Guía docente de la asignatura

Asignatura	ALMACENAMIENTO ESCALABLE		
Materia	BIG DATA		
Módulo	BIG DATA		
Titulación	MÁSTER EN INGENIERÍA INFORMÁTICA		
Plan	510	**Código**	53204
Periodo de impartición	1er CUATRIMESTRE	**Tipo/Carácter**	OPTATIVA
Nivel/Ciclo	MASTER	**Curso**	2º
Créditos ECTS	6 ECTS		
Lengua en que se imparte	Español		
Profesor/es responsable/s	ANÍBAL BREGÓN BREGÓN FERNANDO DÍAZ GÓMEZ MIGUEL A. MARTÍNEZ PRIETO		
Datos de contacto (E-mail, teléfono…)	anibal@infor.uva.es fdiaz@infor.uva.es migumar2@infor.uva.es		
Horario de tutorías	Véase http://www.uva.es/export/sites/uva/2.docencia/2.02.mastersoficiales/2.02.01.ofertaeducativa/2.02.01.01.alfabetica/Ingenieria-Informatica/		
Departamento	Informática (ATC, CCIA, LSI)		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

La asignatura *Almacenamiento Escalable* se encuadra dentro del módulo *Big Data* y ofrece al alumno los conocimientos fundamentales para entender el reto que supone preservar grandes colecciones de datos y las tecnologías más destacadas que existen para abordar dicho reto.

La escalabilidad del almacenamiento de Big Data está directamente relacionada con la utilización de sistemas de archivos distribuidos. Este tipo de infraestructura permite añadir recursos de almacenamiento con los que afrontar las necesidades crecientes que presentan los sistemas informáticos que gestionan grandes colecciones de datos. En la actualidad, el sistema de archivos HDFS *(Hadoop Distributed File System)* es la solución más utilizada en el ámbito Big Data y, por tanto, la referencia a manejar en este ámbito. Los recursos que proporciona HDFS han servido para el desarrollo de soluciones de almacenamiento de más alto nivel: las bases de datos NoSQL (*Not Only SQL*). Aunque existen múltiples tipos de soluciones NoSQL, todas ellas

comparten una naturaleza distribuida y que, por tanto, garantiza su escalabilidad. A lo largo de esta asignatura se profundizará en HDFS y en las bases de datos NoSQL más utilizadas en diferentes áreas centradas en la explotación de Big Data. Finalmente, cabe destacar que a pesar de las bondades de esta infraestructura tecnológica, los resultados de la explotación de Big Data están directamente relacionados con la calidad de los datos almacenados. Este supone que, en la mayoría de los casos, los datos no se almacenan directamente en la base de datos seleccionada, sino que precisan de un proceso previo de “limpieza”, en el que los sistemas de almacenamiento también juegan un papel importante.

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 sistemas de ficheros distribuidos y se presentará HDFS tanto a nivel teórico como práctico. En el segundo 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. Este aspecto define el tema principal del tercer y el último bloque, en el que se abordarán los principios fundamentales de la tecnología NoSQL y se presentarán algunos de los sistemas de bases no relacionales más destacados en el ámbito del Big Data.

1.2 Relación con otras materias

El almacenamiento de 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:

- La relación con la *Tecnología para el Big Data* (semestre 3) es obvia, dado que las tecnologías presentadas en dicha asignatura describen los ecosistemas de referencia en el ámbito del Big Data y, por tanto, complementan las diferentes tecnologías que se presentarán en la asignatura actual.
- Los contenidos planteados en la asignatura *Técnicas Escalables de Análisis de Datos* (semestre 1) introducen al alumno en el procesamiento de Big Data y en las tecnologías que pueden utilizarse para obtener conocimiento de estas grandes colecciones de datos. La relación con la asignatura actual es directa dado que el procesamiento de Big Data está supeditado a la forma en la que se almacenan los datos y a los mecanismos disponibles para su utilización.
- Finalmente, la asignatura actual también está relacionada con *Big Data: Inteligencia de Negocios* (semestre 3). Al igual que en el caso anterior, el desarrollo de cualquier servicio de inteligencia de negocio depende de la forma en la que el Big Data haya sido almacenado.

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

CG1. Capacidad para proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la ingeniería informática.

CG4. Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la Ingeniería en Informática.

CG8. Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos.

CG9. Capacidad para comprender y aplicar la responsabilidad ética, la legislación y la deontología profesional de la actividad de la profesión de Ingeniero en Informática.

2.2 Específicas

Específicas de Tecnologías Informáticas:

CET1. Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.

CET5. Capacidad para analizar las necesidades de información que se plantean en un entorno y llevar a cabo en todas sus etapas el proceso de construcción de un sistema de información.

CET6. Capacidad para diseñar y evaluar sistemas operativos y servidores, y aplicaciones y sistemas basados en computación distribuida.

Específicas de Big Data:

CBD2. Capacidad de diseñar, parametrizar o construir sistemas complejos de inteligencia de negocio y asegurar su mantenimiento trabajando sobre herramientas específicas.

CBD4. Capacidad de implementar sistemas de descubrimiento de conocimiento en grandes bases de datos distribuidas.

CBD5. Capacidad de analizar, diseñar y construir o configurar sistemas de almacenamiento escalable y procesamiento escalable

3. Objetivos

- Modelar e implementar soluciones eficientes de Big data en diferentes áreas de aplicación, utilizando apropiadamente los algoritmos y las estructuras de datos seleccionadas.
- Entender como el uso de sistemas de ficheros distribuidos es aplicable al Big Data, cómo almacenar y consultar Big Data utilizando los entornos actualmente disponibles.
- Aplicar bases de datos no relacionales, las técnicas para almacenar y procesar grandes volúmenes de datos estructurados y no estructurados.
- Ser capaz de entender los retos que supone el almacenamiento de Big Data y como todos ellos pasan por utilizar técnicas de distribución de datos.

- Ser capaz de comprender los principios fundamentales de los sistemas de ficheros distribuidos y ponerlos en práctica con HDFS.
- Ser capaz de entender las capacidades específicas de los modelos principales de almacenamiento NoSQL y sus diferencias respecto al modelo relacional.
- Ser capaz de identificar un problema Big Data y elegir el mejor modelo de almacenamiento NoSQL para afrontarlo.
- Ser capaz de utilizar algunas de las bases de datos NoSQL más demandadas en los escenarios Big Data actuales.

4. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas (T/M)	16	Estudio y trabajo autónomo individual (conocer, comprender, plantear dudas, experimentar)	45
Laboratorios (L)	36	Estudio y trabajo autónomo individual (preparación de prácticas)	45
Trabajos tutelados		Desarrollo trabajos tutelados	
Seminarios (S)	6	Preparación presentaciones	
Tutorías activas	2		
Evaluación*			
Total presencial	**60**	**Total no presencial**	**90**

* Evaluación: Se incluye en las actividades de Laboratorio y Seminarios.

5. Bloques temáticos

Bloque 1: Arquitectura Hadoop y HDFS

Carga de trabajo en créditos ECTS: 2

a. Contextualización y justificación

Este bloque temático trata de introducir los elementos fundamentales de la arquitectura Hadoop 2.x y el papel fundamental que el Sistema de Ficheros Distribuido de Hadoop juega en el framework Map-Reduce que proporciona Hadoop. Partiendo de los requisitos de almacenamiento escalable, se introducen los elementos básicos de un sistema Hadoop, centrándose este bloque temático en las nociones y cuestiones de diseño más relevantes de su sistema de ficheros HDFS. Se practicará también con las herramientas e interfaces que el sistema proporciona para manipular y gestionar el sistema de ficheros HDFS.

b. Objetivos de aprendizaje

- Conocer los requisitos de un sistema de almacenamiento escalable
- Identificar los elementos fundamentales de una arquitectura Hadoop 2.x
- Entender las fundamentos del sistema de ficheros distribuido de Hadoop (HDFS)
- Manejar el sistema de ficheros HDFS

c. Contenidos

Introducción: sistemas de ficheros distribuidos, requisitos del almacenamiento escalable, visión general de la arquitectura Hadoop 2.x

HDFS (Hadoop Distributed File System): diseño y arquitectura de HDFS, conceptos fundamentales de HDFS, interacción con HDFS, Dataflow, Blocks y Replicación.

Procesos Hadoop: Name node, Secondary name node, Job tracker, Task tracker, Data node.

Administración, monitorización y mantenimiento HDFS

d. Métodos docentes

Ver anexo: métodos docentes.

e. Plan de trabajo

Se proporcionará al comienzo de la asignatura.

f. Evaluación

Ver apartado 7.

g. Bibliografía básica

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

h. Bibliografía complementaria

i. Recursos necesarios

El alumno deberá tener acceso a un ordenador personal
Aula virtual de la asignatura
Software: máquina virtual con distribución de Cloudera para Hadoop.

Bloque 2: Transformación y Carga de Big Data

Carga de trabajo en créditos ECTS: 2

a. Contextualización y justificación

Una vez conocidos los fundamentos del almacenamiento HDFS, y sus bondades para la gestión de colecciones Big Data, este segundo bloque de la asignatura se centra en las necesidades de “limpieza” que presentan estas colecciones de datos. Es habitual que esta etapa previa al procesamiento se lleve a cabo utilizando un proceso ETL *(Extract, Transform, Load)* o alguna de sus variantes como ELT *(Extract, Load, Transform)*. En este bloque se revisarán los conceptos fundamentales de las etapas de transformación y carga de los datos y se presentarán las diferentes estrategias (ETL vs. ELT) que se pueden utilizar para su integración en el ámbito de un proyecto Big Data. Todos estos contenidos se expondrán desde una perspectiva teórico-práctica que utilizará como referencias las herramientas *Apache Pig* y *Apache Hive*, cuyos fundamentos y usos principales serán también presentados.

b. Objetivos de aprendizaje

- Comprender la necesidad de realizar un proceso de “limpieza” de Big Data previo a su puesta en explotación.
- Conocer e integrar los procesos de transformación y carga de los datos en el flujo de trabajo de un proyecto Big Data.
- Conocer los principios fundamentales y las características funcionales de las herramientas *Apache Pig* y *Apache Hive*.
- Adquirir las destrezas necesarias para utilizar las herramientas Apache Pig y Apache Hive en el ámbito de un proyecto Big Data.

c. Contenidos

Procesos ETL: conceptos básicos, proceso de transformación, proceso de carga, ETL vs. ELT.
Apache Pig: introducción, lenguaje Pig Latin, Pig en los procesos ETL.
Apache Hive: introducción, lenguaje HiveQL (DDL/DML), tuning, HIve en los proceso ETL.

d. Métodos docentes

Ver anexo: métodos docentes.

e. Plan de trabajo

Se proporcionará al comienzo de la asignatura.

f. Evaluación

Ver apartado 7.

g. Bibliografía básica

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.

h. Bibliografía complementaria

i. Recursos necesarios

El alumno deberá tener acceso a un ordenador personal
Aula virtual de la asignatura
Software: máquina virtual con distribución de Cloudera para Hadoop.

Bloque 3: Bases de Datos NoSQL

Carga de trabajo en créditos ECTS: 2

a. Contextualización y justificación

En este último bloque temático de la asignatura se presentarán las ideas fundamentales sobre almacenamiento NoSQL y la importancia que tienen dentro del framework de Hadoop. Partiendo de las ideas básicas sobre almacenamiento escalable presentadas en el primer bloque, se introducen los conceptos fundamentales, ventajas y desventajas, y principales tipos de bases de datos NoSQL. Posteriormente se profundizará en algunas de las más importantes categorías de bases de datos NoSQL según su forma de almacenar los datos, como son el almacenamiento clave-valor y el almacenamiento documental. Para cada una de estas categorías se realizarán prácticas con algunas de las bases de datos más relevantes, como pueden ser MongoDB o Apache Cassandra. Para terminar, se presentarán otras categorías y aproximaciones para bases de datos NoSQL, como por ejemplo las de almacenamiento tabular o las orientadas a grafos, entre otras.

b. Objetivos de aprendizaje

- Ser capaz de entender los retos que supone el almacenamiento de Big Data y como todos ellos pasan por utilizar técnicas de distribución de datos.
- Ser capaz de entender las capacidades específicas de los modelos principales de almacenamiento NoSQL y sus diferencias respecto al modelo relacional.

- Ser capaz de identificar un problema Big Data y elegir el mejor modelo de almacenamiento NoSQL para afrontarlo.
- Ser capaz de utilizar algunas de las bases de datos NoSQL más demandadas en los escenarios Big Data actuales.

c. Contenidos

Almacenamiento en BBDD NoSQL: introducción, conceptos fundamentales, principales tipos de BBDD NoSQL.
Almacenamiento clave-valor: introducción, ideas principales, Cassandra.
Almacenamiento documental: introducción, ideas principales, MongoDB.
Otras aproximaciones de almacenamiento NoSQL.

d. Métodos docentes

Ver anexo: métodos docentes.

e. Plan de trabajo

Se proporcionará al comienzo de la asignatura.

f. Evaluación

Ver apartado 7.

g. Bibliografía básica

TIWARI, SHASHANK, “Professional NoSQL”, Wiley/Wrox. 2011.
STRAUCH, CHRISTOF, “NoSQL Databases”. Stuttgart Media University. 2012
CHODOROW, K. “MongoDB: The Definitive Guide”, 2nd Edition, O'Reilly Media. 2013.
CARPENTER, J., HEWITT, E. “Cassandra: The Definitive Guide”, 2nd Edition, O'Reilly Media. 2016.

h. Bibliografía complementaria

VAISH, G., “Getting Started with NoSQL”. Packt Publishing. 2013.

i. Recursos necesarios

El alumno deberá tener acceso a un ordenador personal
Aula virtual de la asignatura
Software: máquina virtual con distribución de Cloudera para Hadoop.

6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Bloque 1: Arquitectura Hadoop y HDFS	2	Semanas 1-5
Bloque 2: Transformación y Carga de Big Data	2	Semanas 6-10
Bloque 3: Bases de Datos NoSQL	2	Semanas 11-15

7. Tabla resumen de los instrumentos, procedimientos y sistemas de evaluación/calificación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Bloque 1: entrega y defensa de la práctica 1	1/3	Semana 5
Bloque 2: entrega y defensa de la práctica 2	1/3	Semana 10
Bloque 3: entrega y defensa de la práctica 3	1/3	Semana 15

Convocatoria ordinaria:

Para que el alumno resulte APTO en esta asignatura deberá haber entregado/presentado los trabajos relativos a las 3 entregas mostradas en el cuadro anterior, y haber obtenido en todas y cada una de ellas una calificación de 5.0 puntos (sobre 10) o superior.

Convocatoria extraordinaria:

Para la realización de la parte práctica será necesario contactar con los profesores y fijar una planificación para el desarrollo de la misma. En todo caso, será necesario entregar y defender un trabajo práctico y obtener una calificación de 5.0 puntos (sobre 10) o superior.

8. Anexo: Métodos docentes

ACTIVIDAD	METODOLOGÍA
Clase de teoría	• Clase magistral participativa • Estudio de casos en aula • Resolución de problemas
Clase práctica	• Realización de un trabajo práctico guiado por el profesor. • Clase magistral participativa • Resolución de casos prácticos con apoyo informático • Realización de un proyecto en grupo (2/3 alumnos) guiado por el profesor, siguiendo un enfoque colaborativo.
Seminarios	• Talleres de aprendizaje • Sesiones de debate entre alumnos y profesor sobre su aprendizaje, las técnicas estudiadas y su aplicación práctica a casos reales.

9. Anexo: Cronograma de actividades previstas

El cronograma de actividades aparecerá en el Aula Virtual al inicio de la asignatura.