# Assignment 1

**Question 1:**

Consider the following data items where each item represents a record on a table; each record is have 3 values (Table name, PK, AttributeValue). The database schema of the tables can be described as follow:

T1(**A1**,A2)

T2(**A3**,A1)

The data items are as follow:

(T1, 1, x)
(T1, 2, x)
(T1, 3, x)
(T1, 4, y)
(T1, 5, y)
(T1, 6, y)
(T1, 7, z)
(T1, 8, z)
(T2, A, 1)
(T2, B, 2)
(T2, C, 3)

1- Write a map/reduce program to do an inner join between T1 and T2 where A1 in T1 is a foreign key in T2.

   The query results should be:
   (A, 1,X)
   (B, 2,X)
   (C, 3,X)

2- Write a map/reduce program to do a full outer join between T1 and T2 where A1 in T1 is a foreign key in T2.

   The query results should be:

   (A, 1,X)
   (B, 2,X)
   (C, 3,X)
   (null, 4, y)
   (null, 5, y)
   (null, 6, y)
   (null, 7, z)
   (null, 8, z)

3- Write a map/reduce program to find out find out the difference between two attributes. For example : A1[T1] – A1[T2] , The result would [4,5,6,7,8]

**Question 2:**

Given friendships with weighted relationships like:

(P1, P2, 0.8)
(P1, P3, 0.7)
(P3, P4, 0.9)
(P2, P4, 0.6)
(P2, P5, 0.5)
(P5, P6, 0.4)
(P4, P7, 0.3)
(P7, P8, 0.2)
(P6, P8, 0.1)

Find the friends of friends for a given person, aggregating the weights of the paths by summing them. For example, find the friends of friends for P1.

To calculate the aggregated weights, sum the weights along each path.

**Question 3:**

Given you have a file containing key/value pairs stored on Hadoop Distributed File System. Write a map/reduce program to sort the data in this file using Hadoop sorting mechanism by the key and store the results on Hadoop Distributed File System.

**Important notes:**

- This is a group assignment of 4 members and the members should be from the same group/lab.
- You should create a team from your section. So, all your team members should in one of the following section groups:
  - S1
  - S2, S3
  - ALL
- The team with less than 4 members will not be discussed (If you have a problem creating your team please contact your TA).
- All team members should work and fully understand everything in the assignment even if you distributed the questions, you should understand your colleague's questions.
- The assignment will be discussed in the week starting on Saturday, 19th of April. No late submission is allowed.
- Your submission should include a four **.jar** files for the 4 problems in the assignment.
- Add all jar files to a folder and compress it to a .zip. Rename the .zip file to be GroupNum_firstStudID_SecondStudID_ThirdStudID_FourthStudID. The compressed file would be the file to be delivered.
- Do not share your code with anyone, so that no other student would take your files and submit it under their names.
- Any cheating will be graded ZERO for both teams.
- Each team should discuss the assignment with his/her lab TA. Any team member who misses attending the discussion will take zero.