Máster en Business Intelligence & Data Science

Edición 2017/2018

Universidad de Alcalá

Asignatura: ESCALABILIDAD EN SOLUCIONES DE DATOS

PROFESOR: Javier Rodríguez Rodríguez

OBJETIVOS

El objetivo general de la asignatura es el de adquirir los conocimientos y habilidades necesarias para comprender la problemática de escalar almacenes de datos NoSQL. Nos centraremos para ello en MongoDB y Cassandra como ejemplos paradigmáticos de cómo se persigue escalabilidad con diferentes estilos arquitectónicos.

Los objetivos concretos de la asignatura son los siguientes resultados del aprendizaje:

1. Comprensión de los conceptos básicos que se asocian al rendimiento y escalabilidad de almacenes de datos.
2. Interiorizar el impacto de los índices en el rendimiento de consultas.
3. Comprender la repercusión de la replicación en la disponibilidad.
4. Comprensión del impacto del particionamiento y redundancia de datos en la escalabilidad.
5. Ser capaz de diseñar índices para colecciones de documentos en MongoDB y estudiar su impacto en el rendimiento.
6. Comprender y ensayar las particularidades de la arquitectura de replicación y *sharding* de MongoDB.
7. Comprender y ensayar la arquitectura de clúster de Cassandra.
8. Impacto del modelado de datos en la escalabilidad de Cassandra.

METODOLOGÍA

Se usará una combinación de descripción teórica de los conceptos objetivo complementada con ejercicios prácticos en clase que permitan a los alumnos experimentar de primera mano los conocimientos adquiridos. Se realizará al menos una videoconferencia para seguimiento y refuerzo en la que plantear dudas y consultas.

Los ejercicios se realizarán en instalaciones virtualizadas con entornos preconfigurados de MongoDB y Cassandra que se suministrarán con antelación.

La asignatura no pretende formar en habilidades prácticas propias de un administrador de sistemas NoSQL sino que está diseñada para comprender la complejidad inherente a instalaciones complejas de bases de datos NoSQL y mostrar las herramientas necesarias para garantizar verdadera escalabilidad.

PROGRAMA

Bloque 0:
Contenido: Conceptos generales de escalabilidad en bases de datos no relacionales.

Estrategias para escalabilidad.

Bloque 1:
Contenido: Rendimiento y escalabilidad con MongoDB.
Actividades: Diseño de índices y verificación de impacto. Implantación redundante de MongoDB. Buenas prácticas para MongoDB.
Materiales: VM con MongoDB y Ops Manager

Bloque 2:
Contenido: Escalabilidad con Cassandra: comprensión del cluster.
Actividades: Uso de CCM para simular un clúster de Cassandra en entornos monopuesto.
Materiales: VM con Cassandra y CCM

Bloque 3:
Contenido: Modelado de datos escalable en Cassandra. Integración de Spark con Cassandra.
Actividades: Modelado en CQL y analítica básica mediante Spark.
Materiales: VM con Cassandra y Spark

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	*Nivel alto*	*Nivel medio*	*Nivel bajo*
O1 – Escalar MongoDB	Diseñar y dimensionar un sharded cluster	Optimizar consultas mediante planes de ejecución	Indexar consultas
O2 – Escalar Cassandra	Diseñar modelos de datos escalables	Distribuir adecuadamente los datos en el clúster	Entender los mecanismos de replicación y consistencia

Modelo de evaluación

Elemento	*Peso*
Práctica	100%

PROFESORADO

Javier Rodríguez Rodríguez es Director de Soluciones en OpenSistemas y responsable de proyectos bajo arquitecturas innovadoras para el tratamiento, visualización y exposición de datos en múltiples sectores. Acumula más de una década de experiencia en arquitecturas empresariales y actualmente está especializado en tecnologías NoSQL, streaming y API Management. Licenciado en CC. Físicas por la UCM. Postgrado en Data Science por U-Tad (UCJC). Participa en distintos programas de master y cursos de innovación digital.

https://www.linkedin.com/in/mezco/

Mail: javier.rodriguezro@uah.es