Bookstore Database Design

Part 2

Manisha Goyal (mg7609)

Rhea Chandok (rc5397)

Sanam Palsule (sp7940)

Table of Contents

Data Lake Integration and Dataset Compilation

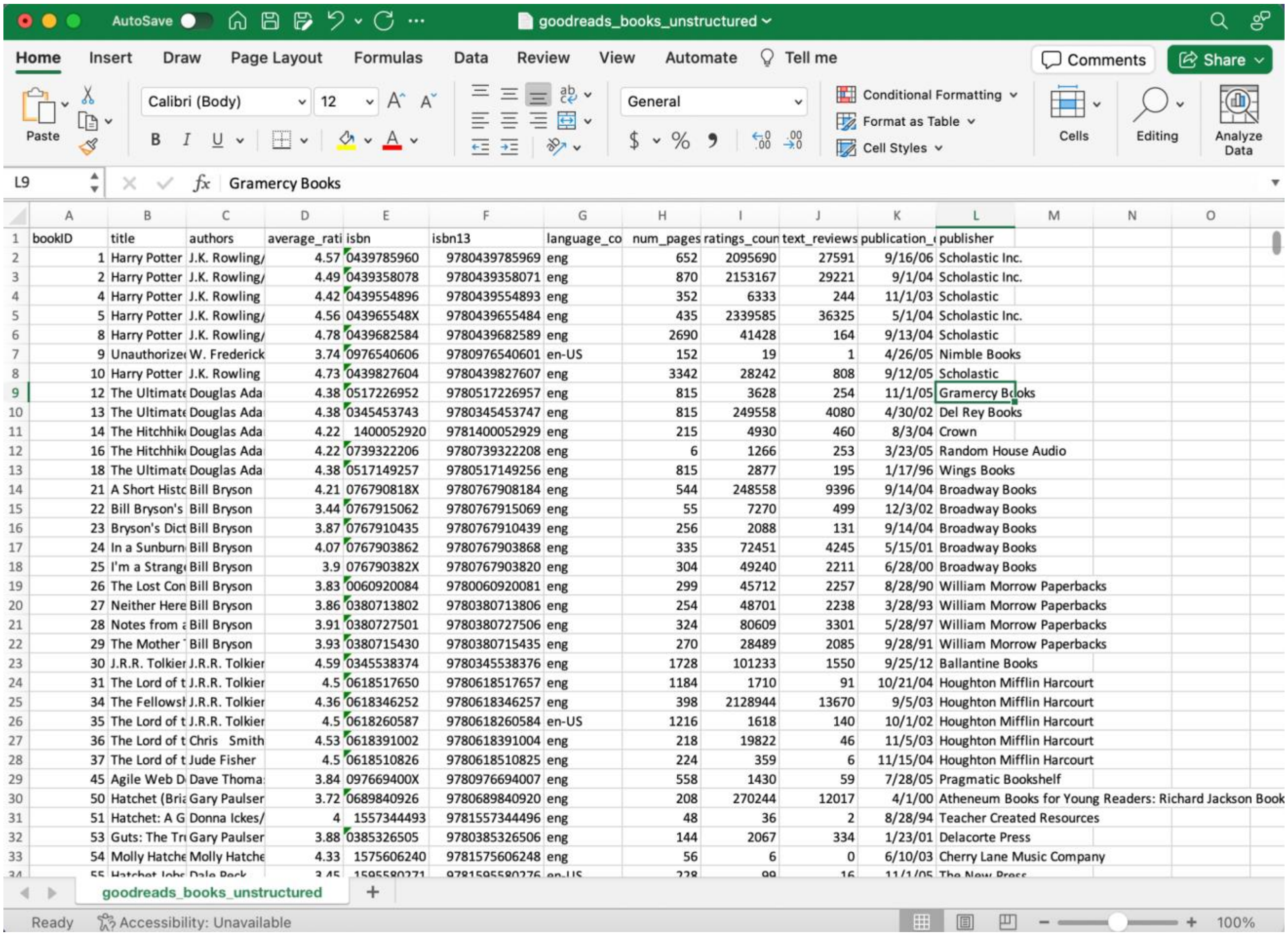
A data lake is an integral part of modern data architectures, especially when dealing with vast and varied datasets. In the context of our bookstore's EDA, the data lake serves as a centralized repository where all types of data—structured, semi-structured, and unstructured—are stored in their raw form. The flexibility of a data lake allows for the storage and analysis of large volumes of unstructured data, which in our case includes Good Reads and Amazon book reviews.

Composition of the Data Lake

Unstructured Datasets

1. Good Reads Book Reviews:

This dataset comprises extensive user-generated reviews and ratings from the Good Reads platform. It offers a wealth of qualitative data, such as reader sentiment and engagement, which are not readily available in structured datasets.

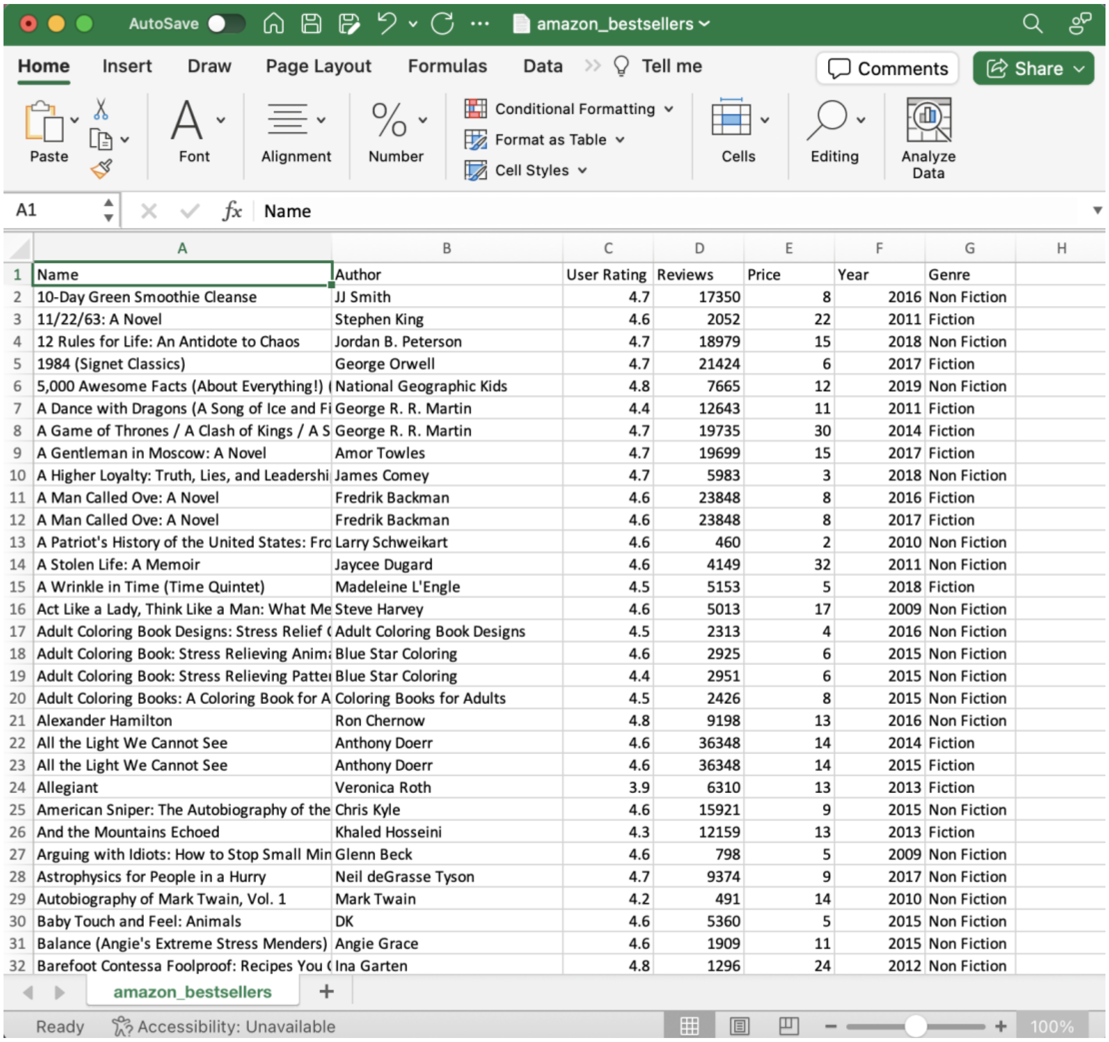


L9 fx Gramercy Books

	A	B	C	D	E	F	G	H	I	J	K	L
1	bookID	title	authors	average_rati	isbn	isbn13	language_co	num_pages	ratings_coun	text_reviews	publication_	publisher
2	1	Harry Potter	J.K. Rowling/	4.57	0439785960	9780439785969	eng	652	2095690	27591	9/16/06	Scholastic Inc.
3	2	Harry Potter	J.K. Rowling/	4.49	0439358078	9780439358071	eng	870	2153167	29221	9/1/04	Scholastic Inc.
4	4	Harry Potter	J.K. Rowling	4.42	0439554896	9780439554893	eng	352	6333	244	11/1/03	Scholastic
5	5	Harry Potter	J.K. Rowling/	4.56	043965548X	9780439655484	eng	435	2339585	36325	5/1/04	Scholastic Inc.
6	8	Harry Potter	J.K. Rowling/	4.78	0439682584	9780439682589	eng	2690	41428	164	9/13/04	Scholastic
7	9	Unauthorize	W. Frederick	3.74	0976540606	9780976540601	en-US	152	19	1	4/26/05	Nimble Books
8	10	Harry Potter	J.K. Rowling	4.73	0439827604	9780439827607	eng	3342	28242	808	9/12/05	Scholastic
9	12	The Ultimate	Douglas Ada	4.38	0517226952	9780517226957	eng	815	3628	254	11/1/05	Gramercy Books
10	13	The Ultimate	Douglas Ada	4.38	0345453743	9780345453747	eng	815	249558	4080	4/30/02	Del Rey Books
11	14	The Hitchhik	Douglas Ada	4.22	1400052920	9781400052929	eng	215	4930	460	8/3/04	Crown
12	16	The Hitchhik	Douglas Ada	4.22	0739322206	9780739322208	eng	6	1266	253	3/23/05	Random House Audio
13	18	The Ultimate	Douglas Ada	4.38	0517149257	9780517149256	eng	815	2877	195	1/17/96	Wings Books
14	21	A Short Histc	Bill Bryson	4.21	076790818X	9780767908184	eng	544	248558	9396	9/14/04	Broadway Books
15	22	Bill Bryson's	Bill Bryson	3.44	0767915062	9780767915069	eng	55	7270	499	12/3/02	Broadway Books
16	23	Bryson's Dict	Bill Bryson	3.87	0767910435	9780767910439	eng	256	2088	131	9/14/04	Broadway Books
17	24	In a Sunburn	Bill Bryson	4.07	0767903862	9780767903868	eng	335	72451	4245	5/15/01	Broadway Books
18	25	I'm a Strang	Bill Bryson	3.9	076790382X	9780767903820	eng	304	49240	2211	6/28/00	Broadway Books
19	26	The Lost Con	Bill Bryson	3.83	0060920084	9780060920081	eng	299	45712	2257	8/28/90	William Morrow Paperbacks
20	27	Neither Here	Bill Bryson	3.86	0380713802	9780380713806	eng	254	48701	2238	3/28/93	William Morrow Paperbacks
21	28	Notes from a	Bill Bryson	3.91	0380727501	9780380727506	eng	324	80609	3301	5/28/97	William Morrow Paperbacks
22	29	The Mother	Bill Bryson	3.93	0380715430	9780380715435	eng	270	28489	2085	9/28/91	William Morrow Paperbacks
23	30	J.R.R. Tolkier	J.R.R. Tolkier	4.59	0345538374	9780345538376	eng	1728	101233	1550	9/25/12	Ballantine Books
24	31	The Lord of t	J.R.R. Tolkier	4.5	0618517650	9780618517657	eng	1184	1710	91	10/21/04	Houghton Mifflin Harcourt
25	34	The Fellowsh	J.R.R. Tolkier	4.36	0618346252	9780618346257	eng	398	2128944	13670	9/5/03	Houghton Mifflin Harcourt
26	35	The Lord of t	J.R.R. Tolkier	4.5	0618260587	9780618260584	en-US	1216	1618	140	10/1/02	Houghton Mifflin Harcourt
27	36	The Lord of t	Chris Smith	4.53	0618391002	9780618391004	eng	218	19822	46	11/5/03	Houghton Mifflin Harcourt
28	37	The Lord of t	Jude Fisher	4.5	0618510826	9780618510825	eng	224	359	6	11/15/04	Houghton Mifflin Harcourt
29	45	Agile Web D	Dave Thoma	3.84	097669400X	9780976694007	eng	558	1430	59	7/28/05	Pragmatic Bookshelf
30	50	Hatchet (Bri	Gary Paulser	3.72	0689840926	9780689840920	eng	208	270244	12017	4/1/00	Atheneum Books for Young Readers: Richard Jackson Book
31	51	Hatchet: A G	Donna Ickes/	4	1557344493	9781557344496	eng	48	36	2	8/28/94	Teacher Created Resources
32	53	Guts: The Tr	Gary Paulser	3.88	0385326505	9780385326506	eng	144	2067	334	1/23/01	Delacorte Press
33	54	Molly Hatche	Molly Hatche	4.33	1575606240	9781575606248	eng	56	6	0	6/10/03	Cherry Lane Music Company
34	55	Hatchet Johs	Dale Peck	3.45	1595580271	9781595580276	en-US	228	99	16	11/1/05	The New Press

2. Amazon Book Reviews:

Like the Good Reads data, this dataset contains customer reviews and ratings from Amazon. It provides a market perspective in comparison to the Good Reads data, encompassing a wide range of customer demographics and preferences.

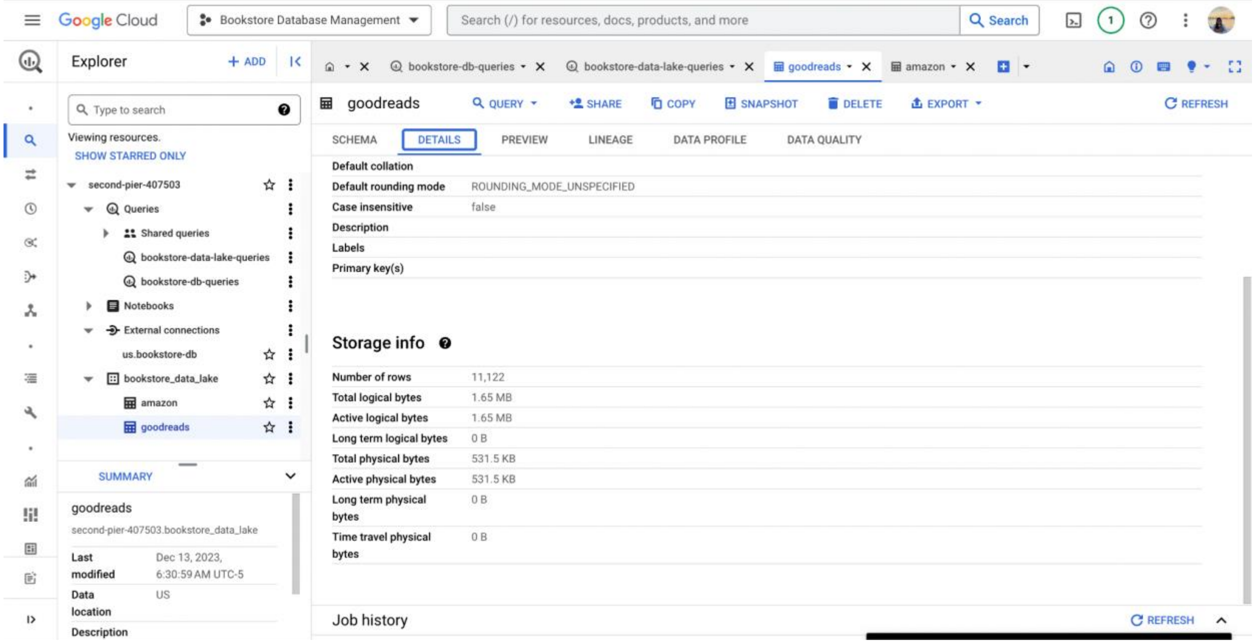


	A	B	C	D	E	F	G	H
1	Name	Author	User Rating	Reviews	Price	Year	Genre	
2	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction	
3	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction	
4	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction	
5	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction	
6	5,000 Awesome Facts (About Everything!) (	National Geographic Kids	4.8	7665	12	2019	Non Fiction	
7	A Dance with Dragons (A Song of Ice and Fi	George R. R. Martin	4.4	12643	11	2011	Fiction	
8	A Game of Thrones / A Clash of Kings / A S	George R. R. Martin	4.7	19735	30	2014	Fiction	
9	A Gentleman in Moscow: A Novel	Amor Towles	4.7	19699	15	2017	Fiction	
10	A Higher Loyalty: Truth, Lies, and Leadershi	James Comey	4.7	5983	3	2018	Non Fiction	
11	A Man Called Ove: A Novel	Fredrik Backman	4.6	23848	8	2016	Fiction	
12	A Man Called Ove: A Novel	Fredrik Backman	4.6	23848	8	2017	Fiction	
13	A Patriot's History of the United States: Fro	Larry Schweikart	4.6	460	2	2010	Non Fiction	
14	A Stolen Life: A Memoir	Jaycee Dugard	4.6	4149	32	2011	Non Fiction	
15	A Wrinkle in Time (Time Quintet)	Madeleine L'Engle	4.5	5153	5	2018	Fiction	
16	Act Like a Lady, Think Like a Man: What Me	Steve Harvey	4.6	5013	17	2009	Non Fiction	
17	Adult Coloring Book Designs: Stress Relief (	Adult Coloring Book Designs	4.5	2313	4	2016	Non Fiction	
18	Adult Coloring Book: Stress Relieving Anim:	Blue Star Coloring	4.6	2925	6	2015	Non Fiction	
19	Adult Coloring Book: Stress Relieving Patte	Blue Star Coloring	4.4	2951	6	2015	Non Fiction	
20	Adult Coloring Books: A Coloring Book for A	Coloring Books for Adults	4.5	2426	8	2015	Non Fiction	
21	Alexander Hamilton	Ron Chernow	4.8	9198	13	2016	Non Fiction	
22	All the Light We Cannot See	Anthony Doerr	4.6	36348	14	2014	Fiction	
23	All the Light We Cannot See	Anthony Doerr	4.6	36348	14	2015	Fiction	
24	Allegiant	Veronica Roth	3.9	6310	13	2013	Fiction	
25	American Sniper: The Autobiography of the	Chris Kyle	4.6	15921	9	2015	Non Fiction	
26	And the Mountains Echoed	Khaled Hosseini	4.3	12159	13	2013	Fiction	
27	Arguing with Idiots: How to Stop Small Min	Glenn Beck	4.6	798	5	2009	Non Fiction	
28	Astrophysics for People in a Hurry	Neil deGrasse Tyson	4.7	9374	9	2017	Non Fiction	
29	Autobiography of Mark Twain, Vol. 1	Mark Twain	4.2	491	14	2010	Non Fiction	
30	Baby Touch and Feel: Animals	DK	4.6	5360	5	2015	Non Fiction	
31	Balance (Angie's Extreme Stress Menders)	Angie Grace	4.6	1909	11	2015	Non Fiction	
32	Barefoot Contessa Foolproof: Recipes You (	Ina Garten	4.8	1296	24	2012	Non Fiction	

Processing and Analysis of Unstructured Data

Data Ingestion

The first step involves ingesting the raw unstructured data from Good Reads and Amazon into the data lake. This process is designed to handle high volumes of data efficiently while preserving its original form for accurate analysis.



goodreads

second-pier-407503.bookstore_data_lake

Last modified: Dec 13, 2023, 6:30:59 AM UTC-5

Data location: US

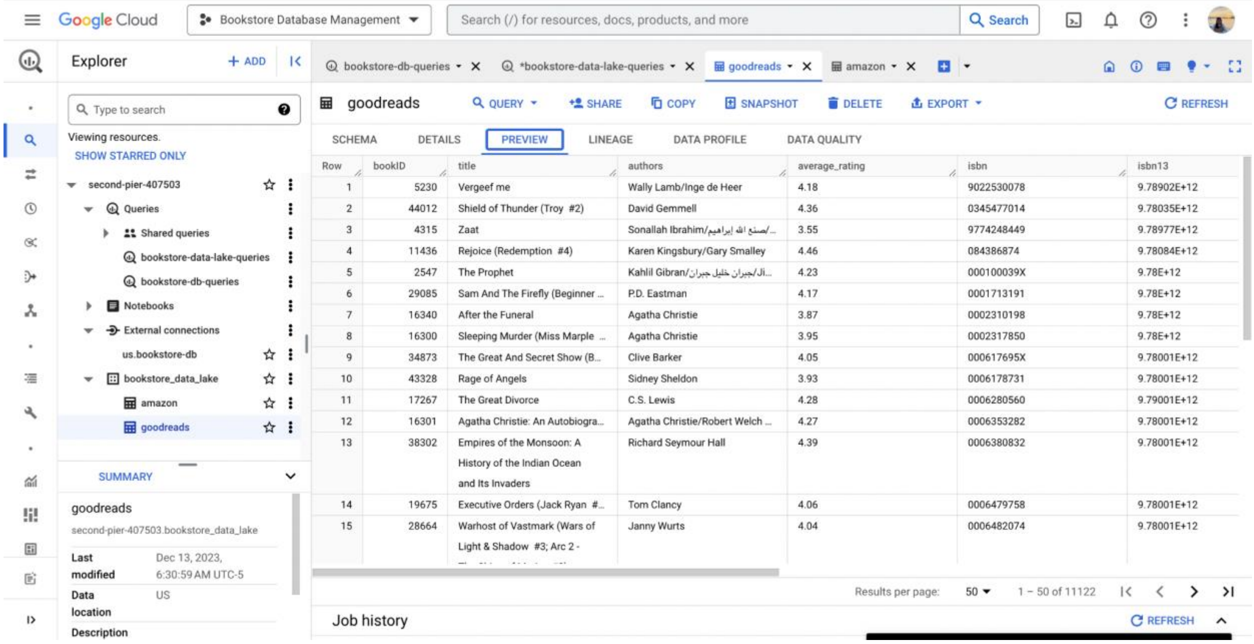
Description

Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	
Labels	
Primary key(s)	

Storage info

Number of rows	11,122
Total logical bytes	1.65 MB
Active logical bytes	1.65 MB
Long term logical bytes	0 B
Total physical bytes	531.5 KB
Active physical bytes	531.5 KB
Long term physical bytes	0 B
Time travel physical bytes	0 B

Job history



goodreads

second-pier-407503.bookstore_data_lake

Last modified: Dec 13, 2023, 6:30:59 AM UTC-5

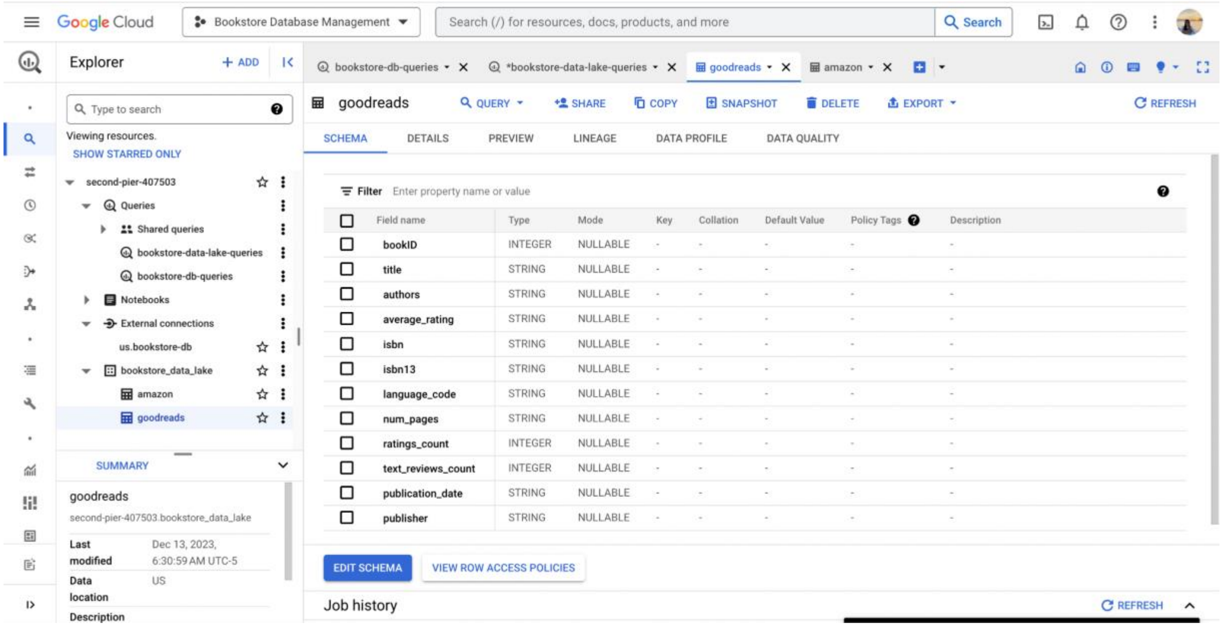
Data location: US

Description

Row	bookID	title	authors	average_rating	isbn	isbn13
1	5230	Vergeef me	Wally Lamb/Inge de Heer	4.18	9022530078	9.78902E+12
2	44012	Shield of Thunder (Troy #2)	David Gemmell	4.36	0345477014	9.78035E+12
3	4315	Zaat	Sonallah Ibrahim/صنع الله إبراهيم/...	3.55	9774248449	9.78977E+12
4	11436	Rejoice (Redemption #4)	Karen Kingsbury/Gary Smalley	4.46	084386874	9.78084E+12
5	2547	The Prophet	Kahlil Gibran/جبران خليل جبران/Ji...	4.23	000100039X	9.78E+12
6	29085	Sam And The Firefly (Beginner ...	P.D. Eastman	4.17	0001713191	9.78E+12
7	16340	After the Funeral	Agatha Christie	3.87	0002310198	9.78E+12
8	16300	Sleeping Murder (Miss Marple ...	Agatha Christie	3.95	0002317850	9.78E+12
9	34873	The Great And Secret Show (B...	Clive Barker	4.05	000617695X	9.78001E+12
10	43328	Rage of Angels	Sidney Sheldon	3.93	0006178731	9.78001E+12
11	17267	The Great Divorce	C.S. Lewis	4.28	0006280560	9.79001E+12
12	16301	Agatha Christie: An Autobiogra...	Agatha Christie/Robert Welch ...	4.27	0006353282	9.78001E+12
13	38302	Empires of the Monsoon: A History of the Indian Ocean and Its Invaders	Richard Seymour Hall	4.39	0006380832	9.78001E+12
14	19675	Executive Orders (Jack Ryan #...	Tom Clancy	4.06	0006479758	9.78001E+12
15	28664	Warhost of Vastmark (Wars of Light & Shadow #3; Arc 2 -	Janny Wurts	4.04	0006482074	9.78001E+12

Results per page: 50 1 – 50 of 11122

Job history

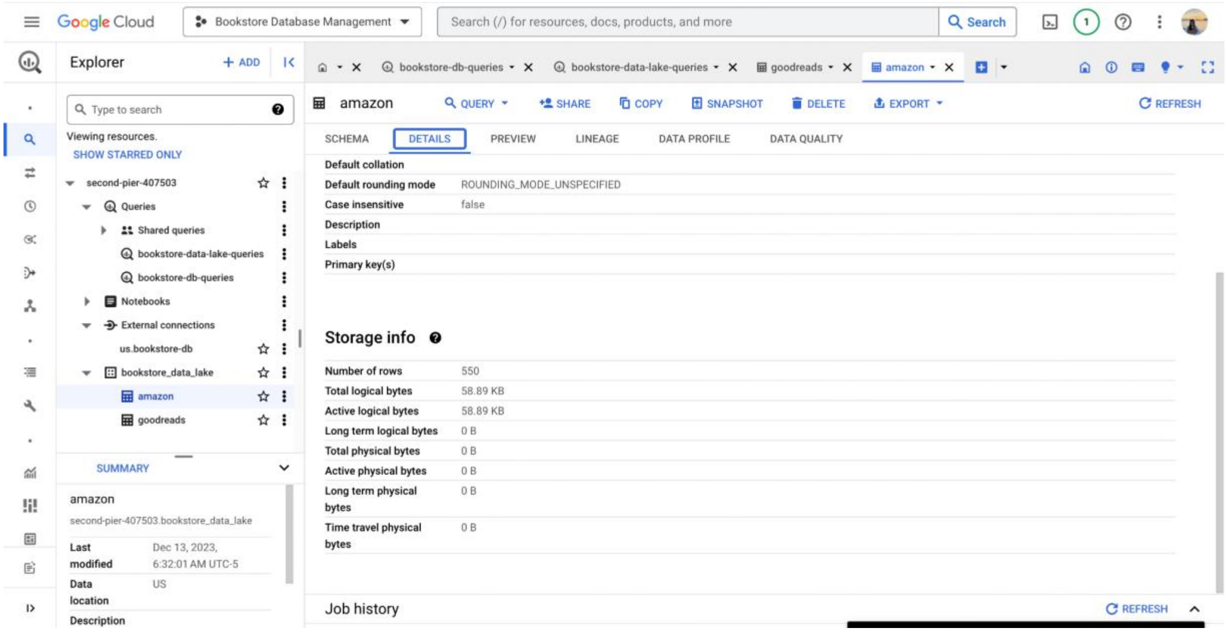


goodreads

	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
☐	bookID	INTEGER	NULLABLE	-	-	-	-	-
☐	title	STRING	NULLABLE	-	-	-	-	-
☐	authors	STRING	NULLABLE	-	-	-	-	-
☐	average_rating	STRING	NULLABLE	-	-	-	-	-
☐	isbn	STRING	NULLABLE	-	-	-	-	-
☐	isbn13	STRING	NULLABLE	-	-	-	-	-
☐	language_code	STRING	NULLABLE	-	-	-	-	-
☐	num_pages	STRING	NULLABLE	-	-	-	-	-
☐	ratings_count	INTEGER	NULLABLE	-	-	-	-	-
☐	text_reviews_count	INTEGER	NULLABLE	-	-	-	-	-
☐	publication_date	STRING	NULLABLE	-	-	-	-	-
☐	publisher	STRING	NULLABLE	-	-	-	-	-

goodreads
second-pier-407503.bookstore_data_lake

Last modified	Dec 13, 2023, 6:30:59 AM UTC-5
Data location	US
Description	



amazon

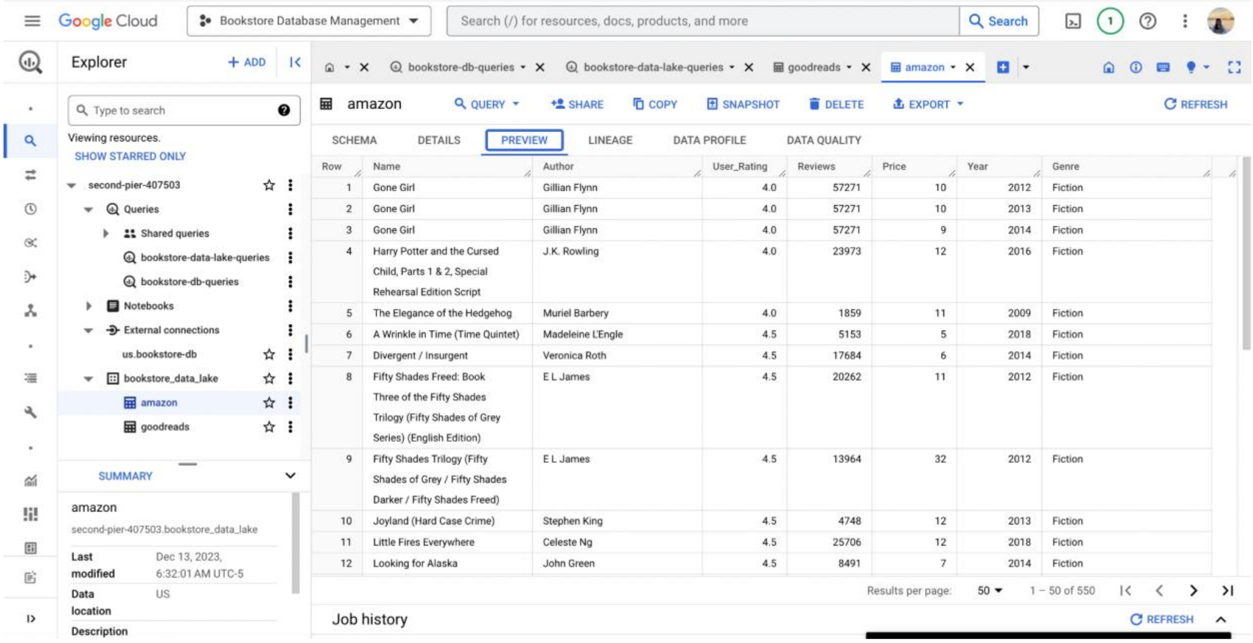
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	
Labels	
Primary key(s)	

Storage info

Number of rows	550
Total logical bytes	58.89 KB
Active logical bytes	58.89 KB
Long term logical bytes	0 B
Total physical bytes	0 B
Active physical bytes	0 B
Long term physical bytes	0 B
Time travel physical bytes	0 B

amazon
second-pier-407503.bookstore_data_lake

Last modified	Dec 13, 2023, 6:32:01 AM UTC-5
Data location	US
Description	



Type to search

Viewing resources.

SHOW STARRED ONLY

- second-pier-407503
  - Queries
    - Shared queries
    - bookstore-data-lake-queries
    - bookstore-db-queries
  - Notebooks
  - External connections
    - us.bookstore-db
  - bookstore_data_lake
    - amazon
    - goodreads

SUMMARY

amazon

second-pier-407503.bookstore_data_lake

Last modified	Dec 13, 2023, 6:32:01 AM UTC-5
Data location	US
Description	

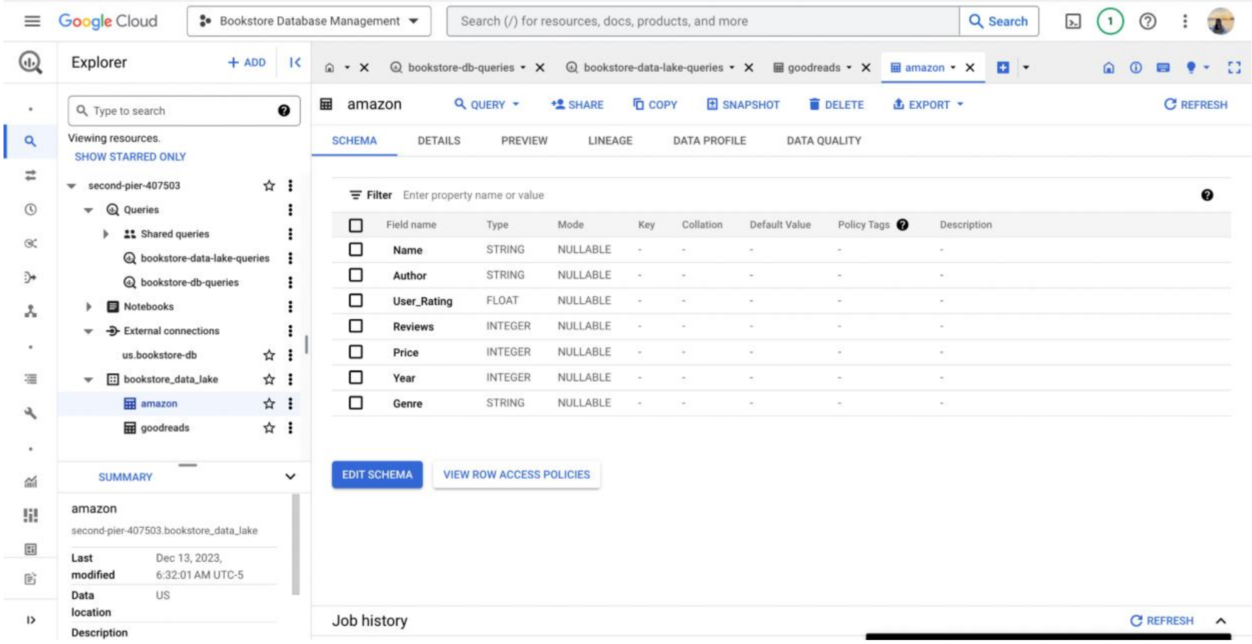
amazon

Row	Name	Author	User_Rating	Reviews	Price	Year	Genre
1	Gone Girl	Gillian Flynn	4.0	57271	10	2012	Fiction
2	Gone Girl	Gillian Flynn	4.0	57271	10	2013	Fiction
3	Gone Girl	Gillian Flynn	4.0	57271	9	2014	Fiction
4	Harry Potter and the Cursed Child, Parts 1 & 2, Special Rehearsal Edition Script	J.K. Rowling	4.0	23973	12	2016	Fiction
5	The Elegance of the Hedgehog	Muriel Barbery	4.0	1859	11	2009	Fiction
6	A Wrinkle in Time (Time Quintet)	Madeleine L'Engle	4.5	5153	5	2018	Fiction
7	Divergent / Insurgent	Veronica Roth	4.5	17684	6	2014	Fiction
8	Fifty Shades Freed: Book Three of the Fifty Shades Trilogy (Fifty Shades of Grey Series) (English Edition)	E L James	4.5	20262	11	2012	Fiction
9	Fifty Shades Trilogy (Fifty Shades of Grey / Fifty Shades Darker / Fifty Shades Freed)	E L James	4.5	13964	32	2012	Fiction
10	Joyland (Hard Case Crime)	Stephen King	4.5	4748	12	2013	Fiction
11	Little Fires Everywhere	Celeste Ng	4.5	25706	12	2018	Fiction
12	Looking for Alaska	John Green	4.5	8491	7	2014	Fiction

Results per page: 50 1 – 50 of 550

Job history

REFRESH



Type to search

Viewing resources.

SHOW STARRED ONLY

- second-pier-407503
  - Queries
    - Shared queries
    - bookstore-data-lake-queries
    - bookstore-db-queries
  - Notebooks
  - External connections
    - us.bookstore-db
  - bookstore_data_lake
    - amazon
    - goodreads

SUMMARY

amazon

second-pier-407503.bookstore_data_lake

Last modified	Dec 13, 2023, 6:32:01 AM UTC-5
Data location	US
Description	

amazon

Filter Enter property name or value

☐	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
☐	Name	STRING	NULLABLE	-	-	-	-	-
☐	Author	STRING	NULLABLE	-	-	-	-	-
☐	User_Rating	FLOAT	NULLABLE	-	-	-	-	-
☐	Reviews	INTEGER	NULLABLE	-	-	-	-	-
☐	Price	INTEGER	NULLABLE	-	-	-	-	-
☐	Year	INTEGER	NULLABLE	-	-	-	-	-
☐	Genre	STRING	NULLABLE	-	-	-	-	-

EDIT SCHEMA VIEW ROW ACCESS POLICIES

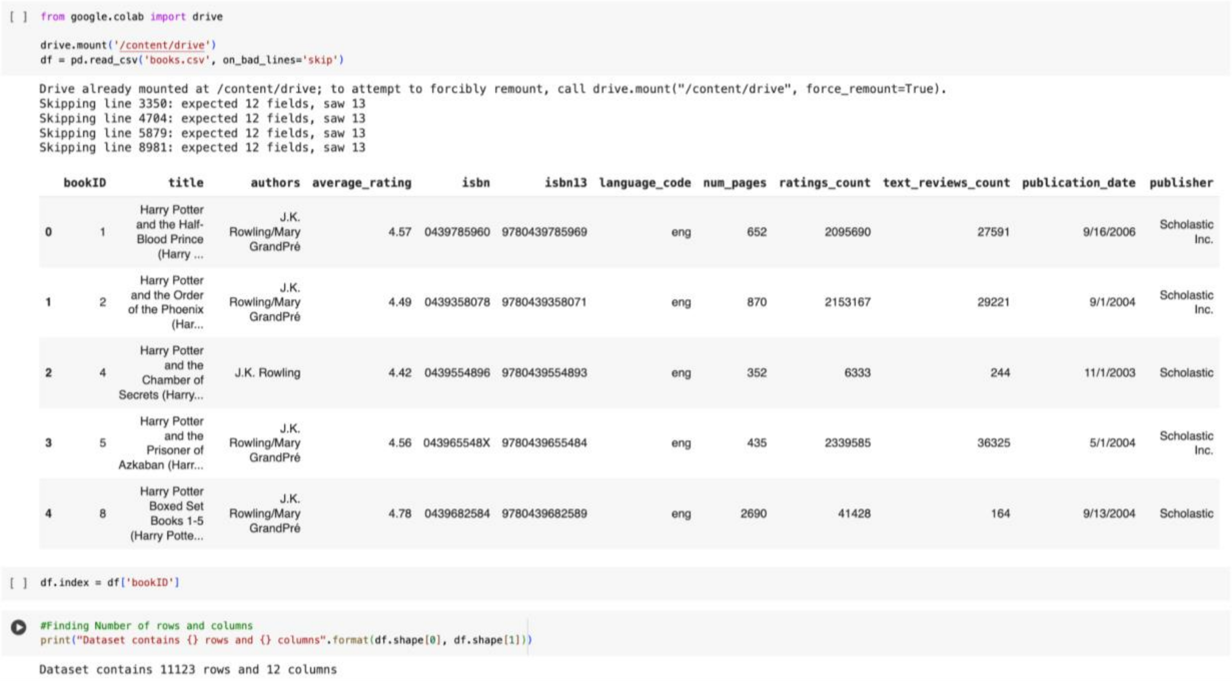
Job history

REFRESH

Data Transformation and Analysis

Using feature engineering and data analytics, we extract meaningful insights from the reviews. This includes identifying key themes, review scores, and customer preferences. The transformation process converts unstructured data into a more semi-structured format that can be easily integrated with the existing EDA.

Basic Data Exploration



```
[ ] from google.colab import drive

    drive.mount('/content/drive')
    df = pd.read_csv('books.csv', on_bad_lines='skip')
```

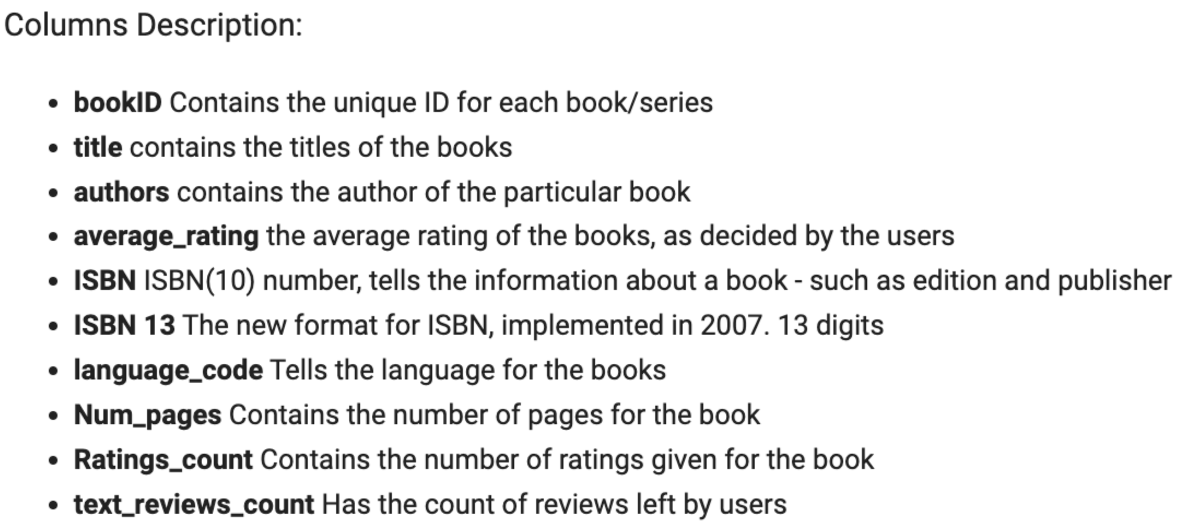
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Skipping line 3350: expected 12 fields, saw 13
Skipping line 4704: expected 12 fields, saw 13
Skipping line 5879: expected 12 fields, saw 13
Skipping line 8981: expected 12 fields, saw 13
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
0	1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling/Mary GrandPré	4.57	0439785960	9780439785969	eng	652	2095690	27591	9/16/2006	Scholastic Inc.
1	2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling/Mary GrandPré	4.49	0439358078	9780439358071	eng	870	2153167	29221	9/1/2004	Scholastic Inc.
2	4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	0439554896	9780439554893	eng	352	6333	244	11/1/2003	Scholastic
3	5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling/Mary GrandPré	4.56	043965548X	9780439655484	eng	435	2339585	36325	5/1/2004	Scholastic Inc.
4	8	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling/Mary GrandPré	4.78	0439682584	9780439682589	eng	2690	41428	164	9/13/2004	Scholastic

```
[ ] df.index = df['bookID']
```

```
#Finding Number of rows and columns
print("Dataset contains {} rows and {} columns".format(df.shape[0], df.shape[1]))
```

```
Dataset contains 11123 rows and 12 columns
```

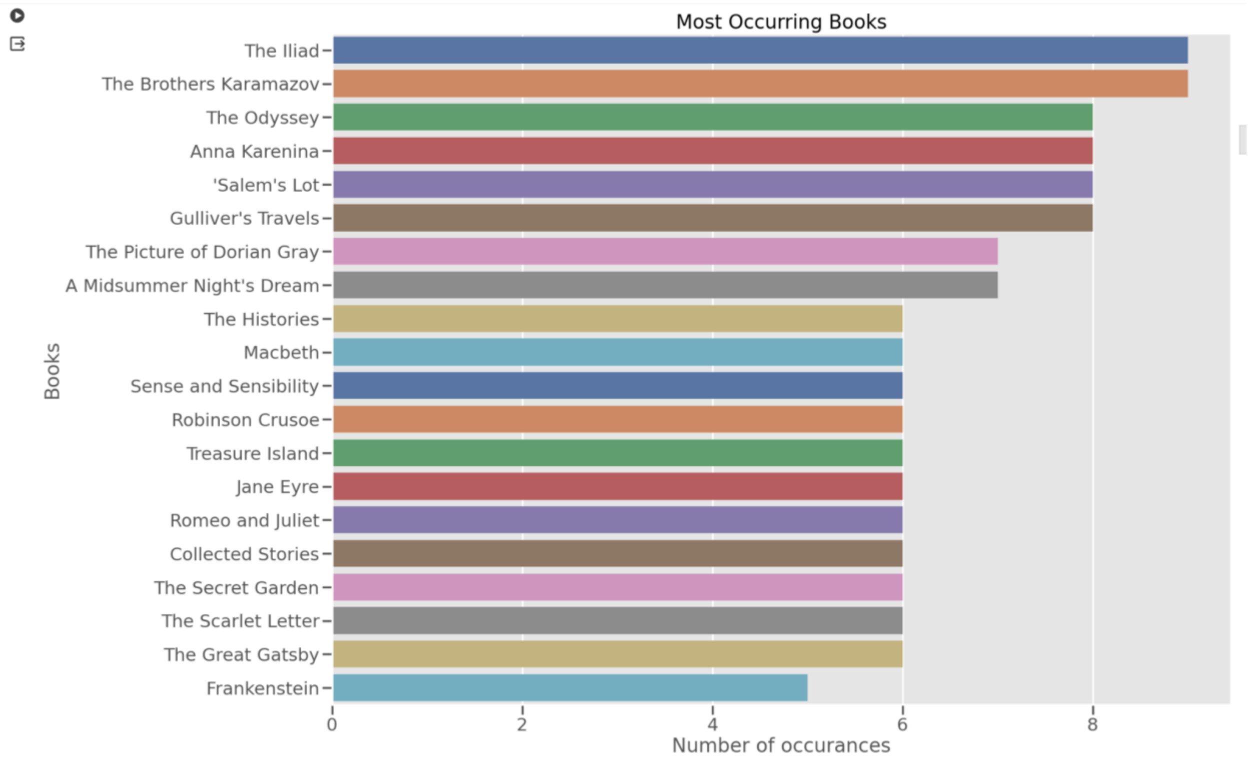


Columns Description:

- **bookID** Contains the unique ID for each book/series
- **title** contains the titles of the books
- **authors** contains the author of the particular book
- **average_rating** the average rating of the books, as decided by the users
- **ISBN** ISBN(10) number, tells the information about a book - such as edition and publisher
- **ISBN 13** The new format for ISBN, implemented in 2007. 13 digits
- **language_code** Tells the language for the books
- **Num_pages** Contains the number of pages for the book
- **Ratings_count** Contains the number of ratings given for the book
- **text_reviews_count** Has the count of reviews left by users

Exploratory Data Analysis

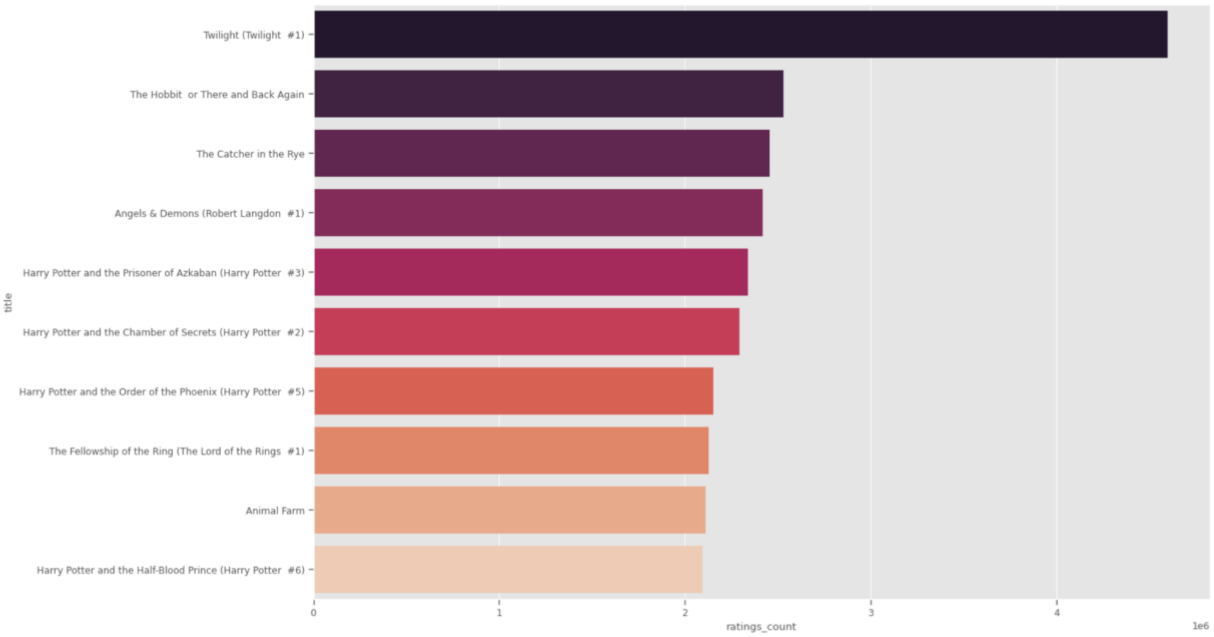
1. Which books have the most occurrences in the list?



We can see that **The Iliad** and **The Brothers Karamazov** have the greatest number of occurrences with the same name in the data.

These books have come up in this database repeatedly, with various publication editions. From the list, we can see that most of the books from the given chart are either old, steadfast classics or books which are usually assigned to schools.

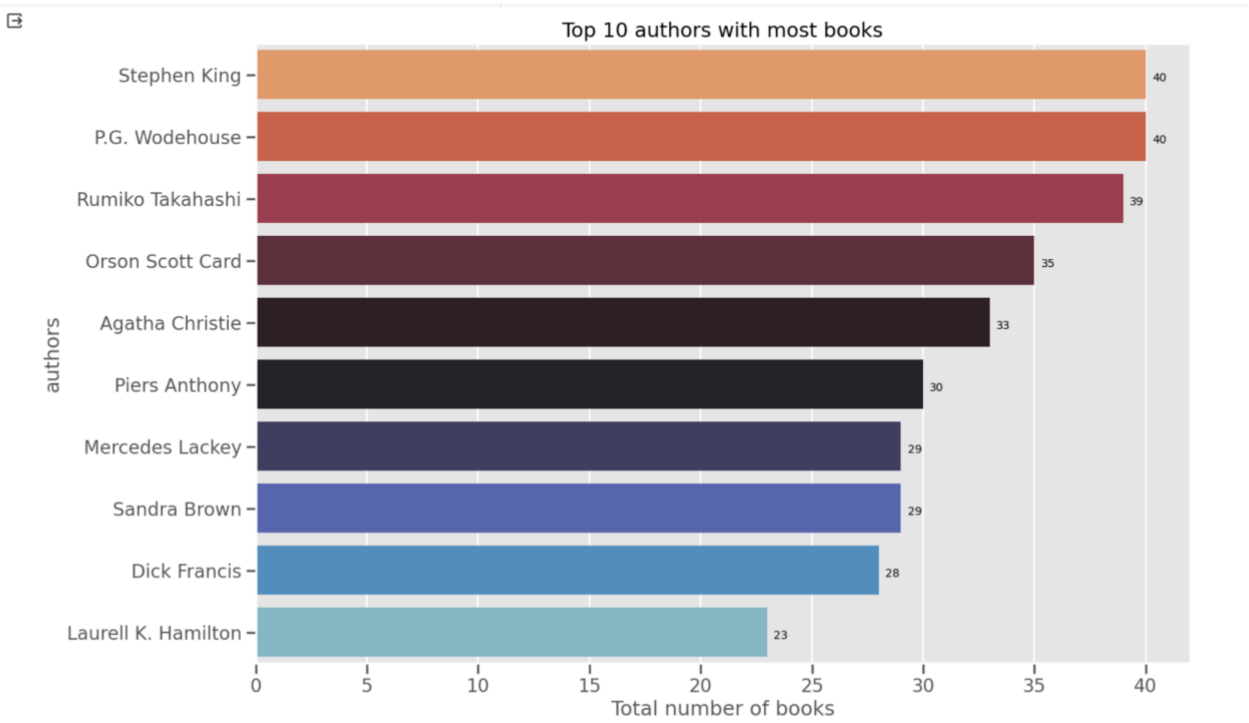
2. Which are the top 10 most rated books?



We can see that the beginning books of the series usually have most of the ratings, i.e., **Harry Potter and the Sorcerer's stone**, **Twilight #1**, **The Hobbit**, **Angels and demons #1**.

Twilight's first book dominates the section by having more than 5000000 ratings. In fact, apart from a few, such as Catcher in the Rye and Animal Farm, all the books seem to be from a series of books, implying perhaps that once people begin a series, they are likely to complete it.

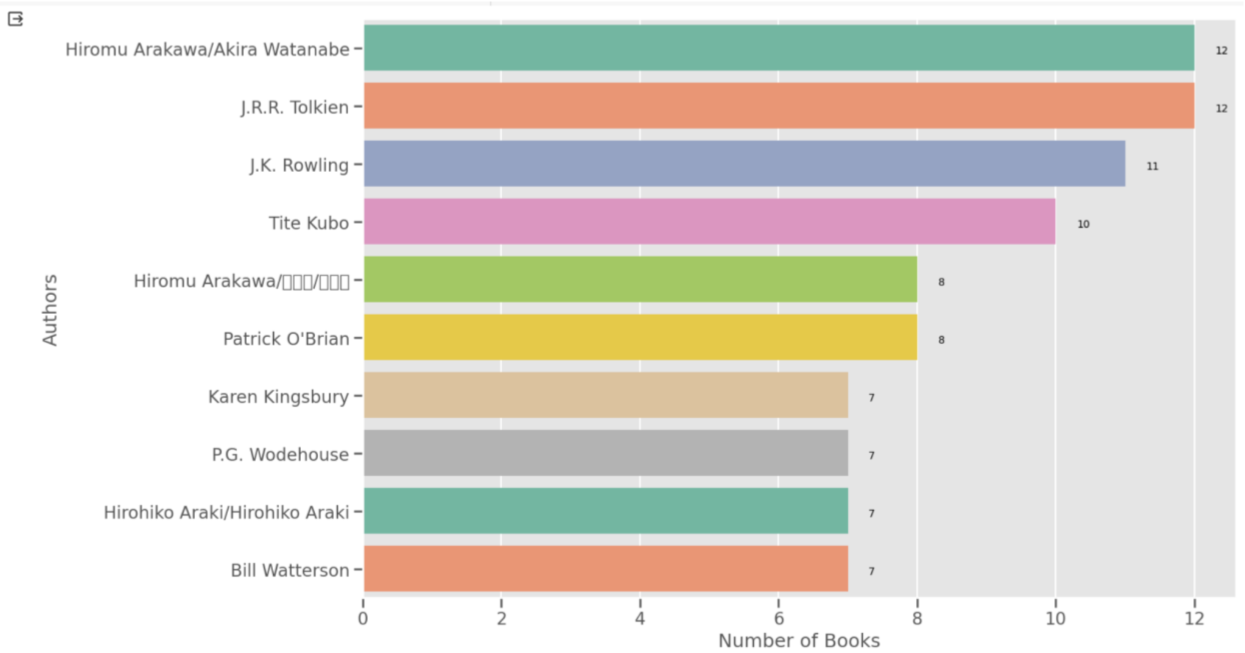
3. Which are the authors with most books?



We can see from the above plot that Stephen King has the greatest number of books in the list - although a lot of them might be just various publications for the same book, since his work has been here for quite a while, spanning decades.

From the names in the list, we can again gather that most of the authors have either been writing for decades, writing numerous books from time to time, or are authors who are regarded as the 'classics' in our history.

4. Which are the top 10 highly rated authors?



We can infer from the plot that the popular author J.R.R Tolkien has the highest ratings for his books, making his average consistency rather impressive.

Integration with Structured Data

The enriched data is then combined with the structured data in the EDA. For instance, ratings result from book ratings can be linked with corresponding book records in the database. This integration enriches the structured data with qualitative insights, offering a more comprehensive view of customer preferences and market trends.

Leveraging Insights in the EDA

Enhanced Decision Making

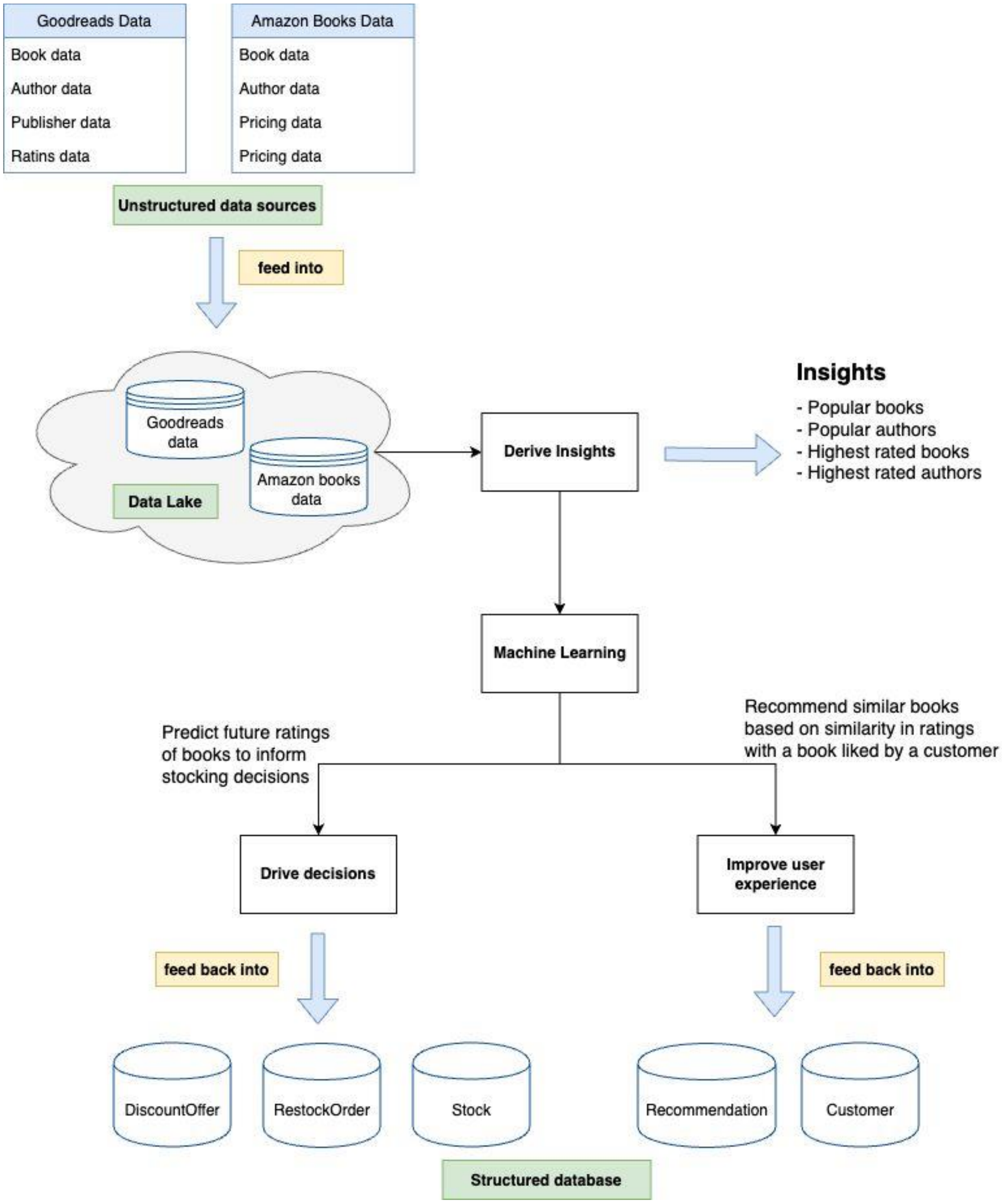
The insights gathered from unstructured data feed into various aspects of the bookstore's operation. For example, highly rated books can inform inventory decisions, highlighting which titles to stock more heavily.

Personalized Recommendations

Analysis of customer reviews helps in fine-tuning the recommendation algorithms, allowing for more personalized and relevant suggestions to bookstore customers.

Market Trend Analysis

By analyzing the sentiments and themes in book reviews, the bookstore can identify emerging trends and reader interests, aiding in strategic planning for stock selection and marketing campaigns.



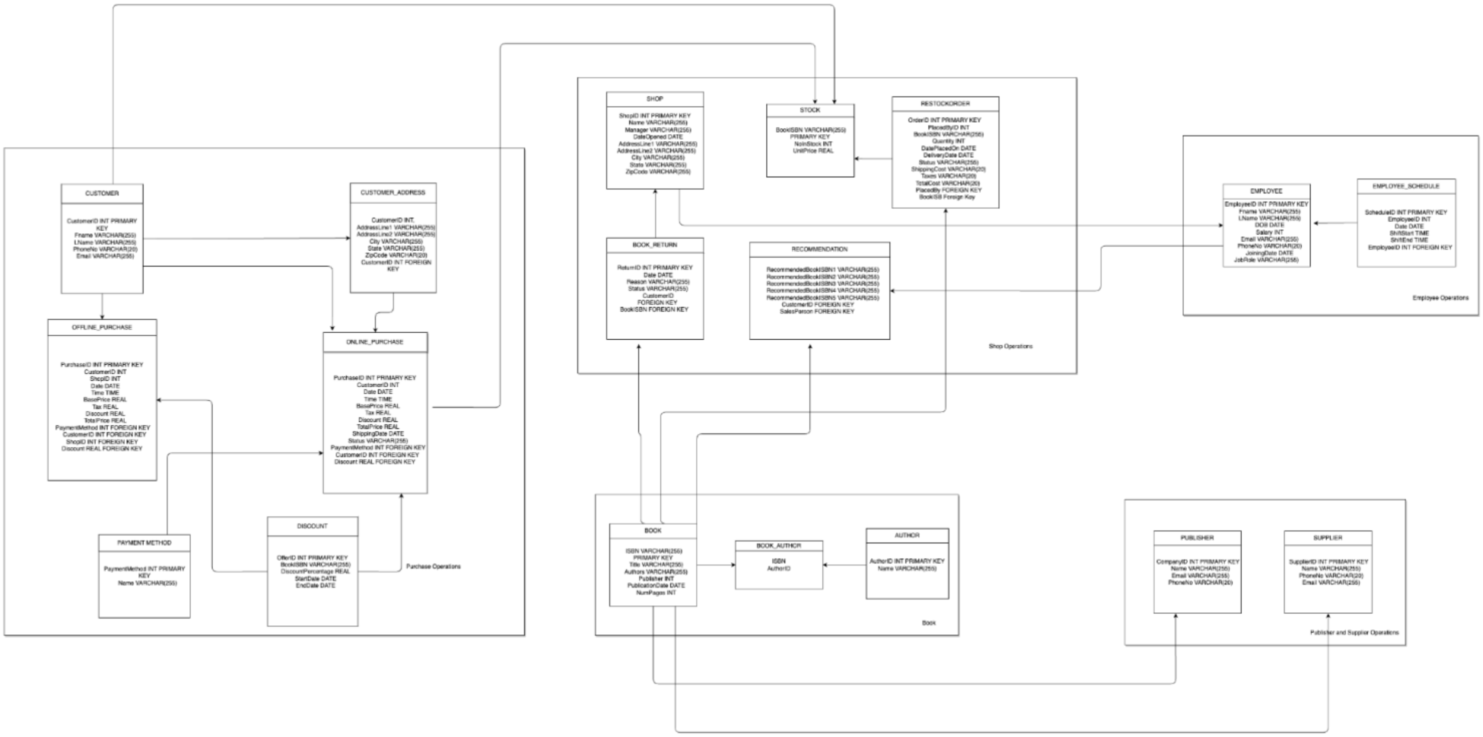
The integration of a data lake into our bookstore's EDA is a strategic move towards embracing big data analytics. By effectively collecting, processing, and leveraging unstructured data from Good Reads and Amazon book reviews, we significantly enhance our understanding of customer preferences and market dynamics. This, in turn, enables more informed decision-making, improves customer engagement, and drives business growth.

Logical Database Schema Creation and Optimization

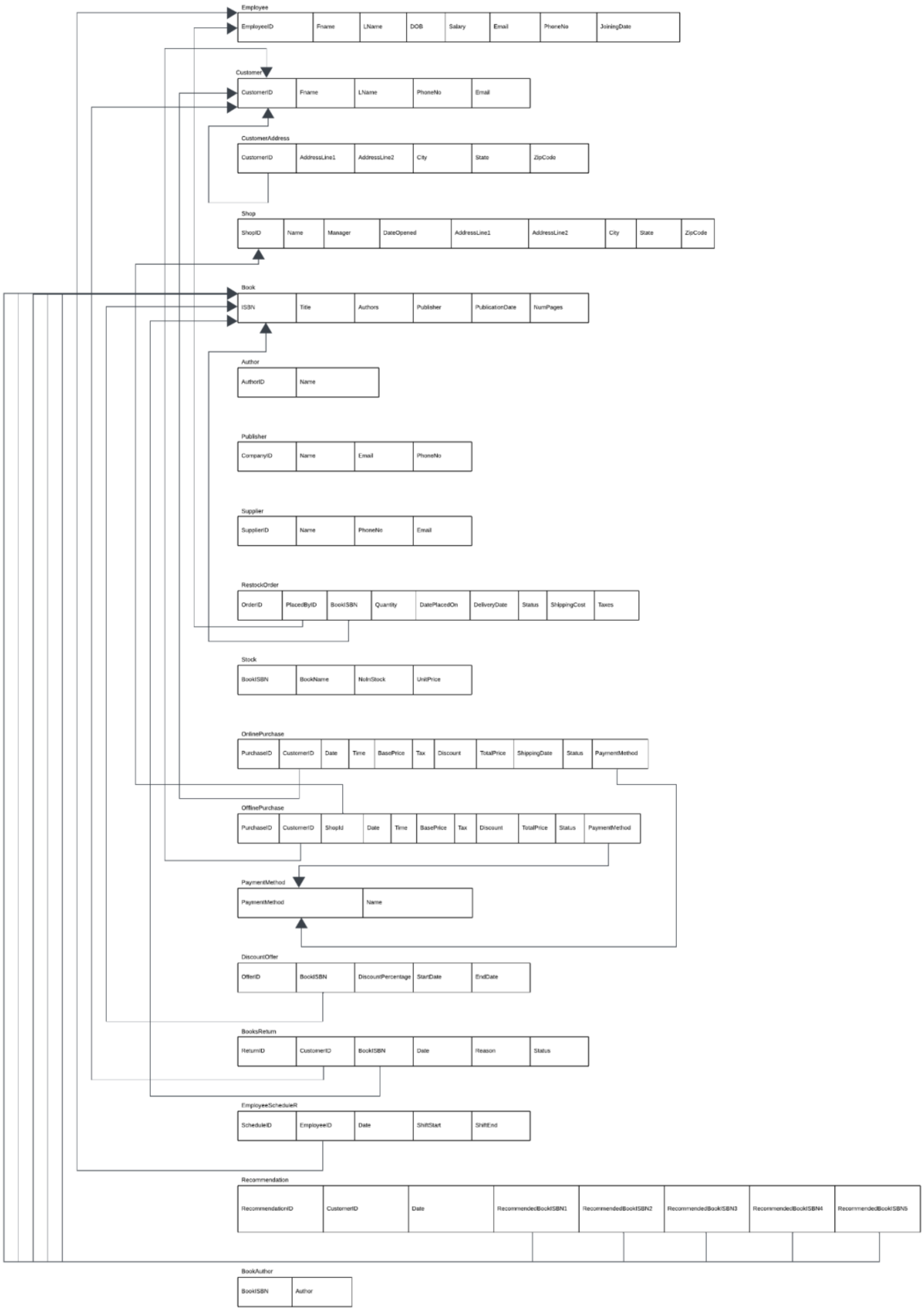
Following the conceptual model established in Part 1 of the project, we have developed a logical database schema designed to efficiently manage and inter-relate the structured and unstructured data within our bookstore's hybrid data model. This schema is a crucial step in transforming our conceptual understanding of the data into a practical, optimized database structure that supports both the operational needs of the bookstore and the analytical insights required for decision-making.

Logical Schema Creation

We began by translating the entity-relationship (ER) conceptual model into a logical schema. This involved defining tables, columns, data types, and constraints that correspond to the entities and relationships outlined in the conceptual model.



Relational Schema Creation



Employee

EmployeeID	Fname	LName	DOB	Salary	Email	PhoneNo	JoiningDate

Customer

CustomerID	Fname	LName	PhoneNo	Email

CustomerAddress

CustomerID	AddressLine1	AddressLine2	City	State	ZipCode

Shop

ShopID	Name	Manager	DateOpened	AddressLine1	AddressLine2	City	State	ZipCode

Book

ISBN	Title	Authors	Publisher	PublicationDate	NumPages

Author

AuthorID	Name

Publisher

CompanyID	Name	Email	PhoneNo

Supplier

SupplierID	Name	PhoneNo	Email

RestockOrder

OrderID	PlacedByID	BookISBN	Quantity	DatePlacedOn	DeliveryDate	Status	ShippingCost	Taxes

Stock

BookISBN	BookName	NoInStock	UnitPrice

OnlinePurchase

PurchaseID	CustomerID	Date	Time	BasePrice	Tax	Discount	TotalPrice	ShippingDate	Status	PaymentMethod

OfflinePurchase

PurchaseID	CustomerID	ShopId	Date	Time	BasePrice	Tax	Discount	TotalPrice	Status	PaymentMethod

PaymentMethod

PaymentMethod	Name

DiscountOffer

OfferID	BookISBN	DiscountPercentage	StartDate	EndDate

BooksReturn

ReturnID	CustomerID	BookISBN	Date	Reason	Status

EmployeeScheduleR

ScheduleID	EmployeeID	Date	ShiftStart	ShiftEnd

Recommendation

RecommendationID	CustomerID	Date	RecommendedBookISBN1	RecommendedBookISBN2	RecommendedBookISBN3	RecommendedBookISBN4	RecommendedBookISBN5

BookAuthor

BookISBN	Author

1. **Identified Entities and Attributes**: Reviewed the ERD and identified each entity along with its attributes. Ensured a clear understanding of the data that needed to be stored.

2. **Defined Primary Keys**: For each entity, designated a primary key. The primary key uniquely identified each record in the table.

3. **Handled Relationships**: Examined relationships between entities and identified foreign keys that connected tables. Ensured that the foreign key in one table matched the primary key in the related table. Reflected cardinality and participation constraints in the schema.

4. **Created Tables**: For each entity in the ERD, created a corresponding table in the relational schema. Included all attributes and their respective data types. Specified primary keys and foreign keys as needed.

5. **Handled Many-to-Many Relationships**: If the ERD included many-to-many relationships, created an associative (junction) table. This table included the primary keys of the entities involved in the relationship, effectively resolving the many-to-many relationship.

6. **Established Naming Conventions**: Established naming conventions for tables and columns to enhance readability and maintainability.

7. **Reviewed and Refined**: Reviewed the relational schema to ensure it accurately represented the information in the ERD. Refined the schema as needed, addressing any potential issues or optimizations.

Schema Optimization – Normalization

The normalization process for the bookstore database was crucial to enhance data organization, eliminate redundancy, and maintain data integrity. The database underwent normalization up to the third normal form (3NF), ensuring a balance between minimizing redundancy and optimizing performance.

First Normal Form (1NF):

- Ensured that each table has a primary key for unique identification - EmployeeID, CustomerID, ISBN, ShopID etc.
- Verified that the values in each column, such as ISBN, Title, and EmployeeID, are atomic, avoiding multi-valued attributes.

Second Normal Form (2NF):

- Removed partial dependencies by ensuring that all non-key columns are fully functionally dependent on the primary key.

Book and Author:

In this initial structure, the Book table has a composite primary key (ISBN, AuthorID), where ISBN uniquely identifies a book, but an author can co-author a book, leading to duplicate ISBN values. Additionally, the AuthorName attribute is included in the Book table, even though it is dependent only on the AuthorID.

- The Book table has a single-column primary key (ISBN), ensuring each record is uniquely identified.
- The AuthorName attribute, which was partially dependent on the composite key (ISBN, AuthorID), has been removed from the Book table and is now fully dependent on the AuthorID in the Author table.

This structure eliminates partial dependencies and adheres to 2NF, improving data integrity and reducing redundancy in the database.

Third Normal Form (3NF):

- Eliminated transitive dependencies, ensuring that non-key columns are only dependent on the primary key.

Book and Publisher:

- In the original structure, the Book table had a Publisher Name, Phone No, and Email ID. The name phone number and email id of publisher are trasnitive dependencies that refer only to the Publisher.

We removed the transitive dependencies by creating a separate table for the dependent attributes.

- The publisher table is created with attributes CompanyID, Name, PhoneNo and EmailID and CompanyID is now included in Book as foreign key.

This structure adheres to 3NF by eliminating transitive dependencies, promoting data integrity, and minimizing redundancy in the database.

Denormalization for Query Optimization

The Schedule details (ScheduleID, Date, ShiftStart, ShiftEnd) were originally part of the Employee entity and were moved to a separate Schedule entity. Denormalization involves intentionally introducing redundancy or breaking normalization rules to improve query performance or simplify certain types of queries.

The schedule information doesn't change frequently and is not critical for every employee-related operation. By separating schedule details into a Schedule entity, we reduce the need to update the Employee entity for schedule changes, thereby potentially improving performance.

Extensions for Unstructured Data

To bridge the gap between structured and unstructured data, the database schema was extended to incorporate reference points to unstructured datasets.

A dedicated **Recommendations table** has been established to capture the personalized book recommendations for each customer. These recommendations are meticulously curated through the utilization of our machine learning model, which operates on unstructured data, extracting valuable insights to enhance the customer experience.

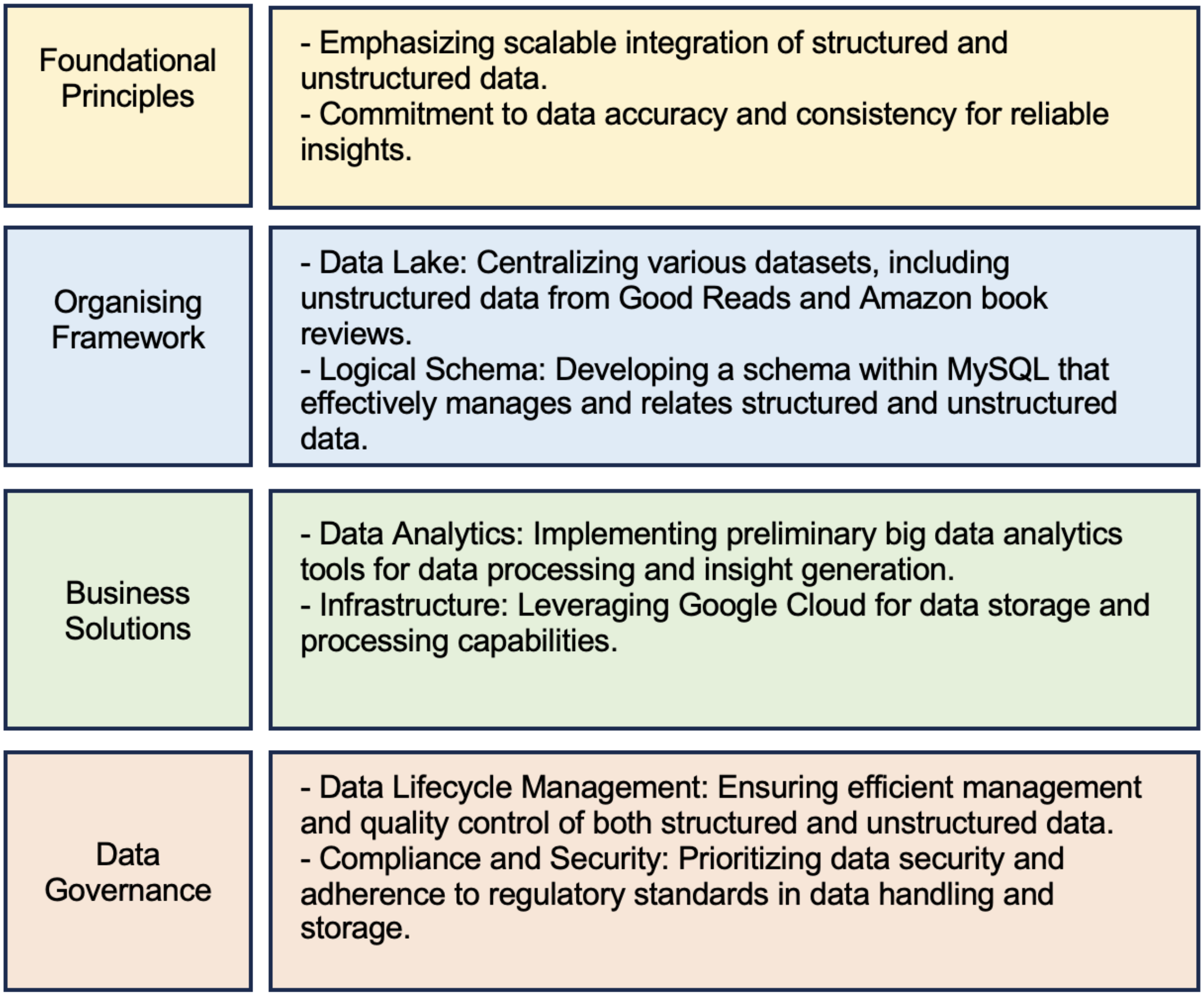
Simultaneously, the **Stock table** undergoes dynamic updates informed by the predictions generated by our machine learning model. Specifically, the model analyzes unstructured data to predict book ratings, and subsequently, the Stock table is selectively updated to restock only those books that receive higher ratings. This strategic approach ensures that our inventory aligns with customer preferences and promotes the availability of books that are likely to garner greater interest and satisfaction among our clientele.

These extensions enable a comprehensive understanding of the bookstore data by integrating insights from both structured and unstructured sources. The resulting database structure supports a wide range of queries and analyses, enhancing the overall flexibility and utility of the system.

Each optimization step was taken with the dual goals of maintaining data integrity and enhancing query performance. The schema was also optimized to support the integration with machine learning models, ensuring that data feeds into these models efficiently and accurately.

In summary, the creation and optimization of the logical database schema are pivotal to the success of our bookstore's data architecture. The structured data is now well-defined, normalized, and optimized for performance, while still being flexible enough to integrate with the unstructured data from our data lake. This hybrid data model supports a comprehensive suite of analytics and machine learning capabilities, setting the foundation for insightful decision-making and a robust, data-driven business strategy.

Reference Architecture



Foundational Principles	- Emphasizing scalable integration of structured and unstructured data. - Commitment to data accuracy and consistency for reliable insights.
Organising Framework	- Data Lake: Centralizing various datasets, including unstructured data from Good Reads and Amazon book reviews. - Logical Schema: Developing a schema within MySQL that effectively manages and relates structured and unstructured data.
Business Solutions	- Data Analytics: Implementing preliminary big data analytics tools for data processing and insight generation. - Infrastructure: Leveraging Google Cloud for data storage and processing capabilities.
Data Governance	- Data Lifecycle Management: Ensuring efficient management and quality control of both structured and unstructured data. - Compliance and Security: Prioritizing data security and adherence to regulatory standards in data handling and storage.

Leveraging Google Cloud for Hybrid Data Management

We have selected Google Cloud as the cloud platform for storing and managing our hybrid data for the bookstore.

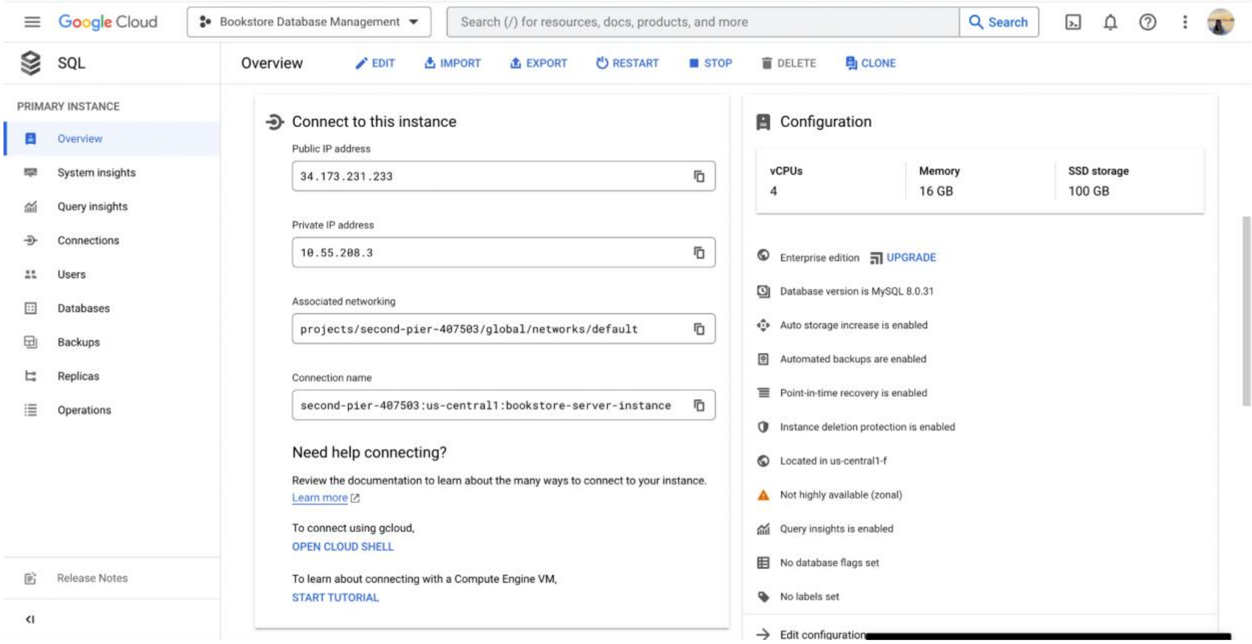
Justification

1. Compatibility: Google Cloud's compatibility with our existing tools (e.g., Google BigQuery, Google Collaboratory) ensures seamless integration and workflow.

2. Scalability: Google Cloud offers robust scalability options, essential for handling our growing data needs and fluctuating traffic.

3. Advanced Analytics and ML Support: Google Cloud's advanced analytics and machine learning services align well with our requirement for sophisticated data processing and model development.

4. Security and Reliability: It provides strong security features and reliability, which are critical for our bookstore's data integrity and availability.



Connect to this instance

Public IP address: 34.173.231.233

Private IP address: 10.55.208.3

Associated networking: projects/second-pier-407503/global/networks/default

Connection name: second-pier-407503:us-central1:bookstore-server-instance

Need help connecting?

Review the documentation to learn about the many ways to connect to your instance. Learn more

To connect using gcloud, OPEN CLOUD SHELL

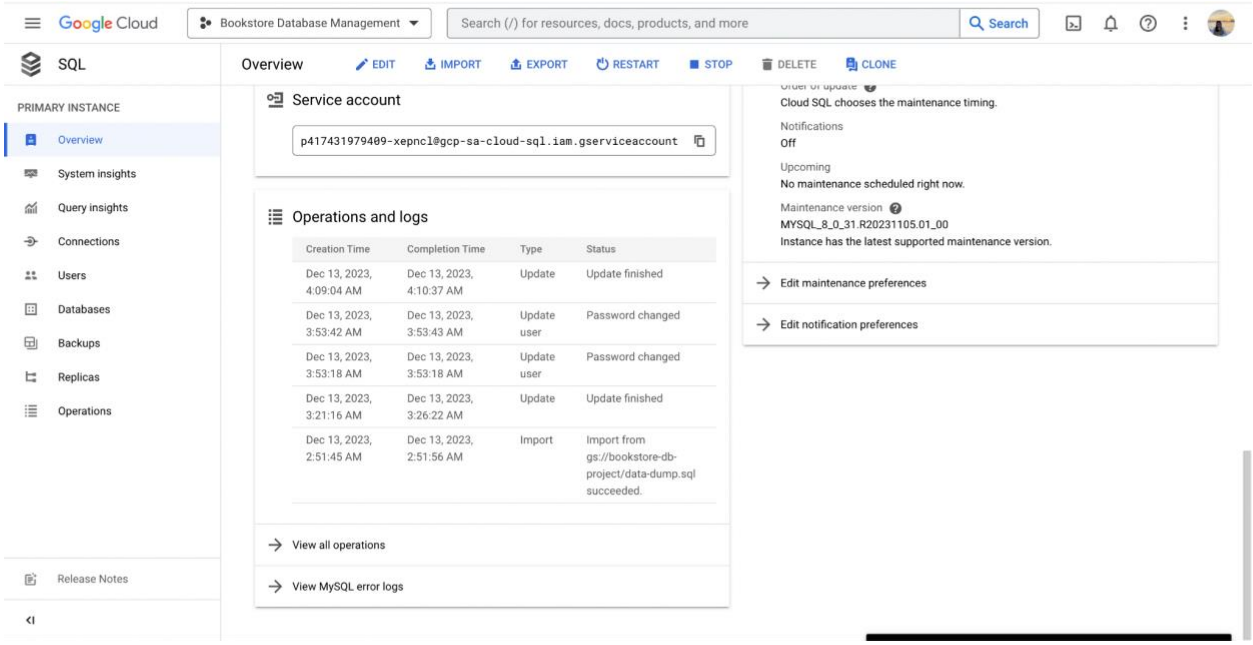
To learn about connecting with a Compute Engine VM, START TUTORIAL

Configuration

vCPUs	Memory	SSD storage
4	16 GB	100 GB

- Enterprise edition UPGRADE
- Database version is MySQL 8.0.31
- Auto storage increase is enabled
- Automated backups are enabled
- Point-in-time recovery is enabled
- Instance deletion protection is enabled
- Located in us-central1-f
- Not highly available (zonal)
- Query insights is enabled
- No database flags set
- No labels set

Edit configuration [redacted]



Service account

p417431979409-xepncl@gcp-sa-cloud-sql.iam.gserviceaccount

Operations and logs

Creation Time	Completion Time	Type	Status
Dec 13, 2023, 4:09:04 AM	Dec 13, 2023, 4:10:37 AM	Update	Update finished
Dec 13, 2023, 3:53:42 AM	Dec 13, 2023, 3:53:43 AM	Update user	Password changed
Dec 13, 2023, 3:53:18 AM	Dec 13, 2023, 3:53:18 AM	Update user	Password changed
Dec 13, 2023, 3:21:16 AM	Dec 13, 2023, 3:26:22 AM	Update	Update finished
Dec 13, 2023, 2:51:45 AM	Dec 13, 2023, 2:51:56 AM	Import	Import from gs://bookstore-db-project/data-dump.sql succeeded.

Order of update

Cloud SQL chooses the maintenance timing.

Notifications

Off

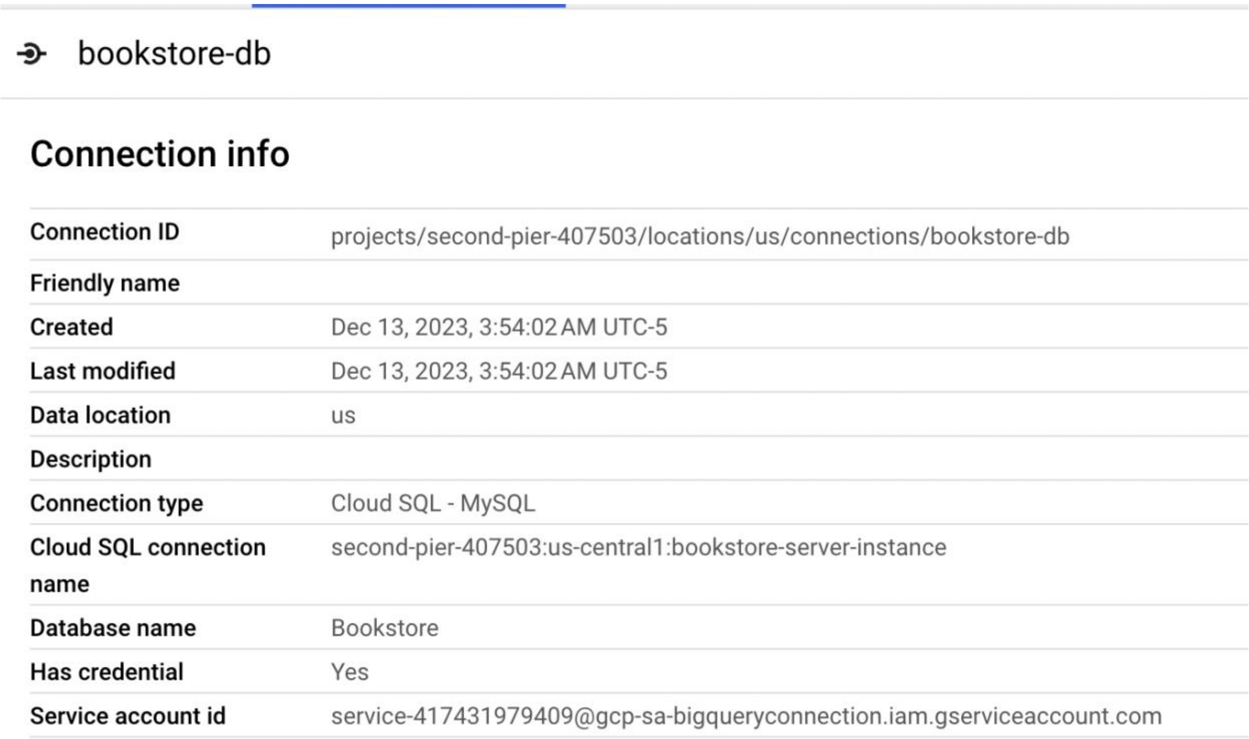
Upcoming

No maintenance scheduled right now.

Maintenance version

MYSQL_8_0_31.R20231105.01_00

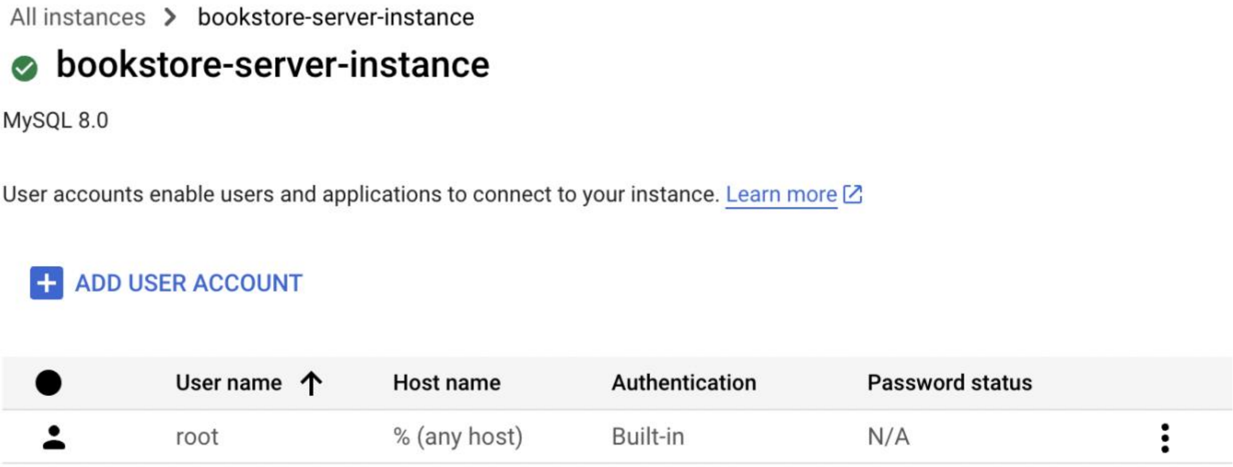
Instance has the latest supported maintenance version.



bookstore-db

Connection info

Connection ID	projects/second-pier-407503/locations/us/connections/bookstore-db
Friendly name	
Created	Dec 13, 2023, 3:54:02 AM UTC-5
Last modified	Dec 13, 2023, 3:54:02 AM UTC-5
Data location	us
Description	
Connection type	Cloud SQL - MySQL
Cloud SQL connection name	second-pier-407503:us-central1:bookstore-server-instance
Database name	Bookstore
Has credential	Yes
Service account id	service-417431979409@gcp-sa-bigqueryconnection.iam.gserviceaccount.com

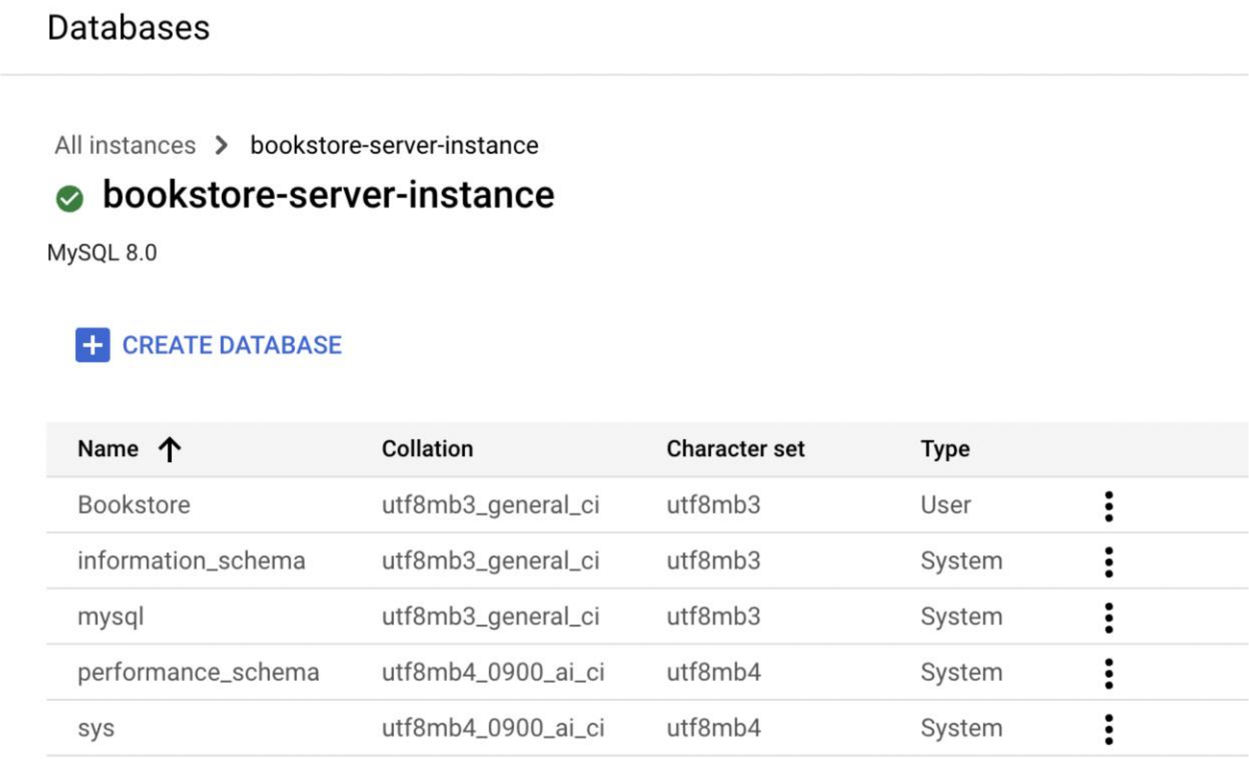


bookstore-server-instance

MySQL 8.0

User accounts enable users and applications to connect to your instance. Learn more

●	User name ↑	Host name	Authentication	Password status	
	root	% (any host)	Built-in	N/A	⋮



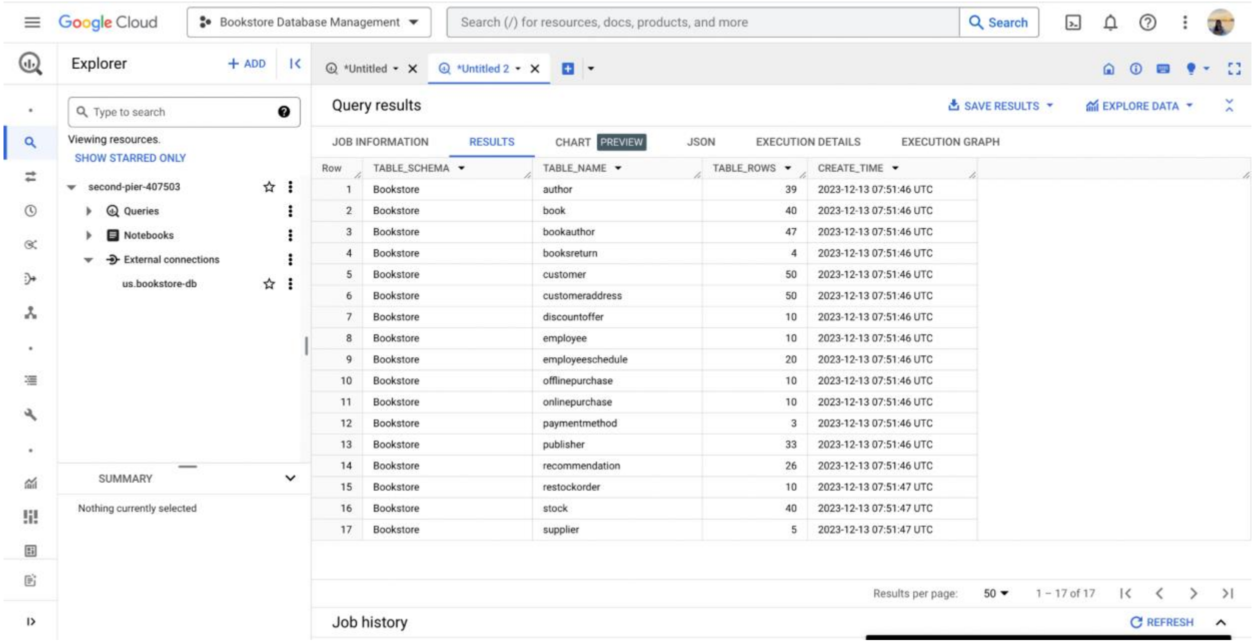
Databases

All instances > bookstore-server-instance

bookstore-server-instance

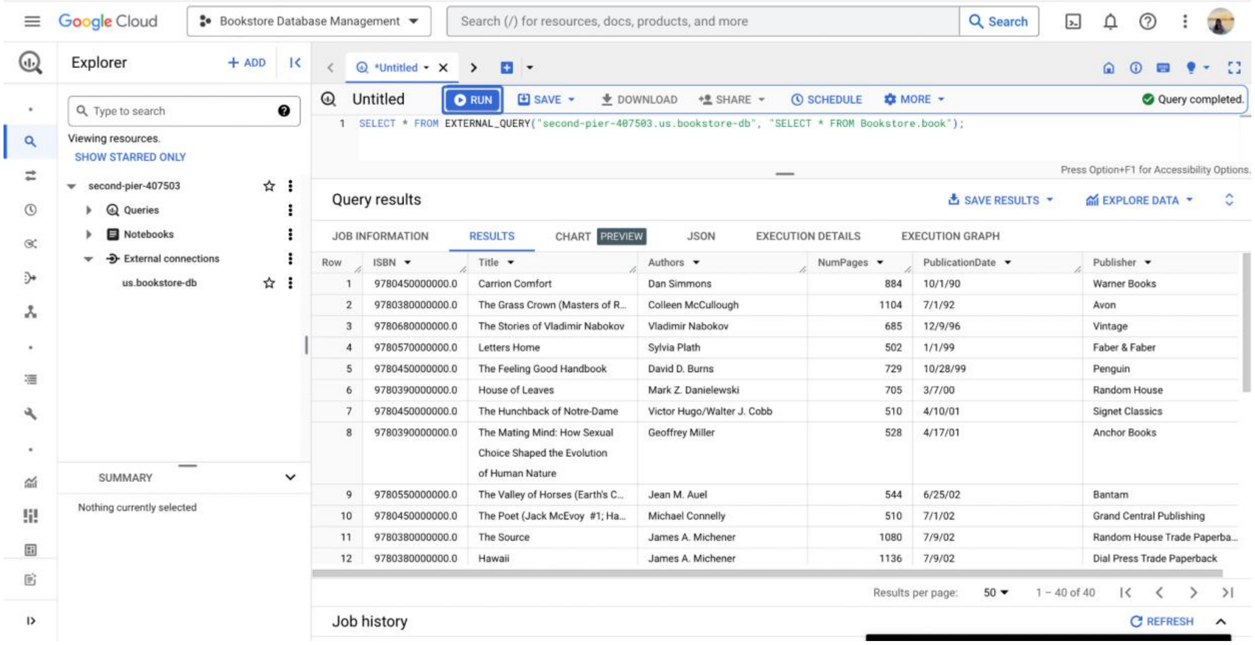
MySQL 8.0

Name ↑	Collation	Character set	Type
Bookstore	utf8mb3_general_ci	utf8mb3	User
information_schema	utf8mb3_general_ci	utf8mb3	System
mysql	utf8mb3_general_ci	utf8mb3	System
performance_schema	utf8mb4_0900_ai_ci	utf8mb4	System
sys	utf8mb4_0900_ai_ci	utf8mb4	System



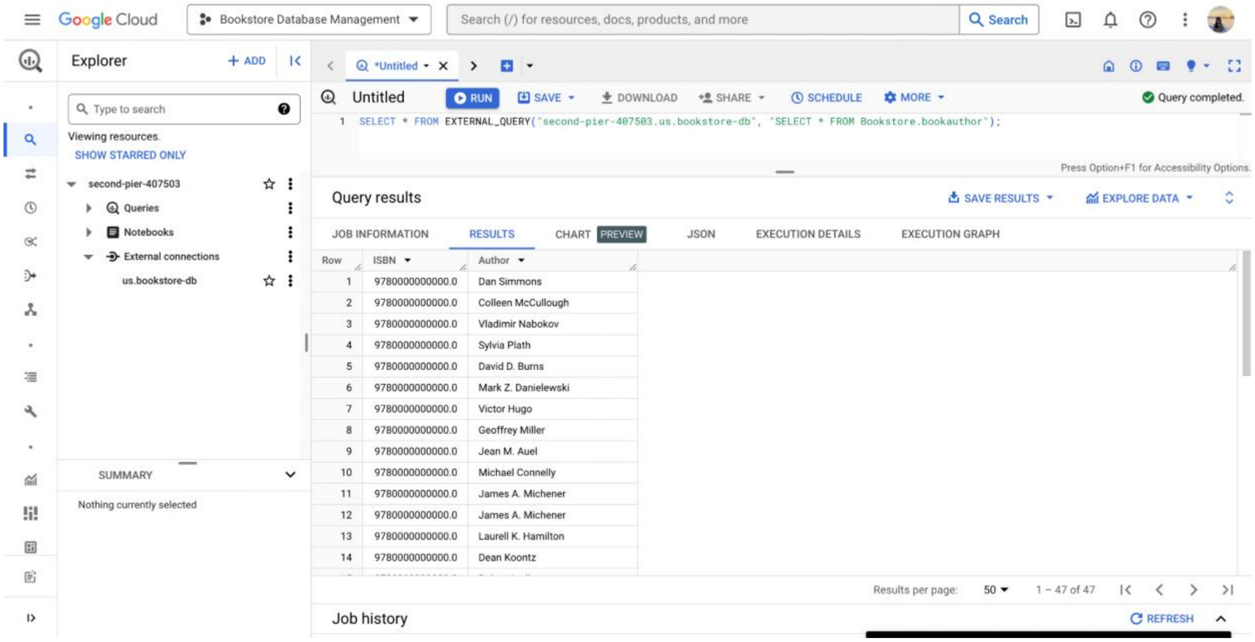
Query results

Row	TABLE_SCHEMA	TABLE_NAME	TABLE_ROWS	CREATE_TIME
1	Bookstore	author	39	2023-12-13 07:51:46 UTC
2	Bookstore	book	40	2023-12-13 07:51:46 UTC
3	Bookstore	bookauthor	47	2023-12-13 07:51:46 UTC
4	Bookstore	booksreturn	4	2023-12-13 07:51:46 UTC
5	Bookstore	customer	50	2023-12-13 07:51:46 UTC
6	Bookstore	customeraddress	50	2023-12-13 07:51:46 UTC
7	Bookstore	discountoffer	10	2023-12-13 07:51:46 UTC
8	Bookstore	employee	10	2023-12-13 07:51:46 UTC
9	Bookstore	employeeschedule	20	2023-12-13 07:51:46 UTC
10	Bookstore	offlinepurchase	10	2023-12-13 07:51:46 UTC
11	Bookstore	onlinepurchase	10	2023-12-13 07:51:46 UTC
12	Bookstore	paymentmethod	3	2023-12-13 07:51:46 UTC
13	Bookstore	publisher	33	2023-12-13 07:51:46 UTC
14	Bookstore	recommendation	26	2023-12-13 07:51:46 UTC
15	Bookstore	restockorder	10	2023-12-13 07:51:47 UTC
16	Bookstore	stock	40	2023-12-13 07:51:47 UTC
17	Bookstore	supplier	5	2023-12-13 07:51:47 UTC



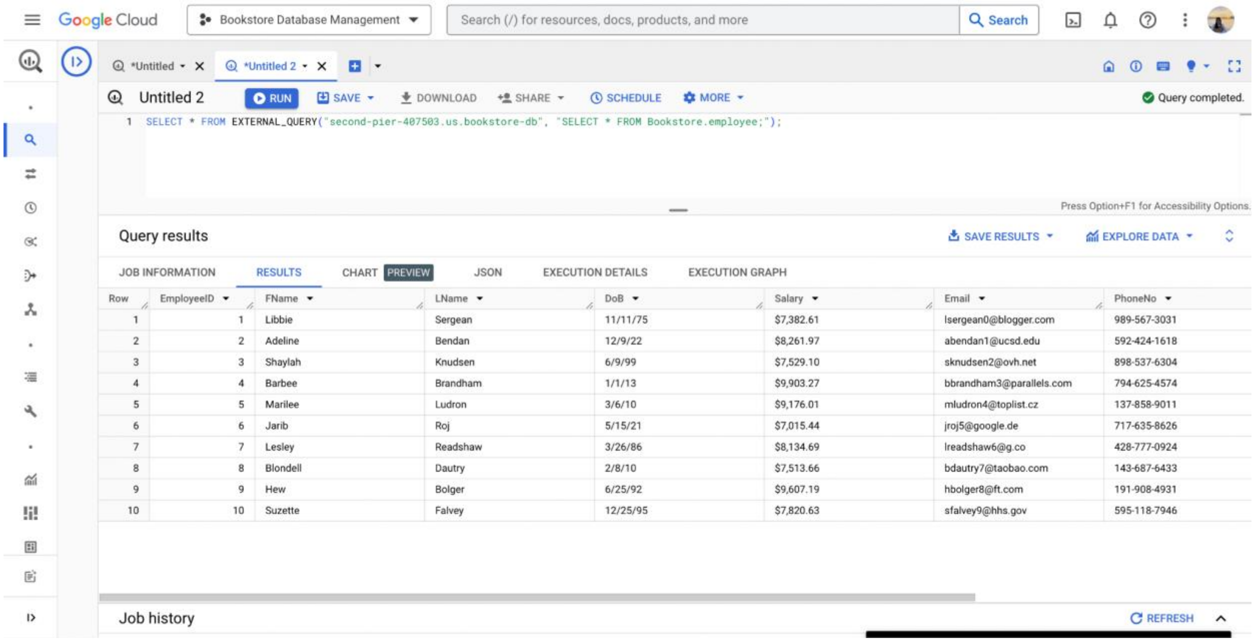
```
SELECT * FROM EXTERNAL_QUERY("second-pier-407503.us.bookstore-db", "SELECT * FROM Bookstore.book");
```

Row	ISBN	Title	Authors	NumPages	PublicationDate	Publisher
1	9780450000000.0	Carrion Comfort	Dan Simmons	884	10/1/90	Warner Books
2	9780380000000.0	The Grass Crown (Masters of R...	Colleen McCullough	1104	7/1/92	Avon
3	9780680000000.0	The Stories of Vladimir Nabokov	Vladimir Nabokov	685	12/9/96	Vintage
4	9780570000000.0	Letters Home	Sylvia Plath	502	1/1/99	Faber & Faber
5	9780450000000.0	The Feeling Good Handbook	David D. Burns	729	10/28/99	Penguin
6	9780390000000.0	House of Leaves	Mark Z. Danielewski	705	3/7/00	Random House
7	9780450000000.0	The Hunchback of Notre-Dame	Victor Hugo/Walter J. Cobb	510	4/10/01	Signet Classics
8	9780390000000.0	The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature	Geoffrey Miller	528	4/17/01	Anchor Books
9	9780550000000.0	The Valley of Horses (Earth's C...	Jean M. Auel	544	6/25/02	Bantam
10	9780450000000.0	The Poet (Jack McEvoy #1; Ha...	Michael Connelly	510	7/1/02	Grand Central Publishing
11	9780380000000.0	The Source	James A. Michener	1080	7/9/02	Random House Trade Paperba...
12	9780380000000.0	Hawaii	James A. Michener	1136	7/9/02	Dial Press Trade Paperback



```
SELECT * FROM EXTERNAL_QUERY("second-pier-407503.us.bookstore-db", "SELECT * FROM Bookstore.bookauthor");
```

Row	ISBN	Author
1	9780000000000.0	Dan Simmons
2	9780000000000.0	Colleen McCullough
3	9780000000000.0	Vladimir Nabokov
4	9780000000000.0	Sylvia Plath
5	9780000000000.0	David D. Burns
6	9780000000000.0	Mark Z. Danielewski
7	9780000000000.0	Victor Hugo
8	9780000000000.0	Geoffrey Miller
9	9780000000000.0	Jean M. Auel
10	9780000000000.0	Michael Connelly
11	9780000000000.0	James A. Michener
12	9780000000000.0	James A. Michener
13	9780000000000.0	Laurell K. Hamilton
14	9780000000000.0	Dean Koontz



```
SELECT * FROM EXTERNAL_QUERY("second-pier-407503.us.bookstore-db", "SELECT * FROM Bookstore.employee;");
```

Row	EmployeeID	FName	LName	DoB	Salary	Email	PhoneNo
1	1	Libbie	Sergean	11/11/75	$7,382.61	lsergean0@blogger.com	989-567-3031
2	2	Adeline	Bendan	12/9/22	$8,261.97	abendan1@ucsd.edu	592-424-1618
3	3	Shaylah	Knudsen	6/9/99	$7,529.10	sknudsen2@ovh.net	898-537-6304
4	4	Barbee	Brandham	1/1/13	$9,903.27	bbrandham3@parallels.com	794-625-4574
5	5	Marilee	Ludron	3/6/10	$9,176.01	mludron4@toplist.cz	137-858-9011
6	6	Jarib	Roj	5/15/21	$7,015.44	jroj5@google.de	717-635-8626
7	7	Lesley	Readshaw	3/26/86	$8,134.69	lreadshaw6@g.co	428-777-0924
8	8	Blondell	Dautry	2/8/10	$7,513.66	bdautry7@taobao.com	143-687-6433
9	9	Hew	Bolger	6/25/92	$9,607.19	hbolger8@ft.com	191-908-4931
10	10	Suzette	Falvey	12/25/95	$7,820.63	sfalvey9@hhs.gov	595-118-7946

Summary

In summary, the integration of a data lake, the development of a comprehensive reference architecture, and the strategic choice of Google Cloud as our cloud platform mark significant advancements in our bookstore's data management capabilities. These efforts not only optimize our existing logical schema but also position us to effectively harness the power of both structured and unstructured data to drive business growth and enhance customer satisfaction.