**Requirements**

To create database for Books list available on Amazon.com and Goodreads. We used python web scraping techniques to pull data from Amazon and used available Goodreads csv file .

In Design Documents you will have

**Source File Specification**

1. File Name Specification – goodreads.csv
2. Online source – amazon.com

Detail Records will tell you what data you are getting from source, what data type, is it mandatory or not and the length of the column.

Goodreads.csv data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **bookID** | **title** | **authors** | **average\_rating** | **isbn** | **isbn13** | **language\_code** | **# num\_pages** | **ratings\_count** | **text\_reviews\_count** | **extra** |
| **1** | Harry Potter and the Half-Blood Prince (Harry Potter  #6) | J.K. Rowling-Mary GrandPré | 4.56 | 439785960 | 9780439785969 | eng | 652 | 1944099 | 26249 |  |
| **2** | Harry Potter and the Order of the Phoenix (Harry Potter  #5) | J.K. Rowling-Mary GrandPré | 4.49 | 439358078 | 9780439358071 | eng | 870 | 1996446 | 27613 |  |
| **3** | Harry Potter and the Sorcerer's Stone (Harry Potter  #1) | J.K. Rowling-Mary GrandPré | 4.47 | 439554934 | 9780439554930 | eng | 320 | 5629932 | 70390 |  |
| **4** | Harry Potter and the Chamber of Secrets (Harry Potter  #2) | J.K. Rowling | 4.41 | 439554896 | 9780439554893 | eng | 352 | 6267 | 272 |  |
| **5** | Harry Potter and the Prisoner of Azkaban (Harry | J.K. Rowling-Mary GrandPré | 4.55 | 043965548X | 9780439655484 | eng | 435 | 2149872 | 33964 |  |

REPORT THIS AD

REPORT THIS AD

The sample data from amazon.com

[{'title': 'Where the Crawdads Sing', 'rating': '4.8 out of 5 stars', 'author': 'Delia Owens', 'review count': 19379}, {'title': 'Disney: Storybook Collection Advent Calendar', 'rating': 0, 'author': 'IglooBooks', 'review count': 0}, {'title': 'The New Girl: A Novel (Gabriel Allon Book 19)', 'rating': '4.6 out of 5 stars', 'author': 'Daniel Silva', 'review count': 630}, {'title': 'Educated: A Memoir', 'rating': '4.7 out of 5 stars', 'author': 'Tara Westover', 'review count': 8741}, {'title': 'One Good Deed', 'rating': '4.1 out of 5 stars', 'author': 'David Baldacci', 'review count': 314}, {'title': "Summer of '69", 'rating': '4.3 out of 5 stars', 'author': 'Elin Hilderbrand', 'review count': 589}, {'title': 'The Best Thing', 'rating': '4.7 out of 5 stars', 'author': 'Mariana Zapata', 'review count': 293}, {'title': 'Dog Man: For Whom the Ball Rolls: From the Creator of Captain Underpants (Dog Man #7)', 'rating': '4.8 out of 5 stars', 'author': 'Dav Pilkey', 'review count': 58}, {'title': 'Under Currents: A Novel', 'rating': '4.5 out of 5 stars', 'author': 'Nora Roberts', 'review count': 567}, {'title': 'The Pigeon HAS to Go to School!', 'rating': '4.8 out of 5 stars', 'author': 'Mo Willems', 'review count': 103}, {'title': 'Here To Stay', 'rating': '4.2 out of 5 stars', 'author': 'Mark Edwards', 'review count': 432}, {'title': 'What You Did', 'rating': '4.3 out of 5 stars', 'author': 'Claire McGowan', 'review count': 1061},

Now using the ETL process now we loaded the data into Staging Tables. We had to discard some information which was there in the goodreads.csv to have what we require.

Amazon.com had lots of information available but we only require very few information as part of this project. So we only scraped what we require for the project like Title, author, rating and review count. Also we had some null values from the pages and we replaced it with “N/A” .

**Staging Table Load**

Goodreads :

Amazon :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Book\_id** | Author | Rating | Review Count | Title |
| **0** | Delia Owens | 4.8 out of 5 stars | 19379 | Where the Crawdads Sing |
| **1** | IglooBooks | N/A | N/A | Disney: Storybook Collection Advent Calendar |
| **2** | Daniel Silva | 4.6 out of 5 stars | 630 | The New Girl: A Novel (Gabriel Allon Book 19) |
| **3** | Tara Westover | 4.7 out of 5 stars | 8741 | Educated: A Memoir |
| **4** | David Baldacci | 4.1 out of 5 stars | 314 | One Good Deed |
| **5** | Elin Hilderbrand | 4.3 out of 5 stars | 589 | Summer of '69 |

Now our Staging Data is ready, and we need to load them into corresponding Target Tables.

We have used Postgres as database. The tables are: all\_books and goodredas

.