Final Report

By Beaver Win Team

Team Members

1. Tenghuan Li

   liten@oregonstate.edu

2. Guangyu Zhang

   zhangg3@oregonstate.edu

3. Abdullah Alsalman

   alsalma2@oregonstate.edu

4. Thekra Alrabiah

   alrabiat@oregonstate.edu

5. Hongjin Wang

   wanghong@oregonstate.edu

6. Hui Xu

   xuh2@oregonstate.edu

CS 540 / Winter 2020

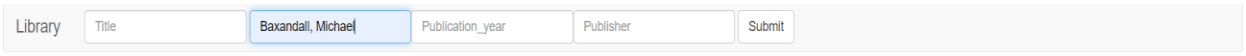
Database Management Systems

Book Information Management System Project

Project Description

The main objective of this project is to provide a shared utilization environment to manage a books database. This project gives us complete information about the storage books database. We can enter the record of new books and retrieve the details of books available in the dataset. Throughout the project, the focus has been on making the database organized and easy to use. The project is very useful for those who want to know about the online Library System as an example.

We collected enough datasets from 20 resources for our project which consists of over 30,000,000 records/tuples. From these databases, we decided the structure of the data and we organized it to have a standard format. Moreover, we have chosen to work on python as a language and frameworks to build our database's project. Therefore, we design the front-end as the initial design.



Title	Author	Publication Year	Publisher	source
German wood statuettes 1500-1800.	Baxandall, Michael	1967.	H.M.S.O.,	Greenglass_Storage_Monographs
German wood statuettes 1500-1800.	Baxandall, Michael	1967.	H.M.S.O.,	amazon_com_extras BX-Books
German wood statuettes 1500-1800.	Baxandall, Michael	1967.	H.M.S.O.,	bookpack_list1365
Painting and experience in fifteenth century Italy	Baxandall, Michael	1974.	Oxford University Press,	Greenglass_Storage_Monographs
Painting and experience in fifteenth century Italy : a primer in the social history of pictorial style / Michael Baxandall.	Baxandall, Michael	1974.	Oxford University Press,	amazon_com_extras BX-Books
Painting and experience in fifteenth century Italy : a primer in the social history of pictorial style / Michael Baxandall.	Baxandall, Michael	1974.	Oxford University Press,	Classroom Library 1286
Shadows and enlightenment / Michael Baxandall.	Baxandall, Michael	c1995.	Yale University Press,	Greenglass_Storage_Monographs
Shadows and enlightenment / Michael Baxandall.	Baxandall, Michael	c1995.	Yale University Press,	bookpack_list1365
The limewood sculptors of Renaissance Germany / Mi	Baxandall, Michael	1980.	Yale University Press,	Greenglass_Storage_Monographs
The limewood sculptors of Renaissance Germany / Michael Baxandall.	Baxandall, Michael	1980.	Yale University Press,	bookpack_list1365
The limewood sculptors of Renaissance Germany / Michael Baxandall.	Baxandall, Michael	1980.	Yale University Press,	Classroom Library 1286

Problems Statement

In the Book Information Management System Project, we faced some challenges that we mentioned in the midterm report and we successfully overcame these challenges. First, we used 20 different sources to follow the project requirements and we used over 30,000,000 tuples which are all related for books information. However, working with a big database, it was hard to deal with massive data in a short time. In our case, we spent a long time opening or importing a database. Therefore, we got access permission to the school server in class and was helpful because the processing speed on that was good enough for us.

The second challenge that we faced is that we worked to join and filter what attributes are useful in our datasets. For instance, one source may contain several tables which are books.csv, author.csv, publisher.csv. However, another source has only one table contains all attributes in one table. In this example, if we delete one attribute, we will lose the functional dependency. Otherwise, we cannot get one of the useful attributes. Therefore, it is important to be conscious to do join operations.

Next, the integration of unity. In 20 sources, there are some same tuples but with a different author, book id, publication_year, publisher_name, source_num, title. It is difficult to have a unity view or table to delete the duplicate tuples. Also, some ids are not continuous in the table. The problem is the integration among different tables. In addition, some similar tuples have the same book_title, but the publish years are different. A guess is that these books are different edition versions. Therefore, we can only take one of them.

The last thing was the null value or lost value processing. Some tuples have null values in non-primary attributes. And some non-English characters (the main part is the title)are lost during import. The problem is how to find these tuples and delete them.

To solve these issues, we divided the database to thirty files and each one of the groups uploaded five files after organizing its information. We used PHPMyAdmin server in tools.engr.oregonstate.edu website to do our work and it was helpful especially with deleting the duplicate tuples. In addition, We have read some papers or blogs about clean big data, and integrate data. We collaborated to use some tools to help us clean those dirty data.

Project Implementation

In implementing the whole system, it mainly uses the MVC design pattern to design the entire system. MVC design pattern can separate an application into three main logical components: the model, the view, and the controller. Each of these components is built to handle specific development aspects of an application. MVC is one of the most frequently used industry-standard web development frameworks to create scalable and extensible projects. In the term of View, it uses Html, CSS, and CSS library named Bootstrap to construct the whole of the front-end. Bootstrap is the world's most popular framework for building responsive, mobile-first sites, with BootstrapCDN. There are many good-looking components for us to call. We mainly adopt headers, forms, and tables from bootstrap to construct the whole of the frond-end. In the Model and Controller, it adopts Python and Django. Django is a high-level Python web framework that enables rapid development of secure and maintainable websites, It helps us interact with the front end, and manipulates the data returned by the front end to generate a table, and presents the data on the table. Django also connects our backend with the database. MYSQL Database is used as a database as it easy to maintain and retrieve records by simple queries. For the project features, we add some features to make the project easier. This feature allows you to

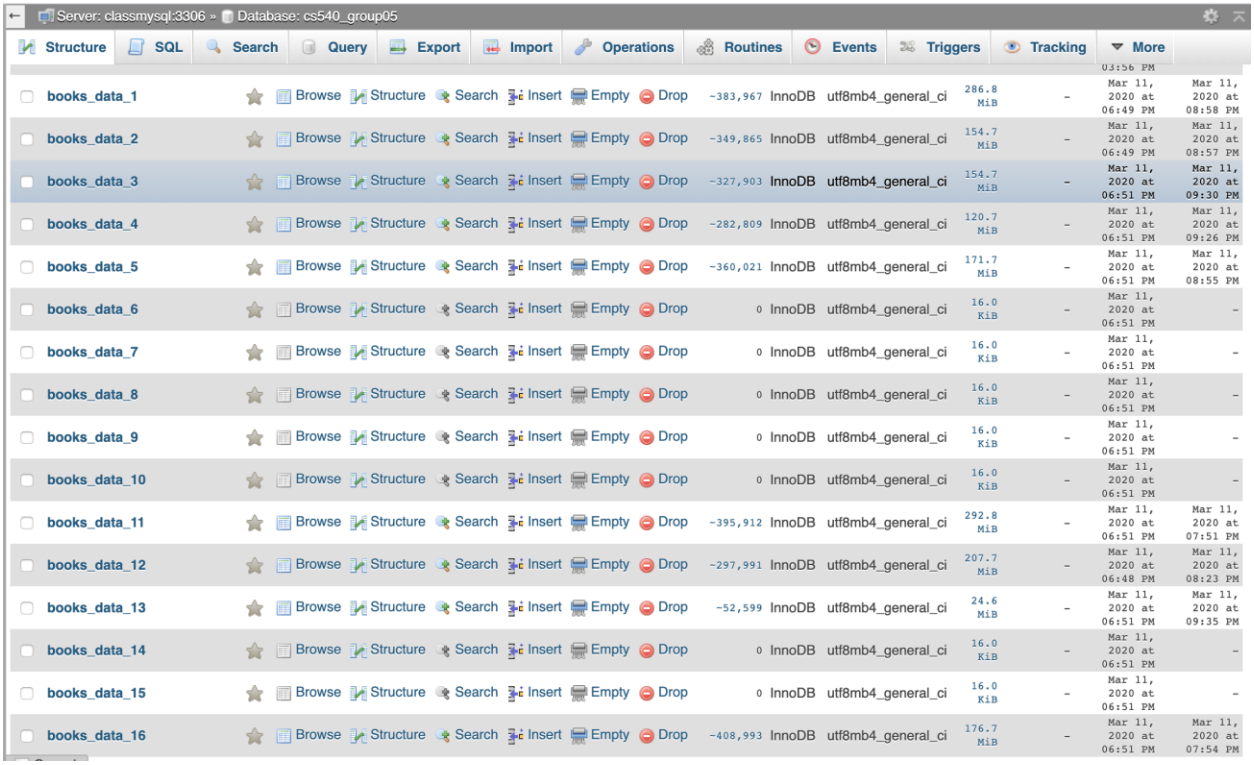
find books. The users can search for books based on book id, book name, publication or by author name. Thus, the database is unable to not allow two books having the same book id.

Data Set

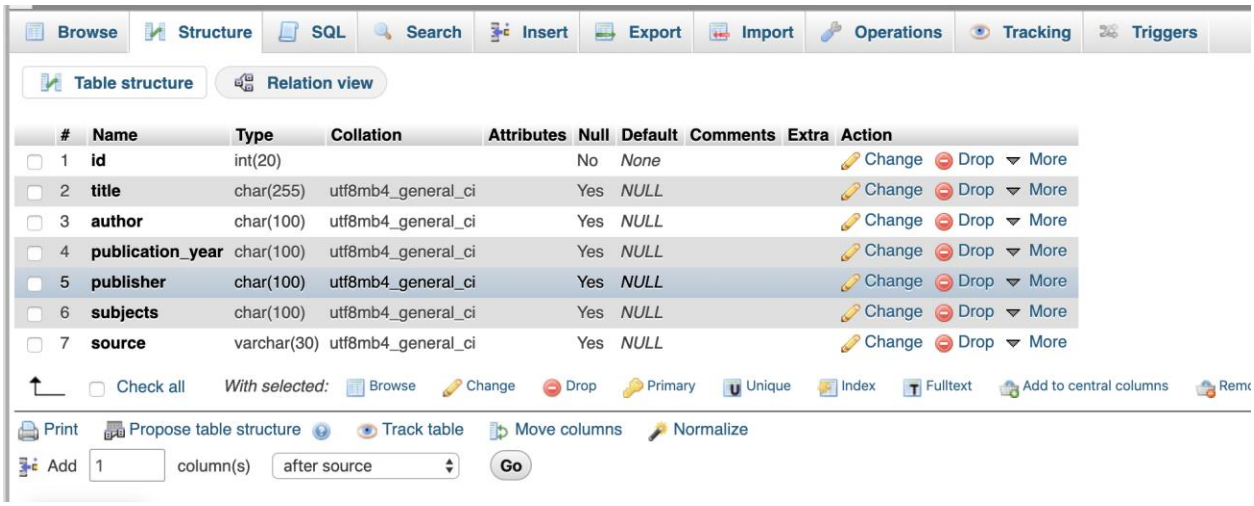
For the whole of the dataset, it collects data from 20 different sources, including the Amazon data set, OSU valley library, data world, and kindle dataset, and over 30,000,000 tuples which are all related to books information. There are many useless attributes and duplicates in a source, so it only integrates the id, title, author, publication_year, publisher_name, source_name, title. We spilt the big table into smaller tables (smaller than 500M) in order to import to the server database. Then, we filter the datasets. Keep the useful data, delete useless attributes. Next, we create tables on server, and design the relationship and dependency of each table. Last, we import the datasets to server. For instance, the column 'id' is the primary key which can determine each tuple in tables. For the global view, we process the datasets in advance, and make sure every table has same column so that we can just union all the tables together in this step. We wrote a python tool to delete useful columns and to delete duplicates of datasets by MySQL. We build scheme mapping for all sources, and we built an entity match for all resources. For example, "publication_year", "publicationyear", and "publication_date" can map each other. In the end, we add a new attribute "source" which can mark the source of the data.

System Design

We designed various tables to cope with the information and the screenshots below show our work to filter the database that we've collected. Screenshot that showing the tables



Screenshot that Showing table's columns



Screenshot that showing rows in a database

3296803	A legacy of spies / John le Carré.	Le Carré, John, 1931	[2017]	Viking,	Smiley George Fictit	Greenglass_Storage_Monographs
3296801	Secrets in death / J.D. Robb.	Robb, J. D., 1950-	2017.	St. Martin's Press,	Dallas Eve Fictitiou	Greenglass_Storage_Monographs
3296799	Dark legacy / Christine Feehan.	Feehan, Christine	2017.	Berkley,	Vampires Fiction, Pa	Greenglass_Storage_Monographs
3296721	The International student of the world problem of	*NULL*	[1936-1938]	American Issue Pub. Co.,	Youth Alcohol use Pe	Greenglass_Storage_Monographs
3296720	The International student of alcoholic drink in li	*NULL*	[1938-1939]	American Issue Pub. Co.,	Youth Alcohol use Pe	Greenglass_Storage_Monographs
3296717	The International student.	*NULL*	*NULL*	Intercollegiate Association, etc.]	Temperance Periodica	Greenglass_Storage_Monographs
3296695	Computadoras para todos / Jaime A. Restrepo.	Restrepo, Jaime A.	2017.	Vintage Espanol,	Computers, Computer	Greenglass_Storage_Monographs
3296667	Becoming animal : an earthly cosmology / David Abr	Abram, David, 1957-	p2017.	Tantor Media,	Anthropology Philoso	Greenglass_Storage_Monographs
3296614	Swamp Fox; the life and campaigns of General Franc	Bass, Robert D. (Rob	[©1959]	Holt	Marion Francis 1732	Greenglass_Storage_Monographs
3296546	Studies in English commerce and exploration in the	Rowland, Albert Lind	1924.	Press of the University of Pennsylvania,	Company of Merchants	Greenglass_Storage_Monographs
3296543	Sing, unburied, sing : a novel / Jesmyn Ward.	Ward, Jesmyn	2017.	Scribner,	African American fam	Greenglass_Storage_Monographs
3296542	The summer that made us / Robyn Carr.	Carr, Robyn	[2017].	Mira Books,	Sisters Fiction, Dom	Greenglass_Storage_Monographs
3296541	My fair lover / Nicole Jordan.	Jordan, Nicole	2017.	Ballantine Books,	Nobility Fiction, Tr	Greenglass_Storage_Monographs
3296264	Dress codes for small towns / Courtney Stevens.	Stevens, Courtney C.	[2017]	Harper Teen, an imprint of HarperCollinsPublishers...	Sex role Juvenile fi	Greenglass_Storage_Monographs
3296263	15 things not to do with a baby / Margaret McAllis	McAllister, Margaret	2015.	Frances Lincoln Children's Books,	Infants Juvenile fic	Greenglass_Storage_Monographs
3296261	The murder book / Jonathan Kellerman.	Kellerman, Jonathan	[2016]	Ballantine Books,	Delaware Alex Fictit	Greenglass_Storage_Monographs
3296260	The world of tomorrow / Brendan Mathews.	Mathews, Brendan.	2017.	Little, Brown and Company,	Brothers Fiction, Ir	Greenglass_Storage_Monographs
3296259	A catered costume party : a mystery with recipes /	Crawford, Isis	2017.	Kensington Books,	Simmons Bernie Ficti	Greenglass_Storage_Monographs
3296258	Breakfast in bed / Rochelle Alers.	Alers, Rochelle	[2017]	Dafina Books/Kensington Publishing Corp.,	African American wom	Greenglass_Storage_Monographs
3296221	The third policeman : a novel / by Flann O'Brien ;	O'Brien, Flann, 1911	1999, ©1967.	Dalkey Archive Press,	Murder Fiction, Dete	Greenglass_Storage_Monographs
3296220	My life in China = Wo de Zhongguo sheng huo / a fi	*NULL*	c2016.	Distributed exclusively by MVD Visual,	Eng Yau King, Eng Ke	Greenglass_Storage_Monographs
3296201	Faces and sounds / Pete Holmes.	Holmes, Pete, 1979-	[p2017]	Comedy Dynamics,	Stand up comedy, Ame	Greenglass_Storage_Monographs

Console

The screenshot that showing the structure that shows tables, rows, and size of the database

after the normalization

Table	Rows	Type	Collation	Size	Overhead	Creation	Last update
books_library	943,557	View	---	-	-	-	-
book_data[21]	262,065	InnoDB	utf8mb4_general_ci	117.7 MiB	-	Mar 07, 2020 at 10:27 AM	-
book_data[16]	335,833	InnoDB	utf8mb4_general_ci	138.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[9]	392,939	InnoDB	utf8mb4_general_ci	160.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[28]	392,187	InnoDB	utf8mb4_general_ci	162.7 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[8]	395,739	InnoDB	utf8mb4_general_ci	165.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[29]	391,036	InnoDB	utf8mb4_general_ci	163.7 MiB	-	Mar 07, 2020 at 10:31 AM	-
book_data[30]	402,210	InnoDB	utf8mb4_general_ci	165.7 MiB	-	Mar 07, 2020 at 10:31 AM	-
book_data[19]	403,333	InnoDB	utf8mb4_general_ci	166.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[22]	401,129	InnoDB	utf8mb4_general_ci	168.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[6]	398,765	InnoDB	utf8mb4_general_ci	168.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[27]	406,912	InnoDB	utf8mb4_general_ci	166.7 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[10]	393,089	InnoDB	utf8mb4_general_ci	169.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[14]	417,561	InnoDB	utf8mb4_general_ci	168.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[18]	412,671	InnoDB	utf8mb4_general_ci	169.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[15]	411,028	InnoDB	utf8mb4_general_ci	169.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[24]	416,826	InnoDB	utf8mb4_general_ci	169.7 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[11]	421,034	InnoDB	utf8mb4_general_ci	276.8 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[5]	409,070	InnoDB	utf8mb4_general_ci	170.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[7]	420,934	InnoDB	utf8mb4_general_ci	170.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[2]	401,796	InnoDB	utf8mb4_general_ci	170.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[3]	403,695	InnoDB	utf8mb4_general_ci	176.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[17]	402,652	InnoDB	utf8mb4_general_ci	176.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[4]	403,381	InnoDB	utf8mb4_general_ci	178.7 MiB	-	Mar 07, 2020 at 10:29 AM	-
book_data[25]	402,046	InnoDB	utf8mb4_general_ci	281.9 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[13]	418,414	InnoDB	utf8mb4_general_ci	178.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[12]	421,667	InnoDB	utf8mb4_general_ci	181.7 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[20]	417,152	InnoDB	utf8mb4_general_ci	291.8 MiB	-	Mar 07, 2020 at 10:28 AM	-
book_data[26]	402,870	InnoDB	utf8mb4_general_ci	280.9 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[23]	466,248	InnoDB	utf8mb4_general_ci	194.7 MiB	-	Mar 07, 2020 at 10:30 AM	-
book_data[1]	541,761	InnoDB	utf8mb4_general_ci	342.9 MiB	-	Mar 07, 2020 at 10:23 AM	-
book_data[31]	745,606	InnoDB	utf8mb4_general_ci	153.7 MiB	-	Mar 07, 2020 at 11:04 AM	-
42 table(s)	13,855,261	InnoDB	utf8mb4_general_ci	5.7 GiB	0 B	Mar 01, 2020 at 03:56 PM	Mar 07, 2020 at 01:24 PM