UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

TEMARIO Big Data - Programación Distribuida y Data Science Aplicada

Unidad 1: Conceptos generales de big data
Repaso de conceptos de big data.
Repaso de la plataforma Hadoop para soluciones distribuidas.
Repaso de conceptos de bases de datos NoSQL y su relación a las arquitecturas distribuidas.
Rol del profesional de big data en las organizaciones (tales como data scientist, data engineer, etc).
Historia de la programación distribuida.

Unidad 2: Arquitectura distribuida
Introducción al concepto de cluster para aplicaciones relacionadas con problemáticas de big data.
Implementación de distribución y replicación de datos. Ventajas y desventajas.
Problemáticas y técnicas de escalabilidad en big data. Planificación a futuro.
Introducción a arquitectura lambda

Unidad 3: Machine learning y algoritmos de data mining
Problemáticas de clasificación y algoritmos tales como Naive Bayes, Decision Tree, etc
Problemáticas de regresión y su diferencia con clasificación.
Problemáticas de clustering y algoritmos tales como k-means y variantes.
Reducción dimensional, usos y aplicaciones.

Unidad 4: Lenguajes y Motores de procesamiento
Nociones generales de Python para ambientes distribuidos.
Primitivas fundamentales de Spark y Map reduce.
Procesamiento real time vs procesamiento batch en herramientas distribuidas.

Unidad 5: Validación de resultados y testing
Técnicas de validación como (por ejemplo, cross validation, split validation).
Interpretación y visualización de resultados.
Técnicas de fabricación de variables artificiales.