IIC2440 — Procesamiento de Datos Masivos

Programa de Curso

8 de Marzo de 2023

Página de Curso

Importante: la página del curso está en `https://github.com/IIC2440/Syllabus-2023-1`. Ahí se puede acceder a la información de contacto del profesor y los ayudantes.

Descripción

Los sistemas de bases de datos forman parte del núcleo del desarrollo de aplicaciones comerciales modernas, y son indispensables para cualquier aplicación que requiera almacenar información. Sin embargo, en la actualidad nos vemos enfrentados a cantidades de datos que las técnicas tradicionales no pueden solucionar, y es ahí donde surge la necesidad de tener herramientas especializadas para el procesamiento de datos masivos. El propósito de este curso es introducir al alumno en las técnicas y modelos de datos que han surgido este último tiempo para hacer frente al problema de manejar grandes volúmenes de datos.

Objetivo General

Durante el curso, el alumno aprenderá diversas técnicas utilizadas en la actualidad para manejar grandes cantidades de datos. Estas técnicas serán abordadas desde distintas perspectivas.

1. El alumno será capaz de comprender las técnicas de indexación que hacen posible que los sistemas de bases de datos puedan extraer información de manera eficiente.
2. El alumno entenderá los fundamentos de los paradigmas y modelos de datos más utilizados para procesar grandes cantidades de información en la actualidad, como por ejemplo *Data Warehousing*, *Map-Reduce* y *Graph Databases*.
3. El alumno aprenderá técnicas del área de Minería de Datos orientadas al procesamiento de grandes cantidades de información.
4. El alumno aprenderá a utilizar herramientas modernas de manejo de datos, como por ejemplo Apache Spark y herramientas en la nube, que permiten trabajar con datos no estructurados y que están almacenados de forma distribuida.

Metodología

El curso se reúne una vez a la semana, de 3:30 a 6:20. La clase se divide en distintos bloques. Normalmente los bloques tienen este orden, pero puede variar de semana a semana:

- Exposición de resultados, algoritmos o análisis.
- Clase presencial.
- Trabajo personal, que puede ser evaluado o no (ver calendario).

Ayudantías. Por el momento no se prevee realizar ayudantías, pero podríamos calendarizarlas durante el semestre, siempre en el horario del Viernes al módulo 5.

Calendario Semanal. Todos los lunes por la mañana se avisará la estructura de la clase, y se entregarán links al material correspondiente.

Evaluación

El curso se evalúa a través de:

- Cuatro trabajos en clase evaluados.
- Dos tareas grupales.
- Una evaluación de mitad de semestre.
- Un examen final.

La nota del curso se calcula de la siguiente forma. Si C es el promedio de las clases evaluadas, T es el promedio de las tareas grupales, M es la nota de la evaluación de mitad de semestre y E es la nota del examen, entonces el promedio final del curso es $0,35T + 0,35C + 0,2E + 0,1M$.

Las fechas de las evaluaciones serán anunciadas oportunamente en la página web del curso.

Contenidos

1. **Tecnología de bases de datos relacionales (repaso)**
   a) Lenguaje de consultas
   b) Índices
   c) Algoritmos
2. **Modelos de Big Data**
   a) Data Warehousing
   b) Técnicas en la Nube
3. **Procesamiento de datos distribuidos**
   a) Algoritmos en entornos distribuidos
   b) El paradigma Map-Reduce
   c) Apache Spark
4. **Minería de Grandes Volúmenes de Datos**
   a) Modelos de canasta y reglas de asociación
   b) Heurísticas, algoritmo Apriori
   c) Minhash y Locally Sensitive Hashing
   d) Manejo de texto con shingling
5. **Manejo de streams**
   a) Filtros de Bloom
   b) Conteos de elementos en Streams
6. **Grafos y redes sociales**
   a) Herramientas orientadas a grafos
   b) Algoritmos de Centralidad
   c) Detección de Comunidades

Otros

El Departamento de Ciencias de la Computación adopta una política de tolerancia-cero frente a copias o plagios. Se sugiere revisar las políticas y penalidades que el departamento establece ante estas acciones. Recuerda también que la universidad y la escuela están suscritas a un código de honor, lo que nos incluye a profesor, ayudantes y alumnos.

Con respecto a copias y plagios, una reflexión. ¿cuál es la razón por la cuál tomas este curso, en una universidad que cuenta con un grupo de investigación en datos de nivel mundial? Los ejercicios de este curso

están pensados para que puedas ir aprendiendo a medida que te vamos evaluando. Siempre vamos a estar dispuestos a contestar todas tus dudas. ¡Aprovecha esta oportunidad para aprender!

El curso tiene dos canales de comunicación oficiales: Las clases y la página Web. Se asume que que toda la información que es entregada por ambos canales llega a todos los alumnos. Por lo mismo, se sugiere a los alumnos revisar la página Web constantemente.

Las clases del curso son obligatorias. En caso de faltar a una clase es responsabilidad del alumno ponerse al día con los contenidos. No se borran evaluaciones, pero se aceptará rendir evaluaciones de forma atrasada, cuando la situación lo amerite.

Bibliografía mínima

1. Rajaraman, Anand, and Jeffrey David Ullman. Mining of massive datasets. Cambridge University Press, 2011.
2. Aggarwal, Charu C. Data mining: the textbook. Springer, 2015.
3. Johannes Gehrke, Raghu Ramakrishnan. Database Management Systems. McGraw-Hill, 2003.
4. Jules Damji, Brooke Wenig, Tathagata Das, Denny Lee. Learning Spark. O'Reilly, 2022.