DATA ENGINEERING PLATFORMS (MSCA 31012)

sbharadwaj@uchicago.edu | jchan530@uchicago.edu

Submissions

- Team : One Submission/per team (PPT) sent to sbharadwaj@uchicago.com
- Individual : Submission on Canvas. SQL query used along with any assumptions & outputs (first few rows) for each question. A knitted HTML/pdf file generated from a RMarkdown AND/OR Jupyter Notebook that addresses all questions

Part A (Team): Business use case , datasets and tools relating to the final project

1. Create a presentation of about 4 – 5 slides which addresses the following about your final project. **– { 0 Points }**
   - Executive summary
   - Business Use Case
   - Data/tools you plan on using

Data

Students have the flexibility to can use any public or in-house (such as IRI/Nielsen) dataset. Additional references to datasets can be found under modules > final project (assignment document).

Note : To work on the IRI dataset, all members of the team need to sign the NDA form found under modules>Final Project>datasets>IRI> IRI NDA.pdf and send signed documentation over to gguevara@uchicago.edu

Part B (Individual): Manipulating, Categorizing, Sorting and Grouping & Summarizing Data

Data (Sakila dataset)

- We will use the Sakila database schema which can be found at: http://dev.mysql.com/doc/index-other.html/
- Full documentation: http://dev.mysql.com/doc/sakila/en/ https://downloads.mysql.com/docs/sakila-en.pdf (Data Dictionary)

```
/*********************************************
** File:      Assignment2-PartB.sql
** Desc:    Manipulating, Categorizing, Sorting and Grouping & Summarizing Data
** Author:
** Date:
*********************************************/

########  QUESTION 1 ########                                    – { 10 Points }

# a) Show the list of databases.

# b) Select sakila database.

# c) Show all tables in the sakila database.

# d) Show each of the columns along with their data types for the actor table.

# e) Show the total number of records in the actor table.

# f) What is the first name and last name of all the actors in the actor table ?

# g) Insert your first name and middle initial ( in the last name column ) into the actors
table.

# h) Update your middle initial with your last name in the actors table.

# i) Delete the record from the actor table where the first name matches your first name.

# j) Create a table payment_type with the following specifications and appropriate data
types

#       Table Name : “Payment_type”
#       Primary Key:  "payment_type_id”
#       Column: “Type”
#       Insert following rows in to the table:
#       1, “Credit Card” ; 2, “Cash”; 3, “Paypal” ; 4 , “Cheque”

# k) Rename table payment_type to payment_types.

# l) Drop the table payment_types.
```

######## QUESTION 2 ######## **– { 10 Points }**

a) List all the movies (title & description) that are rated PG-13 ?

b) List all movies that are either PG OR PG-13 using IN operator ?

c) Report all payments greater than and equal to 2$ and Less than equal to 7$?
Note : write 2 separate queries conditional operator and BETWEEN keyword

d) List all addresses that have phone number that contain digits 589, start with 140 or end with 589
Note : write 3 different queries

e) List all staff members (first name, last name, email) whose password is NULL ?

f) Select all films that have title names like ZOO and rental duration greater than or equal to 4

g) What is the cost of renting the movie ACADEMY DINOSAUR for 2 weeks ?
Note : use of column alias

h) List all unique districts where the customers, staff, and stores are located
Note : check for NOT NULL values

i) List the top 10 newest customers across all stores

######## QUESTION 3 ######## **– { 10 Points }**

a) Show total number of movies

b) What is the minimum payment received and max payment received across all transactions ?

c) Number of customers that rented movies between Feb-2005 & May-2005 (based on paymentDate).

d) List all movies where replacement_cost is greater than 15$ or rental_duration is between 6 & 10 days

e) What is the total amount spent by customers for movies in the year 2005 ?

f) What is the average replacement cost across all movies ?

g) What is the standard deviation of rental rate across all movies ?

h) What is the midrange of the rental duration for all movies

######## QUESTION 4 ######## **– { 10 Points }**

a) Customers sorted by first Name and last name in ascending order.

b) Count of movies that are either G/NC-17/PG-13/PG/R grouped by rating.

c) Number of addresses in each district.

d) Find the movies where rental rate is greater than 1$ and order result set by descending order.

e) Top 2 movies that are rated R with the highest replacement cost ?

f) Find the most frequently occurring (mode) rental rate across products.

g) Find the top 2 movies with movie length greater than 50mins and which has commentaries as a special features.

h) List the years which has more than 2 movies released.

Part C (Individual): Combining Data, Nested Queries, Views and Indexes, Transforming Data

```
/*********************************************
** File:   Assignment2-PartC.sql
** Desc: Combining Data, Nested Queries, Views and Indexes, Transforming Data
** Author:
** Date:
*********************************************/
```

######## QUESTION 1 ######## **– { 20 Points }**

a) List the actors (firstName, lastName) who acted in more then 25 movies.

Note: Also show the count of movies against each actor

b) List the actors who have worked in the German language movies.

Note: Please execute the below SQL before answering this question.

```
SET SQL_SAFE_UPDATES=0;
UPDATE film SET language_id=6 WHERE title LIKE "%ACADEMY%";
```

c) List the actors who acted in horror movies.
Note: Show the count of movies against each actor in the result set.

d) List all customers who rented more than 3 horror movies.

e) List all customers who rented the movie which starred SCARLETT BENING

f) Which customers residing at postal code 62703 rented movies that were Documentaries.

g) Find all the addresses where the second address line is not empty (i.e., contains some text), and return these second addresses sorted.

h) How many films involve a "Crocodile" and a "Shark" based on film description ?

i) List the actors who played in a film involving a "Crocodile" and a "Shark", along with the release year of the movie, sorted by the actors' last names.

j) Find all the film categories in which there are between 55 and 65 films. Return the names of categories and the number of films per category, sorted from highest to lowest by the number of films.

k) In which of the film categories is the average difference between the film replacement cost and the rental rate larger than 17$?

l) Many DVD stores produce a daily list of overdue rentals so that customers can be contacted and asked to return their overdue DVDs. To create such a list, search the rental table for films with a return date that is NULL and where the rental date is further in the past than the rental duration specified in the film table. If so, the film is overdue and we should produce the name of the film along with the customer name and phone number.

m) Find the list of all customers and staff given a store id
Note : use a set operator, do not remove duplicates

######## QUESTION 2 ######## **– { 10 Points }**
a) List actors and customers whose first name is the same as the first name of the actor with ID 8.

b) List customers and payment amounts, with payments greater than average the payment amount

c) List customers who have rented movies atleast once
Note: use IN clause

d) Find the floor of the maximum, minimum and average payment amount

######## QUESTION 3 ######## **– { 5 Points }**
a) Create a view called actors_portfolio which contains information about actors and films (including titles and category).

b) Describe the structure of the view and query the view to get information on the actor ADAM GRANT

c) Insert a new movie titled Data Hero in Sci-Fi Category starring ADAM GRANT

######## QUESTION 4 ######## **– { 5 Points }**
a) Extract the street number (characters 1 through 4) from customer addressLine1

b) Find out actors whose last name starts with character A, B or C.

c) Find film titles that contains exactly 10 characters

d) Format a payment_date using the following format e.g "22/1/2016"

e) Find the number of days between two date values rental_date & return_date

######## QUESTION 5 ######## **– { 20 Points }**
Provide 5 additional queries and indicate the specific business use cases they address.