Proyecto docente

Asignatura	Arquitecturas Big Data		
Materia	Tecnologías Informáticas para el Big Data		
Titulación	Máster Universitario en Inteligencia de Negocio y Big Data en Entornos Seguros		
Plan		**Código**	
Periodo de impartición	Primer Cuatrimestre	**Tipo/Carácter**	Obligatoria
Nivel/Ciclo	Máster	**Curso**	2018-2019
Créditos ECTS	3		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	Miguel Ángel Martínez Prieto y Fernando Díaz		
Datos de contacto (e-mail, teléfono…)	Escuela de Ingeniería Informática (Segovia) Plaza de Santa Eulalia 9 y 11, 40005 Segovia Teléfono: 98342300 e-mails: migumar2@infor.uva.es y fdiaz@infor.uva.es		
Horario de tutorías	Disponible en http://www.inf5g.uva.es/?q=node/20		
Coordinador			
Departamento	Informática (ATC, CCIA, LSI)		

1. Situación / Sentido de la asignatura

1.1 Contextualización

La asignatura Almacenamiento Escalable se encuadra dentro de la materia Tecnologías Informáticas para el Big Data y ofrece al alumno los conocimientos fundamentales para entender el reto que supone entender y diseñar una arquitectura Big Data y las tecnologías más destacadas que existen para abordar dicho reto.

La creciente preocupación actual, tanto de empresas como de particulares, por la gestión de sus datos es enorme. El volumen de datos que se generan actualmente está sufriendo un crecimiento exponencial que está llevando de la mano la creación de nuevas arquitecturas encargadas de almacenar cualquier tipo de dato, estructurado, semi-estructurado y no estructurado (cabe destacar el crecimiento de los datos no estructurados, el cual ronda un 63 % por año). En este ámbito comienza a surgir una nueva arquitectura, llamada Data Lake, mediante la cual se persigue almacenar y procesar cualquier tipo de datos y tratando de mejorar su tratamiento para prevenir problemas de ambigüedades en dichos datos. Esta arquitectura tiene grandes puntos de ruptura con las arquitecturas tradicionales, tales como los Data Warehouse. En esta asignatura se presentará el concepto y las ideas principales sobre Data Lakes y se realizará una aproximación práctica a su desarrollo e implementación.

En resumen, esta asignatura se divide en tres bloques temáticos diseñados para que el alumno obtenga los conocimientos necesarios para poder tomar decisiones efectivas de almacenamiento de Big Data. En el primer bloque se introducirán los conceptos principales sobre modelos arquitectónicos para Big Data. Además, se presentará el concepto de Data Lake y se aprenderá a modelar e implementar los componentes fundamentales de un Data Lake utilizando tecnologías de referencia en el ecosistema Big Data. En el segundo bloque, se presentarán varias herramientas destinadas al transporte de datos, que se responsabilizan de la *ingesta* (desde las fuentes de datos externas hacia HDFS) y la *carga* (desde HDFS hacia los sistemas de gestión) de datos. En el tercer bloque se motivará la importancia del preprocesamiento de los datos y se presentarán algunas de las herramientas más utilizadas para transformar y cargar los datos “en bruto”, como paso previo a su almacenamiento definitivo.

1.2 Relación con otras asignaturas

La Arquitectura Big Data es un aspecto transversal a cualquier sistema informático que gestione grandes colecciones de datos. Por lo tanto, los contenidos impartidos en esta asignatura están relacionados de forma directa con otras asignaturas del plan de estudios, en particular con Almacenamiento Escalable e Infraestructura para el Big Data.

1.3 Prerrequisitos

Se recomienda que el alumno, en sus estudios de grado, haya adquirido un mínimo de competencias en relación con el uso, configuración y administración, y conocimiento de los lenguajes de programación utilizados en sistemas operativos, sistemas distribuidos y sistemas de bases de datos.

2. Competencias

2.1 Generales del título

CG1. Adquisición de competencias teóricas y prácticas para el análisis y diseño de soluciones empresariales en Big Data (almacenamiento y procesamiento de grandes volúmenes de información heterogénea).

2.2 Específicas materia

CBD2. Capacidad de analizar, diseñar y construir o configurar sistemas de almacenamiento escalable y procesamiento escalable

3. Resultados de aprendizaje

Al finalizar la asignatura, el alumno será capaz de …

- Conocer los modelos arquitectónicos de referencia para el diseño e implementación de sistemas Big Data.
- Conocer el concepto de Data Lake y comprender las características básicas y responsabilidades de sus componentes arquitectónicos.
- Aprender a modelar e implementar los componentes fundamentales de un Data Lake utilizando tecnologías de referencia en el ecosistema Big Data.
- Conocer los fundamentos del sistema de ficheros distribuido de Hadoop (HDFS).
- Aprender a modelar e implementar flujos de ingesta de datos con servicios como Flume o Kafka.
- Aprender a implementar tareas individuales de transformación de datos utilizando Pig o Hive y a construir workflows complejos mediante Oozie.

4. Contenido / Programa de la asignatura

4.1 Unidades docentes (bloques de contenidos)

- Modelos Arquitectónicos: Introducción; Arquitectura Lambda; Arquitectura Kappa; Data Lakes.
- Ingesta y Almacenamiento de Datos: Introducción; Flume; Kakfa; HDFS.
- Transformación y Exploración de Datos: Introducción; Pig; Hive; Oozie.

4.2 Bibliografía

CAPRIOLO, E., WAMPLER, D., RUTHERGLEN, J. “Programming Hive”. 1st Ed. O'Reilly Media. 2012.

GATES, A. “Programming Pig”. 1st Ed. O'Reilly Media. 2011.

KIMBALL, R., CASERTA, J. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data”. 1st Ed. Wiley & Sons. 2004.

WHITE, T. “Hadoop: The Definitive Guide”. 4th Ed. O'Reilly Media. 2015

5. Metodología de enseñanza y dedicación del estudiante a la asignatura

Actividad Formativa	Competencias relacionadas	Horas	Presencialidad (%)
Clases, conferencias y técnicas expositivas	CG1, CBD2	12	0
Actividades autónomas y en grupo (trabajos y lecturas dirigidas)	CG1, CBD2	45	0
Pruebas de seguimiento y exposición de trabajos	CG1, CBD2	10	50
Tutoría individual, participación en foros y otros medios colaborativos	CG1, CBD2	8	0

6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Modelos Arquitectónicos	0,6	Noviembre 2018
Ingesta y Almacenamiento de Datos	1,0	Noviembre 2018
Transformación y Exploración de Datos	1,4	Noviembre 2018

7. Evaluación

Instrumento / Procedimiento	Peso primera convocatoria	Peso segunda convocatoria
Evaluación sumativa, que incluye pruebas parciales individuales y prueba final	20%	20%
Realización de trabajos, proyectos, resolución de problemas y casos	60%	60%
Participación en foros y otros medios participativos	20%	20%

8. Recursos de aprendizaje y apoyo tutorial del curso online

Transparencias.
Enunciados de ejercicios.
Cuestionarios de autoevaluación.
Páginas Webs relacionadas
Bibliografía disponible en la Biblioteca
Tutorías individualizadas o en grupo a demanda de los alumnos.

9. Consideraciones / Comentarios adicionales