Título: Ciencia de Datos a través del Big Data

Duración: 20 horas

Introducción:
Este curso ofrece una inmersión en el mundo del Big Data, desde los conceptos fundamentales hasta las últimas herramientas de análisis. Los participantes explorarán los principios esenciales de Big Data, aprendiendo a gestionar grandes volúmenes de datos. Además, se sumergirán en el ecosistema de Hadoop, dominando el arte del procesamiento y almacenamiento distribuido en el Hadoop Distributed File System (HDFS).

Una vez establecidas las bases, el curso se adentrará en Apache Spark, ofreciendo una visión detallada de esta potente y popular plataforma de procesamiento en memoria y su papel fundamental en el análisis de Big Data. Desde la manipulación de datos hasta aplicaciones reales, los estudiantes adquirirán habilidades prácticas para abordar los desafíos más complejos en el ámbito del análisis de datos.

Programa detallado del curso:

Introducción a Big Data (2 horas)

- Clase 1 (2 horas):
  - Conceptos básicos de Big Data
  - Qué es y qué no es Big Data

- Clase 2 (2 horas):
  - Desafíos en el manejo de grandes volúmenes de datos
  - Ejemplos de Big Data
  - Privacidad de datos

Fundamentos de MapReduce y Hadoop (6 horas)

- Clase 3 (2 horas):
  - Principios de Big Data y escalabilidad
  - Arquitectura de MapReduce
  - Programación MapReduce: Map, Reduce y Shuffle

- Clase 4 (2 horas):
  - Introducción a Hadoop y su ecosistema
  - Implementación de un programa MapReduce básico

- Clase 5 (2 horas):
  - Conceptos básicos de HDFS (Hadoop Distributed File System)
  - Optimización de tareas MapReduce
  - Administración de clústeres Hadoop

Procesamiento avanzado de datos con Apache Spark (10 horas)

- Clase 6 (2 horas):
  - Introducción a Apache Spark y sus características
  - Diferencias entre Hadoop y Spark

- Clase 7 (2 horas):
  - Programación en Spark
  - Operaciones básicas en Spark: map, reduce, filter, etc.

- Clase 8 (2 horas):
  - Spark SQL y DataFrames: procesamiento de datos estructurados

- Clases 9 y 10 (4 horas):
  - Ejemplos con ML en Spark

Método de evaluación:

- Cuestionarios para los aspectos teóricos
- Implementación de proyectos para los aspectos prácticos

Bibliografía:

- Triguero, I., & Galar, M. (2023). Large-Scale Data Analytics with Python and Spark: A Hands-on Guide to Implementing Machine Learning Solutions. Cambridge University Press.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). Big data preprocessing. Cham: Springer.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, No. 2018). Cham: Springer.
- García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.
- Chambers, B., & Zaharia, M. (2018). Spark: The definitive guide: Big data processing made simple. " O'Reilly Media, Inc.".
- Hamstra, M., & Zaharia, M. (2013). Learning Spark: lightning-fast big data analytics. O'Reilly & Associates.
- White, T. (2012). Hadoop: The definitive guide. " O'Reilly Media, Inc.".