# ANÀLISI DE DADES AMB BIG DATA:

## *Fonaments, gestió i base de dades NoSQL*

Universitat Oberta de Catalunya

Amb el conveni de col·laboració del Consorci per a la Formació Contínua de Catalunya

# 1. Objectius del curs

- Entendre la situació actual del Big Data en varis escenaris de mercat.
- Entendre el flux de processament de les dades en entorns Big Data.
- Entendre l'entorn Hadoop, Spark en processament Batch i Streaming.
- Conèixer els algoritmes clàssics d'Intel·ligència Artificial proporcionats en un entorn Big Data.
- Aprendre a presentar informació utilitzant tecnologies Big Data.
- Aprendre a dissenyar un projecte Big Data i entendre la seva problemàtica.

# 2. Criteris d'accés dels participants.

Donat que els objectius del curs són avançats el curs queda restringit a participants que estiguin en possessió de títol de llicenciat, graduat, diplomat i enginyer. En aquest cas, el camp del Big Data és altament multidisciplinari i per tant no es limita l'accés dels participants segons àrea de formació.

# 3. Criteris de selecció dels participants

La selecció de participants al curs consta de en 4 etapes: prova d'accés, Currículum Vitae, entrevista personal i disponibilitat horària per la realització del curs.

- **Prova d'accés.** La prova d'accés consta d'entre 20 i 30 preguntes de conceptes bàsics necessaris per la realització del curs. Entre ells, coneixements bàsics d'arquitectura de computadors, programació, paral·lelisme i xarxes d'ordenadors.
- **Currículum Vitae.** Valoració personalitzada del currículum.
- **Entrevista Personal.**
- **Disponibilitat horària.** Donada la naturalesa del curs i la durada d'aquest, és molt important que els alumnes disposin de la suficient dedicació horària. Com es descriu a posteriori en aquest document el curs consta de 150 hores i estarà distribuït en 5 mesos aproximadament. Això correspon a una dedicació aproximada d'entre 5 hores setmanals..

L'orde de prelació es decidirà bàsicament en funció dels punts anteriors. Es prioritzaran els participant estigui actualment empleats i en particular en negocis amb necessitats de tractament de dades Big Data.

# 4. Inici i durada del curs

Hores totals: 150 hores,

Data inici 11/12/19
Data fi: 13/05/20

NOTA: Durant la setmana del 25 de desembre a l'1 de gener no es realitzarà formació.

Distribució d'hores: 6 hores presencials,  145 hores de tele-formació

Les 5 hores presencials es repartiran en primer LAB i en l'últim. El LAB 1, que constarà de 5 hores, està destinat a la presentació del curs, instal·lació de material i descripció de l'entorn a utilitzar. En el LAB 14 es presentarà el projecte realitzar per cada alumne. Tots els alumnes assistiran a la sessió.

# 5. Com apuntar-se

Per apuntar-se al curs s'ha d'enviar una sol·licitud al correu corporate@uoc.edu adjuntant un CV i breu carta de motivació. Seguidament, s'informarà del dia de la prova i l'entrevista personal. S'acceptaran registres fins el 29 de Novembre del 2019.

# 6. Cost del curs

El curs està totalment subvencionat pel *Consorci per a la Formació Contínua de Catalunya*.

# 7. Espais formatius, instal·lacions i equipaments

## a. Espai formatiu

El curs proposat es realitzarà principalment on-line a través del sistema d'aules i recursos de la UOC.

## b. Equipament

Tal com s'especifica a continuació el curs es realitzarà utilitzant la versió educativa de l'entorn Cloudera QuickStarts (https://es.cloudera.com/) que permet instal·lar la totalitat del software que requereix el curs amb facilitat. El sistema requereix d'un mínim de 8GB de RAM, un processador i5 o superior i 512 GB de disc per a realitzar les activitats i projectes proposat al curs. El curs es pot realitzar sense problemes amb sistemes operatius Windows, Linux o MacOS.

# 8. Mòduls formatius

Els diferents mòduls del curs han estat dissenyats per assolir competències en la major part de tecnologies que s'utilitzen actualment en entorns Big Data. El curs assumeix que el participant té coneixements bàsics de programació i per tant la primera part del curs està orientada a repassar i ampliar aquest coneixements per tal de poder desenvolupar el curs amb normalitat. En cas que els participants ja tinguessin les competències de programació necessàries assolides, les hores del LAB 2 es traslladarien a la LAB 14 (realització d'un projecte Big Data). Tal com s'especifica a continuació cada sessió de laboratori (LAB) està composta d'un vídeo explicatiu (Webinnar offline d'entre 1 i 2 hores), una lectura de continguts i una avaluació dels material explicats en el laboratori.

# 9. Resum LABs proposats

LAB 1: Introducció al desenvolupament en Big Data [5 hores, virtual]
LAB 2: Programació en Python: eines avançades i anàlisis de dades [5 hores]
LAB 3: Infraestructures distribuïdes per a Big data: Apache Hadoop [15 hores]
LAB 4: Streaming de dades en alt rendiment: Apache Flume i Kafka [15 hores]
LAB 5: Projecte de Big Data. Infraestructures distribuïdes per a Big Data [15 hores]
LAB 6: Processament distribuït de dades en entorns Big data amb Spark [25 hores]
LAB 7  Visualització de dades y dashboards [10 hores]
LAB 8: Processament distribuït de dades i visualització en Big Data [15 hores]
LAB 9  Anàlisi de dades i aprenentatge automàtic amb Spark [15 hores]
LAB 10:Projecte de Big Data. Anàlisi de dades en Big Data i projecte final [30 hores]

# 10. Planificació de les diferents LABs

LAB 1: 1a setmana
LAB 2: 2a setmana
LAB 3: 3a i 4a setmana
LAB 4: 5a i 6a setmana
LAB 5: 7a i 8a setmana
LAB 6: 9a-11a setmana
LAB 7: 12a i 13 setmana
LAB 8: 14a i 15a setmana
LAB 9: 16a i 17a setmana
LAB 10: 18a-21a setmana

# 11. Descripció detallada del les diferents LABs

**LAB 1: Introducció al desenvolupament en Big Data [5 hores, virtual]**

- Presentació del curs
- Presentació dels alumnes
- Presentació dels continguts del curs
- Presentació de l'entorn formatiu de la UOC

- Instal·lació i validació de les eines de Big Data (software) que es faran servir durant el curs: Python 3.x, Editor i espai de treball, Cloudera Docker, Git

**LAB 2: Programació en Python: eines avançades i anàlisis de dades [5 hores]**

- Revisió llenguatge Python. Introducció, literals, operadors, variables, entrada/sortida bàsica, Condicions, iteracions, operacions, lògiques i de bit, llistes i col·leccions, ordenació, funcions, tuplas i diccionaris, mòduls, paquets, cadenes. [1 hora]
- Revisió llibreries avançades per a anàlisi de dades (Pandas, etc.) [1 hora]
- Presentació del projecte i anàlisis inicial [3 hores]

**LAB 3: Infraestructures distribuïdes per a Big data: Apache Hadoop [15 hores]**

- Lectura de materials [7 hores]
- Introducció a Hadoop i al seu ecosistema
- Arquitectura Hadoop i HDFS
- Importació de dades relacionals amb Apache Sqoop
- Introducció a Impala i Hive
- Modelatge de dades i gestió amb Impala i Hive
- Formats de dades distribuïdes
- Particionat de dades
- Realització d'activitats [7 hores]
- Pràctiques de laboratori amb les Docker Cloudera
- Lab1 i 2. Introducció a Apache Hadoop
- Lab3. Arquitectura Hadoop i HDFS
- Lab4. Importació de dades relacionals amb Apache Sqoop
- Lab5. Introducció a Impala i Hive
- Lab6. Modelatge de dades i gestió amb Impala i Hive
- Lab7. Formats de dades distribuïdes
- Lab8. Particionat de dades
- Auto-avaluació [1 hora]

**LAB 4: Streaming de dades en alt rendiment: Apache Flume i Kafka [15 hores]**

- Webinar offline [1 hora]
- Captura de dades en streaming amb Flume
- Ús de Apache Kafka
- Lectura de materials [8 hores]
- Ingestió de dades en Streaming
- Ús d'Apache Kafka
- Realització d'activitats [5 hores]
- Pràctiques de laboratori amb les VM Cloudera
- Lab9. Ingestió de dades en Streaming
- Lab10. Apache Spark
- Avaluació [1 hora]
- Correcció de les pràctiques del laboratori de Big Data de Cloudera

**LAB 5: Projecte de Big Data. Infraestructures distribuïdes per a Big Data [15 hores]**

- Lectura de materials [1 hores]
- Estudi de les rúbriques del projecte
- Ús de Cloudera Academy
- Realització d'activitats [12 hores]
- Realitzar un projecte de creació infraestructura per a Big Data amb càrrega i distribució de dades de diferents orígens.
- Avaluació [1 hora]
- Feedback de correcció del projecte realitzat.
- Avaluació i co-avaluació.

**LAB 6: Processament distribuït de dades en entorns Big data amb Spark [25 hores]**

- Webinar [1 hores, semi-presencial]
- Introducció a Apache Spark: RDD i agregació de dades
- Desenvolupament d'aplicacions amb Spark
- Lectura de materials [11 hores]
- Introducció a Spark
- Treballant amb RDDs a Spark
- Agregació de dades a RDDs Parells
- Desenvolupament d'aplicacions Spark
- Processament Paral·lel amb Spark
- Persistència de RDD amb Spark
- Realització d'activitats [12 hores]
- Lab10. Introducció a Spark
- Lab11. Treballant amb RDDs a Spark
- Lab12. Agregació de dades a RDDs Parells
- Lab13. Desenvolupament d'aplicacions Spark
- Lab14. Processament Paral·lel amb Spark
- Lab15. Persistència de RDD amb Spark
- Avaluació [1 hora]
- Correcció de les pràctiques del laboratori de Big Data de Cloudera

**LAB 7: Visualització de dades y dashboards [10 hores]**

- Webinar offline [1 hores]
- Aprenentatge de Visualització de dades
- Ús de Plotly amb Python i Jupyter NB
- Kibana, visualització de dades massives
- Grafana. gràfics en temps reals
- Lectura de materials [8 hores]
- Aprenentatge de Visualització de dades
- Manual de Plotly
- Documentació oficial de Kibana

- Documentació oficial de Grafana
- Realització d'activitats [3 hores]
- Exercicis amb Pyhton, Jupyter i Plotly
- Kibana, visualització de dades massives
- Grafana. gràfics en temps reals
- Avaluació [1 hora]
- Correcció dels exercicis proposats

**LAB 8: Processament distribuït de dades i visualització en Big Data [15 hores]**

- Webinar offline [1/2 hora]
- Presentació del projecte
- Lectura de materials [1 1/2 hores]
- Estudi de les rúbriques del projecte
- Ús de Cloudera Academy
- Realització d'activitats [12 hores]
- Realitzar un projecte de creació infraestrura per a Big Data amb creació de RDD, accions, transformacions amb Spark, i visualització de dades.
- Avaluació [1 hora]
- Feedback de correcció del projecte realitzat.
- Avaluació i coavaluació.

**LAB 9: Anàlisi de dades i aprenentatge automàtic amb Spark [15 hores]**

- Webinars [1 hores]
- Aprenentatge de SparkQL
- Aprenentatge de SparkML
- Lectura de materials de programació [11 hores]
- Lab 16. Algorismes de Spark. Spark ML
- Lab 17. SparkQL i DataFrames
- Realització d'activitats [2 hores]
- Lab 16. Algorismes de Spark. Spark ML
- Lab 17. SparkQL i DataFrames
- Avaluació [1 hora]
- Correcció d'exercicis

**LAB 10: Projecte de Big Data. Anàlisi de dades en Big Data i projecte final [30 hores]**

- Webinar offline [1hora]
- Presentació del projecte. Segons la dinàmica dels participants el projecte serà en grup o individual.
- Realització d'activitats [28 hores]
- Desenvolupament del projecte
- Redacció de la memòria
- Preparació de la presentació
- Avaluació i defensa del projecte final [1h, presencial]

- Presentació del projecte
- Avaluació i co-avaluació.