Fall 2016

Big Data Management

Learning objectives of Lecture 6 (5.12.2016)

Department of Computer Science
University of Helsinki

General Instruction: After the sixth lecture of big data course, you should be able to answer the following questions.

1. Show how to use a MapReduce program to compute the co-occurrence matrix for a large collection of documents. For example, given three documents: D1: (Donald Trump President), D2 (USA President Trump), D3: (Donald Trump USA) , the co-occurrence matrix is :

	Donald	Trump	President	USA
Donald	2	2	1	1
Trump	2	3	2	2
President	1	2	2	1
USA	1	2	1	2

2. Show how to use a MapReduce program to compute the join of two tables. For example, given two tables X and Y, the join results are shown in the table Z:

Table X:

A	B
1	ab
1	cd
4	ef

Table Y:

A	C
1	b
2	d
4	c

Table Z:

A	B	C
1	ab	b
1	cd	b
4	ef	c

Considering the following three cases, design three different MapReduce programs to efficiently handle them.

(1) Both table X and Y are big tables and they are distributed in multiple nodes in a computer cluster;

(2) Table X is a big table, but table Y is very small;

(3) Table X is a big table, but table Y is a medium-size one. (Hint: consider to use a bloom filter in this case)