Course Title:	**Big data framework & technologies**
Course Code:	CSE523/1
Program:	Computer science Engineering
Department:	Computer Engineering
Course coordinator:	Dhekra CHERMITI
Institution:	Private Higher School of Engineers of Gafsa (ESIP)

A. Course Identification

1. Credit hours: **3 (1.5-1.5-0)**
2. Course type **a.** College ☐ Department ■ Others ☐ **b.** Fundamental ■ Transversal ☐ Optional ☐
3. Level/year at which this course is offered: 3.1/3
4. Pre-requisites for this course (if any): CSE131, CSE323

1. Mode of Instruction (mark all that apply)

No	Mode of Instruction	Contact Hours	Self-study	Total workload
1	**Traditional classroom**	…..	33	78
2	**Blended**	45		
3	**E-learning**	.....		
4	**Distance learning**	.....		
5	**Other ()**	.....		

2. Contact Hours (based on academic semester)

No	Activity	Contact Hours
1	**Lecture**	30
2	**Laboratory/Studio**	15
3	**Tutorial**	-
4	**Others** (specify)	-
	Total	45

B. Course Objectives and Learning Outcomes

1. Course Description

This course introduces students to Apache Spark and its core components. It covers the different versions of Spark (Scala, Python, Java), programming with Resilient Distributed Datasets (RDDs), working with Spark SQL, and handling graphs with GraphX. Students will also explore cluster architectures (Standalone, Apache Mesos, Hadoop YARN) and learn how to develop applications using SparkML and MLlib for machine learning.

2. Course Main Objective

By the end of this course, students will be able to:

✓ Understand the fundamental concepts of Apache Spark.
✓ Explore Spark's different modules and compare them with the Apache Hadoop environment.
✓ Develop distributed applications using Apache Spark.
✓ Handle large datasets using SQL queries in Spark.
✓ Set up and configure a Spark cluster for real-world applications.
✓ Process real-time data streams using Spark Streaming.
✓ Manipulate and analyze graphs using the GraphX API.

1. Course Learning Outcomes

CLOs		Aligned PLOs
1	**Knowledge and Understanding**	
1.1	✓ Acquire the fundamental concepts of spark.	PLO.K1
1.2	✓ Explore the different modules of Spark and its relationship with the Apache Hadoop environment.	PLO.K2
2	**Skills**	
2.2	✓ Handle datasets using SQL queries in Spark. ✓ Use the GraphX API to manipulate and analyze graph data.	PLO.S2
2.3	✓ Evaluate and analyze real-time data streams with Spark Streaming.	PLO.S3
2.6	✓ Develop distributed applications using Apache Spark.	PLO. S6

C. Course Content

No	List of Topics	Contact Hours
1	Chapter 1: Introduction to Apache Spark & Framework Overview 1. What is Apache Spark**?** 2. History of the Framework and its evolution**.** 3. Versions of Spark: Scala, Python, and Java**.** 4. Overview of Spark Modules: Spark Core, Spark SQL, Spark Streaming, MLlib, and GraphX**.**	4

	5. Comparison with Apache Hadoop: Differences in architecture, speed, and efficiency.	
2	**Chapter 2**: Working with Resilient Distributed Datasets (RDDs) 1. Understanding RDDs: What they are and how they work. 2. Creating, transforming, and reusing RDDs. 3. Using Accumulators and Broadcast Variables for distributed computing.	3
3	**Chapter 3:** Processing Structured Data with Spark SQL 1. Introduction to Spark SQL, DataFrames, and Datasets. 2. Converting between RDDs and DataFrames. 3. Connecting Spark to different data sources (JSON, Parquet, CSV, JDBC, NoSQL). 4. Optimizing performance when querying large datasets.	6
4	**Chapter 4:** Deploying Spark Applications on a Cluster 1. Setting up different cluster architectures: Standalone, Apache Mesos, Hadoop YARN. 2. Using Spark-submit to deploy applications. 3. Configuring cluster resources: memory, CPU, and workload balancing.	6
5	**Chapter 5**: Real-Time Data Processing with Spark Streaming 1. How Spark Streaming processes real-time data. 2. Understanding Discretized Streams (DStreams). 3. Working with Kafka, Flume, HDFS, and TCP sockets for streaming data. 4. Comparing Spark Streaming and Apache Storm for real-time analytics.	5
6	**Chapter 6:** Analyzing Graph Data with GraphX 1. What is GraphX, and where is it used? 2. Building graphs in Spark using Vertex and Edge RDDs. 3. Applying graph algorithms like PageRank and Connected Components	3
7	**Chapter 7:** Introduction to Machine Learning with Spark MLlib 1. Overview of machine learning in Spark. 2. Understanding classification, regression, and clustering. 3. Using MLlib for scalable machine learning with big data	3
Total		30

C.1Practical work Content

No	List of Topics	Contact Hours
1	Lab 1: Installing Apache Spark & Setting Up Environment	3
2	Lab 2: Programming with RDDs: Creating, Transforming & Using Broadcast Variables	3
3	Lab 3: Manipulating Structured Data with Spark SQL	3
4	Lab 4: Handling Graphs with GraphX: Building Graphs & Applying Graph Algorithms	3
5	Lab 5: Introduction to SparkML & MLlib	3
Total		15

D. Teaching and Assessment

1. Alignment of Course Learning Outcomes with Teaching Strategies and Assessment Methods

Code	Course Learning Outcomes	Teaching Strategies	Assessment Methods
1.0	**Knowledge and Understanding**		
PLO.K1	✓ Acquire the fundamental concepts of spark.	Lecturing	Assignments, Quizzes , Exams,
PLO.K2	✓ Explore the different modules of Spark and its relationship with the Apache Hadoop environment.		
2.0	**Skills**		
PLO.S2	✓ Handle datasets using SQL queries in Spark. ✓ Use the GraphX API to manipulate and analyze graph data.	Lecturing	Assignments, Quizzes , Exams,
PLO.S3	✓ Evaluate and analyze real-time data streams with Spark Streaming.		
PLO. S6	✓ Develop distributed applications using Apache Spark.		

2. Assessment Tasks for Students

#	Assessment task*	Week Due	Percentage of Total Assessment Score
1	Practical Work (written or oral)	Weekly	25%
2	Quizzes, Homework assignments	-	-
3	First mid Term	8	25%

#	Assessment task*	Week Due	Percentage of Total Assessment Score
4	Final Exam	16	50%

E. Student Academic Counselling and Support

Arrangements for availability of faculty and teaching staff for individual student consultations and academic advice:
- Office hours - Blackboard interface

F. Learning Resources and Facilities

1. Learning Resources

Required Textbooks	1. Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee – Learning Spark: Lightning-Fast Data Analytics, 2nd Edition, O'Reilly Media, 2020. 2. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia – High-Performance Spark: Best Practices for Scaling and Optimizing Apache Spark, O'Reilly Media, 2017 3. Ben Lorica, Mike Loukides – What You Need to Know About Apache Spark, 2015
Essential References Materials	Apache Spark
Electronic Materials	1. Coursera & Udemy – Apache Spark & Big Data Processing Courses 2. MIT OpenCourseWare (OCW) – Big Data Analytics with Apache Spark 3. GitHub Repositories
Other Learning Materials	NA

2. Facilities Required

Item	Resources
Accommodation (Classrooms, laboratories, demonstration rooms/labs, etc.)	**classroom board software …**
Technology Resources (AV, data show, Smart Board, software, etc.)	**data show;**

G. Course Quality Evaluation

Evaluation Areas/Issues	Evaluators	Evaluation Methods
Effectiveness of teaching and assessment.	Students, Faculty, Program Leaders, Peer Reviewer	Direct/Indirect
Extent of achievement of course learning outcomes.	Faculty, Program Leaders, Peer Reviewer	Direct, Indirect
Quality of Learning resources	Faculty, Program Leaders, Peer Reviewer	Direct, Indirect
Teaching and learning quality and effectiveness.	Students, Faculty Program Leaders, Peer Reviewer	Direct, Indirect

H. Specification Approval Data

Council / Committee	Computer Engineering Council
Date	11/09/2023