

## Información general del curso

Nombre de la asignatura	Bases de Datos Avanzadas - Big Data
Créditos académicos	3
Horas presenciales por semana	3 horas
Horas de trabajo no presencial por semana	6 horas

## BASES DE DATOS AVANZADAS - BIG DATA

Las Bases de Datos Relacionales (BD) se consideran una tecnología genérica, y su éxito se evidencia en la utilización de las mismas. Prácticamente todas las actividades económicas requieren el procesamiento de información. Sin embargo, el auge en el uso de Internet desde el principio de este siglo, y otros factores como el uso de dispositivos que generan datos todo el tiempo, ha creado desafíos en tres dimensiones: volumen, velocidad y variedad (denominados las tres Vs), aunque a menudo se hablan de más.

El *volumen* es un factor clave ya que en el mundo se generan 2.5 quintillones de bytes de datos diariamente. Debido al incremento en el volumen de datos que se están generando (sobre todo de forma automatizada), el 90% de los datos que existen en el mundo han sido creados en los últimos 2 años. La *velocidad* se debe a que es necesario procesar flujos de datos (streams) sin demora y de forma confiable (por ejemplo, para detectar fraude en las transacciones de las tarjetas de crédito). La *variedad* se refiere al hecho que los datos generados hoy-en-día son cada vez más heterogéneos y complejos: comprenden texto libre, datos semi-estructurados, BD relacionales e incluso flujos de video.

Las herramientas convencionales, aunque siguen siendo vigentes para la mayoría de los casos de uso, se ven limitadas ante semejantes desafíos.

## Objetivos de aprendizaje

### Objetivo General

Aprender de temas avanzados de BDs relacionales, en especial las distribuidas, y familiarizarse con tecnologías NoSQL, que representan un nuevo paradigma en el almacenamiento y la recuperación de la información.

### Objetivos específicos

Al finalizar la asignatura el estudiante debe estar en capacidad de:

- Identificar, proponer y emplear factores involucrados relacionados con las arquitecturas de las BDs distribuidas, la fragmentación de datos, la optimización de consultas, y la integración de datos.
- Apreciar las ventajas y limitaciones tanto como de las BD relacionales como las tecnologías NoSQL.
- Desarrollar una sensibilidad de la diversidad de aplicaciones de procesamiento de datos que han surgido en los últimos años, y sus requerimientos no funcionales, como la escalabilidad y la disponibilidad.
- Desarrollar nociones para determinar en cuáles casos es más apropiada la tecnología tradicional, relacional, o la tecnología NoSQL.
- Realizar prácticas en el laboratorio usando herramientas disponibles en el mercado y de código abierto.

### Evaluación

Para estudiantes de la Especialización en Desarrollo de Bases de Datos, quienes cursan 11 sesiones:

- 18% Quizzes
- 54% Tres Laboratorios (18% cada uno)
- 18% Presentación Big Data/NoSQL
- 10% Nota de Clase

Para estudiantes de la Maestría en Ingeniería y Analítica de Datos, quienes cursan 16 sesiones:

- 10% Quizzes
- 30% Tres Laboratorios (10% cada uno)
- 10% Presentación Big Data/NoSQL
- 25% Proyecto Final NoSQL
- 20% Examen Final
- 5% Nota de Clase

### Dinámica de clase/ Metodología

- Clases magistrales
- Desarrollo de ejercicios teóricos
- Prácticas en laboratorio
- Lecturas asignadas por el profesor
- Proyecto sobre un caso práctico

## Cronograma del curso

### Sesión 1: Introducción

- Introducción al curso
- Bases de Datos: Pasado, presente y futuro
- Introducción al álgebra relacional

### Sesión 2: Arquitecturas de Bases de Datos Distribuidas y Fragmentación de Datos

- Principios de los sistemas Distribuidos
- Arquitecturas de Bases de Datos Distribuidas
  - Shared memory, shared disc, shared nothing
- Fragmentación y localización de datos
  - Horizontal primaria
  - Horizontal derivada
  - Vertical
  - Distribución de datos

### Sesión 3: Optimización de Consultas: El Caso Centralizado

- El caso centralizado
  - Análisis de sintaxis
  - Traducción a árbol de operadores lógicos
  - Optimización lógica
  - Optimización física
- Debate: Arquitecturas de Bases de Datos Distribuidas

### Sesión 4: Optimización de Consultas: El Caso Distribuido

- El caso distribuido
  - Fragmentación y Alocación
  - Localización de datos

### Sesión 5: Laboratorio de BD distribuidas (dirigido por Rafael Hernández)

### Sesión 6: Introducción a Big Data

- ¿El fin de la talla única?
- Las Tres Vs (Volumen, Variedad, Velocidad)
- Las redes de sensores y el Internet de las Cosas
- Modelos NoSQL
  - Almacenes llave-valor
  - Almacenes de documentos
  - Almacenes de columnas
  - BD de grafos
- Procesamiento Masivo de Datos
  - Los Data Appliances
  - MapReduce

**Sesión 7: Laboratorio Hadoop (en Python)**

**Sesión 8: Modelando Datos con Grafos**

- Neo4j
- Tutorial de consultas con Cypher

**Sesión 9: Laboratorio Neo4j**

**Sesión 10: Teorema CAP, BASE**

- Transacciones
  - Propiedades ACID
  - Tipos de transacciones
  - Protocolo Two Phase Commit
  - Teorema CAP
  - BASE: Una alternativa a ACID
- Indexación de datos en sistemas escalables
  - Distributed Hash Tables
  - Un caso del mundo real: Amazon Dynamo

**Sesión 11: Charlas: Big Data en el mundo real**

**Sesión 12: Tutoría Proyectos NoSQL**

**Sesión 13: Tutoría Proyectos NoSQL**

**Sesión 14: Tutoría Proyectos NoSQL**

**Sesión 15: Presentación Proyectos NoSQL**

**Sesión 16: Examen final**

## Referencias bibliográficas

1. Tamer Ozsu, M., & Valduriez, P. (2011). *Principles of distributed database systems*. Springer.
2. Ullman, J. D., Garcia-Molina, H., & Widom, J. (2001). *Database systems: the complete book*. Upper Saddle River: Prentice Hall.
3. Kossmann, D. (2000). The state of the art in distributed query processing. *ACM Computing Surveys (CSUR)*, 32(4), 422-469.
4. DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007, October). Dynamo: Amazon's highly available key-value store. In *SOSP* (Vol. 7, pp. 205-220).
5. Ghemawat, S., Gobiuff, H., & Leung, S. T. (2003, October). The Google File System. In *ACM SIGOPS Operating Systems Review* (Vol. 37, No. 5, pp. 29-43). ACM.
6. Pritchett, D. (2008). Base: An acid alternative. *ACM Queue*, 6(3), 48-55.
7. Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Communications of the ACM*, 53(1), 72-77.
8. Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., ... & Widom, J. (2004). Stream: The Stanford data stream management system. *Book chapter*.