# Big Data Management and Analytics

| | | |
|---|---|---|
| Titel | Big Data Management and Analytics | |
| Typ | Vorlesung mit Übung | |
| Credits | 6 ECTS | |
| Lehrform/SWS | 3V + 2Ü | |
| Sprache | Englisch | |
| Modulniveau | Master | |
| Arbeitsaufwand | Präsenzstunden | 75 Stunden |
| | Eigenstudium | 105 Stunden |
| | Gesamtaufwand | 180 Stunden |
| Angestrebte Lernergebnisse | Die Studierenden sind in der Lage die Anwendungsgebiete für Big Data Systeme von Anwendungen herkömmlicher Datenbanksysteme abzugrenzen und die Vorteile der verschiedene Big Data Systeme wiederzugeben. Die Studierenden verstehen die Funktionsweise von parallelen Batch-Systemen und parallelen Streaming Systemen und können diese installieren und verwenden, um massive Datensätze zu verarbeiten. Des Weiteren werden die Studenten in die Lage versetzt Techniken für die Analyse großer Datenbestände anzuwenden und passende Verfahren für eine gegebene Anwendung auszuwählen. | |
| Intended Learning Outcomes | The students are able to distinguish application areas for big data systems from those of ordinary database systems. Furthermore, the students understand the advantages of various big data systems. They are able to reproduce the functionalities of parallel batch processing systems and parallel streaming systems and are able to use these systems to implement big data applications. Additionally, students have the ability to apply techniques for large scale data analysis to massive datasets and select the best fitting solution approach for a given problem setting. | |
| Inhalt | Einführung in das Themengebiet Big Data: Datenquellen, Eigenschaften von Daten, modellgetriebene und datengetriebene Problemlösungen, neue Hardware Architekturen.<br><br>NoSQL Datenbanken: BASE Konzept, Abgrenzung zum ACID Prinzip, CAP Theorem, Einordnung existierender Systeme, Wide-Column Stores, Document Stores, Key-Value Stores, Graph-Datenbanken<br><br>Systeme zur Batch-Verarbeitung: verteilte Filesysteme, | |

| | |
|---|---|
| | Map-Reduce Programmierung, Systemarchitektur von Map-Reduce Systemen, Apache SPARK, parallele Data Mining Algorithmen mit Map-Reduce<br><br>Stream Processing: Streaming-Modelle, parallele Streaming Systeme (z.B. Spark Streaming, Apache Flink, Apache Storm,..), Analysealgorithmen für Streams<br><br>Datenanalyse Methoden für massive Datenbestände: Methoden für hochdimensionale Daten(z.B. PCA mit Power Iterations, CUR), Textdaten (z.B. Shingling), Graphdaten (z.B. Pagerank, Random Walk with Repeat) Shingling, Feature Reduction, Node Rankings<br><br>Optional: Analysemethoden für neue Hardwarearchitekturen (z.B. GPU-Computing) |
| Contents | Introduction to BIG Data: Data sources, characteristics of big data, model driven and data driven problem solutions, new hardware environments<br><br>NoSQL databases: BASE concept, differences to ACID, CAP theorem, classification of big data systems, wide-column Stores, document stores, key-value stores, graph databases<br><br>Massive batch processing systems: distributed file systems, map-reduce programming, system architecture of batch processing systems, Apache SPARK, parallel data mining algorithms based on map reduce<br><br>Stream processing: streaming models, parallel streaming systems (for examples Spark Streaming, Apache Flink, Apache Storm,..), data analysis on massive streams<br><br>Data analytics for massive data sets: high-dimensional data (e.g. PCA via power iterations, CUR), text data (e.g. shingling), graph data (e.g. page rank, random walk with repeat) |
| Prüfung | Prüfungsleistung (benotet):<br>-Klausur: 90-180 min<br><br>Wiederholungsklausur zu Ende des Semesters.<br>Details werden zu Beginn des Moduls bekannt gegeben.<br><br>In der Klausur weisen die Studierenden nach, inwieweit sie die vorgestellten Modelle, Methoden und Algorithmen verstanden haben, komprimiert wiedergeben, anwenden sowie auf verwandte |

| | |
|---|---|
| | Problemstellungen übertragen können. In der Klausur werden 7 bis 10 Aufgaben gestellt, die eine eigenständige Anwendung der Modelle und Verfahren aus der Vorlesung erfordern (wie z.B. programmieren in Apache Spark, Zentralitätsmaße in Netzwerken, exponential histograms..) |
| Examination | Examination requirements (graded):<br>- written exam: 90-180 min<br><br>A makeup exam will be offered at the end of the semester, details will be announced at the beginning of the course.<br><br>Within the written exam, students demonstrate that they understand the presented processes, models, and methods, that they can reproduce and apply them as well as that they can transfer and extend models and methods to similar problems. The written exam consist of 7 to 9 assignments, which require independent application of models, and methods presented in the lecture (e.g. coding Apache Spark, computing centrality measures, exponential histograms..). |
| Literatur/Literature | Jure Leskovec, Anand Rajaraman, Jeff Ullman: Mining of Massive Datasets, Cambridge University Press; 2014 |
| Medienformen | Beamer-Präsentation, Tafelpräsentation |
| Media | slides show, blackboard presentation |
| Lehr- und Lernmethode | Vorlesung, Übung, Aufgaben zum Selbststudium.<br>Das Modul besteht aus einer Vorlesung und Übungen in kleinen Gruppen (als Tutorübungen).<br>In den Hausaufgaben, die freiwillig abzugeben sind, analysieren die Studierenden die in der Vorlesung vorgestellten Prozesse, Modelle und Verfahren, wenden diese auf konkrete Daten an und erweitern diese für ähnliche Problemstellungen. In den Hausaufgaben werden selbständig anspruchsvolle Übungsaufgaben bearbeitet, die ähnlich zu den Klausuraufgaben sind (siehe oben) und daher zur Vorbereitung darauf dienen.<br>In der Übung werden mögliche Lösungsstrategien diskutiert. |
| Teaching and Learning Methods | Lecture, tutorial, assignments for individual study.<br>Within the assignments (the submission is optional) students analyze the processes, models, and methods presented in the corresponding lectures, apply them to real data, and extend these to similar problems. The assignments consist of demanding problems similar to the assignments in the written exam (for details see |

| | |
|---|---|
| | above) and serve as a preparation for the exam. Within the tutorials possible approaches for solutions of the assignments will be discussed. |
| Turnus | Wintersemester |
| Modulverantwortlicher | Prof- Dr. Matthias Schubert |
| Dozenten | Prof. Dr. Matthias Schubert |