**Part 1) Data Engineering (10%)**

This exam includes three tasks for data engineering. Complete all tasks and submit the following items to MyCourseVille:

* DE1.ipynb, DE2.ipynb, DE3.ipynb – source codes for all three tasks
* DE.docx – a capture screen of each question. It cannot be graded if we cannot map your answer to the question.
* All files must be renamed as “{student\_id}\_{firstname}\_{filename}\_Part1”, e.g.,
  + 6030133421\_Chaiyatad\_Part1\_DE.docx
  + 6030133421\_Chaiyatad\_Part1\_DE1.ipynb
  + 6030133421\_Chaiyatad\_Part1\_DE2.ipynb
  + 6030133421\_Chaiyatad\_Part1\_DE3.ipynb
  + Finally, zip all above files to be “6030133421\_Chaiyatad\_Part1.zip”

# Task1: Data Extraction [2 points]

Write a data extraction program to extract the following data from Wikipedia:

<https://en.wikipedia.org/wiki/Chulalongkorn_University>

Extract and print number of references in the “References” section (shown below). For example, if there are 99 references, **your program** will print 99 as the result.

Graphical user interface, text, application, email

Description automatically generated

# Task 2: Data Ingestion [2 points]

Write a program to connect to a kafka broker at 34.126.125.187:9092 and receive a message from topic “topic1”. The message is in JSON format. Print out the value of field “id”.

# Task 3: Spark

Use **Apache Spark** to analyze the 17K Mobile Strategy Games dataset on Kaggle. This dataset includes data of 17007 strategy games on the Apple App Store. It was collected on the 3rd of August 2019, using the iTunes API and the App Store sitemap. It can be downloaded from the link below:

<https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games/download>

You can refer to the following page for more details:

<https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games>

# Attribute Information

|  |  |
| --- | --- |
| URL | The URL |
| ID | The assigned unique ID |
| Name | The name |
| Subtitle | The secondary text under the name |
| Icon URL | URL to 512px x 512px jpg icon |
| Average User Rating | The average rating by users |
| User Rating Count | Number of ratings internationally, null means it is below 5 |
| Price | Price in USD (0 means free game, > 0 means paid game) |
| In-app Purchases | Prices of available in-app purchases |
| Description | App description |
| Developer | App Developer |
| Age Rating | Either 4+, 9+, 12+ or 17+ |
| Languages | ISO2A language codes (e.g. EN = English) |
| Size | Size of the app in bytes |
| Primary Genre | The main genre |
| Genres | Genres of the app |
| Original Release Date | When it was released |
| Current Version Release Date | When it was last updated |

# Tasks (You must use Spark; the code using Pandas CANNOT be scored!).

1. Import data and preprocess the data [2 points]

\*\*\*

Before continue, prepare your data as followed:

* Drop missing values in all columns

df = df.na.drop()

(where df is the dataframe containing data you import)

* Change the following columns to the proper formats:
  + “Average User Rating” and “Price” to double
  + “User Rating Count” and “Size” to int

\*\*\*

* 1. How many rows and columns?
  2. (Screen capture) Preview top 6 rows

1. Explore the data [4 points]
   1. What is the average size of all games?
   2. How much does it cost for the game with the most expensive price?
   3. Based on the average of ‘Average User Rating’, is it true that paid games score much better than free games by large margin (> 1)?
   4. Based on the average of ‘User Rating Count’, is it true that free games have more rating counts than paid games by large margin (> 1000)?