TICS575 - Técnicas para Big Data

Programa del Curso

Marzo de 2021

Profesor: Adrián Soto Suárez
adrian.soto@uai.cl

Clases: L1 - L2

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 el concepto de *Big Data*. 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.

- El alumno aprenderá diversos lenguajes de consultas utilizados por sistemas modernos de bases de datos. La idea es partir recordando conceptos del lenguaje de consultas SQL para luego entender el funcionamiento de los motores NoSQL.
- El alumno además entenderá a un nivel conceptual cómo funcionan los modelos de datos más utilizados en la actualidad, como por ejemplo orientado a documentos, bases de datos de grafos, modelos *key-value*, entre otros.
- 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.
- El alumno aprenderá a utilizar técnicas modernas de manejo de datos, como por ejemplo Apache Spark, que permiten trabajar con datos no estructurados y que están almacenados en *clusters*.
- El alumno conocerá el resultados y técnicas del estado del arte en torno al manejo de datos, como por ejemplo la cota AGM y los algoritmos de join *Worst-case Optimal*.

Finalmente, se espera que el alumno desarrolle la capacidad de entender bajo qué contexto usar cada una de las herramientas tratadas en este curso.

Contenidos

Técnicas tradicionales de manejo de datos

1. El modelo relacional y SQL.
2. Índices en el modelo relacional.
3. Limitaciones del modelo relacional.

El modelo de datos orientado a documentos

4. ¿Por qué bases de datos de documentos?

5. Manejo de documentos de tipo JSON.
6. Lenguajes de consulta de bases de datos de documentos.
7. Índices invertidos y TF-IDF.
8. Búsqueda por texto.

Bases de datos de Grafos

9. El modelo de *Property Graphs*.
10. Lenguajes de consultas de grafos.
11. Analítica de Grafos.
12. Pregel.

Manejo de datos no estructurados

13. *Map-Reduce*.
14. Hadoop.
15. Apache Spark.
16. Análisis de redes con GraphX.

Otros Tópicos

17. Herramientas en la nube y APIs.
18. Web Scraping.
19. Bases de datos *Key-Value*.
20. Bases de datos en memoria.
21. La cota AGM y algoritmos de join WCO.

Metodología

Durante las clases se enseñarán todos los contenidos teóricos sobre cómo funcionan los distintos motores y lenguajes de consultas. Asociado a cada tópico habrá una sesión práctica que hará que el alumno viva la experiencia de trabajar con estos sistemas. Finalmente, los contenidos más importantes tendrán asociado temas de proyecto que los alumnos deberán desarrollar en grupos.
Existe la posibilidad de no tener un segmento práctico en algunas semanas específicas.

Evaluación

La evaluación se realizará en base a:

- **Tareas**: durante el curso habrá entre 4 y 5 tareas asociadas al trabajo práctico durante las clases. De estas tareas se borrará la de peor nota. Así, el promedio de las tareas restantas corresponde a la nota de tareas (**NT**).
- **Presentaciones**: habrán dos rondas de presentaciones durante el periodo de clases. En estas presentaciones, los alumnos en grupos van a exponer sobre un tema que profundiza el contenido de las clases. Además, cada alumno debe dar *feedback* a sus compañeros de las presentaciones. Este *feedback* será evaluado como parte de la nota de cada una de las presentaciones. La nota de presentaciones (**NP**) será el promedio de las notas de ambas presentaciones.
- Proyecto final: habrá un proyecto final en que los alumnos profundizarán alguno de los contenidos vistos en clases. La dinámica será la misma que en las presentaciones del semestre pero se espera un trabajo con mayor nivel de profundidad. La nota del proyecto final será (**NPF**).

Para aprobar el curso **NT**, **NP** y **NPF** deben ser superiores a 4.0. Además, el promedio final será:

$$0{,}35 \cdot \mathrm{NT} + 0{,}4 \cdot \mathrm{NP} + 0{,}25 \cdot \mathrm{NPF}$$

Bibliografía

- Database Management Systems (Johannes Gehrke & Raghu Ramakrishnan).
- MongoDB: The Definitive Guide (Shannon Bradshaw, Eoin Brazil & Kristina Chodorow).
- Graph Databases (Ian Robinson, Jim Webber & Emil Eifrem).

- SPARK: The Definitive Guide (Bill Chambers & Matei Zaharia).
- Material de clases y *papers* que serán subidos a la página del curso.