Distributed Data Management SS 2015
Prof. Dr.-Ing. Sebastian Michel
MSc. Evica Milchevski, MSc. Kiril Panev
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
Sheet 4: Handout 28.05.2015, Presentation 09.06.2015

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN
http://dbis.informatik.uni-kl.de

Assignment 1: Analytics with PIG and Min-Hashing (1 P.)

(a) Reconsider the weather data CSV file from exercise sheet 2. The data stored in the file are of the following format: ***month/day/year;station;hour;temperature***, here is an example:

Example:

1/1/2000;1;1;1.588586391772654
2/1/2000;2;3;1.981028401924819
2/1/2000;2;4;1.875632896548555
...

Implement a PIG script that is computing for each hour of the day the first and third quartile of the temperature. To do so, implement a user defined function (UDF) in Java that is computing the quartiles. Test it in your own Hadoop/PIG installation (you can also use the Hortonworks virtual machine) and demonstrate the solution in the exercise session. Note that you are not allowed to use any external libraries that already implement a UDF's for computing the quartiles.

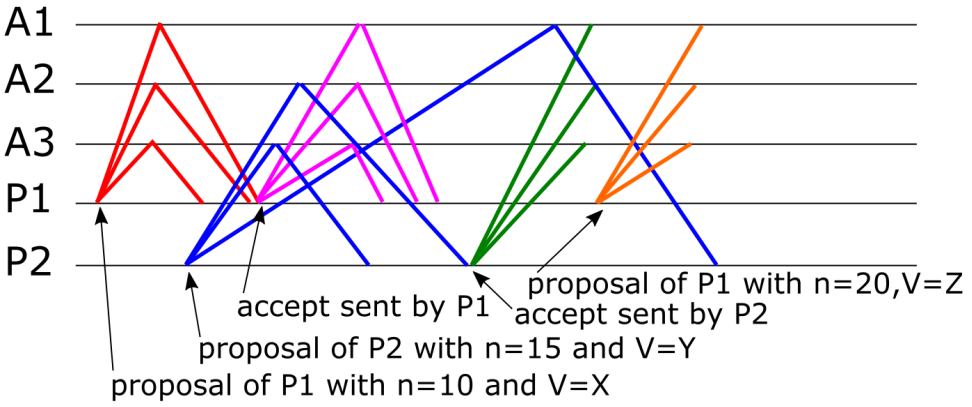
(b) Consider the file docs.csv, provided on the course website, with the following structure:

document_id;term_id

The file contains pairs of documents and the terms that appear in them. In this task you will need to implement Min-Hashing using PIG. First define a UDF that will be applied on the input and computes for the given terms their hash outputs. To make it simple, use only 2 hash functions. Then, use PIG to compute the documents' min-hash signatures and all pairs of documents that have at least one overlapping element in their min-hash signatures.

Assignment 2: Consensus with Paxos (1 P.)

(a) Given the communication across 2 proposers (i.e., P_1 and P_2) and 3 acceptors (i.e., A_1, A_2, and A_3), depicted in the following illustration. First, describe how Paxos acts in each of these communications indicated, and what is the state of the acceptors after each received message. Then, explain a possible way the Paxos consensus algorithm, as described in the lecture, proceeds in this situation (i.e., how the accept and propose messages of P2, respectively, P1 are handled. The horizontal lines represent the evolving time. Assume that in the beginning the acceptors did not see or accept any proposal.



(b) For each of the following states of Paxos accepters, verify if this state can happen when having at least 2 proposers and the starting state of the acceptors is to have no value chosen. If not explain why this state cannot happen. If yes, write an explicit example of communication between the introduced proposers and the three acceptors that leads to this state.

Scenario 1:

A_1	A_2	A_3
$Np = 5$	$Np = 5$	$Np = 6$
$Na = 5$	$Na = 5$	$Na = 6$
$Va = X$	$Va = X$	$Va = Y$

Scenario 2:

A_1	A_2	A_3
$Np = 5$	$Np = 6$	$Np = 7$
$Na = 5$	$Na = 6$	$Na = 6$
$Va = X$	$Va = Y$	$Va = Y$

Assignment 3: Investigating NoSQL Systems (1 P.)

We have briefly discussed in the lecture the wide spectrum of available NoSQL solutions. The website under the URL `http://nosql-database.org/` lists quite many of these.

Familiarize yourself with each of the following systems and its core characteristics.

x	**system**
0	Cassandra
1	MongoDB
2	Riak
3	Neo4J

For one of the systems, during the exercise session you will be asked to present a short overview essay, where you describe the type of system it represents, the core characteristics, prominent "customers" (if any), etc. This essay is to be orally presented using, if required, the blackboard.

To find the system you should present, take the upper-case first character of your last name and take its ASCII code modulo 4. According to the result, x, the system to consider in your essay is given in the above table.