PROGRAMA DE CURSO

Código	Nombre
IN 5528	INTRODUCCIÓN A BIG DATA
Nombre en Inglés	
Introduction to Big Data	

SCT	Unidades Docentes	Horas de Cátedra	Horas Docencia Auxiliar	Horas de Trabajo Personal
6	10	3,0	1,5	5,5

Requisitos	Carácter del Curso
IN3501 Tecnologías de Información y Comunicaciones para la Gestión o CC3001 Algoritmos y Estructura de Datos o AUTOR	Electivo de la carrera Ingeniería Civil Industrial

Competencias a la que tributa el curso

Competencias de Egreso

- Concebir soluciones a los problemas que surgen en las organizaciones, utilizando los conocimientos provenientes de las tecnologías de información y comunicaciones.
- Comunicar ideas y resultados de trabajos profesionales o de investigación, en forma escrita y oral.
- Gestionar su auto-aprendizaje en el desarrollo del conocimiento de su profesión, adaptándose a los cambios del entorno.

Propósito del Curso

Hoy en día, se están obteniendo grandes cantidades de datos tanto estructurados como no estructurados, ante esto, nace la necesidad de entregar a los alumnos de ingeniería industrial conocimientos relacionados con Big Data para que puedan concebir nuevas soluciones a problemas que puedan surgir en todo tipo de organizaciones. Para llevar a cabo lo anterior, el cuerpo docente impartirá tanto clases teóricas como prácticas, además de diversas actividades que realizarán los estudiantes durante el curso haciéndolos agentes activos durante la realización del curso.

Resultados de Aprendizaje

Al finalizar el curso, el estudiante es capaz de:

1. Utilizar la arquitectura Lambda para poder construir una solución a un problema de Big Data.
2. Tener una intuición de aprendizaje de máquinas, sabiendo calcular, interpretar y presentar los resultados de la performance de un modelo.
3. Usar herramientas de visualización de datos.

Metodología Docente	Evaluación General
Este curso tiene una connotación teórico-práctica, por lo que, la metodología de trabajo consiste en: • Clases de Cátedra. • Clases Auxiliares. • Desarrollo de Tareas Grupales. • Presentaciones Orales. • Lecturas y Análisis de Casos.	El curso consta de 2 notas: tareas (NT) y controles (NC). El cálculo de esas notas se efectúa de la siguiente forma: $$CN = \frac{\sum_i C_i}{3} \qquad NT = \frac{\sum_i (w_i * P_i)}{n}$$ Donde C_i es la nota del control i, P_i es la nota de la tarea i y w_i, la ponderación que tiene cada una de ellas. • Las tareas son grupales y la nota P_i se calcula en base a una coevaluación: Si R es la nota obtenida en la tarea y el grupo es de m personas, cada integrante debe dar una nota a cada uno de sus compañeros, incluyéndose a sí mismo, sin superar $R * m$ puntos en total. Entonces, si se denota α_{jk} a la nota que el integrante j le da al integrante k, la nota de la tarea i para el integrante k está dada por: $$P_i = \frac{\sum_j \alpha_{jk}}{m}$$ Sujeto a: $\sum_j \alpha_{jk} = R * m$ • El alumno puede eximirse de dar el examen si $$NT \geq 5.5 \; y \; NC \geq 5.5$$ En este caso, la nota final (NF) corresponde al promedio simple entre NC y NT. • Al final del curso, se puede aplicar un factor $\alpha \in (1,1.03]$ sobre NF cuando la participación del alumno es excepcional. • En caso de que el alumno rinda el examen, la nota final se calcula de la siguiente forma. $$((0,6 * NC) + (0,4 * EX)) * 0,5 + (NT * 0,5)$$ • La condición para aprobar el curso es: $$NT \geq 4.0 \; y \; NC \geq 4.0$$

UNIDADES TEMÁTICAS

Número	Nombre de la Unidad	Duración en Semanas
1	Introducción a Big Data	3

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Un nuevo paradigma para Big Data a. Escalando con bases de datos tradicionales. b. Bases de datos NoSQL. c. Primeros principios. d. ¿Qué es lo que deseamos en un sistema Big Data? e. Problemas con arquitecturas completamente incrementales. f. Arquitectura Lambda. g. Tendencias en tecnología. 2. Data Science y 'Data Products' basados en objetivos	Aprendizaje de los principios de los sistemas de datos y dar una visión general de la Arquitectura Lambda.	1,2,3

Número	Nombre de la Unidad	Duración en Semanas
2	*Batch Layer*	4

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Modelo de datos para Big Data a. Las propiedades de los datos. b. Modelo basado en hechos para representar la data. c. Esquemas gráficos. 2. Almacenamiento de datos en la *Batch Layer* a. Requisitos de almacenamientos para el *dataset* maestro. b. Escogiendo una solución de almacenamiento. c. Como funciona un sistema de archivos distribuidos. d. Almacenando un *dataset* maestro en un sistema de archivos distribuidos. 3. Aprendizaje de Máquinas a. ¿Qué es y cuándo surge? b. Herramientas. c. Tipos de Aprendizaje de Máquinas. d. Aplicaciones. e. Indicadores que permitan evaluar la performance de un modelo f. Visualización de la performance de un modelo. 4. Batch Layer a. Computación en la *Batch Layer* b. Algoritmos recomputados vs algoritmos incrementales. c. Escalabilidad en la *Batch Layer.* d. MapReduce: un paradigma para Big Data.	Aprendizaje sobre el modelamiento de un *master dataset*, utilizando procesamiento por *batch* para crear vistas arbitrarias de sus datos, así como las ventajas y desventajas entre el procesamiento gradual y por lotes. Además, se adquiere una noción general de aprendizaje de máquinas, junto con los indicadores que permiten evaluar la performance de la performance de un modelo.	1,2,4

Número	Nombre de la Unidad	Duración en Semanas
3	*Serving Layer*	1

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Serving Layer a. Métricas de rendimiento para la *Serving Layer.* b. Solución para la el problema de normalización /desnormalización. c. Requisitos para una base de datos *Serving Layer.* d. Contrastando con una solución completamente incremental.	Aprendizaje acerca de las bases de datos especializadas que sólo se escriben de forma masiva y que éstas son dramáticamente más simples que las tradicionales, dándoles un excelente rendimiento, funcionamiento, y las propiedades de robustez.	1,2,5

Número	Nombre de la Unidad	Duración en Semanas
4	*Speed Layer*	2

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Vistas en tiempo real a. Computando en vistas en tiempo real. b. Almacenando vistas en tiempo real. c. Desafíos de la computación incremental. d. Actualizaciones asíncronas vs síncronas. 2. Mining Text Unsupervised Data Mining 3. Procesamiento en *stream* y *queuing*. a. *Queuing*. b. Procesamiento en *stream*.	Aprendizaje de las bases de datos NoSQL y procesamiento en stream. Además de procesamiento en *batch* incrementales, las variantes de la arquitectura básica Lambda, y cómo obtener el máximo provecho de sus recursos.	1,2,5

Número	Nombre de la Unidad	Duración en Semanas
5	*Arquitectura Lambda en Profundidad*	1

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Arquitectura Lambda en profundidad. a. Definición de un sistema de datos. b. *Batch y Serving Layers*. c. *Speed Layer*. d. *Query Layer*.	Procesamiento en *batch* incrementales, las variantes de la arquitectura básica Lambda, y cómo obtener el máximo provecho de sus recursos.	1

Número	Nombre de la Unidad	Duración en Semanas
6	*Visualización*	1

Contenidos	Indicador de Logro	Referencias a la Bibliografía
1. Visualización con librería D3. 2. Visualización de grafos.	Aprendizaje sobre herramientas que se utilizan para visualización en Big Data.	6,7

Bibliografía General
1. Nathan Marz y James Warren. “Big Data: Principles and best practices of scalable realtime data systems”. Manning 2015. 2. Foster Provost y Tom Fawcell. “Data Science for Business: What you need to know about data mining and data-analytic thinking”. 2013 3. Bernard Marr. “Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance”. 2015 4. Tom White. “Hadoop: The Definitive Guide”. O’Reilly 2015 5. Mat Brown. “Learning Apache Cassandra”. 2015 6. Corey L. Lanum “Visualizing Graph Data” 2016 7. Elijah Meeks “D3.js” 2015 8. Henry Brink, Joseph W. Richards, Mark Fetherolf “Real World Machine Learning” 2016

Vigencia desde:	Otoño 2015
Elaborado por:	Juan Velásquez y Felipe Vildoso
Validado por:	
Revisado por:	Unidad de Gestión Curricular, SGD