



Universidad
Nacional
de Córdoba



Facultad
de Matemática,
Astronomía, Física
y Computación

EXP-UNC 23014/2016

VISTO

El contenido de la Res. HCD N° 207/2002, donde se detalla la nómina de materias optativas para la Licenciatura en Ciencias de la Computación; y

CONSIDERANDO

Que el Coordinador de la Comisión Asesora de Computación, Damián BARSOTTI, propone incorporar la materia "Programación Distribuída sobre Grandes Volúmenes de Datos" a la nómina de materias optativas de la Licenciatura en Ciencias de la Computación;

Que mediante Res. HCS N° 122/2002 se ha delegado en este cuerpo la facultad de modificar la nómina de materias optativas del Plan de Estudios de la Licenciatura en Ciencias de la Computación;

Que se cuenta con el aval del Consejo de Grado.

Por ello,

EL CONSEJO DIRECTIVO
DE LA FACULTAD DE MATEMÁTICA, ASTRONOMÍA, FÍSICA Y COMPUTACIÓN

RESUELVE:

ARTÍCULO 1°: Modificar la nómina de materias optativas del Plan de Estudios de la Licenciatura en Ciencias de la Computación, incorporando a la misma la materia "Programación Distribuída sobre Grandes Volúmenes de Datos".

ARTÍCULO 2°: Fijar como programa, correlativas y carga horaria de la materia, los detallados en el Anexo que forma parte de esta Resolución.

ARTÍCULO 3°: Remitir a la Secretaría de Asuntos Académicos de la Universidad esta Resolución para su conocimiento y efectos, en cumplimiento con lo establecido en el Art. 2° de la Res. HCS N° 122/2002.

ARTÍCULO 4°: Notifíquese, publíquese y archívese.

DADA EN LA SALA DE SESIONES DEL CONSEJO DIRECTIVO DE LA FACULTAD DE MATEMÁTICA, ASTRONOMÍA, FÍSICA Y COMPUTACIÓN A VEINTITRÉS DÍAS DEL MES DE MAYO DE DOS MIL DIECISEÍS.

RESOLUCIÓN CD N° 159/2016


Dra. SILVIA PATRICIA SILVETTI
SECRETARIA GENERAL
FaMAF


Dra. Ing. MIRTA IRIONDO
DECANA
FaMAF



ANEXO

Programa de Materia Optativa de la Licenciatura en Ciencias de la Computación

PROGRAMA DE ASIGNATURA	
ASIGNATURA: Programación Distribuida Sobre Grandes Volúmenes de Datos	AÑO: 2016
CHARACTER: Optativa	UBICACIÓN EN LA CARRERA: 5° año 2° cuatrimestre
CARRERA: Licenciatura en Ciencias de la Computación	
REGIMEN: Cuatrimestral	CARGA HORARIA: 120 horas

FUNDAMENTACIÓN Y OBJETIVOS

FUNDAMENTACIÓN: Un fenómeno observado en las tecnologías de la computación actuales es que la capacidad de procesamiento crece con menor velocidad que el volumen de datos, haciendo que sea imposible albergarlos y analizarlos en una sola computadora. Una solución a este problema es distribuir los datos y el cómputo sobre un gran número de dispositivos. El uso de este modelo implica no sólo la implementación de tecnologías de administración de grandes grupos de computadoras, sino principalmente un cambio en el paradigma de programación. Es así que los algoritmos para problemas dados deben ser adaptados a ambientes distribuidos mediante el uso de datos específicos y teniendo en cuenta la paralelización del cómputo, la partición de datos y la tolerancia a fallas de los dispositivos informáticos.

Este curso pretende abarcar tanto los aspectos tecnológicos como los metodológicos de la programación en ambientes distribuidos masivos, mediante el estudio teórico y práctico de las nuevas tecnologías que ayudan al proceso de construcción de esta clase de programas. Es de especial interés para aquellos estudiantes de posgrado y profesionales del ramo interesados en las aplicaciones tecnológicas y en la investigación interdisciplinaria.

OBJETIVOS: Al finalizar la materia los estudiantes estarán en condiciones de:

- Comprender los lineamientos básicos de los métodos de programación en ambientes masivamente distribuidos, en particular los empleados en el análisis de grandes volúmenes de datos.
- Comprender las tecnologías de administración de tareas y datos en ambientes masivamente distribuidos.
- Comprender los lineamientos básicos de técnicas de aprendizaje automático, en particular las empleadas en aprendizaje de grandes volúmenes de datos.
- Comprender y utilizar librerías específicas para el procesamiento de grandes volúmenes de datos.
- Analizar un problema que involucre grandes volúmenes de datos y diseñar experimentos de procesamiento automático sobre dichos datos.

CONTENIDO

1. Introducción

Marco histórico. Necesidad de análisis sobre grandes volúmenes de información. Sistemas de archivos y gestores de clusters. Data flow vs. network programming tradicional. Necesidad de patrones de programación. Arquitecturas para cálculo distribuido.

2. Map Reduce (MR):

Concepto general. Ejemplos de uso. Ejecución de MR. Límites de MR.

1. Estructuras de datos recuperables: Creación. Transformaciones. Acciones. Evaluación lazy. Persistencia. Distintas interfaces de programación.

2. Datos indexados: Fusión. Agrupamiento. Union. Ordenamiento. Acciones. Partición entre nodos.
3. Acceso a Datos: Archivos. Formatos. Sistemas de archivos. Datos estructurados. Base de datos. Web scraping. Anatomía de Datos. Limpieza de Datos.
4. Herramientas de programación: Acumuladores. Variables broadcast. Operaciones por partición. Comunicación con programas externos. Operaciones numéricas.
5. Cluster de computadoras: Arquitecturas. Gestores de clusters. Técnicas de depuración y puesta a punto.
6. Análisis de Datos y Machine Learning (ML).
 - a) Aprendizaje no supervisado: Análisis exploratorio de datos, reglas de asociación, clustering, visualización.
 - b) Aprendizaje supervisado: Clasificación: Árboles de decisión. Naive Bayes. Support vector machines. Random forest. Regresión: Regresión lineal. Regresión logística.
 - c) Filtrado Colaborativo: Mínimos cuadrados alternos.
 - d) Librerías de ML: Casos de uso. Ejemplos.
7. Tópicos avanzados: Streaming. SQL. Grafos. Estadística con R.

BIBLIOGRAFÍA

BIBLIOGRAFÍA BÁSICA

- Learning Spark: Lightning-Fast Big Data Analysis. Karau, H. and Konwinski, A. and Wendell, P. and Zaharia, M. "O'Reilly Media, Inc.", 2015.
- Machine Learning. Tom M. Mitchell. "McGraw-Hill", 1997.

BIBLIOGRAFÍA COMPLEMENTARIA

- Spark Programming Guide, <http://spark.apache.org/docs/latest/programming-guide.html>. Apache Software Foundation, 2015.

EVALUACIÓN

FORMAS DE EVALUACIÓN

Aprobación de 3 trabajos de laboratorio y la realización de un proyecto final con coloquio.

REGULARIDAD

- Aprobar al menos el 60 % de los Trabajos de Laboratorio.

PROMOCIÓN

- Aprobar todos los Trabajos de Laboratorio.
- Aprobar un proyecto final Integrador con coloquio.

CORRELATIVIDADES

Para cursar:

- Algoritmos y Estructuras de Datos II (aprobada).
- Probabilidad y Estadística (regularizada).
- Redes y Sistemas Distribuidos (regularizada).
- Base de Datos (regularizada).

Para rendir:

- Probabilidad y Estadística (aprobada).
- Redes y Sistemas Distribuidos (aprobada).
- Base de Datos (aprobada).

