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	621	**Código**	54547
Periodo de impartición	1º Cuatrimestre	**Tipo/Carácter**	Obligatoria
Nivel/Ciclo	Máster	**Curso**	1
Créditos ECTS	3		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	Miguel Ángel Martínez Prieto y Fernando Díaz Gómez		
Datos de contacto (e-mail, teléfono…)	Escuela de Ingeniería Informática Plaza de la Universidad 1, 40005 Segovia migumar2@infor.uva.es / fdiaz@infor.uva.es		
Horario de tutorías			
Coordinador			
Departamento	Informática (ATC, CCIA, LSI)		
Web			
Descripción General			

1. Situación / Sentido de la asignatura

1.1 Contextualización

La asignatura Arquitecturas Big Data se encuadra dentro de la materia Tecnologías Informáticas para el Big Data y ofrece al alumno los conocimientos fundamentales para comprender 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 al desarrollo e implementación de *dataflows* utilizando sus recursos.

Esta asignatura se divide en tres bloques temáticos diseñados para que el alumno obtenga los conocimientos necesarios para poder tomar decisiones efectivas de extracción, almacenamiento y transformación 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 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 de transformar el Big Data para satisfacer las necesidades particulares establecidas en los diferentes dominios de aplicación. Para ello, se presentarán algunas de las herramientas más utilizadas para transformar y cargar los datos “en bruto”, como paso previo a su explotación.

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, Infraestructura para el Big Data y Modelos de Programación 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.
- Aprender a implementar tareas individuales de transformación de datos utilizando Pig o Hive y a construir *dataflows* complejos mediante Oozie.

4. Contenido / Programa de la asignatura

4.1 Unidades docentes (bloques de contenidos)

- **Modelos Arquitectónicos:**
  - Data Lakes: Conceptos Básicos; Arquitectura; Despliegue.
  - Dataflows: Análisis; Selección de las Fuentes de Datos; Diseño del *Dataflow*.
- **Extracción y Almacenamiento de Big Data:**
  - Apache Flume: Fundamentos; Arquitectura; Configuración; Fuentes y Sumideros.
  - HDFS: Introducción; Conceptos HDFS; Operaciones Básicas.
- **Transformación de Big Data:**
  - Apache Pig: Conceptos Básicos; Pig Latin; Pig en *Dataflows*.
  - Apache Hive: Introducción; Lenguaje HiveQL (DDL/DML); Tuning; Hive en *Dataflows*.
  - Apache Oozie: Introducción; Workflows; Coordinadores.

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 ECTS	Semana 8 (noviembre 2019)
Extracción y Almacenamiento de Big Data	1,0 ECTS	Semanas 9 - 10 (noviembre 2019)
Transformación de Big Data	1,4 ECTS	Semanas 10 – 11 (noviembre – diciembre, 2019)

7. Evaluación

Instrumento / Procedimiento	Peso primera convocatoria	Peso segunda convocatoria
Realización de trabajos, proyectos, resolución de problemas y casos	80%	80%
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

Cualquier información de interés para el desarrollo de la asignatura, que no haya sido recogida en esta guía docente, será publicada con antelación en el curso correspondiente del Campus Virtual.