Procesamiento Masivo de Datos (BIG DATA)

PLANIFICACIÓN DE CURSO

Segundo Semestre académico 2023 - Docencia Presencial

I. ACTIVIDAD CURRICULAR Y CARGA HORARIA

Asignatura:	Procesamiento Masivo de Datos	Código: COM4002-1
Semestre de la Carrera:	Octavo semestre	
Carrera:	Ingeniería Civil en Computación	
Escuela:	Escuela de Ingeniería	
Docente(s):	ALEX DI GENOVA	
Ayudante(s):	POR DEFINIR	
Horario:	Cátedra: Lunes y miercoles de 12:00-13:30 horas. Ayudantía: Jueves 16:15-17:45 horas.	

Créditos SCT: 6	
Carga horaria semestral[1]: 180	horas
Carga horaria semanal:	13 horas

Tiempo de trabajo sincrónico semanal: 4,5	horas
Tiempo de trabajo asincrónico semanal: 8,5	horas

II. RESULTADOS U OBJETIVOS DE APRENDIZAJE ESPERADOS ESTE SEMESTRE

1)	Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
2)	Implementar comunicación entre procesos (OpenMP, pthreads) y máquinas (MPI)
3)	Utilizar Nextflow/Hadoop para distribuir tareas computacionales básicas
4)	Conocer la taxonomía de modelos de datos NoSQL, sus lenguajes de consulta. Construir y manipular una base de datos NO-SQL

[1] *Considere que 1 crédito SCT equivale a 30 horas de trabajo total (presencial/sincrónico y autónomo/asincrónico) en el semestre.*

5) Construir y manipular una base de datos distribuida.

III. UNIDADES, CONTENIDOS Y ACTIVIDADES

UNIDAD 1: **Distribución y Paralelismo**				
Semana	Contenidos	Actividades de enseñanza y aprendizaje		Actividades de evaluación diagnóstica, formativa y/o sumativa
		Tiempo sincrónico	Tiempo asincrónico (trabajo autónomo del o la estudiante)	
1 (28/08)	• Introducción: necesidad de manejar datos masivos	3 horas (Cátedra)	6,75 horas	
2 (04/09)	• Sistemas distribuidos: introducción, objetivos de un sistema distribuido	4,5 horas (Catedra y Ayudantía)	8,5 horas	
3 (11/09)	• Fundamentos de procesamiento paralelo y computación distribuida	4,5 horas (Cátedra y Ayudantía)	8,5 horas	
4 (25/09)	• Arquitecturas y programación en computación distribuida	4,5 horas (Cátedra y Ayudantía)	8,5 horas	

UNIDAD 2: **Modelamiento de Procesamiento Distribuido**				
Semana	Contenidos	Actividades de enseñanza y aprendizaje		Actividades de evaluación diagnóstica, formativa y/o sumativa
		Tiempo sincrónico	Tiempo asincrónico (trabajo	

			autónomo del o la estudiante)	
5 (02/010)	• Métodos tradicionales de comunicación distribuida (MPI/OpenMP/RPC)	4,5 horas (Cátedra y Ayudantía)	8,5horas	Tarea 1
6 (09/10)	• Métodos tradicionales de comunicación distribuida (MPI/OpenMP/RPC)	4,5 horas (Cátedra y Ayudantía)	8,5 horas	Control 1 (11/10)
7 (16/10)	• Procesamiento distribuido moderno: Nextflow, MapReduce, Hadoop,	4,5 horas (Cátedra y Ayudantía)	8,5 horas	
8 (23/10)	• Procesamiento distribuido moderno: Nextflow, MapReduce, Hadoop	4,5 horas (Cátedra y Ayudantía)	8,5 horas	

UNIDAD 3**: Modelos de Almacenamiento Escalable**				
Semana	Contenidos	Actividades de enseñanza y aprendizaje		Actividades de evaluación diagnóstica, formativa y/o sumativa
		Tiempo sincrónico	Tiempo asincrónico (trabajo autónomo del o la estudiante)	
9 (30/10)	• Motores NOSQL	4,5 horas (Cátedra y Ayudantía)	8,5 horas	
10 (6/11)	• Arquitecturas NOSQL	4,5 horas (Cátedra y Ayudantía)	8,5 horas	Control 2 (8/11)

11 (13/11)	• Implementación de bases de datos NOSQL	4,5 horas (Cátedra y Ayudantía)	8,5 horas	

UNIDAD 4**: Bases de datos Distribuidas**				
Semana	Contenidos	Actividades de enseñanza y aprendizaje		Actividades de evaluación diagnóstica, formativa y/o sumativa
		Tiempo sincrónico	Tiempo asincrónico (trabajo autónomo del o la estudiante)	
12 (20/11)	• Introducción a las bases de datos distribuidas	4,5 horas (Cátedra y Ayudantía)	8,5 horas	
13 (27/11)	• Estrategias de data-placement: sharding (partitioning), replication, duplication	4,5 horas (Cátedra y Ayudantía)	8,5 horas	Tarea2
14 (04/12)	• Procesamiento de consultas distribuidas y su optimización	4,5 horas (Cátedra y Ayudantía)	8,5 horas	Control 3 (06/12)
11/12				Control Recuperativo (13/12)

IV. CONDICIONES Y POLÍTICAS DE EVALUACIÓN

Se evaluará el aprendizaje del contenido presentado en las cátedras y en las ayudantías, mediante dos actividades complementarias (tareas, ejercicios) y tres controles de cátedra. Las ponderaciones de cada instancia de evaluación son las siguientes:

1. Calificaciones en actividades complementarias 25%.
2. Calificaciones en controles de cátedra 75%.

La Nota Final del curso se calculará considerando las ponderaciones anteriores. La aprobación de la asignatura está sujeta a las condiciones Nota Cátedra ≥ 4.0 y Nota de Actividades Complementarias ≥ 4.0. Por lo tanto, La aprobación no está sujeta a la Nota Final. En caso de que un estudiante repruebe por una de las 2 condiciones, pero su Nota Final sea mayor a 4,0; se le asignará en el Acta como nota final un 3,9.

Estudiantes que se ausenten a un control tendrán la oportunidad de recuperarlo durante el periodo correspondiente al final del semestre. El control recuperativo es de carácter acumulativo, por lo tanto, contendrá contenido de las cuatro unidades del curso. Adicionalmente, alumnos que quieran remplazar una calificación en un control o actividades complementarias, también podrán rendir el control recuperativo.

Un/a estudiante que cometa plagio sobtendrá un 1,0 en la evaluación y el caso será informado a Escuela de Ingeniería.

V. BIBLIOGRAFÍA Y RECURSOS OBLIGATORIOS

- S. Tanenbaum, M. Van Steen. Distributed Systems: Principles and Paradigms (2nd Edition). Prentice Hall, 2006
- P. J. Sadalage, M. Fowler. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley Professional, 2012

VI. BIBLIOGRAFÍA Y RECURSOS COMPLEMENTARIOS

- K. Hwang, J. Dongarra, G. C. Fox. Distributed and Cloud Computing: From Parallel Processing to the Internet of Things (1st Edition). Morgan Kaufmann, 2011
- M. T. Özsu, P. Valduriez. Principles of Distributed Database Systems. Springer, 2011.
- T. White. Hadoop: The Definitive Guide. O'Reilly, 2012
- 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.