Microsoft - DAT278x

From Graph to Knowledge Graph - Algorithms, Theory and Applications

LAB 1 – Graph

Construct an Author Collaboration Graph – Instructions

In this lab, you will construct a weighted and undirected author collaboration graph from a subset of the Microsoft Academic Graph (MAG). You will also explore this author collaboration graph to examine its properties.

The input to the exercise is a set of academic publications, each of which is associated its authors. The input data is formatted as a tab-separated value file with two columns. The first column represents the paper Id and the second column represents one of its associated authors' Ids. For example, below lists three papers' author information.

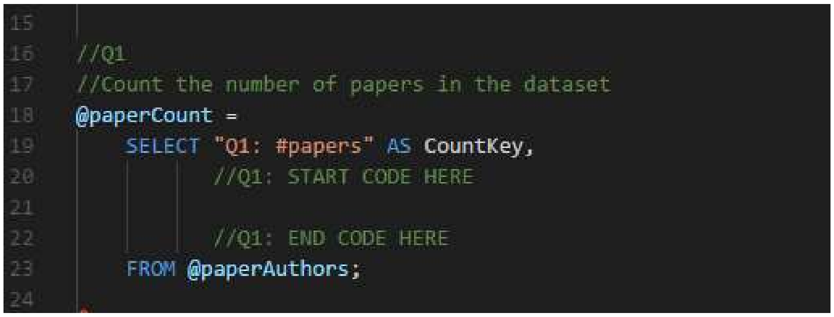
PaperId	AuthorId
120346	2604707830
120346	2708960186
161269	2034139186
161269	1984078314
161269	2079572972
161269	2163455850
672965	2403142114
672965	2630491343

The goal is to build a weighted and undirected author collaboration graph. Each edge is linked between two authors if they publish a paper together. The weight of each edge is defined as the number of collaborations between two authors. In this process, we will ignore papers with single authors.

The output could be the weighted edge list of the constructed graph. Note that you can also use different data structures to represent the graph.

In this Lab assignment, please follow the Azure Data Lake (ADL) setup guide to upload the input academic publication data (Paper_authors.tsv) into the ADL folder (course-kg/input/). The full file path is "/course-kg/input/Paper_authors.tsv".

The associated U-SQL script is “DAT278x_Lab1_Graph.usql”. In this Lab, you will fill in the missing code in this script to answer the questions. The missing code is surrounded with “//Q1: START CODE HERE” and “//Q1: END CODE HERE” if taking Lab Question 1 as an example in the following code segment.

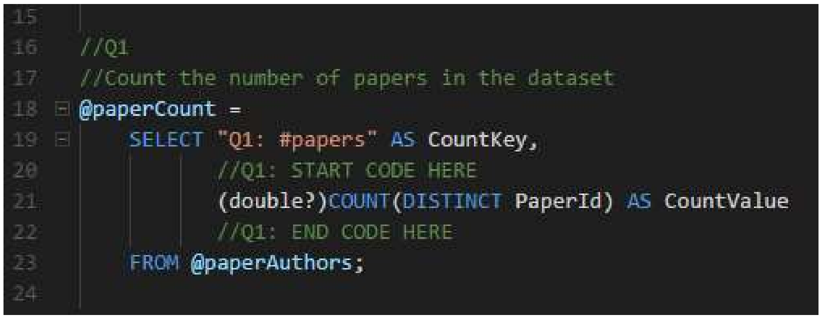


```
//Q1
//Count the number of papers in the dataset
@paperCount =
    SELECT "Q1: #papers" AS CountKey,
           //Q1: START CODE HERE

           //Q1: END CODE HERE
    FROM @paperAuthors;

```

After filling in the missing code “`(double?)COUNT(DISTINCT PaperId) AS CountValue`”, this part of the code will be completed as follows:



```
//Q1
//Count the number of papers in the dataset
@paperCount =
    SELECT "Q1: #papers" AS CountKey,
           //Q1: START CODE HERE
           (double?)COUNT(DISTINCT PaperId) AS CountValue
           //Q1: END CODE HERE
    FROM @paperAuthors;

```

To run the script, follow the Azure Data Lake setup guide. After filling in the missing code lines for all questions and then running the script, the author collaboration graph will be constructed and the answers to the questions will be output.