DATA ENGINEERING PLATFORMS (MSCA 31012)

sbharadwaj@uchicago.edu | salerno@uchicago.edu | jchan530@uchicago.edu

Submissions

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

Part A : Software installations, data extraction, cleaning & transformation

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). **– { 40 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) Neo4J

2. Run the following data preparation steps on the dataset below and submit relevant screenshots as word or pdf document. **– { 20 Points }**

   Note: Dataset sandyrelated.csv is uploaded as part of this assignment.

   a. Import the data into OpenRefine and create a new project "SandyCleanup"
   b. Remove columns where majority of the cells are empty or have "Unspecified" or "NA" values
      (Do not remove the columns that are needed to complete the rest of this exercise)
   c. Trim white spaces on all address related columns and transform addresses into title case
   d. Convert City to title case, then Cluster and Merge the column
   e. Clean up Descriptor Column - Cluster and Merge following text categories:
      1. "Other Water problem(WZZ)", "Other Water problem(QZZ)" as "Other Water Problem"

2. "Commercial 421 A/B Exemptions" as "Commercial Exemption"
3. "Commercial Exemption" "Commercial Other Exemption" as "Commercial Exemption"
4. "Personal DRIE Exemption", "Personal SCHE Exemption", "Personal DHE Exemption" as "Personal Exemption"

f. Clean up Location Type - Cluster and Merge following text categories:
1. "Comercial", "Commercial", "Store/Commercial" as "Commercial"
2. "RESIDENTIAL BUILDING", "Residential Building", "Residence" as "Residential"
3. "Club/Bar/Restaurant", "Bar/Restaurant", "Restaurant" as "Club/Bar/Restaurant"
4. "3+ Family Apt. Building", "3+ Family Apartment Building" as "3+ Family Apartment"
5. "Street/Sidewalk", "Street and Sidewalk" as "Street/Sidewalk"

g. Online web services such as the following can be used to fetch the address given a geocode:
https://developers.google.com/maps/documentation/geocoding/intro#ReverseGeocoding
https://developer.mapquest.com/documentation/open/nominatim-search/

Web Service API Examples:
https://maps.googleapis.com/maps/api/geocode/json?latlng=40.714224,-73.961452
http://nominatim.openstreetmap.org/reverse?format=json&lat=40.714224&lon=-73.961452

- Formulate the URL expression in OpenRefine using GREL that would fetch the complete JSON results from this web service API **(You need not invoke the API or download the data from the web service call. If you want to invoke the API, use limited set of rows)**

h. Look for any other clean up opportunities and execute the clean up on this dataset

i. Export final project into a CSV file on your local computer. Please follow the best practices for file naming.

3. Create and submit a one-page memo summarizing the below content from https://gartner.uchicago.edu {CNET id} **– { 10 Points }**

- Modern Data Management Requires a Balance Between Collecting Data and Connecting to Data

Note : Target audience is the executive management.

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 **– { 10 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: **– { 10 Points }**

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

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

Physician Name	**Physician's Office**	**Patient Name**	**Patient Address**	**Appointment Date**	**Surgery**
Helen Pearson	Chicago Ave, Chicago	Joe Korn	Randolph Street, Chicago	3/7/2017	Tendon Repair
Helen Pearson	Chicago Ave, Chicago	Gillian White	Illinois Street, Chicago	3/22/2017	Skin Graft
Olga Kay	Clark Street, Chicago	Joe Korn	Randolph Street, Chicago	6/13/2016	Sentinel Node Biopsy
Robert Smith	Madison Street, Chicago	Jill Bell	Huron Street, Chicago	6/13/2017	Tendon Repair
Robert Smith	Madison Street, Chicago	Jill Bell	Huron Street, Chicago	6/14/2017	Skin Graft
Wei Jing	Adams Street, Chicago	Mike Li	Lake Street, Chicago	6/13/2017	Knee Arthroscopy
Jay Patel	Monroe Street, Chicago	Gillian White	Illinois Street, Chicago	8/15/2017	Sentinel Node Biopsy
Jay Patel	Monroe Street, Chicago	Ian MacKay	Dearborn Street, Chicago	1/4/2016	Hepatic Resection
Jay Patel	Monroe Street, Chicago	Ian MacKay	Dearborn Street, Chicago	1/5/2018	Liver Transplant
Helen Pearson	Chicago Ave, Chicago	Sheela Nupur	Monroe Street, Chicago	1/4/2016	Knee Arthroscopy

Wei Jing	Adams Street, Chicago	Joe Korn	Randolph Street, Chicago	2/12/2016	Skin Graft
Wei Jing	Adams Street, Chicago	Mike Li	Lake Street, Chicago	4/15/2018	Skin Graft

3. Design a data model that can be used to track information for a movie production studio. The data points captured by the business is below: **– { 10 Points }**

a. The names of the movies
b. The year which a movie was produced
c. The rating for the movie (e.g. G, PG, PG-13, R, etc.)
d. The first and last names of the producer for each movie (assume that there is only one producer per movie)
e. The first and last names of each actor in each movie
f. Keep track of which actors were starring actors in the movie and which were supporting actors
g. The amount of money each actor was paid for making the movie
h. The names and addresses of the theatres where each movie was shown (there can be many theatres, possibly thousands, where each movie was shown)
i. The number of tickets sold for each movie at each theatre
j. The price per ticket at each theatre - for the purpose of this assignment you should assume that a theatre charges the same amount of money for every ticket that it sells.

Please submit a PPT 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) .