BDAPRO: Big Data Analytics Project (6 PJ) 0434 L 484

Content:

Both the sciences and industry are currently undergoing a profound transformation: large-scale, diverse data sets - derived from sensors, the web, or via crowd sourcing - present a huge opportunity for data-driven decision making. This data poses new challenges in a variety of dimensions: in its unprecedented volume, in the speed at which it is generated (its velocity) and in the variety of data sources that need to be integrated. A whole new breed of systems and paradigms is currently developed to be able to cope with that these challenges. The field of Big Data Analytics deals with the technological means of gaining insights from huge amounts of data. Students will conduct projects that deal with applying data mining algorithms to large datasets. For that, students will learn to use so called Parallel Processing Platforms (e.g. Flink, Spark, Hadoop, HBase), systems that execute parallel computations with terabytes of data on clusters of up to several thousand machines. At the start of the project, a student will receive a topic as well as some information material. The team, with the assistance of the lecturer, will decide on a project environment with the suitable tools for team work, project communication, development and testing. Next, the problem will have to be analyzed, modelled and decomposed into individual components, from which tasks are derived that are subsequently assigned to smaller teams or individuals. At weekly project meetings, the project team presents progress and milestones that have been reached. In consultation with the lecturer, it is decided which further steps to take. The project is concluded with a final report as well as a final presentation which includes a demonstration of the prototype.

In this course you will learn to systematically analyze a current issue in the information management area and to develop and implement a problem-oriented solution as part of a team. You will learn tocooperate as team member and to contribute to project organization, quality assurance and documentation. The quality of your solution has to be proven through analysis, systematic experiments and test cases. Examples of IMPRO projects carried out in recent semesters are a tool used to analyse Web 2.0 Forum data, an online multiplayer game for mobile phones, implementation and analysis of new join methods for a cloud computing platform or the development of data mining operations on the massively parallel system Hadoop as part of the Apache open source project Mahout.

After the course, students will be able to understand methods for large scale data analytics and to solve large scale data analytics problems. They will be capable of designing and implementing large scale data analytics solutions in a collaborative team. Students will conduct projects that deal with applying data mining algorithms to large datasets. For that, students will learn to use so called Parallel Processing Platforms, systems that execute parallel computations with terabytes of data on clusters of up to several thousand machines.

At the start of the project, a student will receive a topic as well as some information material. The team, with the assistance of the lecturer, will decide on a project environment with the suitable tools for team work, project communication, development and testing. Next, the problem will have to be analyzed, modelled and decomposed into individual components, from which tasks are derived that are subsequently assigned to smaller teams or individuals. At weekly project meetings, the project team presents progress and milestones that have been reached. In consultation with the lecturer, it is decided which further steps to take. The project is concluded with a final report, a project poster as well as a final presentation which includes a demonstration of the prototype.

Target group:

This course addresses master students with a focus on database systems and information management after the first (master) term in “Informatik”, “Technische Informatik”, “Wirschaftsingenieurwesen”. (If capacity is available, it will be open also for other faculties).

Prerequisite:

Knowledge from the complete Bachelor program (Informatik or Technische Informatik) is required, as well as experiences in programming, software development, linear algebra and statistics. Depending on the topic, additional prerequisites may be required, e.g. „IDB – Implementation of Database Systems".

Solid programming skills in at least one of the following programming languages: Java, C++, Scala, Python.

Basic knowledge in functional programming.

Basic knowledge in distributed source control management systems (Git, Mercurial) and software processes like Scrum.

Registration:

Students are required to register via the DIMA course registration tool before the start of the first lecture (http://www.dima.tu-berlin.de). Within the first six weeks after commencement of the lecture, students will have to register for the course at **QISPOS (university examination protocol tool) and ISIS (course organization tool)** in addition to the registration at the DIMA course registration tool.

Contributions:

Prüfungsform: Portfolioprüfung

The overall grade for the module consists of the results of the course work ('portfolio exam'). The following are included in the final grade:

1. Prototype with test cases and documentation (30p.)
2. Experiment design and execution (20p.)
3. Intermediate presentation (10p.)
4. Experiments analysis (20p.)
5. Final presentation (20p.)

The final grade according to § 47 (2) AllgStuPO will be calculated with the faculty grading table 2.
(Die Gesamtnote gemäß § 47 (2) AllgStuPO wird nach dem Notenschlüssel 2 der Fakultät IV ermittelt.)

Prüfungselement	Gewicht
(Deliverable assessment) Experiments analysis	20
(Deliverable assessment) Final presentation	20
(Deliverable assessment) Intermediate presentation	10
(Learning process review) Experiment design and execution	20
(Learning process review) Prototype with test cases and documentation	30

Short Comment:

Duration of this module is one term.
Das Modul kann in 1 Semester(n) abgeschlossen werden.
Das Modul ist auf 24 Teilnehmer begrenzt.
The lab capacity limits this course to max. 24 participants.

Contact persons:

Bonaventura Del Monte, Jonas Traub, Dr. Alireza Rezaei Mahdiraji, Behrouz Derakhshan, Dr. Quoc Cuong To, Prof. Dr. Volker Markl