



UNIVERSITAT D'ANDORRA

Pla docent mòdul

Pla d'estudis	Màster en Analítica de Dades		
Mòdul	M4: Captura, preprocessat i emmagatzematge de dades		
Codi mòdul		Tipus	Obligatori
Semestre	2		
Coordinació del Mòdul A/e	Joan Ribera Monfort jriberam@uda.ad		
Modalitat	Semipresencial i virtual		
Llengua de docència	Català		
Crèdits	10		
Recomanació de Matrícula	Cap		
Seminari	Captura i preprocessat de dades		
Professorat responsable	Joan Ribera Monfort jriberam@uda.ad		
Professorat Col·laborador (Legislació)	Agda Castillo Massoni acastillom@uda.ad		
Seminari	Emmagatzematge de dades		
Professorat responsable	Paula Rodó i Vila prodo@uda.ad		

Contingut

Pla d'estudis	1
Presentació del Mòdul	2
Continguts del Mòdul	2
Seminari: Captura i preprocessat de dades	2
Seminari: Emmagatzematge de dades	3
Activitats d'avaluació del mòdul	5
Recursos del mòdul	7
Bibliografia	7
Altres Recursos	7
Observacions del Mòdul	8

Presentació del Mòdul

L'objectiu d'aquest mòdul és adquirir els coneixements i competències sobre les fases prèvies a l'anàlisi de dades, concretament de la captació, el preprocessament i l'emmagatzematge.

Primer, es tractaran els principals mètodes de captura de dades, com són l'*scripting*, els fluxos d'extracció, el *web scraping* o l'accés a API.

Posteriorment, es parlarà sobre les funcions de transformació i integració de dades provinents de diferents fonts per, finalment, emmagatzemar aquestes dades en diferents sistemes *NoSQL*, com poden ser els clau-valor, els orientats a documents, o els orientats a grafs. En l'àmbit empresarial, es valoraran les diferències entre els sistemes de bases de dades analítics i els operacionals, i s'introduirà el concepte de factoria d'informació corporativa.

A més, en aquest mòdul, es tractarà el terme de *DataOps*, abreviatura *Data Operations* i que tracta d'agrupar els diferents grups i persones de l'organització, per d'aconseguir, de manera àgil i seguint un conjunt de bones pràctiques, els següents objectius: integració de les dades, validació, gestió de les metadades i observabilitat.

En definitiva, en aquest mòdul, l'estudiant aprendrà les diferents tasques necessàries per arribar a obtenir les dades netes necessàries per a l'anàlisi. A més, un altre objectiu és que l'estudiant conegui i actuï seguint els principis ètics de captació de dades i les polítiques de protecció de dades.

Continguts del Mòdul

Seminari: Captura i preprocessat de dades

Fases del procés de captura de dades
● Fonts de dades ● Formats de dades ● Tècniques de captura de dades ● Aplicació en un entorn real



Pla docent mòdul

Preprocessat de dades

- Governança, dades mestres i qualitat de dades a nivell organitzatiu
- Especificitats de l'anàlisi: discretització, reducció de les dimensions, entre altres

Legislació aplicable al tractament de dades

- Llei Qualificada de Protecció de Dades
- Reglament Europeu General de Protecció de Dades
- Llei sobre drets d'autor i drets veïns

Seminari: Emmagatzematge de dades

Introducció a les bases de dades analítiques

- Introducció a les bases de dades analítiques
- Sistemes analítics vs. Operacionals
- La factoria d'informació corporativa
- Tendències actuals

Introducció als entorns d'emmagatzematge distribuïts

- Diferents tipus: blocs, fitxers i objectes
- La persistència políglota
- Emmagatzematge local distribuït
- Entorn integrats de dades en el núvol
- Flexibilitat, escalabilitat i rendiment
- Exemples de sistemes: IPFS, RAID, Ceph o Gluster, entre d'altres

Característiques i arquitectures de les bases de dades **NoSQL**

- Introducció a les bases de dades NoSQL
- Motivació i característiques dels models d'agregació



UNIVERSITAT D'ANDORRA

Pla docent mòdul

- BD clau-valor
- BD documental
- BD orientada a grafs
- Modelat de dades en entorns NoSQL



UNIVERSITAT D'ANDORRA

Pla docent mòdul

Activitats d'avaluació del mòdul

			MÒDUL					Avaluaci ó Total	
			Seminari Captura i preprocessament de dades			Seminari Emmagatzematge de dades			REP TE
			Treball Virtual 1	Treball Virtual 2	Treball Legislaci ó	Treball Virtual 3	Treball Virtual 4		
Resultats d'Aprenentatge (RA)	Específics	MAD-E003-03: Dissenya repositoris de dades per a emmagatzemar dades heterogènies.				35 %	35 %	30 %	100%
		MAD-E003-04: Usa gestors de bases de dades relacionals i no relacionals perquè les dades puguin ser analitzades.				35 %	35 %	30 %	100%
		MAD-E004-01: Identifica les dades essencials que cal capturar	30%	30%				40 %	100%
		MAD-E004-02: Transforma les dades capturades en dades útils per l'anàlisi a realitzar.	30 %	30 %				40 %	100%
		MAD-E004-03: Aplica tècniques àgils per a la integració i validació de les dades, això com la gestió de les metadades associades.	30 %	30 %				40 %	100%



UNIVERSITAT D'ANDORRA

Pla docent mòdul

	Transversals	MAD-T004: Actuar segons l'ètica i la responsabilitat social com a ciutadà i com a professional			80%			20 %	100%
--	--------------	---	--	--	-----	--	--	------	------

IMPORTANT: és obligatori el lliurament i presentació (defensa) del repte per aprovar el mòdul.



Recursos del mòdul

Bibliografia

Material del professorat

- Apunts del professorat.
- Material de l'aula virtual.

Seminari: Captura i preprocessat de dades

Bibliografia de referència

- **Data Wrangling with Python, tips and tools to make your life easier.** Jaqueline Kazil i Katherine Jarmul (2016).
- **Data Science from Scratch. First principles with Python.** Joel Grus (2019).
- **Cleaning data for effective Data Science.** David Mertz (2021).
- Articles en xarxa publicats pel professorat

Bibliografia complementària

- **"Web Scraping with Python. Collecting more data from the modern web".** Ryan Mitchell (2018)

Seminari: Emmagatzematge de dades

Bibliografia de referència

- **Introducción al Data Warehouse.** Abelló A.; Curto J.; Rius A.; Serra M.; Samos J.; Vidal J. Fundación para la Universitat Oberta de Catalunya.
- **Introducción a NoSQL.** M.E. Rodríguez, P. Urbón & J. Conesa i Caralt, UOC (2021)
- **Graph Databases.** I. Robinson, J. Webber & E. Eifren. O'Reilly (2013)
- Articles en xarxa publicats pel professorat

Altres Recursos

Bibliografia complementària

- **The Data warehouse toolkit: the complete guide to dimensional modeling.** Kimball, R. Nueva York. Wiley. (2002)



UNIVERSITAT D'ANDORRA

Pla docent mòdul

- **NoSQL Distilled. A brief Guide to the Emerging World of Polyglot Persistence.** P.J. Sadalage & M. Fowler. Pearson Education. (2013)

Observacions del Mòdul

- **No s'acceptaran lliuraments fora de termini.**
- Si un estudiant no realitza alguna de les activitats avaluables, la qualificació dels resultats d'aprenentatge associats a l'activitat no realitzada, serà igual a zero.
- Tots els lliuraments avaluables s'efectuaran al campus virtual de l'UdA i s'hauran de lliurar en el format indicat a l'enunciat de l'activitat.



Els continguts d'aquesta obra estan subjectes a una llicència de Reconeixement - No comercial - Sense obres derivades 3.0 de Creative Commons. Se'n permet la reproducció, distribució i comunicació pública sempre que se'n citi l'autor i no se'n faci un ús comercial. La llicència completa es pot consultar a: <http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode>.