INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

DPT INFORMATIQUE
COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
INSA Campus LyonTech - 7 Avenue Jean Capelle
Batiment Blaise Pascal - 69621 VILLEURBANNE
Phone 0472438892
Web site : http://if.insa-lyon.fr

Conferences and Seminars

Foundation of data engineering

IDENTIFICATION

CODE : IF-5-S1-EC-OT7
ECTS : 6.0

HOURS

Lectures :	0.0 h
Seminars :	48.0 h
Laboratory :	0.0 h
Project :	0.0 h
Teacher-student contact :	48.0 h
Personal work :	48.0 h
Total :	96.0 h

ASSESSMENT METHOD

Oral&written examination

TEACHING AIDS

https://dataengineering.training

TEACHING LANGUAGE

English

CONTACT

M. TOMMASINI Riccardo
riccardo.tommasini@insa-lyon.fr

AIMS

The course aims at giving an overview of Data Engineering foundational concepts, i.e., data modeling, collection, transformation, and wrangling. The course discusses the role of the data engineers in the data science/AI life-cycle is. It shows (i) how to gather data from various sources (databases, repositories, web services, and even websites); (ii) how to design data pipelines to perform (Streaming) and how to scale them up; (iii) how to clean, enhance, and augment data to feed data science pipelines

CONTENT

After completing this course, the student can:
* Collect data from heterogeneous data sources, such as from websites, social media, web services, SQL and NoSQL databases, structured, JSON and plain text file formats;
* Design models for the data using various techniques, e.g., as ER diagrams fro relational data, property graph for graph data, event sourcing for temporal data.
* Design (streaming) data pipelines to transform data into desired forms and perform data integration;
* Describe the advantages and limitations of various data storage and processing technologies based on the given requirements (e.g., batch processing, real-time processing);
* Assess and measure the data by quality and volume;
* Clean the data by detecting and dealing with issues such as duplicates and missing values;

It is tailored for Msc students and PhDs who would like to strengthen their fundamental understanding of Data Engineering, i.e., Data Modelling, Collection, and Wrangling.

BIBLIOGRAPHY

Database System Concepts 7th Edition Avi Silberschatz Henry F. Korth S. Sudarshan McGraw-Hill ISBN 9780078022159
The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling Third Edition Ralph Kimball Margy Ross
Designing Data-Intensive Applications - Martin Kleppmann, https://dataintensive.net/
Designing Event-Driven Systems, https://www.oreilly.com/library/view/designing-event-driven-systems/9781492038252/
Graph Databases, https://neo4j.com/graph-databases-book/

PRE-REQUISITE

Familiarity with the following concepts is strongly recommended to succeed in the course:
Algorithm and Data Structures
Graphs, Trees, Tables, Lists
Programming Languages
Java and Python
(Relational) Databases and Query Languages
SQL