Guia docent
270653 - OD - Open Data

Última modificació: 14/02/2020

Unitat responsable: Facultat d'Informàtica de Barcelona
Unitat que imparteix: 747 - ESSI - Departament d'Enginyeria de Serveis i Sistemes d'Informació.

Titulació: MÀSTER UNIVERSITARI EN INNOVACIÓ I RECERCA EN INFORMÀTICA (Pla 2012). (Assignatura optativa).

Curs: 2019 **Crèdits ECTS:** 6.0 **Idiomes:** Anglès

PROFESSORAT

Professorat responsable:

Altres:

CAPACITATS PRÈVIES

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

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:
CEC1. Capacitat per aplicar el mètode científic en l'estudi i anàlisi de fenòmens i sistemes en qualsevol àmbit de la Informàtica, així com en la concepció, disseny i implantació de solucions informàtiques innovadores i originals.
CEC3. Capacitat per aplicar solucions innovadores i realitzar avanços en el coneixement que explotin els nous paradigmes de la Informàtica, particularment en entorns distribuïts.

Genèriques:
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.

Transversals:
CTR1. EMPRENEDORIA I INNOVACIÓ: Conèixer i comprendre l'organització d'una empresa i les ciències que regeixen la seva activitat; capacitat de comprendre les regles laborals i les relacions entre la planificació, les estratègies industrials i comercials, la qualitat i el benefici. Desenvolupar la creativitat, l'esperit emprenedor i la tendència a la innovació.
CTR3. TREBALL EN EQUIP: Ser capaç de treballar com a membre d'un equip, ja sigui com a un membre més, ja sigui realitzant tasques de direcció, amb la finalitat de contribuir a desenvolupar projectes d'una manera pragmàtica i amb sentit de la responsabilitat; assumir compromisos tenint en compte els recursos disponibles.

METODOLOGIES DOCENTS

El curs té sessions magistrals i de laboratori.

Magistrals: El professor exposa el tema. Els estudiants segueixen la lliçó, prenen apunts i preparen material addicional fora de classe. També se'ls pot demanar que portin a terme activitats avaluatòries dins d'aquestes sessions.

Laboratori: Principalment, les sessions de laboratori estaran dedicades a la pràctica (amb o sense ordinador) dels conceptes introduïts a les sessions magistrals. Eines rellevants pels conceptes introduïts són presentades i emprades en petits projectes en aquestes sessions.

Projecte: El projecte final intenta posar en comú tots els conceptes vists a classe en un entorn realista.

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

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.Integrate, combine and refine semi-structured or non-structured data into decisional systems
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

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores activitats dirigides	3,0	2.01
Hores grup petit	25,0	16.78
Hores aprenentatge autònom	96,0	64.43
Hores grup gran	25,0	16.78

Dedicació total: 149 h

CONTINGUTS

Introducció i formalització del concepte de Varietat en Big Data i la seva gestió

Descripció:
Definició de les tasques de gestió de dades: des de la perspectiva de les bases de dades i de la representació del coneixement.

Definició de Varietat en el món del Big Data. Heterogeneitats sintàctiques i semàntiques. Efecte de l'heterogeneitat de les dades en les diferents tasques de gestió de dades.

Concepte d'integració de daeds. Definició d'un marc teòric per a la gestió i integració de fonts de dades heterogènies.

Principals components d'un sistema d'integració de dades: fonts, esquema global i mappings.

La necessitat d'un model de dades canònic per a la integració de dades. Definició de model de dades. Característiques essencials dels models canònics de dades.

Els grafs com a solució a la gestió de la varietat

Descripció:
l'Idoneïtat dels grafs com a model de dades canònic en sistemes d'integració de dades.

Principals característiques dels models de dades de grafs. Diferencia amb altres models de dades (especialment amb el model relacional).

Concepte de dades i metadadades en els models de grafs.

Casos d'ús (èmfasi en els beneficis topològics): detecció de frau, aplicacions en bioinformàtica, gestió del trànsit i logística, xarxes socials, etc.

Introducció als principals models de graf: property graph i knowledge graph.

Gestió dels property graph

Descripció:
Estructures de dades. Restriccions d'integritat del model.

Operacions bàsiques. Basades en la topologia, contingut i híbrides.

Llenguatges de consulta per a grafs: GraphQL.

Conceptes de bases de dades graf. Heterogeneïtat de les diferents eines actual. Impacte d'aquestes heterogeneïtats en les principals operacions.

Bases de dades graf distribuïdes. Necessitat. Dificultats. El paradigma thinking like a vertex com estàndar de facto pel processament distribuït de grafs.

Principals algoritmes distribuïts de processament de grafs.

Gestió dels knowledge graph

Descripció:
Estructures de dades. RDF. Origen i relació amb Linked Open Data. Restriccions d'integritat.

Estructures de dades: RDFS i OWL. Relació amb la lògica de primer ordre. Fonaments en Description Logics. Restriccions d'integritat. Raonament.

Operacions bàsiques i llenguatge de consulta. SPARQL i la seva àlgebra. Entailment regimes (raonament).

Triplestores. Diferències amb les bases de dades de grafs. Implementacions natives i basades en l'àlgebra relacional. Impacte d'aquestes decisions en les principals operacions.

Triplestores distribuïts. Necessitats i dificultats. Graph Engine 1.0 com a pardigma de triplestore distribuït.

Principals algoritmes distribuïts.

Comparativa entre ambdos paradigmes i casos d'ús

Descripció:
Diferències entre ambdos paradigmes i casos d'ús.

Recapitulació d'ambdos models. Similituts y diferències. Conceptes exportables entre ambdos models.

Principals casos d'ús. Gestió de metadades: semantificació del Data Lake i governança de dades.

Principals casos d'ús. Explotació de les seves característiques topològiques: recomendadors sobre grafs i mineria de dades.

Visualització. A través de GUI (Gephi) oprogramàtiques (D3.js o GraphLab).

ACTIVITATS

Lectures

Descripció:
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.

Objectius específics:
1, 2, 3, 4, 5

Dedicació: 53h 30m
Grup gran/Teoria: 25h 30m
Aprenentatge autònom: 28h

Hands-on Session

Descripció:
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.

Objectius específics:
3, 4

Dedicació: 88h 30m
Grup petit/Laboratori: 25h 30m
Activitats dirigides: 3h
Aprenentatge autònom: 60h

Final Exam

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

Objectius específics:
1, 2, 3, 5

Dedicació: 10h
Grup gran/Teoria: 2h
Aprenentatge autònom: 8h

SISTEMA DE QUALIFICACIÓ

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

EX = Nota final de l'examen
LAB = Nota ponderada dels laboratoris
EC = Nota de les activitats a les sessions magistrals
P = Projecte

EC = En algunes sessions de teoria es portaran a terme activitats que, abans que acabi la classe, el docent recollirà i avaluarà a posteriori.

LAB: Hi ha tres laboratoris, cada un amb un pes potencial diferent. Els laboratoris s'han de portar a terme en groups assignats pels docents.

C: Projecte final de curs

BIBLIOGRAFIA

Bàsica:

- Lenzerini, M. "Data integration: a theoretical perspective". Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems PODS '02 [en línia]. 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 [en línia]. 2nd ed. Harlow, Essex: Pearson Education Limited, 2013 [Consulta: 16/03/2020]. Disponible a: https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?docID=5174436. ISBN 9781292037301.
- Ozsu, M.T. "A survey of RDF data management systems". Frontiers of computer science [en línia]. 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 línia]. 29-06-2019 [Consulta: 16/03/2020]. Disponible a: https://link.springer.com/openurl.asp?genre=journal&issn=1066-8888.