

# Syllabus

## Different shapes of Data

- **Instructor:** László Salló (salacika@gmail.com office hours: on demand)
- **Credits:** 2 (4 ECTS)
- **Term:** Fall 2017-2018
- **Course level:** [MA/MSc]
- **Prerequisites:** Basic R programming covered in Data Analysis

## Course description

For several decades the world of data was characterized by one strong constant — we store data in relational databases. Since 2010 this landscape has changed radically. One can argue about increased volumes, or the users' appetite for higher variety of data, but the best way to understand the change is to see what business problems are solved with the help of this new landscape.

To understand the magnitude of this change, here is one number: the market once dominated by a few dozen of data systems, now has more than 300 ranked members, 60% of these represents new NoSQL solutions.

There is a lot value in such a variety, but challenges as well: the world has shifted from general solutions to specialized solutions, from closed commercial systems to open source systems, therefore choosing the right solution to the problem became a key element of the business analytics.

This course meant to be a practical course presenting several business scenarios and the appropriate data solutions to support these scenarios. From the simplest relational cubes we will get to the "different shapes of data", the immense variety of new technologies and data concepts all meant to support a new world of information. We will not focus much on individual tools, but instead on complete data processes, from the acquisition of the data until the birth of the business value.

## Who should sign up for this course?

- Business Analytics students without technical background should join to acquire basic skills in data access and data handling (SQL, Bash, NoSQL, KNIME)
- Everybody interested in what a real life data scientist is doing
- Managers willing to become Data Officers

## Course objectives

- To get the students familiar with different aspects of the data world, present an overview of the major data technologies and concepts, study them in the context of business problems
- To stress the importance of business value in the world of data
- To connect the dots: during the Business Analytics program the students have the opportunity to learn several techniques including Big Data, Machine Learning or Exploratory Data Analysis. The aim of this course is to link these techniques in complex data processes (It is however not a prerequisite to know these techniques in advance to complete the course)
- There is also a human aspect of this course: we are aiming to present several job roles from data scientist, through data engineer to chief data officer to give the students job market guidance, useful after graduation.

## Learning outcomes

At the end of the course, students will be able to choose the right data solution for a given problem, knowing the available options and their characteristics and limitations. This should enable Business Analysts to map the needs of the business stakeholders to the right data strategy and ultimately enable the implementation of the right data solution.

## Reading list

Martin Kleppmann (2017). *Designing Data-Intensive Applications - The Big Ideas Behind Reliable, Scalable, and Maintainable Systems* Sebastopol, CA : Oreilly & Associates Inc

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

Web based tutorials will be also available.

## Assessment

Quizzes with Socrative (10%)  
Homework assignments (20%)  
Use case presentation (20%)  
Exam (50%)

## Course schedule

12 Sessions of 100 minutes are planned with 6 lectures and 6 practice labs. Each lecture will have a theoretical introduction followed by a use case presented mostly by guests. After each lecture, we will do a lab session to practice the concepts introduced in the theoretical part. The overall time effort for this course is estimated to 50 hours including the 20 hours of lectures and labs.

## 1. THE VALUE OF DATA & BUSINESS INTELLIGENCE

Business organizations and data. ERP. Business Intelligence. Business Analytics. Data OLTP. Data Warehouses. ELT. OLAP.

## 2. THE EMERGE OF THE NEW QUESTIONS AND CHALLENGES

Old and new business questions. Challenges with RDBMS. Scaling. CAP Theorem. ACID. High Availability. Fault tolerant systems. Replication. The new tools. Cloud. DBaaS.

## 3. INTRODUCING THE NEW PLATFORMS OF DATA ANALYTICS

New trends. Analytical vs operational data problems. NoSQL in brief. Db-engines.com. Key Value Stores (Redis, Dynamo). Demo. Use case review. Data pipelines. Lambda architecture.

## 4. ADVANCED SEARCH

Wide column stores (Cassandra). Demo. Use case review. Search engines. (Solr). Full text search. Open NLP. Geo search. Faceted search. Demo. Use case review.

## 5. SPECIAL TOOLS FOR SPECIAL PROBLEMS

Time series analytics. Time Series DBMS (InfluxDB). Demo. Use case review. Graph theory/Network science. Graph DBMS (Neo4J). Demo. Use case review. Geographical Information Systems.

## 6. THE BIG PLAYER & POLYGLOT PERSISTENCY

Document Stores (MongoDB). Relation to other stores. Demo. Use case review. Polyglot persistency. Best Practices.