

Guia docent

270678 - BDM - Administració de Dades Massives

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).
ERASMUS MUNDUS EN BIG DATA MANAGEMENT AND ANALYTICS-BDMA (Pla 2017). (Assignatura obligatòria).

Curs: 2019

Crèdits ECTS: 6.0

Idiomes: Anglès

PROFESSORAT

Professorat responsable:

Altres:

CAPACITATS PRÈVIES

En ser Big Data Management l'evolució del Data Warehousing, aquest coneixement s'assumeix en aquest curs. Per tant, s'espera coneixement general sobre: disseny de bases de dades relacionals; Arquitectura del sistema de gestió de bases de dades; ETL i OLAP

Específicament, s'espera coneixement 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

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.

CEC2. Capacitat per al modelatge matemàtic, càlcul i disseny experimental en centres tecnològics i d'enginyeria d'empresa, particularment en tasques de recerca i innovació en tots els àmbits de la Informàtica.

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:

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

Transversals:

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.

Bàsiques:

CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incompleta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.



METODOLOGIES DOCENTS

El curs es compona de teoria, problemes i sessions de laboratori.

Teoria: Les classes de teoria comprenen les explicacions professor que constitueixen la part principal del curs. Els estudiants també tindran alguns continguts per llegir i preparar fora de l'aula, i se'ls demanarà que participin en activitats d'aprenentatge cooperatiu per a resoldre alguns problemes.

Laboratori: Hi haurà algunes sessions de laboratori per a presentar diferents tecnologies en un cert entorn d'execució.

Projecte: A final de curs, es farà un petit projecte on s'haurà de mostrar tots els coneixements adquirits durant el curs en una prova de concepte.

OBJECTIUS D'APRENTATGE DE L'ASSIGNATURA

1. Comprendre les diferències i beneficis de la gestió de dades a memòria.
2. Comprendre el fluxe d'execució una consulta distribuïda.
3. Identificar les dificultats de l'escalabilitat i la paral·lelització.
4. Dissenyar una base de dades distribuïda utilitzant eines NOSQL.
5. Produir un programa funcional per a processar Big Data en un entorn Cloud.
6. Gestionar i processar un Stream de dades.
7. Dissenyar l'arquitectura per a un sistema de gestió de Big Data.

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup gran	25,5	17.00
Hores aprenentatge autònom	96,0	64.00
Hores grup petit	25,5	17.00
Hores activitats dirigides	3,0	2.00

Dedicació total: 150 h

CONTINGUTS

Introducció

Descripció:

Big Data, Cloud Computing, Escalabilitat

Disseny de Big Data

Descripció:

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

Gestió de dades distribuïdes

Descripció:

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



Gestió de dades a memòria

Descripció:

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

Processament distribuït de dades

Descripció:

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

Gestió i processament de Streams

Descripció:

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

Arquitectures de Big Data

Descripció:

Centralized and Distributed functional architectures of relational systems; Data Warehousing architectures; Service Oriented Architecture; Lambda architecture

ACTIVITATS

Classes de teoria

Descripció:

En aquestes activitats, el professor introduirà els principals conceptes teòrics de l'assignatura. Juntament amb les exposicions, s'utilitzaran activitats d'aprenentatge cooperatiu. Aquestes demanaran la participació activa dels estudiants i, consegüentment, seran avaluades.

Objectius específics:

1, 2, 3, 5, 6, 7

Dedicació: 50h

Grup gran/Teoria: 25h

Aprenentatge autònom: 25h

Examen

Descripció:

Examen escrit de los conceptos teòric-pràctics introduïts al llarg del curs.

Objectius específics:

1, 2, 3, 4, 5, 6, 7

Dedicació: 19h

Grup gran/Teoria: 2h

Aprenentatge autònom: 17h



Laboratori

Descripció:

Els estudiants utilitzaran diferents eines NOSQL en entorns de proves.

Objectius específics:

2, 3, 4, 5, 6, 7

Dedicació: 81h

Grup petit/Laboratori: 27h

Aprenentatge autònom: 54h

SISTEMA DE QUALIFICACIÓ

Nota final = 50% $\min(10, 1.1 \cdot \frac{\sum(Ci \cdot Wi)}{\sum(Wi)}) + 10\% Pe + 30\% E + 10\% Pr$

Ci = Notes de les activitats col·laboratives i laboratoris

Wi = Pes igual a 1, 2, 4, 6 o 8 (depenent de la rellevància i dificultat de l'activitat / laboratori corresponent)

Pe = Avaluació entre companys

E = Examen

Pr = Projecte

Càlcul de Pe: Els estudiants tindran múltiples companys en les activitats realitzades durant el semestre i aquests l'avaluaran. Basant-se en aquestes avaluacions, el professor assignarà una nota a cada estudiant.

BIBLIOGRAFIA

Bàsica:

- 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 [en línia]. New York ; London: Springer, 2009 [Consulta: 04/03/2020]. Disponible a: <https://link.springer.com/referencework/10.1007/978-0-387-39940-9>. 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.

Complementària:

- Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database system: the complete book [en línia]. 2nd ed. Harlow, Essex: Pearson Education Limited, 2013 [Consulta: 04/03/2020]. Disponible a: <https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?docID=5174436>. ISBN 9781292037301.
- Loshin, D. Master data management [en línia]. Amsterdam ; Boston: Morgan Kaufmann/Elsevier, 2009 [Consulta: 04/03/2020]. Disponible a: <https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?docID=365618>. ISBN 9780080921211.
- Erl, T. Service-oriented architecture: analysis and design for services and microservices. 2nd ed. Boston: Prentice Hall, 2017. ISBN 9780133858587.

RECURSOS

Enllaç web:

- <http://it4bi-dc.ulb.ac.be>. PhD programme
- <http://cs.ulb.ac.be/conferences/ebiss.html>. Summer school