

**FITXA IDENTIFICATIVA****Dades de l'Assignatura**

Codi	36437
Nom	Dades massives
Cicle	Grau
Crèdits ECTS	6.0
Curs acadèmic	2022 - 2023

Titulació/titulacions

Titulació	Centre	Curs	Període
1406 - Grau en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	4	Primer quadrimestre

Matèries

Titulació	Matèria	Caràcter
1406 - Grau en Ciència de Dades	12 - Computació	Obligatòria

Coordinació

Nom	Departament
LIBEROS MASCARELL, ALEJANDRO	242 - Enginyeria Electrònica
RODRIGO BORT, MIGUEL	242 - Enginyeria Electrònica

RESUM

L'assignatura “Dades massives” és una assignatura quadrimestral, consta de 6 crèdits ECTS i s'imparteix durant el 1r quadrimestre del quart curs del Grau en Ciència de Dades de la Universitat de València.

Aquesta assignatura pretén consolidar les competències de l'alumnat en l'adquisició, emmagatzematge i tractament de les dades, i alhora aportar les eines per a l'escalabilitat d'aquestes tècniques al volum de dades massives.

De manera general, en finalitzar el curs l'alumnat ha de ser capaç de:

- Entendre el concepte de dades massives o Big Data, conèixer el seu impacte social i aplicacions.
- Conèixer i saber utilitzar els diferents recursos maquinari i programari per a l'emmagatzematge de dades.
- Conèixer i saber utilitzar les eines maquinari i programari pròpies dels sistemes de fitxers distribuïts.
- Conèixer i saber utilitzar tècniques bàsiques de programació paral·lela per a l'accés i tractament de sistemes de dades distribuïdes.
- Conèixer i saber desenvolupar diferents algorismes avançats per a l'explotació de dades massives fent ús de les capacitats de computació distribuïda de les arquitectures introduïdes.



La informació de contacte amb el professorat responsable està publicada en la web del Departament d'Enginyeria Electrònica (<http://www.uv.es/die>). El material de l'assignatura (apunts, guions de pràctiques, activitats, etc.) estarà a l'abast de l'alumnat a través d'Aula Virtual, la plataforma d'e-learning de la Universitat de València (<http://aulavirtual.uv.es/>).

El professorat farà ús, preferiblement, del correu electrònic per a convocar actes d'avaluació, puntualitzacions sobre aquesta guia docent i altres aspectes rellevants de cara al procés d'ensenyament-aprenentatge. D'igual manera, el professorat podrà introduir tasques d'avaluació contínua o altres aspectes rellevants en el desenvolupament diari a l'aula, així com a l'Aula Virtual de l'assignatura.

Les classes de teoria s'impartiran en castellà i les classes pràctiques i de laboratori segons consta en la fitxa de l'assignatura disponible en la web del grau.

CONEXIMENTS PREVIS

Relació amb altres assignatures de la mateixa titulació

No heu especificat les restriccions de matrícula amb altres assignatures del pla d'estudis.

Altres tipus de requisits

L'assignatura es desenvolupa en l'últim curs del Grau en Ciència de Dades, i per tant suposa una etapa de consolidació i ampliació de gran part dels coneixements treballats durant el grau.

Sense haver-hi requisits previs per a la matriculació de l'assignatura, es recomana haver cursat prèviament les assignatures de Xarxes i Seguretat, Programació Paral·lela i Infraestructura d'Emmagatzematge de Dades. Es recomana també haver treballat prèviament amb els llenguatges de programació treballats durant el grau.

COMPETÈNCIES

1406 - Grau en Ciència de Dades

- (CG02) Capacitat de resoldre problemes amb iniciativa, creativitat, i de comunicar i transmetre coneixements, habilitats i destreses, comprenent la responsabilitat ètica i professional de l'activitat del Científic de Dades.
- (CT03) Habilitat per defensar el seu treball amb rigor i arguments, exposant-ho de forma adequada i precisa, recolzant-se en els mitjans necessaris.
- (CT05) Capacitat per avaluar els avantatges i inconvenients de diferents alternatives metodològiques i/o tecnològiques en diferents àmbits d'aplicació.
- (CE02) Conèixer i aplicar de forma metodològica les tècniques de programació i l'algorísmia necessàries per al processament eficient d'informació i la resolució informàtica de problemes que utilitzen grans volums de dades.



- (CE04) Conèixer i utilitzar els diferents models d'emmagatzematge de dades i els sistemes de gestió de les bases de dades utilitzant llenguatges de programació de definició, consulta i manipulació dels mateixos.
- (CE08) Capacitat per comprendre, seleccionar i utilitzar la infraestructura i tècniques adequades per al tractament de dades massives, atenent a criteris d'eficiència, escalabilitat, seguretat, tolerància a fallades i adequació a l'entorn de producció.
- (CE11) Capacitat per dissenyar i implementar la presa de dades, la seva integració, transformació, selecció, comprovació de la seva qualitat i veracitat a partir de diferents fonts, tenint en compte el seu caràcter, heterogeneïtat i variabilitat.
- (CB2) Que els estudiants sàpiau aplicar els seus coneixements al seu treball o vocació d'una forma professional i posseïsquen les competències que solen demostrar-se per mitjà de l'elaboració i defensa d'arguments i la resolució de problemes dins de la seua àrea d'estudi.
- (CB5) Que els estudiants hagen desenvolupat aquelles habilitats d'aprenentatge necessàries per a emprendre estudis posteriors amb un alt grau d'autonomia.

RESULTATS DE L'APRENTATGE

Els objectius generals de l'assignatura descrits en el resum juntament amb les competències generals, bàsiques, transversals i específiques, resulten en els següents resultats d'aprenentatge (RA).

Identificar i descriure els requeriments d'emmagatzematge i processament dels sistemes de dades massives.

RA1- Conèixer i saber utilitzar els serveis i eines oferits pels sistemes operatius per a l'emmagatzematge de dades (CG02, CB5, CE04).

RA2- Conèixer i saber utilitzar els diferents nivells d'emmagatzematge local, des dels dispositius físics d'emmagatzematge fins als sistemes de fitxers (CG03, CB2, CE08, CE11).

RA4- Conèixer l'estructura bàsica de suport físic usat en "Big Data" (CG03, CB2, CE08, CE11, CT05).

RA9- Identificar i descriure els requeriments d'emmagatzematge i processament dels sistemes de dades massives (CG03, CB2, CT05, CE04).

Descriure i aplicar els nous paradigmes de processament paral·lel i distribuït necessaris en els sistemes de dades massives.

RA5- Conèixer l'estructura de capes de les xarxes de computadors, així com els principals protocols i serveis usats en Internet i en el tractament de dades (CG03, CB5, CE04).

RA6- Conèixer i saber utilitzar els dispositius físics i virtuals necessaris per a crear i mantindre xarxes de processadors (CG02, CB5, CT03, CE02).

RA10- Descriure i aplicar els nous paradigmes de processament paral·lel i distribuït necessaris en els sistemes de dades massives (CG03, CB2, CT03, CE02).

Conèixer l'arquitectura i gestionar els recursos de sistemes de fitxers distribuïts.

RA3- Conèixer les tècniques habituals de virtualització de recursos i saber seleccionar-les i utilitzar-les (CG03, CB2, CT05, CE11).

RA7- Conèixer els riscos derivats de l'obtenció, processament, emmagatzematge i intercanvi de dades (CG02, CB2, CT03, CE04).

RA11- Conèixer l'arquitectura i gestionar els recursos de sistemes de fitxers distribuïts (CG02, CB2, CT03, CE08).



Conèixer i aplicar diferents paradigmes sobre arquitectura, flux de dades i el model de programació de dades massives.

RA8- Seleccionar i aplicar les mesures tècniques que permeten mantindre la seguretat dels sistemes d'obtenció, processament, emmagatzematge i intercanvi de dades (CG02, CB2, CT03, CE08).

RA12- Conèixer i aplicar diferents paradigmes sobre arquitectura, flux de dades i el model de programació de dades massives (CE02, CT03, CB5, CG02).

RA16- Avaluar les prestacions i escalabilitat d'un sistema de processament paral·lel, establint i aplicant les mètriques per a la seua comparació (CG03, CB2, CT03, CE02).

Conèixer i utilitzar els serveis oferits pels sistemes de computació en el núvol per al processament de dades massives.

RA13- Conèixer i utilitzar els serveis oferits pels sistemes de computació en el núvol per al processament de dades massives (CG02, CB2, CT05, CE02).

RA14- Identificar i descriure les arquitectures dels computadors paral·lels i distribuïts (CE04, CT03, CB2, CG03).

RA15- Conèixer i aplicar els paradigmes de programació paral·lela i distribuïda, els models de programació relacionats i els estàndards per al desenvolupament de sistemes d'altres prestacions (CG03, CB2, CT03, CE02).

RA17- Dissenyar i desenvolupar algorismes distribuïts que exploten les capacitats de paral·lelisme de les infraestructures de computació paral·lela i distribuïda (CE02, CT03, CB5, CG02).

RA18- Dissenyar i desenvolupar programes que utilitzen amb eficiència els multiprocessadors i les arquitectures distribuïdes per al processament de dades (CE02, CT03, CB5, CG02).

DESCRIPCIÓ DE CONTINGUTS

1. Introducció a les dades massives

En aquesta unitat es definirà el concepte de dades massives o Big Data, la dependència d'aquesta definició amb els recursos maquinari disponible/necessari. També es tractarà l'impacte d'aquesta tecnologia en els últims anys. (RA4).

2. Emmagatzematge de dades

En aquesta unitat s'aprofundirà en l'emmagatzematge de dades des del punt de vista dels recursos programari i maquinari (físic i virtual) que permeten la definició, consulta, manteniment i manipulació de volums de dades massives. S'introduirà i treballarà el model d'emmagatzematge de dades HDFS (Hadoop Distributed File System). (RA: 1-4).

3. Dades distribuïdes

En aquesta unitat es treballarà sobre les eines maquinari i programari per a la generació, manteniment, accés i tractament de bases de dades distribuïdes en xarxes interconnectades. Es prestarà especial atenció als protocols que asseguren la integritat de les dades en els sistemes de fitxers distribuïts i a la virtualització de xarxes o sistemes de computació en el núvol, entre altres. Es treballarà amb el protocol HDFS així com amb sistemes d'emmagatzematge distribuït en plataformes en el núvol com GoogleCloud (RA: 4-9, 11, 13).



4. Programació distribuïda

En aquesta unitat s'introduiran les tècniques bàsiques per al tractament i processament de dades de manera distribuïda/paral·lela en xarxes de processadors. Es prestarà especial atenció al model de programació MapReduce i a l'entorn Apache Spark. (RA: 5-6, 8-15).

5. Aplicacions i algorismes de programació distribuïda.

En aquesta unitat es consolidaran les competències de l'alumnat en l'explotació de les dades massives mitjançant el disseny i avaluació d'algorismes de programació distribuïda d'alt nivell que exploten les capacitats de paral·lelisme de les infraestructures de computació paral·lela. Es profunditzarà en l'aprenentatge d'Apache Spark sobre bases de dades distribuïdes amb HDFS en plataformes en el núvol com GoogleCloud (RA: 10-18).

VOLUM DE TREBALL

ACTIVITAT	Hores	% Presencial
Classes de teoria	34,00	100
Pràctiques en laboratori	20,00	100
Pràctiques en aula	6,00	100
Elaboració de treballs en grup	20,00	0
Estudi i treball autònom	24,00	0
Lectures de material complementari	6,00	0
Preparació d'activitats d'avaluació	24,00	0
Preparació de classes de teoria	6,00	0
Preparació de classes pràctiques i de problemes	10,00	0
TOTAL	150,00	

METODOLOGIA DOCENT

Classes presencials/síncrones

En les classes es desenvoluparan els continguts de la matèria emprant una metodologia expositiva.

S'instaurarà també la participació de l'alumnat a través de diferents mitjans: es podran emprar qüestionaris en línia, qüestions obertes o qualsevol altra eina per a avaluar el grau de consolidació de la matèria.

Aquesta metodologia expositiva/interactiva s'alternarà amb pràctiques d'aula en les quals es resoldran casos pràctics o s'estudiaran casos d'aplicació entre altres.

Competències més rellevants: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.



Pràctiques de laboratori

Els continguts treballats en classe es consolidaran amb diferents pràctiques de laboratori d'assistència obligatòria on es proposaran diferents exercicis pràctics relacionats amb la matèria. Els resultats d'aquestes pràctiques s'entregaran de manera escalonada al llarg del curs per a la seua avaluació, d'acord amb les indicacions del professorat.

Competències més rellevants: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

Preparació de treballs pràctics i teòrics

A més de les tasques descrites, l'alumnat haurà de fer tasques de manera no presencial associades a la preparació de pràctiques i l'elaboració de memòries i informes en els quals s'ha de prestar especial atenció a la descripció amb rigor i amb capacitat crítica de les diferents metodologies emprades i decisions preses (CT03, CT05). A més, serà imprescindible un estudi individual de la matèria tant de manera contínua com de cara als diferents actes d'avaluació. Durant el curs es podrà a més sol·licitar la consulta de publicacions o l'assistència a xarrades d'interés relacionades amb la matèria.

Competències més rellevants: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG03, CG04, CG05, CG07.

Tutories

Al llarg del curs el professorat atendrà dubtes tant a través d'e-mail, com de tutories en format presencial o per teleconferència. La forma de contacte amb el professorat serà el correu electrònic tal com s'indicarà en Aula Virtual. A més, la informació de contacte amb el professorat responsable està publicada en la web del Departament d'Enginyeria Electrònica (<http://www.uv.es/die>).

Plataforma e-learning i comunicació

El material de l'assignatura (apunts, guions de pràctiques, activitats, etc.) estarà a l'abast de l'alumnat a través d'Aula Virtual, la plataforma d'e-learning de la Universitat de València (<http://aulavirtual.uv.es/>).

El professorat farà ús, preferiblement, del correu electrònic per a convocar actes d'avaluació, puntualitzacions sobre aquesta guia docent i altres aspectes rellevants de cara al procés d'ensenyament-aprenentatge. D'igual manera, el professorat podrà introduir tasques d'avaluació contínua o altres aspectes rellevants en el desenvolupament diari a l'aula, així com a l'Aula Virtual de l'assignatura.

AVALUACIÓ

Pel que fa a l'avaluació es tindran en compte diferents dimensions del procés d'ensenyament-aprenentatge. En primer lloc, el sistema d'avaluació respon a les diferents competències, resultats d'aprenentatge i continguts a treballar durant el curs. En segon lloc, es balancejaran tant aquelles activitats desenvolupades en grup com el treball individual. Finalment, l'avaluació es proposa com a formativa, és a dir, es facilitaran comentaris que afavorisquen l'esmena d'aspectes a millorar detectats durant el curs, ja siga en la interacció diària entre alumnat i professorat, a través de comentaris en Aula Virtual o en sessions de revisió.

Tant en primera com en segona convocatòria, la nota final (NF) respon als diferents sistemes d'avaluació (SE) atesa la següent expressió:

$$NF = SE1 \cdot 0,5 + SE2 \cdot 0,3 + SE3 \cdot 0,2$$



SE1: Prova objectiva individual. (Instruments d'avaluació: Ex1/Ex2)

SE2: Avaluació de pràctiques. (Instruments d'avaluació: Labs/LabEx)

SE3: Avaluació contínua. (Instruments d'avaluació: EjP/TC/altres)

En qualsevol cas: (1) si la nota SE1 té un valor menor a 5/10, NF serà igual a SE1; (2) si la nota SE2 té un valor menor a 5/10, NF serà igual a SE2; (3) Per a superar l'assignatura, NF ha de ser superior a 5/10 tant en primera com en segona convocatòria.

A continuació, es descriuen els diferents instruments d'avaluació:

Ex1/Ex2: Examen individual (SE1). Podrà contindre tant qüestions breus, com de desenvolupament de qüestions teoricopràctiques, problemes, casos d'estudi etc. Es podrà preguntar sobre qualsevol aspecte treballat durant el curs, també podran aparèixer nous problemes relacionats amb la matèria, en considerar-se aquesta una metodologia útil per a valorar la consolidació de les competències i continguts. Aquesta prova es realitzarà d'acord amb el calendari d'exàmens de l'escola, Ex1 correspon a la primera convocatòria i Ex2 a la segona.

La participació en Ex2 serà obligatòria sempre que no se supere l'assignatura en primera convocatòria, en cas contrari la nota en segona convocatòria serà de No Presentat, qualsevol excepció referent a això haurà de ser autoritzada pel professorat.

Competències més rellevants: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.

Labs: Laboratoris (SE2). Durant el curs, i preferiblement en equips de dues persones, es realitzaran pràctiques de laboratori. S'avaluarà tant el desenvolupament a l'aula, com els informes associats a cada pràctica que hauran de discutir els procediments emprats amb rigor i si escau respondre a les preguntes plantejades. A més, es podran sol·licitar tasques de preparació de la pràctica que podran ser igualment avaluables fins a un màxim d'un terç de la nota corresponent a cada pràctica. Per a tindre una nota associada a cada pràctica serà obligatòria l'assistència.

En primera convocatòria, SE2 estarà determinada per l'avaluació contínua dels laboratoris a través dels mètodes descrits anteriorment.

La còpia en qualsevol d'aquestes activitats serà penalitzada de manera estricta podent-se anul·lar totes les notes d'avaluació de laboratori. La no assistència de manera no justificada i reiterada a les sessions de laboratori suposarà una nota de laboratori en primera convocatòria de 0.

Competències més rellevants: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

LabEx: Examen de laboratori (SE2). L'alumnat que no supere en primera convocatòria la qualificació mínima en SE2 a través del mètode descrit anteriorment, haurà de realitzar un examen de laboratori. En aquest examen s'avaluarà el seu acompliment amb les eines utilitzades durant el curs i la seua capacitat per a interpretar resultats, entre altres competències associades als laboratoris. Aquest examen es realitzarà en segona convocatòria i d'acord amb el calendari oficial.

Competències més rellevants: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB4, CG02, CG03, CG07.

EjP: Exercici parcial individual (SE3). Durant el curs i en horari de classe, es realitzarà una prova per a avaluar la consolidació de continguts i competències, així com donar l'oportunitat a l'alumnat d'enfrontar-se a exercicis similars als que es podrà trobar en Ex1/Ex2. Els continguts aplicables a aquesta prova, així com les normes a seguir i la data es comunicarà durant el curs. En cap cas aquesta prova eliminarà matèria de cara a Ex1/Ex2. El pes d'aquest exercici serà d'un 10% de la nota final.

Competències més rellevants: CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.



TC: Tasques col·laboratives (SE3). L'avaluació contínua es veurà completada a través de diferents tasques col·laboratives que es podran plantejar tant per a la seua realització a l'aula com de forma no presencial. Aquestes tasques es realitzaran en equip, es treballarà entre altres la coordinació entre diferents membres d'un equip, la discussió per a aconseguir solucions de consens, l'anàlisi de bibliografia rellevant, etc. Es podran emprar tècniques d'avaluació contínua i per parells per a diferenciar les notes de diferents membres d'un equip. El pes d'aquest apartat serà d'un 10% de la nota final.

No es tindran en compte activitats o memòries entregades fora de termini, ni es podran recuperar activitats no realitzades.

Competències més rellevants: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.

Altres (SE3). El professorat es reserva la possibilitat d'afegir altres mètodes d'avaluació contínua com: observació diària, controls d'assistència i participació... que podran substituir/complementar a l'apartat TC.

En segona convocatòria, i atenent a les notes mínimes indicades anteriorment, es podrà calcular la nota final com: $NF = SE1 \cdot 0,6 + SE2 \cdot 0,4$

En qualsevol cas, el sistema d'avaluació es regirà pel que estableix el Reglament d'Avaluació i Qualificació de la Universitat de València per a Graus i Màsters.

REFERÈNCIES

Bàsiques

- Large Scale Machine Learning with Python, B. Sjoardin, A. Boscheti, L. Massaron.
- Big Data: Principles and Best Practices. N. Marz, J. Warren
- Spark, the definitive guide, B. Chambers, M. Zaharia

Complementàries

- Learning Spark, H. Karau, A. Konwinski, P. Wendell, M. Zaharia
- Learning PySpark, T. Drabas, D. Lee
- PySpark Cookbook, D. Lee, T. Drabas