# Big Data Analytics

# Exercise Sheet 2

Prof. Dr.Dr. Lars Schmidt-Thieme, Mohsan Jameel

Information Systems and Machine Learning Lab University of Hildesheim

April 21st, 2016

Submission until April 27th, 2016, to mohsan.jameel@ismll.de

## Exercise 1: Parallel Program Design (5 points)

a) What are the different design consideration while developing parallel programs
b) Explain with example the difference between data partitioning and function partitioning.

## Exercise 2: Measuring Parallel Program Efficiency (5 points)

a) With help of formulas calculate parallel speedup and efficiency for the following two programs? Also categorize the speedup of each program i.e. linear, super linear or sublinear.

| Program | Data size | Ts (sec) | $T_{P2}$(sec) | $T_{P4}$ (sec) | $T_{P8}$ (sec) | $T_{P16}$ (sec) |
|---|---|---|---|---|---|---|
| Task A | 100,000 | 432 | 220 | 140 | 90 | 74 |
| Task B | 500,000 | 610 | 305 | 150 | 75 | 40 |

## Exercise 3: Parallel KMeans Clustering (5 points)

Sketch the parallel algorithm for KMeans cluster. Explain your design choice, such as what are the hotspots, partitioning strategy and any synchronization required. You can present your solution with help of pseudocode, flow chart or block diagram.