
DENOMINACIÓN DE LA ACTIVIDAD: 11088 - Bases de Datos Masivas (Gestión y Análisis)

TIPO DE ACTIVIDAD ACADÉMICA: Asignatura

CARRERA: Licenciatura en Sistemas de Información

PLAN DE ESTUDIOS: 17.13

DOCENTE RESPONSABLE:

Mg. Banchemo, Santiago - Profesor adjunto

EQUIPO DOCENTE:

Lic. Juan Manuel Fernandez - Jefe de Trabajos Prácticos

Lic. Pablo Maximiliano Lulic - Ayudante de Primera

ACTIVIDADES CORRELATIVAS PRECEDENTES:

PARA CURSAR: (11059) Sistemas de Información III, (11078) Base de Datos II y (41408) Modelos, Simulación y Teoría de la Decisión en condición de Regular.

PARA APROBAR: (11059) Sistemas de Información III, (11078) Base de Datos II y (41408) Modelos, Simulación y Teoría de la Decisión en condición de Aprobado.

CARGA HORARIA TOTAL: HORAS SEMANALES: 4 - HORAS TOTALES: 64.

DISTRIBUCIÓN INTERNA DE LA CARGA HORARIA:

[TIPO DE ACTIVIDAD: 2 horas, 50% TEÓRICO.

[TIPO DE ACTIVIDAD: 2 horas, 50% PRÁCTICO.

PERÍODO DE VIGENCIA DEL PRESENTE PROGRAMA: 2023-2024
--

CONTENIDOS MÍNIMOS O DESCRIPTORES

Sistemas de Información para apoyo de decisiones gerenciales: SIG, DSS, EIS. Análisis y Diseño de sistemas de apoyo de decisiones gerenciales. Definición de indicadores. Gestión de Datos Masiva. Bases de Datos multidimensionales (Datamart - Datawarehouse). Metodología para el desarrollo de procesamientos transaccionales (OLTP) y analíticos (OLAP). Análisis de Negocios (Business Analysis). Inteligencia de Negocios (BI). Tableros de comando. Minería de datos (Data Mining) y descubrimiento de conocimiento.

FUNDAMENTACIÓN, OBJETIVOS, COMPETENCIAS

La calidad de la información provista por los sistemas informáticos para la toma de decisiones permite aumentar la eficiencia de las mismas, junto con la reducción de los tiempos de análisis. Para soportar estas tareas, diversas tecnologías se han desarrollado y evolucionado en los últimos años, permitiendo contar con herramientas muy poderosas. Inclusive, los requerimientos de hardware y software y la alta disponibilidad de los mismos las ponen al alcance de muchos potenciales usuarios.

En un primer lugar, el concepto los *data warehouse* como base de un sistema de información gerencial que cumple el rol de integrador de información proveniente de diversas fuentes, muchas - inclusive - funcionalmente distintas (bases de datos corporativas, bases de datos locales, sistemas heredados), brindando una visión integrada. Esta concepción permite obtener información estratégica, ya sea de perfiles, riesgos, rentabilidades, evolución, entre otras. A su vez con la explosión del Big Data y las soluciones a gran escala para dar respuesta en tiempo real, agilizando los procesos de generación de repositorios surgen nuevas herramientas que se muestran como evoluciones de los enfoques tradicionales. Data Lake es un ejemplo claro de esto, donde la propuesta de generación de herramientas de manera más ágil y veloz es la premisa.

Por otro lado, el uso de la minería de datos complementa esta visión permitiendo extraer de forma automática relaciones ocultas en grandes volúmenes de datos, las cuales pueden mostrar aspectos profundos difíciles hallar.

De aquí que diseñar almacenes de datos y explorarlos con técnicas avanzadas habilita la utilización de herramientas de análisis de datos sofisticadas que soportan múltiples instancias de toma de decisiones organizacionales.

OBJETIVOS GENERALES:

Se espera que al completar el curso los estudiantes:

- Comprendan el entorno de análisis de datos para la toma de decisiones en una organización.
- Sean capaces de diseñar e implementar un *data warehouse* y soluciones basadas en OLAP.
- Conozcan las técnicas clásicas de minería de datos y su utilización.
- Desarrollen criterios para la interpretación de los resultados de los procesos de descubrimiento de conocimiento.
- Sean capaces de analizar los requerimientos de un programa de implementación de soluciones de análisis de datos.

CONTENIDOS

Unidad I - Introducción al proceso de descubrimiento de conocimiento

Sistemas de Información para apoyo de decisiones gerenciales. Marco teórico del proceso de descubrimiento de conocimiento. Etapas del proceso. Metodologías de descubrimiento de conocimiento.

Unidad II - Preprocesamiento e integración

Selección de fuentes de información e integración de datos a través de procesos de Extracción, Transformación y Carga (ETL). Métodos de tratamiento de ruido, análisis de outliers, reducción de dimensionalidad, construcción y transformación de variables. Tratamiento de fuentes heterogéneas de información, bases de dato NoSQL: características y tipos.

Unidad III - Bases de datos multidimensionales

Bases de Datos multidimensionales. Esquemas en estrella, copo de nieve y constelación. Metodología para el desarrollo de procesamientos analíticos OLAP. Implementaciones relacionales, multidimensionales e híbridas. Jerarquías y granularidad.

Unidad IV - Conceptos de modelización en minería de datos

Minería de datos (Data Mining) y descubrimiento de conocimiento. Algoritmos de extracción de conocimiento (reglas, árboles, regresión, clasificación, clustering). Interpretación de resultados. Aplicaciones.

Unidad V - Aplicaciones de análisis

Inteligencia de Negocios (BI), definición de indicadores, tableros de comandos (operativos, directivos y estratégicos). Proceso de diseño de un tablero de comandos. Herramientas de implementación de tableros de comandos. Reporting, herramientas para implementación de reportes.

Unidad VI - Infraestructura de Big Data

Ecosistemas de Big Data. Escalabilidad horizontal con AMQP. Descripciones de alto nivel: RabbitMQ. Sistemas de archivos distribuidos HDFS. Frameworks de desarrollo de Big Data: Hadoop MapReduce y Spark.

METODOLOGÍA

El desarrollo es teórico/práctico, con balance entre ambas partes. Las clases teóricas permiten plantear conceptos y modelos que luego serán empleados en la resolución de las actividades prácticas y/o las pruebas en laboratorio.

En las clases prácticas se realizarán tanto ejercicios de diseño de soluciones como implementaciones utilizando software de terceros o propio.

TRABAJOS PRÁCTICOS

En las actividades prácticas se consideran tanto la resolución de ejercicios de análisis de una situación como la resolución propiamente dicha.

Se proveerán entornos basados en contenedores Docker para llevar adelante todas las etapas de implementación práctica. En este sentido, se utilizarán gestores de bases de datos y herramientas de OLAP y OLTP de libre disponibilidad. Como ejemplo de BI se utilizará Pentaho, mientras que para minería de datos se utilizarán los lenguajes de programación R y Python. Se programarán pequeñas piezas de software o extensiones a las existentes.

También se utilizarán herramientas de software libre como Apache Kylin, Apache Hadoop, Spark y RabbitMQ para la conformación de ecosistemas de Big Data.

Para el Trabajo Final, los alumnos presentan su propio proyecto, acordado con los docentes. En éste deben realizar una investigación relacionada a algunos de los tópicos desarrollados en la asignatura o bien una propuesta con estudio experimental de algún enfoque alternativo a las técnicas existentes. En cualquiera de los casos, se debe elaborar un documento con formato de artículo de

investigación (paper) donde se expongan los objetivos, antecedentes, la propuesta, la metodología utilizada y los resultados obtenidos.

REQUISITOS DE APROBACION Y CRITERIOS DE CALIFICACIÓN:

La evaluación consta de 2 (dos) exámenes parciales y un trabajo final integrador (descrito en el apartado anterior) obligatorio. Los exámenes parciales se aprueban con nota 4 (cuatro) o superior mientras que el integrador con 7 (siete) o superior.

CONDICIONES PARA PROMOVER (SIN EL REQUISITO DE EXAMEN FINAL), DE ACUERDO AL ART.23 DEL RÉGIMEN GENERAL DE ESTUDIOS

RESHCS-LUJ:0000996-15

- a) Tener aprobadas las actividades correlativas al finalizar el turno de examen extraordinario de ese cuatrimestre.
- b) Cumplir con un mínimo del 80% de asistencia para todas las actividades.
- c) Aprobar todos los trabajos prácticos previstos en este programa, pudiendo recuperarse hasta un 25% del total por ausencias o aplazos.
- d) Aprobar el 100% de las evaluaciones previstas con un promedio no inferior a seis (6) puntos sin recuperar ninguna.
- e) Aprobar una evaluación integradora de la asignatura con calificación no inferior a siete (7) puntos.

CONDICIONES PARA APROBAR COMO REGULAR (CON REQUISITO DE EXAMEN FINAL) DE ACUERDO AL ART.24 DEL RÉGIMEN GENERAL DE ESTUDIOS

RESHCS-LUJ:0000996-15

- a) Estar en condición de regular en las actividades correlativas al momento de su inscripción al cursado de la asignatura.
- b) Cumplir con un mínimo del 70% de asistencia para todas las actividades.
- c) Aprobar todos los trabajos prácticos previstos en este programa, pudiendo recuperarse hasta un 40% del total por ausencias o aplazos.
- d) Aprobar el 100% de las evaluaciones previstas con un promedio no inferior a cuatro (4) puntos, pudiendo recuperar el 50% de las mismas. Cada evaluación solo podrá recuperarse en una oportunidad. Antes de presentarse a un examen, el estudiante debe tener aprobado el trabajo práctico integrador.

Antes de presentarse a un examen, el estudiante debe tener aprobado el trabajo práctico integrador.

EXAMENES PARA ESTUDIANTES EN CONDICIÓN DE LIBRES

- 1) Para aquellos estudiantes que, habiéndose inscripto oportunamente en la presente actividad hayan quedado en condición de libres por aplicación de los artículos 22, 25, 27, 29 o 32 del Régimen General de Estudios, SI podrán rendir en tal condición la presente actividad.
- 2) Para aquellos estudiantes que no cursaron la asignatura y se presenten en condición de alumnos libres en la Carrera, por aplicación de los artículos 10 o 19 del Régimen General de Estudios, SI podrán rendir en tal condición la presente actividad.
- 3) Para rendir la actividad académica en condición de libre deberán entregar con 15 días de antelación todos los Trabajos Prácticos obligatorios, los cuales serán corregidos y deberán estar aprobados por el equipo docente. Junto con los trabajos prácticos, deberán desarrollar el Trabajo Práctico Integrador previsto antes en este programa, cuyo tema habrán consensuado antes con el equipo docente a efectos de delinear propuesta y alcances de la misma. Por último, una vez aprobadas las instancias anteriores, deberán presentarse en una de las Mesas de Examen previstas institucionalmente y defender oralmente el Trabajo Final Integrador.

BIBLIOGRAFÍA

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. *Data Mining: Concepts and Techniques*.

Daniel T. Larose. 2014. Segunda edición. *Discovering Knowledge in Data: An Introduction to Data Mining*.

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Videla, A., & Williams, J. J. (2012). *RabbitMQ in action: distributed messaging for everyone*. Manning.

Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on Apache Spark. *International Journal of Data Science and Analytics*, 1(3-4), 145-164.

Scott, J. A. (2015). *Getting Started with Apache Spark*. MapR Technologies.

Fundamentals of Database Systems. Ramez Elmasri, Shamkant Navathe. Addison-Wesley, 6ta. Ed., 2010.

Fundamentos de bases de datos. Abraham Silberschatz. McGraw-Hill, 2002.

OLAP Solutions: Building Multidimensional Information Systems. Erik Thomsen. Wiley, 2da Ed., 2002.

Data Warehouses and Olap: Concepts, Architectures and Solutions. Robert Wrembel, Christian Koncilia. IGI Global, 2006.

DW 2.0: The Architecture for the Next Generation of Data Warehousing. William H. Inmon, Derek Strauss, Genia Neushloss, Morgan Kaufmann, 2008.

Introducción a la minería de datos.
José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez. Pearson, 2004.

The Data Warehouse ETL Toolkit . Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Ralph Kimball and Joe Caserta Wiley Publishing, Inc. 2004.

DISPOSICIÓN DE APROBACIÓN: CD