Guía docente
270678 - BDM - Administración de Datos Masivos

Última modificación: 04/02/2025

Unidad responsable: Facultad de Informática de Barcelona
Unidad que imparte: 747 - ESSI - Departamento de Ingenieria de Servicios y Sistemas de Información.

Titulación: MÁSTER UNIVERSITARIO EN INNOVACIÓN E INVESTIGACIÓN EN INFORMÁTICA (Plan 2012). (Asignatura optativa).
MÁSTER UNIVERSITARIO ERASMUS MUNDUS EN GESTIÓN Y ANÁLISIS DE DATOS MASIVOS (BDMA) (Plan 2021). (Asignatura obligatoria).

Curso: 2024 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: BESIM BILALLI

Otros: Segon quadrimestre:
BESIM BILALLI - 11, 12
SERGI NADAL FRANCESCH - 11, 12
UCHECHUKWU FORTUNE NJOKU - 11, 12

CAPACIDADES PREVIAS

Al ser Big Data Management la evolución del Data Warehousing, dicho conocimiento se asume en este curso. Por lo tanto, se espera conocimiento general sobre: diseño de bases de datos relacionales; Arquitectura del sistema de gestión de bases de datos; ETL y OLAP

Específicamente, se espera conocimiento sobre:
- Multidimensional modeling (i.e, star schemas)
- Querying relational databases
- Physical design of relational tables (i.e., partitioning)
- Hash and B-tree indexing
- External sorting algorithms (i.e., merge-sort)
- ACID transactions

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
CEC1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CEC2. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CEC3. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.

Genéricas:
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.

Transversales:
CTR3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo, ya sea como un miembro más, o realizando tareas de dirección con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.

Básicas:
CB7. 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

METODOLOGÍAS DOCENTES

The course comprises theory, and lab sessions.

Theory: Classical theory lectures in conjunction with complementary explanations and problem solving.

Lab: The course contents are applied in a realistic problem in the course project, done in teams, where students will put into practice the kinds of tools studied during the course. Since this course is part of the BDMA Erasmus Mundus master syllabus, this project is conducted jointly with the Viability of Business Projects (VBP) and Debates on Ethics of Big Data (DEBD) courses.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

1.Comprender los principales métodos avanzados de gestión de datos y diseñar e implementar gestores de bases de datos no relacionales, con especial énfasis en sistemas distribuidos.
2.Comprender, diseñar, explicar y llevar a cabo procesamiento paralelo de la información en sistemas distribuidos masivamente.
3.Gestionar y procesar un flujo continuo de datos.
4.Diseñar, implementar y mantener arquitecturas de sistemas que gestionan el ciclo de vida del dato en entornos analíticos.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas grupo pequeño	27,0	18.00
Horas grupo grande	27,0	18.00
Horas aprendizaje autónomo	96,0	64.00

Dedicación total: 150 h

CONTENIDOS

Introducción

Descripción:
Big Data, Cloud Computing, Escalabilidad

Diseño de Big Data

Descripción:
Polyglot systems; Schemaless databases; Key-value stores; Wide-column stores; Document-stores

Gestión de datos distribuidos

Descripción:
Transparency layers; Distributed file systems; File formats; Fragmentation; Replication and synchronization; Sharding; Distributed hash; LSM-Trees

Gestión de datos en memoria

Descripción:
NUMA architectures; Columnar storage; Late reconstruction; Light-weight compression

Procesamiento distribuido de datos

Descripción:
Distributed Query Processing; Sequential access; Pipelining; Parallelism; Synchronization barriers; Multitenancy; MapReduce; Resilient Distributed Datasets; Spark

Gestión y procesamiento de Streams

Descripción:
One-pass algorithms; Sliding window; Stream to relation operations; Micro-batching; Sampling; Filtering; Sketching

Arquitecturas de Big Data

Descripción:
Centralized and Distributed functional architectures of relational systems; Lambda architecture

ACTIVIDADES

Clases de teoría

Descripción:
En estas actividades, el profesor introducirá los principales conceptos teóricos de la asignatura. Juntamente con las exposiciones, se utilizarán actividades de aprendizaje cooperativo. Estas requerirán la participación activa de los estudiantes y, consecuentemente, serán evaluadas.

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CB7. 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
CEC1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CEC3. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CEC2. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CTR3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo, ya sea como un miembro más, o realizando tareas de dirección con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.

Dedicación: 50h
Aprendizaje autónomo: 25h
Grupo grande/Teoría: 25h

Examen

Descripción:
Examen escrito de los conceptos teórico-prácticos introducidos a lo largo del curso.

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CB7. 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
CEC1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CEC3. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CEC2. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CTR3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo, ya sea como un miembro más, o realizando tareas de dirección con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.

Dedicación: 19h
Aprendizaje autónomo: 17h
Grupo grande/Teoría: 2h

Laboratorio

Descripción:
Los estudiantes utilizarán diferentes herramientas NOSQL en entornos de pruebas.

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CB7. 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
CEC1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CEC3. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CEC2. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CTR3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo, ya sea como un miembro más, o realizando tareas de dirección con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.

Dedicación: 81h
Aprendizaje autónomo: 54h
Grupo pequeño/Laboratorio: 27h

SISTEMA DE CALIFICACIÓN

Final Mark = 60%E + 40%L

L = Weighted average of the marks of the lab deliverables and presentations
E = Final exam

BIBLIOGRAFÍA

Básica:
- Özsu, M.T.; Valduriez, P. Principles of distributed database systems. 4th ed. New York: Springer, 2020. ISBN 9783030262525.
- Liu, L.; Özsu, M.T. Encyclopedia of database systems. New York ; London: Springer, 2009. ISBN 9780387399409.
- Sadalage, P.J.; Fowler, M. NoSQL distilled: a brief guide to the emerging world of polygot persistence. Boston, Mass. ; London: Addison-Wesley, 2013. ISBN 9780321826626.
- Plattner, H.; Zeier, A. In-memory data management. 2nd ed. Berlin: Springer, 2012. ISBN 9783642295744.
- Zaharia, M. An architecture for fast and general data processing on large clusters. ACM Books, 2016. ISBN 9781970001563.
- Leskovec, J.; Rajaraman, A.; Ullman, J.D. Mining of massive datasets. 3rd ed. Cambridge: Cambridge University Press, 2020. ISBN 9781108476348.
- Aggarwal, C.C. (ed.). Data streams: models and algorithms. New York: Springer, 2007. ISBN 9780387287591.
- PALMER, Matt. Understanding ETL Data Pipelines for Modern Data Architectures [en línea]. O'Reilly Media, Inc., 2024 [Consulta: 24/02/2025]. Disponible a: https://www.oreilly.com/library/view/understanding-etl/9781098159269/. ISBN 9781098159252.

Complementaria:
- Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database system: the complete book [en línea]. Second edition, Pearson new international edition. Harlow, Essex: Pearson Education Limited, 2014 [Consulta: 14/03/2025]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5174436. ISBN 9781292024479.
- Loshin, D. Master data management. Amsterdam ; Boston: Morgan Kaufmann/Elsevier, 2009. ISBN 9781282285507.

RECURSOS

Enlace web:
- http://cs.ulb.ac.be/conferences/ebiss.html- https://deds.ulb.ac.be