Fakultät für Informatik
Facoltà di Scienze e Tecnologie informatiche
Faculty of Computer Science

COURSE DESCRIPTION – ACADEMIC YEAR 2022/2023

Course title	**Real-Time Big Data Processing**
Course code	73033
Scientific sector	INF/01
Degree	Master in Computational Data Science (LM-18)
Semester	2
Year	1 and 2
Credits	6
Modular	No

Total lecturing hours	40
Total lab hours	20
Attendance	Generally, attendance is not compulsory, but non-attending students have to contact the lecturer at the start of the course to agree on the modalities of the independent study.
Prerequisites	• Programming (Java or Python – knowledge of one is enough) • Databases (SQL – basic knowledge) • Web development (HTML, JavaScript – basic knowledge)
Course page	https://ole.unibz.it/

Specific educational objectives	The course belongs to the type "caratterizzanti – discipline informatiche" for the study path “no curriculum”. The course aims at teaching both scientific foundations and practical aspects of real-time big data processing technologies. The students will learn the basic concepts of such systems and how to use them to solve concrete problems, including real-time data analyses, applications of machine learning and complex event processing over streaming data. Moreover, students will be trained to evaluate the advantages and disadvantages of such technologies in different application contexts.

Lecturer	Francesco Corcoglioniti
Contact	Piazza Domenicani 3, francesco.corcoglioniti@unibz.it
Scientific sector of lecturer	ING-INF/05
Teaching language	English
Office hours	Arrange beforehand by email.
Lecturing assistant (if any)	--
Contact LA	--
Office hours LA	--
List of topics	• Reactive Streams • Messaging System (e.g., Apache Kafka) • Stateful Stream Processing (e.g., Apache Flink) • Micro-batch Stream Processing (e.g., Apache Spark) • Applications over Stream Processing • Semantic Technologies for streaming data

Teaching format	Frontal lectures and hands-on labs (not evaluated). The lectures present the basic concepts, their realization in the open source systems studied in the course (e.g., Kafka, Flink, Spark), and their practical use with concrete examples. The labs permit students to practice the technologies of the course, by solving small tasks as part of guided tutorials often involving complete applications (from data ingestion to web front end) showcasing the use of these technologies.

Learning outcomes	Knowledge and understanding: • D1.1 - Knowledge of the key concepts and technologies of data science disciplines • D1.3 - Knowledge of principles, methods and techniques for processing data in order to make them usable for practical purposes, and understanding of the challenges in this field • D1.4 - Sound basic knowledge of storing, querying and managing large amounts of data and the associated languages, tools and systems • D1.5 - Knowledge of principles and models for the representation, management and processing of complex and heterogeneous data Applying knowledge and understanding: • D2.1 - Practical application and evaluation of tools and techniques in the field of data science • D2.2 - Ability to address and solve a problem using scientific methods Making judgments • D3.2 - Ability to autonomously select the documentation (in the form of books, web, magazines, etc.) needed to keep up to date in a given sector Communication skills • D4.1 - Ability to use English at an advanced level with particular reference to disciplinary terminology • D4.3 - Ability to structure and draft scientific and technical documentation Learning skills • D5.1 - Ability to autonomously extend the knowledge acquired during the course of study • D5.2 - Ability to autonomously keep oneself up to date with the developments of the most important areas of data science • D5.3 - Ability to deal with problems in a systematic and creative way and to appropriate problem solving techniques.

Assessment	The assessment of the course is based on a project which is done during the semester and requires students to solve a concrete problem by using methods and technologies taught in the course (100% of the mark). The project verifies whether the student is able to apply advanced data management techniques to solve concrete problems. The project is assessed through the submission of the solution source

	code and an accompanying written report, followed by an oral exam where the student defends the project with a short presentation including slides and live demo. The exam modalities are the same for attending and non-attending students.
Assessment language	English
Assessment Typology	Monocratic
Evaluation criteria and criteria for awarding marks	The final exam grade is the project mark (100%). Criteria for the evaluation of the project: correctness of the solution, complexity of the project, technologies used in the solution, quality of the report and the presentation.

Required readings	There is no single textbook that covers the entire course. The course material is collected from various textbooks, research papers and online documentation, including the following books: • G. Shapira, T. Palino, R. Sivaram, K. Petty. "Kafka: The Definitive Guide". 2nd edition. O'Reilly Media, Inc. November 2021. • V. Kalavri, F. Hueske. "Stream Processing with Apache Flink". 1st edition. O'Reilly Media, Inc. April 2019. • M. Armbrust. "Learning Spark". 2nd edition. O'Reilly Media, Inc. July 2020. • M. Kleppmann: "Designing Data-Intensive Applications". 1st edition. O'Reilly Media, Inc. March 2017. Subject Librarian: David Gebhardi, David.Gebhardi@unibz.it
Supplementary readings	Additional sources will be announced during the course.
Software used	Languages: Java or Python (students may also use Scala for their projects), SQL, HTML, JavaScript (or TypeScript) Software: Apache Kafka, Apache Flink, Apache Spark, ReactiveX, Docker and Docker Compose