Máster en Ingeniería Informática
62225 - Manipulación y análisis de grandes volúmenes de datos

Guía docente para el curso 2015 - 2016

Curso: , Semestre: , Créditos: 6.0

Información básica

Profesores

- **Sandra Silvia Baldassarri** sandra@unizar.es
- **Eva Mónica Cerezo Bagdasari** ecerezo@unizar.es
- **Sergio Ilarri Artigas** silarri@unizar.es
- **Raquel Trillo Lado** raqueltl@unizar.es

Recomendaciones para cursar esta asignatura

El alumno que curse esta asignatura debería haber cursado asignaturas previas relacionadas con bases de datos y sistemas de información.

Actividades y fechas clave de la asignatura

El calendario de clases, prácticas y exámenes, así como las fechas de entrega de trabajos de evaluación, se anunciará con suficiente antelación.

Inicio

Resultados de aprendizaje que definen la asignatura

El estudiante, para superar esta asignatura, deberá demostrar los siguientes resultados...

1:

Comprender y especificar los requisitos necesarios para la interacción, almacenamiento, transferencia y procesado de grandes volúmenes de datos.

2:

Conocer, comprender y aplicar las técnicas más comunes para la representación, tratamiento, análisis e interacción con repositorios de datos heterogéneos.

3:

Diseñar, desarrollar y evaluar una aplicación que facilite la elaboración y gestión de grandes volúmenes de

datos, conforme a criterios de escalabilidad y normativa existente.

Introducción

Breve presentación de la asignatura

En un entorno cada vez más competitivo, el éxito de una entidad pasa por tener la capacidad de obtener, manipular, procesar y analizar grandes volúmenes de datos, tanto propios como externos (heterogéneos). A partir de este análisis, la empresa es capaz tomar sus decisiones basándose en información actual y precisa. Un correcto diseño de la interacción persona-ordenador es clave para facilitar la recogida y obtención de esta información, así como para comunicar y comprender la información generada. Por otra parte, las técnicas y metodologías de computación y procesamiento en este ámbito se centran principalmente en aspectos de paralelismo y computación intensiva. En esta asignatura se realiza un recorrido por todos los aspectos, desde el diseño hasta la manipulación, procesado y análisis, que representan el reto introducido por grandes volúmenes de datos.

Contexto y competencias

Sentido, contexto, relevancia y objetivos generales de la asignatura

La asignatura y sus resultados previstos responden a los siguientes planteamientos y objetivos:

La asignatura estará centrada en la comprensión, el análisis y la evaluación de los siguientes contenidos:

- Importancia de los datos y de su análisis en diferentes ámbitos de aplicación. Perspectiva del diseñador, usuario y analista.
- Técnicas de interacción y visualización. Diseño centrado en el usuario. Evaluación de la usabilidad de aplicaciones y software. Aplicación al diseño de aplicaciones interactivas.
- Modelos de almacenamiento y procesamiento de grandes volúmenes de datos.
- Sistemas y metodologías para el análisis y manipulación de datos.

Con el desarrollo de la asignatura, y de conformidad con las competencias y resultados de aprendizaje esperados, se pretenden lograr los siguientes objetivos:

- Que el alumno analice, dado un problema que implica grandes volúmenes de datos, los requisitos necesarios para su gestión (almacenamiento, transferencia, procesamiento, visualización e interacción).
- Que el alumno desarrolle los elementos necesarios para integrar fuentes de datos heterogéneas, utilizando técnicas clásicas para la representación, tratamiento, análisis visualización e interacción con repositorios de datos heterogéneos.
- Que el alumno desarrolle una aplicación para un contexto dado, donde sea necesaria la gestión de grandes volúmenes de datos, y teniendo en cuenta criterios de escalabilidad, usabilidad y normativos.

Contexto y sentido de la asignatura en la titulación

En esta asignatura se realiza un recorrido por todos los aspectos que representan el reto introducido por los grandes volúmenes de datos, desde el diseño hasta la manipulación, procesado y análisis.

Los egresados de este master son los profesionales capaces de liderar la implantación de sistemas de información que permitan conseguir este fin, así como auditar su adecuada implantación.

Al superar la asignatura, el estudiante será más competente para...

1:
Afrontar con éxito los siguientes desempeños transversales:

1. Proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la Ingeniería Informática.
2. Dirección de obras e instalaciones de sistemas informáticos, cumpliendo la normativa vigente y

asegurando la calidad del servicio.
3. Elaboración, planificación estratégica, dirección, coordinación y gestión técnica y económica de proyectos en todos los ámbitos de la Ingeniería en Informática siguiendo criterios de calidad y medioambientales.
4. Comprender y aplicar la responsabilidad ética, la legislación y la deontología profesional de la actividad de la profesión de Ingeniero en Informática.
5. Aplicar los principios de la economía y de la gestión de recursos humanos y proyectos, así como la legislación, regulación y normalización de la informática.
6. Aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.
7. Predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad.
8. Transmitir de un modo claro y sin ambigüedades a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.
9. Asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.
10. 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
11. 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.
12. 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.
13. Comunicar conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
14. Poseer las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

2:
Afrontar con éxito los siguientes desempeños relacionados con la Ingeniería Informática:

1. Comprender y saber aplicar el funcionamiento y organización de Internet, las tecnologías y protocolos de redes de nueva generación, los modelos de componentes, software intermediario y servicios.
2. Analizar las necesidades de información que se plantean en un entorno y llevar a cabo en todas sus etapas el proceso de construcción de un sistema de información.
3. Aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento.
4. Conceptualizar, diseñar, desarrollar y evaluar la interacción persona-ordenador de productos, sistemas, aplicaciones y servicios informáticos.

Importancia de los resultados de aprendizaje que se obtienen en la asignatura:

El conjunto de los resultados de aprendizaje se puede resumir diciendo que el alumnado será capaz de diseñar soluciones para la gestión y el análisis de grandes volúmenes de datos en distintos escenarios, escogiendo entre las soluciones tecnológicas existentes de forma adecuada. Esto es de gran importancia hoy en día en el mundo laboral, ya que multitud de empresas privadas e instituciones públicas, así como otras entidades especializadas en ámbitos concretos, cuentan con grandes volúmenes de datos que necesitan gestionar. Términos como *Big Data*, *Data Science*, *Data Analytics*, *Data Warehouses*, *Business Intelligence*, y *Data Mining*, están cobrando especial importancia en los últimos años, dada la necesidad de profesionales en este ámbito.

Evaluación

Actividades de evaluación

El estudiante deberá demostrar que ha alcanzado los resultados de aprendizaje previstos mediante las siguientes actividades de evaluacion

1:
Realización y presentación de trabajos. Estudio de un tema relacionado con la asignatura, elaboración de un informe sobre el mismo, y su presentación en clase. [30%]. Resultados de aprendizaje: 1, 2 y 3

2:
Proyecto. Un proyecto de grupo en el laboratorio, en el que se podrán en práctica los conocimientos y habilidades adquiridos en la asignatura. [40%]. Resultados de aprendizaje: 1, 2 y 3

3:
Prueba final escrita incluyendo preguntas de respuesta corta y de respuesta extensa. [30%]. Resultados de aprendizaje: 1, 2 y 3

Actividades y recursos

Presentación metodológica general

El proceso de aprendizaje que se ha diseñado para esta asignatura se basa en lo siguiente:

Las actividades de enseñanza y aprendizaje presenciales se basan en:

1. **Clase presencial.** Exposición de contenidos mediante presentación o explicación por parte de un profesor (posiblemente incluyendo demostraciones).
2. **Charlas de expertos.** Cuando sea posible, se contará con la exposición de contenidos mediante presentación o explicación por parte de un experto externo a la Universidad.
3. **Seminario**. Período de instrucción basado en contribuciones orales o escritas de los estudiantes.
4. **Aprendizaje basado en problemas**. Enfoque educativo orientado al aprendizaje y a la instrucción en el que los alumnos abordan problemas reales en pequeños grupos y bajo la supervisión de un tutor
5. **Clases prácticas**. Cualquier tipo de actividad de carácter práctico o colaborativo en el aula.
6. **Laboratorio**. Actividades desarrolladas en espacios especiales con equipamiento especializado (laboratorio, aulas informáticas).
7. **Tutoría**. Período de instrucción realizado por un tutor con el objetivo de revisar y discutir los materiales y temas presentados en las clases.
8. **Evaluación**. Conjunto de pruebas escritas, orales, prácticas, proyectos, trabajos, etc. utilizados en la evaluación del progreso del estudiante

Las actividades de enseñanza y aprendizaje no presenciales se basan en:

1. **Trabajos teóricos**. Preparación de seminarios, lecturas, investigaciones, trabajos, memorias, etc. para exponer o entregar en las clases teóricas.
2. **Trabajos prácticos.** Preparación de actividades para exponer o entregar en las clases prácticas.
3. **Estudio teórico.** Estudio de contenidos relacionados con las “clases teóricas”: incluye cualquier actividad de estudio que no se haya computado en el apartado anterior (estudiar exámenes, trabajo en biblioteca, lecturas complementarias, hacer problemas y ejercicios, etc.)
4. **Estudio práctico.** Relacionado con las “clases prácticas”
5. **Actividades complementarias**. Son tutorías no académicas y actividades formativas voluntarias relacionadas con la asignatura, pero no con la preparación de exámenes o con la calificación: lecturas, seminarios, jornadas, vídeos, etc.

Actividades de aprendizaje programadas (Se incluye programa)

El programa que se ofrece al estudiante para ayudarle a lograr los resultados previstos comprende las siguientes actividades...

1: Contenidos programados

- Introducción y motivación al problema de los grandes volúmenes de datos (*Big Data*).
- Almacenamiento de grandes volúmenes de datos:
  - Almacenes de datos (*data warehouses*). Diseño en estrella.
  - Bases de datos NoSQL.
- Gestión de grandes volúmenes de datos:
  - Distribución de los datos.
  - Integración de información con fuentes de datos heterogéneas.
  - Uso de técnicas de representación del conocimiento (ontologías) para la representación de fuentes de datos y su acceso e integración.
  - Técnicas de procesamiento paralelo: MapReduce (Hadoop).
  - Sistemas de gestión de flujos de datos (*data streams*).
  - Otras técnicas: agentes móviles.
- Interacción con grandes volúmenes de datos:
  - Técnicas de visualización.
  - Diseño de interfaces de usuario apropiados.
  - Usabilidad.
- Análisis de grandes volúmenes de datos:
  - Minería de datos.
  - Análisis del sentimiento.
  - Minería de textos.
- Casos de uso y aplicaciones:
  - Datos provenientes de sensores.
  - Datos no estructurados en la web.
  - Sistemas de recomendación.
  - Análisis de blogs y redes sociales.
  - Ciudades inteligentes (*smart cities*).
  - Sistemas de Transporte Inteligentes.

2:

Trabajo del estudiante

La asignatura consta de 6 créditos ECTS que suponen una dedicación estimada por parte del alumno de 150 horas (60 horas presenciales y 90 horas no presenciales) distribuidas del siguiente modo:

- 55 horas, aproximadamente, de actividades presenciales (clases magistrales incluyendo seminarios profesionales, resolución de problemas y casos, y prácticas de laboratorio).
- 65 horas de trabajo en grupo.
- 25 horas de trabajo y estudio individual efectivo.
- 5 horas dedicadas a distintas pruebas de evaluación.

Planificación y calendario

Calendario de sesiones presenciales y presentación de trabajos

La organización docente prevista de las sesiones presenciales en el campus Río Ebro es la siguiente:

- Clases magistrales.
- Resolución de problemas y casos.
- Prácticas de laboratorio.

Los horarios de todas las clases y fechas de las sesiones de prácticas se anunciarán con suficiente antelación a través de las webs del centro y de la asignatura.

Los proyectos propuestos serán entregados al finalizar el cuatrimestre, en las fechas que se señalen.

Bibliografía recomendada por el profesor

Bibliografía básica:

- “Big Data: Análisis de Grandes Volúmenes de Datos en Organizaciones”, Luis Joyanes Aguilar, Marcombo, 2013.
- “Data Warehousing in the Age of Big Data”, Krish Krishnan, The Morgan Kaufmann Series on Business Intelligence, ISBN 978-0124058910, Morgan Kaufmann, 2013.
- “Big Data: Principles and Best Practices of Scalable Realtime Data Systems”, Nathan Marz, James Warren, ISBN 978-1617290343, Manning Publications, 2014.
- "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", David Loshin, ISBN 978-0124173194, Morgan Kaufmann, 2009.
- “The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling”, Ralph Kimball, Margy Ross, John Wiley & Sons, 2011.
- “Interactive Data Visualization: Foundations, Techniques and Applications”, Matthew O. Ward, Georges Grinstein, Daniel Keim, A K Peters/CRC Press, 2010
- Transparencias, enunciados de problemas, casos de estudio y guiones de prácticas que los profesores de la asignatura pondrán a disposición del alumnado mediante la plataforma Moodle 2 del Anillo Digital Docente (http://add.unizar.es).

Bibliografía complementaria:

- “Multidimensional Databases and Data Warehousing”, Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen, Morgan & Claypool Publishers, 2010.
- “Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications”, Elzbieta Malinowski, Esteban Zimányi, Springer, 2008.
- “The Data Warehouse Lifecycle Toolkit” (Second Edition), Ralph Kimball, John Wiley & Sons, 2008.
- “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”, Bing Liu, Springer, 2011.
- “Introduction to Data Mining and its Applications”, S. Sumathi, S. N. Sivanandam, Studies in Computational Intelligence, volume 29, Springer, 2006.
- "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, Bing Liu, ISBN 978-1608458844, Morgan & Claypool Publishers, 2012.
- “Design for information: An introduction to the Histories, Theories and Best Practices Behind Effective Information Visualizations”, Isabel Meirelles, Rockport Publishers, 2013

Referencias bibliográficas de la bibliografía recomendada

- Jensen, Christian S. Multidimensional Databases and Data Warehousing / Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen Morgan & Claypool Publishers, 2010.
- Joyanes Aguilar, Luis. Big data : análisis de grandes volúmenes de datos en organizaciones / Luis Joyanes Aguilar . - 1ª ed. [Barcelona] : Marcombo, 2014
- Kimball, Ralph. The Data Warehouse Lifecycle Toolkit / Ralph Kimball. John Wiley & Sons, 2008
- Kimball, Ralph. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling / Ralph Kimball, Margy Ross John Wiley & Sons, 2011
- Krishnan, Krish. Data warehousing in the age of big data / Krish Krishnan Amsterdam : Morgan Kaufmann is an imprint of Elsevier, cop. 2013
- Liu, Bing. Sentiment Analysis and Opinion Mining : Synthesis Lectures on Human Language Technologies / Bing Liu Morgan & Claypool Publishers, 2012.
- Liu, Bing. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data / Bing Liu Springer, 2011
- Loshin, David. Big data analytics : from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph / David Loshin Amsterdam : Elsevier, cop. 2013
- Malinowski, Elzbieta. Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications / Elzbieta Malinowski, Esteban Zimányi. Springer, 2008
- Marz, Nathan. Big Data: Principles and Best Practices of Scalable Realtime Data Systems / Nathan Marz, James Warren Manning Publications, 2014
- Meirelles, Isabel. Design for information : An introduction to the histories, theories and best practices behind effective information visualizations / Isabel Meirelles Rockport Publishers, 2013
- Sumathi, S.. Introduction to Data Mining and its Applications / S. Sumathi, S. N. Sivanandam Springer, 2006.
- Ward, Matthew O.. Interative data visualization : Foundations, techniques and applications / Matthew O. Ward...[et al.] CRC Press, 2010