UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

SÍLABO

1. INFORMACIÓN GENERAL

1.1 Curso	:	Big Data y Analítica de Datos
1.2 Código	:	CS370
1.3 Programa Académico	:	Computación Científica
1.4 Créditos	:	5
1.5 Prerrequisitos	:	CS271. Bases de Datos I. (4^{to} Sem)
1.6 Tipo de curso	:	Electivo
1.7 Área de Estudios	:	Estudios Generales
1.8 Número de Semanas	:	16
1.9 Horas por semana	:	3 HT; 4 HP;
1.10 Ciclo	:	I
1.11 Modalidad de aprendizaje	:	Presencial
1.12 Periodo Académico	:	2026-I
1.13 Inicio del semestre	:	
1.14 Fin del semestre	:	

2. PROFESORES

Atención previa coordinación con el profesor

3. INTRODUCCIÓN AL CURSO

En la era actual, el volumen, velocidad y variedad de los datos superan las capacidades de los sistemas tradicionales de almacenamiento y procesamiento. Este curso introduce los conceptos y tecnologías fundamentales del Big Data. Cubre sistemas de archivos distribuidos, frameworks de procesamiento paralelo, bases de datos NoSQL y flujo de datos en tiempo real. Los estudiantes aprenderán a diseñar arquitecturas escalables para extraer valor de conjuntos de datos masivos.

4. OBJETIVOS

- Comprender las 5 V's del Big Data y su impacto en el diseño de sistemas.
- Diseñar e implementar soluciones escalables usando el paradigma MapReduce.
- Dominar el uso de sistemas de archivos distribuidos como HDFS.
- Analizar diferentes patrones arquitectónicos para Big Data (Lambda y Kappa).
- Aplicar técnicas de procesamiento de flujos para análisis de datos en tiempo real.

5. RESULTADOS DEL ESTUDIANTE

ABET-3) Habilidad para desarrollar y llevar a cabo experimentos o probar hipótesis, analizar e interpretar datos y utilizar el juicio científico para obtener conclusiones. (Usage)

ABET-2) Habilidad para formular o diseñar un sistema, proceso, procedimiento o programa para el propósito previsto. (Usage)

6. TEMAS

Unidad 1: Conceptos Fundamentales de Sistemas de Bases de Datos (12 horas)	
Resultados esperados: ABET-2,ABET-3	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Propósito y ventajas de los sistemas de bases de datos ▪ Componentes de los sistemas de bases de datos ▪ Diseño de funciones fundamentales de DBMS (por ejemplo, mecanismos de consulta, gestión de transacciones, gestión de búfer, métodos de acceso) ▪ Arquitectura de base de datos, independencia de datos y abstracción de datos ▪ Gestión de transacciones ▪ Normalización ▪ Enfoques para gestionar grandes volúmenes de datos (por ejemplo, sistemas de bases de datos NoSQL, uso de MapReduce) ▪ Cómo soportar aplicaciones solo de CRUD ▪ Bases de datos distribuidas/sistemas basados en la nube ▪ Datos estructurados, semiestructurados y no estructurados ▪ Uso de un lenguaje de consulta declarativo ▪ Sistemas que soportan contenido estructurado y/o de flujo	▪ Identificar al menos cuatro ventajas que proporciona el uso de un sistema de base de datos [Analizar] ▪ Enumerar los componentes de un sistema de base de datos (relacional) [Enumerar] ▪ Seguir una consulta a medida que es procesada por los componentes de un sistema de base de datos (relacional) [Analizar] ▪ Defender el valor de la independencia de datos [Defender] ▪ Componer una consulta simple de selección-proyección-unión en SQL [Componer] ▪ Enumerar las cuatro propiedades de un gestor de transacciones correcto [Enumerar] ▪ Describir las ventajas de eliminar datos duplicados repetidos [Describir] ▪ Esbozar cómo MapReduce usa paralelismo para procesar datos eficientemente [Esquematizar/Esbozar] ▪ Evaluar las diferencias entre bases de datos estructuradas y semiestructuradas/no estructuradas [Evaluar]
Lecturas : [**White2015**], [**Karau2017**]	

Unidad 2: Bases de Datos Distribuidas/Computación en la Nube (12 horas)	
Resultados esperados: ABET-2,ABET-3	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ DBMS Distribuido: • Almacenamiento de datos distribuido • Procesamiento de consultas distribuido • Modelo de transacción distribuida • Soluciones homogéneas y heterogéneas • Bases de datos distribuidas cliente-servidor ▪ DBMS Paralelo: • Arquitecturas de DBMS paralelo: memoria compartida, disco compartido, nada compartido • Aceleración y escalamiento, por ejemplo, uso del modelo de procesamiento MapReduce • Replicación de datos y modelos de consistencia débil	▪ Describir los componentes clave de un DBMS distribuido, incluyendo almacenamiento de datos distribuido, procesamiento de consultas y gestión de transacciones [Describir] ▪ Analizar las ventajas y desventajas entre arquitecturas de DBMS paralelo: memoria compartida, disco compartido y nada compartido [Analizar] ▪ Describir estrategias de replicación de datos y modelos de consistencia débil en sistemas de bases de datos distribuidas [Describir]
Lecturas : [**White2015**], [**Lin2010**]	

Unidad 3: Sistemas NoSQL (12 horas)	
Resultados esperados: ABET-2,ABET-3	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ ¿Por qué NoSQL? (por ejemplo, Desajuste de impedancia entre Aplicación [CRUD] y RDBMS) ▪ Modelo de datos Clave-Valor y Documento ▪ Sistemas de almacenamiento (por ejemplo, sistemas Clave-Valor, Lagos de Datos (*Data Lakes*)) ▪ Modelos de Distribución (Fragmentación y Replicación) ▪ Bases de Datos de Grafos ▪ Modelos de Consistencia (Actualización y Lectura, consistencia de quórum, teorema CAP) ▪ Modelo de procesamiento (por ejemplo, MapReduce, map-reduce multi-etapa, map-reduce incremental) ▪ Estudios de Caso: Sistema de almacenamiento en la nube (por ejemplo, S3); Bases de datos de grafos; ”Cuándo no usar NoSQL”	▪ Desarrollar un caso de uso para el uso de NoSQL sobre RDBMS [Crear] ▪ Describir las características definitorias detrás de los modelos de datos basados en Clave-Valor y Documentos [Describir]
Lecturas : [**Sadalage2012**], [**Kleppmann2017**]	

Unidad 4: Análisis de Datos (12 horas)	
Resultados esperados: ABET-2,ABET-3	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Técnicas exploratorias de datos (motivación, representación, estadísticas descriptivas, visualizaciones) ▪ Ciclo de vida de ciencia de datos: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación, despliegue y aceptación del usuario ▪ Algoritmos de minería de datos y aprendizaje automático: por ejemplo, clasificación, agrupamiento, asociación, regresión ▪ Adquisición y gobernanza de datos ▪ Consideraciones de seguridad y privacidad de datos ▪ Equidad y sesgo de datos ▪ Técnicas de visualización de datos y su uso en análisis de datos ▪ Resolución de Entidades	▪ Describir varios enfoques de exploración de datos, incluyendo visualización, para comprender conjuntos de datos no familiares [Describir] ▪ Aplicar varios enfoques de exploración de datos para comprender conjuntos de datos no familiares [Aplicar] ▪ Describir algoritmos básicos de aprendizaje automático/minería de datos y cuándo son apropiados para su uso [Describir] ▪ Aplicar varios algoritmos de aprendizaje automático/minería de datos [Aplicar] ▪ Describir consideraciones legales y éticas en la adquisición, uso y modificación de conjuntos de datos [Describir] ▪ Describir problemas de equidad y sesgo en la recolección y uso de datos [Describir]
Lecturas : [**zaharia2012resilient**], [**chambers2018spark**], [**Kleppmann2017**], [**Psaltis2017**]	

7. PLAN DE TRABAJO

7.1 Metodología

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.

7.2 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.

7.3 Sesiones Prácticas

Las sesiones prácticas se llevan en clase donde se desarrollan una serie de ejercicios y/o conceptos prácticos mediante planteamiento de problemas, la resolución de problemas, ejercicios puntuales y/o en contextos aplicativos.

8. SISTEMA DE EVALUACIÓN

********* EVALUATION MISSING ********

9. BIBLIOGRAFÍA BÁSICA