DATA ENGINEERING PLATFORMS (MSCA 31012)

ASSIGNMENT 1

Submissions (via Canvas)

- Submit solutions in PDF, PPT, Excel or MS Word document (as applicable). Do not submit zip files.
- Do not submit the cleaned up dataset for the OpenRefine project.

Part A : Software installations

1. **Follow the installation guides uploaded** (or search google for installation instructions) and install the following software on your local computer (submit a **screenshot of your desktop** with shortcuts and validations). **– { 20 Points }**
   1) OpenRefine
   2) MySQL (server + workbench)
   3) Anaconda (Open Data Science Platform : Python)
   4) R-studio
   5) Tableau (https://www.tableau.com/academic/students)
   6) FileZilla Or CyberDuck
   7) MongoDB
   8) GCP (credits added to your account)

Part B : Relational data model and design principles

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/

1. Relational Data Modeling (show all Screen caps in Word docx format) **– { 20 Points }**

   a. Download Sakila dataset and unzip sakila-db.zip file from the URL listed above.
   b. Execute sakila-schema.sql file in the SQL workbench
   c. Reverse Engineer the database and generate the EER diagram using the MySQL workbench
   d. Add a new lookup table: payment_type (1 to Many relationship with payment entity) with the following attributes:
      - payment_type_id (Primary Key) : SMALLINT(6)
      - method - varchar (10)
      - description – varchar (45)

   Add the foreign key payment_type_id in the Payment entity with the following attributes:
      - Payment_type_id (Foreign Key) : SMALLINT(6)

e. For the Payment table fill out the form below:

Table Name: Payment

Field (Attributes)	Primary Key (Y/N)	Foreign Key (Y/N)	Related Table(s) (only enter this for foreign key fields) & Type of relationship between tables

2. Normalization : For the table below: **– { 20 Points }**

a. Provide examples of insertion, deletion, and modification anomalies.

b. Normalize this table to 3NF and list any assumptions.

Physician ID	Physician Name	Physician's Office	Patient ID	Patient Name	Patient Address	Appointment Date	Surgery
1	Helen Pearson	832 E Chicago Ave, Chicago 60648	1000	Joe Korn	821 W Randolph Street, Chicago, 60631	3/7/2017	Tendon Repair
1	Helen Pearson	832 E Chicago Ave, Chicago 60648	1001	Gillian White	4331 Illinois Street, Chicago, 60632	3/22/2017	Skin Graft
2	Olga Kay	3606 N. Clark Street, Chicago 60647	1000	Joe Korn	821 W Randolph Street, Chicago, 60631	6/13/2016	Sentinel Node Biopsy
3	Robert Smith	41 W. Madison Ave, Chicago 60606	1002	Jill Bell	162 E Huron Street, Chicago, 60613	6/13/2017	Tendon Repair
3	Robert Smith	41 W. Madison Ave, Chicago 60606	1003	Jill Bell	162 E Huron Street, Chicago, 60613	6/14/2017	Skin Graft
4	Wei Jing	3606 N. Clark Street, Chicago	1004	Mike Li	4531 W Lake Street, Chicago 60654	6/13/2017	Knee Arthroscopy
5	Jay Patel	41 W. Madison Ave, Chicago 60606	1001	Gillian White	4331 Illinois Street, Chicago, 60632	8/15/2017	Sentinel Node Biopsy
5	Jay Patel	41 W. Madison Ave, Chicago 60606	1006	Ian MacKay	41 N Dearborn Street, Chicago, 60652	1/4/2016	Hepatic Resection
5	Jay Patel	41 W. Madison Ave, Chicago 60606	1006	Ian MacKay	41 N Dearborn Street, Chicago, 60652	1/5/2018	Liver Transplant
2	Helen Pearson	832 E Chicago Ave, Chicago 60648	1007	Sheela Nupur	333 W Monroe Street, Chicago, 60606	1/4/2016	Knee Arthroscopy
4	Wei Jing	3606 N. Clark Street, Chicago 60647	1000	Joe Korn	821 W Randolph Street, Chicago, 60631	2/12/2016	Skin Graft
4	Wei Jing	3606 N. Clark Street, Chicago 60647	1004	Mike Li	4531 W Lake Street, Chicago 60654	4/15/2018	Skin Graft

3. Data Modeling Presentation in PPTX (presentation slide deck): **– { 20 Points }**

Design a data model that can be used for property management and monitoring of single-family homes for investors and owners.

Consider data for the following entities/attributes that need to be captured by business:

a. Home location
b. Age of the house
c. Construction material used
d. Type of residence (apt, condo, etc.)
e. Home layout (number of roomes, sq footage, etc.)
f. Number and Types of Appliances (Heating, Fridge etc.)
g. Name and other details of the renters/leasers/resident (s)
h. Rental Payments made against the house
i. Add other entities (and/or collection of attributes) that you think could add insights for the investors and business users

Please **submit a PPTX with 4 slides** that details the Entity Relationship Diagram (tables/relationships/cardinality/datatypes), short summary of Design considerations (which database, how many users , need for distributed databases, data security, privacy and integrity).

Part C: Data Collection & Preparation

1. This assignment is related to data collection and transformation. (Knit to PDF) **– { 20 Points}**
   a. **Using Public APIs** : Choose any data provider (such as Twitter/YouTube… etc). to collect data and transform it to a clean structured tabular data (sample size of 50 records) using Python
   b. **WebScraping** : Choose a website you want to scrape. Collect some of the data from the website and transform it to a clean structured tabular data (sample size of 50 records) using Python

   Note: Must knit both examples to PDF for grading. Please make sure to produce a clean output file that limits rows printouts for readability.

References:

- https://medium.com/pew-research-center-decoded/using-apis-to-collect-website-data-b7fc340d59e3
- https://towardsdatascience.com/getting-started-with-apis-in-python-to-gather-data-1185796b1ec3
- https://www.dataquest.io/blog/python-api-tutorial/
- https://www.dataquest.io/blog/web-scraping-tutorial-python/
- https://likegeeks.com/python-web-scraping/

Another useful site : https://lmgtfy.com/