GESIS Fall Seminar in Computational Social Science 2022

Syllabus for week 2:
"Big Data Management and Analytics"

Lecturers:	Prof. Dr. Rainer Gemulla	Adrian Kochsiek
Affiliation:	University of Mannheim	University of Mannheim
Email:	rgemulla@uni-mannheim.de	adrian.kochsiek@uni-mannheim.de

Date: September 12-16, 2022
Time: 09:00-16:00

About the Lecturers

Rainer Gemulla is a professor at the university of Mannheim, where he heads the Data Analytics group. His research focuses on methods and systems for scalable data analytics, data mining, and machine learning. Recent examples include work on representation learning for knowledge graphs (e.g., the LibKGE framework) and on efficient parameter management for parallel machine learning (e.g., the NuPS parameter server). Rainer has published numerous papers in top-ranked international conferences (including ACL, AAAI, ICDE, ICDM, ICLR, PVLDB, KDD, SIGMOD) and journals (including TKDD, TODS, VLDB Journal) and frequently serves on program committees or as area chair for conferences related to data analytics. He has been appointed a junior fellow of the German Informatics Society in 2013 and received multiple awards for his work.

Adrian Kochsiek is a PhD student at the university of Mannheim and a member of the Data and Web Science Group. His research focuses on representation learning for large-scale knowledge graphs and he has published his work in top-ranked international conferences (PVLDB, ACL, EMNLP). Next to research, Adrian is responsible for tutorials and assignments of the course "Large Scale Data Management" in the Mannheim Master of Data Science program.

Course Description

This course introduces systems and techniques for storing, querying, and working with datasets that are too large, too complex, or simply too inconvenient to work with on a single machine or programming language. Participants learn the foundations necessary to work with available "Big Data systems" on their own, whether in a local installation or via cloud computing. It is organized in a workshop format, i.e., morning sessions that introduce and discuss key concepts and techniques, followed by practical sessions in which participants gain hands-on experience on selected systems and applications. The course makes use of Python as the main programming language; it's one of the most suitable languages for data science with large, complex datasets.

We start with an introduction (or refresher, depending on the participant's background) of processing structured data (e.g., data frames), first directly within Python, then using a relational database system and the SQL query language for data access. Building on these foundations, the course introduces the large-scale computation engine Apache Spark for pre-processing and analysing data in a scalable fashion. We subsequently introduce and discuss non-relational data representation formats that are suitable for more complex data, most notably JSON (JavaScript Object Notation, for semi-structured data and documents) and, if time permits, RDF (Resource Description Framework, for graph data and knowledge graphs). The course concludes with an introduction into selected NoSQL databases that are useful for managing such data.

Keywords

Big Data, Database Systems, Data Formats, Data Processing

Course Prerequisites

- We assume that participants already have experience with programming (e.g., in Python or R).
- Although we will discuss fundamental aspects of working with data in Python, we highly recommend those with no experience with Python to take part in the course Introduction to CSS with Python (week 1) or familiarize themselves with the very basics of Python.
- Participants are expected to have a working Python environment installed (see below).

Target Group

Participants will find the course useful if:

- They want to work with large and/or complex datasets.
- They want to leverage available data management and processing solutions (either locally or in the cloud) for improved efficiency and ease of use.

Course and Learning Objectives

By the end of the course participants will:

- Understand different data representations (including relational data, semi-structured data, and graph data) and their advantages/disadvantages.
- Be able to process structured data in Python (using Pandas).
- Know how to insert, update, and query structured data in a relational database system using the SQL query language (using MariaDB).
- Be familiar with the Apache Spark framework for performing computations on large datasets.
- Be able to perform basic parallel data processing with Apache Spark.
- Know basic types of NoSQL systems as well as their properties.
- Be able to store, query, and process semi-structured data in a NoSQL database (using MongoDB).

Organizational Structure of the Course

The course is organized in a workshop format with 6 hours per day. Each day, we introduce and discuss key concepts and techniques in the morning, followed by practical sessions in the afternoon. In the latter sessions, participants gain hands-on experience on selected systems and applications through exercises and practical assignments. Lecturers will be available throughout to provide guidance and for individual consultations.

Software and Hardware Requirements

Participants need to bring a laptop and have a Python 3 environment installed. Additional installation instructions (i.e., additional Python packages) will be provided later on. The required data management software (such as MariaDB, Apache Spark or MongoDB) will be hosted on our servers and does not need to be installed locally.

Day-to-Day Schedule

Day 1: Introduction & Relational Data in Python

- Introduction and overview of the course
- Python basics
- Working with structured data within Python using Pandas

Day 2: Relational Databases

- Introduction to relational databases
- Working with structured data in relational databases using the SQL query language (and MariaDB)
- Working with relational databases from within Python
- Efficient relational database design and indexing

Day 3: Apache Spark

- Introduction to the large-scale computation engine Apache Spark and the distributed file system HDFS
- Accessing and storing data in an Apache Spark cluster (both directly and using Python)
- Parallel and efficient processing, transformation, and analysis of large-scale datasets using PySpark

Day 4: Complex Data

- Discussion of non-relational data representation formats, most notably, semi-structured data and graph data
- Introduction to the JavaScript Object Notation (JSON)
- Working with semi-structured and graph data in Python
- Combining structured data and semi-structured data

Day 5: NoSQL databases

- Introduction and overview of NoSQL databases
- When to use relational databases vs. specific types of NoSQL databases
- Working with semi-structured data using the NoSQL database system MongoDB
- Summary and wrap-up

Literature

- Python and Pandas
  - McKinney, Wes. 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.
  - https://pandas.pydata.org/getting_started.html
- SQL
  - Beaulieu, Alan. 2020. *Learning SQL: Generate, manipulate, and retrieve data*. O'Reilly Media.
  - DeBarros, Anthony. 2022. *Practical SQL: A Beginner's Guide to Storytelling with Data*. No Starch Press.
- Spark
  - Karau, Holden, et al. 2015. *Learning spark: lightning-fast big data analysis*. O'Reilly Media, Inc.
  - Ryza, Sandy, et al. 2017. *Advanced analytics with spark: patterns for learning from data at scale*. O'Reilly Media, Inc.
- NoSQL
  - Strauch, Christof. 2011. “NoSQL databases:” Lecture Notes, *Walter Kriha, Ultra-Large Scale Sites,* Stuttgart Media University. https://www.christof-strauch.de/nosqldbs.pdf.
  - https://www.mongodb.com/docs/manual/