



UNIVERSIDAD
POLITÉCNICA
DE MADRID

PROCESO DE
COORDINACIÓN DE LAS
ENSEÑANZAS PR/CL/001



E.T.S. de Ingeniería de
Sistemas Informáticos

ANX-PR/CL/001-01

GUÍA DE APRENDIZAJE

ASIGNATURA

613000129 - Gestión De Sistemas De Datos Masivos

PLAN DE ESTUDIOS

61AH - Máster Universitario En Aprendizaje Automático Y Datos Masivos

CURSO ACADÉMICO Y SEMESTRE

2023/24 - Primer semestre

Índice

Guía de Aprendizaje

1. Datos descriptivos.....	1
2. Profesorado.....	1
3. Conocimientos previos recomendados.....	2
4. Competencias y resultados de aprendizaje.....	2
5. Descripción de la asignatura y temario.....	4
6. Cronograma.....	6
7. Actividades y criterios de evaluación.....	8
8. Recursos didácticos.....	11
9. Otra información.....	13

1. Datos descriptivos

1.1. Datos de la asignatura

Nombre de la asignatura	613000129 - Gestión de Sistemas de Datos Masivos
No de créditos	3 ECTS
Carácter	Obligatoria
Curso	Primer curso
Semestre	Primer semestre
Período de impartición	Septiembre-Enero
Idioma de impartición	Castellano
Titulación	61AH - Máster Universitario en Aprendizaje Automático y Datos Masivos
Centro responsable de la titulación	61 - Escuela Técnica Superior De Ingeniería De Sistemas Informáticos
Curso académico	2023-24

2. Profesorado

2.1. Profesorado implicado en la docencia

Nombre	Despacho	Correo electrónico	Horario de tutorías *
Joaquin Gayoso Cabada	1105	j.gayoso@upm.es	Sin horario. solicitar tutoria vía email
Carlos Badenes Olmedo (Coordinador/a)	1304	carlos.badenes@upm.es	Sin horario. solicitar tutoria vía email

Andrea Jesus Cimmino Arriaga	1304	andreajesus.cimmino@upm. es	Sin horario. solicitar tutoria vía email
---------------------------------	------	--------------------------------	--

* Las horas de tutoría son orientativas y pueden sufrir modificaciones. Se deberá confirmar los horarios de tutorías con el profesorado.

3. Conocimientos previos recomendados

3.1. Asignaturas previas que se recomienda haber cursado

El plan de estudios Máster Universitario en Aprendizaje Automático y Datos Masivos no tiene definidas asignaturas previas recomendadas para esta asignatura.

3.2. Otros conocimientos previos recomendados para cursar la asignatura

- Sistemas de Virtualización Ligera basados en Contenedores (Docker)
- Programación en lenguaje Java
- Programación en lenguaje Python
- Conocimientos básicos de administración/configuración de sistemas y redes TCP/IP
- Bases de Datos Relacionales: lenguaje SQL y SGBD MySQL

4. Competencias y resultados de aprendizaje

4.1. Competencias

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

CB7 - 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

CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

CB9 - 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

CE02 - Gestionar grandes volúmenes de información procedentes de diversas fuentes heterogéneas y efectuar su análisis en tiempo real mediante el diseño de cuadros de mando e informes para obtener conocimiento útil.

CG1 - Capacidad para aplicar el método científico y saber organizar y planificar experimentos con rigor metodológico en el ámbito del aprendizaje automático y los datos masivos

CG2 - Participar en la aplicación de mecanismos de descripción, cuantificación, análisis, interpretación y evaluación de resultados experimentales del ámbito de los datos masivos y el aprendizaje automático

CG3 - Capacidad para reunir e interpretar datos masivos relevantes para emitir juicios que incluyan una reflexión sobre temas importantes de índole científico, social o ético en el ámbito del aprendizaje automático y los datos masivos

CG4 - Capacidad de aplicar iniciativa, integración, colaboración y potenciación de la discusión crítica en el ámbito del trabajo en equipo dentro del ámbito del aprendizaje automático y datos masivos

CG5 - Participar en la transmisión de la información generada, las ideas, los problemas y las soluciones de forma oral y escrita para un público tanto especializado como no especializado

CT1 - Creatividad

CT2 - Organización y planificación

CT3 - Gestión de la información

CT4 - Liderazgo de equipos

CT5 - Trabajo en contextos internacionales

K04 - El alumno compara y explica los diferentes tipos de fuentes de datos, tales como temporales, de flujo y tabulares

K05 - El alumno analiza las distintas arquitecturas para el almacenamiento y procesamiento de datos masivos de altas prestaciones

S02 - El alumno planifica y ejecuta la gestión y el despliegue de infraestructuras de datos masivos

4.2. Resultados del aprendizaje

RA21 - Comprender las teorías, modelos y técnicas actuales para la gestión, utilización y administración de sistemas de almacenamiento de datos heterogéneos

RA22 - Conocer y aplicar las técnicas de integración y transformación de datos

RA23 - Distinguir los tipos de problemas de análisis de datos

RA24 - Aplicar los criterios de diseño de cuadros de mando efectivos

RA25 - Generar informes de explotación mediante herramientas específicas.

5. Descripción de la asignatura y temario

5.1. Descripción de la asignatura

El principal objetivo de la asignatura es presentar los desafíos que supone manejar grandes volúmenes de datos para tareas de extracción de conocimiento y mostrar las técnicas y métodos utilizados para resolverlos. Los 5 retos básicos que se asumen con los datos masivos: Volumen, Velocidad, Variedad, Veracidad, y Valor, definen un escenario completamente distinto al clásico modelo de gestión de datos transaccional y requieren de una organización de datos que no será únicamente relacional. Es necesario, por tanto, conocer enfoques no relacionales para gestionar datos estructurados y no estructurados, y analizar de forma práctica cuáles son las restricciones y compromisos que los sistemas de gestión de datos masivos deben asumir. El alumno, guiado por un proceso de Extracción (E), Transformación (T) y Carga (L) que se ajustará a los clásicos enfoques ETL o más recientes ELT, aprenderá a crear, organizar y preparar grandes colecciones de datos para su posterior análisis y evaluación.

5.2. Temario de la asignatura

1. Ecosistema Big Data
 - 1.1. Arquitecturas de la Información
 - 1.2. Virtualización de Servicios
 - 1.3. Fuentes de Datos
2. Extracción de Datos
 - 2.1. Datos Estructurados
 - 2.2. Datos No Estructurados
 - 2.3. Técnicas de Captura de Información
3. Carga de Datos
 - 3.1. Alta Disponibilidad. Rendimiento
 - 3.2. Particionado y Replicación de Datos
 - 3.3. Datos Distribuidos
4. Transformación de Datos
 - 4.1. Heterogeneidad e Integración de Datos
 - 4.2. Metadatos. Automatización de Procesos
 - 4.3. Visualización de Datos

6. Cronograma

6.1. Cronograma de la asignatura *

Sem	Actividad en aula	Actividad en laboratorio	Tele-enseñanza	Actividades de evaluación
1	<p>T1. Arquitecturas de la Información Duración: 01:00 LM: Actividad del tipo Lección Magistral</p> <p>T2. Datos Estructurados Duración: 01:00 LM: Actividad del tipo Lección Magistral</p>	<p>P1. Virtualización de Microservicios Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p> <p>P2. Procesamiento de Tablas Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p>		
2	<p>T3. Datos No Estructurados Duración: 01:00 LM: Actividad del tipo Lección Magistral</p> <p>T4. Heterogeneidad e Integración de Datos Duración: 01:00 LM: Actividad del tipo Lección Magistral</p>	<p>P3. Procesamiento de Textos Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p> <p>P4. Limpieza y Enriquecimiento de Datos Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p>		<p>Preparación de Corpus TG: Técnica del tipo Trabajo en Grupo Evaluación continua No presencial Duración: 10:00</p>
3	<p>T5. Almacenamiento Orientado a Documentos Duración: 01:00 LM: Actividad del tipo Lección Magistral</p> <p>T6. Visualización de Datos Duración: 01:00 LM: Actividad del tipo Lección Magistral</p>	<p>P5. Explorador de Documentación Científica Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p> <p>P6. Interfaz de Consulta de Datos Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p>		<p>Extensión de Datos TG: Técnica del tipo Trabajo en Grupo Evaluación continua No presencial Duración: 10:00</p>
4	<p>T7. Diseño de Experimentos Duración: 01:00 LM: Actividad del tipo Lección Magistral</p> <p>T8. Metodología y Evaluación Duración: 01:00 LM: Actividad del tipo Lección Magistral</p>	<p>P7. Hackathon Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p> <p>P8. Hackathon Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p>		<p>Almacenamiento de Datos TG: Técnica del tipo Trabajo en Grupo Evaluación continua No presencial Duración: 10:00</p>
5	<p>T9. Preparación de Resultados Duración: 01:00 LM: Actividad del tipo Lección Magistral</p>	<p>P9. Hackathon Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio</p>		<p>Análisis y Presentación de Datos Masivos PG: Técnica del tipo Presentación en Grupo Evaluación continua Presencial Duración: 03:00</p>

6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				Proyecto Recuperación TI: Técnica del tipo Trabajo Individual Evaluación sólo prueba final Presencial Duración: 01:00 Examen Individual EG EX: Técnica del tipo Examen Escrito Evaluación sólo prueba final Presencial Duración: 02:00

Para el cálculo de los valores totales, se estima que por cada crédito ECTS el alumno dedicará dependiendo del plan de estudios, entre 26 y 27 horas de trabajo presencial y no presencial.

7. Actividades y criterios de evaluación

7.1. Actividades de evaluación de la asignatura

7.1.1. Evaluación (progresiva)

Sem.	Descripción	Modalidad	Tipo	Duración	Peso en la nota	Nota mínima	Competencias evaluadas
2	Preparación de Corpus	TG: Técnica del tipo Trabajo en Grupo	No Presencial	10:00	10%	5 / 10	CT1 CT4 CB7 CG3
3	Extensión de Datos	TG: Técnica del tipo Trabajo en Grupo	No Presencial	10:00	10%	5 / 10	CB8 CB6 CG4 CT5 K04
4	Almacenamiento de Datos	TG: Técnica del tipo Trabajo en Grupo	No Presencial	10:00	10%	5 / 10	CT2 CT3 CE02 K05
5	Análisis y Presentación de Datos Masivos	PG: Técnica del tipo Presentación en Grupo	Presencial	03:00	70%	5 / 10	CB9 CB10 CG1 CG2 CG5 S02

7.1.2. Prueba evaluación global

Sem	Descripción	Modalidad	Tipo	Duración	Peso en la nota	Nota mínima	Competencias evaluadas
17	Proyecto Recuperación	TI: Técnica del tipo Trabajo Individual	Presencial	01:00	30%	5 / 10	CB7 CB9 CB10 CG1 CG2 CG3 CG4 CG5 CT1 CT2 CT3 CT4 CT5 S02

17	Examen Individual EG	EX: Técnica del tipo Examen Escrito	Presencial	02:00	70%	5 / 10	CB8 CB6 CE02 K04 K05
----	----------------------	-------------------------------------	------------	-------	-----	--------	----------------------------------

7.1.3. Evaluación convocatoria extraordinaria

Descripción	Modalidad	Tipo	Duración	Peso en la nota	Nota mínima	Competencias evaluadas
Examen Convocatoria Extraordinaria	EX: Técnica del tipo Examen Escrito	Presencial	03:00	100%	5 / 10	CB7 CB8 CB9 CB10 CB6 CG1 CG2 CG3 CG4 CG5 CT1 CT2 CT3 CT4 CT5 CE02 K04 K05 S02

7.2. Criterios de evaluación

Se plantean las siguientes modalidades de evaluación, denominadas progresiva (EP) y global (EG)

1. Evaluación Progresiva (EP)

La asignatura se evaluará preferentemente de forma progresiva mediante la realización de un proyecto colaborativo que evoluciona en tres etapas, junto con un hackathon final donde se realizará el análisis y presentación final de los resultados de acuerdo con lo siguiente:

1. Los proyectos son actividades colaborativas que se realizarán en grupos de hasta 4 alumnos de entre los matriculados de la asignatura

2. Tanto la monitorización como la retroalimentación de los proyectos se harán preferentemente en la plataforma Moodle valorándose específicamente la participación de los alumnos

3. Cada etapa del proyecto cubre 4 capítulos del temario de la asignatura y su peso es el 10% de la nota total. El Hackathon final engloba todo el contenido de la asignatura, haciendo especial énfasis en el análisis crítico de los datos, y su peso es del 70%.

4. Es obligatorio realizar todas las etapas del proyecto y alcanzar en cada uno de ellas una calificación igual o superior al 35% de la valoración total del proyecto. En caso de no poder entregar alguna de las etapas se requerirá documentación que lo justifique

5. En el caso de que en todas las etapas del proyecto se haya alcanzado una nota igual o superior al requisito mínimo el grupo, y por tanto los alumnos que forman el grupo, podrán participar al final del curso en el Hackathon cuyo objetivo es doble:

Por un lado se pretende evaluar si se han alcanzado un nivel mínimo de conocimientos en los contenidos básicos de la asignatura.

En segundo lugar se persigue evaluar la aportación individual del alumno a los proyectos realizados conjuntamente con otros compañeros durante el período lectivo

2. Evaluación Global (EG)

Si no se ha superado la evaluación progresiva debido a que:

- por razones debidamente justificadas no se hubiera entregado alguna de las etapas del proyecto
- porque no se satisface el criterio mínimo en la evaluación de los mismos

el alumno realizará:

- un proyecto individual de recuperación cuyo peso es del 30% en la evaluación sumativa
- una prueba global individual cuyo peso es del 70% en la evaluación sumativa

Resumen Ordinaria/Extraordinaria

Para superar la asignatura en la convocatoria ordinaria de junio se establecen dos situaciones posibles:

1. Obtener un mínimo de 50 puntos sobre los 100 disponibles en el cómputo global de la EP
2. Si el alumno no ha superado la EP obtener una nota mínima igual o superior al 50% de la valoración en la EG

Para poder superar la asignatura en la convocatoria extraordinaria de julio, se establecen los siguientes requisitos:

1. No se evaluará mediante proyectos sino que únicamente se realizará un examen individual que cubrirá todos los aspectos teóricos y prácticos de la asignatura
2. En esta prueba se deberá obtener un mínimo de 50 puntos sobre los 100 disponibles en el cómputo global

8. Recursos didácticos

8.1. Recursos didácticos de la asignatura

Nombre	Tipo	Observaciones
First Course in Database Systems, Jeffrey D. Ullman, Jennifer Widom, ISBN-10: 013600637X. 2007	Bibliografía	
Fundamentos de Bases de Datos, A. Silberschatz, H. Korth, S. Sudarsham, 5a edición, Mcgraw-Hill, 2006	Bibliografía	

Hadoop, The Definitive Guide T. White, 4a edición, O'Reilly, 2015	Bibliografía	
Architecting Data Lakes 2Ed. Ben Sharma. O'Reilly Media 2018	Bibliografía	
Architecting Modern Data Platforms. Kunigk. O'Really 2018	Bibliografía	
Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems. Martin Kleppmann. O'Really 2017	Bibliografía	
The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science. Alex Gorelik. O'Really 2019	Bibliografía	
Big Data: Concepts, Technology, and Architecture, Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H. Gandomi, Wiley, 2021	Bibliografía	
Ethics of Big Data: Balancing Risk and Innovation, Kord Davis, O'Reilly, 2012	Bibliografía	

9. Otra información

9.1. Otra información sobre la asignatura

Actuación ante comportamientos fraudulentos

Dada la naturaleza de los conocimientos y tecnologías abarcadas en esta asignatura, se plantea el problema de la existencia de innumerables fuentes de información a disposición, desde ideas a desarrollar, pasando por códigos de todo tipo hasta aplicaciones completas. Es por esto que se premiará especialmente la originalidad y el esfuerzo propios, sobre el uso de materiales no propios. El uso de materiales ajenos de cualquier naturaleza (código, ideas, etc.) deberá ser debidamente declarado públicamente e identificado claramente, reconociendo su extensión y citando las fuentes de su autoría original. En caso contrario se considerará como plagio. Si se detecta plagio en algún proyecto, los alumnos involucrados perderán la nota que hubiera obtenido con anterioridad

Actuación ante detección de fraudes o copias/plagio

Los derechos y deberes de los estudiantes universitarios están desarrollados en los Estatutos de la Universidad Politécnica de Madrid (BOCM de 15 de noviembre de 2010) y en el Estatuto del Estudiante Universitario (RD 1791/2010 de 30 de diciembre). El artículo 124 a) de los Estatutos de la UPM fija como deber del estudiante "Seguir con responsabilidad y aprovechamiento el proceso de formación, adquisición de conocimientos, y aprendizaje correspondiente a su condición de universitario" y el artículo 13 del Estatuto del Estudiante Universitario, en el punto d) especifica también como deber del estudiante universitario "abstenerse de la utilización o cooperación en procedimientos fraudulentos en las pruebas de evaluación, en los trabajos que se realicen o en documentos oficiales de la universidad"

En el caso de que en el desarrollo de las pruebas de evaluación se aprecie el incumplimiento de los deberes como estudiante universitario, el coordinador de la asignatura podrá ponerlo en conocimiento del Director o Decano del Centro, que de acuerdo con lo establecido en el artículo 74 (n) de los Estatutos de la UPM tiene competencias para "Proponer la iniciación del procedimiento disciplinario a cualquier miembro de la Escuela o Facultad, por propia iniciativa o a instancia de la Comisión de Gobierno" al Rector, en los términos previstos en los estatutos y normas de aplicación