Guía docente
270653 - OD - Open Data

Última modificación: 04/02/2021

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).

Curso: 2020 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: OSCAR ROMERO MORAL

Otros: Segon quadrimestre:
OSCAR ROMERO MORAL - 11, 12

CAPACIDADES PREVIAS

The student must be familiar with basics on databases and data modeling. Programming skills are also mandatory.

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.
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:
CG4. Capacidad para la dirección general y técnica de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería 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.

Transversales:
CTR1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y comprender la organización de una empresa y las ciencias que rigen su actividad; capacidad de comprender las reglas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Desarrollar la creatividad, el espíritu emprendedor y la tendencia a la innovación. 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.
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.

METODOLOGÍAS DOCENTES

El curso tiene sesiones magistrales y de laboratorio.

Magistrales: El profesor expone el tema. Los estudiantes siguen la lección. toman apuntes y preparan material adicional fuera de clase. También se les pide que participen en actividades evaluatorias durante estas sesiones.

Laboratorio: Principalmente, las sesiones de laboratorio estan dedicadas a la práctica (con o sin ordenador) de los conceptos introducidos en las sesiones magistrales. Herramientas relevantes para dichos conceptos se presentan en estas sesiones y se pide llevar a cabo pequeños proyectos en que haya que usarlas.

Proyecto: El proyecto del curso pone en práctica los conocimientos del curso en un entorno realista.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

1.Determine how to apply graph formalisms to solve the Variety challenge (data integration)
2.Master the main semantic-aware formalisms to enable semantic modeling
3.Determine how to apply graph formalisms to solve the Variety challenge (data integration)
4.Reinforce team work capabilities in order to develop innovative solutions by means of complementing the organization data with external data
5.Perform graph data processing both in centralized and distributed environments

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Grupo grande/Teoría	25,0	16.78
Grupo pequeño/Laboratorio	25,0	16.78
Actividades dirigidas	3,0	2.01
Aprendizaje autónomo	96,0	64.43

Dedicación total: 149 h

CONTENIDOS

Introducción y formalización del concepto de Variedad en Big Data y su gestión

Descripción:
Definición de las tareas de gestión de datos: desde la perspectiva de las bases de datos y de la representación del conocimiento.

Definición de Variedad en el concepto de Big Data. Heterogeneidades sintácticas y semánticas. Efecto de las heterogeneidades de datos en las diferentes tareas de gestión de datos identificadas.

Concepto de integración de datos. Definición de un marco teórico para la gestión e integración de fuentes de datos heterogéneas.

Principales componentes de un sistema de integración de datos: fuentes, esquema global o de integración y mappings.

La necesidad de un modelo de datos canónico para la integración de datos. Definición de modelo de datos. Características esenciales de los modelos de datos canónicos.

Los grafos como solución a la gestión de la variedad

Descripción:
La idoneidad de los grafos como modelo de datos canónico en sistemas de integración de datos.

Principales características de los modelos de datos en grafo. Diferencia con otros modelos de datos (especialmente con el modelo relacional).

Concepto de dato y metadato y su formalización en los modelos de grafos.

Casos de uso (especial énfasis en los beneficios topológicos de los grafos): detección de fraude, aplicaciones en bioinformática, gestión del tráfico y logística, redes sociales, etc.

Introducción a los principales modelos de grafos: property graph y knowledge graph.

Gestión de los property graphs

Descripción:
Estructuras de datos. Restricciones de integridad del modelo.

Operaciones básicas. Basadas en la topología, en el contenido y mixtas.

Lenguajes de consulta de datos en grafo. GraphQL.

Concepto de graph database. Heterogeneidad de las herramientas a la hora de implementar las estructuras internas del grafo. Impacto de dichas decisiones en las operaciones básicas presentadas.

Distributed graph database. Necesidad. Dificultades. Paradigma thinking like a vertex como estándar de facto para procesamiento de grafos distribuidos.

Principales algoritmos de procesamiento distribuido.

Gestión de los knowledge graphs

Descripción:
Estructuras de datos: lenguajes RDF. Origen y relación con Linked Open Data. Restricciones de integridad de cada lenguaje.

Estructuras de datos: RDFS y OWL. Relación con la lógica de primer orden. Fundamentos en Description Logics. Restricciones de integridad de cada lenguaje. Concepto de razonamiento.

Operaciones básicas y lenguaje de consulta. SPARQL y álgebra subyacente. Entailment regimes (razonamiento).

Concepto de triplestore. Diferencias con una graph database. Implementaciones nativas. Implementaciones basadas en el modelo relacional. Impacto de dichas decisiones en las operaciones básicas presentadas.

Distributed triplestore. Necesidad. Dificultades. Graph Engine 1.0 como paradigma de triplestore distribuido.

Principales algoritmos distribuidos.

Diferencias entre ambos paradigmas y casos de uso

Descripción:
Recapitulación de ambos modelos. Similitudes y diferencias. Conceptos exportables entre ambos modelos.

Principales casos de uso. Gestión de metadatos: semantificación del Data Lake y gobernanza de datos.

Principales casos de uso. Explotación de sus características topológicas: recomendadores sobre grafos y minería de datos en grafo.

Visualización. A través de GUI (Gephi) o programáticas (D3.js o GraphLab).

ACTIVIDADES

Lectures

Descripción:
During lectures the main concepts will be discussed. Lectures will combine master lectures and active / cooperative learning activities. The student is meant to have a pro-active attitude during active / cooperative learning activities. During master lectures, the student is meant to listen, take notes and ask questions.

Objetivos específicos:
1, 2, 3, 4, 5

Competencias relacionadas:
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CG4. Capacidad para la dirección general y técnica de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería Informática.
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.
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.
CTR1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y comprender la organización de una empresa y las ciencias que rigen su actividad; capacidad de comprender las reglas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Desarrollar la creatividad, el espíritu emprendedor y la tendencia a la innovación. 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.

Dedicación: 53h 30m
Grupo grande/Teoría: 25h 30m
Aprendizaje autónomo: 28h

Hands-on Session

Descripción:
The student will be asked to practice the different concepts introduced in the lectures. This includes problem solving either on the computer or on paper.

Objetivos específicos:
3, 4

Competencias relacionadas:
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CG4. Capacidad para la dirección general y técnica de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería Informática.
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.
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.
CTR1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y comprender la organización de una empresa y las ciencias que rigen su actividad; capacidad de comprender las reglas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Desarrollar la creatividad, el espíritu emprendedor y la tendencia a la innovación. 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.

Dedicación: 88h 30m
Grupo pequeño/Laboratorio: 25h 30m
Actividades dirigidas: 3h
Aprendizaje autónomo: 60h

Final Exam

Descripción:
Written exam of the theoretical concepts introduced along the course.

Objetivos específicos:
1, 2, 3, 5

Competencias relacionadas:
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.
CG4. Capacidad para la dirección general y técnica de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería Informática.
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.
CTR1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y comprender la organización de una empresa y las ciencias que rigen su actividad; capacidad de comprender las reglas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Desarrollar la creatividad, el espíritu emprendedor y la tendencia a la innovación. 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.

Dedicación: 10h
Grupo grande/Teoría: 2h
Aprendizaje autónomo: 8h

SISTEMA DE CALIFICACIÓN

Nota final = 10% EC + 40% EX + 40% LAB + 10% P

EX = Nota examen final
LAB = Nota ponderada de los laboratorios
EC = Nota de las actividades llevadas a cabo en las sesiones magistrales
C = Peer evaluation

EC = En algunas sesiones de teoría el docente propondrá algunas actividades que seran recogidas, antes del final de la clase, y evaluadas a posteriori.

LAB: Hay tres sesiones de prácticas, cada una con un peso potencial. Los laboratorios se haran en grupos asignados por el docente.

P: Proyecto final de curso

BIBLIOGRAFÍA

Básica:

- Lenzerini, M. "Data integration: a theoretical perspective". Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems PODS '02 [en línea]. June 2002, pp. 233¿246 [Consulta: 16/03/2020]. Disponible a: https://dl-acm-org.recursos.biblioteca.upc.edu/doi/abs/10.1145/543613.543644.- Aggarwal, C.C.; Wang, H. Managing and mining graph data. New York: Springer, 2010. ISBN 9781441960443.
- Baader, F. von. The description logic handbook: theory, implementation, and applications. 2nd ed. Cambridge: Cambridge University Press, 2007. ISBN 9780521876254.
- Abiteboul, S. Web data management. New York: Cambridge University Press, 2012. ISBN 9781107012431.
- Pan, J.Z.[et al.] (eds.). Ontology-driven software development. Berlin: Springer, 2013. ISBN 9783642312250.
- Groppe, S. Data management and query processing in semantic web databases. New York: Springer, 2011. ISBN 9783642193569.
- Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database systems: the complete book. 2nd ed. Harlow, Essex: Pearson Education Limited, 2013. ISBN 9781292037301.
- Ozsu, M.T. "A survey of RDF data management systems". Frontiers of computer science [en línea]. vol. 10, issue 3, june 2016, pp. 418¿432(2016) [Consulta: 16/03/2020]. Disponible a: https://link.springer.com/journal/volumesAndIssues/11704.- Sahu, S.; Mhedhbi, A.; Salihoglu, S.; Lin, J.; Özsu, M.T. "The ubiquity of large graphs and surprising challenges of graph processing: extended survey". The VLDB Journal [en línea]. 29-06-2019 [Consulta: 16/03/2020]. Disponible a: https://link.springer.com/openurl.asp?genre=journal&issn=1066-8888.