Institut für Informationssysteme
Technische Universität Braunschweig

Technische Universität Braunschweig
Institut für Informationssysteme
http://www.ifis.cs.tu-bs.de

Wolf-Tilo Balke, Christoph Lofi

Distributed Data Management

Sheet 12 (until **29.07.2010** – Two weeks)

This exercise is optional and will provide only **bonus points**. You may hand it in via Email to lofi@ifis.cs.tu-bs.de .

(**35 points bonus**)

Using the HBase and Hadoop appliance used in the last exercise, develop a **Java program** which does the following:

- Import the following **CSV** file into an according **HBase** table: http://www.ifis.cs.tu-bs.de/webfm_send/517
  - The file is an excerpt from IMDB, the columns are as follows:
    Movie Name, Movie Year, Average IMDB Rating (0-10), Number of Rating Votes, Name of Director
- Create a Hadoop **Map & Reduce Task** which generates a statistic containing a row for each year containing the year and the average movie rating of all movies released in that year and having more than 200 rating votes, e.g. (“2011”,”7.0”); (“2012”,”6.1”); …
- Copy & Paste the statistic into your solation, also attach the Java program.

Refer to the HBase, Hadoop and Cloudera help and tutorial web pages for assistance.
You may use the pre-configured Eclipse installation delivered by the Cloudera VM.