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 (Segovia) Campus María Zambrano Plaza de la Universidad 1, 40005 Segovia Teléfono: 98342300 (ext. 2419) e-mails: migumar2@infor.uva.es / fdiaz@infor.uva.es		
Horario de tutorías	Disponible en http://www.inf5g.uva.es/node/765		
Coordinador			
Departamento			
Web	Informática (ATC, CCIA, LSI)		
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, así como algunas de las tecnologías más utilizadas para construir flujos de datos *(dataflows)* de datos sobre estas arquitecturas.

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. En este contexto, el *Data Lake* se ha convertido en la arquitectura de referencia para almacenar y procesar cualquier tipo de datos, así como para su exploración y explotación en entornos analíticos. 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 (ETLSs)* 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 de Transformación. En el primer bloque se introducirá el Data Lake, describiendo cada una de sus capas y analizando sus responsabilidades. Además, se revisarán algunos de los modelos arquitectónicos utilizados para la implementación del Data Lake en entornos de procesamiento por lotes *(batch)* y/o tiempo real *(streaming)*, aunque en esta asignatura nos centraremos, principalmente, en el primero de ellos. En el segundo bloque, se introducirá HDFS como tecnología de referencia para el almacenamiento de Big Data en entornos Hadoop. También se presentarán algunas de las herramientas utilizadas para el transporte de datos hacia y desde el Data Lake. Finalmente, el tercer bloque abordará las necesidades de transformación de datos en el Data Lake. Esta responsabilidad es fundamental para transformar los datos en crudo *(raw data)* que se mantiene en el Data Lake en los datos de negocio *(smart data)* que se requieren en los diferentes entornos de explotación en los que, hoy en día, se está utilizando Big Data como base para la toma de decisiones.

1.2 Relación con otras asignaturas

Las arquitecturas 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 el concepto de Data Lake y comprender las características básicas y responsabilidades de sus componentes arquitectónicos.
- Conocer los modelos arquitectónicos de referencia la construcción de Data Lakes.
- 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

Bloque 1: "Data Lakes"

Carga de trabajo en créditos ECTS: 0,6

a. Contextualización y justificación

¿Qué es Big Data? La sencillez de esta pregunta esconde la complejidad de una respuesta que depende, fuertemente, del contexto en el que plantee. En esta asignatura el Big Data se entiende como parte del contexto tecnológico de una organización y, por tanto, tenemos que ser capaces de comprender las implicaciones que tiene en este contexto y los recursos de los que disponemos para poder explotarlo en el entorno de negocio de la organización.

El Data Lake es la piedra angular en el proceso de cambio que requieren las organizaciones para poder utilizar las grandes colecciones de datos (Big Data) de las que disponen. No es un sistema gestor de bases de datos, ni un sistema de archivos, ni tampoco un entorno de computación. Sin embargo, da cabida a todas estas piezas, además de a todas aquellas que garantizan el auto-servicio de sus usuarios, entendiendo como tal a todos los miembros de la organización que utilicen el Big Data de la organización para satisfacer sus (diferentes) necesidades.

En este bloque se establecen los fundamentos teóricos necesarios para abordar el resto de la asignatura. Para ello, se planteará una pequeña motivación a las necesidades tecnológicas que plantea la gestión de Big Data en una organización y se justificará la necesidad del Data Lake para satisfacerlas. A continuación, se analizará la arquitectura del Data Lake, estableciendo la relación entre sus capas y las asignaturas que conforman el módulo "Big Data", incluyendo esta. Para finalizar, revisaremos los modelos arquitectónicos de despliegue del Data Lake más utilizados e introduciremos una metodología de ciclo de vida para proyectos Big Data, que establece las diferentes etapas por las que pasa un proyecto desde su análisis hasta su puesta en explotación.

b. Objetivos de aprendizaje

- Comprender qué es Big Data y porqué es un reto su gestión.
- Conocer el concepto de Data Lake, sus capas y cada una de sus responsabilidades.
- Entender cómo el Data Lake debe desplegarse y utilizarse en un proyecto Big Data.
- Aprender una metodología de ciclo de vida para la realización de proyectos Big Data.

c. Contenidos y materiales de aprendizaje

- Big Data: definición, contexto histórico, tecnologías y ciclo de vida.
- Data Lakes: definición, hoja de ruta, arquitectura y gestión de datos.
- Ciclo de Vida (de un proyecto Big Data): metodología, inicio, análisis, ETL, evaluación y despliegue.

d. Métodos docentes

- Material de lectura básica.
- Material audiovisual asíncrono.
- Material complementario.
- Actividades individuales y/o colaborativas (Tareas).

- Aprendizaje basado en proyectos (Proyecto).
- Tutorías online síncronas.

e. Bibliografía básica

GORELIK, A. *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. 2019. O'Reilly.
SHARMA, B..*Architecting Data Lakes*. O'Reilly, 2018.

f. Bibliografía complementaria

CHAPMA, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., WIRTH, R.. *CRISP-DM 1.0: Step-by-step data mining guide*. 2000
KIMBALL, R. y CASERTA, J..*The Data Warehouse ETL Toolkit*. John Wiley&Sons, 2004.
PRESS, G. *A Very Short History Of Big Data*. 2013. https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.
MARZ, N. y WARREN, J. *Big Data principles and best practices of scalable realtime data systems*. Manning,2015.
MATTHEWS, K. *The difference between a data swamp and a data lake? 5 signs.* 2019. https://www.information-age.com/data-swamp-data-lake-123481597/.

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
  - Asíncronas: foro, email.
  - Síncronas: sala de videoconferencia.

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
0,6	Semana 10

Bloque 2: "Capa de Almacenamiento (HDFS)"

Carga de trabajo en créditos ECTS: 1,0

a. Contextualización y justificación

El sistema HDFS, junto con la capa de entrada/salida de datos, proporciona la infraestructura básica de almacenamiento escalable en arquitecturas Big Data basadas en Hadoop. HDFS es un sistema de archivos distribuido de alto rendimiento basado en Java y que su vez, se basa en el sistema de archivos UNIX subyacente. La capa de entrada/salida de datos es la principal responsable de los diferentes formatos de archivo, técnicas de compresión y serialización de datos para el almacenamiento de Hadoop.

En este bloque temático se pretende dar a conocer las características del sistema HDFS, sus componentes arquitectónicos fundamentales y las funcionalidades básicas soportadas. Una vez introducidas las nociones fundamentales de HDFS, es importante identificar qué papel juega esta capa de almacenamiento dentro de la pila Hadoop en el contexto del desarrollo de aplicaciones Big Data. Se debe justificar la necesidad de introducir componentes de integración de datos, tales como el cliente HDFS, HUE, Flume o Sqoop, entre otros. Los componentes de integración de datos actúan como una capa transversal a toda la pila Hadoop y está formada por las herramientas, software o código de usuario que ayudan a integrar el almacenamiento de datos con las acciones o vistas de usuario. Es decir, desde el punto de vista de los usuarios, estos componentes proporcionan una vista unificada de los datos dentro de la pila Hadoop a través de las diferentes carpetas distribuidas, en diferentes ficheros y formatos de datos. A su vez, la Capa de Almacenamiento (HDFS), junto con los componentes de integración de datos, proporcionan a los usuarios y/o aplicaciones Big Data, la infraestructura básica sobre la que se asienta su desarrollo, para lo cual se hace uso de componentes de más alto nivel, típicamente interfaces de acceso a los datos (basados en lenguajes de consulta SQL, NoSQL o directamente APIs de programación) y/o los diferentes motores de procesamiento de datos disponibles en Hadoop (basados en Map-Reduce, Spark, HBASE, etc.)..

b. Objetivos de aprendizaje

- Conocer los principios de diseño del sistema de ficheros HDFS en el contexto de las aplicaciones Big Data.
- Conocer la arquitectura básica del sistema HDFS.
- Identificar el papel de la capa de almacenamiento soportada por HDFS en una arquitectura Big Data basada en Hadoop.
- Utilizar la interfaz básica de mandatos del cliente HDFS.
- Conocer la arquitectura de Flume y el enfoque de ingestión de datos en HDFS dirigido por eventos.
- Plantear pequeños flujos de ingestión de datos en HDFS basados en Flume.

c. Contenidos y materiales de aprendizaje

- Introducción a HDFS y su papel en una arquitectura Big Data basada en Hadoop.
- Arquitectura de HDFS.
- Componentes de integración de datos en HDFS (cliente HDFS, Flume, Sqoop, HUE, etc.).
- Apache Flume e ingestión de datos en HDFS basada en Flume.

d. Métodos docentes

- Material de lectura básica.
- Material audiovisual asíncrono.
- Material complementario.
- Actividades individuales y/o colaborativas (Tareas).
- Aprendizaje basado en proyectos (Proyecto).
- Tutorías online síncronas.

e. Bibliografía básica

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

SINGH, C. y KUMAR, M. “Mastering Hadoop 3: Big data processing at scale to unlock unique business insights”. Packt Publishing, 2019.

f. Bibliografía complementaria

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
  - Asíncronas: foro, email.
  - Síncronas: sala de videoconferencia.

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
1,0	Semanas 11 y 12

Bloque 3: "Capa de Transformación"

Carga de trabajo en créditos ECTS: 1,4

a. Contextualización y justificación

La transformación del *raw data* en *smart data* es el objetivo principal de un proyecto Big Data. Dicho de otra forma, la transformación se encarga de convertir los datos en bruto, de los que dispone la organización, en datos orientados al negocio y, para ello, implementa diferentes fases centradas en la limpieza, alineamiento o integración de los datos.

La capa de transformación, dentro del Data Lake, proporciona los recursos necesarios para implementar las fases anteriores. Para ello, se apoya sobre las capas inferiores del Data Lake (infraestructura, almacenamiento y procesamiento) y proporciona diferentes herramientas que permiten abordar las necesidades habituales de transformación. En este bloque presentaremos Apache Pig y Apache Hive, como ejemplos de herramientas de transformación de Big Data, que se desarrollan sobre los modelos ya conocido de la capa de computación. Apache Pig ofrece un lenguaje procedimental que permite refinar los datos de forma incremental hasta alcanzar el *smart data* deseado. Apache Hive implementa una vista tabular de los datos almacenados en el Data Lake y ofrece un lenguaje declarativo (tipo SQL) para realizar las operaciones de transformación. Finalmente, revisaremos Apache Oozie, como tecnología de referencia para orquestar diferentes procesos de transformación que, a su vez, pueden haberse implementado con cualquiera de las dos tecnologías anteriores, con otras de más bajo nivel (MapReduce o Spark) o, incluso, con soluciones más tradicionales como *scripts* desarrollados para *bash* o utilizando lenguajes como Python.

b. Objetivos de aprendizaje

- Comprender la complejidad de transformar *raw data* en *smart data*.
- Conocer los principios fundamentales de Apache Pig y aprender a utilizarlos para implementar procesos de transformación de datos en el ámbito de un Data Lake.
- Conocer los principios fundamentales de Apache Hive y aprender a utilizarlos para implementar procesos de transformación de datos en el ámbito de un Data Lake.
- Conocer los principios fundamentales de Apache Oozie y aprender a utilizarlos para construir los *dataflows* requeridos en los proyectos Big Data.

c. Contenidos y materiales de aprendizaje

- Computación Big Data: conceptos básicos, YARN, MapReduce y Spark.
- Apache Hive: introducción, gestión y manipulación de datos.
- Apache Pig: introducción, modelo de datos y Pig Latin.
- Apache Oozie: introducción, *workflows*, acciones, nodos de control y configuración.

d. Métodos docentes

- Material de lectura básica.
- Material audiovisual asíncrono.
- Material complementario.
- Actividades individuales y/o colaborativas (Tareas).
- Aprendizaje basado en proyectos (Proyecto).
- Tutorías online síncronas.

e. 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.
ISLAM, M.K. y SRINIVASAN, A. Apache Oozie. 1st Ed. O'Reilly Media. 2015.
WHITE, T. Hadoop: The Definitive Guide. O'Reilly. 2015.

f. Bibliografía complementaria

KREPS, J. *Questioning the Lambda Architecture*. O'Reilly, 2014. https://www.oreilly.com/ideas/questioning-the-lambda-architecture.
LIN, J. *The Lambda and the Kappa*. IEEE Internet Computing 21(5), pp. 60-66. 2017.
MARZ, N. *How to beat the CAP theorem*. http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html. 2011.
OLSTON, C.; REED, B.; SRIVASTAVA, U.; KUMAR, R. y TOMKINS, A. *Pig Latin: A Not-so-foreign Language for Data Processing*. En Proceedings of SIGMOD,pp. 1099--1110, 2008.
THUSOO, A.; SARMA, J.S.; JAIN, N.; SHAO, Z.; CHAKKA, P.; ANTHONY, S.; LIU, H.; WYCKOFF, P. y MURTHY, R. *Hive: a Warehousing Solution over a Map-Reduce Framework*. Proceedings of the VLDB Endowment VLDB Endowment, 2(2), 1626-1629, 2009.

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
  - Asíncronas: foro, email.
  - Síncronas: sala de videoconferencia.

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
1,4	Semanas 12, 13 y 14

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. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES	HORAS
Horas de tutoría síncrona o asíncrona	6
Horas de lectura y reproducción materiales	20
Horas de trabajo autónomo individual	20
Horas de trabajo colaborativo	27
Horas de actividades de evaluación	2
Total	75

7. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Data Lakes	0,6	Semana 10
Capa de Almacenamiento (HDFS)	1,0	Semanas 11 y 12
Capa de Transformación	1,4	Semanas 12, 13 y 14

8. Evaluación

Instrumento / Procedimiento	Peso primera convocatoria	Peso segunda convocatoria
Evaluación sumativa, que incluye pruebas parciales individuales y prueba final	30%	30%
Realización de trabajos, proyectos, resolución de problemas y casos	50%	50%
Participación en foros y otros medios participativos	20%	20%

Criterios / Comentarios a la evaluación

- **Convocatoria ordinaria:** en la convocatoria ordinaria se planteará la resolución de pequeños supuestos prácticos para cada uno de los bloques que componen la asignatura. En aquellos bloques no susceptibles de ser evaluados mediante estos supuestos, se realizará un cuestionario evaluable online. Todas estas actividades sumarán un 30% de la nota. Además, se desarrollará un pequeño proyecto Big Data (por equipos) en el que se pondrán en práctica todos los conocimientos adquiridos durante la asignatura. Esta actividad supondrá el 50% de la nota. Finalmente, se propondrá un trabajo de documentación que supondrá el 20% de la nota de la asignatura. Para aprobar la asignatura en convocatoria ordinaria será necesario obtener una nota de 5 sobre 10 en cada una de las partes que componen la evaluación de la asignatura. Aquellos alumnos que no lleguen a la nota de 5 sobre 10 en alguna de las parte de evaluación, tendrán que presentarse a esas partes suspensas en la convocatoria extraordinaria de la asignatura.

- **Convocatoria extraordinaria:** sólo será necesario presentarse a la convocatoria extraordinaria para aquellas partes suspensas (nota inferior a 5 sobre 10) en la convocatoria ordinaria. No se permitirá que los alumnos se presenten, en convocatoria extraordinaria, a aquellas partes ya aprobadas en la convocatoria ordinaria. Para aprobar la asignatura será necesario obtener una nota de 5 sobre 10 en cada una de las partes a las que el alumno tenga que presentarse en la convocatoria extraordinaria.

9. Consideraciones / Comentarios adicionales