Guía docente
270955 - BDM - Administración de Datos Masivos

Última modificación: 31/01/2024

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 CIENCIA DE DATOS (Plan 2021). (Asignatura obligatoria).

Curso: 2023 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: ALBERTO ABELLO GAMAZO

Otros: Segon quadrimestre:
ALBERTO ABELLO GAMAZO - 11, 12
BESIM BILALLI - 11, 12
SERGI NADAL FRANCESCH - 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:
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE4. Aplicar métodos escalables de almacenamiento y procesamiento paralelo de datos, incluyendo flujos continuos de datos, una vez identificados los más apropiados para un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos

Genéricas:
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos

Transversales:
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

Básicas:
CB10. 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.
CB6. 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.
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
CB8. 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.
CB9. 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.

METODOLOGÍAS DOCENTES

The course comprises theory, and lab sessions.

Theory: Inverted class techniques will be used, which require that the student works on the provided multimedia materials before the class. Then, theory lectures comprise the teacher's complementary explanations and problem solving.

Lab: There will be a project done in teams where students will put into practice the kinds of tools studied during the course. This will be evaluated in two deliverables and individual tests.

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. Se requerirá la participación activa de los estudiantes.

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE4. Aplicar métodos escalables de almacenamiento y procesamiento paralelo de datos, incluyendo flujos continuos de datos, una vez identificados los más apropiados para un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CB6. 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.
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
CB8. 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.
CB9. 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.
CB10. 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.

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

Examen

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

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE4. Aplicar métodos escalables de almacenamiento y procesamiento paralelo de datos, incluyendo flujos continuos de datos, una vez identificados los más apropiados para un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CB6. 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.
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
CB8. 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.
CB9. 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.
CB10. 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.

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

Laboratorio

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

Objetivos específicos:
1, 2, 3, 4

Competencias relacionadas:
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE4. Aplicar métodos escalables de almacenamiento y procesamiento paralelo de datos, incluyendo flujos continuos de datos, una vez identificados los más apropiados para un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CB6. 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.
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
CB8. 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.
CB9. 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.
CB10. 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.

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

SISTEMA DE CALIFICACIÓN

Final Mark = min(10 ; 60%E + 40%L + 10%P)

L = Weighted average of the marks of the lab deliverables and tests
E = Final exam
P = Participation in the class

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: Springer, 2009. ISBN 9780387399409.
- Sadalage, P.J.; Fowler, M. NoSQL distilled: a brief guide to the emerging world of polygot persistence. Boston, Mas; 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. [s.l]: 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.

Complementaria:
- Garcia-Molina, Hector; Ullman, Jeffrey D; Widom, Jennifer. Database systems : the complete book. Second edition, Pearson new international edition. Essex: Pearson Education Limited, [2014]. 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