



NYU

Book Recommendation System

Atharva Moroney [amm9801]

Sachin Malepati [sm9449]

Vipul Goyal [vg2134]

Deepali Chugh [dc4600]

05.04.2022


Abstract

We used the UCSD GoodReads Dataset

Aim of the project -

- Gain valuable insights from the Data
- Find reading and publishing trends
- Develop a Book Recommendation System

Motivation

- Bibliophiles  Book Recommendation Systems.
- Such kind of systems are used by majority of book sellers.
- Book Recommendation Systems increase the visibility of Books.
- Better Customer Experience
- Insights and Analytics help understand the trends among readers.

Goodness

- To find the top-rated books overall, we only consider the books which have number of ratings greater than the average number of ratings
- We discard the interactions in which the books are present on the user's shelf but they have not read it yet
- Our trends match the trends published online

Data Sources

UCSD GoodReads Dataset: <https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home>

Books	Authors	Genres	Shelves	Book Reviews
Metadata and information about ~2.4 million books	Metadata about authors	Information about book genres	Complete user-book interaction map containing ~229 million interactions	on the books by various users ~15 million reviews
Format: JSON	Format: JSON	Format: JSON		Format: JSON
Size: 9.2 GB	Size: 105 MB	Size: 200 MB	Format: CSV	Size: 16.7 GB
			Size: 4.5 GB	
Total Size of the Dataset : 32 GB				

Data Sample: Books

Dataset contains the metadata for the Goodreads books.

Data profiling:

- Total books = ~9M
- Average description len = ~175 letters
- Average number of pages = ~183
- E-Books=~390K
- Genre-wise total books

```
{'isbn': '',  
  'text_reviews_count': '7',  
  'series': ['189911'],  
  'country_code': 'US',  
  'language_code': 'eng',  
  'popular_shelves': [{'count': '58', 'name': 'to-read'},  
                     {'count': '15', 'name': 'fantasy'}],  
  'asin': 'B00071IKUY',  
  'is_ebook': 'false',  
  'average_rating': '4.03',  
  'kindle_asin': '',  
  'similar_books': ['19997', '828466', '1569323', '425389', '1176674',  
                   '262740', '3743837'],  
  'description': 'Omnibus book club edition containing the Ladies of Madrigyn and  
  'format': 'Hardcover',  
  'link': 'https://www.goodreads.com/book/show/7327624-the-unschooled-wizard',  
  'authors': [{'author_id': '10333', 'role': ''}],  
  'publisher': 'Nelson Doubleday, Inc.',  
  'num_pages': '600',  
  'publication_day': '',  
  'isbn13': '',  
  'publication_month': '',  
  'edition_information': 'Book Club Edition',  
  'publication_year': '1987',  
  'url': 'https://www.goodreads.com/book/show/7327624-the-unschooled-wizard',  
  'image_url': 'https://images.gr-assets.com/books/1304100136m/7327624.jpg',  
  'book_id': '7327624',  
  'ratings_count': '140',  
  'work_id': '8948723',  
  'title': 'The Unschooled Wizard (Sun Wolf and Starhawk, #1-2)',  
  'title_without_series': 'The Unschooled Wizard (Sun Wolf and Starhawk, #1-2)'}
```

Data Sample: Authors

Dataset contains the metadata for the Goodreads Authors.

Data profiling:

Total Authors: ~ 830k

Number of Authors between the ratings:

- 0 to 1 : 32146
- 1 to 2 : 3824
- 2 to 3 : 31787
- 3 to 4 : 436220
- 4 to 5 : 325547

```
{'average_rating': '3.98',  
  'author_id': '604031',  
  'text_reviews_count': '7',  
  'name': 'Ronald J. Fields',  
  'ratings_count': '49'}
```

Data Sample: Genres

Dataset contains the metadata for the Goodreads Book Genres.

Data profiling:

- 10 unique genres
- Books with missing Genres: ~ 409k
- Total Records: ~2.3 Million

```
{'book_id': '7327624',  
  'genres': {'fantasy, paranormal': 31,  
             'fiction': 8,  
             'mystery, thriller, crime': 1,  
             'poetry': 1}}
```


Data Sample: Shelves

Dataset contains complete user-book interaction map

Data profiling:

- Total 3 files
- Total records = ~229 Million
- Unique books = ~2.4 Million
- Unique authors = ~836K

```
{'user_id': '8842281e1d1347389f2ab93d60773d4d',  
  'book_id': '25735618',  
  'review_id': 'ea74f2b6645b7d16f3ede2aca10226f0',  
  'is_read': True,  
  'rating': 0,  
  'date_added': 'Fri Aug 25 13:55:10 -0700 2017',  
  'date_updated': 'Tue Oct 17 23:53:44 -0700 2017',  
  'read_at': '',  
  'started_at': 'Tue Oct 17 09:23:10 -0700 2017'}
```

Data Sample: Reviews

Dataset contains the metadata for the Goodreads books.

Data profiling:

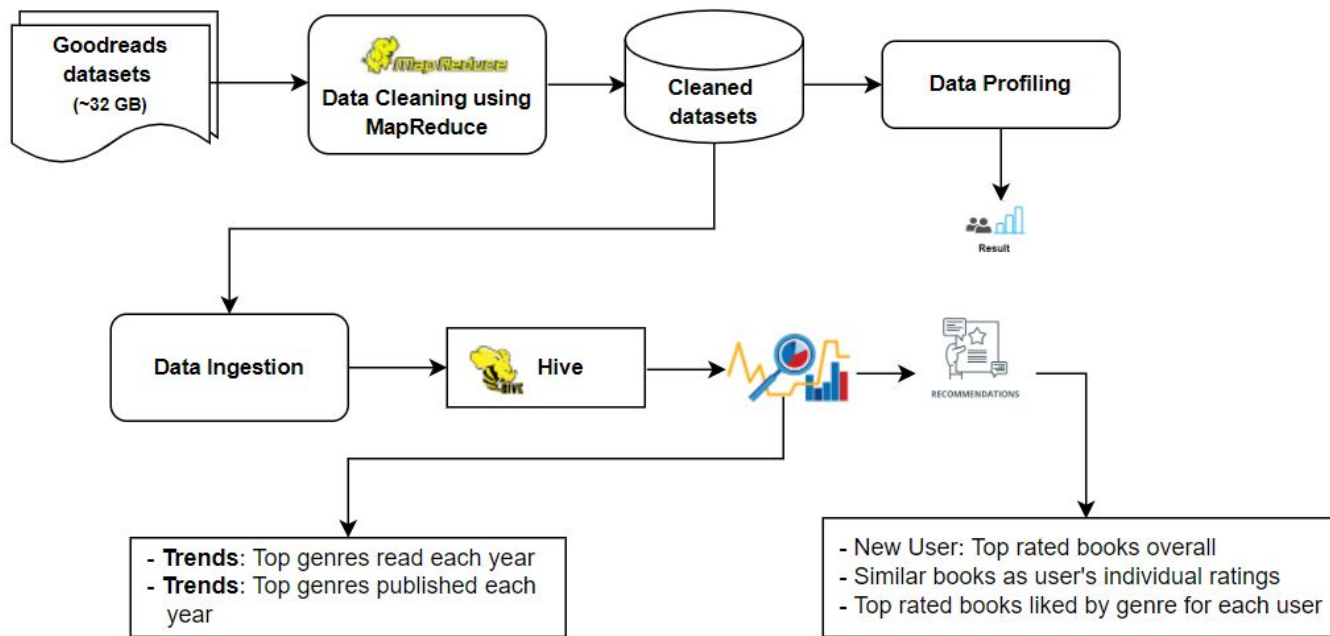
Total Records: ~15 Million

Number of unique users: 465k

Number of unique books reviewed: ~2M

```
{'user_id': '8842281e1d1347389f2ab93d60773d4d',  
  'book_id': '4986701',  
  'review_id': 'bb7de32f9fadc36627e61aaef7a93142',  
  'rating': 4,  
  'review_text': 'Found the Goodreads down image in this',  
  'date_added': 'Thu Aug 04 10:02:02 -0700 2011',  
  'date_updated': 'Thu Aug 04 10:02:02 -0700 2011',  
  'read_at': '',  
  'started_at': '',  
  'n_votes': 6,  
  'n_comments': 4}
```

Design Diagram



Coding Challenge 1

In some datasets, multiple levels of nested JSON objects were present inside the arrays.

Multiple MapReduce programs had to be written

```
public class BookReviewReducer
    extends Reducer<IntWritable, Text, NullWritable, Text> {

    @Override
    public void reduce(IntWritable key, Iterable<Text> values, Context context)
        throws IOException, InterruptedException {

        for (Text value : values) {
            Object obj = JSONValue.parse(value.toString());
            JSONObject data = (JSONObject)obj;
            String book_id = (String)data.get("book_id");
            JSONArray authors = (JSONArray) data.get("authors");

            Iterator<JSONObject> iterator = authors.iterator();

            while(iterator.hasNext()) {
                JSONObject obj2 = iterator.next();
                String author_id = (String) obj2.get("author_id");
                context.write(NullWritable.get(), new Text(book_id+","+author_id));
            }

            if(authors.size() != 0) {
                context.getCounter(Author.COUNT).increment(1);
            }
        }
    }
}
```

Coding Challenge 2

Multiple intermediate temporary tables and/or table joins had to be created because Hive doesn't support nested subqueries.

```
select distinct t4.book_id, t6.genre, t4.average_rating,
substr(t4.title, 1, 50) title from books t4 join (
  select t2.similar_book_id from similar_books t2 join (
    select t1.book_id from interactions t1
    where t1.user_id='f7fe5196ae6a346eb1c1e00a21d5693c' and t1.rating in (
      select max(t0.rating) from interactions t0
      where t0.user_id='f7fe5196ae6a346eb1c1e00a21d5693c')) t3
    on t2.book_id=t3.book_id ) t5 on t4.book_id=t5.similar_book_id
  join genres_new t6 on t4.book_id=t6.book_id
where ratings_count > 2133.88 order by t4.average_rating desc limit 20;
```

```
insert into reading_trends select t1.read_at_year, t2.genre, count(distinct t3.book_id, t3.user_id)
from reviews t1 join genres_new t2 on t1.book_id = t2.book_id
join interactions t3 on t1.book_id = t3.book_id
group by t1.read_at_year, t2.genre

select distinct t1.* from reading_trends t1 join (
select read_at_year c1, max(count) c2 from reading_trends group by read_at_year) t2
on t1.read_at_year = t2.c1 and t1.count = t2.c2 where t1.read_at_year >= 1960 and t1.read_at_year <= 2017;
```

Coding Challenge 3

In the books file, the 'similar_books' field contained book_ids separated by a string inside a JSON Array.

Separate Hive tables were created to handle such kind of fields.

```
'similar_books': ['19997', '828466', '1569323', '425389', '1176674', '262740', '3743837',  
'880461', '2292726', '1883810', '1808197', '625150', '1988046', '390170',  
'2620131', '383106', '1597281'],
```

```
for (Text value : values) {  
    Object obj = JSONValue.parse(value.toString());  
    JSONObject data = (JSONObject)obj;  
    String book_id = (String)data.get("book_id");  
    JSONArray similar_books = (JSONArray) data.get("similar_books");  
    Iterator<String> iterator = similar_books.iterator();  
    while(iterator.hasNext()) {  
        context.write(NullWritable.get(), new Text(book_id+","+iterator.next()));  
    }  
}
```

Results: Recommendation Type 1

New user recommendations: Top rated Books Overall

book_id	genre	average_rating	title
24812	comics	4.82	The Complete Calvin and Hobbes
11221285	fiction	4.78	The Way of Kings Part 2 (The Stormlight Archive #1.2)
8	fiction	4.77	Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5)
20343865	romance	4.77	Words of Radiance (The Stormlight Archive #2)
11543195	romance	4.77	Words of Radiance (The Stormlight Archive #2)
20150777	romance	4.77	Words of Radiance (The Stormlight Archive #2)
17332218	romance	4.77	Words of Radiance (The Stormlight Archive #2)
54741	comics	4.76	Toda Mafalda
27272698	fiction	4.76	Lodestar (Keeper of the Lost Cities #5)
95602	romance	4.76	Mark of the Lion Trilogy
5031805	history	4.76	ESV Study Bible
165068	children	4.75	The Jesus Storybook Bible: Every Story Whispers His Name
24814	comics	4.75	It's a Magical World: A Calvin and Hobbes Collection
6314759	romance	4.74	Harry Potter Boxset (Harry Potter #1-7)
28787784	romance	4.74	Harry Potter: The Complete Collection
11825646	romance	4.74	Percy Jackson Collection: Percy Jackson and the Lightning Thief the Last Olympi
862041	romance	4.74	Harry Potter Boxset (Harry Potter #1-7)
28787664	romance	4.74	Harry Potter: The Complete Collection (1-7)
70489	comics	4.74	There's Treasure Everywhere: A Calvin and Hobbes Collection
59715	comics	4.73	The Authoritative Calvin and Hobbes: A Calvin and Hobbes Treasury

20 rows selected (86.877 seconds)

Results: Recommendation Type 2

Books similar to the ones user has rated the highest

t4.book_id	t6.genre	t4.average_rating	title
17927395	romance	4.71	A Court of Mist and Fury (A Court of Thorns and Ro
27422533	romance	4.65	Wildfire (Hidden Legacy #3)
8062063	romance	4.63	Fullmetal Alchemist Vol. 24 (Fullmetal Alchemist
22299763	romance	4.62	Crooked Kingdom (Six of Crows #2)
9832370	fiction	4.59	BookRags Summary: A Storm of Swords
6585201	fiction	4.54	Changes (The Dresden Files #12)
12369942	romance	4.53	Endless (The Violet Eden Chapters #4)
12119529	romance	4.52	Magic Breaks (Kate Daniels #7)
1070527	comics	4.52	Avatar Volume 1: The Last Airbender (Avatar #1)
13061289	romance	4.5	Lying Season (Experiment in Terror #4)
7743175	romance	4.5	A Memory of Light (Wheel of Time #14)
11544421	romance	4.49	Magic Rises (Kate Daniels #6)
16164271	comics	4.49	Locke & Key Vol. 6: Alpha & Omega
13643021	romance	4.49	Pretty Guardian Sailor Moon Vol. 9 (Pretty Soldie
28862528	romance	4.49	Saga Vol. 6 (Saga #6)
17167166	romance	4.49	Crown of Midnight (Throne of Glass #2)
13605723	romance	4.49	Sentinel (Covenant #5)
17950614	romance	4.48	UnDivided (Unwind #4)
17333171	romance	4.47	Magic Shifts (Kate Daniels #8)
2767793	romance	4.46	The Hero of Ages (Mistborn #3)

20 rows selected (143.99 seconds)

Results: Recommendation Type 3

Top books in the genres most liked by the user

book_id	genre	average_rating	title
10042900	comics	5.0	Failure Incompetence: Aborted Jokes and Abandoned
10010348	history	5.0	Yankee Doodle Discord: A Walk with Planet Eris Thr
10041312	romance	5.0	Sunday Awakening
10042975	comics	5.0	Through The Wood Beneath The Moon
1002467	history	5.0	Queen's Mate: Three Women of Power in France on th
10094543	children	5.0	The Treasure-Hunt Three and Judge MIA's Decree
10093564	poetry	5.0	A Book of Verses
10021824	children	5.0	How To Do Everything
10137948	poetry	5.0	Indelible Marks
10085889	fiction	5.0	The Fun Room
9949000	non	5.0	Broadway Yearbook 1999-2000
10106581	fiction	5.0	Lope de Vega: Monster of Nature
9977815	poetry	5.0	In Confidence
1001463	fiction	5.0	Ru 486: Misconceptions Myths and Morals
10024738	history	5.0	Your Positive Potential: Action Steps for Self-Emp
10000294	children	5.0	The Hoopicopter
10000373	children	5.0	The Tablecloth
10105406	history	5.0	Stretch
9996906	fiction	5.0	I Figli dello Spazio
10000132	children	5.0	The Iron Chicken

20 rows selected (167.127 seconds)

Results: Recommendation Type 4

Top books in the genre - “Children”

book_id	genre	average_rating	_c3
165068	children	4.75	The Jesus Storybook Bible: Every Story Whispers Hi
13135293	children	4.68	Rangers Apprentice Bundle Books 1-8 (Ranger's Appr
10517686	children	4.67	One Direction: Forever Young: Our Official X Facto
8129	children	4.58	L.M. Montgomery's Anne of Green Gables
7846067	children	4.58	We are in a Book! (Elephant & Piggie #13)
8346300	children	4.55	Harry Potter: A Pop-Up Book: Based on the Film Phe
8319728	children	4.54	Beautiful Oops!
17290220	children	4.54	Rosie Revere Engineer
23497854	children	4.53	Island of Graves (Unwanteds #6)
181400	children	4.52	The Tale of Three Trees
397	children	4.52	The Gettysburg Address
967662	children	4.51	You Are Mine (Wemmicksville #2)
4732276	children	4.51	The Book Whisperer: Awakening the Inner Reader in
7869212	children	4.5	The Remarkable Soul of a Woman
24819508	children	4.49	Finding Winnie: The True Story of the World's Most
129909	children	4.49	The Boy Who Was Raised as a Dog: And Other Stories
452718	children	4.49	Disney's The Little Mermaid: Classic Storybook
99110	children	4.49	The Complete Tales and Poems of Winnie-the-Pooh (W
385250	children	4.49	The Jolly Postman or Other People's Letters
129511	children	4.49	Taking Charge of Your Fertility: The Definitive Gu

Results: Insights

Reading Trends

1996	fiction	6481301
1997	fiction	6848509
1998	fiction	7324989
1999	fiction	8062962
2000	fiction	8936125
2001	fiction	8884865
2002	fiction	9118705
2003	fiction	9949189
2004	fiction	10286284
2005	romance	11436770
2006	romance	13248651
2007	romance	16103676
2008	romance	20243010
2009	romance	23418511
2010	romance	27223387
2011	romance	31922249
2012	romance	37106771
2013	romance	41583074
2014	romance	44986169
2015	romance	47676002
2016	romance	48709995
2017	romance	47165184

Publishing Trends

1997	fiction	4839
1998	fiction	5536
1999	fiction	5971
2000	fiction	7044
2001	fiction	7410
2002	fiction	7928
2003	fiction	8780
2004	fiction	9203
2005	fiction	10611
2006	fiction	11827
2007	fiction	13791
2008	romance	16245
2009	romance	20745
2010	romance	26236
2011	romance	37348
2012	romance	51260
2013	romance	58870
2014	romance	55071
2015	romance	45623
2016	romance	39009
2017	romance	24087

Obstacles

- Experienced multiple failures during the data upload to HPC Cluster

Summary

- Processed **~32 GB** of data using MapReduce.
- Performed data **cleaning, profiling** and ingestion into Hive.
- Extracted **top-rated books** overall and genre-wise.
- Developed a Book Recommendation System giving **4 types of recommendations** to the users.
- Analysed reading and publishing **trends**.

References

- Mengting Wan, Julian McAuley, "[Item Recommendation on Monotonic Behavior Chains](#)", in RecSys'18.
- Mengting Wan, Rishabh Misra, Napa Nakashole, Julian McAuley, "[Fine-Grained Spoiler Detection from Large-Scale Review Corpora](#)", In, ACL'19.
- Dataset: <https://sites.google.com/enq.ucsd.edu/ucsdbookgraph/home>

Thank you!!!