Guía didáctica

ASIGNATURA: Procesamiento de datos masivos

Título: *Máster Universitario en Big Data y Ciencia de Datos*
Materia: *Big Data*
Créditos: *6 ECTS*
Código: *03MBID*
Curso: Octubre *I 2024*

Índice

1. Organización general

1.1. Datos de la asignatura

MATERIA	***Big Data***
ASIGNATURA	*Procesamiento de datos masivos* **6 ECTS**
Carácter	obligatoria
Cuatrimestre	Primero
Idioma en que se imparte	Castellano
Requisitos previos	No existen
Dedicación al estudio por ECTS	**25 horas**

1.2. Equipo docente

Profesores	**Grupo A:** **Dr. Óscar Garibo Orts** *oscar.garibo@professor.universidadviu.com* **Grupo B:** **Dr. Óscar Garibo Orts** *oscar.garibo@professor.universidadviu.com*

1.3. Introducción a la asignatura

Esta asignatura permite al alumnado especializarse en el procesamiento masivo de datos en entornos basados en Internet y la nube (*cloud computing*). En los últimos años se ha generado una gran cantidad de información gracias a Internet. Esta información en muchos casos es compartida y procesada por personas, pero también por organizaciones en la Industria 4.0 y robots (sensores, servidores, etc.) en el Internet de las cosas (IoT).

Cuando se tienen datos masivos, las tecnologías tradicionales no son capaces de procesar la información y se necesita utilizar tecnologías más novedosas denominadas Big Data. Estas tecnologías suelen ejecutarse de forma distribuida sobre un *cluster* de computadores que se puede desplegar en un *data center* o adquirirlo como un servicio en la nube.

El objetivo principal de esta materia es dar a conocer diferentes técnicas de procesamiento de grandes cantidades de información tanto en sistemas locales como en la nube, instruyendo al alumno en su utilización para el procesamiento del denominado Big Data, principalmente, a través del modelo de procesamiento MapReduce y los ecosistemas Hadoop, Spark y Flink.

Objetivos generales:

Los objetivos propios de la asignatura Procesamiento de datos masivos son:

- Conocer los mecanismos de virtualización ligera y el uso de contenedores.
- Conocer los principales modelos del Cloud Computing, así como los principales proveedores *cloud* públicos y herramientas para el despliegue de nubes privadas/híbridas.
- Ser capaz de desplegar un *cluster* virtual para el procesamiento del Big Data usando Hadoop.
- Conocer el modelo de programación MapReduce e implementar algoritmos para resolver problemas simples en los principales *frameworks*, Hadoop MapReduce, Spark y Flink.
- Utilizar herramientas y lenguajes de alto nivel para el procesamiento masivo de datos en Hadoop y Spark.
- Utilizar Apache Spark como motor de análisis unificado para el procesamiento masivo de datos, combinando SQL, procesamiento de flujos y analítica compleja.

1.4. Competencias y resultados de aprendizaje

COMPETENCIAS GENERALES

CG.1.- Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo o aplicación de ideas, a menudo en un contexto de investigación

CG.2.- Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CG.3.- Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

CG.4.- Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

COMPETENCIAS ESPECÍFICAS DE LA ASIGNATURA

CE01- Conocer los fundamentos de la ingeniería de datos masivos para su modelado, gestión, procesamiento y análisis.

CE02- Utilizar técnicas y herramientas de programación especializada en analítica y procesamiento de datos en entornos de Big Data.

CE03- Aplicar diferentes modelos de almacenamiento de datos masivos, así como sistemas de bases de datos, para su procesamiento en infraestructuras distribuidas.

CE04- Resolver problemas reales de clasificación, modelización e interpretación de un conjunto de datos en el contexto de Big Data.

CE05- Entender las implicaciones legales, morales y éticas en lo referente al uso de datos personales en el contexto de Big Data.

CE14- Analizar los resultados de los modelos de análisis de datos en un contexto de toma de decisiones.

CE15- Identificar la solución Big Data óptima para un problema, en términos de eficiencia, eficacia e interpretación de resultados.

CE16- Diseñar estrategias de visualización de resultados y presentación de conclusiones obtenidos en el proceso de análisis de datos en un contexto de toma de decisiones.

RESULTADOS DE APRENDIZAJE

Al finalizar esta asignatura se espera que el estudiante sea capaz de:

RA.2.- Conocer los principios básicos de las tecnologías más utilizadas en la implementación de los diferentes tipos de análisis.

RA.3.- Implementar las tecnologías más utilizadas para el procesamiento de los diferentes tipos de análisis.

2. Contenidos/temario

Tema 1: Introducción al *Cloud Computing*

1.1.- Computación Ubicua

1.2.- Computación en la nube: definiciones

1.3.- Definición, modelos de servicio y modelos de despliegue

1.4.- Infraestructura *cloud*

1.5.- Tipos de nubes

1.6.- Despliegue de aplicaciones Big Data

Tema 2: Introducción al Big Data

2.1.- Definición y conceptos sobre Big Data

2.2. Modelos de datos y procesamiento en Big Data

2.3.- Modelo de procesamiento MapReduce

2.4.- Principales *frameworks* en Big Data

2.5.- Patrones de diseño en Big Data

Tema 3: Ecosistema Hadoop: Procesamiento distribuido del Big Data con Hadoop.

3.1.- Arquitectura de Hadoop: almacenamiento con HDFS, gestión de recursos con YARN.

3.2.- Programación MapReduce en Hadoop

3.3.- Hadoop Streaming

3.4.- Tecnologías de alto nivel

Tema 4: Ecosistema Spark: Procesamiento *in-memory*

4.1. Persistencia en memoria y disco

4.2.- Introducción y operaciones básicas en Spark

4.3.- Trabajando con DataFrames y DataSets

4.4.- Introducción a los RDD

4.5.- Programación en *streaming*

4.6.- Extensiones de Spark: Spark MLlib, R sobre Spark, GraphX

4.7. Visualización de datos y tecnologías de alto nivel

Tema 5: Ecosistema Flink: Introducción al procesamiento streaming con Apache Flink

5.1.- Arquitectura de Flink

5.2.- Procesamiento de datos masivos 5.3.- Tecnologías de alto nivel

3. Metodología

La metodología de la Universidad Internacional de Valencia (VIU) se caracteriza por una apuesta decidida en un modelo de carácter e-presencial. Así, siguiendo lo estipulado en el calendario de actividades docentes del Título, se impartirán en directo un conjunto de sesiones, que, además, quedarán grabadas para su posterior visionado por parte de aquellos estudiantes que lo necesitasen. En todo caso, se recomienda acudir, en la medida de lo posible, a dichas sesiones, facilitando así el intercambio de experiencias y dudas con el docente.

En lo que se refiere a las metodologías específicas de enseñanza-aprendizaje, serán aplicadas por el docente en función de los contenidos de la asignatura y de las necesidades pedagógicas de los estudiantes. De manera general, se impartirán contenidos teóricos y, en el ámbito de las clases prácticas se podrá realizar la resolución de problemas, el estudio de casos y/o la simulación.

Por otro lado, la Universidad y sus docentes ofrecen un acompañamiento continuo al estudiante, poniendo a su disposición foros de dudas y tutorías para resolver las consultas de carácter académico que el estudiante pueda tener. Es importante señalar que resulta fundamental el trabajo autónomo del estudiante para lograr una adecuada consecución de los objetivos formativos previstos para la asignatura.

4. Actividades formativas

Durante el desarrollo de cada una de las asignaturas se programan una serie de actividades de aprendizaje que ayudan a los estudiantes a consolidar los conocimientos trabajados. A continuación, se relacionan las actividades que forman parte de la asignatura:

1. ***1. Clases expositivas***

Sesiones dedicadas al desarrollo de los contenidos mediante una metodología de lección magistral. El profesor expone los contenidos de forma que el alumno pueda participar en dicho espacio para interactuar y realizar cuestiones.

2. ***2. Sesiones con expertos en el aula***

Participación de expertos en la materia dedicadas a ofrecer sus experiencias profesionales en el ámbito de la materia de estudio, ya sea en diseño de proyectos o en cuestiones técnicas.

3. Observación y evaluación de recursos didácticos audiovisuales

Los estudiantes visualiza recursos didácticos audiovisuales como complemento a las sesiones del profesor. El docente dispone al alumnado los recursos para después evaluar con una prueba.

4. Estudio y seguimiento de material interactivo

Los estudiantes disponen de un documento sobre los contenidos de la asignatura en un formato multimedia que complementan el texto base.

5. Clases prácticas sobre

El profesor diseña una serie de sesiones prácticas aplicando diferentes metodologías como el estudio de casos, resolución de problemas, simulación, trabajo cooperativo, diseño de proyectos y

trabajo práctico sobre el uso de herramientas propias del Big Data y la Ciencia de Datos. En dichas sesiones el alumno puede recibir comentarios sobre los ejercicios propuestos.

6. Prácticas observacionales

El profesor propone una serie de recurso audiovisuales complementarios para que los alumnos puedan observarlos y así adquirir conocimientos que puedan ser aplicados en otras actividades de carácter práctico.

7. Actividades de seguimiento de la asignatura

El profesor puede proponer en las sesiones programadas una serie de actividades como la exposición de trabajos sobre un tema y tratar de comentar sobre la participación de los alumnos. También proponer el desarrollo de resúmenes, mapas conceptuales, *one minute paper*, test de autoevaluación…, o, también, un espacio de reflexión sobre el aprendizaje y el análisis crítico.

8. Tutorías

Los estudiantes pueden solicitar tutorías por correo para la orientación sobre dudas de la asignatura.

9. Lectura, análisis y estudio del manual de la asignatura

Trabajo autónomo del estudiante sobre el estudio del texto base.

10. Lectura, análisis y estudio de material complementario

Trabajo autónomo del estudiante sobre el estudio de otros materiales dispuestos por el profesor.

11. Desarrollo de actividades del portafolio

Trabajo autónomo del estudiante sobre el desarrollo de actividades prácticas evaluables.

12. Trabajo cooperativo

Los estudiantes trabajan conjuntamente para la resolución de actividades propuestas en clase, del portafolio, con el objetivo de fomentar la interacción y el trabajo en grupo.

13. Prueba objetiva final

Como parte de la evaluación de cada una de las asignaturas (a excepción de las prácticas y el Trabajo fin de título), se realiza una prueba (examen final). Esta prueba se realiza en tiempo real (con los medios de control antifraude especificados) y tiene como objetivo evidenciar el nivel de adquisición de conocimientos y desarrollo de competencias por parte de los estudiantes. Esta actividad, por su definición, tiene carácter síncrono.

El examen constará de alrededor de 20 preguntas de tipo test, junto con 5 a 10 preguntas de respuestas cortas. El tiempo establecido para el examen será de 90 minutos.

5. Planificación de las sesiones

Sesión	Fecha	Contenido/Tema
SESIÓN 1	18/11/2024	Tutoría colectiva. Tema 1: Introducción al *Cloud Computing* Òscar Garibo
SESIÓN 2	20/11/2024	Tema 1: Introducción al *Cloud Computing* Tema 2: Introducción al Big Data Òscar Garibo
SESIÓN 3	25/11/2024	Tema 2: Introducción al Big Data Tema 3: Ecosistema Hadoop Òscar Garibo
SESIÓN 4	27/11/2024	Tema 3: Ecosistema Hadoop Òscar Garibo
SESIÓN 5	02/12/2024	Tema 3: Ecosistema Hadoop Actividad guiada Òscar Garibo
SESIÓN 6	04/12/2024	Tema 3: Ecosistema Hadoop Actividad guiada Òscar Garibo
SESIÓN 7	09/12/2024	Tema 4: Ecosistema Spark Òscar Garibo
SESIÓN 8	11/12/2024	Tema 4: Ecosistema Spark Òscar Garibo
SESIÓN 9	13/12/2024	Tema 4: Ecosistema Spark Actividad guiada Òscar Garibo
SESIÓN 10	16/12/2024	Tema 4: Ecosistema Spark Actividad guiada Òscar Garibo
SESIÓN 11	18/12/2024	Tema 4: Ecosistema Spark Actividad guiada Òscar Garibo
SESIÓN 12	08/01/2025	Tema 5: Ecosistema Flink Òscar Garibo
SESIÓN 13	10/01/2025	Resumen Final: Ecosistema Big Data. Debate en grupos. Tutoría colectiva. Òscar Garibo.

NOTA: El horario es único de 20:00 a 22:00

Fechas de realización de la prueba		
Franja	**A**	**B**
1ª Convocatoria	Lunes 13 de enero de 2025 de 12:00 a 14:00	Lunes 13 de enero de 2025 de 20:00 a 22:00
2ª Convocatoria	Miércoles 26 de marzo de 2025 de 12:00 a 14:00	Miércoles 26 de marzo de 2025 de 20:00 a 22:00

Fechas de entrega del portafolio	
1ª Convocatoria	**Actividad 1:** Recomendada: Viernes 20 de diciembre de 2024 hasta las 23:59 Fecha última: Lunes 13 de enero de 2025 hasta las 23:59 **Actividad 2:** Recomendada: Lunes 30 de diciembre de 2024 hasta las 23:59 Fecha última: Lunes 13 de enero de 2025 hasta las 23:59 **Foro debate:** Recomendada: Lunes 9 de diciembre de 2024 hasta las 23:59 Fecha última: Lunes 9 de diciembre de 2024 hasta las 23:59 **Solamente disponible en 1era convocatoria** **Test autoevaluación (vídeo):** **Solamente disponible**: Lunes 2 de diciembre de 2024 hasta las 23:59 **Solamente disponible en 1era convocatoria**
2ª Convocatoria	**Actividad 1, Actividad 2 (no hay test ni foro debate):** Miércoles 26 de marzo de 2025 hasta las 23:59

Atención: Para aprobar el portafolio, además de obtener un mínimo de 50% de aprobado, debe haber aprobado la Actividad 1 y la Actividad 2.

6. Evaluación

6.1. Sistema de evaluación

El Modelo de Evaluación de estudiantes en la Universidad se sustenta en los principios del Espacio Europeo de Educación Superior (EEES), y está adaptado a la estructura de formación virtual propia de esta Universidad. De este modo, se dirige a la evaluación de competencias.

Sistema de Evaluación	Ponderación
Portafolio*	**60 %**
Actividad 1 (20%): Implementar programas Big Data utilizando el framework Hadoop **Actividad 2** (20%): Implementar programas Big Data utilizando el framework Apache Spark **Foro debate**** (10%): Participar en el sesión de debate en grupo sobre un tema relacionado con la asignatura	

Actividad Evaluación Continua** (10%): Test autoevaluación sobre vídeo relacionado a un tema relacionado a la asignatura	
Sistema de Evaluación	**Ponderación**
Prueba final*	**40 %**
Prueba sumativa y final teórico-práctica (preguntas abiertas, preguntas de prueba objetiva, examen truncado, etc.). El examen constará de preguntas de tipo test, junto con preguntas de respuestas cortas. El tiempo establecido para el examen será de 90 minutos.	

***Es requisito indispensable para superar la asignatura aprobar cada apartado (portafolio y prueba final)** con un mínimo de 5 para ponderar las calificaciones. **Además, en el portafolio debe tener aprobadas la Actividad 1 y la Actividad 2.**

**** Solamente disponibles en primera convocatoria**

Los enunciados y especificaciones propias de las distintas actividades serán aportados por el docente, a través del Campus Virtual, a lo largo de la impartición de la asignatura.

Atendiendo a la Normativa de Evaluación de la Universidad, se tendrá en cuenta que la utilización de **contenido de autoría ajena** al propio estudiante debe ser citada adecuadamente en los trabajos entregados. Los casos de plagio serán sancionados con suspenso (0) de la actividad en la que se detecte. Asimismo, el uso de **medios fraudulentos durante las pruebas de evaluación** implicará un suspenso (0) y podrá implicar la apertura de un expediente disciplinario.

6.2. Sistema de calificación

La calificación de la asignatura se establecerá en los siguientes cómputos y términos:

Nivel de aprendizaje	Calificación numérica	Calificación cualitativa
Muy competente	9,0 - 10	Sobresaliente
Competente	7,0 - 8,9	Notable
Aceptable	5,0 -6,9	Aprobado
Aún no competente	0,0 -4,9	Suspenso

Sin detrimento de lo anterior, el estudiante dispondrá de una **rúbrica simplificada** en el aula que mostrará los aspectos que valorará el docente, como así también los **niveles de desempeño que tendrá en cuenta para calificar las actividades vinculadas a cada resultado de aprendizaje.**

La mención de «**Matrícula de Honor**» podrá ser otorgada a estudiantes que hayan obtenido una calificación igual o superior a 9.0. Su número no podrá exceder del cinco por ciento de los estudiantes matriculados en una materia en el correspondiente curso académico, salvo que el número de estudiantes matriculados sea inferior a 20, en cuyo caso se podrá conceder una sola «Matrícula de Honor.

7. Bibliografía

7.1. Bibliografía de referencia

- Erl T., Puttini R. & Mahmood Z. (2013). Cloud Computing, Concepts, Technology & Architecture. Ed. Prentice-Hall.
- White, T. (2015). Hadoop: The Definitive Guide, Storage and Analysis at Internet Scale, 4a edición. O'Reilly.
- Chambers B. & Zaharia, M. (2018). Spark: The Definitive Guide: Big Data Processing Made Simple, 1a edición. O'Reilly.

7.2. Bibliografía complementaria

- Hueske F. & Kalavri V. (2019). Stream Processing with Apache Flink, 1a edición. O'Reilly.
- Zaharia M., Karau H., Konwinski, A. & Wendell, P. (2015). Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly.
- Karau, H. & Warren, R. (2017). High Performance Spark: Best practices for scaling and optimizing Apache Spark, 1a edición. O'Reilly
- Murthy, A. C., Vavilapalli, V. K., Eadline, D., & Markham, J. (2014). Apache Hadoop YARN: moving beyond MapReduce and batch processing with Apache Hadoop 2. Pearson Education.
- Lin, J., & Dyer, C. (2010). Data-intensive text processing with MapReduce. Synthesis Lectures on Human Language Technologies, 3(1), 1-177.
- Miner, D., & Shook, A. (2012). MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems. " O'Reilly Media, Inc.".