Big data: architectures and data analytics

MapReduce - Exercises
Exercise #27

- Categorization rules
  - Input:
    - A large textual file containing a set of records
      - Each line contains the information about one single user
      - Each line has the format
        - `UserId, Name, Surname, Gender, YearOfBirth, City, Education`
    - A small file with a set of business rules that are used to assign each user to a category
      - Each line contains a business rule with the format
        - `Gender=<value> and DateOfBirth=<value> -> Category`
      - Rules are mutually exclusive

Exercise #27

- Output:
  - One record for each user with the following format
    - The original information about the user plus the category assigned to the user by means of the business rules
    - Since the rules are mutually exclusive, there is only one rule applicable for each user
    - If no rules is applicable/satisfied by a user, assign the user to the "Unknown" category
Exercise #27 - Example

- Users
  - User#1, John, Smith, M, 1934, New York, Bachelor
  - User#2, Paul, Jones, M, 1956, Dallas, College
  - User#3, Jenny, Smith, F, 1934, Philadelphia, Bachelor
  - User#4, Laura, White, F, 1926, New York, Doctorate

- Business rules
  - Gender=M and YearOfBirth=1934 -> Category#1
  - Gender=M and YearOfBirth=1956 -> Category#3
  - Gender=F and YearOfBirth=1934 -> Category#2
  - Gender=F and YearOfBirth=1956 -> Category#3

Exercise #27 - Example

- Output
  - User#1, John, Smith, M, 1934, New York, Bachelor, Category#1
  - User#2, Paul, Jones, M, 1956, Dallas, College, Category#3
  - User#3, Jenny, Smith, F, 1934, Los Angeles, Bachelor, Category#2
  - User#4, Laura, White, F, 1926, New York, Doctorate, Unknown
Exercise #28

- Mapping Question-Answer(s)
  - Input:
    - A large textual file containing a set of questions
      - Each line contain one question
      - Each line has the format
        - QuestionId, Timestamp, TextOfTheQuestion
    - A large textual file containing a set of answers
      - Each line contain one answer
      - Each line has the format
        - AnswerId, QuestionId, Timestamp, TextOfTheAnswer

- Output:
  - One line for each pair (question, answer) with the following format
    - QuestionId, TextOfTheQuestion, AnswerId, TextOfTheAnswer
Exercise #28 - Example

- Questions
  - Q1, 2015-01-01, What is ..?
  - Q2, 2015-01-03, Who invented ..

- Answers
  - A1, Q1, 2015-01-02, It is ..
  - A2, Q2, 2015-01-03, John Smith
  - A3, Q1, 2015-01-05, I think it is ..

Exercise #28 - Example

- Output
  - Q1, What is ..?, A1, It is ..
  - Q1, What is ..?, A3, I think it is ..
  - Q2, Who invented ..?, A2, John Smith
Exercise #29

- User selection
  - Input:
    - A large textual file containing a set of records
      - Each line contains the information about one single user
      - Each line has the format
        - UserId, Name, Surname, Gender, YearOfBirth, City, Education
    - A large textual file with pairs (UserId, MovieGenre)
      - Each line contains pair UserId, MovieGenre with the format
        - UserId, MovieGenre
      - It means that UserId likes movies of genre MovieGenre

Exercise #29

- Output:
  - One record for each user that likes both Commedia and Adventure movies
  - Each output record contains only Gender and YearOfBirth of the selected user
    - Gender, YearOfBirth
  - Duplicate pairs must not be removed
Exercise #29 - Example

- **Users**
  - User#1, John, Smith, M, 1934, New York, Bachelor
  - User#2, Paul, Jones, M, 1956, Dallas, College
  - User#3, Jenny, Smith, F, 1934, Philadelphia, Bachelor

- **Likes**
  - User#1, Commedia
  - User#1, Adventure
  - User#1, Drama
  - User#2, Commedia
  - User#2, Crime
  - User#3, Commedia
  - User#3, Horror
  - User#3, Adventure

Exercise #29 - Example

- **Output**
  - M, 1934
  - F, 1934