CS 520 Data Integration, Warehousing, and Provenance

Course Description:

This course introduces the basic concepts of **data integration**, **data warehousing**, and **provenance**. We will learn how to resolve structural heterogeneity through schema matching and mapping. The course introduces techniques for querying several heterogeneous datasources at once (**data integration**) and translating data between databases with different data representations (**data exchange**). Furthermore, we will cover the data-warehouse paradigm including the **Extract-Transform-Load (ETL)** process, the **data cube** model and its relational representations (such as snowflake and star schema), and efficient processing of analytical queries. This will be contrasted with **Big Data analytics** approaches that (besides other differences) significantly reduce the upfront cost of analytics. When feeding data through complex processing pipelines such as data exchange transformations or ETL workflows, it is easy to loose track of the **origin of data**. In the last part of the course we therefore cover techniques for representing and keeping track of the origin and creation process of data - aka its **provenance**.

The course is emphasizing practical skills through a series of homework assignments that help students develop a strong background in data integration systems and techniques. At the same time, it also addresses the underlying formalisms. For example, we will discuss the logic based languages used for schema mapping and the dimensional data model as well as their practical application (e.g., developing an ETL workflow with rapid miner and creating a mapping between two example schemata). The literature reviews will familiarize students with data integration and provenance research.

Course Material:

The following text book will be helpful for following the course and studying the presented material. Given the lack of text books on the topic, the part on data provenance will require literature study.

Doan, Halevy, and Ives. **Principles of Data Integration**, 1st Edition, Morgan Kaufmann, 2012

One of the following standard text books on databases in general may be helpful. However, this is not required reading material.

Elmasri and Navathe. **Fundamentals of Database Systems**, 6th Edition , Addison-Wesley , 2003
Ramakrishnan and Gehrke. **Database Management Systems**, 3nd Edition , McGraw-Hill , 2002
Silberschatz, Korth, and Sudarshan. **Database System Concepts**, 6th Edition , McGraw Hill , 2010
Garcia-Molina, Ullman, and Widom. **Database Systems: The Complete Book**, 2nd Edition, Prentice Hall, 2008

The slides will be made available on the course webpage.

Prerequisites:

- *Courses:* CS425

Course Details:

The following topics will be covered in the course:

- Introduction (3 hours)
  - Heterogeneity of data
  - Uncertainty and incompleteness
  - Autonomous and distributed data sources
  - Structured vs. unstructured data
- Preprocessing and Cleaning (4 hours)
  - Entity resolution
  - Data fusion
  - Cleaning
- Data Integration (10 hours)
  - Mediated schemata and query rewrite
  - Schema matching
  - Schema mappings
- Data Exchange (5 hours)
  - Data exchange transformations
  - Universal solutions
- Data Warehousing (10 hours)
  - Extract-transform-load (ETL)
  - Data cubes
  - Star- and snowflake schemas
  - Efficient analytics (OLAP) and relationship to transactional relational systems (OLTP)
- Big Data Analytics (3 hours)
  - Big Data analytics platforms and programming models
  - Differences between Big Data analytics and traditional warehousing approaches
  - Big Data integration
- Data Provenance (10 hours)
  - Why- and Where-provenance
  - Provenance polynomials
  - Provenance in data integration
  - Provenance for missing answers

Workload

The workload will consist of

1. A final exam covering the topics of the course.
2. Several homework assignments. Over the course of the assignments the students will build a small data integration pipeline - starting from preprocessing, over schema matching and mapping, to transformation and query rewrite.
3. Review of one or two research papers related to data integration or provenance.

Course Objectives:

After attending the course students should:

- Understand the problems that arise with querying heterogeneous and autonomous data sources
- Understand the differences and similarities between the data integration/exchange, data warehouse, and Big Data analytics approaches
- Be able to build parts of a small data integration pipeline by "glueing" existing systems with new code
- Have learned formal languages for expressing schema mappings
- Understand the difference between virtual and materialized integration (data integration vs. data exchange)
- Understand the concept of data provenance and know how to compute provenance