Big data Course Project

By Himani Anil Deshpande

INTRODUCTION

As an avid reader and an ambitious engineer, I was always fascinated by the way any user data can help in predicting the future choices that we can make. I have been parts of Online book clubs and have been following reviews on good reads and Amazon Books, which regularly provides me with recommendations based on the books I have reviewed. So, as part of the Big data Project I decided to use this opportunity to build one for myself. In this project, I have focused on analyzing the data to find the Top/Best Authors, Publishers, Books, etc. I have also tried to build a user-based recommendation system using PySpark. The recommendation system is an algorithm based on Machine Learning concepts which suggests relevant items based on the history of the user's usage.

I am using the Book-Crossing Dataset [1] which was mined and collated by Cai-Nicolas Ziegler and provided to us by Institut für Informatik. The data set contains approximately 279k users without their personal information but identified by a unique id (i.e User-ID). The dataset is provided as 3 CSV files

- BX-Users: Defines details of a User and contains its Demographic data
- BX-Books: Defines the Book entity, which is uniquely identified by ISBN number and contains other information about books like the Title, author, Publisher, image URL’s, etc
- BX-Book-Ratings: Defines the ratings provided by the user to all the books it has reviewed.

BACKGROUND

My obsession with reading has been rooted in the mythological stories that my mother and grandmother used to tell me as a child. There were times when I as a teenager with no technological knowledge I used to fear that due to digital copies of Books in epub or rar files as well as a paper like feel of books through Kindle, I wouldn’t be able to feel or experience the smell of a new book.

Now, I understand that the written format nor the words will ever be lost. As an engineer, I understand the digital and audio formats of the books are an important way in which to preserve the written word. With the advent of e-commerce industries, the recommendation algorithms have not only improved the revenue but also the quality of user experience. Amazon provides recommendation based on our previous purchase and our browsing history. They also provide us the opportunity to improve our recommendation by excluding any purchases we have made in past.

Big data Course Project

By Himani Anil Deshpande

METHODODLOGIES

The data pipeline that I have followed

- Data Collection
- Data Storage
- Data Loading
- VPC Network
- Hadoop Cluster
- Data Preprocessing
- Data Visualization using Jupyter Notebook

The tools I have used:

- Google Cloud Platform
- Python
- Pyspark and PYSpark SQL
- Google Cloud Storage
- Virtual Private Cloud
- DataProc Cluster

Data Collection, Storage and Loading

As mentioned, I used the Book Crossing dataset from this link. I used the CSV Dump file for download, after unzipping I uploaded the 3 CSV files into Google Cloud storage.

The steps I followed creating a Project named hdeshpa-final-project in FA21-BL-INFO-I535 are

- Go to Navigation Menu
- Cloud Storage in Storage
- Browser
- Enable Cloud Storage API (it will pop up)
- Create Bucket
- Give the Bucket Name and Region

Upload the 3 CSV files using Upload Files and browse to location where you have kept the files

VPC(Virtual Private Cloud Network)

A Virtual Private Cloud (VPC) network is a virtual version of a physical network, implemented inside of Google's production network. I got error while creating a DataProc cluster, where I need to set firewall rules for VM-to-VM Communication.

Big data Course Project

By Himani Anil Deshpande

Using https://cloud.google.com/dataproc/docs/concepts/network which was given in the error warning, I set up the auto mode VPC network

The steps I followed

- Navigation Menu
- VPC Network in Pinned Section
- Click Create VPC Network
- Give Name as final-vpc-hdeshpa
- Click auto in Subnets
- Keep Dynamic routing rule as Regional
- Keep MTU as default value of 1460
- Click create

Instantiating Hadoop Cluster

I decided on using Cloud DataProc which manages the Hadoop and Spark Services and allows me to create, manage and shut down clusters quickly. Also gives me open-source tools like Machine learning, querying, etc.

I decided to use a single Master node with no worker nodes and resize if I get issues. Resize the cluster is one of the advantages of dataProc clusters and can be easily done.

The steps I followed:

- Navigation Menu
- DataProc in BigData Section
- Enable Cloud Dataproc API(it pops up)
- Click Create Cluster and give Cluster name as hdeshpa-final-cluster
- Keep Region (us-east1) same as the Cloud Storage bucket
- Keep Cluster Type as Single Node
- Enable Component Gateway and select Jupyter Notebook in Optional Components
- Series N1 and Machine Type ad n1standard-2 (2 vCPU and 7.5 GB Memeory)
- Customize the Cluster by Selecting the Primary network as the VPC network created(final-vpc-hdeshpa)
- Choose the Cloud storage staging Bucket as the bucket you created or let it create a new bucket. I choose to create a new Bucket
- Let Everything else be at default configurations
- Click Create

Data Preprocessing and Vizualizations

I needed to preprocess the data which I did using Pandas library in Python using Jupyter Notebook.

Once the Cluster is Up and Running, I followed the below steps :

Big data Course Project

By Himani Anil Deshpande

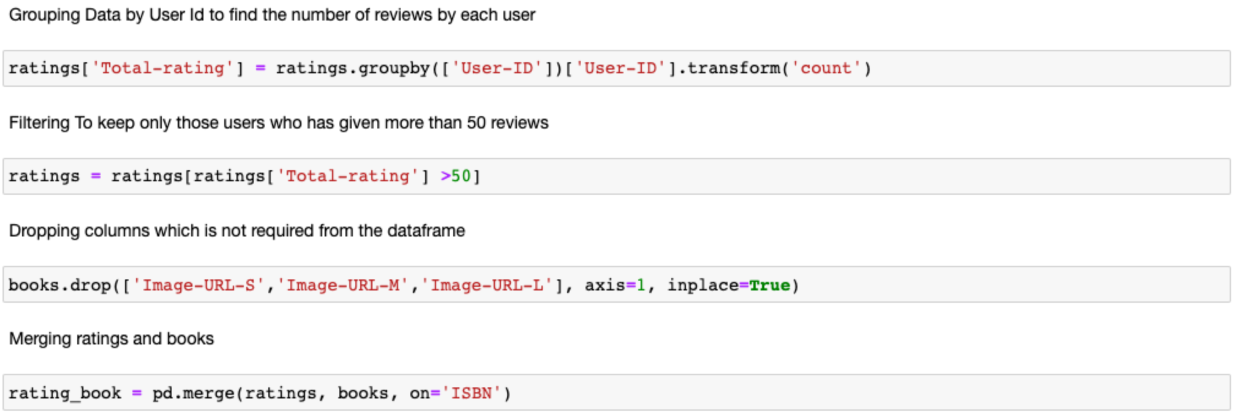
- Click on the Cluster name (hdeshpa-final-cluster)
- Go to Web Interfaces
- Click on Jupyter, which will open the Notebook in another notebook
- Create a folder IPYNB in Local Disk
- Click on New and Create a PySpark Notebook, it will open the notebook in new window

I have created 2 Jupyter notebooks

- ➔ Book Visualizations
- ➔ Book Recommendations

In Book Visualizations, I have used Pandas library, seaborn and Matplotlib graphics libraries for plotting the visualizations. I have performed the following preprocessing actions:

- ➔ Taken users who have given ratings/reviewed above 50 books
- ➔ In Year of Publication there are a lot of 0 values and also there is a non-numeric values like "DK Publishing Inc" and "Gallimard" which I have removed while getting the years when more than 2000 books were published.
- ➔ I have also not used 'Image-URL-S','Image-URL-M','Image-URL-L' columns while working in this book however, the raw data file will have these saved.
- ➔ Created two Columns one containing the number of reviews given by each user and the other contains the number of times a book was reviewed.
- ➔ Dropped Duplicates records after merging the datasets



Grouping Data by User Id to find the number of reviews by each user

```
ratings['Total-rating'] = ratings.groupby(['User-ID'])['User-ID'].transform('count')
```

Filtering To keep only those users who has given more than 50 reviews

```
ratings = ratings[ratings['Total-rating'] >50]
```

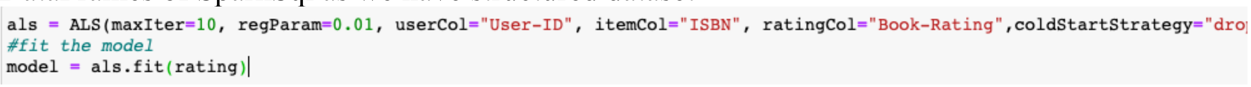
Dropping columns which is not required from the dataframe

```
books.drop(['Image-URL-S','Image-URL-M','Image-URL-L'], axis=1, inplace=True)
```

Merging ratings and books

```
rating_book = pd.merge(ratings, books, on='ISBN')
```

In Book Recommendations I have created a Collaborative Filtering Model to give us predictions. This Book uses PySpark Libraries to create a Session which is a combination of entry-way for the querying as well as other Spark functionalities. We have used abstractions like Datasets and DataFrames of SparkSql as we have structured dataset



```
als = ALS(maxIter=10, regParam=0.01, userCol="User-ID", itemCol="ISBN", ratingCol="Book-Rating",coldStartStrategy="dro
#fit the model
model = als.fit(rating)
```

Big data Course Project

By Himani Anil Deshpande

RESULTS

After preprocessing, I have analyzed the data to find

- Most reviewed books(Fig 1)
- Top 10 Books (Fig 2)
- Top 10 Publishers (Fig3)
- Top 10 Authors of all time(Fig 4)
- All the years when more than 2000 books were published(fig 5)
- Word Cloud of Book Titles(fig 6)
- Recommendations of books for user 276688 (Fig 7)

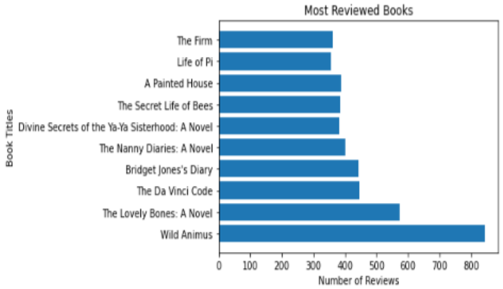


Fig 1

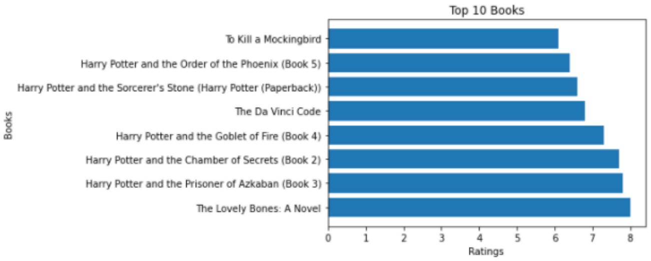


Fig 2

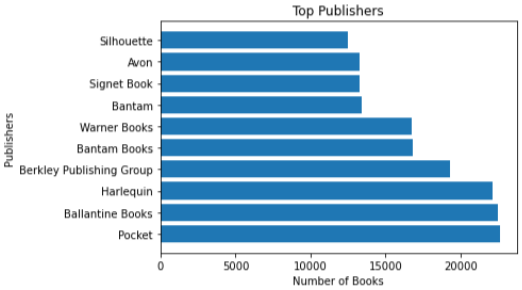


Fig 3

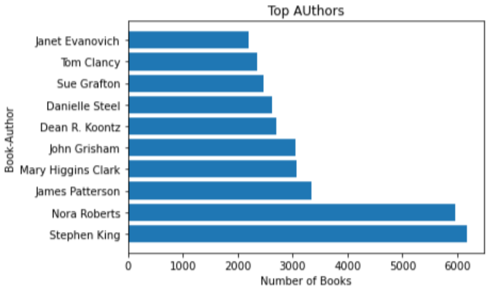


Fig 4

Big data Course Project

By Himani Anil Deshpande

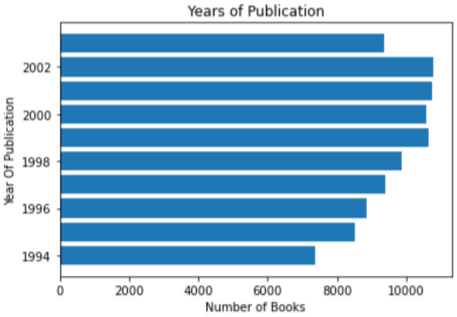


Fig 5

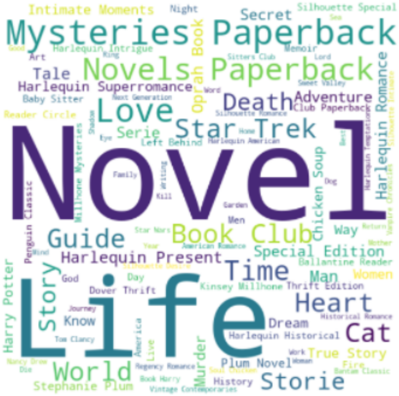
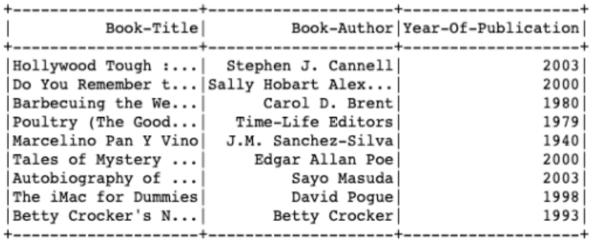


Fig 6



```
+--------------------+--------------------+-------------------+
|          Book-Title|         Book-Author|Year-Of-Publication|
+--------------------+--------------------+-------------------+
|Hollywood Tough :...|  Stephen J. Cannell|               2003|
|Do You Remember t...|Sally Hobart Alex...|               2000|
|Barbecuing the We...|      Carol D. Brent|               1980|
|Poultry (The Good...|   Time-Life Editors|               1979|
|Marcelino Pan Y Vino|  J.M. Sanchez-Silva|               1940|
|Tales of Mystery ...|     Edgar Allan Poe|               2000|
|Autobiography of ...|         Sayo Masuda|               2003|
|The iMac for Dummies|         David Pogue|               1998|
|Betty Crocker's N...|       Betty Crocker|               1993|
+--------------------+--------------------+-------------------+
```

Fig 7

Discussion(Interpreting the results)

Visualizations:

I have created 6 visualizations

➔ Most Reviewed Books

This shows that Wild Animus is one of the books which has been reviewed by 844 users which is followed closely by Lovely Bones: A Novel with 589 reviews. These are the two most popular reading choices according to most users.

➔ Top 10 Books:

The Lovely Bones: A Novel is the most liked novel by the users. It is quickly followed by Harry Potter Book 2 and 3. We can easily see that the Harry Potter in one of the most like Books across the world.

➔ Top 10 Authors:

Stephen King and Nora Roberts are the first and second most like authors in the world. Most of the users across the world have read these two authors. We can see other authors famous authors like Danielle Steel and John Grisham are also in this list.

➔ Top 10 Publishers:

Publishers like Pocket, Ballantine Books and Harlequin have published highest number of books.

➔ All the years when more than 2000 books were published

We can see that in the year of 2002, the highest number of books were published, followed by 2001. The years 1994-2003 produced the highest number of books.

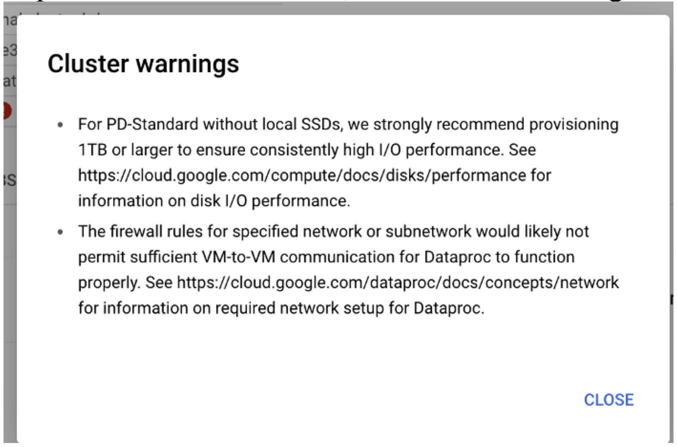
➔ WordCloud on Book Titles

Big data Course Project

By Himani Anil Deshpande

The word cloud of book titles shows that the title Novel, Life, Heart, Mysteries, paperback, Love have the highest importance in the book title. These are the most frequent important words used in the Book Title.

One of the major issues I faced was trying to launch a DataProc Cluster, when I researched on the issue, I found that I need to create a VPC (Virtual Private Cloud) to use it.
Then Another issue was to create a pandas Data Frame which wasn't reading the 3 CSV files from the cloud storage properly, on my research that I need to use encoding="latin1" as the dataset was encoded in Latin. Initially, I was going to use BigQuery for loading the datasets, However I got and issue while uploading the CSV files which was due to incorrect values like "" in place of Null values. So, I used a cloud storage instead.



Cluster warnings

- For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.
- The firewall rules for specified network or subnetwork would likely not permit sufficient VM-to-VM communication for Dataproc to function properly. See https://cloud.google.com/dataproc/docs/concepts/network for information on required network setup for Dataproc.

My previous experience with ML and Data Cleaning helped me while creating the book recommendation system and the Visualizations. Taking Inspiration from the course work I was able to create a Cloud storage, create a cluster with a VM and Virtual Private Cloud network, PySpark and Spark SQL for querying and PySpark MLlib library for creating a recommendation model. The model is based on collaborative filtering where conceptually two users have same tastes if they have similarly rated identical objects, this is performed by ALS i.e Alternating Least squares algorithm. The book recommendation that I created shows me the top 10 recommendations that the user Id "276688" can purchase/read. (Figure 7)

CONCLUSION

I have created a Book recommendation system which is important part of current e-commerce industry. Generally, any recommendation system/algorithm will prove to have a huge impact on revenue of the business as well as provide an easy user-experience. I was able to create an end-to-end pipeline of a Big Data Management using the Book Crossing dataset. I understand the importance of creating a pipeline to handle the data as well as the various datastores on the basis of which I need to store and analyze the data. It is critical to point out that with proper analysis and management of resources we can handle the Management of any Companies Data. I am confident that after this course I will be able to handle the role of a Data Manager.

REFERENCES:

- https://cloud.google.com/dataproc/docs/tutorials/jupyter-notebook
- https://cloud.google.com/dataproc/docs/concepts/components/jupyter#console
- https://cloud.google.com/vpc/docs/using-vpc
- https://cloud.google.com/dataproc/docs/troubleshooting
- https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network
- https://spark.apache.org/docs/1.1.0/mllib-collaborative-filtering.html
- https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.SparkSession.html
- http://www2.informatik.uni-freiburg.de/~cziegler/BX/

Citation:

1. Improving Recommendation Lists Through Topic Diversification, Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; Proceedings of the 14th International World Wide Web Conference (WWW '05), May 10-14, 2005, Chiba, Japan. To appear.