CMU MSP 36602 Pig HW Due 10:00 PM Tue March 26

EvilCorp cares nothing about our privacy. The have used various machine learning techniques to link the users ids from two different companies. Our task is to use Apache Pig to make a dataset from which we could compute a paired t-test of the per-customer total purchases from the two companies. The input datasets are:

company1.tab, **company2.tab**: tab-separated with columns for customer id, purchase amount, date (days since Jan. 1, 1970), and store number. There is one row per customer/purchase.

evilCorp.tab: tab-separated with columns company1Id and company2Id. Each row indicates a unique customer for which there is a high likelihood that the two ids from the two companies represent the same person.

You only need to work in Pig local mode.

Place all three input files in the “evil” subfolder of your working directory. Create the output in a subfolder of evil that the user can specify using a parameter named “out”. The columns of the output must be company1 id, sum of purchases from company 1, company2 id, and sum of purchases from company 2. The output must be in the form of a csv file (with no column headers).

Follow the template given in **pigHW.pig**.

Avoid any unnecessary computation. Don’t worry if the output shows a very small rounding error due to limitations of the IEEE754 floating point standard.