Apache Hive practise questions

Dataset – Books.csv

It is a book dataset 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 – ISBN identifier of the book

BookTitle – title of the book

BookAuthor - author of the book

YearOfPublication – year in which book was published

Publisher – the publisher who published the book

ImageURLS – the URL of small size image

ImageURLM– the URL of medium size image

ImageURLL– the URL of large size image

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.