READ ME

The data in each of the columns are enclosed within double quotes in the raw datasets. Hence, it's necessary to perform data cleansing to remove these double quotes.

There are multiple ways to perform data cleansing, using Map Reduce pattern algorithms, PIG etc.
I'm going to take the HIVE approach, as it's the most widely used technique.
So, in this approach, we'll be using HIVE to clean the data, as well as perform the necessary analysis to get the solutions.

1) **Cleaning and Loading BX-Book-Ratings.csv**

CREATING TABLE TO LOAD DATA WITH DOUBLEQUOTES:

```
create table if not exists bookratings
(userid string, isbn string, bookrating string)
 ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  WITH SERDEPROPERTIES (
       "separatorChar" = "\;",
       "quoteChar"     = '\"'
    )
```

Note: SERDE is used to remove the double quotes from the raw dataset. When SERDE is used in HIVE, by default, the **datatypes of the columns are converted to String.**
Use "describe bookratings" query to check the datatype of the columns.

LOADING DATA INTO TABLE:

```
load data local inpath "BX-Book-Ratings.csv" into table bookratings
```

Note: Since the datatype is by default String due to SERDE, we need **to cast String to BigInt while querying for analysis.**

SAMPLE QUERY FOR TYPE CAST:

```
select distinct(rating) from bookratings
order by cast(rating as bigint)
```

Note: If you don't type cast, the query will run. But, the data will be ordered lexically. (Run the above query without casting, and you'll see the difference.)

2) **Cleaning and Loading BX-Books.csv**

CREATING TABLE TO LOAD DATA WITH DOUBLEQUOTES:

```
create table if not exists bookstable
(isbn string, title string, author string, year string, publisher string,urls
string,urlm string,urll string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    WITH SERDEPROPERTIES (
      "separatorChar" = "\;",
      "quoteChar"     = '\"'
    )
tblproperties ("skip.header.line.count"="1")
```

Note: SERDE is used to remove the double quotes from the raw dataset. When SERDE is used in HIVE, by default, the **datatypes of the columns are converted to String.**
Use "describe bookstable" query to check the datatype of the columns.

LOADING DATA INTO TABLE:

```
load data local inpath "BX-Books.csv" into table bookstable
```

Note: Since the datatype is by default String due to SERDE, we need **to cast String to BigInt while querying for analysis.**

3) **BX-Users.csv**

We do not need this dataset to solve the given problem statements.

Check "Queries.pdf" for the solutions.