Universidad Nacional Mayor de San Marcos
Escuela Profesional de
Ciencia de la Computación
Silabo del curso
Periodo Académico 2018-II

1. Código del curso y nombre: CS3700. Big Data (Obligatorio)
2. Créditos: 3
3. Horas de Teoría y Laboratorio: 1 HT; 4 HL; (15 semanas)
4. Docente(s)

Atención previa coordinación con el profesor

5. Bibliografía

[Bal+08] Shumeet Baluja et al. "Video Suggestion and Discovery for Youtube: Taking Random Walks Through the View Graph". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China: ACM, 2008, pp. 895–904. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367618. URL: http://doi.acm.org/10.1145/1367497.1367618.

[BVS13] Rajkumar Buyya, Christian Vecchiola, and S. Thamarai Selvi. *Mastering Cloud Computing: Foundations and Applications Programming*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013. ISBN: 9780124095397, 9780124114548.

[Cou+11] George Coulouris et al. *Distributed Systems: Concepts and Design*. 5th. USA: Addison-Wesley Publishing Company, 2011. ISBN: 0132143011, 9780132143011.

[HDF11] Kai Hwang, Jack Dongarra, and Geoffrey C. Fox. *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN: 0123858801, 9780123858801.

[Low+12] Yucheng Low et al. "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud". In: *Proc. VLDB Endow.* 5.8 (Apr. 2012), pp. 716–727. ISSN: 2150-8097. DOI: 10.14778/2212351.2212354. URL: http://dx.doi.org/10.14778/2212351.2212354.

[Mal+10] Grzegorz Malewicz et al. "Pregel: A System for Large-scale Graph Processing". In: *ACM SIGMOD Record*. SIGMOD '10 (2010), pp. 135–146. DOI: 10.1145/1807167.1807184. URL: http://doi.acm.org/10.1145/1807167.1807184.

6. Información del curso

(a) **Breve descripción del curso** En la actualidad conocer enfoques escalables para procesar y almacenar grande volumenes de información (terabytes, petabytes e inclusive exabytes) es fundamental en cursos de ciencia de la computación. Cada dia, cada hora, cada minuto se genera gran cantidad de información la cual necesitá ser procesada, almacenada, analisada.

(b) **Prerrequisitos:**

- CS2702. Bases de Datos II. (5^{to} Sem)
- CS3P01. Computación Paralela y Distribuída. (7^{mo} Sem)

(c) **Tipo de Curso:** Obligatorio

(d) **Modalidad:** Presencial

7. Objetivos específicos del curso.

- Que el alumno sea capaz de crear aplicaciones paralelas para procesar grandes volumenes de información.

- Que el alumno sea capaz de comparar las alternativas para el procesamiento de big data.
- Que el alumno sea capaz de proponer arquitecturas para una aplicación escalable.

8. Contribución a los resultados (*Outcomes*)

a) Aplicar conocimientos de computación y de matemáticas apropiadas para la disciplina. (**Usar**)

b) Analizar problemas e identificar y definir los requerimientos computacionales apropiados para su solución. (**Usar**)

i) Utilizar técnicas y herramientas actuales necesarias para la práctica de la computación. (**Usar**)

j) Aplicar la base matemática, principios de algoritmos y la teoría de la Ciencia de la Computación en el modelamiento y diseño de sistemas computacionales de tal manera que demuestre comprensión de los puntos de equilibrio involucrados en la opción escogida. (**Usar**)

9. Competencias (IEEE)

C2. Capacidad para tener una perspectiva crítica y creativa para identificar y resolver problemas utilizando el pensamiento computacional.⇒ **Outcome a,b**

C16. Capacidad para identificar temas avanzados de computación y de la comprensión de las fronteras de la disciplina.⇒ **Outcome i**

CS2. Identificar y analizar los criterios y especificaciones apropiadas a los problemas específicos, y planificar estrategias para su solución.⇒ **Outcome i,b**

CS3. Analizar el grado en que un sistema basado en el ordenador cumple con los criterios definidos para su uso actual y futuro desarrollo.⇒ **Outcome j**

CS6. Evaluar los sistemas en términos de atributos de calidad en general y las posibles ventajas y desventajas que se presentan en el problema dado.⇒ **Outcome j**

10. Lista de temas a estudiar en el curso

1. Introducción a Big Data
2. Hadoop
3. Procesamiento de Grafos en larga escala

11. Metodologia y Evaluación
Metodología:

Sesiones Teóricas:
Las sesiones de teoría se llevan a cabo en clases magistrales donde se realizarán actividades que propicien un aprendizaje activo, con dinámicas que permitan a los estudiantes interiorizar los conceptos.

Sesiones de Laboratorio:
Para verificar que los alumnos hayan alcanzado el logro planteado para cada una de las unidades de aprendizaje, realizarán actividades que les permita aplicar los conocimientos adquiridos durante las sesiones de teoría y se les propondrá retos que permitan evaluar el desempeño de los alumnos.

Exposiciones individuales o grupales:
Se fomenta la participación individual y en equipo para exponer sus ideas, motivándolos con puntos adicionales en las diferentes etapas de la evaluación del curso.

Lecturas:
A lo largo del curso se proporcionan diferentes lecturas, las cuales son evaluadas. El promedio de las notas de las lecturas es considerado como la nota de una práctica calificada. El uso del campus virtual UTEC Online permite a cada estudiante acceder a la información del curso, e interactuar fuera de aula con el profesor y con los otros estudiantes.
Sistema de Evaluación:

12. Contenido

Unidad 1: Introducción a Big Data (15)	
Competences esperadas: C2, C4	
Objetivos de Aprendizaje	**Tópicos**
• Explicar el concepto de Cloud Computing desde el punto de vista de Big Data[Familiarizarse] • Explicar el concepto de los Sistema de Archivos Distribuidos [Familiarizarse] • Explicar el concepto del modelo de programación MapReduce[Familiarizarse]	• Visión global sobre Cloud Computing • Visión global sobre Sistema de Archivos Distribuidos • Visión global sobre el modelo de programación MapReduce
Lecturas : [Cou+11]	

Unidad 2: Hadoop (15)	
Competences esperadas: C2, C4	
Objetivos de Aprendizaje	**Tópicos**
• Entender y explicar la suite de Hadoop. [Familiarizarse] • Implementar soluciones usando el modelo de programación MapReduce. [Usar] • Entender la forma como se guardan los datos en el HDFS. [Familiarizarse]	• Visión global de Hadoop. • Historia. • Estructura de Hadoop. • HDFS, Hadoop Distributed File System. • Modelo de Programación MapReduce
Lecturas : [HDF11], [BVS13]	

Unidad 3: Procesamiento de Grafos en larga escala (10)	
Competences esperadas: C16	
Objetivos de Aprendizaje	**Tópicos**
• Entender y explicar la arquitectura del proyecto Pregel. [Familiarizarse] • Entender la arquitectura del proyecto GraphLab. [Familiarizarse] • Entender la arquitectura del proyecto Giraph. [Familiarizarse] • Implementar soluciones usando Pregel, GraphLab o Giraph. [Usar]	• Pregel: A System for Large-scale Graph Processing. • Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. • Apache Giraph is an iterative graph processing system built for high scalability.
Lecturas : [Low+12], [Mal+10], [Bal+08]	