Universidad de Puerto Rico
Recinto de Río Piedras
Facultad de Administración de Empresas
Instituto de Estadística y Sistemas Computadorizados de Información

PRONTUARIO

A. Título

Conceptos y aplicaciones de Big Data

B. Codificación del curso

SICI 5015

C. Cantidad de Horas/Créditos

Tres (3) horas semanales, 3 créditos

D. Pre-requisitos
Haber aprobado los siguientes cursos: SICI 4015 - Diseño físico e implantación utilizando bases de datos (o su equivalente) y ESTA 3042 – Estadística Aplicada y Analítica de Datos II

E. Descripción del Curso

Estudio y uso práctico de la tecnología, las herramientas y técnicas para el manejo y análisis de datos de gran volumen y de variedad de formatos (Big Data). Tecnología para captura y almacenamiento distribuído o elástico de datos de gran tamaño recibidos tanto en batch, como streaming. Procesamiento en paralelo y herramientas para análisis de este tipo de datos. Aprendizaje automático (machine learning) y su uso para análisis de big data. Algoritmos para analísis de big data tales como predictivos y prescriptivos. Análisis de datos semi estructurados y no estructurados tales como texto e imágenes.

F. Objetivos de aprendizaje

Al finalizar el curso el estudiante podrá:

1. Explicar qué es Big Data y cómo se diferencia del procesamiento y análisis tradicional de datos y su rol en la ventaja competitiva de la empresa.

2. Explicar los aspectos que deben considerarse al planificar y adoptar Big Data: Privacidad, seguridad, fuentes de datos y su calidad, gobernanza de datos, el cliclo de vida de analítica de datos, tipos de análisis, modos de análisis.
3. Describir la tecnología necesaria para almacenar y analizar Big Data.
4. Explicar cómo se utiliza procesamiento distribuído y paralelo para procesar Big Data.
5. Utilizar herramientas para capturar y almacenar datos de Big Data
6. Hacer análisis descriptivo de Big Data utilizando herramientas para procesamiento distribuído y en paralelo (Ejemplo MapReduce, Pig)
7. Hacer consultas interactivas básicas de Big Data utilizando Hive y/o base de datos NoSQL.
8. Explicar qué es "machine learning".
9. Hacer análisis predictivo básico de Big Data utilizando técnicas de aprendizaje supervisado como clasificación.
10. Explicar qué es "deep learning"
11. Hacer análisis de texto y "sentiment analysis".
12. Realizar otros análisis de Big Data tales como: sistemas de recomendación, "image classification ", análisis de redes sociales y análisis de grafos

G. Bosquejo de Contenido y distribución del tiempo

Horas	Tópico
Parte I – Introduccción a Big Data	
3	1. Discusión del prontuario 2. Introducción a Big Data y Analítica de Datos a. La gran acumulación de datos y la analítica de datos b. Las características de Big Data c. Tipos de análisis de datos (estadística, regresión, clasificación, clustering, análisis de texto,…) d. Modos de analítica (batch, tiempo real e interactivo) e. Aspectos a considerar al planificar y adoptar Big Data: Privacidad, seguridad, fuentes de datos y su calidad, gobernanza de datos, el cliclo de vida de analítica de datos f. Ejemplos de aplicaciones de Big Data
Parte II – La infraestructura para Big Data – Almacenamiento y Procesamiento	

Horas	Tópico
1.5	3. La infraestructura para Big Data a. El ecosistema de Hadoop b. Otros productos para Big Data
1.5	4. Almacenamiento de Big Data a. Sistemas de archivos distribuídos, características: clusters, sharding y replication.
3	5. Procesamiento de Big Data a. Paralelismo y procesamiento distribuído para Big Data
1.5	6. Instalación y configuración de una estiba de software para Big Data
Parte III – Análisis en Batch	
1.5	7. Herramientas para captura de datos
3	8. Lenguajes de programación para procesamiento paralelo y distribuído (Ej. Pig)
Parte IV – Queries interactivos NoSQL y Bases de datos para Big Data	
3	9. Bases de datos para Big Data (Ej. Hive)
3	10. NoSQL
1.5	**Primer examen**
Parte V – Conceptos de Aprendizaje automático (“Machine Learning”)	
1.5	11. Introduccion al aprendizaje automático, terminología a. Categorías de aprendizaje automático: b. Clasificacion supervisada c. Clasificacion no supervisada d. Aprendizaje por refuerzo (“Reinforcement learning”) e. Fases en el proceso de “machine learning”
1.5	12. Herramientas de aprendizaje automático para Big Data
Parte VI. Clasificacion supervisada	
1.5	13. Clasificacion supervisada a. Particionamiento de datos b. Metodos de evaluación c. Técnicas de clasificación supervisada
3	14. Arboles de Decisión a. Particionamiento recursivo, ganancia de información y entropía b. Aplicación y validación de arboles de decisión c. Random Forest

Horas	Tópico
3	15. Redes neurales artificiales a. Conceptos de redes neurales: Perceptron, capas de la red neural, función de activación, descenso de gradiente y "backpropagation". b. Aplicación y validación de redes neurales
Parte VII. Deep learning	
1.5	16. Definicion y terminología de "Deep learning" a. Retos de Deep learning b. Herramientas para aplicar Deep Learning a Big Data
1.5	17. Redes neurales profundas y otras arquitecturas de redes: "Convolutional neural networks", redes recurrentes, autoencoder, "generative adversarial networks" 18. Aplicación y evaluación de Deep learning con grandes cantidades de datos
Parte VIII. Analisis de imágenes	
3	19. Procesamiento de imágenes 20. "Convolutional neural networks" CNN a. Arquitectura b. Selección de hiperparametros c. Construccion del modelo d. Aplicación y evaluación de CNN usando imágenes
Parte IX. Analisis de texto	
3	21. Pasos en el análisis de texto a. Representacion de documentos 22. Clasificación de documentos a. "sentiment analysis" 23. Redes neurales recurrentes (RNN) a. Backpropagation a través del tiempo b. Long short-term memory c. Redes neurales recurrentes distribuidas d. Aplicación y evaluación de RNN usando datos de texto
Parte X. Reglas de asociación	
1.5	24. Reglas de asociación a. Medición de "Confidence", "support" y "lift" b. Algoritmo "a priori" c. Evaluación de resultados
1.5	**Segundo examen**

H. Técnicas instruccionales

A. Lista mínima de estrategias instruccionales

1. Estrategia instruccional principal:
    a. El curso enfatizará el enfoque de “Project Based Learning”. Los estudiantes practicarán los conceptos y las técnicas mayormente mediante ejercicios y proyectos, como una manera de profundizar en el aprendizaje y de apoyar el desarrollo de un nivel adecuado de destreza. Muchas de las actividades de práctica se llevarán a cabo en el salón de clases. Otras se llevarán a cabo en sesiones fuera del salón de clases supervisadas por el profesor.
2. Otras estrategias instruccionales:
    a. La participación activa de los estudiantes es muy importante para lograr los objetivos del curso. El profesor deberá promover dicha participación.
    b. Las estrategias instruccionales incluirán el uso de la tecnología para apoyar y hacer más efectivo y eficiente el proceso de enseñanza y aprendizaje. Por ejemplo, se utilizarán proyectores digitales para presentar el material a ser discutido. Además, se utilizará el acceso a Internet para presentar material que ilustre los temas discutidos, así como también realizar ejercicios en clase usando herramientas computadorizadas de Big Data.
    c. Se enfatizará la aplicación práctica de los conceptos y técnicas sin descuidar los aspectos teóricos.
    d. La preparación de asignaciones fuera del salón de clase será una parte importante de las estrategias instruccionales de este curso.

I. Recursos de aprendizaje e instalaciones mínimas requeridas

1. El estudiante deberá tener acceso a una computadora personal con acceso a Internet ya sea en un laboratorio de computadoras o en su casa.
2. Salón de clases equipado con pizarras, acceso al Internet y proyector que se pueda conectar a una computadora personal para desplegar visuales en una pantalla electrónica y una computadora por estudiante.
3. Acceso a uno o más servidores que provea acceso a las herramientas para recolectar, preparar, almacenar y analizar Big Data.

J. Proyectos

Se trabajarán un proyecto de varias fases:

1. **Analítica de datos big data.** Cada equipo de tres estudiantes trabajará un proyecto de Big Data.

Nota: La realización de estos proyectos requerirá experiencia práctica fuera del horario de la clase.

K. Técnicas de evaluación

Métodos y distribución de pesos sugeridos:

Examen Parcial #1	30%
Examen Parcial #2	30%
Proyecto	25%
Asignaciones	15%
	100%

L. Acomodo Razonable

Según la Ley de Servicios Educativos para Personas con Impedimentos (Ley 51 del 7 de junio de 1996), todo estudiante que requiera acomodo razonable deberá notificarlo al profesor el primer día de clases.

Los estudiantes que reciban servicios de Rehabilitación Vocacional deben comunicarse con el profesor al inicio del semestre para planificar el acomodo razonable y el equipo asistido necesario conforme a las recomendaciones de la Oficina de Asuntos para las Personas con Impedimento (OAPI) del Decanato de Estudiantes.

M. Normativa sobre discrimen por sexo y género e Integridad Académica

Normativa sobre discrimen por sexo y género en modalidad de violencia sexual

La Universidad de Puerto Rico prohíbe el discrimen por razón de sexo y género en todas sus modalidades, incluyendo el hostigamiento sexual. Según la Política Institucional contra el Hostigamiento Sexual en la Universidad de Puerto Rico, Certificación Num. 130, 2014-2015 de la Junta de Gobierno, si un estudiante está siendo o fue afectado por conductas relacionadas a hostigamiento sexual, puede acudir ante la Oficina de la Procuradoría Estudiantil, el Decanato de Estudiantes o la Coordinadora de Cumplimiento con Título IX para orientación y/o presentar una queja.

Conducta Estudiantil Sujeta a Sanciones Disciplinarias

Los actos de deshonestidad académica están sujetos a sanciones disciplinarias, según establece el Reglamento General de Estudiantes de la Universidad de Puerto Rico, Certificación 13, 2009-2010, Parte VI, Artículo 6.2.3

No se permite en momento alguno el uso de teléfonos celulares o cualquier otro artefacto electrónico no autorizado previamente. El profesor podrá tomar las medidas disciplinarias que considere pertinentes para evitar su uso.

N. Sistema de Calificaciones

A, B, C, D, F

O. Libro de Texto

Bahga, A.; Madisetti, V. 2016. Big Data Science & Analytics – A Hands-On Approach. Arshdeep Bahga & Vijay Madisetti. ISBN: 978-0-13-429107-9

Lecturas

El material a cubrir en clase será reforzado por lecturas que estarán disponibles en Moodle y/o la Biblioteca de la Facultad.

P. Bibliografía

Dasgupta, N. (2018). Practical Big Data Analytics – Hands-On Techniques To Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL and R. Packt.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113. (https://www.usenix.org/legacy/events/osdi04/tech/full_papers/dean/dean.pdf)

Dev, D. (2017) Deep learning With Hadoop – Build, Implement and Scale Distributed Deep learning Models for Large-scale Datasets. Packt.

Erl, T.; Khattak, W.; Buhler, P. 2016. Big Data Fundamentals – Concepts, Drivers & Techniques. Prentice Hall.

EMC Education Services. (2015) Data Science and Big Data Analytics – Discovering, analyzing, visualizing and presenting data. Wiley.

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). *The Google file system* (Vol. 37, No. 5, pp. 29-43). ACM.

Jansen, S. (2018) Hands-On Machine Learning for AlgorithmicTrading – Design and Implement Investment Strategies Based on Smart Algorithms that Learn from Data Using Python. Packt.

Olston, C., B. Reed, U. Srivastava, R. Kumar and A. Tomkins (2008, June). Pig Latin: A Not-So-Foreign Language for Data Processing C. In . *ACM SIGMOD 2008 International Conference on Management of Data, Vancouver, Canada*. (http://www.dcs.bbk.ac.uk/~dell/teaching/cc/paper/sigmod08/p1099-olston.pdf)

Quddus, J. (2018) Machine Learning with Apache Spark Quick Start Guide – Uncover Patterns, Derive Actionable Insights and Learn from Big Data Using MLlib.

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., ... & Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment*, *2*(2), 1626-1629. (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.303.2572&rep=rep1&type=pdf)

White, T. (2015). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.". Fourth Edition. (http://www.academia.edu/download/34613212/_HadoopThe.Definitive.Guide3rd_Early.Release_.Tom.White.%E6%96%87%E5%AD%97%E7%89%88_1.pdf)

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud*, *10*(10-10), 95.

(http://static.usenix.org/legacy/events/hotcloud10/tech/full_papers/Zaharia.pdf
)

Q. Referencias electrónicas
Portal sobre temas de: Machine Learning, Data Science, Data Mining, Big Data, Analytics y AI: https://www.kdnuggets.com/

Preparado por: Prof. María Teresa Jiménez y Dra. Roxana Aparicio