FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

PROGRAMA DE CURSO

Código	Nombre
CC66I	Big Data
Nombre en Inglés	

SCT	Equivalencia	Horas de Cátedra	Horas Docencia Auxiliar	Horas de Trabajo Personal
3	CC66F	24	0	72

Requisitos	Carácter del Curso
Exclusivo Magister en TI.	Electivo

Propósito del Curso

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 sobre los nuevos almacenes de datos que son diseñados para escalar a través de múltiples máquinas. Entender cómo funcionan (de alto nivel) los sitios web como “Facebook”, “Twitter”, y “Google”.

Resultados de Aprendizaje

- 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 los nuevos almacenes de datos (NoSQL) que son diseñados para escalar a través de múltiples máquinas.
- Entender cómo funcionan (en alto nivel) los sitios de web como “Facebook” y “Twitter”.
- Entender cómo funcionan motores de búsqueda como “Google”.

Metodología Docente	Evaluación General
Clases expositivas de 90 minutos cada una y sesiones prácticas de 90 minutos.	100% laboratorios

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Unidades Temáticas

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 1	Introducción	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introducción al curso • Introducción a los desafíos de escala en Twitter.	• Entender que es "Big Data" • Entender porque se necesitan sistemas distribuidos. • Entender los desafíos y las metas en diseñar sistemas distribuidos.	[HDF11]

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 2	HDFS / Apache Hadoop	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Diseño de sistemas distribuidos. • Sistemas de archivos distribuidos: GFS y HDFS. • Infraestructuras modernas de procesamiento distribuido: MapReduce/Hadoop .	• Aprender cómo funcionan sistemas de archivos distribuidos. • Aprender programar y ejecutar tareas de MapReduce/Hadoop en un ambiente distribuido.	[DG04] [W12]

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 3	Apache Pig	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• El lenguaje de scripting: Apache Pig.	• Aprender programar en PIG	

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 4	Apache Spark	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Infraestructuras modernas de procesamiento distribuido.	• Entender las diferencias principales entre MapReduce y Spark. • Aprender diseñar y programar tareas en Spark.	

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 5	Recuperación de Información: Indexación	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Métodos para hacer crawling • Índices invertidos, compresión	• Aprender sobre los desafíos asociados con crawling. • Entender cómo funciona la indexación de motores de búsqueda como Google. • Ver los principios de compresión de índices invertidos • Crear un motor de búsqueda de texto escalable con Lucene.	[BP98] [BR11]

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 6	Recuperación de Información: Ranking	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• TF-IDF (revisitado) • Vector Space Model y Similaridad coseno. • PageRank	• Aprender técnicas de ranking para documentos de texto. • Poder implementar un algoritmo de PageRank y ver cómo afecta los resultados de una búsqueda.	[BP98] [BR11]

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 7	NoSQL: Key-Value, Tabular	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• Introducción a “NoSQL” • Garantías de bases de datos distribuidas: CAP. • Formas de distribuir los datos • Tipos de sistemas “NoSQL”	• Entender la motivación de NoSQL • Aprender sobre las técnicas más importantes que usan estos sistemas. • Poder cargar y consultar datos en Cassandra: un sistema NoSQL.	[D07] [TS06] [OV11] [SF12]

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Número	Nombre de la Unidad	Duración en Semanas
Clase N° 8	NoSQL: Documentos, Grafos	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
• MongoDB	• Poder cargar y consultar datos en MongoDB: un sistema "NoSQL".	

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).

[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:	Primavera 2018
Elaborado por:	Aidan Hogan