

| Titulació                   | Tipus | Curs | Semestre |
|-----------------------------|-------|------|----------|
| 2503758 Enginyeria de Dades | OB    | 3    | 2        |

### Professor/a de contacte

Nom: Antonio Espinosa Morales

Correu electrònic: antoniomiguel.espinosa@uab.cat

### Utilització d'idiomes a l'assignatura

Llengua vehicular majoritària: espanyol (spa)

Grup íntegre en anglès: No

Grup íntegre en català: No

Grup íntegre en espanyol: No

### Equip docent extern a la UAB

Carles Carrillo

Pere Pons

Ramon Grau

Ramon Grau

### Prerequisits

Encara que no hi ha prerequisits formalment establerts i és responsabilitat de la pròpia assignatura proporcionar als i les alumnes un mitjà per adquirir els coneixements descrits a l'apartat de continguts de l'assignatura. És recomanable: un bon coneixement de programació, del funcionament d'un computador, del sistema operatiu a nivell d'usuari programador i dels sistemes de bases de dades relacionals i no relacionals

### Objectius

L'objectiu d'aquesta assignatura és conèixer els conceptes fonamentals dels Sistemes Distribuïts i els paradigmes de processament massiu de dades. Es veuran models de programació per aplicacions de processament "batch", en memòria i en "streaming". Arquitectures d'aplicacions i revisió de conceptes rellevants respecte al tractament de les dades com integritat, accesibilitat, fiabilitat, consistència i seguretat en el processament de dades a gran escala.

### Competències

- Concebre, dissenyar i implementar aplicacions eficients per a l'anàlisi i la gestió de dades massives.
- Concebre, dissenyar i implementar sistemes d'emmagatzematge de dades de forma eficient i segura.
- Planificar i gestionar el temps i els recursos disponibles.
- Que els estudiants tinguin la capacitat de reunir i interpretar dades rellevants (normalment dins de la seva àrea d'estudi) per emetre judicis que incloguin una reflexió sobre temes destacats d'indole social, científica o ètica.

## Resultats d'aprenentatge

1. Aplicar tècniques per automatitzar la resposta de les aplicacions a situacions dinàmiques (fiabilitat, escalabilitat, emergències...).
2. Desenvolupar aplicacions que processin dades a gran escala utilitzant paradigmes batch i streaming.
3. Planificar i gestionar el temps i els recursos disponibles.
4. Que els estudiants tinguin la capacitat de reunir i interpretar dades rellevants (normalment dins de la seva àrea d'estudi) per emetre judicis que incloguin una reflexió sobre temes destacats d'indole social, científica o ètica.

## Continguts

1. Introducció a les aplicacions massives de dades
2. Conceptes fonamentals del tractament de dades en entorns massius: fiabilitat, escalabilitat, sostenibilitat. Models de dades i llenguatges de consulta.
3. Gestió de grans volums de dades. Data warehousing. Principis dels sistemes Data warehousing, Business intelligence, modelització multidimensional, operadors OLAP, esquema en estrella, procés ETL
4. Introducció a les bases de dades a memòria amb Redis
5. Aplicacions amb grans volums de dades amb eines Apache. Introducció a les eines i ecosistema Apache Spark per a processar grans volums de dades. Utilització de Dataframes i MLib

## Metodologia

En el desenvolupament de l'assignatura es podran diferenciar quatre tipus d'activitats docents:

- Clases teòriques: exposició de la part teòrica de cada tema del programa. L'estructura típica d'una classe d'aquest tipus serà la següent: en primer lloc es farà una introducció on es presentaran breument els objectius de l'exposició i els continguts a tractar. A continuació es desgranaran els continguts objecte d'estudi, incloent exposicions narratives, desenvolupaments formals que proporcionin els fonaments teòrics, i intercalant exemples, que il·lustrin l'aplicació dels continguts exposats. Finalment, el/la professor/a exposarà les conclusions dels continguts. Durant tot el curs hi hauran avaluacions continuades de grups de temes.
- Sessions de laboratori. La part pràctica dels temes teòrics quedarà completada amb sessions al laboratori, on l'alumne desenvoluparà una sèrie de programes i haurà d'intentar resoldre problemes concrets que seran publicats al començar el temari. Alguns d'aquests exercicis s'hauran d'entregar al campus virtual en les dates especificades. Les pràctiques es desenvoluparan en grups de dos o tres alumnes. Les classes inclouen un conjunt de 6 sessions al laboratori, de 2 hores de durada, on l'alumne realitzarà el desenvolupament dels enunciats publicats.
- Problemes: revisió de casos pràctics. Al final de l'assignatura es presentarà una llista de casos pràctics amb reptes de dades relacionats amb objectius de negoci a realitzar. Els estudiants treballaran en grup per presentar les conclusions del seu estudi de forma oral a les sessions finals i es realitzarà una avaluació en grup dels treballs presentats.

Aquest plantejament del treball està orientat a promoure un aprenentatge actiu i a desenvolupar les competències de capacitat d'organització i planificació, comunicació oral i escrita, treball en equip i raonament crític. La qualitat dels exercicis realitzats, de la seva presentació i del seu funcionament es valorarà especialment.

Nota: es reservaran 15 minuts d'una classe, dins del calendari establert pel centre/titulació, per a la complementació per part de l'alumnat de les enquestes d'avaluació de l'actuació del professorat i d'avaluació de l'assignatura/mòdul.

## Activitats formatives

| Títol                     | Hores | ECTS | Resultats d'aprenentatge |
|---------------------------|-------|------|--------------------------|
| Tipus: Dirigides          |       |      |                          |
| Problemes                 | 9     | 0,36 | 1, 3, 4                  |
| Pràctiques                | 12    | 0,48 | 2, 3, 4                  |
| Teoria                    | 20    | 0,8  | 3, 4                     |
| Tipus: Autònomes          |       |      |                          |
| Estudi autònom            | 30    | 1,2  | 3, 4                     |
| Preparació de laboratoris | 32    | 1,28 | 2, 3, 4                  |
| Preparació de problemes   | 20    | 0,8  | 1, 3, 4                  |

## Avaluació

L'objectiu del procés d'avaluació és verificar que l'alumne ha assolit els coneixements i habilitats definits en els objectius de l'assignatura, així com les competències.

S'avaluaran quatre tipus d'activitats de manera independent on la suma ponderada d'elles donarà la nota final. Aquestes activitats són:

Teoria (T)

Resolució de les pràctiques de laboratori (PL)

Realització d'un informe sobre un cas pràctic d'estudi (PA).

La part de Teoria (T) s'avaluarà amb dos controls parcials individuals durant tot el curs. La nota final de Teoria sortirà de la suma ponderada dels dos controls ( $0.5 \cdot \text{Control 1} + 0.5 \cdot \text{Control 2}$ ). Hi haurà una segona oportunitat per recuperar aquesta part el dia que tenim assignat a la setmana de exàmens de juny. Es podran recuperar les parts que no hagin estat superades en els controls parcials de teoria. La nota mínima per aprovar aquesta part és  $\geq 4,5$ . Cada parcial ha de tenir una nota mínima de 4 per poder fer promig.

La part de Resolució de pràctiques de laboratori (PL) s'avaluarà de manera grupal. Estan planificats 4 lliuraments. La nota final sortirà de la suma ponderada dels lliuraments. Per aprovar les PL la nota mínima haurà de ser  $\geq 4,5$ . Només hi ha una única oportunitat (no es pot recuperar aquesta part). És obligatori presentar tots els entregables de pràctiques per aprovar l'assignatura.

Els casos d'estudi es presentaran de forma oral i s'avaluaran de forma creuada per tota la classe i pel professor. El valor d'aquests exercicis és del 10% de la nota final i donada la seva naturalesa i objectiu no són recuperables.

La nota final de l'assignatura serà la suma ponderada de les notes de cadascuna de les quatre activitats: 60% de Teoria, 10% Casos d'estudi i 30% de Resolució de pràctiques de laboratori. El resultat haurà de ser  $\geq 5$ .

En cas de no superar l'assignatura per no arribar a la puntuació mínima en algun dels apartats (Teoria o Pràctiques de Laboratori), tot i que al fer la mitjana ponderada la nota final fos igual o superior a 5 la nota que es posarà a l'expedient serà de 4,5.

En cas de que la mitjana no arribi a 5 la nota que figurarà a l'expedient serà la nota mitjana obtinguda numèricament.

Si l'alumne entrega qualsevol activitat, s'entén que es presenta a l'assignatura i serà avaluat/da. Si no entrega cap activitat, es pot considerar No evaluable.

Atorgar una qualificació de matrícula d'honor és decisió del professorat responsable de l'assignatura. La normativa de la UAB indica que les MH només es podran concedir a estudiants que hagin obtingut una qualificació final igual o superior a 9.00. Es pot atorgar fins a un 5% de MH del total d'estudiants matriculats.

Les dates d'avaluació continuada i lliurament de treballs es publicaran al campus virtual i poden estar subjectes a possibles canvis de programació per motius d'adaptació a possibles incidències; sempre s'informarà al campus virtual sobre aquests canvis ja que s'entén que el CV és el mecanisme habitual d'intercanvi d'informació entre el professorat i els/les estudiants.

Per a cada activitat d'avaluació, s'indicarà un lloc, data i hora de revisió en la que l'estudiant podrà revisar l'activitat amb el/la professor/a. En aquest context, es podran fer reclamacions sobre la nota de l'activitat, que seran avaluades pel professorat responsable de l'assignatura. Si l'estudiant no es presenta a aquesta revisió, no es revisarà posteriorment aquesta activitat.

Nota sobre plagis:

Sense perjudici d'altres mesures disciplinàries que s'estimin oportunes, i d'acord amb la normativa acadèmica vigent, les irregularitats comeses per un o una estudiant que puguin conduir a una variació de la qualificació en una activitat evaluable es qualificaran amb un zero (0). Les activitats d'avaluació qualificades d'aquesta forma i per aquest procediment no seran recuperables. Si és necessari superar qualsevol d'aquestes activitats d'avaluació per aprovar l'assignatura, aquesta assignatura quedarà suspesa directament, sense oportunitat de recuperar-la en el mateix curs. Aquestes irregularitats inclouen, entre d'altres:

la còpia total o parcial d'una pràctica, informe, o qualsevol altra activitat d'avaluació; deixar copiar; presentar un treball de grup no fet íntegrament pels i les membres del grup (aplicat a tots els i les membres, no solament als que no han treballat); presentar com a propis materials elaborats per un tercer, encara que siguin traduccions o adaptacions, i en general treballs amb elements no originals i exclusius de l'estudiant; tenir dispositius de comunicació (com telèfons mòbils, smart watches, bolígrafs amb càmera, etc.) accessibles durant les proves d'avaluació teórico-pràctiques individuals (exàmens); parlar amb companys o companyes durant les proves d'avaluació teórico-pràctiques individuals (exàmens); copiar o intentar copiar d'altres alumnes durant les proves d'avaluació teórico-pràctiques (exàmens); usar o intentar usar escrits relacionats amb la matèria durant la realització de les proves d'avaluació teórico-pràctiques (exàmens), quan aquests no hagin estat explícitament permesos.

En cas de no superar l'assignatura degut a que alguna de les activitats d'avaluació no arriba a la nota mínima requerida, la nota numèrica de l'expedient serà el valor menor entre 4.5 i la mitjana ponderada de les notes. Amb les excepcions de que s'atorgarà la qualificació de "No Avaluable" als i les estudiants que no participin en cap de les activitats d'avaluació, i de que la nota numèrica de l'expedient serà el valor menor entre 3.0 i la mitjana ponderada de les notes en cas que l'estudiant hagi comès irregularitats en un acte d'avaluació (i per tant no serà possible l'aprobat per compensació). En edicions futures d'aquesta assignatura, a l'estudiant que hagi comès irregularitats en un acte d'avaluació no se li convalidarà cap de les activitats d'avaluació realitzades.

En resum: copiar, deixar copiar o plagiar (o l'intent de) en qualsevol de les activitats d'avaluació equival a un SUSPENS, no compensable i sense convalidacions de parts de l'assignatura en cursos posteriors.

## Activitats d'avaluació

| Títol                        | Pes | Hores | ECTS | Resultats d'aprenentatge |
|------------------------------|-----|-------|------|--------------------------|
| Control individual parcial 1 | 30% | 2     | 0,08 | 1, 3, 4                  |
| Control individual parcial 2 | 30% | 2     | 0,08 | 1, 3, 4                  |
| Estudi de casos pràctics     | 10% | 5     | 0,2  | 1, 3, 4                  |

## **Bibliografia**

Designing Data intensive applications - Martin Kleppmann, O'Reilly, 2017

The Data warehouse ETL toolkit - Ralph Kimball, Joe Caserta. Wiley, 2004

Spark, the definitive guide, Big data processing made simple. Bill Chambers and Matei Zaharia, O'Reilly, 2018

Learning Spark - Lightning fast data analysis - Holden Karau, Andi Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly, 2015

Beginning Scala - Layka, Vishal. Apress; 2nd ed. 2015.

Redis in Action - Josiah L. Carlson. Manning, 2013.

## **Programari**

Talend Open Studio for Data Integration

Redis

Apache Spark

Jupyter Notebook

Ubuntu Linux