ANX-PR/CL/001-01

GUÍA DE APRENDIZAJE

ASIGNATURA

105001031 - Infraestructuras De Big Data

PLAN DE ESTUDIOS

10CD - Grado En Ciencia De Datos E Inteligencia Artificial

CURSO ACADÉMICO Y SEMESTRE

2022/23 - Segundo semestre

Índice

Guía de Aprendizaje

1. Datos descriptivos

1.1. Datos de la asignatura

Nombre de la asignatura	105001031 - Infraestructuras de Big Data
No de créditos	6 ECTS
Carácter	Obligatoria
Curso	Tercero curso
Semestre	Sexto semestre
Período de impartición	Febrero-Junio
Idioma de impartición	Castellano
Titulación	10CD - Grado en Ciencia de Datos e Inteligencia Artificial
Centro responsable de la titulación	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Curso académico	2022-23

2. Profesorado

2.1. Profesorado implicado en la docencia

Nombre	Despacho	Correo electrónico	Horario de tutorías *
Santiago Eibe Garcia (Coordinador/a)	2311	santiago.eibe@upm.es	Sin horario. Consultar las tutorías en el Aula Virtual de la asignatura. Concertar cita previa mediante correo electrónico (seibe@fi.upm.es)

Raul Alonso Calvo	2315	raul.alonso@upm.es	Sin horario. Sin horario. Consultar las tutorías en el Aula Virtual de la asignatura

* 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

- ProgramaciÓn Para Ciencia De Datos

- Bases De Datos I

- Arquitecturas Para El Procesamiento Masivo De Datos

- RepresentaciÓn E Intercambio De Datos

3.2. Otros conocimientos previos recomendados para cursar la asignatura

- Sistemas de Virtualización Ligera basados en contenedores (Docker)

- IDE: Eclipse, IntelliJ IDEA o similar

- Programación en lenguaje Java

- Ofimática básica

- Conocimiento y experiencia con MySQL en particular

- 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

CB02 - Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio

CB03 - Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética

CE05 - Capacidad de diseñar e implementar los procesos de selección, limpieza, transformación, integración y verificación de la calidad de los datos de cara a su posterior tratamiento.

CE06 - Capacidad para describir los fundamentos de las infraestructuras de gestión e intercambio de datos: hardware, sistemas operativos, bases de datos, redes de computadores.

CE07 - Capacidad de diseñar e implementar sistemas de información (incluyendo modelos de datos y estrategias de gestión de datos) dimensionados para gestionar el volumen, velocidad y variedad de los datos, de forma adecuada para su almacenamiento, procesamiento y acceso para tratamientos posteriores.

CE08 - Poseer las destrezas para aplicar las tecnologías actuales de computación de altas prestaciones para diseñar e implementar nuevas aplicaciones de ciencia de datos.

CG06 - Identificar y utilizar las tecnologías de la información y las comunicaciones más adecuadas en el ámbito de la ingeniería.

CG07 - Capacidad para integrar aspectos sociales, ambientales, económicos y éticos inherentes a la ingeniería, analizando sus impactos, y comprometiéndose con la búsqueda de soluciones a retos del desarrollo sostenible.

4.2. Resultados del aprendizaje

RA112 - RA-APID-10 Conocer y manejar los conceptos asociados a bases de datos no relacionales

RA107 - RA-APID-8 Conocer las infraestructuras y plataformas paralelas de procesamiento de datos

RA108 - RA-APID-9 Dimensionar sistemas informáticos para gestionar el volumen, velocidad y variedad de los datos

RA113 - RA-APID-11 Ser capaz de implementar y gestionar una base de datos en un gestor no relacional

RA122 - RA-APID-13 Emplear tecnologías e infraestructuras para el desarrollo y despliegue de servicios distribuidos, seguros, escalables, elásticos, altamente disponibles y consistentes

RA106 - RA-APID-7 Conocer y saber utilizar las técnicas fundamentales de computación de altas prestaciones

RA127 - RA128 - Conocer las infraestructuras y plataformas paralelas de procesamiento de datos

RA124 - RA130 - Conocer y manejar datos en streaming / complex event processing

5. Descripción de la asignatura y temario

5.1. Descripción de la asignatura

El principal objetivo de la asignatura es presentar la problemática asociada con la implantación de una infraestructura de ***Big Data*** caracterizada en las siguientes dimensiones:

- Validez: integridad y corrección de datos
- Variabilidad: sistemas dinámicos, cambiantes
- Volatilidad: cambios en la dimensión temporal
- Vulnerabilidad: robustez ante fallos y/o ataques
- Visualización: visualizar la utilidad de la información

Ello se traduce en los 5 retos básicos (Volumen, Velocidad, Variedad, Veracidad, Valor) que debe enfrentar toda infraestructura de Big Data y que, en conjuto, definen un escenario completamente distinto al clásico modelo de gestión de datos OLTP, mayoritariamente relacional. Se pretende poner en contraste ambos modelos mostrando las dificultades que los volúmenes masivos de datos plantean a la diversidad de sistemas de gestión de datos en la actualidad. En particular, se analizará la casuística con la que se enfrentan los **procesos *ETL*/*ELT*/*Streaming ETL*** en estos entornos

5.2. Temario de la asignatura

1. Problemática Actual de los Sistemas de Gestión de Datos
   1.1. Revisión de la Arquitectura Tradicional de los Sistemas Gestores de Bases de Datos
   1.2. Funcionalidades básicas de un SGBD(R)
2. Problemática del Almacenamiento de Datos
   2.1. Jerarquía de Almacenamiento
   2.2. Técnicas de Indexación. Rendimiento
3. Procesamiento de Interrogaciones
   3.1. Preprocesamiento y Análisis Sintáctico. Árboles de Interrogación
   3.2. Optimización. Análisis de Costes. Estimación
   3.3. Ejecución de Consultas. Operaciones en un Plan de Ejecución
4. Caso de Estudio: Relacional o No Relacional (esa es la cuestión)
   4.1. Arquitectura y Administración Básicas
   4.2. Concepto de Motor de Almacenamiento
   4.3. Funcionalidades Avanzadas. Soporte PostRelacional
   4.4. Comparativa MySQL/MariaDB vs NoSQL
5. Definición de Big Data: Problemática 5 Vs
   5.1. Almacenamiento vs Procesamiento (Computación)
   5.2. Alta Disponibilidad. Heterogeneidad e Integración de Datos
   5.3. Particionado y Replicación de Datos
   5.4. Datos Distribuidos
   5.5. Visualización de Datos
6. OLAP: Introducción al Data Warehouse
   6.1. Principios Básicos del Data Warehouse
   6.2. Data Warehouse vs Data Lake
   6.3. Modelo Multidimensional. BD del Data Warehouse
   6.4. Módulos y Funcionalidad de un Data Warehouse
   6.5. Técnicas y Herramientas OLAP

6.6. Caso de Estudio: Pentaho BI Server

7. ETL. Tipos y Problemática

7.1. Funcionalidad ETL

7.2. Taxonomía: ETL vs ELT vs Streaming ETL

7.3. Metadatos. Automatización de Procesos

7.4. Estudio de Herramientas ETL. Spoon, Talend

8. ETLs en los Ecosistemas de Big Data

8.1. Ecosistema de Hadoop: Arquitectura y Componentes

8.2. Caso de Estudio: Hive

9. Apache Spark

9.1. Abstracción Resilient Distributed Datasets (RDD)

9.2. Modelo de Computación In-Memory. Spark vs Hadoop

9.3. SparkSQL. DataFrames

10. Procesamiento de Flujos de Datos: Streaming

10.1. Problemática Streaming ETL

10.2. DSMS: Data Streams Management Systems

10.3. Visualización de Datos en Tiempo Real

10.4. Framework Spark Streaming

10.5. Caso de Estudio: Kafka

6. Cronograma

6.1. Cronograma de la asignatura *

Sem	Actividad en aula	Actividad en laboratorio	Tele-enseñanza	Actividades de evaluación
1	**1. Problemática Actual de los Sistemas de Gestión de Datos** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
2	**1. Problemática Actual de los Sistemas de Gestión de Datos** Duración: 02:00 LM: Actividad del tipo Lección Magistral **2. Almacenamiento de Datos** Duración: 02:00 LM: Actividad del tipo Lección Magistral			
3	**3. Procesamiento de Interrogaciones** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
4		**4. Caso de Estudio: Relacional vs No Relacional** Duración: 04:00 PL: Actividad del tipo Prácticas de Laboratorio		
5	**5. Volúmenes Masivos de Datos. Big Data** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
6	**5. Volúmenes Masivos de Datos. Big Data** Duración: 04:00 LM: Actividad del tipo Lección Magistral			**Operacional Relacional vs NoSQL** TG: Técnica del tipo Trabajo en Grupo Evaluación continua No presencial Duración: 40:00
7	**6. OLAP. Introducción al Data Warehouse** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
8	**6. Infraestructura Data Warehouse** Duración: 02:00 LM: Actividad del tipo Lección Magistral **7. ETL. Tipos y Problemáticas** Duración: 02:00 LM: Actividad del tipo Lección Magistral			
9	**7. ETL. Tipos y Problemáticas** Duración: 02:00 LM: Actividad del tipo Lección Magistral	**7. ETL. Sakila Star** Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio		

10	**8. ETLs en los Ecosistemas de Big Data** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
11	**8. ETLs en los Ecosistemas de Big Data** Duración: 02:00 LM: Actividad del tipo Lección Magistral	**8. Caso de Estudio: Hive** Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio		
12	**9. Apache Spark** Duración: 04:00 LM: Actividad del tipo Lección Magistral			**Ecosistemas de Big Data** TG: Técnica del tipo Trabajo en Grupo Evaluación continua No presencial Duración: 54:00
13	**9. Apache Spark** Duración: 02:00 LM: Actividad del tipo Lección Magistral	**9. Apache Spark** Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio		
14	**10. Procesamiento de Flujos de Datos: Streaming** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
15	**10. Procesamiento de Flujos de Datos: Streaming** Duración: 02:00 LM: Actividad del tipo Lección Magistral	**10. Procesamiento de Flujos de Datos: Streaming** Duración: 02:00 PL: Actividad del tipo Prácticas de Laboratorio		**Proyecto Recuperación** TI: Técnica del tipo Trabajo Individual Evaluación sólo prueba final No presencial Duración: 25:00
16	**10. Procesamiento de Flujos de Datos: Streaming** Duración: 04:00 LM: Actividad del tipo Lección Magistral			
17				**Nota Participación NPa** OT: Otras técnicas evaluativas Evaluación continua y sólo prueba final Presencial Duración: 00:00 **Examen Individual EP** EX: Técnica del tipo Examen Escrito Evaluación continua Presencial Duración: 02:00 **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.

* El cronograma sigue una planificación teórica de la asignatura y puede sufrir modificaciones durante el curso derivadas de la situación creada por la COVID-19.

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
6	Operacional Relacional vs NoSQL	TG: Técnica del tipo Trabajo en Grupo	No Presencial	40:00	25%	3.5 / 10	CG06 CE06
12	Ecosistemas de Big Data	TG: Técnica del tipo Trabajo en Grupo	No Presencial	54:00	35%	3.5 / 10	CE08 CG06 CE05 CE06 CE07
17	Nota Participación NPa	OT: Otras técnicas evaluativas	Presencial	00:00	10%	0 / 10	
17	Examen Individual EP	EX: Técnica del tipo Examen Escrito	Presencial	02:00	30%	5 / 10	CE05 CE06

7.1.2. Prueba evaluación global

Sem	Descripción	Modalidad	Tipo	Duración	Peso en la nota	Nota mínima	Competencias evaluadas
15	Proyecto Recuperación	TI: Técnica del tipo Trabajo Individual	No Presencial	25:00	30%	5 / 10	
17	Nota Participación NPa	OT: Otras técnicas evaluativas	Presencial	00:00	10%	0 / 10	
17	Examen Individual EG	EX: Técnica del tipo Examen Escrito	Presencial	02:00	60%	5 / 10	CE08 CG06 CE05 CE06 CE07

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	02:00	100%	5 / 10	CE05 CE06

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 **2 proyectos colaborativos P1, P2** más un **examen individual** de acuerdo con lo siguiente:

1. Los proyectos son **actividades colaborativas** que se realizarán obligatoriamente durante el período lectivo y 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. Dicha participación en las actividades de la asignatura, tanto en el aula como a través de la plataforma de teleenseñanza, se valorará como **NPa**. El peso de este item es del **10%** del total
4. El proyecto **P1** cubre los 5 primeros capítulos del temario de la asignatura y su peso es el **25%** de la nota total. El proyecto **P2** abarca los 5 restantes siendo el peso el **35%**. Por consiguiente, los proyectos suponen el **60%** del total de la nota de la evaluación progresiva (ver tabla de evaluación sumativa apartado 7.1.1)
5. Es obligatorio realizar ambos proyectos y alcanzar en cada uno de ellos una calificación igual o superior al **35%** de la valoración total del proyecto. En caso de no poder entregar alguno de los proyectos se requerirá documentación que lo justifique
6. En el caso de que en ambos proyectos se haya alcanzado una nota igual o superior al requisito mínimo el alumno realizará al final del curso un **examen individual** 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 y/o aquéllos no cubiertos suficientemente por los proyectos
    - 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 alguno de los proyectos
- porque no se satisface el criterio mínimo en la evaluación de los mismos

el alumno realizará:

- un proyecto individual de recuperación de la nota de proyectos cuyo peso es del **30%** en la evaluación sumativa (ver tabla apartado 7.1.2)
- una prueba global individual cuyo peso es del **60%** en la evaluación sumativa (ver tabla apartado 7.1.2)

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
Building the Data Warehouse. W.H. Immon. 1996. Willey	Bibliografía	
Managing the Data Warehouse. W.H. Immune. 1997. Willey	Bibliografía	
Building the Operational Data Store. W. H. Immon. 1999. Willey	Bibliografía	
Exploration Data Warehouse. W. Immon. 2000. Willey	Bibliografía	
The Data Warehouse Lifecycle Toolkit. R. Kimball. 2000. Willey	Bibliografía	
Improving Data Warehouse and Business Information Quality. Methods for Reducing Cost and Increasing Profits. L. English. 1999 Willey	Bibliografía	
Principles of Data Base Systems (Second Edition), Jeffrey D. Ullman, Ed. Computer Science Press, Rockville, Maryland, 1982	Bibliografía	
First Course in Database Systems, Jeffrey D. Ullman, Jennifer Widom, ISBN-10: 013600637X. 2007	Bibliografía	
Sistemas de Bases de Datos, R. Elmasri y S.B.Navathe, 2ª edición, Addison-Wesley Iberoamericana, 1997	Bibliografía	

Fundamentos de Bases de Datos, A. Silberschatz, H. Korth, S. Sudarsham, 5ª edición, Mcgraw-Hill, 2006	Bibliografía	
Database Systems-A Practical Approach to Design, Implementation and Management. 4th ed., Connolly, T., Begg, C. AND Strachan, A., 2004. Addison-Wesley	Bibliografía	
Fundamentals of Database Systems, 5th ed ., Elmasri, R. Navathe, S., 2006.. Addison-Wesley	Bibliografía	
MySQL Administrator?s Bible, Sheeri Cabral, Keith Murphy. Wiley Publishing 2009	Bibliografía	
The Complete Reference MySQL, Vikram Vaswani.McGraw-Hill/Osborne 2007	Bibliografía	
Database Systems: The Complete Book, Hector Garcia-Molina, Jeff Ullman, and Jennifer Widom. (DS-CB), 2008, 2nd edition	Bibliografía	
Mondrian 3.0 Technical Guide	Bibliografía	
Pentaho Data Integration 2ªEd María Carina Roldán	Bibliografía	
Pentaho Data Integration 4 Cookbook A.S.Pulvirenti, M.C. Roldán	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	
Real-Time Analytics. Byron Ellis. Wiley 2014	Bibliografía	
Spark - The Definitive Guide: Big Data Processing Made Simple. Bill Chambers and Matei Zaharia. O'Really 2018	Bibliografía	
The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science. Alex Gorelik. O'Really 2019	Bibliografía	
Streaming Data. Understanding The Real-Time Pipeline. Andrew G. Psaltis. Manning 2017	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