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 CC3201 Bases de Datos y IN4901 Práctica Profesional I 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.
- Emprender e innovar en el desarrollo de soluciones a problemas de ingeniería, demostrando iniciativa y capacidad de toma de decisión.
- 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 utilizar la arquitectura Lambda para poder construir una solución a un problema de Big Data. Los conocimientos adquiridos para este fin son los siguientes:

- Principios de los sistemas de datos y dar una visión general de la Arquitectura Lambda.
- *Batch Layer* de la arquitectura Lambda. Se aprenderá 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.
- *Serving Layer*, la que proporciona baja latencia de acceso a las vistas producidas por la *Batch Layer*. Se aprenderá 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.

- *Speed Layer*, la que compensa la alta latencia de la *Batch Layer* para poder entregar resultados actualizados para las consultas. Se aprenderá acerca de las bases de datos NoSQL y procesamiento en stream.
- Procesamiento en *batch* incrementales, las variantes de la arquitectura básica Lambda, y cómo obtener el máximo provecho de sus recursos.

Metodología Docente	Evaluación General
Este curso tiene una connotación teórico-práctica. Está compuesto por cátedras y auxiliares, las cuales condujeran a la construcción de una solución de *Big Data*. Se aplicará la teoría a la resolución de casos simples, que permitirán al alumno prepararse para desarrollar con éxito las distintas actividades del curso. Esta metodología tiene como objetivo primario entregar la práctica necesaria para llevar con éxito el diseño e implementación de las tareas a realizar durante el curso. En este sentido, se realizarán controles para evaluar el aprendizaje logrado en el curso, además de otras actividades complementarias como lecturas o presentaciones sobre temáticas relevantes para el curso.	El curso consiste de 2 notas, tareas (NT) y controles (NC). El cálculo de esas notas se efectúa de la siguiente forma: • NC = Promedio de controles (∑ Ci)/n, donde Ci son las notas de los controles. • NT = Promedio de las entregas parciales (∑ wi*Pi)/n, donde Pi son las notas de las tareas y wi la ponderación que tiene cada una de ellas. • El alumno puede eximirse de dar el examen si el promedio actual de controles es mayor o igual a 5.5 y la nota de tareas (NT) es mayor o igual que 5.5. En este caso, la nota final corresponde al promedio simple entre NC y NT • 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+NP*0,5 • La condición para aprobar el curso es: NP >= 4.0 y NC >= 4.0

UNIDADES TEMÁTICAS

Número	Nombre de la Unidad	Duración en Semanas
1	Introducción a Big Data	2

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.	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*	6

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	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.	1,2,4

archivos distribuidos. 3. *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.		

Número	Nombre de la Unidad	Duración en Semanas
3	*Serving Layer*	2
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*	5
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.	Aprendizaje de las bases de datos NoSQL y procesamiento en stream. Además de procesamiento en *batch* incrementales, las variantes	1,2,5

c. Desafíos de la computación incremental. d. Actualizaciones asíncronas vs síncronas. 2. Procesamiento en *stream* y *queuing*. a. *Queuing*. b. Procesamiento en *stream*. 3. Procesamiento en *stream* con *micro-batch* a. Logrando la semántica exactamente de una vez. b. Conceptos fundamentales. c. Extendiendo diagrama para procesamiento en *micro-batch*. d. Terminando la *Speed Layer*. 4. Arquitectura Lambda en profundidad. a. Definición de un sistema de datos. b. *Batch y Serving Layers*. c. *Speed Layer*. d. *Query Layer*.	de la arquitectura básica Lambda, y cómo obtener el máximo provecho de sus recursos.	

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

Vigencia desde:	Otoño 2016
Elaborado por:	Juan Velásquez y Felipe Vildoso
Validado por:	
Revisado por:	Comisión de Docencia DII