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

Al finalizar el curso el alumno entenderá los fundamentos del procesamiento masivo de datos, enfocado a la forma en que esto es realizado por grandes empresas como Google, Facebook, Twitter, Amazon, o grandes laboratorios de investigación como CERN.

The goal of the course will be to provide a solid foundation in the traditional design aspects relating to Distributed Computing and Distributed Databases, showing how they have influenced modern developments in cloud computing, including distributed data storage (e.g., NoSQL storage techniques) and data processing abstractions (e.g., MapReduce/Hadoop, Pregel/Giraph).

Practical assignments will reinforce course material and provide introductory hands-on programming experience developing distributed applications for handling massive data.

Metodología Docente	Evaluación General
Clases expositivas de 90 minutos cada una y sesiones prácticas de 90 minutos.	La evaluación contemplará Controles, Examen y Actividades Complementarias.

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Unidades Temáticas

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

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Course Introduction/Motivation for handling massive data. • Introduction to Distributed Systems. • Goals of a Distributed System • Fundamentals of parallel processing and distributed computing. • Traditional distributed computing architectures. • Modern distributed computing instantiations.	Understand the fundamental design principles of Distributed Systems (fault tolerance, performance, horizontal scale, economy). See why parallel processing and distributed storage are key to handling massive data. Learn about the different types of Distributed Systems: learn about traditional abstractions (P2P, client-server, grid, OSI layers), and see how they influence modern instantiations (Bittorrent, Internet, Web, cloud computing, Google, etc.).	[TS06, cap. 1,2] [KDF11, cap. 1,2]

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

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Traditional distributed communication methods (RPC/MPI) • Modern distributed processing abstractions, focusing on MapReduce with introduction to others (including Graph) • Course assignments using Java RMI for introductory distributed communication methods • Course assignments using Hadoop (MapReduce) to parallelize and solve a basic distributed computing task	Learn basics of distributed communication (RPC/MPI): gain hands-on experience of communicating between machines through introductory Java RMI assignment (intermediate Java skills a prerequisite). Learn modern distributed (cloud) computation abstractions, including MapReduce and Pregel (as used by Google et al.): gain hands-on experience with introductory programming assignments using Hadoop and Giraph (intermediate Java skills a prerequisite).	[TS06, cap. 2] [W12] [MABDHLC10] [OV11, cap. 18]

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

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introduction to Distributed Databases • Goals of distributed databases • Data-placement strategies: sharding (partitioning), replication, duplication • Fundamentals of distributed query processing and optimization • Programming assignment to build a basic distributed database using Java RMI and on-disk files	Learn the fundamentals of Distributed Databases, including the trade-offs between fault-tolerance, scalability, performance and economy. Learn the different ways in which data can be spread out and managed over multiple machines, so as to enable efficient and reliable querying over massive data. Understand the different types of guarantees a distributed database can make, and their formal limitations. Build a basic distributed database with hash-based sharding and replication: assignment will use Java RMI (intermediate Java skills a prerequisite).	[TS06, cap. 6,7] [OV11]

Número	Nombre de la Unidad	Duración en Semanas
4	Modelos de Almacenamiento Escalable	4

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introduction to modern NoSQL stores • Taxonomy of different NoSQL data models: key-value stores, column stores, document stores, graph stores, relational/tabular stores, triple stores, inverted indexes. • Brief introduction to modern NoSQL query and imperative languages.	Cover the taxonomy of current NoSQL stores commonly used for large-scale data management in cluster/cloud computing environments. Compare and contrast the strengths and weaknesses of different data models employed by stores. Learn about the different query languages employed by different stores.	[SF12]

• Programming assignment combining MapReduce (Hadoop) and a NoSQL store (TBD) for a distributed processing task.	Hands-on assignment will combine a NoSQL store and MapReduce to solve a distributed processing task, bringing together aspects of the different modules (intermediate Java skills a prerequisite).	

Bibliografía
• [TS06] A. S. Tanenbaum, M. Van Steen. Distributed Systems: Principles and Paradigms (2nd Edition). Prentice Hall, 2006. • [MABDHLC10] G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, G. Czajkowski. Pregel: a system for large-scale graph processing. SIGMOD Conference 2010: 135-146. • [KDF11] K. Hwang, J. Dongarra, G. C. Fox. Distributed and Cloud Computing: From Parallel Processing to the Internet of Things (1st Edition). Morgan Kaufmann, 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:	Otoño 2014
Elaborado por:	Pablo Barcelo / Aidan Hogan