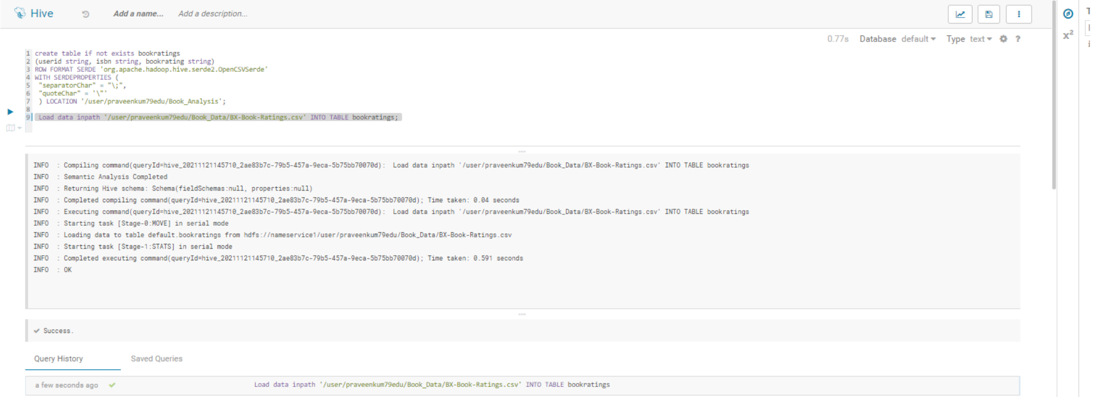
Book Project

Create and Load data in tables

Create Book Ratings table and Load data into it

```
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" = '\"'
 ) LOCATION '/user/praveenkum79edu/Book_Analysis';

 Load data inpath '/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv' INTO TABLE bookratings;
```



```
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" = '\"'
 ) LOCATION '/user/praveenkum79edu/Book_Analysis';

 Load data inpath '/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv' INTO TABLE bookratings;
```

```
INFO  : Compiling command(queryId=hive_20211121145710_2ae83b7c-79b5-457a-9eca-5b75bb70070d):  Load data inpath '/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv' INTO TABLE bookratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211121145710_2ae83b7c-79b5-457a-9eca-5b75bb70070d); Time taken: 0.04 seconds
INFO  : Executing command(queryId=hive_20211121145710_2ae83b7c-79b5-457a-9eca-5b75bb70070d):  Load data inpath '/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv' INTO TABLE bookratings
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.bookratings from hdfs://nameservice1/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211121145710_2ae83b7c-79b5-457a-9eca-5b75bb70070d); Time taken: 0.591 seconds
INFO  : OK
```

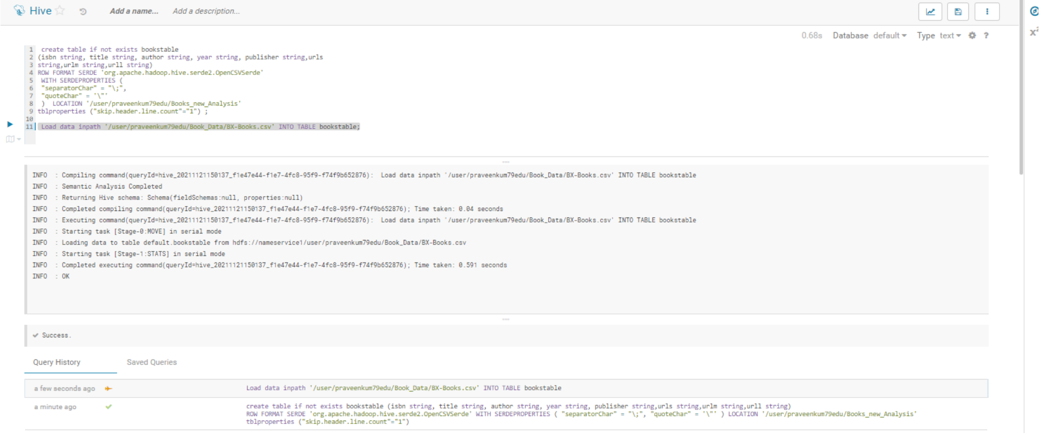
✓ Success.

```
Load data inpath '/user/praveenkum79edu/Book_Data/BX-Book-Ratings.csv' INTO TABLE bookratings
```

Create Books Table and load data in to it

```
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" = '\"'
 )  LOCATION '/user/praveenkum79edu/Books_new_Analysis'
tblproperties ("skip.header.line.count"="1") ;

 Load data inpath '/user/praveenkum79edu/Book_Data/BX-Books.csv' INTO TABLE bookstable;
```



```
 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" = '\"'
 )  LOCATION '/user/praveenkum79edu/Books_new_Analysis'
tblproperties ("skip.header.line.count"="1") ;

 Load data inpath '/user/praveenkum79edu/Book_Data/BX-Books.csv' INTO TABLE bookstable;
```

```
INFO  : Compiling command(queryId=hive_20211121150137_f1e47e44-f1e7-4fc8-95f9-f74f9b652876):  Load data inpath '/user/praveenkum79edu/Book_Data/BX-Books.csv' INTO TABLE bookstable
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211121150137_f1e47e44-f1e7-4fc8-95f9-f74f9b652876); Time taken: 0.04 seconds
INFO  : Executing command(queryId=hive_20211121150137_f1e47e44-f1e7-4fc8-95f9-f74f9b652876):  Load data inpath '/user/praveenkum79edu/Book_Data/BX-Books.csv' INTO TABLE bookstable
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.bookstable from hdfs://nameservice1/user/praveenkum79edu/Book_Data/BX-Books.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211121150137_f1e47e44-f1e7-4fc8-95f9-f74f9b652876); Time taken: 0.591 seconds
INFO  : OK
```

✓ Success.

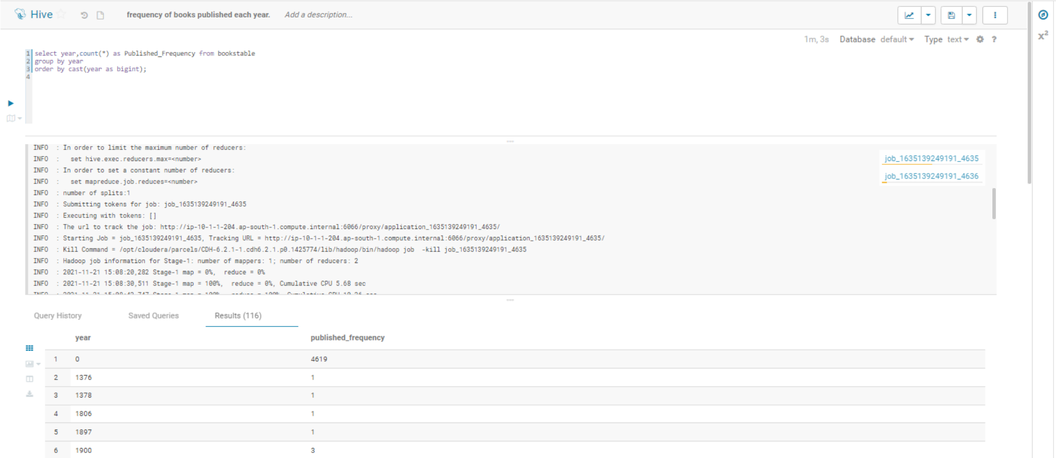
```
Load data inpath '/user/praveenkum79edu/Book_Data/BX-Books.csv' INTO TABLE bookstable

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" = '\"' ) LOCATION '/user/praveenkum79edu/Books_new_Analysis'
tblproperties ("skip.header.line.count"="1")
```

Problem Statement

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

```
select year,count(*) as Published_Frequency from bookstable
group by year
order by cast(year as bigint);
```



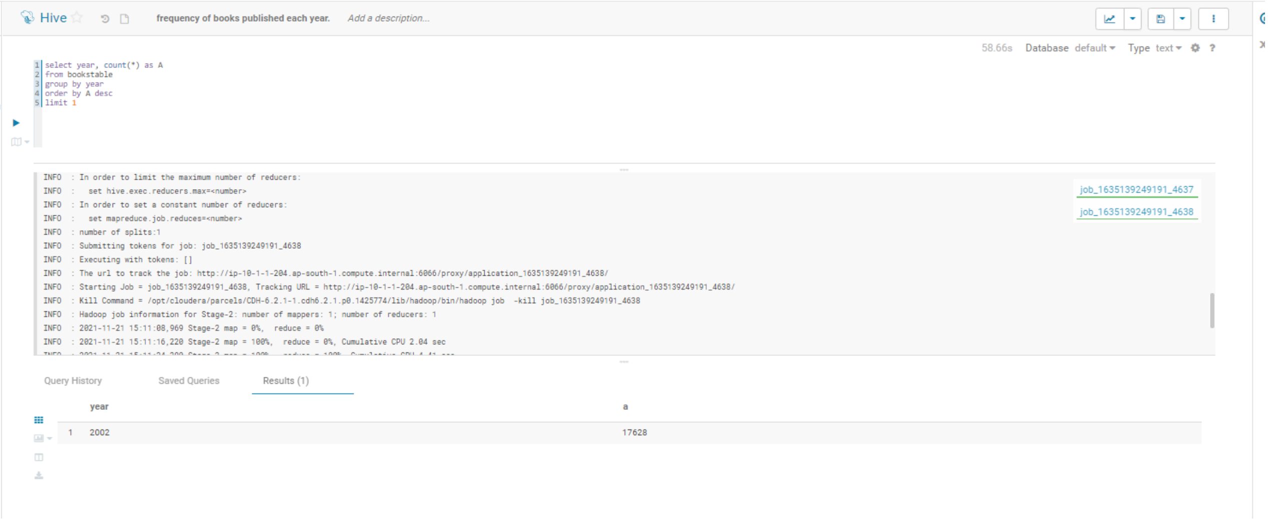
```
select year,count(*) as Published_Frequency from bookstable
group by year
order by cast(year as bigint);
```

```
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1635139249191_4635
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4635/
INFO  : Starting Job = job_1635139249191_4635, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4635/
INFO  : Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1635139249191_4635
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
INFO  : 2021-11-21 15:08:20,282 Stage-1 map = 0%,  reduce = 0%
INFO  : 2021-11-21 15:08:30,511 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.68 sec
```

	year	published_frequency
1	0	4619
2	1376	1
3	1378	1
4	1806	1
5	1897	1
6	1900	3

- **Find out in which year maximum number of books were published**

```
select year, count(*) as A
from bookstable
group by year
order by A desc
limit 1
```



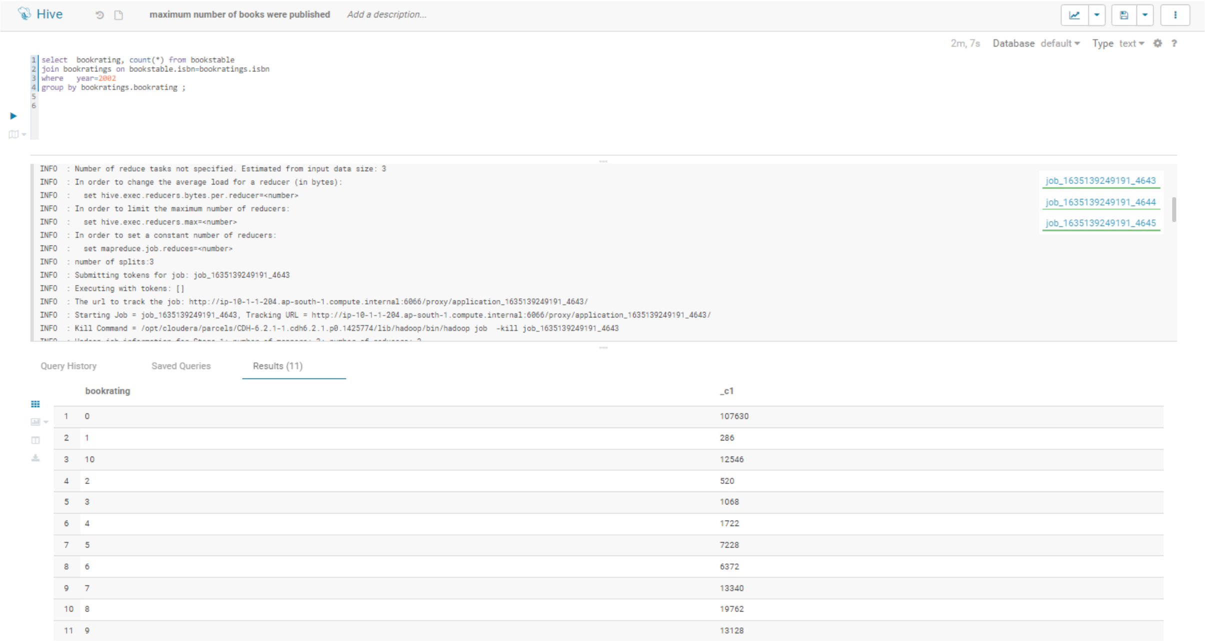
```
select year, count(*) as A
from bookstable
group by year
order by A desc
limit 1
```

```
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1635139249191_4638
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4638/
INFO  : Starting Job = job_1635139249191_4638, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4638/
INFO  : Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1635139249191_4638
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2021-11-21 15:11:08,969 Stage-2 map = 0%,  reduce = 0%
INFO  : 2021-11-21 15:11:16,220 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.04 sec
```

	year	a
1	2002	17628

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

```
select  bookrating, count(*) from bookstable
join bookratings on bookstable.isbn=bookratings.isbn
where   year=2002
group by bookratings.bookrating ;
```



```
select  bookrating, count(*) from bookstable
join bookratings on bookstable.isbn=bookratings.isbn
where   year=2002
group by bookratings.bookrating ;
```

```
INFO  : Number of reduce tasks not specified. Estimated from input data size: 3
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:3
INFO  : Submitting tokens for job: job_1635139249191_4643
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4643/
INFO  : Starting Job = job_1635139249191_4643, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4643/
INFO  : Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1635139249191_4643
```

	bookrating	_c1
1	0	107630
2	1	286
3	10	12546
4	2	520
5	3	1068
6	4	1722
7	5	7228
8	6	6372
9	7	13340
10	8	19762
11	9	13128