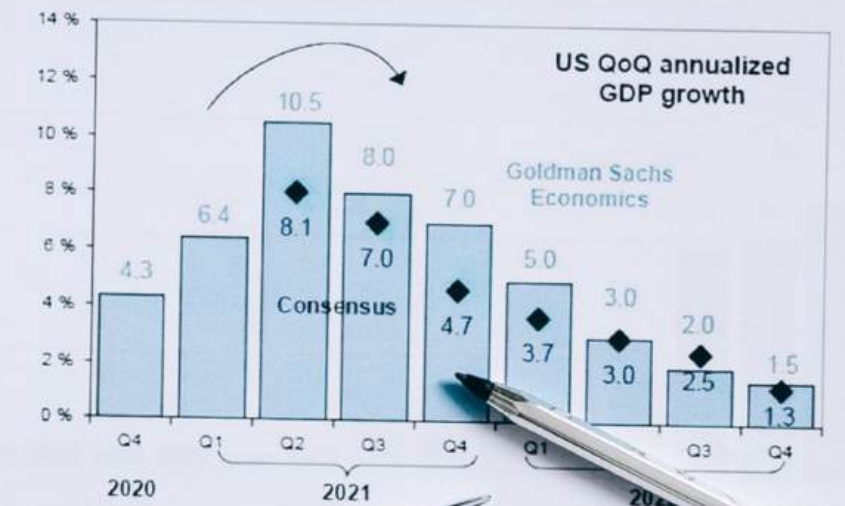


○ ○ ○ ○

Exhibit 3: US economic growth expected to peak this quarter
as of May 5, 2021



PROCESS AND ANALYZE DATA WITH THE CLOUD

○ ○ ○ ○

NỘI DUNG

1. Giới thiệu tổng quan
2. Dữ liệu
3. Xử lý dữ liệu
4. Phân tích kinh doanh
5. Kết quả



1. TỔNG QUAN

TMA Solutions Bình Định

Được thành lập vào năm 1997, TMA là tập đoàn công nghệ hàng đầu tại Việt Nam, gần 4,000 kỹ sư tài năng.

Lĩnh vực

- Ngân hàng và tài chính
- Viễn thông
- Chính phủ và dịch vụ công
- Sản xuất và kinh doanh
- Giáo dục và đào tạo

Chi nhánh

7 chi nhánh tại Việt Nam
6 chi nhánh tại các quốc gia Mỹ, Úc, Canada, Đức, Nhật Bản và Singapore.

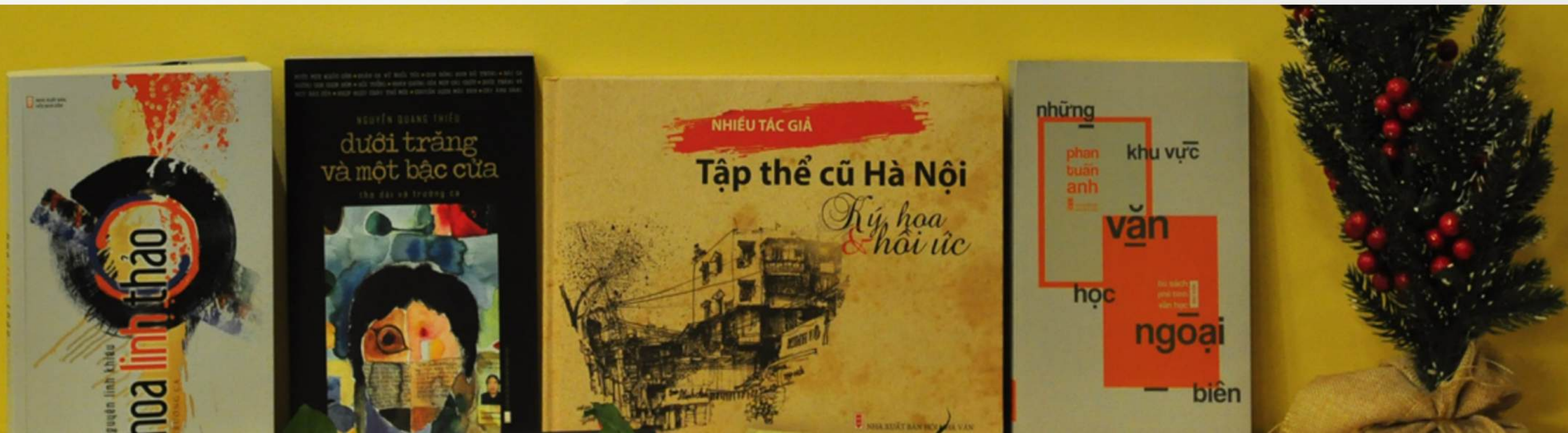
Khách hàng

Từ hơn 30 quốc gia khác nhau trên toàn thế giới.

1. TỔNG QUAN

Khách hàng

Một nhà xuất bản sách tìm những quyển sách để xuất bản có thể tăng doanh thu cho công ty của mình.



Bảng Users

```
User-ID: string  
Location: string  
Age: string
```

Bảng Ratings

```
User-ID: integer  
ISBN: string  
Book-Rating: integer
```

Bảng Books

```
ISBN: string  
Book-Title: string  
Book-Author: string  
Year-Of-Publication: integer  
Publisher: string  
Image-URL-S: string  
Image-URL-M: string  
Image-URL-L: string
```

2. DỮ LIỆU



3. XỬ LÝ DỮ LIỆU



BẢNG BOOKS

ISBN	Book-Title	Book-Author	Year-Of...	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
B00009ANY9	Cane River	Lalita Tademy	2001	Warner Books	> http://image...	> http://images....	> http://image...
000636988X	> How Not to Be a Perfect ...	Libby Purves	1986	> HarperCollins ...	> http://image...	> http://images....	> http://image...
000649840X	Angelas Ashes	Frank Mccourt	0	Harpercollins Uk	> http://image...	> http://images....	> http://image...
0002005018	Clara Callan	> Richard Bruc...	2001	> HarperFlamin...	> http://image...	> http://images....	> http://image...
002026478X	> AGE OF INNOCENCE (M...	Edith Wharton	1993	Scribner	> http://image...	> http://images....	> http://image...
0002245663	> Forms of Devotion: Storie...	> Diane Schoe...	1998	> Harpercollins ...	> http://image...	> http://images....	> http://image...
0002258579	Resurrection Club	> Christopher ...	1999	Harpercollins Uk	> http://image...	> http://images....	> http://image...
0002261529	Temptation	Dermot Bolger	2000	Harpercollins	> http://image...	> http://images....	> http://image...
002542730X	> Politically Correct Bedtim...	James Finn Garner	1994	> John Wiley &a...	> http://image...	> http://images....	> http://image...
002542730x	> Politically Correct Bedtim...	James Finn Garner	1994	> John Wiley &a...	> http://image...	> http://images....	> http://image...
0002713276	Bell' Antonio	Vitaliano Brancati	1994	Harpercollins	> http://image...	> http://images....	> http://image...




```
# Đọc CSV với mã hóa ISO-8859-1
df_books = spark.read.format("csv") \
    .option("header", "true") \
    .option("delimiter", ";") \
    .option("encoding", "ISO-8859-1") \
    .load("dbfs:/FileStore/books/books.csv")
display(df_books)
```

Der illustrierte Mann. Erzählungen.
Der König in Gelb.
Fahrenheit 451
Die Mars- Chroniken. Roman in Erzählungen.
Das Böse kommt auf leisen Sohlen.
Läwenzahnwein. Roman.
Das Kind von morgen. Erzählungen.
Die Mechanismen der Freude. Erzählungen.
Familientreffen. Erzählungen.
Der Tod kommt schnell in Mexico. Erzählungen.

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

XÓA CỘT



```
06:39 PM (<1s) 5
columns_to_drop = ["Image-URL-L", "Image-URL-M", "Image-URL-S", "Year-Of-Publication"]
df_books = df_books.drop(*columns_to_drop)

df_books: pyspark.sql.dataframe.DataFrame
  ISBN: string
  Book-Title: string
  Book-Author: string
  Publisher: string
```

XÓA DỮ LIỆU TRỐNG VÀ TRÙNG

```
df_books = df_books.dropna(subset='ISBN')
df_books = df_books.withColumn('ISBN', upper(trim(col('ISBN'))))
df_books = df_books.dropDuplicates(['ISBN'])
display(df_books)
```

3. XỬ LÝ DỮ LIỆU



SỬA LỖI ĐỊNH DẠNG

```
columns_to_process = ["Book-Title", "Publisher", "Book-Author"]
for col_name in columns_to_process:
    df_books = df_books.withColumn(col_name, regexp_replace(regexp_replace
                                                                (col_name, r'^\x00-\x7F', ''), r'\?(?!\\s)', 'o'))
display(df_books)
```

```
df_books = df_books.withColumn('Book-Title', regexp_replace
                                (regexp_replace(col('Book-Title'), r"^[a-zA-Z0-9.\s()\\"]", " "), r"\s+", " "))
columns = ["Publisher", "Book-Author"]
for column in columns:
    df_books = df_books.withColumn(column, regexp_replace
                                    (regexp_replace(col(column), r"^[a-zA-Z().]", " "), r"\s+", " "))
display(df_books)
```

3257203659	Der illustrierte Mann. Erzählungen.	Ray Bradbury
3257207522	Der König in Gelb.	Raymond Chandler
3257208626	Fahrenheit 451	Ray Bradbury
3257208634	> Die Mars- Chroniken. Roman in Erzähl...	Ray Bradbury
3257208669	Das Böse kommt auf leisen Sohlen.	Ray Bradbury



^A _C ISBN	^A _C Book-Title	^A _C Book-Author
3257203659	Der illustrierte Mann. Erzählungen.	Ray Bradbury

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

SAU XỬ LÝ

df_books: pyspark.sql.dataframe.DataFrame

ISBN: string
Book-Title: string
Book-Author: string
Publisher: string

	ISBN	Book-Title	Book-Author	Publisher
1	000220083X	AMPHIBIANS AND REPTILES	Trevor Beebee	Trafalgar Square
2	000225946X	Swimmer	Bill Broady	Flamingo
3	000255710X	The Danakil Diary: Journeys Through Abyssinia, 1930-34	Wilfred Thesiger	Harpercollins Canada
4	000470973X	Dogs Photo Guide (Collins Gem)	Jane Edie	HarperCollins UK
5	000614330X	The Grove of Eagles	Winston Graham	HarperCollins Publishers
6	000617616X	At Close Quarters	Gerald Seymour	HarperCollins Publishers
7	000617695X	The Great & Secret Show	Barker Clive	HarperCollins Publishers
8	000636988X	> How Not to Be a Perfect Mother: The Crafty Mother's Guide to a Quiet Life (...)	Libby Purves	HarperCollins Publishers
9	000637610X	You Got an Ology	Maureen Lipman	HarperCollins Publishers
10	000639194X	Firewing	Kenneth Oppel	HarperCollins Canada



3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

BẢNG USERS

```
# Đọc file Users
df_users = spark.read.format("csv") \
    .option("header", "true") \
    .option("delimiter", ";") \
    .option("encoding", "ISO-8859-1") \
    .load("dbfs:/FileStore/books/users.csv")
display(df_users)
```



	^A _C User-ID	^A _C Location	^A _C Age
1	1	nyc, new york, usa	NULL
2	2	stockton, california, usa	18
3	3	moscow, yukon territory, russia	NULL
4	4	porto, v.n.gaia, portugal	17
5	5	farnborough, hants, united kingdom	NULL
6	6	santa monica, california, usa	61
7	7	washington, dc, usa	NULL
8	8	timmins, ontario, canada	NULL
9	9	germantown, tennessee, usa	NULL
10	10	albacete, wisconsin, spain	26
11	11	melbourne, victoria, australia	14
12	12	fort bragg, california, usa	NULL
13	13	barcelona, barcelona, spain	26
14	14	mediapolis, iowa, usa	NULL
15	15	calgary, alberta, canada	NULL

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

XÓA DỮ LIỆU TRỐNG VÀ TRÙNG

▶ ▼ ✓ 06:39 PM (4s)

```
df_users = df_users.dropna(subset=["User-ID"])  
df_users = df_users.dropDuplicates(['User-ID'])
```


3. XỬ LÝ DỮ LIỆU

○○○○

CHUẨN HÓA CỘT

```
columns_to_process = ["Location"]
for col_name in columns_to_process:
    df_users = df_users.withColumn(col_name,upper(trim(regex_replace(col_name, r'^\x00-\x7F',''))))
display(df_users)
```

▶ (2) Spark Jobs

▶ df_users: pyspark.sql.dataframe.DataFrame = [User-ID: string, Location: string ... 1 more field]

Table ▾ +

	^A _C User-ID	^A _C Location	¹ ₃ Age
1	100010	EATON RAPIDS, MICHIGAN, USA	null
2	100014	SAN JOSE, CALIFORNIA, USA	25
3	100021	ALEXANDRIA, VIRGINIA, USA	null
4	100062	GLENSIDE, PENNSYLVANIA, USA	16
5	100070	CAIRO, ALASKA, ANGOLA	null
6	100090	LONDON, ENGLAND, UNITED KINGDOM	29

3. XỬ LÝ DỮ LIỆU

○○○○

TÁCH CỘT

▶ ▼ ✓ 06:39 PM (3s) 15

```
df_users = df_users.withColumn("Country", split(col("Location"), ",")[2]).drop('Location')
display(df_users)
```

▶ (2) Spark Jobs

▼ 📄 df_users: pyspark.sql.dataframe.DataFrame

User-ID: string
Age: integer
Country: string

	^A _C User-ID	^A _C Location	¹ ₃ Age
1	100010	EATON RAPIDS, MICHIGAN, USA	null
2	100014	SAN JOSE, CALIFORNIA, USA	25
3	100021	ALEXANDRIA, VIRGINIA, USA	null
4	100062	GLENSIDE, PENNSYLVANIA, USA	16
5	100070	CAIRO, ALASKA, ANGOLA	null




	^A _C User-ID	¹ ₃ Age	^A _C Country
1	100010	null	USA
2	100014	25	USA
3	100021	null	USA
4	100062	16	USA
5	100070	null	ANGOLA
6	100090	29	UNITED KINGDO...

3. XỬ LÝ DỮ LIỆU




XỬ LÝ DỮ LIỆU TRỐNG

```
df_users = df_users.replace("", "N/A", subset = "Country")  
display(df_users)
```

	^A _C User-ID	¹ ₃ Age	^A _C Country 
1	10114	23	
2	101146	79	
3	101527	null	
4	101837	21	
5	103436	35	
6	10436	null	
7	104458	null	



	^A _C User-ID	¹ ₃ Age	^A _C Country 
2941	107499	33	N/A
2942	108866	null	N/A
2943	109126	39	N/A
2944	109180	60	N/A
2945	10954	null	N/A

○ ○ ○ ○

A 15x4 grid of dots representing a 4x15 matrix. The dots are arranged in 4 columns and 15 rows, with some dots missing in the first two columns.

A 15x4 grid of blue dots. There are 4 columns and 15 rows of dots. Each column contains 15 dots, and each row contains 4 dots. The dots are arranged in a regular grid pattern.

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

XÓA CỘT

```
df_users = df_users.drop("Age")
```

► (2) Spark Jobs

▼  df_users: pyspark.sql.dataframe.DataFrame
User-ID: string
Country: string

Table ▼ +

	^A _C User-ID	^A _C Country
1	100010	USA
2	100014	USA
3	100021	USA
4	100029	Germany
5	100031	USA
6	100062	USA
7	100070	Angola

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

BẢNG USERS

	^B _C User-ID	^B _C Country
1	100010	USA
2	100014	USA
3	100021	USA
4	100029	Germany
5	100031	USA
6	100062	USA
7	100070	Angola
8	10009	Australia
9	100090	United Kingdom
10	10010	Belgium
11	100102	Indonesia
12	100116	USA
13	10012	Australia

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

BẢNG RATINGS

```
09:03 PM (2s)

# Đọc file Ratings
df_ratings = spark.read.format("csv") \
    .option("header", "true") \
    .option("delimiter", ";") \
    .load("dbfs:/FileStore/books/ratings.csv")
display(df_ratings)
```



df_ratings: pyspark.sql.dataframe.DataFrame

User-ID: string
ISBN: string
Book-Rating: string

Table ▾ +

	^A _C User-ID	^A _C ISBN	^A _C Book-Rating
1	276725	034545104X	0
2	276726	0155061224	5
3	276727	0446520802	0
4	276729	052165615X	3
5	276729	0521795028	6
6	276733	2080674722	0
7	276736	3257224281	8
8	276737	0600570967	6
9	276744	038550120X	7

3. XỬ LÝ DỮ LIỆU

○ ○ ○ ○

XÓA DỮ LIỆU TRÙNG



09:03 PM (6s)

26

```
df_ratings = df_ratings.dropDuplicates(['User-ID', 'ISBN'])  
display(df_ratings)
```

CHUYỂN ĐỔI KIỂU DỮ LIỆU



09:03 PM (<1s)

28

```
df_ratings = df_ratings.withColumn("Book-Rating", df_ratings["Book-Rating"].cast("int"))  
display(df_ratings)
```

▼ df_ratings: pyspark.sql.dataframe.DataFrame

User-ID: string

ISBN: string

Book-Rating: integer

○ ○ ○ ○

RÀNG BUỘC DỮ LIỆU

09:03 PM (<1s)

29

```
df_ratings = df_ratings.join(df_books, df_ratings.ISBN == df_books.ISBN, how="leftsemi") \
    .join(df_users, df_ratings["User-ID"] == df_users["User-ID"], how="leftsemi")
```

▼  df_ratings: pyspark.sql.dataframe.DataFrame

User-ID: string

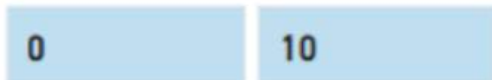
ISBN: string

Book-Rating: integer

4. PHÂN TÍCH KINH DOANH

BOOKS REVIEW REPORT

Average Ratings



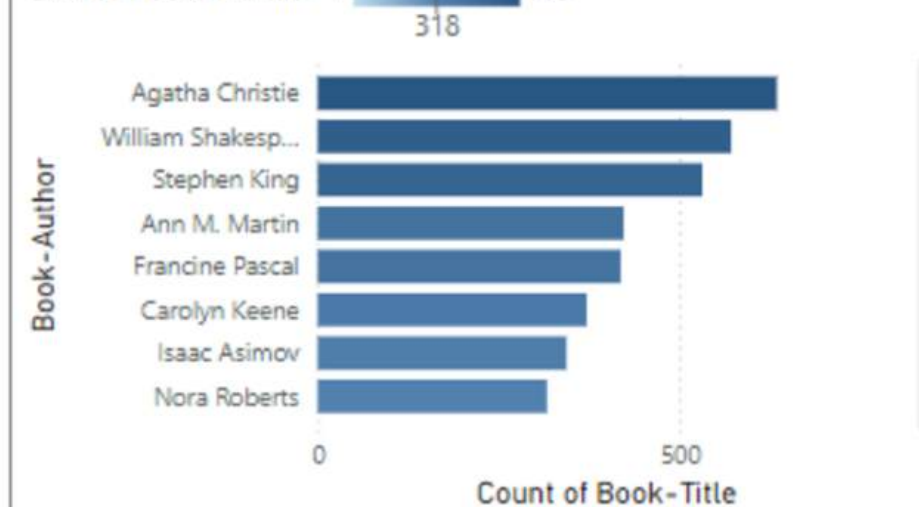
1.02M

Count of Book-Rating

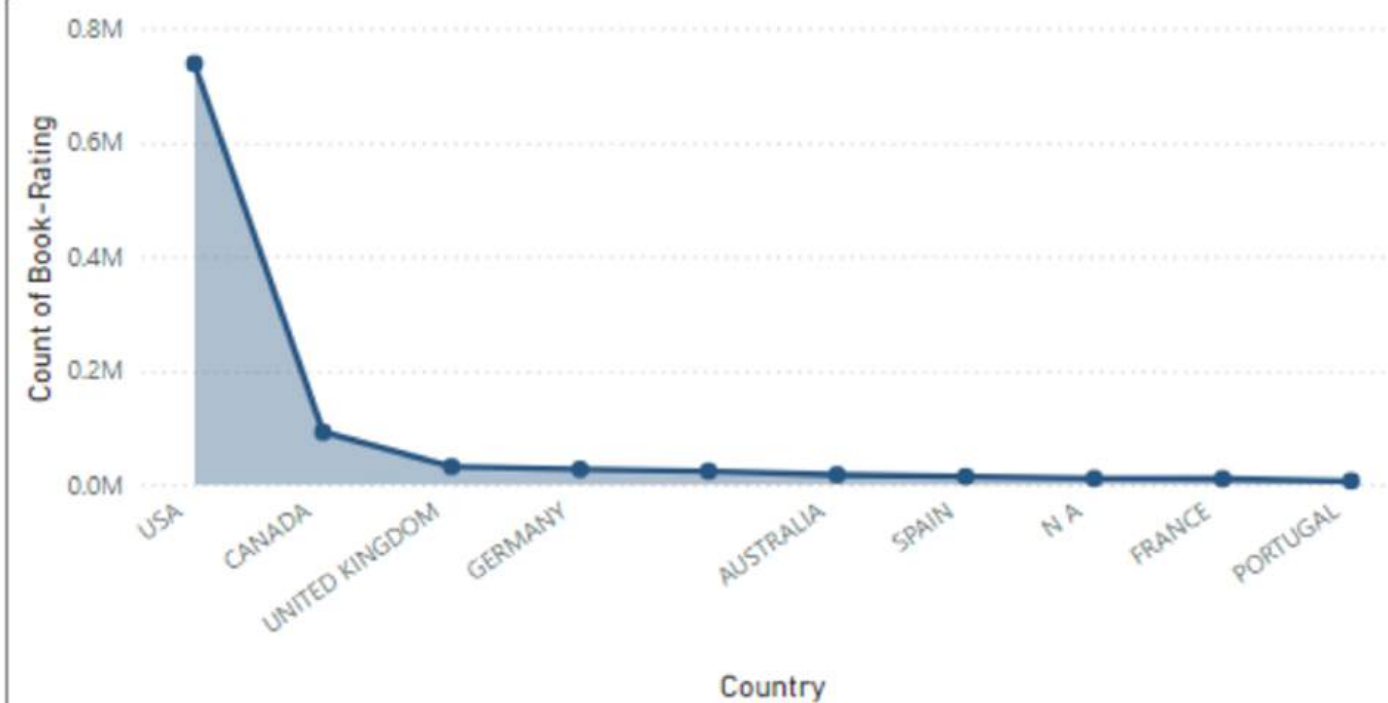
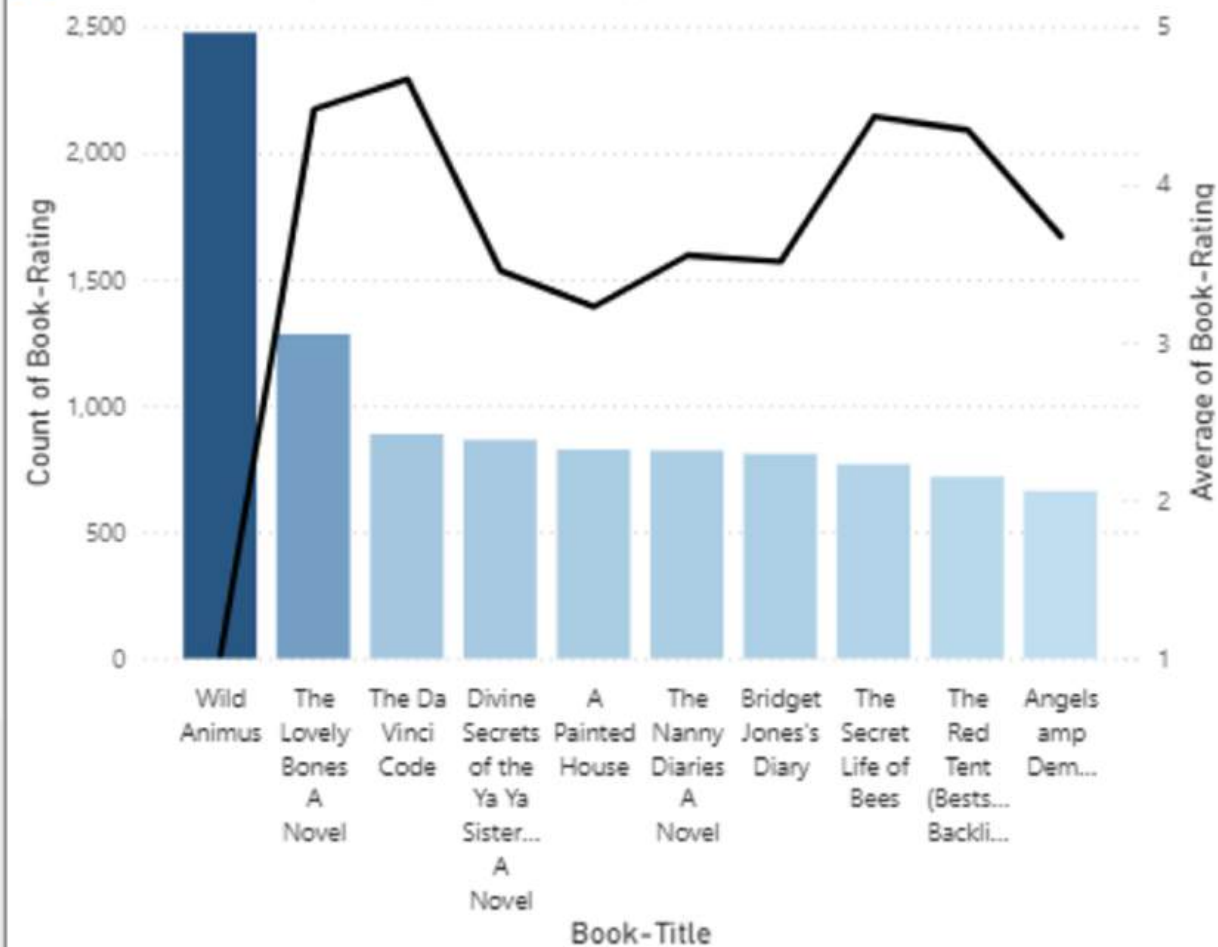
2.84

Average of Book-Rating

Count of Book-Author 1 635



Count of Book-Rating ● Average of Book-Rating



o o o o

5. KẾT QUẢ VỀ BẢN THÂN

Kỹ năng kỹ thuật

- Công cụ và nền tảng phân tích dữ liệu PySpark, DataBricks, và Azure
- Phân tích và trực quan hóa dữ liệu Power BI, hoặc Matplotlib, Seaborn trong Python

Kinh nghiệm thực tế

- Áp dụng các kỹ thuật phân tích dữ liệu vào các bài toán thực tế trong doanh nghiệp.
- Kinh nghiệm làm việc trong môi trường công nghiệp với áp lực và yêu cầu cao.
- Hiểu biết về các phương pháp và quy trình làm việc chuyên nghiệp trong ngành.



RESULTS

o o o o

o o o o

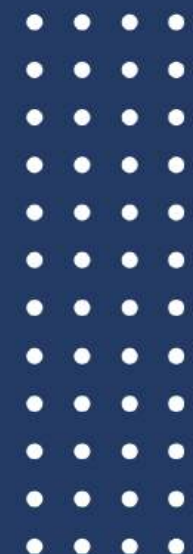
5. KẾT QUẢ VỀ DỰ ÁN

- Hoàn thành các yêu cầu và mục tiêu đã đề ra cho dự án xử lý và phân tích dữ liệu.
- Xây dựng được các báo cáo và dashboard cung cấp thông tin hỗ trợ quyết định kinh doanh cho nhà xuất bản sách.

A large, dark blue curved line sweeps across the right side of the slide. In the lower right, a magnifying glass with a silver rim is positioned over the word 'RESULTS'. The word is in a bold, red, sans-serif font. Below the word, four small white circles are arranged horizontally.

RESULTS

o o o o



THANK YOU

