CIENCIA DE LA COMPUTACIÓN

BIG DATA

3 CRÉDITOS

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

CE2A

ÍNDICE

ASIGNATURA 3

DATOS GENERALES 3

Ciclo: 9º 3

Créditos: tres (3) créditos 3

Horas de teoría: una (1) semanal 3

Horas de práctica: cuatro (4) semanales 3

Duración del período: dieciséis (16) semanas 3

Condición: 3

Modalidad: Virtual 3

Requisitos: 3

PROFESORES 3

Profesor coordinador del curso 3

Profesor(es) instructor(es) del curso 3

INTRODUCCIÓN AL CURSO 3

OBJETIVOS 4

COMPETENCIAS Y CRITERIOS DE DESEMPEÑO 4

RESULTADOS DE APRENDIZAJE 5

TEMAS 5

PLAN DE TRABAJO 6

Metodología 6

Sesiones de teoría 6

Sesiones de práctica 6

SISTEMA DE EVALUACIÓN 6

REFERENCIAS BIBLIOGRÁFICAS 7

UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA
SILABO 2021-1

1. ASIGNATURA

CS3700 - Big Data

2. DATOS GENERALES

2.1 Ciclo: 9º
2.2 Créditos: tres (3) créditos
2.3 Horas de teoría: una (1) semanal
2.4 Horas de práctica: cuatro (4) semanales
2.5 Duración del período: dieciséis (16) semanas
2.6 Condición:
- Obligatorio para Ciencia de la Computación

2.7 Modalidad: Virtual
2.8 Requisitos:
- CS2702 – Bases de Datos II
- CS3P01 – Computación Paralela y Distribuida

3. PROFESORES

3.1 Profesor coordinador del curso

Frizzi Alejandra San Roman Salazar(fsanroman@utec.edu.pe)
Horario de atención: lunes y viernes de 10:00 a 11:00 am

3.2 Profesor(es) instructor(es) del curso

Wilder Nina Choquehuayta (wnina@utec.edu.pe)
Horario de atención: lunes y viernes de 10:00 a 11:00 am

4. INTRODUCCIÓN AL CURSO

El curso, de naturaleza teórico-práctica está diseñado para que los estudiantes de la carrera de Ciencia de la Computación profundicen su conocimiento para almacenamiento, análisis, procesamiento y gestión de grandes volúmenes de información (terabytes, petabytes e inclusive exabytes). Cada día, cada hora, cada minuto se genera gran cantidad de información la cual necesita ser procesada, almacenada, analizada en tiempo real o próximo a real.

Los temas principales que se revisarán en este curso son relacionados al manejo de herramientas como hadoop, hdfs, mapreduce, yarn, spark que permiten al alumno trabajar con grandes conjuntos de datos.

5. OBJETIVOS

- **Sesión 1**: Identificar conceptos de big data y aplicaciones en diferentes entornos: médico, ambiental, comercial, entre otros.

- **Sesión 2**: Identificar conceptos y utilizar herramientas relacionadas con big data: Hadoop.

- **Sesión 3**: Identificar conceptos y utilizar herramientas relacionadas con big data: HDFS.

- **Sesión 4**: Identificar conceptos y utilizar herramientas relacionadas con big data: MapReduce.

- **Sesión 5**: Identificar conceptos y utilizar herramientas relacionadas con big data: Spark.

- **Sesión 6**: Modelar e implementar soluciones de big data escalables para datos de cualquier naturaleza utilizando las herramientas como HDFS, Spark, HBase, Cassandra entre otras.

- **Sesión 7**: Identificar y utilizar herramientas relacionadas con big data utilizando procesamiento de grafos a larga escala.

6. COMPETENCIAS Y CRITERIOS DE DESEMPEÑO

Los criterios de desempeño que se van a trabajar en este curso son:

1.3. Aplica conocimientos de computación apropiados para la solución de problemas definidos y sus requerimientos en la disciplina del programa (*nivel 3*).

2.4. Resuelve problemas de computación y otras disciplinas relevantes en el dominio (*nivel 3*).

3.2. Diseña, implementa y evalúa soluciones a problemas complejos de computación (*nivel 2*).

4.1. Crea, selecciona, adapta y aplica técnicas, recursos y herramientas modernas para la práctica de la computación y comprende sus limitaciones (*nivel 3*).

7. RESULTADOS DE APRENDIZAJE

Al final del curso de Big Data se espera que el estudiante sea capaz:

RA1. Construir aplicaciones paralelas para el procesamiento de grandes volúmenes de datos

RA2. Desarrollar aplicaciones que satisfagan requerimientos del mundo real.

RA3. Inferir modelos enfocados al paradigma map-reduce para el procesamiento de grandes volúmenes de datos.

RA4. Adaptar paradigmas y frameworks para resolver problemas escalables y complejos.

8. TEMAS

1. **Introducción a Big Data**
   - 1.1. Características de Big Data
   - 1.2. Ecosistemas y arquitectura de Big Data
   - 1.3. Visión global sobre sistema de Archivos Distribuidos y Procesamiento
   - 1.4. Infraestructura On Premise y Cloud
   - 1.5. Aplicaciones de Big Data

2. **Hadoop**
   - 2.1. Visión global de Hadoop
   - 2.2. Estructura y seguridad de Hadoop
   - 2.3. HDFS: Hadoop Distributed File System
   - 2.4. Modelo de Programación MapReduce
   - 2.5. Aplicaciones del uso de MapReduce
   - 2.6. Procesamiento e ingesta de datos en lote y real time

3. **Spark**
   - 3.1. Arquitectura de Spark
   - 3.2. Spark RDD, DataFrame y Dataset
   - 3.3. Spark SQL
   - 3.4. Spark Streaming
   - 3.5. Spark Machine Learning Library
   - 3.6. Spark GraphX

4. **Procesamiento de Grafos en larga escala**
   - 4.1. Pregel: A System for Large-scale Graph Processing
   - 4.2. Distributed GraphLab
   - 4.3. Apache Giraph is an iterative graph processing system built for high scalability.

9. PLAN DE TRABAJO

9.1 Metodología

Este curso presenta por metodología activa el aprendizaje clásico y el aprendizaje basado en problemas; ambos son fundamentales para introducir al estudiante a los conceptos básicos y afianzar la base necesaria para los siguientes cursos de carrera.

9.2 Sesiones de teoría

Las sesiones teóricas serán desarrolladas bajo la estructura de clase magistral. El desarrollo de las sesiones teóricas están focalizadas en el estudiante, a través de la participación activa con el uso de preguntas abiertas y cerradas.

9.3 Sesiones de práctica

Las sesiones prácticas se desarrollarán a través de una metodología activa generando el aprendizaje práctico por parte del estudiante. Las sesiones de práctica se caracterizan por el desarrollo de ejercicios modelos y aplicados en base a los conceptos teóricos aprendidos.

10. SISTEMA DE EVALUACIÓN

	TEORÍA	**PRÁCTICA Y/O LABORATORIO**
EVALUACIÓN *La ponderación de la evaluación se hará si ambas partes están aprobadas	Evaluación Continua **C1** (5%) Práctica Calificada **PC1** (5%) Práctica Calificada **PC2** (5%) Práctica Calificada **PC3** (5%) Práctica Calificada **PC4** (5%) Práctica Calificada **PC5** (5%)	Laboratorio **L1** (14%) Proyecto **P1** (56%)
	30%	70%
	100%	

Las rúbricas que permitirán medir las actividades más significativas del curso y que, además se relacionan con la evaluación de las competencias del estudiante son: enlace

11. REFERENCIAS BIBLIOGRÁFICAS

Shumeet Baluja. "*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.

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.

George Coulouris et al. *Distributed Systems: Concepts and Design. 5th.* USA: Addison-Wesley Publishing Company, 2011. isbn: 0132143011, 9780132143011.

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.

Yucheng Low. "*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.

Grzegorz Malewicz. "*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.