Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Computer Science and Engineering

BACHELOR PROJECT ASSIGNMENT

Student: **Michal Stanke**

Study programme: Open Informatics
Specialisation: Software Systems

Title of Bachelor Project: **Fulltext search in HBase database**

Guidelines:

1. Implement a full-text search in texts (e.g. URLs) stored in a big-table HBase database. We suggest the use of MapR Hadoop Distribution.
2. Prepare a set of "real-world" text (e.g. URLs similar to real ones with representative structure of domains, paths and parameters) or use existing open datasets. The volume of the database should be in tens or hundreds gigabytes.
3. Design a suitable table scheme to decompose the text and store it in a searchable way.
4. Find appropriate and openly available indexing engine to index and search the text stored in designed schema and implement a CLI as an interface to the search engine and related instrumentation.
5. Determine the performance of the system in load time (including any indexing) and query time. Benchmark with full scan and possibly other alternate solutions.

Bibliography/Sources:

Nick Dimiduk, Amandeep Khurana: HBase in Action; Manning Publications; 1 edition (November 17, 2012)

Lars George; HBase: The Definitive Guide; O'Reilly Media; 1 edition (September 23, 2011)

MapR User Documentation: available at http://doc.mapr.com/display/MapR/Home, accessed Jan 2015

Bachelor Project Supervisor: Ing. Martin Rehák, Ph.D.

Valid until the end of the summer semester of academic year 2015/2016

L.S.

doc. Ing. Filip Železný, Ph.D.
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, March 25, 2015