National University of Engineering (UNI)
School of Computer Science
Syllabus 2026-I

1. COURSE

CS370. Big Data (Mandatory)

2. GENERAL INFORMATION

2.1 Course	:	CS370. Big Data
2.2 Semester	:	9^{th} Semester.
2.3 Credits	:	3
2.4 Hours	:	1 HT; 4 HP;
2.5 Duration of the period	:	16 weeks
2.6 Type of course	:	Mandatory
2.7 Learning modality	:	Face to face
2.8 Prerrequisites	:	CS3P1. Parallel and Distributed Computing . (8^{th} Sem)

3. PROFESSORS

Meetings after coordination with the professor

4. INTRODUCTION TO THE COURSE

In the current era, the volume, velocity, and variety of data exceed the capabilities of traditional storage and processing systems. This course introduces the fundamental concepts and technologies of Big Data. It covers distributed file systems, parallel processing frameworks, NoSQL databases, and real-time data streaming. Students will learn to design scalable architectures to extract value from massive datasets.

5. GOALS

- Understand the 5 V's of Big Data and their impact on system design.
- Design and implement scalable solutions using the MapReduce paradigm.
- Master the use of distributed file systems like HDFS.
- Analyze different architectural patterns for Big Data (Lambda and Kappa).
- Apply stream processing techniques for real-time data analysis.

6. COMPETENCES

AG-C11) Use of Tools: Applies modern computing tools in problem solving. (Usage)

AG-C09) Design and Development of Solutions: Designs, implements, and evaluates solutions for complex computing problems. (Usage)

7. TOPICS

Unit 1: Conceptos Fundamentales de Sistemas de Bases de Datos (12 hours)	
Competences Expected: AG-C09,AG-C11	
Topics	**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]
Readings : [Whi15], [Kar+17]	

Unit 2: Bases de Datos Distribuidas/Computación en la Nube (12 hours)	
Competences Expected: AG-C09,AG-C11	
Topics	**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]
Readings : [Whi15], [LD10]	

Unit 3: Sistemas NoSQL (12 hours)	
Competences Expected: AG-C09,AG-C11	
Topics	**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]
Readings : [SF12], [Kle17]	

Unit 4: Análisis de Datos (12 hours)	
Competences Expected: AG-C09,AG-C11	
Topics	**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]
Readings : [Mat+12], [BM18], [Kle17], [Psa17]	

8. WORKPLAN

8.1 Methodology

Individual and team participation is encouraged to present their ideas, motivating them with additional points in the different stages of the course evaluation.

8.2 Theory Sessions

The theory sessions are held in master classes with activities including active learning and roleplay to allow students to internalize the concepts.

8.3 Practical Sessions

The practical sessions are held in class where a series of exercises and/or practical concepts are developed through problem solving, problem solving, specific exercises and/or in application contexts.

9. EVALUATION SYSTEM

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

10. BASIC BIBLIOGRAPHY

[LD10] Jimmy Lin and Chris Dyer. *Data-Intensive Text Processing with MapReduce*. Morgan and Claypool Publishers, 2010.

[Mat+12] Zaharia Matei et al. "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing". In: *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association, 2012.

[SF12] Pramod J. Sadalage and Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley, 2012.

[Whi15] Tom White. *Hadoop: The Definitive Guide*. 4th. O'Reilly Media, 2015.

[Kar+17] Holden Karau et al. *Learning Spark: Lightning-Fast Big Data Analysis*. O'Reilly Media, 2017.

[Kle17] Martin Kleppmann. *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media, 2017.

[Psa17] Andrew Psaltis. *Streaming Data: Understanding the real-time pipeline*. Manning Publications, 2017.

[BM18] Chambers Bill and Zaharia Matei. *Spark: The Definitive Guide: Big Data Processing Made Simple*. O'Reilly Media, 2018.