

ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

593000509 - Big Data Applications For Iot

DEGREE PROGRAMME

59AH - Master Universitario En Internet Of Things (IOT)

ACADEMIC YEAR & SEMESTER

2024/25 - Semester 2

Index

Learning guide

1. Description.....	1
2. Faculty.....	1
3. Prior knowledge recommended to take the subject.....	2
4. Skills and learning outcomes	2
5. Brief description of the subject and syllabus.....	3
6. Schedule.....	5
7. Activities and assessment criteria.....	7
8. Teaching resources.....	11
9. Other information.....	12

1. Description

1.1. Subject details

Name of the subject	593000509 - Big Data Applications For Iot
No of credits	4.5 ECTS
Type	Compulsory
Academic year of the programme	First year
Semester of tuition	Semester 2
Tuition period	February-June
Tuition languages	English
Degree programme	59AH - Master Universitario en Internet Of Things (Iot)
Centre	59 - Escuela Tecnica Superior De Ingenieria Y Sistemas De Telecomunicacion
Academic year	2024-25

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Ivan Pau De La Cruz		ivan.pau@upm.es	Sin horario.
Bonifacio Alberto Mozo Velasco (Subject coordinator)	ETSISI- 4313	a.mozo@upm.es	Sin horario.
Javier Martin Rueda		javier.martin@upm.es	Sin horario.

* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

3. Prior knowledge recommended to take the subject

3.1. Recommended (passed) subjects

- Distributed Systems For Iot

3.2. Other recommended learning outcomes

- There are no entry requirements although it is highly recommended to have solid foundations on programming languages (such as C, Java or Python) and data structures

4. Skills and learning outcomes *

4.1. Skills to be learned

CB07 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB09 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

CE.11 - Diseñar aplicaciones capaces de procesar grandes volúmenes de información y flujos de datos usados en aplicaciones IoT incluyendo el almacenamiento, la gestión y el análisis de la información disponible

CG02 - Los alumnos serán capaces de aplicar métodos y tecnologías avanzadas que les permitan abordar necesidades y problemas en aplicaciones IoT

CG04 - Los alumnos tendrán la capacidad de aplicar criterios de eficiencia, escalabilidad, fiabilidad y seguridad en distintos ámbitos de aplicaciones inteligentes y sistemas ciberfísicos, tales como Smart Living, Smart Cities o eHealth

CT.01 - Capacidad de uso de la lengua inglesa para el trabajo en contextos internacionales

CT.03 - Creatividad, iniciativa y capacidad emprendedora

CT.04 - Capacidad para la elaboración, planificación, coordinación y gestión técnica y económica de proyectos siguiendo criterios éticos, de calidad y medioambientales

4.2. Learning outcomes

RA36 - To extract information and knowledge for an organization from a massive data set

RA34 - To apply the most appropriate management and analysis techniques for a massive data set

RA44 - To select the most appropriate architecture and storage for a massive set of heterogeneous data within the scope of IoT

* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

5. Brief description of the subject and syllabus

5.1. Brief description of the subject

This course provides a comprehensive introduction to the Big Data processing problem. IoT applications can quickly generate massive amounts of different types of data which are constantly flowing, and which can be stored and analyzed to gather valuable insights. Adequate architectures, distributed tools and strategies are required to collect, inject, distribute, store, analyze, and visualize these data at scale.

Some of the distributed tools which are used to collect, inject, and distribute IoT data are studied in other courses of the master's degree, and this course leverages that base to focus on the processing and analysis stages. Different paradigms are described considering both batch and online strategies.

Some of the most popular distributed computing and data processing platforms in the current Big Data scenario are presented and studied: Apache Hadoop and its ecosystem, Apache Spark SQL/Streaming/ML, and the Elastic ELK stack.

Additionally, the student will learn the foundations of Machine Learning techniques and how to apply them in Big Data use cases.

The course includes practical group activities to build representative cases of Big Data systems mainly using Unix-type systems, Python programming, and cloud environments.

5.2. Syllabus

1. Introduction to Big Data
 - 1.1. Concepts, challenges and models
 - 1.2. Architecture of IoT big data applications
2. The Elastic Stack
 - 2.1. Data collection and transformation
 - 2.2. Data storage and retrieval
 - 2.3. Data visualization
 - 2.4. Application of the ELK stack to IoT applications
3. The Apache Hadoop ecosystem
 - 3.1. The HDFS filesystem
 - 3.2. The Map-Reduce paradigm
 - 3.3. Cluster resource management
 - 3.4. Hadoop core
 - 3.5. Overview of the Hadoop ecosystem
4. Distributed Computing with Apache Spark.
 - 4.1. Resilient Distributed Datasets (RDDs)
 - 4.2. The Spark API
 - 4.3. Structured processing with Spark SQL
 - 4.4. Structured streaming processing with Spark
5. Introduction to Machine Learning

6. Schedule

6.1. Subject schedule*

Week	Type 1 activities	Type 2 activities	Distant / On-line	Assessment activities
1	Unit 1. Introduction to Big-Data. Introduction to Big Data Processing architectures Duration: 03:30 Lecture			
2	Unit 2. ELK Stack Duration: 03:15 Lecture			Test #1 Big Data concepts and architectures Problem-solving test Progressive assessment and Global Examination Presential Duration: 00:15
3	Unit 2. ELK Stack Duration: 03:30 Lecture			
4	Unit 3. Hadoop Map-Reduce ecosystem Duration: 03:05 Lecture			Test #2 ELK stack Problem-solving test Progressive assessment and Global Examination Presential Duration: 00:25
5	Unit 4. Apache Spark. Spark SQL. Spark Streaming Duration: 03:10 Lecture			Test #3 Hadoop and Map-Reduce Problem-solving test Progressive assessment and Global Examination Presential Duration: 00:20
6	Unit 4. Apache Spark. Spark SQL. Spark Streaming Duration: 03:10 Lecture			Test #4 Spark SQL/Streaming Problem-solving test Progressive assessment and Global Examination Presential Duration: 00:20 Spark SQL/Streaming practical assignment Group work Progressive assessment and Global Examination Not Presential Duration: 00:00
7	Unit 5. Apache Spark. RDDs Duration: 03:30 Lecture			

8	Unit 6. Apache Spark. Map-Reduce API Duration: 01:00 Lecture	Parallelization of applications using Apache Spark Duration: 02:30 Laboratory assignments		
9	Unit 7. Apache Spark RDDs. Extended API Duration: 01:00 Lecture	Parallelization of applications using Apache Spark Duration: 02:30 Laboratory assignments		
10	Unit 8. Introduction to Machine Learning Duration: 03:30 Lecture			
11	Unit 9. Supervised Machine Learning Duration: 01:30 Lecture	Unit 10. Spark Practical Assignment: Parallel implementation in Spark of a ML based Attack detector Duration: 02:00 Laboratory assignments		
12		Unit 10. Spark Practical Assignment: Parallel implementation in Spark of a ML based Attack detector Duration: 03:30 Laboratory assignments		Parallel implementation in Spark of a ML based Attack detector Group work Progressive assessment and Global Examination Not Presential Duration: 00:00
13				Individual Exam Spark RDDs & ML Written test Progressive assessment and Global Examination Presential Duration: 03:30
14				
15				
16				
17				

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

7. Activities and assessment criteria

7.1. Assessment activities

7.1.1. Assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
2	Test #1 Big Data concepts and architectures	Problem-solving test	Face-to-face	00:15	5%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
4	Test #2 ELK stack	Problem-solving test	Face-to-face	00:25	20%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
5	Test #3 Hadoop and Map-Reduce	Problem-solving test	Face-to-face	00:20	7%	0 / 10	CT.04 CG04 CB09 CG02 CT.03 CE.11 CB07 CT.01
6	Test #4 Spark SQL/Streaming	Problem-solving test	Face-to-face	00:20	8%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
6	Spark SQL/Streaming practical assignment	Group work	No Presential	00:00	10%	5 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03

							CE.11
12	Parallel implementation in Spark of a ML based Attack detector	Group work	No Presential	00:00	30%	5 / 10	CB07 CT.01 CG04 CB09 CG02 CT.03 CE.11
13	Individual Exam Spark RDDs & ML	Written test	Face-to-face	03:30	20%	5 / 10	CB07 CT.01 CG04 CB09 CG02 CT.03 CE.11

7.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
2	Test #1 Big Data concepts and architectures	Problem-solving test	Face-to-face	00:15	5%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
4	Test #2 ELK stack	Problem-solving test	Face-to-face	00:25	20%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
5	Test #3 Hadoop and Map-Reduce	Problem-solving test	Face-to-face	00:20	7%	0 / 10	CT.04 CG04 CB09 CG02 CT.03 CE.11 CB07 CT.01
6	Test #4 Spark SQL/Streaming	Problem-solving test	Face-to-face	00:20	8%	0 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11

6	Spark SQL/Streaming practical assignment	Group work	No Presential	00:00	10%	5 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
12	Parallel implementation in Spark of a ML based Attack detector	Group work	No Presential	00:00	30%	5 / 10	CB07 CT.01 CG04 CB09 CG02 CT.03 CE.11
13	Individual Exam Spark RDDs & ML	Written test	Face-to-face	03:30	20%	5 / 10	CB07 CT.01 CG04 CB09 CG02 CT.03 CE.11

7.1.3. Referred (re-sit) examination

Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
The student will take all individual evaluations of Units 1, 2, 3 and 4	Problem-solving test	Face-to-face	01:30	40%	5 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11
The student will hand the practical assignment "Parallel implementation in Spark of a ML based Attack detector" up to the day of the written exam	Group work	Face-to-face	00:00	30%	5 / 10	CG04 CB09 CG02 CT.03 CE.11 CB07 CT.01

The student will take an individual written exam covering Units 5-10	Written test	Face-to-face	02:00	20%	5 / 10	CB07 CT.01 CG04 CB09 CG02 CT.03 CE.11
The student will hand the practical assignment "Spark SQL/Streaming" up to the day of the written exam	Group work	Face-to-face	00:00	10%	5 / 10	CB07 CT.01 CT.04 CG04 CB09 CG02 CT.03 CE.11

7.2. Assessment criteria

There will be several individual exams throughout the course, and some practical assignments which will be done in groups. See tables above for details.

8. Teaching resources

8.1. Teaching resources for the subject

Name	Type	Notes
Course slides	Bibliography	Course slides
Hadoop	Bibliography	Hadoop: The Definitive Guide. White, Tom. O'Reilly Media, 2015. ISBN-13: 9781491901632.
Spark	Bibliography	Spark - The Definitive Guide: Big data processing made simple. Chambers, Bill; Zaharia, Matei. O'Reilly Media, 2018. ISBN-13: 9781491912218.
Spark 2	Bibliography	Learning Spark: Lightning-fast Data Analytics. Damji, Jules; Wenig, Brooke; Das, Tathagata; Lee, Denny. O'Reilly Media, 2020. ISBN-13: 9781492050049.
Kafka	Bibliography	Kafka - The Definitive Guide: Real-Time Data and Stream Processing at Scale. Shapira, Gwen; Palino, Todd; Sivaram, Rajini; Petty, Krit. O'Reilly Media, 2021. ISBN-13: 9781492043089
Elasticsearch	Bibliography	Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. Gormley, Clinton; Tong, Zachary. O'Reilly Media, 2015. ISBN-13: 9781449358549
Virtual Machines	Equipment	Virtual machines with an environment for Big Data practices: ELK stack, Hadoop, Spark, Kafka, Visual Studio Code, git, simulators, etc.

9. Other information

9.1. Other information about the subject

Sustainable Development Goals:

This subject is aligned with Goal 4 Quality Education that aims to "Guarantee inclusive, equitable and quality education and promote lifelong learning opportunities for all" that defined within the Sustainable Development Goals (SDGs) established by the United Nations Organization (UN).