Universidad Tecnológica de la Mixteca

Clave DGP: 200089

Maestría en Ciencia de Datos

PROGRAMA DE ESTUDIOS

NOMBRE DE LA ASIGNATURA
Datos Masivos

SEMESTRE	CLAVE DE LA ASIGNATURA	TOTAL DE HORAS
Tercer semestre	**371032**	**35 Mediación docente** **65 Estudio independiente**

OBJETIVO(S) GENERAL(ES) DE LA ASIGNATURA
Estudiar, analizar y comprender a profundidad las herramientas computacionales para el almacenamiento y procesamiento de bases de datos masivas, mediante el enfoque del cómputo distribuido.

TEMAS Y SUBTEMAS

1 Introducción a los datos masivos.

1.1 Definición y características (Volumen, Velocidad, Variedad, Valor y Veracidad).
1.2 Importancia y privacidad de los datos masivos.
1.3 Arquitectura: integración y flujo de datos.
1.4 Almacenamiento clave-valor
1.5 Arquitecturas

2. Almacenamiento masivo.

2.1 Hadoop File System
2.2 Amazon S3
2.3 Parquet

3. Procesamiento distribuido

3.1 MapReduce
3.2 Hadoop
3.3 Yarn
3.4 Spark o Flink (Pyspark)
3.5 AWS Elastic MapReduce

4. Aprendizaje máquina con datos masivos

4.1 Pytorch y Tensorflow en Sagemarker
4.2 API Rest de Yarn
4.3 AWS Cloud 9
4.4 Sagemarker R

ACTIVIDADES DE APRENDIZAJE
Sesiones dirigidas por parte del profesor en la que se presentan los conceptos poniendo énfasis en las herramientas de cómputo. Se realizarán algunos ejemplos al finalizar cada capítulo con el objetivo de que se refuerce la teoría vista en cada tema. Algunas herramientas son comerciales por lo que se recomienda auxiliarse de guías para su exposición.

CRITERIOS Y PROCEDIMIENTOS DE EVALUACIÓN Y ACREDITACIÓN
Exámenes parciales y final. Tareas Simulaciones en computadora. Proyectos. Esto tendrá una equivalencia del 100% en la calificación final del semestre

BIBLIOGRAFÍA (TIPO, TÍTULO, AUTOR, EDITORIAL Y AÑO)

Básica:

1. Nathan Marz y James Warren, Big Data: Principles and Best Practices of Scalable Real-Time Data Systems.

2. Ortega, J.M. (2023). Big data, machine learning y data science en python. Editorial. RA-MA, S.A. 669 pp
3. Rajaraman, A. and Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.

Consulta:

1. Learning Spark: Lightning-Fast Data Analytics" por Holden Karau, Andy Konwinski, Patrick Wendell y Matei Zaharia
2. Hadoop: The Definitive Guide" por Tom White

PERFIL PROFESIONAL DEL DOCENTE
Doctorado en Ciencias de la Computación, o áreas afines, con especialidad en Inteligencia artificial y/o Ciencia de datos.

Vo.Bo
M.T.C.A. ERIK GERMÁN RAMOS PÉREZ
COORDINADOR DE LA UNIVERSIDAD VIRTUAL

AUTORIZÓ
L.I. MARIO ALBERTO MORENO ROCHA
VICE-RECTOR ACADÉMICO