NATIONAL UNIVERSITY OF SAN MARCOS

FACULTY OF MATHEMATICAL SCIENCES

SYLLABUS

1. GENERAL INFORMATION

1.1 Course name	:	Big Data and Analytics
1.2 Code	:	CS370
1.3 Academic Program	:	Scientific Computing
1.4 Credits	:	5
1.5 Prerrequisites	:	CS271. Data Management. (4^{th} Sem)
1.6 Type of Course	:	Elective
1.7 Area of Studies	:	Estudios Generales
1.8 Number of Weeks	:	16
1.9 Hours per week	:	3 HT; 4 HP;
1.10 Cycle	:	I
1.11 Learning modality	:	Face to face
1.12 Academic Period	:	2026-I
1.13 Semester Start Date	:	
1.14 Semester End Date	:	

2. PROFESSORS

Meetings after coordination with the professor

3. 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.

4. 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.

5. COMPETENCES

ABET-3) An ability to develop and conduct experiments or test hypotheses, analyze and interpret data, and use scientific judgment to draw conclusions. (Usage)

ABET-2) An ability to formulate or design a system, process, procedure, or program for the intended purpose. (Usage)

6. TOPICS

Unit 1: Core Database System Concepts (12 hours)	
Competences Expected: ABET-2,ABET-3	
Topics	**Learning Outcomes**
• Purpose and advantages of database systems • Components of database systems • Design of core DBMS functions (e.g., query mechanisms, transaction management, buffer management, access methods) • Database architecture, data independence, and data abstraction • Transaction management • Normalization • Approaches for managing large volumes of data (e.g., NoSQL database systems, use of MapReduce) • How to support CRUD-only applications • Distributed databases/cloud-based systems • Structured, semi-structured, and unstructured data • Use of a declarative query language • Systems supporting structured and/or stream content	• Identify at least four advantages that using a database system provides [Analyze] • Enumerate the components of a (relational) database system [Enumerate] • Follow a query as it is processed by the components of a (relational) database system [Analyze] • Defend the value of data independence [Defend] • Compose a simple select-project-join query in SQL [Compose] • Enumerate the four properties of a correct transaction manager [Enumerate] • Describe the advantages for eliminating duplicate repeated data [Describe] • Outline how MapReduce uses parallelism to process data efficiently [Outline] • Evaluate the differences between structured and semi/unstructured databases [Evaluate]
Readings : [**White2015**], [**Karau2017**]	

Unit 2: Distributed Databases/Cloud Computing (12 hours)	
Competences Expected: ABET-2,ABET-3	
Topics	**Learning Outcomes**
• Distributed DBMS: – Distributed data storage – Distributed query processing – Distributed transaction model – Homogeneous and heterogeneous solutions – Client-server distributed databases • Parallel DBMS: – Parallel DBMS architectures: shared memory, shared disk, shared nothing – Speedup and scale-up, e.g., use of the MapReduce processing model – Data replication and weak consistency models	• Describe the key components of a distributed DBMS, including distributed data storage, query processing, and transaction management [Describe] • Analyze the tradeoffs between parallel DBMS architectures: shared memory, shared disk, and shared nothing [Analyze] • Describe data replication strategies and weak consistency models in distributed database systems [Describe]
Readings : [**White2015**], [**Lin2010**]	

Unit 3: NoSQL Systems (12 hours)	
Competences Expected: ABET-2,ABET-3	
Topics	**Learning Outcomes**
• Why NoSQL? (e.g., Impedance mismatch between Application [CRUD] and RDBMS) • Key-Value and Document data model • Storage systems (e.g., Key-Value systems, Data Lakes) • Distribution Models (Sharding and Replication) • Graph Databases • Consistency Models (Update and Read, Quorum consistency, CAP theorem) • Processing model (e.g., Map-Reduce, multi-stage map-reduce, incremental map-reduce) • Case Studies: Cloud storage system (e.g., S3); Graph databases; "When not to use NoSQL"	• Develop a use case for the use of NoSQL over RDBMS [Create] • Describe the defining characteristics behind Key-Value and Document-based data models [Describe]
Readings : [**Sadalage2012**], [**Kleppmann2017**]	

Unit 4: Data Analytics (12 hours)	
Competences Expected: ABET-2,ABET-3	
Topics	**Learning Outcomes**
• Exploratory data techniques (motivation, representation, descriptive statistics, visualizations) • Data science lifecycle: business understanding, data understanding, data preparation, modeling, evaluation, deployment, and user acceptance • Data mining and machine learning algorithms: e.g., classification, clustering, association, regression • Data acquisition and governance • Data security and privacy considerations • Data fairness and bias • Data visualization techniques and their use in data analytics • Entity Resolution	• Describe several data exploration approaches, including visualization, to understanding unfamiliar datasets [Describe] • Apply several data exploration approaches to understanding unfamiliar datasets [Apply] • Describe basic machine learning/data mining algorithms and when they are appropriate for use [Describe] • Apply several machine learning/data mining algorithms [Apply] • Describe legal and ethical considerations in acquiring, using, and modifying datasets [Describe] • Describe issues of fairness and bias in data collection and usage [Describe]
Readings : [**zaharia2012resilient**], [**chambers2018spark**], [**Kleppmann2017**], [**Psaltis2017**]	

7. WORKPLAN

7.1 Methodology

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

7.2 Theory Sessions

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

7.3 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.

8. EVALUATION SYSTEM

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

9. BASIC BIBLIOGRAPHY