Guía Docente: Entornos y Ecosistemas de Soporte Big Data

DATOS GENERALES	
Facultad	Facultad de Ciencias y Tecnología
Titulación	Máster en Análisis Inteligente de Datos (Big Data)
Plan de estudios	2018
Materia	Tecnologías de computación en Big Data
Carácter	Obligatorio
Período de impartición	Segundo Trimestre
Curso	Primero
Nivel/Ciclo	Máster
Créditos ECTS	6
Lengua en la que se imparte	Castellano
Prerrequisitos	No se prevén requisitos previos; por tanto, los requisitos serán los propios del título.

DATOS DEL PROFESORADO			
Profesor Responsable	Miguel García Medina	**Correo electrónico**	miguel.garcia.medina@ui1.es
Área		**Facultad**	Facultad de Ciencias y Tecnología
Perfil Profesional 2.0	LinkedIn		

CONTEXTUALIZACIÓN Y JUSTIFICACIÓN DE LA ASIGNATURA	
Asignaturas de la materia	• Entornos y Ecosistemas de Soporte Big Data
Contexto y sentido de la asignatura en la titulación y perfil profesional	Tal y como ha sido estudiado en otras asignaturas de la titulación, los sistemas de procesado de datos masivos orientados a la distribución en disco propiciaron que las organizaciones hicieran frente al problema del procesado de volúmenes masivos de datos. Gracias a ellos se pudo extraer valor de los datos de la organización y ofrecer servicios de valor añadido antes no posibles. Pese a su potencial como herramientas para el procesado de datos, y su uso en una gran multitud de proyectos académicos y comerciales, existen sistemas de procesado masivo de datos alternativos que han conseguido resultados más notables que los sistemas de procesado de datos masivos orientados a disco en determinado tipo de escenarios. Uno de estos sistemas son los sistemas de procesado de datos masivos orientados a la distribución de datos en memoria principal. Pese a que generalmente los computadores cuentan con menor capacidad de almacenamiento en la memoria principal, ésta siempre ha sido tradicionalmente mucho más rápida en operaciones de lectura y escritura. Su velocidad de acceso a datos hacen de estos sistemas especialmente interesantes para situaciones en las que un conjunto de datos ha de ser accedido en múltiples ocasiones durante una tarea de cómputo o algoritmo. Ésta es precisamente la situación en la cual están emplazados un gran número de algoritmos de inteligencia artificial, aprendizaje automático, y análisis de datos. Es por ello por lo que los sistemas de procesado de datos masivos orientados a la distribución de datos en memoria principal han ganado especial relevancia en esos ámbitos, relegando los sistemas de procesado de datos masivos orientados a disco a un rol más secundario. En esta asignatura, el estudiante estudiará a nivel arquitectónico, teórico, y práctico uno de estos sistemas de procesado de datos masivos orientados a la distribución de datos en memoria principal: Apache Spark. Este se empleará como ejemplo para estudiar las particularidades de este tipo sistemas y como práctica de uno de los sistemas de procesado de datos más populares en la actualidad y altamente demandado en el ámbito laboral.

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE Y RESULTADOS DE APRENDIZAJE	
Competencias de la asignatura	• CB6: Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación. • CB7: Saber aplicar los conocimientos adquiridos y ser capaz de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio. • CB8: Ser capaz de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios. • CB9: Ser capaz de transmitir sus conclusiones, y los conocimientos y fundamentos que las sustentan, a públicos especializados y no especializados de un modo claro y sin ambigüedades. • CB10: 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. • CG2: Ser capaz de permanecer eficaz dentro de un medio cambiante, así como a la hora de enfrentarse con nuevas tareas, retos y personas. • CG4: Ser capaz de proponer soluciones imaginativas y originales así como ser capaz de promover la innovación e identificación de alternativas contrapuestas a los métodos y enfoques tradicionales en el contexto del análisis de datos masivos o bigdata. • CG5: Diseñar y desarrollar la implementación y puesta en marcha de proyectos de bigdata en diferentes áreas de aplicación social y profesional. • CG6: Ser capaz de integrarse en equipos de trabajo o investigación multidisciplinares de manera eficaz y colaborativa. • CE01: Comprender, aplicar y analizar arquitecturas y técnicas propias de bigdata para el análisis de datos estáticos y dinámicos, estructurados y no estructurados. • CE02: Identificar y utilizar herramientas software especializadas para el tratamiento de grandes volúmenes de datos en distintos contextos. • CE03: Saber diseñar y desarrollar soluciones en lenguajes y entornos de programación especializados en big data. • CE05: Diseñar, desarrollar y probar soluciones bigdata adaptadas para la captación, almacenamiento y tratamiento de grandes volúmenes de datos procedentes de diferentes contextos. • CE06: Comprender y utilizar técnicas avanzadas de visualización de datos y de experiencia de usuario para el diseño e implementación de interfaces adaptadas al usuario en los procesos de análisis de bigdata en distintos contextos.
Resultados de aprendizaje de la asignatura	• Conocer adecuadamente los principales entornos y ecosistemas utilizados para el análisis masivo de datos. • Análisis y explotación de datos en entornos y ecosistemas especializados en Big Data. • Planificar y diseñar un proyecto de Big Data apoyado en el escenario más adecuado para un fin.

PROGRAMACION DE CONTENIDOS	
Breve descripción de la asignatura	En esta asignatura se verá entre otros: • Ecosistema Sparck. Fundamentos. • Almacenamiento de datos en Spark. • Fuentes de datos y consultas en Spark. • MLlib. Algoritmos e hipótesis. • GraphX. Algoritmos de grafos • Desarrollo de caso práctico
Contenidos	**Unidad didáctica 1: Los sistemas de procesado masivo de datos orientados a la distribución en memoria: Apache Spark** 1. La distribución de datos en disco y en memoria principal 2. ¿Qué es Apache Spark? 3. Arquitectura 4. Modelo de distribución de datos: RDD 5. Lenguajes de programación para el desarrollo en Apache Spark 6. Iniciando Apache Spark y primeras operaciones **Unidad 2: API de bajo nivel para la carga, transformación y guardado en datos en Apache Spark** 1. Carga de datos en RDD 2. Operaciones de transformación de datos sobre RDD únicos 3. Operaciones de transformación de datos sobre RDD múltiples 4. Variables compartidas 5. Integración de código local y distribuido 6. Guardado de datos **Unidad 3: Modelado y consulta de datos estructurados en Spark** 1. Dataframes y Datasets 2. Transformación de Dataframes 3. ¿Qué es Spark SQL? 4. Creación de tablas, vistas y bases de datos con Spark SQL 5. Consultas en Spark SQL **Unidad 4: Análisis inteligente de datos en Spark** 1. Paralelización de algoritmos de análisis de datos: modelos 2. Preparando Datasets en Spark

	3. ¿Qué son MLlib y GraphX? 4. Soporte para la clasificación, regresión y aprendizaje no supervisado 5. Soporte para los sistemas de recomendación 6. Soporte para operar con grafos **Unidad 5: Procesamiento de series de datos en tiempo real** 1. ¿Por qué necesitamos procesar series de datos? 2. ¿Qué es Spark Streaming? 3. Modelos de procesado de flujos de datos 4. Operaciones y transformaciones sobre flujos de datos **Unidad 6: Otras entornos y ecosistemas de soporte Big Data** 1. Apache Hive 2. Apache Pig 3. Apache Flink 4. Apache Kafka 5. Apache Mahout 6. Nueva versión Apache Spark 3.0

METODOLOGÍA	
Actividades formativas	Para cada una de las seis unidades didácticas de la asignatura, se plantearán una serie de actividades de carácter tanto formativo como sumativo con el fin de adquirir competencias y resultados de aprendizaje de la asignatura. Las actividades propuestas comprenderán: • **Estudio de casos prácticos**. Estos se emplearán para la adquisición de competencias prácticas y que lleven al estudiante a situaciones similares a las encontradas en el mundo real. En estos casos prácticos, el estudiante aplicará los conceptos teóricos introducidos en la correspondiente unidad para la resolución de un problema. Estas actividades van encaminadas a la adquisición de resultados de aprendizaje de mayor complejidad cognitiva. • **Contenidos teóricos**. Texto o materiales audiovisuales empleados para la introducción de nuevos conceptos en cada unidad didáctica y los aspectos más teóricos de la materia. Al final de cada unidad se adjuntarán actividades de autoevaluación que permitirán al estudiante comparar su proceso de aprendizaje actual con los resultados esperados. Además, se podrán sugerir lecturas o resolución de ejercicios teóricos para facilitar el aprendizaje del estudiante. • **Trabajos de investigación**. Investigación y exposición de contenidos teóricos llevados a cabo de forma individual o colaborativa para fomentar la discusión y el aprendizaje cooperativo. • **Foros de debate**. Los estudiantes debatirán sobre el impacto de determinados temas de la asignatura sobre aspectos de la vida cotidiana. Estos foros buscan fomentar el pensamiento crítico y el aprendizaje cooperativo. • **Cuestionarios**: Cuestionarios evaluables que servirán para poner a prueba los conocimientos adquiridos.

EVALUACIÓN	
Sistema evaluativo	*En caso de que la situación sanitaria impida la realización presencial de los exámenes con todas las garantías, la Universidad Isabel I celebrará dichas pruebas en modalidad online. Para la realización de dichos exámenes, la universidad incorporará la herramienta de proctoring a nuestra plataforma tecnopedagógica, con el objetivo de garantizar los procesos de autentificación del alumno, como el control del entorno durante el desarrollo de las pruebas de evaluación. A su vez, la Universidad Isabel I pondrá a disposición del alumnado una Unidad de Exámenes Online específica para ofrecer apoyo técnico durante todo el proceso y así solventar todas las incidencias que se puedan presentar.* El sistema de evaluación se basará en una selección de las pruebas de evaluación más adecuadas para el tipo de competencias que se trabajen. El sistema de calificaciones estará acorde con la legislación vigente *(Real Decreto 1125/2003, de 5 de septiembre, por el que se establece el sistema europeo de créditos y el sistema de calificaciones en las titulaciones universitarias de carácter oficial y de validez en todo el territorio nacional)*. El sistema de evaluación de la Universidad Isabel I queda configurado de la siguiente manera: **Sistema de evaluación convocatoria ordinaria** **Opción 1. Evaluación continua** Los estudiantes que opten por esta vía de evaluación deberán realizar el **seguimiento de la evaluación continua (EC)** y podrán obtener hasta un **60 %** de la calificación final a

través de las actividades que se plantean en la evaluación continua.

Además, deberán realizar un **examen final presencial (EX)** que supondrá el **40 %** restante. Esta prueba tiene una parte dedicada al control de la identidad de los estudiantes que consiste en la verificación del trabajo realizado durante la evaluación continua y otra parte en la que realizan diferentes pruebas teórico-prácticas para evaluar las competencias previstas en cada asignatura.

Para la aplicación de los porcentajes correspondientes, el estudiante debe haber obtenido una nota mínima de un 4 en cada una de las partes de las que consta el sistema de evaluación continua.

Se considerará que el estudiante supera la asignatura en la convocatoria ordinaria por el sistema de evaluación continua, siempre y cuando al aplicar los porcentajes correspondientes se alcance una calificación mínima de un 5.

Opción 2. Prueba de evaluación de competencias

Los estudiantes que opten por esta vía de evaluación deberán realizar una **prueba de evaluación de competencias (PEC)** y un **examen final presencial (EX)**.

La **PEC** se propone como una prueba que el docente plantea con el objetivo de evaluar en qué medida el estudiante adquiere las competencias definidas en su asignatura. Dicha prueba podrá ser de diversa tipología, ajustándose a las características de la asignatura y garantizando la evaluación de los resultados de aprendizaje definidos. Esta prueba supone el 50 % de la calificación final.

El **examen final presencial**, supondrá el **50 %** de la calificación final. Esta prueba tiene una parte dedicada al control de la identidad de los estudiantes que consiste en la verificación del seguimiento de las actividades formativas desarrolladas en el aula virtual y otra parte en la que realizan diferentes pruebas teórico-prácticas para evaluar las competencias previstas en cada asignatura.

Al igual que con el sistema de evaluación anterior, para la aplicación de los porcentajes correspondientes el estudiante debe haber obtenido una puntuación mínima de un 4 en cada una de las partes de las que consta la opción de prueba de evaluación de competencias.

Se considerará que el estudiante supera la asignatura en la convocatoria ordinaria por el sistema de la prueba de evaluación de competencias siempre y cuando al aplicar los porcentajes correspondientes se alcance una calificación mínima de un 5.

Sistema de evaluación convocatoria extraordinaria

Todos los estudiantes, independientemente de la opción seleccionada, que no superen las pruebas evaluativas en la convocatoria ordinaria tendrán derecho a una convocatoria extraordinaria.

La convocatoria extraordinaria completa consistirá en la realización de una **prueba de evaluación de competencias** que supondrá el **50 %** de la calificación final y un **examen final presencial** cuya calificación será el **50 %** de la calificación final.

Para la aplicación de los porcentajes correspondientes, el estudiante debe haber obtenido una nota mínima de un 4 en cada una de las partes de las que consta el sistema de evaluación de la convocatoria extraordinaria.

Los estudiantes que hayan suspendido todas las pruebas evaluativas en convocatoria ordinaria (evaluación continua o prueba de evaluación de competencias y examen final) o

	no se hayan presentado deberán realizar la convocatoria extraordinaria completa, como se recoge en el párrafo anterior. En caso de que hayan alcanzado una puntuación mínima de un 4 en alguna de las pruebas evaluativas de la convocatoria ordinaria (evaluación continua o prueba de evaluación de competencias y examen final), se considerará su calificación para la convocatoria extraordinaria, debiendo el estudiante presentarse a la prueba que no haya alcanzado dicha puntuación o que no haya realizado. En el caso de que el alumno obtenga una puntuación que oscile entre el 4 y el 4,9 en las dos partes de que se compone la convocatoria ordinaria (EC o PEC y examen), solo se considerará para la convocatoria extraordinaria la nota obtenida en la evaluación continua o prueba de evaluación de competencias ordinaria (en función del sistema de evaluación elegido), debiendo el alumno realizar el examen extraordinario para poder superar la asignatura. Al igual que en la convocatoria ordinaria, se entenderá que el alumno ha superado la materia en convocatoria extraordinaria si, aplicando los porcentajes correspondientes, se alcanza una calificación mínima de un 5.

BIBLIOGRAFÍA Y OTROS RECURSOS	
Bibliografía básica	• Zaharia, M., Chambers, B. (2018). Spark: The Definitive Guide. (1ª edición). O'Reilly. Este es un libro prácticamente de cabecera para cualquiera que desee adentrarse a fondo en el empleo de Apache Spark como herramienta de procesado masivo de datos. Escrito por algunos de los ingenieros involucrados en la creación de la herramienta, el libro se adentra tanto en la API a bajo nivel como en los proyectos satélite que ofrecen funcionalidad extra como Spark SQL, Spark Streaming, GraphX, o MLlib. • Karau, H., Warren, R. (2017). High performance Spark: Best practices for scaling & optimizing Apache Spark. (1ª edición). O'Reilly. Otro libro generalista sobre Apache Spark, aunque más centrado en el empleo eficiente de las operaciones disponibles en el sistema para construir aplicaciones de Spark altamente eficientes. Trata alguno de los proyectos satélite del proyecto Spark, aunque su análisis dentro del libro no es exhaustivo.
Bibliografía complementaria	• Ze?evi?, P., Bona?i, M. (2016). Spark in Action. (1ª edición). Manning Publications. • Frampton, M. (2015). Mastering Apache Spark (1ª edición). Packt Publishing. • Nandi, A. (2015). Spark for Python Developers. (1ª edición). Packt Publishing. • Ryza, S., Laserson, U., Owen, S., Wills, J. (2017). Advanced Analytics with Spark: Patterns for Learning from Data at Scale. (2ª edición). O'Reilly. • Malak, M., East, R. (2016). Spark GraphX in Action. (1ª edición). Manning Publications. • Macías, M., Gómez, M., Tous, R., Torres, J. (2015). Introducción a Apache Spark. (1ª edición). UOC editorial. • Nabi, Z. (2016). Pro Spark Streaming: The Zen of Real-time Analytics using Apache Spark. (1ª edición). Apress. • Guller, M. (2015). Big Data Analytics with Spark: A practitioner's guide to using Spark for large scale data analysis. (1ª edición). Apress.
Otros recursos	• Enlace oficial al software de Apache Spark y documentación asociada: https://spark.apache.org/ • Enlace del proyecto MLlib dentro de Apache Spark: https://spark.apache.org/mllib/ • Enlace del proyecto Graphx dentro de Apache Spark: https://spark.apache.org/graphx/ • Enlace del proyecto Spark Streaming dentro de Apache Spark: https://spark.apache.org/streaming/ • Enlace del proyecto Spark SQL dentro de Apache Spark: https://spark.apache.org/sql/ • Enlace oficial al software de Apache Hive y documentación asociada: https://hive.apache.org/ • Enlace oficial al software de Apache Kafka y documentación asociada: https://kafka.apache.org/ • Enlace oficial al software de Apache Pig y documentación asociada: https://pig.apache.org/ • Enlace oficial al software de Apache Flink y documentación asociada: https://flink.apache.org/ • Enlace oficial al software de Apache Mahout y documentación asociada: https://mahout.apache.org