1. **Preparing dataset.**

**Step 1:**

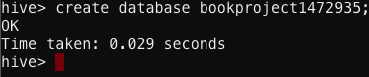
**Create a new database bookproject1472935 and use it.**

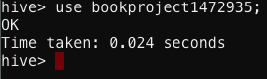
**Command:**

Create database bookproject1472935;

Use bookproject1472935;

**Screenshot:**



****

**Step 2:**

**Creating a table and loading to the database just created. (BX-Books.csv)**

**Command:**

CREATE TABLE books1472935 (ISBN STRING, Title STRING, Author STRING, Year\_of\_Pub STRING, Publisher STRING, Image\_URL\_S STRING, Image\_URL\_M STRING, Image\_URL\_L STRING ) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

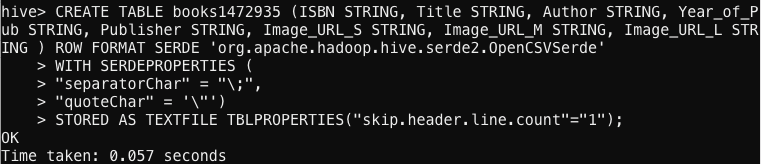
"separatorChar" = "\;",

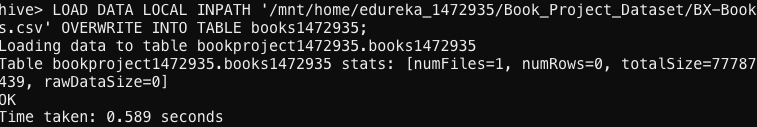
"quoteChar" = '\"')

STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/mnt/home/edureka\_1472935/Book\_Project\_Dataset/BX-Books.csv' OVERWRITE INTO TABLE books1472935;

**Screenshot:**

****

****

**Step 3:**

**Creating a table and loading to the database just created. (BX-Book-Ratings.csv)**

**Command:**

CREATE TABLE bookratings1472935 (UserID STRING, ISBN STRING, BookRating STRING)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

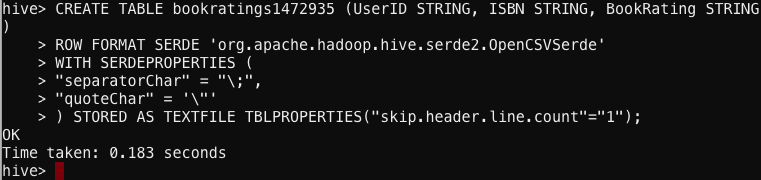
"separatorChar" = "\;",

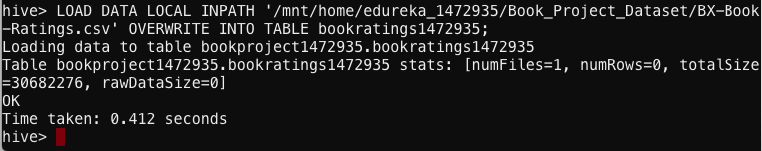
"quoteChar" = '\"'

) STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/mnt/home/edureka\_1472935/Book\_Project\_Dataset/BX-Book-Ratings.csv' OVERWRITE INTO TABLE bookratings1472935;

**Screenshots:**

****

****

**Solutions for the problem statements:**

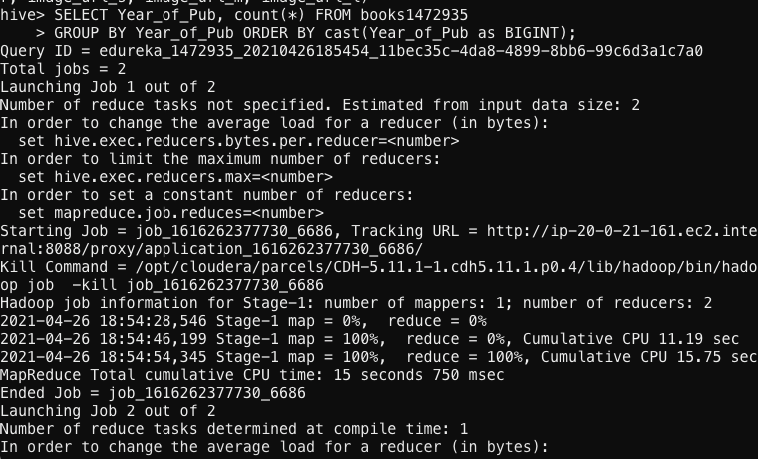
1. **A. Find out the frequency of books published each year. (Hint: Use Boooks.csv file for this)**

**Command:**

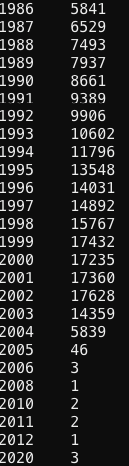
SELECT Year\_of\_Pub, count(\*) FROM books1472935

GROUP BY Year\_of\_Pub ORDER BY cast(Year\_of\_Pub as BIGINT);

**Screenshot:**

****

**Results:**

****

1. **B. Find out in which year maximum number of books were published**

**Command:**

SELECT Year\_of\_Pub, count(\*) as A

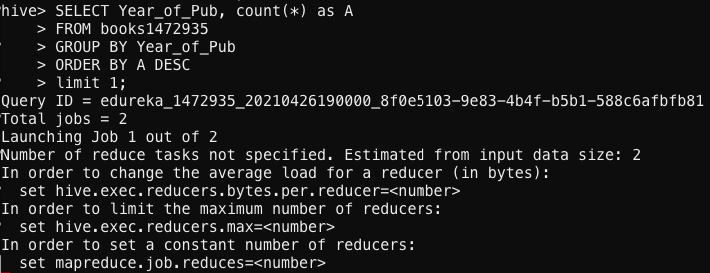
FROM books1472935

GROUP BY Year\_of\_Pub

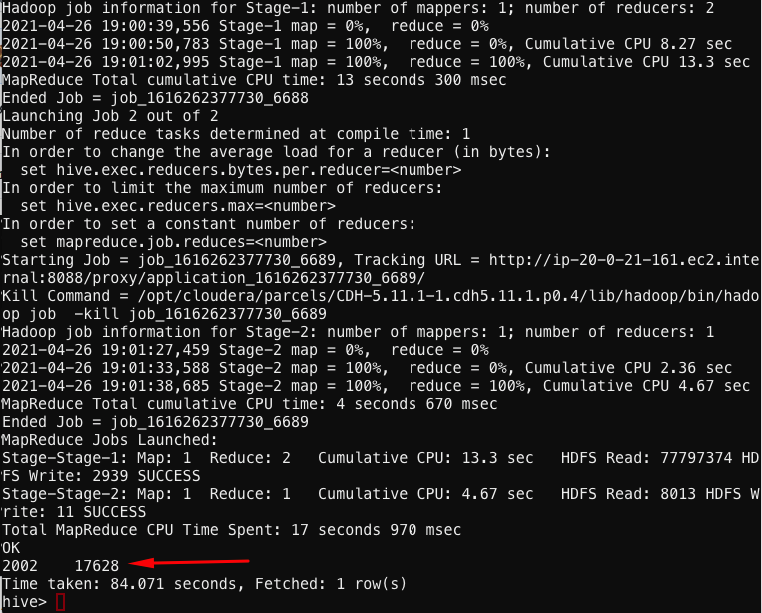
ORDER BY A DESC

limit 1;

**Screenshot:**

****

**Results:**



1. **C. Find out how many books were published based on ranking in the year 2002. ( Hint: Use Book.csv and Book-Ratings.csv)**

**Command:**

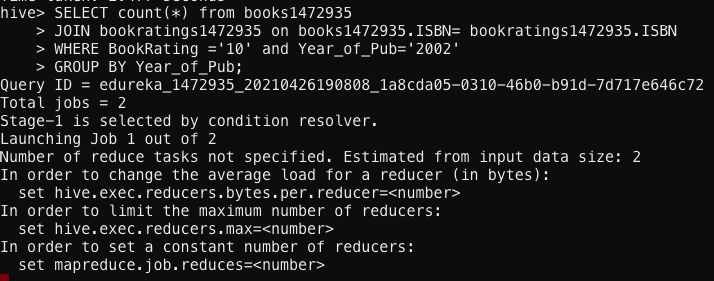
SELECT count(\*) from books1472935

JOIN bookratings1472935 on books1472935.ISBN= bookratings1472935.ISBN

WHERE BookRating ='10' and Year\_of\_Pub='2002'

GROUP BY Year\_of\_Pub;

**Screenshots:**

****

**Results:**

