Proyecto docente

Asignatura	Almacenamiento Escalable		
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**	1
Créditos ECTS	3		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	Aníbal Bregón y Fernando Díaz		
Datos de contacto (e-mail, teléfono…)	Escuela de Ingeniería Informática (Segovia) Campus María Zambrano Plaza de la Universidad, 1, 40005 Segovia Teléfono: 98342300 (ext. 5696) e-mails: anibal@infor.uva.es y fdiaz@infor.uva.es		
Horario de tutorías	Disponible en http://www.inf5g.uva.es/node/765		
Coordinador	Aníbal Bregón		
Departamento	Informática (ATC, CCIA, LSI)		
Web	Página de la asignatura en https://ubuvirtual.ubu.es/		
Descripción General	La asignatura 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, haciendo especial hincapié en las tecnologías no relacionales, o NoSQL.		

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 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.

En resumen, esta asignatura se divide en dos grande 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 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 asignaturas

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, en particular con Arquitecturas Big Data 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 …

- 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 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. Contenido / Programa de la asignatura

Bloque 1: "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, proporcionan la infraestructura básica de almacenamiento escalable en aplicaciones 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 conocer, más en detalle, el diseño y funcionamiento de los componentes internos de HDFS, introducir los procedimientos de administración, monitorización y mantenimiento de HDFS, revisar las nuevas funcionalidades introducidas en Hadoop 3 y conocer los diferentes formatos de ficheros existentes, centrándonos en los formatos adecuados para Big Data.

b. Objetivos de aprendizaje

- Comprender el diseño interno de los componentes del sistema HDFS
- Conocer las nuevas funcionalidades del sistema HDFS en Hadoop 3 en relación con la gestión de datos
- Realizar tareas básicas de configuración, administración y monitorización del sistema HDFS
- Identificar qué formatos de ficheros almacenados en HDFS son adecuados para las aplicaciones Big Data en un *framework* basado en Hadoop

c. Contenidos y materiales de aprendizaje

- Detalles internos de la arquitectura de HDFS
- Nuevas funcionalidades de la gestión de datos en HDFS
- Configuración, administración y monitorización de HDFS
- Formatos de ficheros, compresión y serialización de datos

d. Métodos docentes

- Material de lectura básica
- Material complementario
- Cuestionarios de evaluación
- Actividades individuales y/o colaborativas (Tareas)
- Material audiovisual asíncrono
- 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:
    - Asíncronos: foros, emails
    - Sala videoconferencia

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
0,5	Semana 1
0,5	Semana 2

4. Contenido / Programa de la asignatura

Bloque 2: "Fundamentos de NoSQL"

Carga de trabajo en créditos ECTS: 0,5

a. Contextualización y justificación

Existen dos grandes formas de escalar bases de datos relaciones: el escalado vertical y el escalado horizontal. El escalado vertical se centra en aumentar las capacidades de escalabilidad añadiendo recursos hardware de más capacidad, lo cual acaba llegando a un punto a partir del cual no es posible (o no es económicamente viable) escalar. En cambio, el escalado horizontal se centra en aumentar el número de máquinas y distribuir los datos entre éstas. Sin embargo, esto genera problemas adicionales relacionados con el particionado de los datos. Veremos como el uso de sharding permite tener escalabilidad horizontal y mejora el rendimiento, pero aún así puede tener problemas debido a su complejidad. Además, si el sharding se utiliza con replicación basada en la localización, se mejora la escalabilidad y la disponibilidad, pero aparecen nuevos problemas que pueden comprometer la consistencia de los datos. Estas ideas darán pie al teorema CAP, que será el teorema central de este bloque.

El teorema CAP, nombre que viene de las siglas de sus 3 propiedades (Consistency, Availability y Partition tollerance), establece que sólo podemos tener 2 de esas 3 propiedades al mismo tiempo. Esto tuvo como resultado bases de datos con propiedades ACID relajadas, conocidas como propiedades BASE (Basically Available, Soft-state, Eventually consistent). BASE es una alternativa flexible a ACID para aquellos almacenes de datos que no requieren una adherencia estricta a las transacciones. Todas estas ideas dieron lugar a un tipo de bases de datos que siguen las propiedades BASE: las bases de datos NoSQL.

NoSQL define un conjunto de bases de datos no relacionales, que siguen las propiedades BASE, que escalan horizontalmente y que son baratas y fáciles de implementar. Existen 4 grupos principales de bases de datos NoSQL: clave-valor, orientadas a columnas, orientadas a documentos y basadas en grafos. En este bloque se presentarán las ideas fundamentales de estos 4 grupos de bases de datos NoSQL.

b. Objetivos de aprendizaje

- Conocer el problema de la escalabilidad y distribución de bases de datos
- Conocer el teorema CAP y las propiedades BASE
- Conocer las principales características que definen las bases de datos NoSQL
- Conocer los principales tipos de bases de datos NoSQL

c. Contenidos y materiales de aprendizaje

- Escalando bases de datos
- El teorema CAP
- Propiedades BASE
- Bases de datos NoSQL
- Taxonomía de bases de datos NoSQL (Clave-valor, Orientadas a columnas, Orientadas a documentos, Basadas en grafos)
- NewSQL

d. Métodos docentes

- Material de lectura básica
- Material complementario
- Cuestionarios de evaluación
- Material audiovisual asíncrono
- Tutorías online síncronas
- Seminarios

e. Bibliografía básica

- TIWARI, SHASHANK, “Professional NoSQL”, Wiley/Wrox. 2011.
- STRAUCH, CHRISTOF, “NoSQL Databases”. Stuttgart Media University. 2012.
- VAISH, G., “Getting Started with NoSQL”. Packt Publishing. 2013.
- BREWER, E. (2000). Towards Robust Distributed System. Symposium on Principles of Distributed Computing (PODC).

f. Bibliografía complementaria

- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. 2006. Bigtable: a distributed storage system for structured data. In Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7 (OSDI '06), Vol. 7. USENIX Association, Berkeley, CA, USA, 15-15
- Eric A. Brewer. CAP twelve years later: How the "rules" have changed, February 2012
- S. Gilbert and N. Lynch. Perspectives on the CAP Theorem, February 2012
- Dan Pritchett. 2008. BASE: An Acid Alternative. Queue 6, 3 (May 2008), 48-55. DOI=10.1145/1394127. 1394128

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
    - Asíncronos: foros, emails
    - Sala videoconferencia

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
0,5	Semana 2

4. Contenido / Programa de la asignatura

Bloque 3: "HBase"

Carga de trabajo en créditos ECTS: 0,5

a. Contextualización y justificación

HBase es una base de datos distribuida, de código abierto y orientada a columnas. Está basada en BigTable de Google y una de sus grandes ventajas es que forma parte del proyecto Hadoop de la Fundación de Software Apache y se ejecuta sobre HDFS, aprovechando las características de este último. El modelo de datos de HBase organiza los datos en filas (rows) que se identifican de manera única mediante un rowkey. Cada fila puede tener una o varias familias de columnas, las cuales están formadas por columnas. Las columnas vienen definidas mediante pares clave-valor, más un timestamp, que marca el instante en el que se introdujeron los datos (formato posix). HBase tiene tres componentes principales un maestro activo (HMaster), una serie de servidores de regiones (HRegionServers) y Zookeeper. El HMaster se encarga principalmente de asignar regiones a los servidores de regiones, monitorizar los servidores de regiones y la administración de los cambios en el esquema (como creación de tablas y familias de columnas). Los HRegionServers se utilizan para realizar la lectura y escritura de los datos. Por último, Zookeeper es el encargado de coordinar de manera distribuida para mantener el estado de los servidores dentro del clúster. Este tercer bloque de la asignatura se centrará en explicar todos estos conceptos de HBase, así como su uso práctico con ejemplos sencillos y una versión simplificada de un ejemplo real.

b. Objetivos de aprendizaje

- Conocer las características básicas de HBase, así como su modelo de datos.
- Conocer la arquitectura de HBase y cómo se realizan las operaciones principales.
- Comenzar a trabajar con HBase y empezar a familiarizarse con el Shell de HBase mediante ejemplos sencillos.

c. Contenidos y materiales de aprendizaje

- HBase vs HDFS
- Modelo de datos
- Modelo físico
- Arquitectura de HBase
- HBase – ejercicios prácticos

d. Métodos docentes

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

e. Bibliografía básica

- HBase: The Definitive Guide, por Lars George, O'Reilly Media. Agosto de 2011

f. Bibliografía complementaria

- HBase in Action, por Nicholas Dimiduk and Amandeep Khurana. Noviembre de 2012. ISBN 978161729052

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
    - Asíncronos: foros, emails
    - Sala videoconferencia

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
0,5	Semana 3

4. Contenido / Programa de la asignatura

Bloque 4: "Cassandra"

Carga de trabajo en créditos ECTS: 1,0

a. Contextualización y justificación

En este cuarto bloque de la asignatura se presenta en detalle una de las bases de datos NoSQL más utilizadas y más exitosas del mercado, Cassandra. Cassandra es una base de datos orientada a columnas creada por Facebook combinando las ideas de Google Bigtable y Amazon Dynamo. Cassandra ofrece una muy alta disponibilidad, capacidad de escalar de manera lineal y sin punto de fallo único. La capacidad de Cassandra puede ser incrementada tan sólo añadiendo nuevos nodos al clúster y además permite realizar escrituras de una manera muy rápida. Es compatible con Hadoop y, tanto el factor de replicación como el nivel de consistencia de los datos son completamente configurables (permitiendo tener niveles de consistencia desde débil hasta fuerte). Este bloque estará compuesto de cuatro partes, dos de ellas con contenidos teóricos para entender el funcionamiento y el modelo de datos en Cassandra, otra parte se centrará en la instalación de Cassandra y el último bloque del tema se centrará en los aspectos más prácticos del manejo y uso de Cassandra.

b. Objetivos de aprendizaje

- Conocer las características básicas de Cassandra, así como su modelo de datos.
- Conocer la arquitectura de Cassandra y cómo se realizan sus operaciones principales, como son el particionado, la replicación, la escritura y la definición del nivel de consistencia.
- Aprender cómo se realiza la instalación de Cassandra.
- Comenzar a trabajar con Cassandra y empezar a familiarizarse con el CQL mediante ejemplos sencillos.
- Aprender, desde un punto de vista práctico, a crear una base de datos en Cassandra, entender las distintas opciones para crear la primary key en Cassandra y aprender a crear índices.

c. Contenidos y materiales de aprendizaje

- Introducción a Cassandra
- Modelo de datos en Cassandra
- Cassandra vs HBase
- Arquitectura de Cassandra (particionado, replicación, snitches, escritura, lectura, etc...)
- Instalación y configuración
- Ejercicios prácticos

d. Métodos docentes

- Material de lectura básica
- Material complementario
- Actividades individuales y/o colaborativas (Tareas)
- Cuestionarios de evaluación
- Material audiovisual asíncrono
- Tutorías online síncronas

e. Bibliografía básica

- CARPENTER, J., HEWITT, E. “Cassandra: The Definitive Guide”, 2nd Edition, O'Reilly Media. 2016.
- Avinash Lakshman and Prashant Malik. 2010. Cassandra: a decentralized structured storage system. SIGOPS Oper. Syst. Rev. 44, 2 (April 2010), 35-40. DOI=10.1145/1773912.1773922

f. Bibliografía complementaria

g. Recursos necesarios

- Plataforma Moodle
- Herramientas de comunicación:
    - Asíncronos: foros, emails
    - Sala videoconferencia

h. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO (detallar orden semanas)
0,5	Semana 4
0,5	Semana 5

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

7. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
HDFS	1,0	Semanas 1 y 2
Fundamentos de NoSQL	0,5	Semana 2
HBase	0,5	Semana 3
Cassandra	1,0	Semanas 4 y 5

8. Evaluación

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

Criterios / Comentarios a la evaluación
• **Convocatoria ordinaria:** En la convocatoria ordinaria se realizarán cuestionarios y se pedirán proyectos y/o trabajos para cada uno de los bloques que componen la asignatura. Para aquellas partes más teóricas no susceptibles de ser evaluados mediante trabajos prácticos, se realizarán cuestionarios evaluables individuales online. Además de esto, también 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