**Part 3) Data Engineering (10%)**

This exam includes **three tasks** for data engineering. Complete all tasks, answer all questions in the exam (MyCourseVille) and leave the following items on your desktop:

* The files must be renamed as follows:
  + {StudentID}\_{FirstName}\_Part3\_DE{Task}.ipynb – source codes for all three tasks
  + {StudentID}\_{FirstName}\_Part3\_DE.docx – a capture screen of each question. It **cannot** be graded if we **cannot** map your answer to the question.

# Task1: Data Extraction [4 points]

The redis contains the movie database, which has the following structure:

|  |  |  |  |
| --- | --- | --- | --- |
| **Key Pattern** | **Type** | **Description** | **Example** |
| m:xxx | Hash | Information of movie id ‘xxx’ including ‘title’, ‘runtime’, and ‘year’ | **Key**  m:1  **Value**  title = The Shawshank Redemption  runtime = 142 min  year = 1994  votes = 2343110  director = Frank Darabont  rating = 9.3 |
| g:ZZZZ | Set | Set of movie keys in ‘ZZZZ’ genre | **Key**  g:Drama  **Value**  { m:1, m:11, m:14, m:102, m:293, … } |

Write a program to access a redis (host=“redis”) and answer the following question:

1. What is the title of movie id 52?
   1. Avengers: Endgame
   2. The Dark Knight
   3. Guardians of the Galaxy
   4. Harry Potter and the Prisoner of Azkaban
   5. Inception
   6. Toy Story
2. How many movies are there in 'Thriller' genre?
   1. 18
   2. 45
   3. 99
   4. 121
   5. 132
   6. 407
3. How many genres are there in the database?
   1. 21
   2. 37
   3. 42
   4. 56
   5. 63
   6. 74
4. What is the movie with the highest rating in ‘Action’ genre?
   1. 8.5
   2. 8.6
   3. 8.7
   4. 8.8
   5. 8.9
   6. 9.0

# Task 2: Data Ingestion [2 points]

Write a program to connect to a kafka broker at (host= “kafka:9094”) and receive a message from topic “seq”. The driver will generate messages in JSON format with 2 fields, ‘seq’ and ‘code’. The seq field in the message will start from 0 and increase toward the maximum value n and then go back to 0 again. For each message, the code field will be a random integer, unique for each seq number. For example: suppose the maximum value n is 2, the messages will be generated as followed:

{ ‘seq’: 0, ‘code’: 309 }

{ ‘seq’: 1, ‘code’: 183 }

{ ‘seq’: 2, ‘code’: 482 }

{ ‘seq’: 0, ‘code’: 309 }

{ ‘seq’: 1, ‘code’: 183 }

…

Answer the following questions:

1. What is the maximum sequence number?
   1. 4
   2. 5
   3. 6
   4. 7
   5. 8
   6. 9
2. Which sequence number has the maximum code?
   1. 0
   2. 1
   3. 2
   4. 3
   5. 4
   6. 5
   7. 6
   8. 7
   9. 8
   10. 9
3. What is the code for sequence number 1
   1. 471
   2. 102
   3. 335
   4. 836
   5. 901
   6. 284
   7. 494
   8. 625
   9. 548
   10. 790

# Task 3: Spark [4 points]

Use Apache Spark to analyze the airbnb dataset (airbnb\_data.csv). This dataset includes data of Airbnb in New York City. It has been cleaned and ready to be analyzed. It contains the following columns:

# Column Information

|  |  |  |
| --- | --- | --- |
| id | integer | Airbnb's unique identifier for the listing |
| name | string | Name of the listing |
| host\_id | integer | Airbnb's unique identifier for the host/user |
| host\_name | string | Name of the host. Usually just the first name(s). |
| neighbourhood\_group | string | District where the listing located |
| neighbourhood | string | Sub-district where the listing located |
| room\_type | string | Type of the listing (Entire home/apt, Private room, Shared room, Hotel room) |
| construction\_year | integer | Year of construction of this listing |
| price | integer | daily price in USD |
| service\_fee | integer | One-time charge for cleaning |
| minimum\_nights | integer | minimum number of night stay for the listing |
| number\_of\_reviews | integer | The number of reviews the listing has |
| reviews\_per\_month | double | The number of reviews the listing has over the lifetime of the listing |
| availability\_365 | integer | The availability of the listing 365 days in the future as determined by the calendar. |

# Answer the following questions:

1. How many rooms are there in this dataset?
   1. 38247
   2. 46822
   3. 59462
   4. 62011
   5. 74928
   6. 85040
2. Calculate average service\_fee
   1. 46.74
   2. 62.96
   3. 88.36
   4. 107.54
   5. 125.22
   6. 144.07
3. How many rooms are the most expensive price?
   1. 3
   2. 15
   3. 24
   4. 47
   5. 61
   6. 89
4. Which neighbourhood\_group that has the most listings?
   1. Bronx
   2. Brooklyn
   3. Manhattan
   4. Queens
   5. Staten Island