Lab Hive

(i) Enter the following URL in Google Chrome:
https://classroom-vc.gre.ac.uk

Launch the Hadoop VM image, following the instructions from Lab 5.

(ii) Once the Hadoop desktop has loaded, right click on the desktop and start a terminal

(iii) At the command prompt type:

```
start-dfs.sh
start-yarn.sh
```

to start Hadoop.

(iv) Change directory into your workspace directory

```
cd ../workspace
```

(v) Copy the Lab7 directory from `Datasets` into your workspace directory

(vi) Change current directory to the `workspace/Lab7` directory

(vii) Copy the book.csv and prize.csv files onto the HDFS

```
hdfs dfs –put book.csv

hdfs dfs –put prize.csv
```

(viii) The book.csv file has the following column headings:

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

(ix) Type hive at the command prompt to start the Hive CLI

(x) At the hive command prompt, create a table for the book.csv data:

```
create table book(ISBN STRING, Title STRING, Author STRING, Year INT, Publisher STRING, URL1 STRING, URL2 STRING, URL3 STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

(xi) Next load the data into the book table:

```
LOAD DATA INPATH 'book.csv' OVERWRITE INTO TABLE book;
```

(xii) Likewise create a table prize and load the data from prize.csv into this table. The prize.csv has the following column headings:

year	isbn	Title	Author	Publisher	NumberSold

(you should create the table on your own!)

(xiii) Run the following queries and comment on what information they give:

```
SELECT * from book;

SELECT year, count(title) from book group by year;

SELECT year, count(title) from book WHERE year < 1990 group by year;

SELECT isbn, title, author, year FROM book WHERE author LIKE '%Michael%' AND year > 1990;

SELECT title, author, year FROM book WHERE title LIKE '%Green%' ;

SELECT MAX(year)FROM book;

SELECT Author, count(Author) FROM book GROUP BY Author;
```

(xiv) To write the results of your query to a file try the following:

```
INSERT OVERWRITE LOCAL DIRECTORY './hiveResult'
SELECT Author, count(Author)
FROM book
GROUP BY Author;
```

Once you have run this query you can **cd** to **/hiveResult**

In the directory **hiveResult** will be a file which will contain the results of your query.

(xv) Run the following two queries :

```
SELECT DISTINCT prize.Author, book.title from book JOIN prize ON
(prize.Author = book.Author) order BY prize.Author;
```

```
SELECT prize.Author, count(distinct book.title) from book JOIN prize
ON (prize.Author = book.Author) GROUP BY prize.Author;
```

Do you understand the results?

Notes:

To exit the hive shell type quit

```
$ hive> quit;
```

The user may wish to run one or more queries (semicolon separated) from the terminal and then have the hive CLI exit immediately after completion. The CLI accepts a -e command argument that enables this feature:

```
$ hive -e "SELECT * FROM book LIMIT 3";
```

Hive can execute one or more queries that were saved to a file using the **-f** file argument. By convention, saved Hive query files use the **.q** or **.sql** extension.

```
$ hive -f /path/to/file/withqueries.sql
```