Facultad de Ingeniería y Ciencias
Escuela de Informática y Telecomunicaciones

Descriptor de asignatura

Big data

1. Identificación de la asignatura:

Nombre de la Asignatura: Big data	
Códigos: CDI-2003	Créditos: 5
Duración: Semestral	Ubicación en el plan de estudios: Semestre 5
Requisitos: CIT-2307 Bases de datos	
Sesiones cátedras semanales: 2 cátedras	
Sesiones de Ayudantía: 1	

2. Descripción de la asignatura:

Este curso explora los conceptos fundamentales para optimizar la gestión y recuperación de grandes volúmenes de datos. Se abordan métodos para distribuir el almacenamiento y mejorar el acceso a la información en sistemas de alta demanda, así como estrategias para mantener datos en memoria para un procesamiento más rápido. Además, se analizan enfoques para organizar datos de manera eficiente, comparando estructuras enfocadas en el análisis histórico y en la flexibilidad de grandes repositorios de datos.

3. Resultados de Aprendizaje:

1. Identifica la jerarquía de memoria y los tiempos de acceso, para optimizar la recuperación de información mediante índices y estructuras de almacenamiento eficientes.
2. Evalúa y selecciona tecnologías de almacenamiento NoSQL y de última generación, en función de los requisitos de escalabilidad y consistencia del sistema.
3. Aplica algoritmos de similitud, para mejorar la búsqueda y análisis de datos en grandes volúmenes de información.
4. Diseña arquitecturas de almacenamiento distribuido, incluyendo sistemas en memoria y de archivos, para mejorar el procesamiento y la disponibilidad de datos.
5. Diferencia entre Data Warehouses y Data Lakes, comprendiendo sus propiedades, procesos de transformación (ETL/ELT) y el uso de tecnologías habilitantes.
6. Participa en equipos de trabajo, planificando, coordinando y ejecutando tareas con liderazgo y responsabilidad, comunicándose efectivamente y elaborando informes técnicos que reflejen procedimientos, resultados y análisis del trabajo realizado.

4. Unidades Temáticas:

Unidad 1: Fundamentos de recuperación y almacenamiento de información
- Jerarquía de memoria
- Tiempos de acceso
- Índices

Unidad 2: Almacenamiento distribuido y tecnologías NoSQL
- Arquitecturas de almacenamiento distribuido (en memoria, DFS)
- NoSQL y tecnologías de última generación (VoltDB, Cockroach, Spanner)

Unidad 3: Algoritmos de similitud y procesamiento de datos
- Algoritmos de similitud (distancia coseno, embeddings, Jaccard, etc.)

Unidad 4: Data warehouse y data lake en ecosistemas cloud
- Data warehouse vs. data lake: conceptos, propiedades
- ETL/ELT
- Cloud storage y tecnologías habilitantes (Delta Lake, Hudi, S3)

5. Descripción general del método de enseñanza

La metodología contempla dos clases semanales donde la cátedra es complementada por la presentación y discusión de artículos científicos o textos de actualidad relacionados con los temas expuestos en la cátedra. Esto busca desarrollar la curiosidad de el/la estudiante con respecto a las temáticas del curso, permitiendo a su vez desarrollar habilidades relacionadas con la búsqueda de información y el autoaprendizaje continuo. Además, se busca que se ponga en práctica los conocimientos aprendidos por medio de actividades.

6. Descripción general de la modalidad de evaluación

Se contempla la realización de trabajos prácticos, dos pruebas solemnes y un examen. La nota final (NF) del curso se calculará a partir de una nota de presentación (NP) y la nota del examen (NE). Asimismo, para el cálculo de la NP participan las notas de las pruebas solemnes (S1 y S2) y la nota de actividades (NA) que involucra el trabajo práctico de al menos 3 tareas.

Para aprobar el curso, se debe haber rendido todas las evaluaciones asociadas a la nota de actividades y tener una nota (NA) mayor igual a 4,0. En caso contrario, reprueba con la calificación de la nota de actividades.

Según la regla general, para aprobar el curso debe tenerse que NF ≥ 4,0 y para presentarse a Examen NP ≥ 3,5. La inasistencia a una prueba solemne implica reemplazo de su nota con la NE.