Universitat Autònoma de Barcelona

2019/2020

Sistemas Distribuidos

Código: 44212
Créditos ECTS: 6

Titulación	Tipo	Curso	Semestre
4313136 Modelización para la Ciencia y la Ingeniería / Modelling for Science and Engineering	OT	0	1

Contacto

Nombre: Antonio Espinosa Morales

Correo electrónico: AntonioMiguel.Espinosa@uab.cat

Uso de idiomas

Lengua vehicular mayoritaria: inglés (eng)

Equipo docente

Daniel Franco Puntes

Prerequisitos

Se recomienda tener conocimientos de programación Python y conocer el uso de sistemas Linux para el desarrollo de proyectos.

Objetivos y contextualización

Los objectivos del módul son los siguientes:¡

- Dar soluciones a problemas de análisis de datos con herramientas de código abierto
- Dar soluciones a problemas de análisis de datos como Linux, mysql y Spark
- Entender las limitaciones de las herramientas de gestión de datos para seleccionar las herramientas necesarias para un determinado problema
- Aprender metodologías de consulta en gestores de datos de cada tecnología
- Utilizar herramientas de Computación Cloud para solucionar problemas de análisis de datos
- Aplicar una metodología de análisis de datos para resolver problemas prácticos

Al final de las sesiones de teoría y de laboratorio, los estudiantes deberían tener suficientes conocimientos para enteder los requerimientos de un problema de análisis de datos en un contexto industrial. Deben poder elegir una combinación de herramientas y diseñar una solución para un problema de datos concreto

Competencias

- Analizar y evaluar arquitecturas de computadores paralelos y distribuidos, así como desarrollar y optimizar software avanzado para las mismas
- Participar en proyectos de investigación y equipos de trabajo en el ámbito de la ingeniería de la información y el cómputo de altas prestaciones.
- 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.
- 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.

- 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.
- 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.
- 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.

Resultados de aprendizaje

1. Aplicar diversas técnicas de tratamiento y análisis de los datos para preparar estos análisis en sistemas distribuidos.
2. Aplicar los conocimientos adquiridos en el diseño de sistemas de almacenamiento distribuido, para diseñar aplicaciones intensivas de datos y cómputo.
3. Conocer las características técnicas de distribución y gestión de datos y sus implicaciones de coste en entornos distribuidos.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Seleccionar tanto la plataforma distribuida, como el lenguaje más adecuado, a la hora de generar la propuesta de solución a un problema de cómputo distribuido.

Contenido

T1: Introducción a los Sistemas Distribuidos y los sistemas de Procesamiento de grandes conjuntos de datos (4 horas)

T2: Procesamiento de datos en Linux y gestión de workflows (10 horas)

- Arquitectura de sistemas
- Sistemas de Ficheros
- Herramientas de procesamiento de texto
- Gestión de workflows en Linux

T3: Bases de datos relacionales con Mysql (10 horas)

- Modelo de datos relacional
- Modelización de datos
- SQL y problem solving con queries

T4: Procesamiento de datos con las herramientas Apache(10 horas)

- Limitaciones del modelo relacional con grandes conjuntos de datos
- Consistencia débil y modelos relacionados
- Ecosistema de herramientas Apache
- Resolver problemas de gestión de datos con Hadoop, Hive y Spark

T5: Cloud computing (4 hours)

- Introduccion al cloud computing
- Análisis de datos con proveedores de Cloud: AWS / Azure

Metodología

La metodologia de trabajo combinan el desarrollo en classe y sesiones de resolución de problemas en les sessions de laboratori

Actividades

Título	Horas	ECTS	Resultados de aprendizaje
Tipo: Dirigidas			
Laboratorio	24	0,96	1, 2, 8, 7, 5, 4
Teoria	38	1,52	1, 2, 3, 6, 7, 9, 4
Tipo: Autónomas			
Desarrollo de ejercicios prácticos	62	2,48	1, 2, 7, 5

Evaluación

La avaluación de la assignatura se realitzará considerabdi la combinación del trabajo desarrollado en les sessiones de laboratori y el examen final

Actividades de evaluación

Título	Peso	Horas	ECTS	Resultados de aprendizaje
Exámen	30%	2	0,08	1, 2, 8, 4
Laboratorio Cloud Computing	10%	4	0,16	1, 3, 7
Laboratorio Linux	20%	6	0,24	1, 3, 7, 5
Laboratorio Mysql	20%	6	0,24	1, 2, 3, 6, 7, 9, 4
Laboratorio Spark	20%	8	0,32	1, 2, 3, 7, 9, 4

Bibliografía

A. Wittig, M. Wittig. "Amazon Web Services in Action", Manning, 2nd Edition, 2018.

G. Coulouris, J. Dollimore and T. Kinderg, "Distributed Systems. Concepts and design ", Addison-Wesley, 5th edition, 2012.

Bell, Charles; Kindahl, Mats; Thalmann, Lars. "MySQL High Availability". O'Reilly, 2010.

Chang, Fay, et al. "Bigtable: A Distributed Storage System for Structured Data." OSDI, 2006

Dewitt, David, and Jim Gray. "Parallel Database Systems: The Future of High Performance Database Processing." Communications of the ACM 35, no. 6 (1992): 85-98

Schwartz, Baron; Zaitsev, Peter; Tkachenko, Vadim; Zawodny, Jeremy D.; Lentz, Arjen; Balling, Derek J. "High Performance MySQL", O'Reilly, 2008.

Seyed M. M. "Saied" Tahaghoghi and Hugh E. Williams. Learning MySQL. O'Reilly, 2006

Nathan Haines. "Beginning Ubuntu for Windows and Mac Users". Apress 2015. *recurso electrónico en la biblioteca de la UAB*

William E. Shotts. "The Linux Command Line". Second Internet Edition. 2013. http://linuxcommand.org/tlcl.php

Petar Zecevic, Marko Bonaci. "Spark in Action". First Edition. Manning. 2017

V. Layka, D. Pollak. "Beginning Scala". Apress. 2015. *recurs electrónico en la biblioteca de la UAB*

Dan C. Marinescu. "Cloud Computing. Theory and Practice". Morgan-Kaufmann. 2018.

R. Buyya, R. N. Calheiros, A. V. Dastjerdi. "Big data. Principles and paradigms". Morgan-Kaufmann. 2016.