DS 410 - Miniproject
Kakao

Dongyeon Kang, Jay Patel, Jiho Lee,
Jongmin Chung, Junho Lee, Seongeun Kim,
Sunwoo Kim, Yong Jun Choi

Book recommendations

PennState

Data science question and dataset acquisition.

Introduce the question you wish to answer. How did you decide to choose this problem?
→ Our project tries to address the challenge of personalizing book recommendations. The question arose from our collective interest in literature and the potential to leverage data analytics to enhance readers' experiences. We aim to develop a system that can recommend books to users in a way that enhances their reading satisfaction.

Introduce your dataset. How did you find it, how did you prepare it for analyses, how big is it, what challenges did you face in preparing it.
→ The dataset for our study is the Amazon Books Reviews dataset, obtained from Kaggle. This dataset, which is approximately 2.86GB, includes a wide range of user reviews spanning from 1996 to 2014. Preparing this dataset for analysis involved big preprocessing to manage its size and complexity, including cleaning the data and extracting relevant features for our analysis.

Talk about what you ultimately wish to do with this data in terms of modeling, analyses.
→ Our ultimate objective with this dataset is to create a recommendation system using machine learning techniques. We aim to provide genre-specific book suggestions tailored to each user. This work involves not only just predicting user preferences but also identifying trends within genres.

Requirement for ICDS

Why was this problem intractable for your local mode? How would ICDS Roar-Collab assist you?
→ Our project involves analyzing a huge Amazon book review dataset (2.86GB), which is too large and complex for ordinary computers. Running sophisticated machine learning algorithms like Alternating Least Squares (ALS) requires a lot of computing power. ICDS Roar-Collab gives us access to high-performance computing, allowing us to handle big datasets more effectively and run complicated models quicker than we could with just a regular computer.

What steps did you take to make this run on ICDS. Mention any unique challenges that you faced and how you overcame them.
→
Handling large dataset

Exploring the Dataset for Enhanced Book Recommendation

1.Research Objective

- Our team is focused on tailoring book recommendations to individual preferences, sparked by our passion for literature and data analytics. We are building a system to suggest books that not only fit user tastes but also elevate their reading experience

2.Data Overview

- We are utilizing the Amazon Books Reviews dataset from Kaggle, featuring 2.86GB of user reviews from 1996 to 2014. Our preparation process involved significant preprocessing to streamline the dataset, including data cleaning and feature extraction, to ready it for in-depth analysis.

2.Analytical Goals

- Our goal is to leverage machine learning to craft a recommendation engine. We strive to offer personalized genre-specific book recommendations, aiming to accurately predict user preferences and detect emerging trends across book genres.

Methods

Introduce your methods/modeling strategies and the rationale for using them

- Method Used: ALS(Alternative Least Squares)

- Rationale

- scalability : ALS can handle large datasets comprising of book reviews

- model: ALS performs as optimal method using matrix factorization to find out the pattern or relationship between items(book) and user feedback(review)

Introduction

Topic : Book recommendation system

1. **How did we decided to choose this problem**
2. **Primary goal of utilizing the dataset**
3. **What are we going to talk about**
    - Dataset acquisition
    - Data preprocessing
    - Requirement for ICDS
    - Methods
    - Results

Abstract / Motivation

- This project aims to develop a genre-specific book recommendation system using the Amazon book review dataset.

- Leveraging machine learning, particularly Alternating Least Squares (ALS), our system delivers personalized book recommendations tailored to user preferences.

- Our team is deeply passionate about reading and literature. Inspired by our coursework in data analytics and machine learning, we are excited to take on the challenge of creating a book recommendation system customized to the diverse range of literary genres available.

Dataset Acquisition

About our dataset

- Amazon Books Reviews dataset - Kaggle
- 2.86GB, includes a wide range of user reviews

Challenge

- Due to its volume, processing the dataset on local machine was inefficient and time-consuming

Data Preprocessing

- Sample Dataset Creation
- Data Merging
- Data Storage

Data Pre-processing

1. Sample Dataset Creation

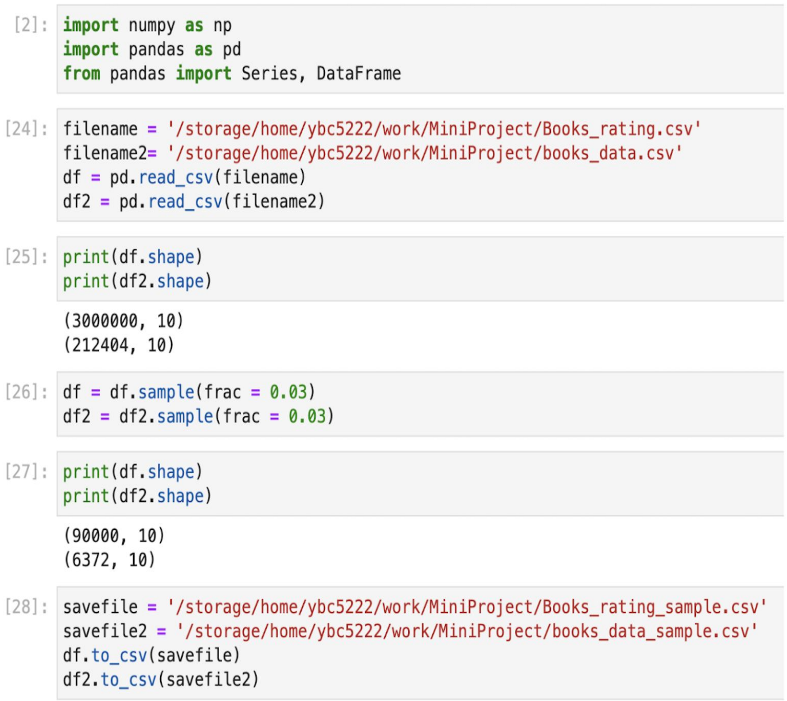
- Generated a 3% sample from the original dataset to streamline our development and test processing

2. Data Merging

- Merged two dataset by matching book titles to combine ratings with book data

3. Data Storage

- Saved out new dataset into CSV file for further analysis and model training



```
[2]:  import numpy as np
      import pandas as pd
      from pandas import Series, DataFrame
```

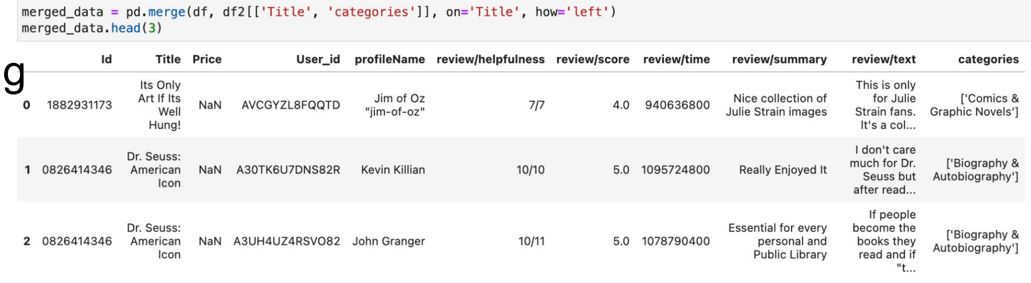
```
[24]: filename = '/storage/home/ybc5222/work/MiniProject/Books_rating.csv'
      filename2= '/storage/home/ybc5222/work/MiniProject/books_data.csv'
      df = pd.read_csv(filename)
      df2 = pd.read_csv(filename2)
```

```
[25]: print(df.shape)
      print(df2.shape)

      (3000000, 10)
      (212404, 10)
```

```
[26]: df = df.sample(frac = 0.03)
      df2 = df2.sample(frac = 0.03)
```

```
[27]: print(df.shape)
      print(df2.shape)

      (90000, 10)
      (6372, 10)
```

```
[28]: savefile = '/storage/home/ybc5222/work/MiniProject/Books_rating_sample.csv'
      savefile2 = '/storage/home/ybc5222/work/MiniProject/books_data_sample.csv'
      df.to_csv(savefile)
      df2.to_csv(savefile2)
```



```
merged_data = pd.merge(df, df2[['Title', 'categories']], on='Title', how='left')
merged_data.head(3)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text	categories
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...	['Comics & Graphic Novels']
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...	['Biography & Autobiography']
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...	['Biography & Autobiography']

Data Pre-processing (Cont’)

- **Data merging & sample data**
  : Merged the two datasets using a common key to enrich the raw rating data with additional contextual information.

- **Sample data**
  : Utilized sample dataset by creating a 10 percent random value from the full dataset to enhance efficiency and explore the data structure distribution and relationship for prototyping models

- **Data cleaning**
  : Removed null values in essential columns like Review/score, Title, and User_id to enhance the data quality

- **String Indexing**
  : User_id and Title are indexed to numeric values to be utilized for ALS

Methods

- Method Used: **ALS**(Alternative Least Squares)

- Rationale

- scalability : ALS can handle large datasets comprising of book reviews

- model: ALS performs as optimal method using matrix factorization to find out the pattern or relationship between items(book) and user feedback(review)

Methods

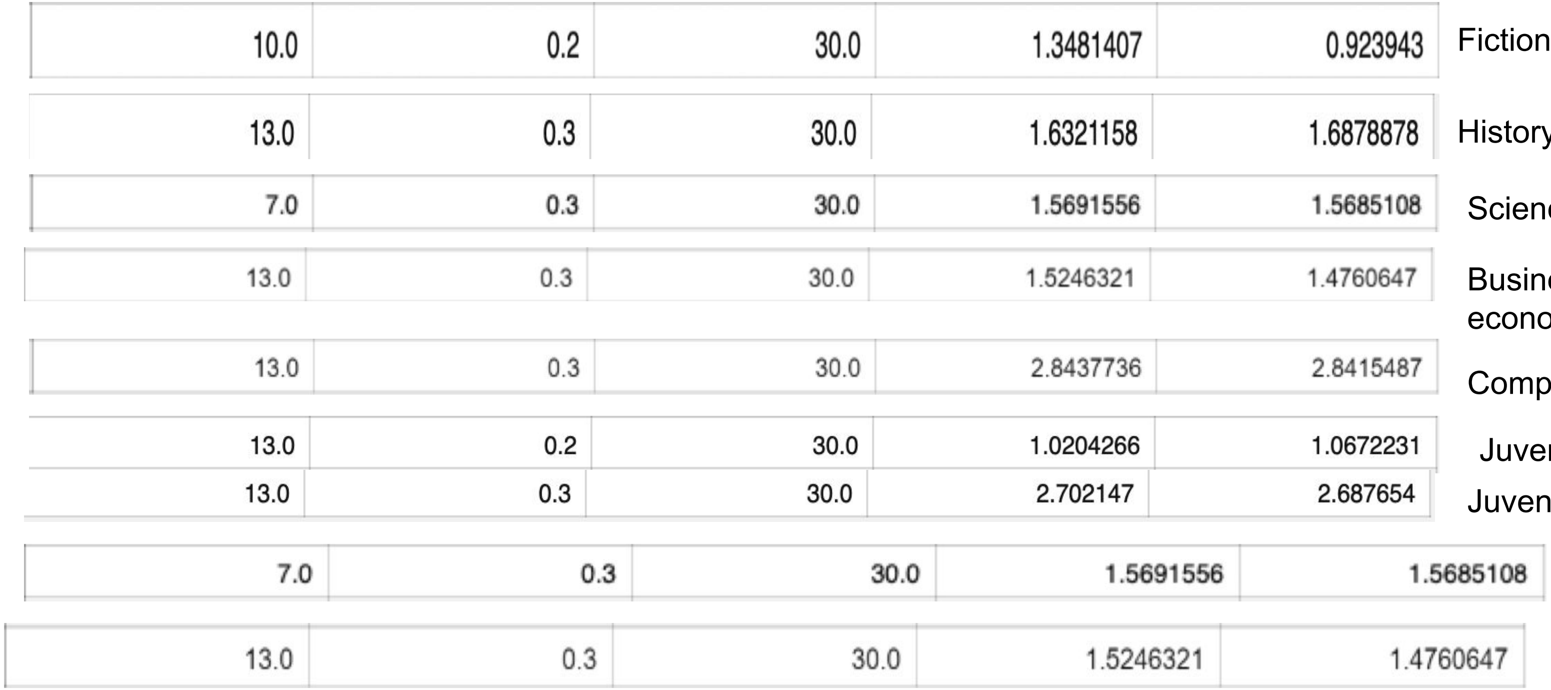
- Model Training : The ALS algorithm from the lecture is applied to train data

- Model Evaluation :
  - trained model is used to predict ratings for a validation dataset
  then, the predictions are compared to actual ratings to compute
  RMSE(Root Mean Square Error) to provide model accuracy

- Hyperparameter Tuning:
  - utilized various combinations of hyperparameters
  (rank, regularization, iterations)

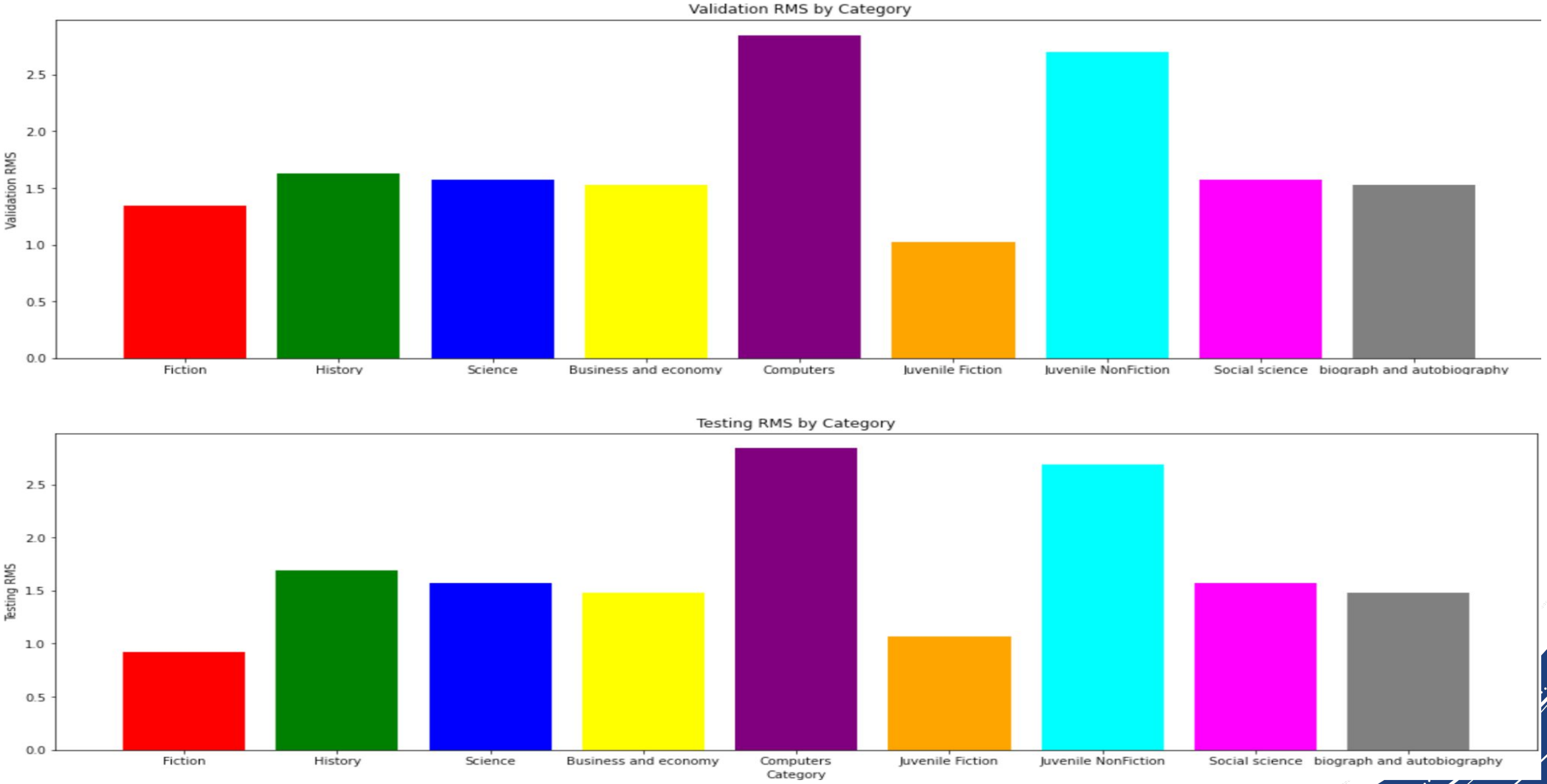
Results

Provide visualizations of your processed data, your models, your methods, anything that assists with answering your science question.

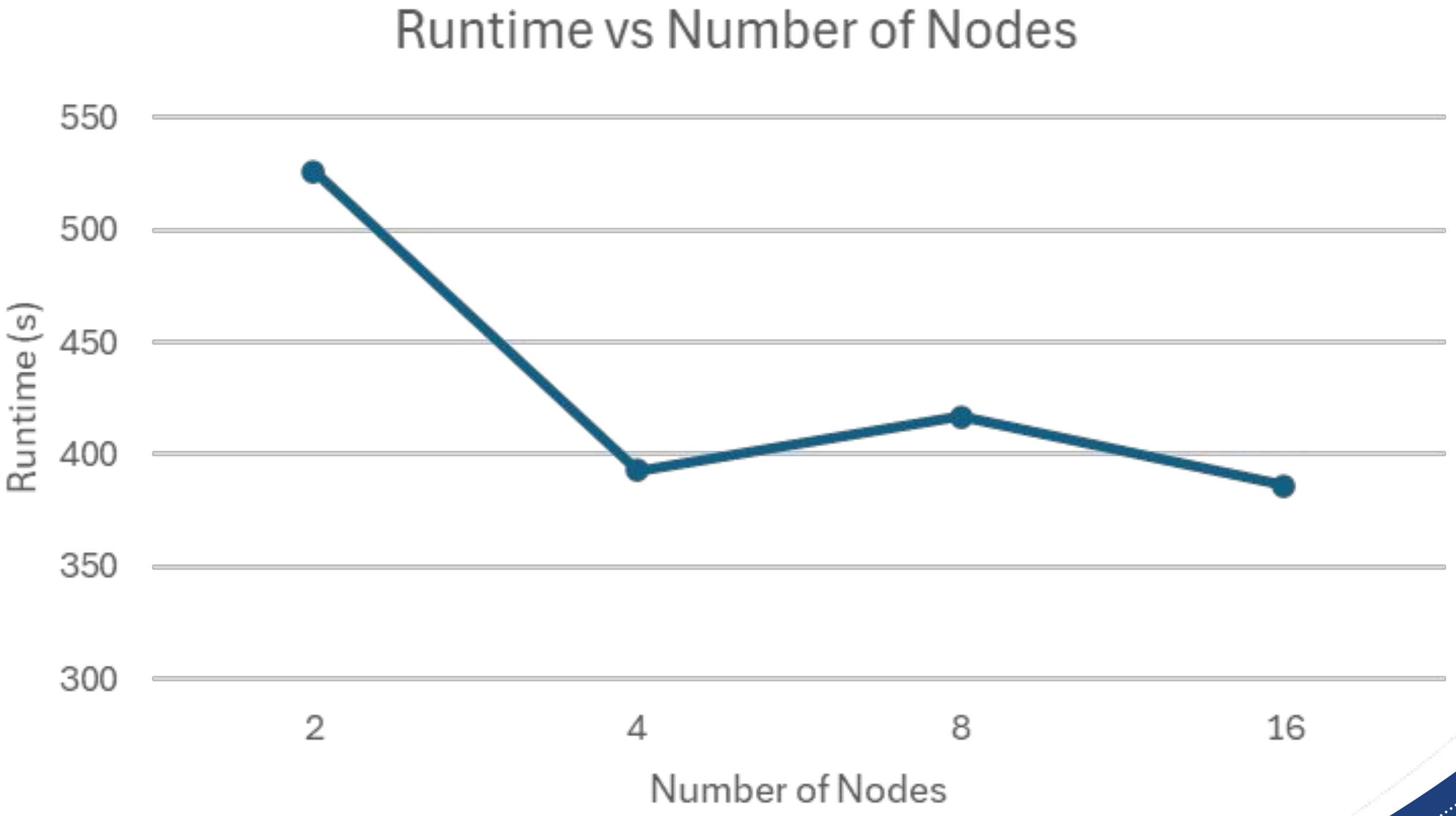


10.0	0.2	30.0	1.3481407	0.923943	Fiction
13.0	0.3	30.0	1.6321158	1.6878878	History
7.0	0.3	30.0	1.5691556	1.5685108	Science
13.0	0.3	30.0	1.5246321	1.4760647	Business and economy
13.0	0.3	30.0	2.8437736	2.8415487	Computers
13.0	0.2	30.0	1.0204266	1.0672231	Juvenile Fiction
13.0	0.3	30.0	2.702147	2.687654	Juvenile NonFiction
7.0	0.3	30.0	1.5691556	1.5685108	Social science
13.0	0.3	30.0	1.5246321	1.4760647	Biography & Autobiography

Results


Validation RMS by Category
Validation RMS
0.0
0.5
1.0
1.5
2.0
2.5
Fiction
History
Science
Business and economy
Computers
Juvenile Fiction
Juvenile NonFiction
Social science
biograph and autobiography
Testing RMS by Category
Testing RMS
0.0
0.5
1.0
1.5
2.0
2.5
Fiction
History
Science
Business and economy
Computers
Juvenile Fiction
Juvenile NonFiction
Social science
biograph and autobiography
Category

Results


Runtime vs Number of Nodes
550
500
450
400
350
300
Runtime (s)
2
4
8
16
Number of Nodes

Requirement for ICDS

Why ICDS?

- Local mode has memory limitation and time-consuming for handling this large dataset
- Offered enhanced processing power and improved data management

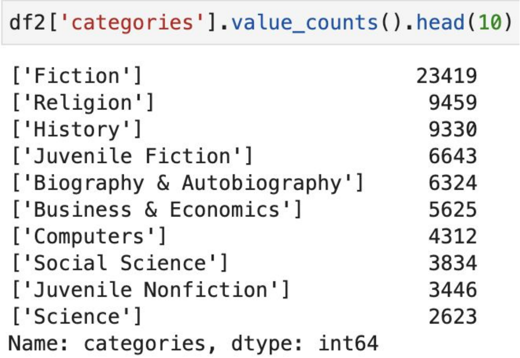
Encountered Errors:

- java.lang.OutOfMemoryError: Java heap space" indicates that the Java Virtual Machine (JVM) has run out of memory while executing the ALS algorithm.

- ALS algorithm matrix exception error

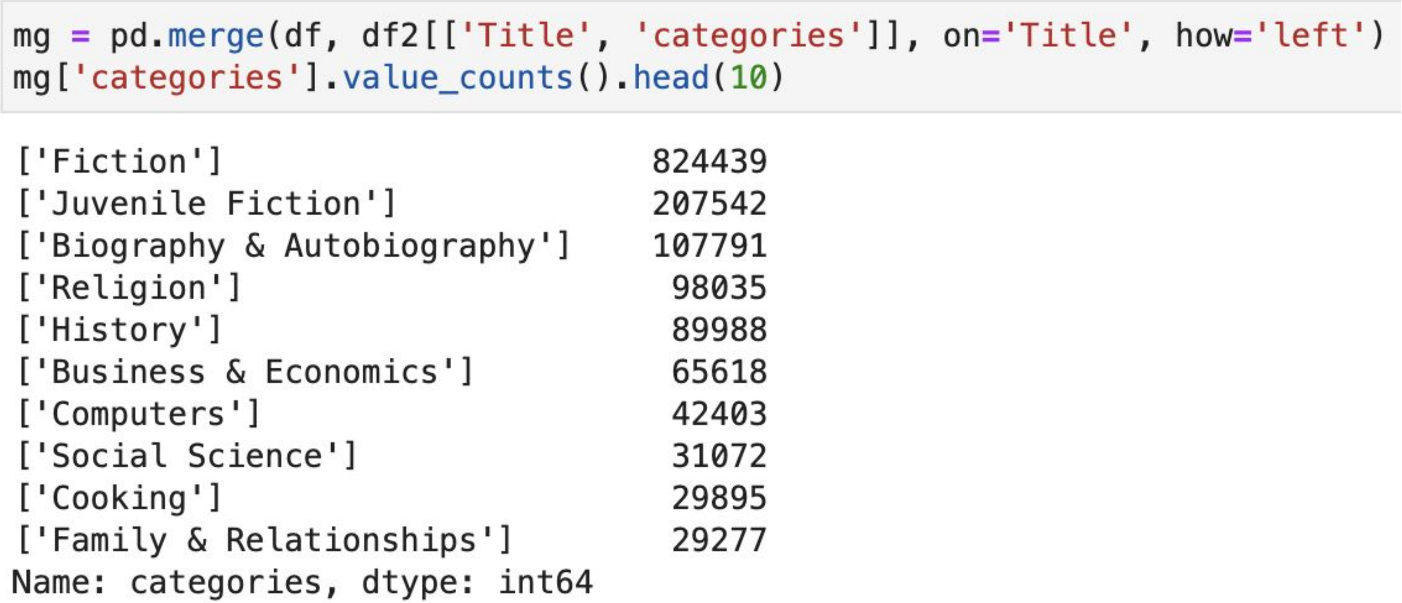
Reason : The dataset was too large when running locally or in cluster mode.

Requirement for ICDS



```
df2['categories'].value_counts().head(10)
```

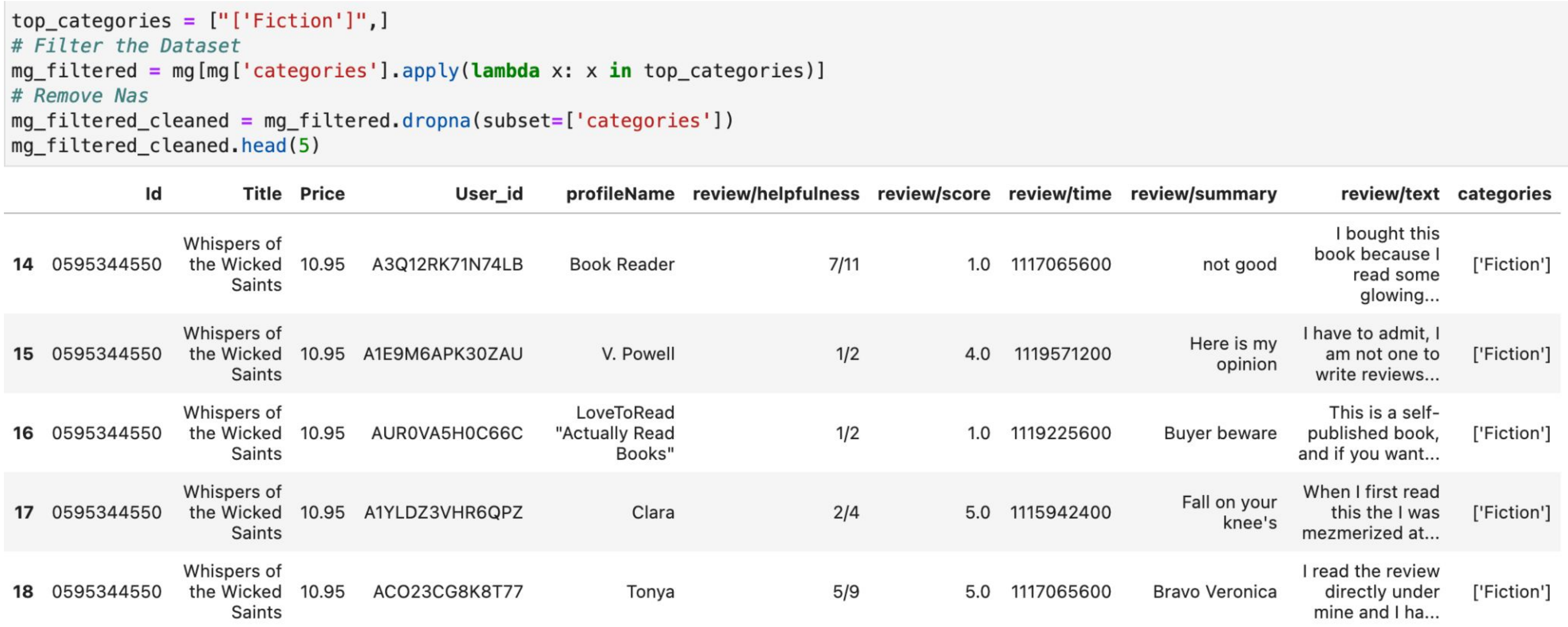
```
['Fiction']                           23419
['Religion']                           9459
['History']                            9330
['Juvenile Fiction']                   6643
['Biography & Autobiography']          6324
['Business & Economics']               5625
['Computers']                          4312
['Social Science']                     3834
['Juvenile Nonfiction']                3446
['Science']                            2623
Name: categories, dtype: int64
```



```
mg = pd.merge(df, df2[['Title', 'categories']], on='Title', how='left')
mg['categories'].value_counts().head(10)
```

```
['Fiction']                      824439
['Juvenile Fiction']             207542
['Biography & Autobiography']    107791
['Religion']                      98035
['History']                       89988
['Business & Economics']          65618
['Computers']                     42403
['Social Science']                31072
['Cooking']                       29895
['Family & Relationships']        29277
Name: categories, dtype: int64
```

Requirement for ICDS



```
top_categories = ["['Fiction']",]
# Filter the Dataset
mg_filtered = mg[mg['categories'].apply(lambda x: x in top_categories)]
# Remove Nas
mg_filtered_cleaned = mg_filtered.dropna(subset=['categories'])
mg_filtered_cleaned.head(5)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text	categories
14	0595344550	Whispers of the Wicked Saints	10.95	A3Q12RK71N74LB	Book Reader	7/11	1.0	1117065600	not good	I bought this book because I read some glowing...	['Fiction']
15	0595344550	Whispers of the Wicked Saints	10.95	A1E9M6APK30ZAU	V. Powell	1/2	4.0	1119571200	Here is my opinion	I have to admit, I am not one to write reviews...	['Fiction']
16	0595344550	Whispers of the Wicked Saints	10.95	AUR0VA5H0C66C	LoveToRead "Actually Read Books"	1/2	1.0	1119225600	Buyer beware	This is a self-published book, and if you want...	['Fiction']
17	0595344550	Whispers of the Wicked Saints	10.95	A1YLDZ3VHR6QPZ	Clara	2/4	5.0	1115942400	Fall on your knee's	When I first read this the I was mezmerized at...	['Fiction']
18	0595344550	Whispers of the Wicked Saints	10.95	ACO23CG8K8T77	Tonya	5/9	5.0	1117065600	Bravo Veronica	I read the review directly under mine and I ha...	['Fiction']

Requirement for ICDS

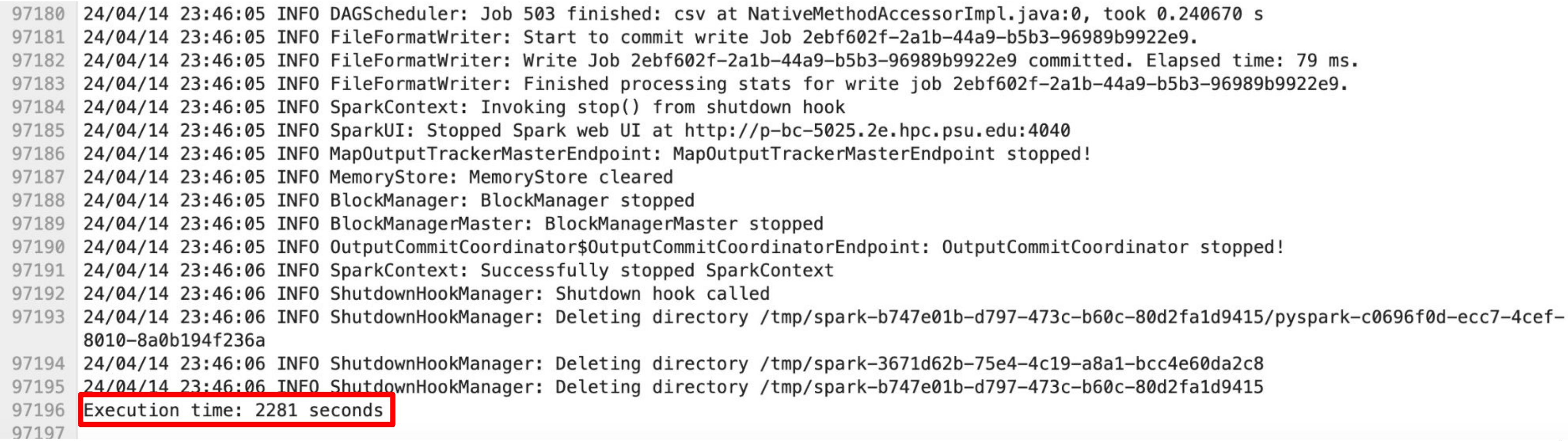
Adaptation to ICDS:

- ICDS clusters with high-memory nodes reduce processing time for computational intensive tasks
- partitioned the large dataset into clusters based on genres
- examined by focusing on the top 10 genres with the highest number of ratings

Result:

we were able to manage the dataset more efficiently and delve into the data without encountering issues while running in cluster mode

Requirement for ICDS



```
24/04/14 23:46:05 INFO DAGScheduler: Job 503 finished: csv at NativeMethodAccessorImpl.java:0, took 0.240670 s
24/04/14 23:46:05 INFO FileFormatWriter: Start to commit write Job 2ebf602f-2a1b-44a9-b5b3-96989b9922e9.
24/04/14 23:46:05 INFO FileFormatWriter: Write Job 2ebf602f-2a1b-44a9-b5b3-96989b9922e9 committed. Elapsed time: 79 ms.
24/04/14 23:46:05 INFO FileFormatWriter: Finished processing stats for write job 2ebf602f-2a1b-44a9-b5b3-96989b9922e9.
24/04/14 23:46:05 INFO SparkContext: Invoking stop() from shutdown hook
24/04/14 23:46:05 INFO SparkUI: Stopped Spark web UI at http://p-bc-5025.2e.hpc.psu.edu:4040
24/04/14 23:46:05 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/04/14 23:46:05 INFO MemoryStore: MemoryStore cleared
24/04/14 23:46:05 INFO BlockManager: BlockManager stopped
24/04/14 23:46:05 INFO BlockManagerMaster: BlockManagerMaster stopped
24/04/14 23:46:05 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/04/14 23:46:06 INFO SparkContext: Successfully stopped SparkContext
24/04/14 23:46:06 INFO ShutdownHookManager: Shutdown hook called
24/04/14 23:46:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-b747e01b-d797-473c-b60c-80d2fa1d9415/pyspark-c0696f0d-ecc7-4cef-8010-8a0b194f236a
24/04/14 23:46:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-3671d62b-75e4-4c19-a8a1-bcc4e60da2c8
24/04/14 23:46:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-b747e01b-d797-473c-b60c-80d2fa1d9415
Execution time: 2281 seconds
```

ICDS utilization

- Combined Usage overall : 30 + times
- Duration of Each Session : ~4 hours
- Number of Cores Used : ~ 30 CPU cores
- Memory Per Core : ~8 GB