MapReduce in Python

In this exercise you will be able to implement relational operators using MapReduce.

Dataset. In this session we will use the Adult dataset[1], containing information about census and their income. You can check the files *adult.names* (located in the resources directory of the Python project) to get a better understanding of the schema of data being used. As input data, we provide you with a SequenceFile dataset (*adult.1000.sf*) where the key is a surrogate ID, and the value is a comma separated set of attributes conforming to the schema in *adult.names*. The following tuple is an example of the file:

('GtdDh4aF', '18,Local-gov,674771,Doctorate,8,Widowed,Wife,Other,Female,44859,8519,31,Yugoslavia')

Furthermore, we provide you with the method *Utils.get(array,attribute)*, which returns the projection for a specific attribute in the array. Note that the array should also contain the key as the first value (see the provided example).

Examples. We provide you with the implementation of the following operators:

- Projection
  - SELECT DISTINCT age, relationship, native_country FROM adult
- Cross Product
  - SELECT external.*, internal.*
    FROM adult as internal, adult as external
    WHERE external.native_country = "Italy" AND internal.native_country = "Ecuador"

Running the program. Using *python3*, execute the *Main.py* method and pass as parameter the desired operator.

Task. Implement the following operators, considering the following examples:

- Selection
  - SELECT * FROM adult WHERE workclass = "Private"
- Grouping[2]
  - SELECT native_country, list(capital_gain) FROM adult GROUP BY native_country
- Aggregation
  - SELECT native_country, SUM(capital_gain) FROM adult GROUP BY native_country
- Union
  - SELECT capital_gain FROM adult a1 WHERE native_country = "Italy"
    UNION
    SELECT capital_loss FROM adult a2 WHERE native_country = "Ecuador"
- Difference (based on one attribute)
  - SELECT age FROM adult a1 WHERE native_country = "Italy" EXCEPT
    SELECT age FROM adult a2 WHERE native_country = "Ecuador"
- Intersection (based on one attribute)
  - SELECT age FROM adult a1 WHERE native_country = "Italy" INTERSECT
    SELECT age FROM adult a2 WHERE native_country = "Ecuador"
- Join
  - SELECT external.*, internal.*
    FROM adult as internal INNER JOIN adult as external ON internal.marital_status = external.marital_status
    WHERE external.native_country = "Italy" AND internal.native_country = "Ecuador"

[1] https://archive.ics.uci.edu/ml/datasets/Adult

[2] Note this operation does not exist in standard SQL