id	código do melhor livro de um determinado trabalho literário	Integer	
work_id	código de cada trabalho literário	Integer	
books_count	Número de livros diferentes associados a um determinado trabalho literário	Integer	
isbn	Número de 10 dígitos que identifica cada livro	String	
isbn13	Número de 13 dígitos que identifica cada livro. Funciona de forma semelhante ao Isbn mas identifica também a língua ou o país.	String	
authors	Autor/es de cada livro	String	
original_publication_year	Ano de publicação original do livro	Integer	
original_tittle	Título original do livro	String	
tittle	Nome do livro	String	
language_code	Código do idioma em que o livro foi originalmente escrito	String	
average_rating	Classificação média do livro atribuída pelos utilizadores do Goodreads	Float	
ratings_count	Número total de avaliações para um determinado livro	Integer	
work_ratings_count	Número total de avaliações para uma determinada obra	Integer	
work_text_reviews_count	Número de comentários para uma determinada obra	Integer	
ratings_1	Quantidade de avaliações com 1 estrela	Integer	
ratings_2	Quantidade de avaliações com 2 estrela	Integer	
ratings_3	Quantidade de avaliações com 3 estrela	Integer	
ratings_4	Quantidade de avaliações com 4 estrela	Integer	
ratings_5	Quantidade de avaliações com 5 estrela	Integer	
image_url	URL da imagem da capa do livro	String	
small_image_url	URL da imagem da capa do livro em formato pequeno	String	

Nodo	User		
Descrição	Representa um utilizador na base de dados.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
user_id	Identificador único do utilizador	Integer	PK

Nodo	Tag		
Descrição	Representa uma tag associada a um livro.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
tag_id	código identificador de cada tag	Integer	PK
tag_name	nome da tag respetiva	String	

- Relacionamentos

Relação	Rated		
Descrição	Define a relação entre um utilizador e um livro quando o utilizador atribui uma avaliação ao livro.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
rating	código do utilizador	Integer	-

Relação	Has_Tag		
Descrição	Define a relação entre um livro e uma tag.		
Observações			
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
count	Número de utilizadores que atribuíram a tag ao livro.	Integer	-

Relação	To_read		
Descrição	Define a relação entre um utilizador e um livro quando o utilizador adiciona o livro à lista de leitura.		
Observações	Esta relação indica que o utilizador pretende ler o livro. Não possui propriedades adicionais.		
Campos			
Nome	Descrição	Tipos de dados	Restrições de Domínio
-	-	-	-

Questões analíticas

Pergunta 1 - Listar os livros por ordem crescente pelo respetivo ano de publicação

```
MATCH (b:Book)
RETURN b
ORDER BY b.original_publication_year ASC;
```

Resultado da Query:

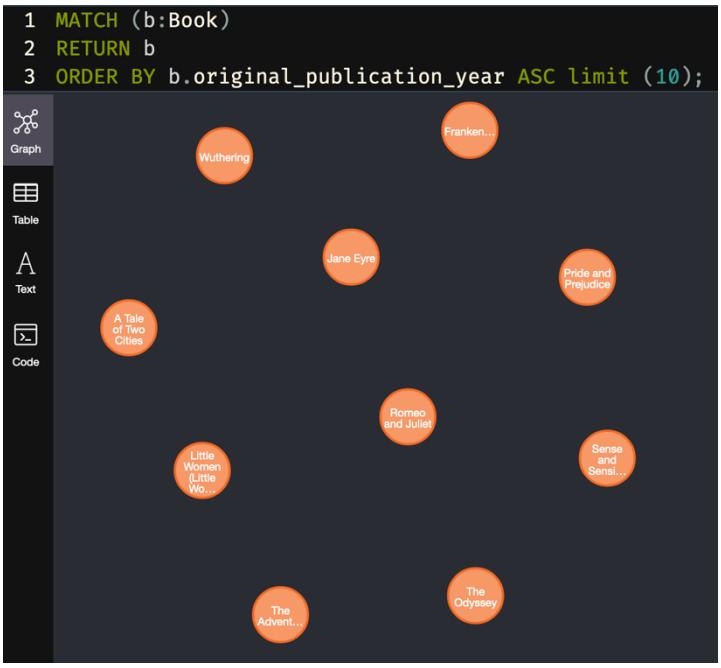


Figura 22- Resposta com limite de 10 à questão 1

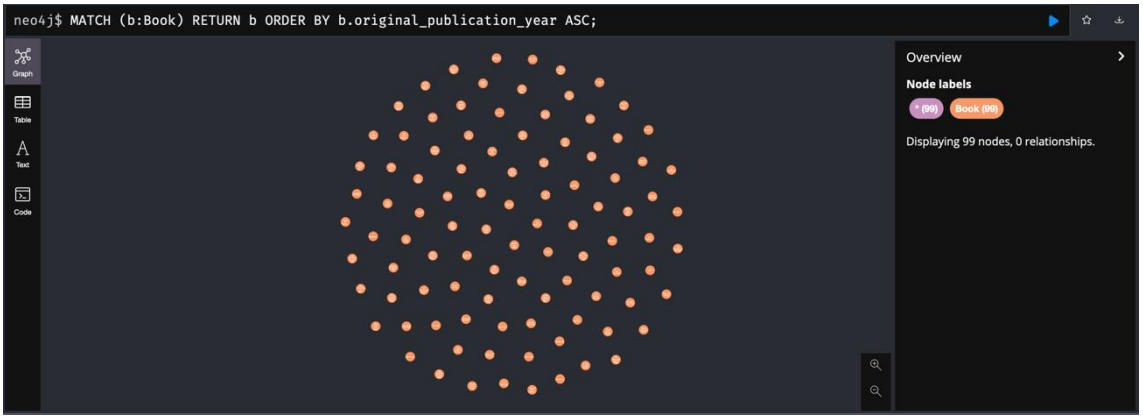


Figura 23- Resposta sem limite à questão 1

Pergunta 2 - Para cada livro com classificação inferior a 4, listar o título que lhe corresponde

```
MATCH (b:Book)<-[:RATED]-(u:User)
WHERE u.rating < 4
RETURN b.title;
```

Resultado da Query:

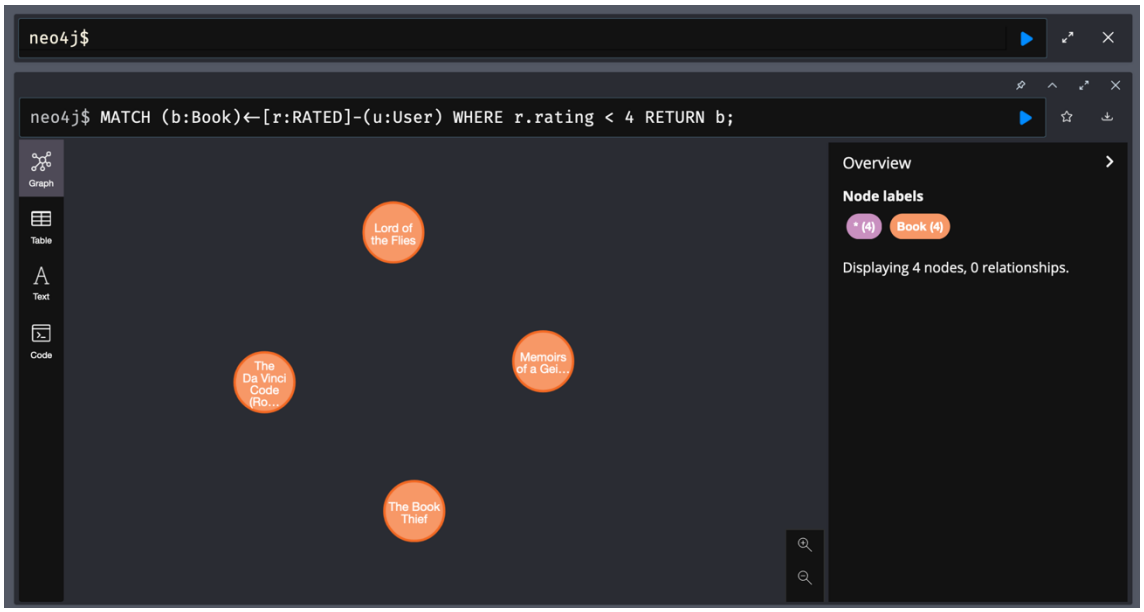


Figura 24- Resposta à questão 2

Pergunta 3 - Quantas avaliações foram feitas pelo utilizador user_id = 8?

```
MATCH (u:User {user_id: 8})-[r:RATED]->(b:Book)
WHERE r.rating > 0
RETURN COUNT(b);
```

Resultado da Query:

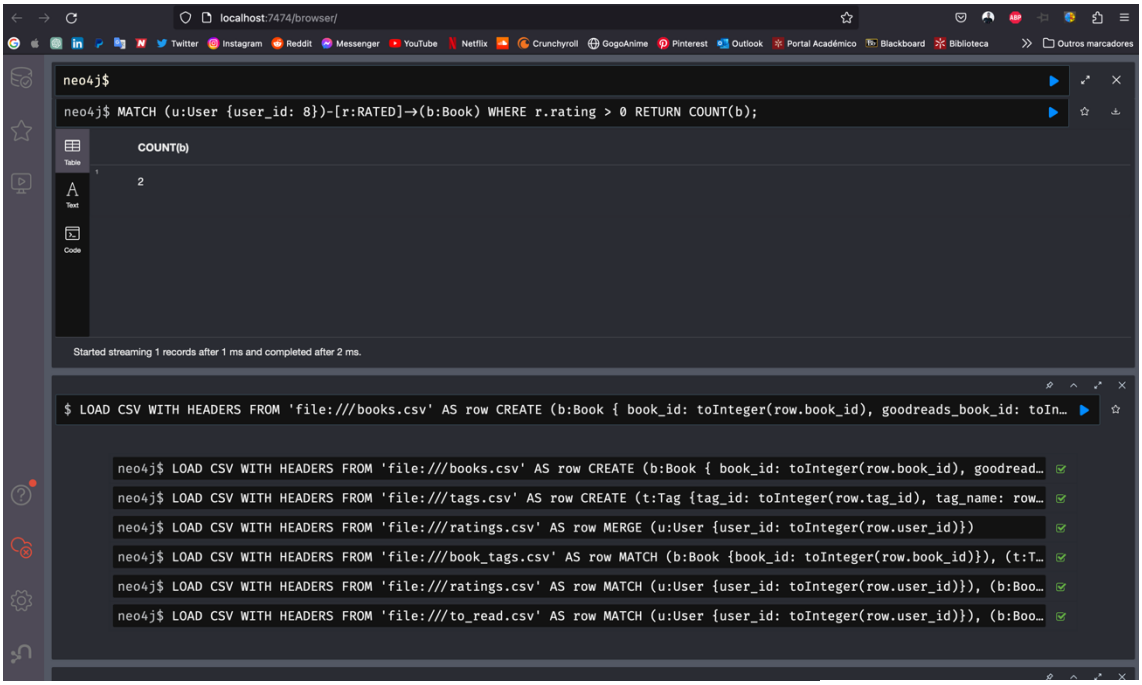


Figura 25- Resposta à questão 3

Pergunta 4 - Listar os vários livros por ordem alfabética do nome do respetivo autor.

```
MATCH (b:Book)
RETURN b
ORDER BY b.authors;
```

Resultado da Query:

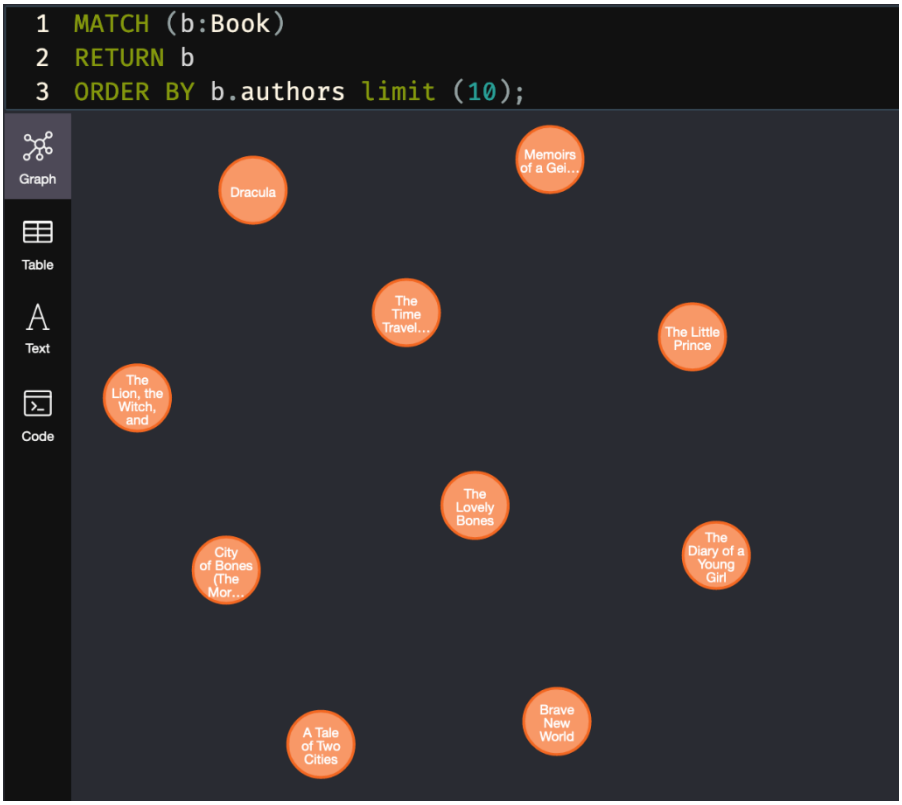


Figura 26- Resposta sem limite à questão 4

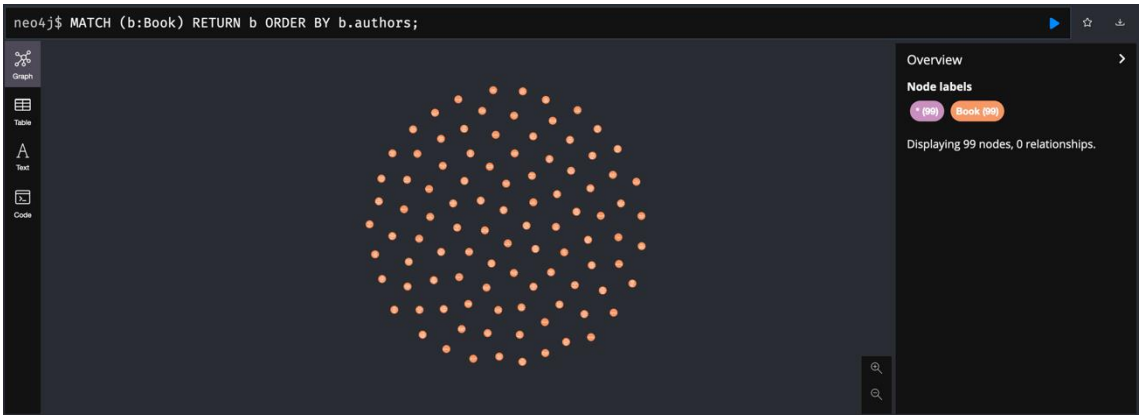


Figura 27- Resposta sem limite à questão 4

Pergunta 5- Book_id, obra original e avaliação média de todos os livros com classificação superior a 4.5 estrelas.

```
MATCH (b:Book)
WHERE b.average_rating > 4.5
RETURN b.book_id, b.original_title, b.average_rating
```

Resultado da Query:

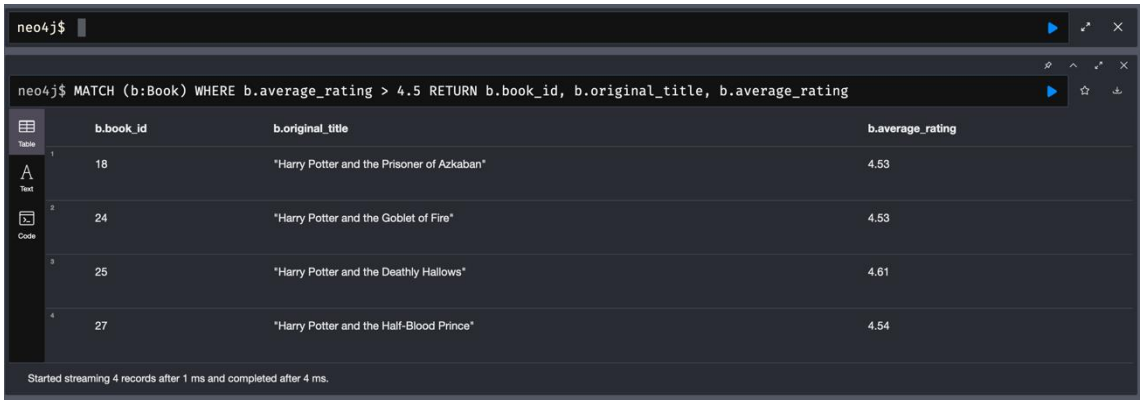


Figura 28- Resposta à questão 5

Pergunta 6- Mostrar autores e o livro original que possuam mais de 4 milhões de avaliações.

```
MATCH (b:Book)
WHERE b.ratings_count > 4000000
RETURN b.authors, b.original_title
```

Resultado da Query:

Figura 29- Resposta à questão 6

Conclusão

A transição do conjunto de dados Goodbooks do MySQL para o NEO4J representou uma etapa fundamental em nosso projeto, permitindo-nos explorar uma nova tecnologia de banco de dados e aprofundar nosso entendimento sobre suas características e funcionalidades. Ao longo desse processo, adquirimos conhecimentos valiosos em modelagem de grafos utilizando a linguagem Cypher e importação de dados através de arquivos CSV. Além disso, tivemos a oportunidade de explorar as vantagens do NEO4J em relação a consultas, armazenamento e relacionamento de dados.

Durante nossa jornada com o NEO4J, ficou evidente que essa ferramenta possui uma abordagem poderosa e eficiente para análise de dados complexos. A representação dos dados em formato de grafo proporcionou uma perspetiva única e valiosa para compreender as relações entre os elementos do conjunto de dados Goodbooks. As consultas em Cypher nos permitiram explorar essas conexões de forma intuitiva e eficaz, oferecendo uma visão mais abrangente e compreensiva dos dados.

Além dos benefícios específicos relacionados à análise de dados complexos, a migração para o NEO4J também ampliou nosso conhecimento e habilidades em relação às ferramentas de gerenciamento de dados. Aprendemos a adaptar um modelo de dados previamente baseado em SQL para uma estrutura de grafos, entendendo as particularidades e as considerações específicas dessa abordagem.

Diante desses resultados, estamos confiantes de que estamos preparados para enfrentar futuros desafios relacionados a projetos de bases de dados. A experiência adquirida ao longo dessa fase do projeto nos proporcionou uma base sólida de conhecimento e habilidades práticas para aplicar em futuros empreendimentos nessa área.

Em suma, a transição do Goodbooks para o NEO4J representou uma etapa enriquecedora em nosso projeto, que nos permitiu expandir nosso conhecimento, explorar uma nova tecnologia e adquirir habilidades relevantes para nossa formação académica e profissional em bases de dados.

Conclusão e comparações finais

Durante o desenvolvimento do projeto, exploramos três tecnologias de gerenciamento de bancos de dados: MySQL, Cassandra e NEO4J. Cada uma dessas tecnologias possui características e funcionalidades distintas, adequadas para diferentes cenários e requisitos.

No MySQL, uma tecnologia de banco de dados relacional, utilizamos a linguagem SQL para modelar e manipular os dados do conjunto Goodbooks. O MySQL é amplamente utilizado, conhecido por sua estabilidade, escalabilidade e confiabilidade. Com sua estrutura de tabelas e relacionamentos, o MySQL é adequado para lidar com dados estruturados e realizar consultas complexas, utilizando recursos como junções, filtragens e ordenações. Das três tecnologias BD presentes neste relatório, esta foi a que o grupo se sentiu mais confortável em utilizar e que requereu menos tempo de aprendizagem, visto que estávamos mais familiarizados. Porém, nesta fase do projeto, foi uma adaptação ao ambiente e à forma como funcionavam os programas que, para nós, eram algo completamente novo.

Já o Cassandra é uma tecnologia de banco de dados distribuído, altamente escalável e projetado para lidar com grandes volumes de dados. Diferente do modelo relacional do MySQL, o Cassandra adota o modelo de armazenamento de pares de chave-valor e não possui suporte a operações de junção e consultas complexas. Em vez disso, o Cassandra é otimizado para consultas rápidas e eficientes, sendo ideal para cenários que exigem alta disponibilidade e escalabilidade horizontal. Sendo CQL (Cassandra Query Language) uma linguagem bem semelhante a MySQL, o processo de codificação pareceu bastante familiar, não tendo requerido grande tempo para aprender. No entanto, devido às limitações do Cassandra em relação a determinados comandos específicos, como o "distinct", foi necessário criar uma tabela separada para cada consulta, o que tornou o processo de consulta mais complexo, na nossa opinião.

Por fim, o NEO4J é uma tecnologia de banco de dados baseada em grafos, que permite modelar e explorar relações complexas entre os dados. Com a linguagem Cypher, utilizamos a abordagem de grafos para representar os relacionamentos entre entidades do conjunto Goodbooks. O NEO4J permite consultas eficientes e profundas explorações dos dados, fornecendo uma perspetiva única sobre as relações presentes no conjunto de dados. É também, a nosso ver, o modelo mais fácil de modelar, visto que se assemelha muito a “atores” e “ações” no pensamento e criação de entidades e relacionamentos. Porém, a sua linguagem é bem diferente de MySQL e CQL, tendo sido necessário um tempo de aprendizagem bem maior para modelar o grafo e criar as consultas em código.

Assim, cada uma das tecnologias utilizadas no projeto possui características distintas que as tornam adequadas para diferentes necessidades. O MySQL é uma escolha sólida para dados estruturados e consultas complexas, o Cassandra destaca-se em cenários de alta disponibilidade e escalabilidade, enquanto o NEO4J é ideal para análises de redes e exploração de relacionamentos complexos. Através desse projeto, pudemos compreender as diferenças entre essas tecnologias e utilizar suas capacidades de forma efetiva, ampliando nosso conhecimento e habilidades em gerenciamento de bancos de dados.

Avaliação Individual da Equipa

Entrega 1 (MySQL)

Pedro Pires [a95549]: N+2
Liandro Cruz [a100436]: N+1
João Ferreira [a105100]: N-1
Gonçalo Cruz [a100639]: N-1
Tomás Marques [a100436]: N-1

Entrega 2 (Cassandra)

Pedro Pires [a95549]: N+3
Liandro Cruz [a100436]: N
João Ferreira [a105100]: N-1
Gonçalo Cruz [a100639]: N-1
Tomás Marques [a100436]: N-1

Entrega 3 (NEO4J)

Pedro Pires [a95549]: N+2
Liandro Cruz [a100436]: N+1
João Ferreira [a105100]: N-1
Gonçalo Cruz [a100639]: N-1
Tomás Marques [a100436]: N-1

Links

Link para download do Dataset: https://github.com/zygmuntz/goodbooks-10k

Link do GoodReads: https://www.goodreads.com/

Link da documentação do NEO4J: https://neo4j.com/docs/cypher-manual/current/clauses/

Nota: O dataset foi enviado junto com o ficheiro deste relatório, uma vez que teve de ser alterado devido a erros no próprio dataset.