ESIROI
ÉCOLE D'INGÉNIEURS
UNIVERSITÉ DE LA RÉUNION

DEPARTMENT : *IT*
SEMESTER : *S7*

## IDENTIFICATION

Course code : ESI-SPI-CI-IN4-S7-UE5-EC1
ECTS : 1

## DURATION

| | |
|---|---|
| Lectures : | 4 |
| Tutorial classes : | 16 |
| Labs : | 4 |
| Total : | 24 |

Project :
Personnal work :

## ASSESSMENT

Continuous assessments and evaluated practical work.

## TEACHING MATERIALS

Slides, guided and practical work sheets.

## LANGUAGE

English

## CONTACT

Understand the technical and algorithmic challenges related to processing massive data volumes ( Big Data ), and master the architectural solutions and distributed frameworks that address them. Through a combined theoretical and practical approach, students will acquire the skills needed to design, deploy, and operate large scale data processing pipelines, considering performance, scalability, and reliability constraints.

Modifié le : 14 juillet 2026

ESIROI - Campus de Terre Sainte
*40 Avenue de Soweto, 97455 Saint-Pierre, Cedex*
Tel.: +262 262 81 83 01 / 02 ; www.esiroi-reunion.fr

## UE : UE7-DATA

## COURSE : Massive Data Systems

# SCOPE OF THE COURSE

### OBJECTIVES

### COMPETENCE AND SKILLS

- Identify and analyze scalability issues in projects involving large volumes of data
- Design and implement a distributed architecture suitable for storing and processing massive data
- Apply parallel and distributed programming paradigms (MapReduce, Spark) to solve large-scale processing problems
- Define, implement, and use descriptive and predictive analysis by leveraging massive data
- Evaluate and compare the performance of parallel and distributed algorithms considering hardware and architectural constraints

# PROGRAM

- **Scalability challenges :**
  - The need for measurement in the context of Big Data : size, capacity, throughput, and latency.
  - Consequences of rapid data growth on computing and storage.
  - Complexity models suited to distributed systems ; the need to prioritize architectural simplicity.
  - Coordination approaches between computing entities ; horizontal and vertical scalability.
- **Hardware architectures for Big Data :**
  - Fast and efficient input/output mechanisms ; memory hierarchy and cache coherence.
  - Parallel computing architectures : multicore, GPU, grid computing, shared and distributed memory, vector processing.
  - Flynn's taxonomy (SISD, SIMD, MISD, MIMD) and implications for architectural choices.
  - Parallel storage hierarchy ; instruction optimization considerations.
- **Frameworks for distributed computing :**
  - Classification of parallel and distributed computing models.
  - Hadoop and the HDFS / YARN ecosystem ; MapReduce paradigm.
  - Apache Spark : RDDs, DataFrames, Spark SQL, Spark Streaming.
  - Distributed messaging systems : Apache Kafka for real-time data streams.
  - Process interaction : communication, coordination, fault tolerance.
- **Distributed data storage :**
  - Distributed storage approaches : HDFS, object storage (S3), NoSQL databases (HBase, Cassandra, MongoDB).

  - Ensuring data quality : consistency, cleaning, representativeness.
  - CAP theorem and trade-offs between consistency, availability, and partition tolerance.
  - Scalability techniques : consistent hashing, partitioning, replication, sharding.
  - Data backup, archiving, and recovery.

- **Parallel and distributed programming :**

  - Concurrency, parallelism, and distributed systems : differences and use cases.
  - Amdahl's Law and Gustafson's Law : theoretical limits of parallelism.
  - Parallel algorithms : load balancing, task and data decomposition.
  - Complexity of parallel and distributed algorithms ; state and side-effect management.

- **Data pipelines and stream processing :**

  - Lambda and Kappa architectures : batch and stream processing.
  - Data pipeline orchestration : Apache Airflow.
  - Use cases : ingestion, transformation, aggregation, and visualization of massive data.

## BIBLIOGRAPHY

- Sakr, S. & Zomaya, A. Y. (Eds.) - *Encyclopedia of Big Data Technologies* - Springer International Publishing, 2019
- Tom White - *Hadoop : The Definitive Guide* - O'Reilly, 4th ed., 2015
- Bill Chambers & Matei Zaharia - *Spark : The Definitive Guide* - O'Reilly, 2018
- Martin Kleppmann - *Designing Data-Intensive Applications* - O'Reilly, 2017
- Jimmy Lin & Chris Dyer - *Data-Intensive Text Processing with MapReduce* - Morgan & Claypool, 2010
- Apache Spark Documentation : `https://spark.apache.org/docs/latest/`
- Apache Kafka Documentation : `https://kafka.apache.org/documentation/`
- Apache Hadoop Documentation : `https://hadoop.apache.org/docs/stable/`
- Apache Airflow Documentation : `https://airflow.apache.org/docs/`

## REQUIREMENTS

- **Probability and Random Variables (S5) :** distributions, expectation, variance — statistical foundations for analyzing and sampling massive data.
- **Stochastic Processes (S6) :** modeling data flows and time series.
- **Data Mining (S6) :** clustering, classification, and association algorithms — directly applied to distributed volumes.
- **Machine Learning (S7, parallel) :** predictive models and ML pipelines to integrate into Big Data architectures.
- **Expected cross-disciplinary skills :**

  - Python programming (pandas, NumPy) and functional programming concepts
  - Proficiency in Linux terminal and basic commands (files, processes, network)

ESIROI
ÉCOLE D'INGÉNIEURS
UNIVERSITÉ DE LA RÉUNION

— Understanding of algorithmics and complexity (Big O notation)
— Knowledge of client-server architectures and relational databases
— Ability to read technical documentation in English

**ESIROI - Campus de Terre Sainte**
*40 Avenue de Soweto, 97455 Saint-Pierre, Cedex*
Tel.: +262 262 81 83 01 / 02 ; www.esiroi-reunion.fr