**BooksCrossing Dataset**

It is book crossing data having all the details of books published in each year.  
  
**Understanding Data**

The data format is comma separated values.

It contains 8 columns made up of following:   
**ISBN   
Book-Title   
Book-Author   
Year-Of-Publication   
Publisher   
Image-URL-S   
Image-URL-M   
Image-URL-L**

**Exploration ideas using Hive**

1) Create a database(library), table(myBooks) and describe the table.   
2) Load the data into the table myBooks.

3) Find the unique books titles.

4) Find how many books are published in every year.

5) Find the books that have been published more than once.

6) Find the top five publishers.  
  
1) Create a database(library), table(myBooks) and describe the table.  
hive> hive> create table myBook13(

> isbn string,booktitle string,bookauthor string,yearofpub string,publisher string,imageS string,imageM string,imageL string)

> row format delimited fields terminated by "\073" stored as textfile;  
hive> describe myBook13;

OK

col\_name data\_type comment

isbn string

booktitle string

bookauthor string

yearofpub string

publisher string

images string

imagem string

imagel string

2) Load the data into the table myBooks.

hive> hive> load data local inpath 'file:///home/cloudera/khasimbabu/HIVE\_Excercise/attachment\_Dataset-Apache-Hive-Assignment-Books.xls' into table myBook13;  
hive> select \* from myBook13 limit 5;

OK

mybook13.isbn mybook13.booktitle mybook13.bookauthor mybook13.yearofpub mybook13.publisher mybook13.images mybook13.imagem mybook13.imagel

"ISBN" "Book-Title" "Book-Author" "Year-Of-Publication" "Publisher" "Image-URL-S" "Image-URL-M" "Image-URL-L"

"0195153448" "Classical Mythology" "Mark P. O. Morford" "2002" "Oxford University Press" "http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg"

"0002005018" "Clara Callan" "Richard Bruce Wright" "2001" "HarperFlamingo Canada" "http://images.amazon.com/images/P/0002005018.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0002005018.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg"

"0060973129" "Decision in Normandy" "Carlo D'Este" "1991" "HarperPerennial" "http://images.amazon.com/images/P/0060973129.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0060973129.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg"

"0374157065" "Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It" "Gina Bari Kolata" "1999" "Farrar Straus Giroux" "http://images.amazon.com/images/P/0374157065.01.THUMBZZZ.jpg" "http://images.amazon.com/images/P/0374157065.01.MZZZZZZZ.jpg" "http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg"  
  
hive> select count(\*) from myBook13;  
  
3) Find the unique books titles.

hive> select count(DISTINCT booktitle) as book\_title\_count from mybook13;  
  
4) Find how many books are published in every year.

hive> select yearofpub,count(booktitle) as booktitlecount from mybook13 group by yearofpub;

5) Find the books that have been published more than once.

hive> select booktitle, count(booktitle) as booktitlecount from myBook13 group by booktitle having booktitlecount>1;  
  
6) Find the top five publishers.  
hive> select publisher,count(publisher) as publishercount from myBook13 group by publisher ORDER by publishercount DESC limit 5;  
  
  
[cloudera@quickstart /]$ hdfs dfs -ls /user/hive/warehouse/retail.db

Found 16 items

drwxrwxrwx - cloudera supergroup 0 2021-01-26 06:51 /user/hive/warehouse/retail.db/customer

drwxrwxrwx - cloudera supergroup 0 2021-01-26 23:10 /user/hive/warehouse/retail.db/mybook11

drwxrwxrwx - cloudera supergroup 0 2021-01-26 23:13 /user/hive/warehouse/retail.db/mybook12

drwxrwxrwx - cloudera supergroup 0 2021-01-26 23:16 /user/hive/warehouse/retail.db/mybook13

drwxrwxrwx - cloudera supergroup 0 2021-01-26 22:30 /user/hive/warehouse/retail.db/mybook5

drwxrwxrwx - cloudera supergroup 0 2021-01-26 22:32 /user/hive/warehouse/retail.db/mybook7

drwxrwxrwx - cloudera supergroup 0 2021-01-26 22:51 /user/hive/warehouse/retail.db/mybook8

drwxrwxrwx - cloudera supergroup 0 2021-01-26 22:57 /user/hive/warehouse/retail.db/mybook9

drwxrwxrwx - cloudera supergroup 0 2021-01-26 18:09 /user/hive/warehouse/retail.db/mybooks

drwxrwxrwx - cloudera supergroup 0 2021-01-26 21:12 /user/hive/warehouse/retail.db/mybooks1

drwxrwxrwx - cloudera supergroup 0 2021-01-26 21:13 /user/hive/warehouse/retail.db/mybooks2

drwxrwxrwx - cloudera supergroup 0 2021-01-26 07:43 /user/hive/warehouse/retail.db/out1

drwxrwxrwx - cloudera supergroup 0 2021-01-26 07:56 /user/hive/warehouse/retail.db/out2

drwxrwxrwx - cloudera supergroup 0 2021-01-26 08:05 /user/hive/warehouse/retail.db/out3

drwxrwxrwx - cloudera supergroup 0 2021-01-26 06:39 /user/hive/warehouse/retail.db/transaction

drwxrwxrwx - cloudera supergroup 0 2021-01-26 10:09 /user/hive/warehouse/retail.db/transactionbycategory

[cloudera@quickstart /]$ hdfs dfs -ls /user/hive/warehouse/retail.db/mybook13

Found 1 items

-rwxrwxrwx 1 cloudera supergroup 141201 2021-01-26 23:16 /user/hive/warehouse/retail.db/mybook13/attachment\_Dataset-Apache-Hive-Assignment-Books.xls