Guía docente

Identificación de la asignatura

Asignatura / Grupo	11638 - Gestión y Almacenamiento de Datos Masivos / 1
Titulación	Máster Universitario en Análisis de Datos Masivos en Economía y Empresa
Créditos	6
Período de impartición	Segundo semestre
Idioma de impartición	Castellano

Profesores

	Horario de atención a los alumnos					
Profesor/a	**Hora de inicio**	**Hora de fin**	**Día**	**Fecha inicial**	**Fecha final**	**Despacho / Edificio**
Juan Bonnín Hernández j.bonnin@uib.cat	Hay que concertar cita previa con el/la profesor/a para hacer una tutoría					

Contextualización

Se ha considerado lo siguiente para esta asignatura dentro de la especialización en tecnologías informáticas del MADM.

Con la proliferación de dispositivos y aplicaciones móviles, estamos ante una era con un crecimiento nunca antes visto de movimiento de datos y de diferentes formatos de datos.

Muchas de las empresas no tienen capacidad operativa ni contemplan el abanico de datos disponibles para mejorar la toma de decisiones y mejorar la inteligencia del negocio. Se pierden oportunidades, productividad e ingresos.

En esta asignatura analizaremos las diferentes propuestas en infraestructuras tecnológicas y técnicas que se han realizado. Especialmente, nos centraremos en una de las propuestas con mayor éxito: Apache Hadoop, un proyecto open source que implementa el modelo de programación paralelo propuesto en MapReduce y otras tecnologías de la familia Apache.

Requisitos

Para cursar esta asignatura se recomienda realizar la asignatura: 11636 - Computació al Núvol (Cloud Computing), donde se impartey configurauna infraestructura básica de análisis en el Cloud. **Ambás se pueden cursar en el mismo cursoacadémico.**

Recomendables

Se recomienda tener nociones de:

* UNIX/Linux (ver requisitos de la asignatura 11636)
* Nociones básicas de Java. Adjuntamos una serie de recursos:

Fecha de publicación: 15/07/2019

* https://docs.oracle.com/javase/tutorial
* https://www.udemy.com/java-tutorial/?locale=es_ES

Competencias

Específicas

* CE2 Capacidad para la administración y gestión de software para el procesamiento de datos masivos
* CE5 Capacidad para la utilización de herramientas disponibles para preparar y ejecutar aplicaciones para datos masivos en la nube.
* CESP2 Capacidad para seleccionar, atendiendo a criterios de eficiencia, escalabilidad, optimación de acceso, corrección de errores y adecuación al entorno de producción, las bases de datos y el paradigma de datos óptimo en soluciones “Big Data”.

Genéricas

* CG1 Saber recuperar datos y extraer conocimiento de grandes volúmenes de datos mediante la aplicación eficiente de técnicas de análisis de datos en diferentes dominios. Adoptar los modos de interacción adecuados según las tareas de usuario que se estén apoyando, en especial en aquellos casos en los que interviene el razonamiento analítico.
* CG4 Comprender y utilizar el lenguaje y las herramientas asociadas al análisis de datos para modelizar y resolver problemas complejos, reconociendo y valorando las situaciones y problemas susceptibles de ser tratados utilizando dichas herramientas y las técnicas asociadas.

Básicas

* Se pueden consultar las competencias básicas que el estudiante tiene que haber adquirido al finalizar el máster en la siguiente dirección: http://estudis.uib.cat/es/master/comp_basiques/

Contenidos

Contenidos temáticos

1. Introducción a Apache Spark
   * SPARK estructura: ventajas y componentes
   * Creación y uso de un reslident distributed datasets (RDD)
   * Datos dentro de SPARK
2. Gestión y almacenamiento de datos
   Dentro del ecosistema de Apache, veremos:
   * Apache Hive
   * Apache Pig
   * Apache Cassandra
3. MapReduce
   Anatomía de un programa enMapReduce
   Streaming en Hadoop
   Chaining jobs
   Ejercicios

Fecha de publicación: 15/07/2019

4. Gestionar MapReduce: gestión, monitorización y pruebas
   Trabajar con counters

   Uso de la interface de Hadoop CLI

   Monitorización de trabajos y su histórico

   Creación de unit-tests

Metodología docente

Actividades de trabajo presencial (1,44 créditos, 36 horas)

Modalidad	Nombre	Tip. agr.	Descripción	Horas
Clases teóricas	Clases teóricas	Grupo grande (G)	Mediante el método expositivo el profesor establecerálos fundamentos teóricos y prácticos sobre los diferentes aspectos tratados en los temas de la asignatura. Para cada tema se dará información sobre el método de trabajo aconsejable y el material didáctico adicional que el alumno deberá de utilizar para preparar de forma autónoma contenido.	20
Clases prácticas	Clases prácticas	Grupo grande (G)	Los alumnos realizarán trabajos guiados por el profesor donde se mostrará el uso de herramientas de trabajo de la asignatura.	6
Clases de laboratorio	Clases de laboratorio	Grupo mediano (M)	Configuración de entornos Apache	10

Al inicio del semestre estará a disposición de los estudiantes el cronograma de la asignatura a través de la plataforma UIBdigital. Este cronograma incluirá al menos las fechas en las que se realizarán las pruebas de evaluación continua y las fechas de entrega de los trabajos. Asimismo, el profesor o la profesora informará a los estudiantes si el plan de trabajo de la asignatura se realizará a través del cronograma o mediante otra vía, incluida la plataforma Aula Digital.

Actividades de trabajo no presencial (4,56 créditos, 114 horas)

Modalidad	Nombre	Descripción	Horas
Estudio y trabajo autónomo individual o en grupo	Práctica1	Gestión de infraestructura	57
Estudio y trabajo autónomo individual o en grupo	Práctica2	Programación bajo paradigma MapReduce	57

Fecha de publicación: 15/07/2019

Riesgos específicos y medidas de protección

Las actividades de aprendizaje de esta asignatura no conllevan riesgos específicos para la seguridad y salud de los alumnos y, por tanto, no es necesario adoptar medidas de protección especiales.

Evaluación del aprendizaje del estudiante

Fraude en elementos de evaluación

De acuerdo con el artículo 33 del Reglamento Académico, "con independencia del procedimiento disciplinario que se pueda seguir contra el estudiante infractor, la realización demostrablemente fraudulenta de alguno de los elementos de evaluación incluidos en guías docentes de las asignaturas comportará, a criterio del profesor, una minusvaloración en su calificación que puede suponer la calificación de «suspenso 0» en la evaluación anual de la asignatura".

Clases prácticas

Modalidad	Clases prácticas
Técnica	Trabajos y proyectos **(no recuperable)**
Descripción	Los alumnos realizarán trabajos guiados por el profesor donde se mostrará el uso de herramientas de trabajo de la asignatura.
Criterios de evaluación	CG1, CG2, CG4

Porcentaje de la calificación final: 10%

Clases de laboratorio

Modalidad	Clases de laboratorio
Técnica	Trabajos y proyectos **(no recuperable)**
Descripción	Configuración de entornos Apache
Criterios de evaluación	CG1, CG2, CG4

Porcentaje de la calificación final: 10%

Práctica1

Modalidad	Estudio y trabajo autónomo individual o en grupo
Técnica	Informes o memorias de prácticas **(recuperable)**
Descripción	Gestión de infraestructura
Criterios de evaluación	CE5, CE2

Porcentaje de la calificación final: 30% con calificación mínima 5

Fecha de publicación: 15/07/2019

Práctica2

Modalidad	Estudio y trabajo autónomo individual o en grupo
Técnica	Informes o memorias de prácticas **(recuperable)**
Descripción	Programación bajo paradigma MapReduce
Criterios de evaluación	CESP2,

Porcentaje de la calificación final: 50% con calificación mínima 5

Recursos, bibliografía y documentación complementaria

Transparencias y apuntes por parte de los profesores

Bibliografía básica

Hadoop Apache: http://hadoop.apache.org/
Mastering Cloud Computing. Foundations and applications programming. Morgan Kaufman. 2013Hadoop in Action. Manning. 2010

Bibliografía complementaria

Professional Hadoop Solutions, Wiley. 2013
Hadoop. The definitive Guide. Storage and Analysis at Internet Scale. O'Really Media. 2012

Fecha de publicación: 15/07/2019