**Syllabus**

**Course title ECBS 5146 – Data Engineering 2: Different Shapes of Data**

**Instructor** Miklós Koren, László Salló

Email korenm@ceu.edu,

Office N13 213 (Budapest), by appointment

**Credits**  2 US credits (4 ECTS credits)

**Module** Data Engineering

**Term**  Fall 2020-2021 2020-2021

**Course level** Master's

**Prerequisites** Data Engineering 1

**Course drop** Course can be dropped free of charge 24 hours after the first session. After this date drop is possible until the course is halfway over (late drop fee applies). No changes are allowed past that date.

**1. Course Description**

**Content.** In Data Engineering 2 we deep dive in advanced topics relevant for Data Analysts, we go beyond the tabular format and discover different shapes of data and the tools supporting these shapes. By the end of the course we aim to build all these new technologies in a comprehensive analytical pipeline.

**Relevance.** In Data Engineering 1 we introduced SQL and the related technologies such as Relational Database Management Systems and classical Data Warehouse architectures. .

**2. Learning Outcomes**

**Key outcomes.** By the end of the course, students will be able to

- Understand the tradeoffs of data architectures

- Work with different data formats and files

- Model advanced data structures

- Have broad understanding on choosing the right technical solution within the 1000s of technologies available these days

- Build a data pipeline for analytics

**Other outcomes.** The course will also help develop skills in the following areas.

|  |  |
| --- | --- |
| **Learning Area** | **Learning Outcome** |
| Critical thinking | Evaluate and compare different data analytics tools and technologies.  Separate important from unimportant features of analytics problems. |
| Quantitative reasoning | Create logical models for relational data.  Understand and apply database normalization.  Classify search algorithms according to algorithmic complexity.  Express the algorithmic complexity of simple algorithms written in pseudocode.  Compare different data structures. |
| Technology skills | Create and query SQL databases.  Load and save flat files in different serialization formats with different character encodings. |
| Interpersonal communication skills | Convey technical concepts verbally.  Collaborate with others on technical tasks.  Communicate design decisions with block diagrams. |
| Management knowledge and skills | Build analytical systems§ with many components.  Organize work components effectively.  Meet deadlines. |
| Cultural sensitivity and diversity | Work together with students of different backgrounds. |
| Ethics and social responsibility | Understand ethical and legal constraints of acquiring and processing data.  Apply good practices of data protection. |

**3. Reading List**

**Required**

None of the books below are required in their entirety. The relevant chapters will be highlighted for each section.

**Recommended**

Martin Kleppmann (2017). Designing Data-Intensive Applications - The Big Ideas Behind Reliable, Scalable, and Maintainable Systems Sebastopol, CA : Oreilly Associates Inc (https://dataintensive.net/buy.html)

Sadalage, Fowler (2013). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence Addison-Wesley Professional

Silberschatz, Kotz and Sudarshan (2011). Database System Concepts, 6th Edition. Chapters 7.1-7.3.

**4. Teaching Method and Learning Activities**

Learning objectives will be achieved through

- The course will involve a mix of theoretical presentations and practice. There will no separate practical seminar, almost every course scheduled has practice incorporated in the class session.

**5. Assessment**

Grading will be based on the total score out of 100, in line with CEU’s standard grading guidelines.

- Questionnaire as Exam (40 percent)

- Take-home Assignment (20 percent)

- Final Project (40 percent)

**6. Technical requirements**

- Personal laptop computer with administrative privileges to install open source software.

- Operating system: Windows 10+ or Mac OS X 10.8+, or Linux 2.6.18+

- Latest Chrome browser

- KNIME Analytics Platform (https://www.knime.com/downloads/download-knime)

- Internet access.

**7. Topic Outline and Schedule**

|  |  |  |
| --- | --- | --- |
| **Session** | **Topics** | **Readings** |
| 1 | Data architecture tradeoffs | Kleppmann (2017), Chapter 1. Fowler, Martin (2015). "DataLake [blog post]" https://martinfowler.com/bliki/DataLake.html |
| 2 | The new tools, polyglot persistence, NoSQL practice | None |
| 3 | Data structures, NoSQL practice | Skerritt, Brandon (2019). "All You Need to Know About Big O Notation [blog post]" https://skerritt.blog/big-o/ |
| 4 | NoSQL practice | None |
| 5 | Advanced data modeling, data pipelines, Knime practice | Silberschatz, Kotz and Sudarshan (2011). Chapters 7.1-7.3. |
| 6 | Knime practice | None |
| 7 | Data serialization for ease of sharing, bash practice | Kleppmann (2017), Chapter 4. |
| 8 | Final quiz, term project preparation | None |

**8. Short Bio of the Instructor**

Miklós Koren is professor of economics at CEU, senior research fellow at the Institute of Economics, and research fellow of the Centre for Economic Policy Research. He is the founder of the Business Analytics MSc program and the CEU MicroData research group. His research focuses on how talent and technology jointly determine business success. Professor Koren has more than two decades of experience with data and coding. He is a certified Carpentries Instructor.

Laszlo Sallo is an IT Manager with Software Engineering background. Over the past 20+ years, he participated in more than 150 international projects. Currently he is serving as Delivery Manager for EPAM Systems, leading a data related project portfolio, combining his skills in Agile Project Management and Data Engineering. Laszlo holds an M.S. in Business Analytics from Central European University (2016). He is instructor of CEU since 2016.