Máster Executive en Business Analytics y Big Data

Edición 2014 / 2015

Asignatura: Bases de Datos No Convencionales
Módulo: Tecnologías de Big Data/Gestión de Datos
Coordinador: Elena García Barriocanal, elenagarcia@campusciff.net
Profesores: Jordi Conesa jordiconesa@campusciff.net
Elena García Barriocanal, elenagarcia@campusciff.net

OBJETIVOS

Cuando se almacenan grandes volúmenes de datos de forma distribuida los esquemas más trabajados hasta el momento, conocidos como esquemas relacionales, dejan de ser útiles por la dispersión de los datos, la ausencia de información estructurada, la dificultad en representar las complejas interrelaciones que existen entre los datos, el tiempo de respuesta requerido o una conjunción de estos factores. Es por este motivo que surgen nuevos paradigmas fuera de las convenciones actuales, a los que se conoce como bases de datos NoSQL.

Esta asignatura nos permite poner los pilares en el tratamiento de Big Data, ya que permite disponer de mecanismos para representar los datos no estructurados, grabarlos, consultarlos, modificarlos y eliminarlos de acuerdo a necesidades específicas de dominios donde se gestiona mucha información y se deben tomar decisiones en base a la misma.

El objetivo fundamental de este módulo es alcanzar una visión global de qué se entiende por una base de datos NoSQL, proporcionando al estudiante una visión de los múltiples conceptos, modelos y herramientas que se pueden agrupar bajo esta denominación.

Los objetivos concretos del módulo consisten en garantizar los siguientes resultados del aprendizaje:

1. Ser capaz de contextualizar los conceptos fundamentales de bases de datos NoSQL y ponerlos en valor.
2. Conocer los distintos modelos NoSQL: saber cómo se organizan los datos en cada uno de ellos y conocer sus principios básicos de diseño.
3. Ser capaz de escoger el mejor modelo de datos para cada problema concreto, o la combinación de ellos en un proyecto informático.
4. Conocer las características principales y el funcionamiento de algunos de los productos NoSQL más relevantes: Cassandra, MongoDB, Riak y Neo4j.

Las competencias que se adquieren en este módulo están estrechamente relacionadas con otras que el estudiante adquiere en los módulos de paralelización de datos, escalibilidad de bases de datos y análisis de redes sociales.

METODOLOGÍA

La asignatura se organiza en torno al concepto de sesión presencial de 5 horas de duración, habitualmente divididas en grupos de 2 horas o 2,5 horas.

En dichas sesiones se tratará en primer lugar los conceptos teóricos imprescindibles para la adquisición de las competencias y la realización de

casos prácticos.

Cada sesión presencial va a acompañada de un caso práctico guiado por el profesor siguiendo métodos activos, por lo que cada estudiante debe acudir a clase con su propio ordenador.

En caso de complejidad en la instalación de determinados productos, se le proporcionará al estudiante las máquinas virtuales oportunas CentOS sobre Virtualbox.

Para la completa adquisición de las competencias por parte del estudiante es fundamental el trabajo personal fuera de las sesiones, completando los casos prácticos según las directrices que se aportarán y realizando las tareas de evaluación continua, que además de ser una herramienta de evaluación constituyen una pieza fundamental para el aprendizaje.

PROGRAMA

Sesión 1: Introducción. Modelo documental.

Introducción
Introducción a NoSQL y persistencia políglota. Introducción a los modelos NoSQL.
Conceptos: Introdución NoSQL, persistencia políglota, teorema CAP, introducción a los modelos de agregación, introducción al modelo de grafo.
Actividades: Presentación

Modelos de agregación: Bases de datos documentales
Conceptos: Introducción al modelo documental, consideraciones de diseño en el modelo documental, modelo de datos de MongoDB, Shell de MongoDB, Operaciones CRUD en MongoDB.
Actividades: Presentación, casos prácticos con MongoDB y trabajo personal práctico sobre las actividades presenciales.

Sesión 2: Modelo clave-valor y modelo. Modelo en grafo.

Modelos de agregación: Bases de datos clave-valor
Conceptos: Introducción al modelo clave-valor, consideraciones de diseño en el modelo clave-valor y introducción a Riak, Modelo de datos de Riak, operaciones CRUD en Riak, el anillo de Riak.
Actividades: Presentación, casos prácticos con Riak y trabajo personal práctico sobre las actividades presenciales.

Modelo en grafo
Conceptos: introducción a los grafos, grafos etiquetados con propiedades,

modelo de grafos, consideraciones de diseño en los modelos en grafo e introducción a Neo4j. Modelo de datos en Neo4j, operaciones CRUD en Neo4j y lenguaje Chypher y ejercicios.
Actividades: Presentación, ejercicios introductorios a los modelos en grafo, ejercicios de consultas en Chyper y trabajo personal práctico sobre las actividades presenciales.

Sesión 3: Modelo por columnas. Caso práctico final.

Modelos de agregación: Bases de datos por columnas
Conceptos: Introducción al modelo de columnas, consideraciones de diseño en el modelo de columnas, introducción a Cassandra. Operaciones CRUD en Cassandra
A*ctividades*: Presentación. Casos prácticos con Cassandra y trabajo personal práctico sobre las actividades presenciales. (I)

Caso práctico final
Actividades: Presentación del caso práctico final. Formación de grupos y trabajo en grupo sobre el caso práctico final.

MATERIALES

Recursos de uso general

- Popescu. NoSQL and Polyglot Persistence. (http://bit.ly/1ggTT4k).
- P.J. Sadalage & M. Fowler (2013). NoSQL Distilled. A brief Guide to the Emerging World of Polyglot Persistence, Pearson Education. (http://bit.ly/1koKhBZ).
- M. Stonebraker & U. Çetintemel (2005). "One Size fits all: An Idea whose time has come and gone", Proceedings of the International Conference on Data Engineering, pp 2-11. (http://cs.brown.edu/people/ugur/fits_all.pdf).
- R. Catell (2010). "Scalable SQL and NoSQL Data Stores". SIGMOD Record 39(4), pp 12-27. (http://dl.acm.org/citation.cfm?id=1978919).
- Joe Celko's (2013). Complete Guide to NoSQL. Elsevier. (http://www.sciencedirect.com/science/book/9780124071926)
- E. Redmond, J Wilson (2012). Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, The Pragmatic Bookshelf. (http://pragprog.com/book/rwdata/seven-databases-in-seven-weeks)
- I. Robinson, J. Webber & E. Eifren (2013). *Graph Databases*. O'Reilly. (http://graphdatabases.com/)
- M. Gyssens& J. Paredaens (1990). "A Graph Oriented Object Database Model", ACM SIGMOD International Conference.

(http://bit.ly/JL2QWc).

Recursos para actividades prácticas:

- Documentación oficial de Riak: http://docs.basho.com/riak/2.0.0pre11
- Little Riak Book (http://littleriakbook.com)
- K. Chodorow (2011). 50 Tips for MongoDB Developers O'Really Media.
- MongoDB Corp (2014) RDBMS to MongoDB Migration Guide. http://www.mongodb.com/lp/white-paper/migration-rdbms-nosql-mongodb
- DATASATAX CORP. (2013) Introduction to Apache Cassandra http://www.odbms.org/wp-content/uploads/2014/06/WP-IntroToCassandra.pdf
- The Neo4j Team. The Neo4j Manual. Neotechnology (http://docs.neo4j.org/chunked/2.0.3/)
- The Neo4j Community. Learn Cypher - the Neo4j query language (http://www.neo4j.org/learn/cypher)
- The Neo4j Team (Last consulted: May 2014). Neo4j Cypher Refcard 2.0 (http://docs.neo4j.org/refcard/2.0/)
-

Otros recursos:

- DATASTAX CORPORATION (2014) Implementing a NoSQL Strategy http://www.odbms.org/2014/06/implementing-nosql-strategy/
- Fowler, M., Sadalage, P (2012) The future is: Polyglot Persistence. http://martinfowler.com/articles/nosql-intro-original.pdf
- Fowler, M. (2012) Introduction to NoSQL http://www.youtube.com/watch?v=qI_g07C_Q5I
- Página Oficial de cursos/tutoriales sobre MongoDB: https://university.mongodb.com/

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	Nivel alto	Nivel medio	Nivel bajo
O1: Conceptos fundamentales	Identificar los casos en los que es conveniente utilizar bases de datos NoSQL	Conocer de las ventajas e inconvenientes de usar bases de datos NoSQL.	Saber qué es NoSQL y los distintos modelos que lo componen.

O2: Conocimiento modelos	Saber adaptar el diseño conceptual en función del modelo de base de datos utilizado	Conocer las características de cada uno de los modelos NoSQL.	Conocer las principales diferencias entre los distintos modelos NoSQL.
O3: Selección modelo	Saber identificar la necesidad de varios modelos en un para un determinado dominio de problema y saber integrar esos modelos	Dado un dominio concreto, saber escoger la combinación de modelos de datos más adecuada para representar sus datos.	Identificar el modelo de datos más adecuado para cada caso concreto.
O4: Funcionamiento productos	Ser capaz de crear una base de datos, poblarla de datos y realizar consultas complejas en cada producto.	Ser capaz de realizar consultas básicas en los distintos productos.	Tener una idea de cómo interaccionar con cada uno de los productos.

Modelo de evaluación

La siguiente tabla detalla los pesos de cada una de las actividades de evaluación:

Elemento	*Peso*
Entrega práctica individual: modelo clave-valor	10%
Entrega práctica individual: modelo documental	10%
Entrega práctica individual: modelo columnas	10%
Entrega práctica individual: modelo grafo	10%
Trabajo en grupo: caso práctico final	60%

El caso práctico final se desarrollará en grupos de 2 o 3 personas y consistirá en el diseño e implementación de una base de datos de acuerdo al concepto de persistencia políglota, en el que intervienen bases de datos de diferentes modelos para adecuarse de la mejor manera posible a las necesidades del dominio. Dicho caso se trabajará en su mayor parte fuera de las sesiones presenciales.

PROFESORADO

Dr. Jordi Conesa Caralt es profesor agregado en los estudios de Informática, Multimedia y Telecomunicaciones de la Universitat Oberta de Catalunya, donde ha ejercido de director del máster de Business Intelligenge durante los

últimos años (desde 2008 hasta 2013). Su carrera investigadora se ha enfocado mayoritariamente al modelado conceptual, las ontologías y la semántica. Actualmente, está derivando al ámbito de learning analitics, para lo cual utiliza bases de datos NoSQL y técnicas de Big Data. Docentemente es responsable de asignaturas en las temáticas de bases de datos (incluyendo no sólo bases de datos relacionales sino también heterogéneas y escalables – NoSQL), semántica y gestión del conocimiento y modelado conceptual.

Dra. Elena García Barriocanal es profesora titular en el Departamento de Ciencias de la Computación de la Universidad de Alcalá y Directora del grupo de investigación Information Engineering, donde se desarrollan proyectos de investigación relacionados con grandes volúmenes de datos, como "CIEN LPS-BIGGER: Línea de Productos Software para Big Data a partir de Aplicaciones Innovadoras en Entornos Reales" o el proyecto Europeo de infraestructuras científicas para agricultura agINFRA (http://www.aginfra.eu/). Elena ha desarrollado su carrera investigadora en el campo de la semántica y las ontologías y actualmente se centra en bases de datos heterogéneas y escalables.