24.03.2025 Week 6 exercises: MapReduce

Exercise 1:

You are given a symmetric social network (like Facebook) where a is a friend of b implies that b is also a friend of a. The input is a dataset D (sharded) containing such pairs of identifiers (a, b)—no assumption is made regarding the order between a and b. Pairs appear exactly once and are not repeated. Find the last names of those users whose first name is "Kanye" and who have at least 300 friends. You can chain MapReduces if you want (but only if you must, and even then, the shortest possible chain).

You don't need to write code—pseudocode is fine as long as it is understandable. Your pseudocode may assume the presence of appropriate primitives (e.g., "`firstname(user_id)`", etc.). The Map function takes as input a tuple `(key = a, value = b)`.

Exercise 2:

For an asymmetrical social network, you are given a dataset D where lines consist of (a, b) which means user a follows user b. Write a MapReduce program (Map and Reduce separately) that outputs the list of all users U who satisfy the following three conditions simultaneously: i) user U has at least 2 million followers, and ii) U follows fewer than 20 other users, and iii) all the users that U follows, also follow U back.

Exercise 3 (Final Exam 2020)**:**

Matrix multiplication is a fundamental operation in machine learning. Assume we now want to perform matrix multiplication over two extremely large matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

Design an $O(1)$-round Map-Reduce program for computing a matrix-matrix product $\boldsymbol{M} = \boldsymbol{AB}$, where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. is an m x n matrix and $\boldsymbol{B} \in \mathbb{R}^{n \times m}$.

Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are represented through mn pairs, and each pair corresponds to $(A, i, j, \boldsymbol{A}[i, j])$ or $(B, i, j, \boldsymbol{B}[i, j])$ (the first three entries are string while the last one is a floating point). The Map function reads each available pair as the input.

Exercise 4 ([10 points] Final Exam 2022 - multiple choice: Only 1 correct answer. No negative marking.)**:**

1) For each of the following questions, compute the total communication cost between the mappers and the reducers, i.e., the total number of (key, value) pairs outputted by the mappers. Use the optimal algorithm that runs in one MapReduce step for each of the tasks. Also, assume that there is no combiner.

(a) [2 points] Word count for a dataset comprising W total words with d distinct words. The mappers receive a single word as input.

( ) d

( ) W

( ) $\frac{W}{d}$

( ) dW

( ) $2W$

(b) [3 points] Matrix multiplication of two matrices of size $m \times n$ and $n \times p$. The mappers read input tuples in the form `<Matrix identifier, row index, column index, value>`.

( ) mp

( ) $n(m+p)$

( ) $2n(m+p)$

( ) mnp

( ) $2mnp$

(c1) [2 points] Computing the INNER JOIN of two relations defined as follows: relation `R1(X, Y)` with four tuples {(5,21), (7,16), (15,3), (3,21)} and relation `R2(Y, Z)` with {(3,1), (4,8), (21,28)}. The mappers read input tuples in the form `<Relation identifier, column-1 value, column-2 value>`.

Recall: The database INNER JOIN operation of two relations outputs concatenated tuples from R1 and R2 having equal values on the common column.

( ) 12

( ) 7

( ) 4

( ) 3

( ) 2

(c2) [1 point] How many output tuples are produced by the reducer(s) in question (c1) ?

( ) 12

( ) 7

( ) 4

( ) 3

( ) 2

(d) [2 points] Difference of two sets X and Y containing a total of x and y elements respectively. The mappers read input tuples in the form `<Set identifier, Value>`.

( ) $x+y$

( ) $x-y$

( ) x

( ) y

( ) xy