Technische Universität Berlin
Database Systems and Information Management Group
Prof. Dr. Volker Markl, Sebastian Schelter, Christoph Boden, Thomas Bodner

AIM3: Scalable Data Mining and Data Analysis
Summer term 2012

Second Assignment

PACTs and Stratosphere

Due on May 16th

Basics of PACTs and Stratosphere

With this assignment, we will exercise the more general PACT programming model of the Stratosphere system.

1. **Average temperature per month**

   Perform the "average temperature per month"computation from the first assignment here using the second-order functions map and reduce in Stratosphere instead of Hadoop.

   Implement the missing parts in *AverageTemperaturePerMonthPact*.

2. **Join books and authors using a PACTs**

   We will peform the join of books and authors from the first assignment here using the second-order function match. Match will read in the different tuple sets and supply all pairs having the same key to a user-defined function which provides a much easier way of joining than plain Hadoop MapReduce.

   Implement the missing parts in *BookAndAuthorJoinPact*.

3. **Produce Cartesian product with Cross PACT**

   Next we deal with the files *crossLeft* and *crossRight* in the folder *src/test/resources/assignment2*. Both files contain data of the format:

   key[TAB]value

   Your task is to implement a PACT program that computes the Cartesian product of the two inputs. The values of the joined tuples shall be summed up:

   key1[TAB]kex2[TAB]value1+value2

4. **Group multiple input sets with CoGroup PACT**

   In this task, we use the files *coGroupLeft* and *coGroupRight* in the folder *src/test/resources/assignment2*. Both files contain data of the format:

   key[TAB]value

   You must implement a PACT program that groups the two inputs on their keys and, in addition, computes an aggregate value whereas the values of the left input shall be added to the running aggregate and the values of the right input substracted from it:

   key1[TAB]key2[TAB]sumOfLeftValues-sumOfRightValues

Deadline

Source code for the exercises is available at *https://github.com/dimalabs/scalable-datamining-class*.

Register in the ISIS information system at *https://www.isis.tu-berlin.de/course/view.php?id=6535* and upload your solution in the form of a patch file until May 16th.