Universidad Nacional de Córdoba
FACULTAD DE MATEMÁTICA ASTRONOMÍA Y FÍSICA

UNC

UNIVERSIDAD NACIONAL DE CÓRDOBA
Facultad de Matemática, Astronomía y Física

PROGRAMA DE CURSO DE POSGRADO

TÍTULO: Programación Distribuida sobre Grandes Volúmenes de Datos	
AÑO: 2017	**CUATRIMESTRE: segundo**
CARGA HORARIA: 120	**No. DE CRÉDITOS: 3**
CARRERA/S: Computación, Matemática, Astronomía, Física, Ingenierías	
DOCENTE ENCARGADO: Damián Barsotti	

PROGRAMA

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. Límitaciones 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. Anátomía 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

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.

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

MODALIDAD DE LA EVALUACIÓN

Aprobación de 3 trabajos de laboratorio y la realización de un proyecto final con coloquio en base a la aplicación de las herramientas aprendidas a un tema de relevancia científica o profesional y de interés del alumno.