1 Parallel Data Processing and Analysis

<table>
<thead>
<tr>
<th>Module</th>
<th>Code</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDM</td>
<td>Parallel Data Processing and Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Credit Points</th>
<th>Workload</th>
<th>Duration</th>
<th>Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>180 h</td>
<td>1 Semester</td>
<td>-</td>
</tr>
</tbody>
</table>

Program | Master Scientific Computing  
Methods | Lecture 2 h + Exercise course 2 h  
Lecturer | Prof. Dr. Artur Andrzejak  
Objectives | To have a firm knowledge of parallel data processing and analysis, in particular of very large data sets.

Content | This module covers the following topics:
- programming paradigms for parallel-distributed data processing, especially Map-Reduce and Spark programming models
- usage of tools like Apache Spark, Hadoop, Pig, Hive, and possibly other frameworks for parallel-distributed data processing
- application cases in parallel data analysis, for example clustering, recommendation, search for similar objects, mining of data streams
- techniques for parallel pre-processing of data
- fundamentals of analysis techniques such as classification, regression, clustering and evaluation of the results
- parallel algorithms for data analysis and their implementations
- theory and practice of scalability and tuning of frameworks.

Learning outcomes | • Knowledge of selected approaches and programming paradigms of parallel data processing
• Knowledge how to use tools for parallel data processing (among others Apache Hadoop and Spark)
• Familiarity with application domains of big data analysis
• Knowledge of methods of parallel pre-processing of data
• Knowledge of methods like classification, regression, clustering and their parallel implementations
• Knowledge of scaling of parallel algorithms.

Prerequisites | None

Suggested previous knowledge | Knowledge of Java/Python and in elementary probability theory / statistics; module IBD can be taken as a complement / extension.

Assessment(s) | Successful participation in the exercises with homework (achieving a minimum score of at least 50% of points) and passing a final exam.

Literature | Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014 (online)
Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, (third edition), 2012