

Gestión de Datos Masivos para Aprendizaje Automático Sergio Ruben D'Arrigo (profesor adjunto con dedicación parcial)
--

Programa:

La materia está orientada a que los estudiantes sean capaces de:

- Conocer la complejidad del tratamiento de datos masivos para la inteligencia artificial, el aprendizaje automático y la analítica avanzada
- Conocer modelos de almacenamiento de datos masivos (data warehouses, data lakes, data *lakehouses*), sus fundamentos, su concepción, diseño e implementación.
- Utilizar estos repositorios analíticos como origen de datos para las tareas de inteligencia artificial, aprendizaje automático y analítica avanzada
- Conocer las nuevas tendencias tecnológicas en la era del Big Data
- Resolver casos de uso concretos aplicando los conocimientos brindados de la materia

Temario:

- Introducción y motivación:
 - Introducción. Importancia de los datos y rol de la analítica. Organizaciones “Data Driven”. Clasificación de los tipos de datos. ¿Qué entendemos por datos masivos?
 - Desafíos para la inteligencia artificial y la analítica. Acceso e ingesta. Procesamiento. Integración.
 - Almacenamiento. Explotación y visualización.
- Modelos y Diseño de Almacenamiento de Datos Masivos:
 - Principales soluciones de almacenamiento de datos masivos. Modelos multidimensionales. Conceptos de Data Warehouse y Data Lake. Arquitectura, modelización, diseño. Linaje y gobierno de datos. Protección de datos personales. Soluciones cloud.
 - Procesos de ingesta de datos masivos:
 - Implementación. Ingesta, almacenamiento y actualización. ETL y ELT. Ingesta en tiempo real.
 - Integridad y calidad.
- Explotación, visualización y análisis de datos masivos:
 - Exploración, BI, Aprendizaje Automático. Utilización de sandbox analíticos.
- Nuevas tendencias para almacenamiento de datos masivos:
 - Variantes de data warehouses. Data *warehouses* y Big Data. Data *lakehouses*.
 - Utilización de IA para extracción de características. Estado del arte.

Bibliografía:

Libros

- DAMA International (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd Ed.). Technics Publications
- Eckerson W. (2011) Performance Dashboards. Measuring, Monitoring and Managing your Business (2da. Ed.). Wiley

- Golfarelli, M. & Rizzi, S.. (2009) Data Warehouse Design, Modern Principles and Methodologies (1ra. Ed.), McGraw-Hill
- Gorelik A. (2019). The Enterprise Big Data Lake – Delivering the Promise of Big Data and Data Science (1ra. Ed.). O'Reilly Media
- Inmon B., Levin M. & Srivstava R. (2021). Building the Data Lakehouse (1ra. Ed.). Technics Publications
- Inmon B. (2016) Data Lake Architecture (1ra. Ed) Technics Publications
- Inmon W. (2005) Building the Data Warehouse (4ta. Ed.). Wiley Publishing
- Kimball R. & Ross M. (2013) The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling (3ra. Ed.). Wiley & Sons
- Kitchin R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & their consequences (1ra. Ed.). SAGE Publications Ltd
- Krishnan K. (2013). Data Warehousing in the Age of Big Data (1ra. Ed.) Morgan Kaufmann.
- Menon P. (2022) Data Lakehouse in Action (1ra. Ed.). Packt Publishing
- Provost F. & Fawcett T. (2013) Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly
- Serra J (2024) Deciphering Data Architectures (1ra. Ed.). O'Reilly
- Sharda R., Delen D. & Turban E..(2017) Business Intelligence, Analytics and Data Science: A Managerial Perspective (4ta. Ed.) Pearson
- Silberschatz A., Korth H. & Sudarshan S. (2020). Database System Concepts (7ma. Ed.), McGraw Hill
- Tan P., Steinbach M., Kumar V. & Karpatne A. (2019) Introduction to Data Mining. (2da Ed.). Pearson
- Vaisman, A. & Zimanyi, E. (2022). Data Warehouse Systems: Design and Implementation (2da. Ed.). Springer-Verlag GmbH

Artículos

- Armbrust, Ghodsi, Xin, Zaharia. (2021) Lakehouse: A New Generation of Open Platforms that Unify DataWarehousing and Advanced Analytics, 11th Annual Conference on Innovative Data Systems Research (CIDR '21), January 11–15, 2021
- Armbrust M., Das T., Sun L., Yavuz B., Zhu S., Murthy M., Torres J., van Hovell H., Ionescu A., Łuszczak A., Switakowski M., Szafranski M., Xiao Li, Ueshin T., Mokhtar M., Boncz P., Ghodsi A., Paranjpye S., Senster P., Xin R. & Zaharia M. (2020). Delta Lake: HighPerformance ACID Table Storage over Cloud Object Stores. Proceedings of the VLDB Endowment, Vol. 13, No. 12
- Costa V. & Santos M.. (2018) Evaluating Several Design Patterns and Trends in Big Data Warehousing Systems. En Advanced Information Systems Engineering . Springer International Publishing AG, part of Springer Nature 2018
- Delen & Ram (2018) Research challenges and opportunities in business analytics.. En Journal of Business Analytics, Vol.1 2018
- Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996) From Data Mining to Knowledge Discovery in Databases., 1996. En AI Magazine Volume 17 Number 3
- Golfarelli, M., & Rizzi, S. (2018). From Star Schemas to Big Data: 20+ Years of Data Warehouse Research. A Comprehensive Guide Through Italian Database Research.
- Khalid, Fakharaldien, Zain & Majid (2015) Big Data Analysis and Storage. Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering, Orlando, Florida, USA, September 10-11, 2015
- Komorowsky M., Marshall D., Scalciccioli J. & Crutain Y. (2016) Exploratory Data Analysis,. En Secondary Analysis of Electronic Health Records, MIT Critical Data

- Melmood H, Gilman E, Cortes M, Kostakos P, Byrne A, Valta K, Tekes S & Riekkilä J (2019) Implementing big data lake from heterogeneous data sources. En 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)