COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

COURSE SYLLABUS
2024-2025

GENERAL INFORMATION

Course information	
Name	Stream Processing
Code	DTC-MIC-525
Degree	Máster en Big Data. Tecnología y Analítica Avanzada
Semester	2nd (Spring)
ECTS credits	3.0
Type	Compulsory
Department	Telematics and Computer Science
Coordinator	Carlos Morrás Ruiz-Falcó

Instructor	
Name	Carlos Morrás Ruiz-Falcó
Department	Telematics and Computing
Office	
e-mail	cmorras@icai.comillas.edu
Phone	
Office hours	Arrange an appointment through email.

DETAILED INFORMATION

Contextualization of the course

Contribution to the professional profile of the degree

The purpose of this course is to provide students with a fundamental understanding and an extensive practical experience on how to process and analyze data in a streaming fashion with the most important tools in the big data ecosystem.
By the end of the course, students will:

- Understand the basic principles of the stream processing.
- Have practical experience with the most important tools in the stream processing and basic knowledge on Scala language
- Have well-formed criteria to choose the most appropriate tool for a given streaming application. (Spark Streaming, Spark Structured Streaming, Kafka Streams, Flink).

Prerequisites

Students willing to take this course should be familiar with Scala and/or Java programming languages. It will also be desirable to know the basic concepts of the streaming framework Kafka.

CONTENTS

Contents

Theory

Unit 1. Introduction

1.1 ¿What is Stream Processing?
1.2 Scala Basics
1.3 Course setup (IntelliJ, Git, Zeppelin, etc.)

Unit 2. Kafka Revisited

2.1 Introduction to Kafka
2.2 Kafka architecture
2.3 Consumers and producers
2.4 Kafka tools

Unit 3. Spark SQL

3.1 Dataframes and Datasets
3.2 Actions and transformations (filter, join, aggregation, etc)
3.3 ¿How to read and write data?

Unit 4. Common stream processing frameworks

4.1 Overview of most common stream processing tools (Spark Streaming, Flink, Kafka Streams, etc)

Unit 5. Spark Structured Streaming

5.1 Introduction
5.2 Input table and result table. Kafka integration.
5.3 Stateless (transformations, filtering) and stateful operations (aggregations, joins)
5.4 Windowing. Late data handling. Watermarking.

Laboratory[1]

Practice 0. Introduction to Scala

In the first lab session, students will become familiar with the Scala language that they will use throughout the course.

Practice 1. Kafka Tool

In this practice, students will be guided to achieve a well understanding of Kafka using command line.

Practice 2. Air Traffic

In this practice, students will work with streaming air traffic data coming from real world and will be responsible of creating the data structures and consume, aggregate, join and transform that streaming data.

Practice 3. TV Audiences

In this practice, students will have to implement a set of functions using TV Audiences as streaming data.

Competences and learning outcomes

Competences[2]

General competences

CG1. Have acquired advanced knowledge and demonstrated, in a research and technological or highly specialized context, a detailed and well-founded understanding of the theoretical and practical aspects, as well as of the work methodology in one or more fields of study.

[1] Practice topics are tentative and may be replaced by others of similar nature.
[2] Competences in English are a free translation of the official Spanish version.

	Haber adquirido conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.
CG2.	Know how to apply and integrate their knowledge, understanding, scientific rationale, and problem-solving skills to new and imprecisely defined environments, including highly specialized multidisciplinary research and professional contexts. *Saber aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.*
CG3.	Know how to evaluate and select the appropriate scientific theory and the precise methodology of their fields of study in order to formulate judgements based on incomplete or limited information, including, when necessary and pertinent, a discussion on the social or ethical responsibility linked to the solution proposed in each case. *Saber evaluar y seleccionar la teoría científica adecuada y la metodología precisa de sus campos de estudio para formular juicios a partir de información incompleta o limitada incluyendo, cuando sea preciso y pertinente, una reflexión sobre la responsabilidad social o ética ligada a la solución que se proponga en cada caso.*
CG4.	Be able to predict and control the evolution of complex situations through the development of new and innovative work methodologies adapted to the scientific/research, technological or specific professional field, in general multidisciplinary, in which they develop their activity. *Ser capaces de predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad.*
CG5.	Be able to transmit in a clear and unambiguous manner, to specialist and non-specialist audiences, results from scientific and technological research or state-of-the-art innovation, as well as the most relevant foundations that support them. *Saber transmitir de un modo claro y sin ambigüedades, a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.*
CG6.	Have developed sufficient autonomy to participate in research projects and scientific or technological collaborations within their thematic area, in interdisciplinary contexts and, where appropriate, with a high knowledge transfer component. *Haber desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro de su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.*
CG7.	Being able to take responsibility for their own professional development and their specialization in one or more fields of study. *Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.*
Specific competences	
CE4.	Know the techniques used to process streaming data, as well as the different platforms, tools, and languages that make it possible. *Conocer las técnicas para procesar flujos de datos en streaming, así como las diferentes plataformas, herramientas y lenguajes que lo hacen posible.*

Learning outcomes
By the end of the course students should: RA1. Have familiarity with Scala language and functional programming principles RA2 List the characteristics and advantages of stream processing frameworks. RA3 Develop code to process streaming data. RA4 Be familiar with Spark Structured Streaming and Kafka RA5 Understand and propose uses of stream processing in general and in the industry in particular. RA6 Have familiarity with some of the most common development tools in the industry RA7 Be capable of addressing simple streaming data projects.

TEACHING METHODOLOGY

General methodological aspects

To ensure useful and practical learning, theoretical classes will be combined with master classes that reflect the reality of the market. Real case studies with also be studied from business and technical perspectives, some of which will be used in practical sessions.

In-class activities	Competences
▪ **Lectures:** The lecturer will introduce the fundamental concepts of each unit, along with some practical recommendations, and will go through worked examples to support the explanation. Active participation will be encouraged by raising open questions to foster discussion and by proposing quizzes and short application exercises to be solved in class.	CG1, CG3, CG4, CG7, CE5
▪ **Practical sessions:** Under the instructor's supervision, students will apply the concepts learned in the lectures to real cases, in order to face and solve implementation problems that typically arise.	CG1, CG2, CG3, CG5, CG6, CG7, CE5
▪ **Tutoring** for groups or individual students will be organized upon request.	–
Out-of-class activities	**Competences**
▪ Personal study of the course material and resolution of the proposed exercises.	CG1, CG3, CG4, CG7, CE5
▪ Practical session preparation to make the most of in-class time.	CG1
▪ Practical results analysis and report writing.	CG2, CG5, CE5

ASSESSMENT AND GRADING CRITERIA

Assessment activities	Grading criteria	Weight
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Critical analysis of numerical exercises' results.	40%
Practical assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to understand results in real environment. ▪ Written communication skills.	60%

GRADING AND COURSE RULES

Grading

Regular assessment

- **Final exam** will account for 40%
- **Lab** (practical assignments) will account for the remaining 60%

In order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 5 out of 10 points, and the laboratory mark must be at least 5 out of 10 points.

Retake

Lab marks will be preserved as long as they result in a passing grade. Otherwise a project will have to be developed and handed in. In addition, all students will take a final exam. The resulting grade will be computed as follows:

- Final exam will account for 40%.
- **Lab** will account for the remaining 60%
  - If the student passed the lab during regular assessment
    - Practical assignments: 60%
  - Otherwise
    - Final project: 60%

As in the regular assessment period, in order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 5 out of 10 points, and the mark of the laboratory must be at least 5 out of 10 points. Otherwise, the final grade will be the lower of the three marks.

Course rules

- Class attendance is mandatory according to Article 93 of the General Regulations (Reglamento General) of Comillas Pontifical University and Article 6 of the Academic Rules (Normas Académicas) of the ICAI School of Engineering. Not complying with this requirement may have the following consequences:
  - Students who fail to attend more than 15% of the lectures may be denied the right to take the final exam during the regular assessment period.
  - Regarding practice, absence to more than 15% of the sessions can result in losing the right to take the final exam of the regular assessment period and the retake. Missed sessions must be made up for credit.
- Students who commit an irregularity in any graded activity will receive a mark of zero in the activity and disciplinary procedure will follow (cf. Article 168 of the General Regulations (Reglamento General) of Comillas Pontifical University).

WORK PLAN AND SCHEDULE[3]

In and out-of-class activities	Date/Periodicity	Deadline
Final exam	After the lecture period	–
Practice sessions	From week 2	–
Review and self-study of the concepts covered in the lectures	After each lesson	–
Practice preparation	Before every lab session	–
Practice report writing	–	One week after the end of each session

STUDENT WORK-TIME SUMMARY			
IN-CLASS HOURS			
Lectures		Practical sessions	
12		3/4	
OUT-OF-CLASS HOURS			
Self-study	Practice preparation	Report writing	Homework assignments
32	9.5	7.5	11
		ECTS credits:	3 (90 hours)

BIBLIOGRAPHY

Basic bibliography

https://spark.apache.org/docs/2.4.0/structured-streaming-programming-guide.html

Complementary bibliography

Programming Scala O'Reilly

Spark structured streaming III

[3] A detailed work plan of the subject can be found in the course summary sheet (see following page). Nevertheless, this schedule is tentative and may vary to accommodate the rhythm of the class.