FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

PROGRAMA DE CURSO

Código	**Nombre**			
CC5212	Procesamiento Masivo de Datos			
Nombre en Inglés				
Massive Data Processing				
SCT	**Unidades Docentes**	**Horas de Cátedra**	**Horas Docencia Auxiliar**	**Horas de Trabajo Personal**
6	10	3	1,5	5,5
Requisitos			Carácter del Curso	
CC3201 Bases de Datos			Electivo	

Resultados de Aprendizaje

El curso se basa en los métodos modernos de gestión, procesamiento, análisis y consulta de datos en gran escala, que se usan en grandes empresas como por ejemplo, Google, Facebook, Twitter, Amazon, etc. El curso tiene tres módulos: aspectos fundamentales de sistemas distribuidos, recuperación de información en gran escala, y bases de datos distribuidas.

Al finalizar el curso el alumno entenderá los fundamentos del procesamiento masivo de datos, enfocado a la forma en que esto es realizado por las grandes empresas.

Los objetivos específicos son:

- Dominar los fundamentos de gestión de datos en gran escala y los fundamentos del procesamiento distribuido de datos.
- Aprender los lenguajes para analizar datos masivos sobre múltiples máquinas.
- Aprender sobres las nuevas bases de datos distribuidas (NoSQL).
- Entender cómo funcionan (en alto nivel) los sitios web como “Facebook” y “Twitter” que manejen volúmenes masivos de datos.
- Entender cómo funcionan motores de búsqueda como “Google” que indexen billones de documentos.

Metodología Docente	**Evaluación General**
Clases expositivas de 90 minutos cada una y sesiones prácticas de 90 minutos. Se imparte el curso en inglés.	La evaluación contemplará. Examen (35%), Proyecto (20%) y Laboratorios (45%).

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Unidades Temáticas

Número	Nombre de la Unidad	Duración en Semanas
1	Introducción y Distribución	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introducción a los desafíos que plantea el procesamiento masivo de datos. • Introducción a los desafíos de escala en Twitter. • Diseño de sistemas distribuidos • Garantías de sistemas distribuidos (CAP). • Java *Remote Method Invocation.*	• Entender porque se necesitan sistemas distribuidos • Entender los desafíos y las metas en diseñar sistemas distribuidos • Aprender a programar un sistema distribuido básico (en Java)	[TS06]

Número	Nombre de la Unidad	Duración en Semanas
2	Procesamiento Distribuido de Datos	4

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Sistemas de archivos distribuidos: Google File System (GFS) • Infraestructuras modernas de procesamiento distribuido: MapReduce • Lenguajes y frameworks de fuente abierta: HDFS/Hadoop/PIG	• Entender los conceptos importantes con respecto a la gestión de datos a gran escala en empresas como Google • Aprender a programar y ejecutar tareas de MapReduce/Hadoop en un ambiente distribuido • Aprender a programar en PIG	[DG04] [W12]

Número	Nombre de la Unidad	Duración en Semanas
3	Recuperación de Información a Gran Escala	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introducción a la recuperación de información. • Crawling. • Índices invertidos y compresión. • Medidas de ranking de importancia: PageRank. • Revisión de medidas de ranking de relevancia: TF-IDF. • Una libraría de fuente abierta: Lucene.	• Entender cómo funciona un motor de búsqueda como "Google" y cómo se puede escalar. • Aprender cómo se puede hacer "ranking" de las páginas web. • Aprender programar un motor de búsqueda usando Lucene y aplicar ranking usando PageRank/TF-IDF.	[BP98] [BR11

Número	Nombre de la Unidad	Duración en Semanas
4	Conceptos de Manejo de Datos Distribuidos	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introducción a "*NoSQL*" • "*Key-Value stores*": Dynamo • "*Tabular stores:*" Bigtable • "*Graph stores:*" Neo4J • Crear un índice de Cassandra	• Entender las ventajas y las desventajas de usar un sistema NoSQL (en comparación con SQL) • Aprender sobre las garantías posibles con una base de datos distribuida • Aprender sobre las tipas de bases de datos NoSQL que existen y sus aplicaciones • Aprender a usar Cassandra	[TS06] [OV11] [SF12] [C08] [D07]

Número	Nombre de la Unidad	Duración en Semanas
5	Proyectos y Conclusión	2

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Aplicar las técnicas del curso a un problema de la elección del alumno. • Revisar los temas del curso • Perspectivas de futuro.	• Reforzar los temas del curso.	

Bibliografía

[BP98] S. Brin and L. Page The Anatomy of a Large-Scale Hypertextual Web Search Engine, Seventh International World-Wide Web Conference (WWW 1998).

[DG04] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters, Google Whitepaper, 2004. (Published in CACM, 2008).

[C08] Fay Chang et al. Bigtable: A Distributed Storage System for Structured Data. ACM Trans. Comput. Syst. 26(2), 2008.

[TS06] A. S. Tanenbaum, M. Van Steen. Distributed Systems: Principles and Paradigms (2nd Edition). Prentice Hall, 2006.

[D07] G. DeCandia et al. Dynamo: Amazon's Highly Available Key-value Store. Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles (SOSP), 2007.

[BR11] R. A. Baeza-Yates, B. A. Ribeiro-Neto: Modern Information Retrieval - the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England 2011.

[OV11] M. T. Özsu, P. Valduriez. Principles of Distributed Database Systems. Springer, 2011.

[W12] T. White. Hadoop: The Definitive Guide. O'Reilly, 2012.

[SF12] P. J. Sadalage, M. Fowler. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley Professional, 2012.

Vigencia desde:	Ultima Revisión: Primavera 2016
Elaborado por:	Aidan Hogan / Pablo Barceló