ADVANCED DATA STRUCTURES AND STORAGE

Bachelor in Data and Business Analytics BDBA SEP-2023
ADSS-DBA.2.M.A

Area Data Science

Number of sessions: 30

Academic year: 23-24

Degree course: SECOND

Number of credits: 6.0

Semester: 2º

Category: COMPULSORY

Language: English

Professor: **EDUARDO RODRÍGUEZ LORENZO**

E-mail: erodriguezl@faculty.ie.edu

EDUARDO RODRÍGUEZ LORENZO

Eduardo Rodriguez Lorenzo is Senior Manager at NETSCOUT and Adjunct Professor at IE School of Science and Technology. He is a technologist specializing in Telecommunication Networks, Cybersecurity, Software Architecture, Data Engineering and Analytics.

He studied at UPM (Universidad Politécnica de Madrid), King's College London and London University.

At NETSCOUT, he leads a global team of Data and Network Engineers with a strong focus on Network Service Assurance, Cybersecurity, Data Engineering and Analytics.

He has gained broad international experience delivering high-value Consulting Services (Customer Experience & Customer Journeys, Business Intelligence, Service Assurance, Data Monetization, Process Engineering...) and Data-driven Solutions (Cloud & Backend Architecture, Data Feeds, Database, Dashboard, Interaction & Visualisation Design) to global Enterprises and Communication Service Providers. He has played an active role in the launch, measurement and optimisation of Mobile Networks for various top international Telcos.

He is a member of the Spanish Charter of Telecommunications Engineers (COIT) where he is an active member of the Telecommunications Policy and Regulation Group and the Digital Transformation Group.

His main interests include Disruptive Technologies, Data Engineering Architectures, Networks, Distributed Systems and Graph technology.

He joined IE University in 2020.

IE | LinkedIn | Twitter

Office Hours

Office hours will be on request. Please contact at:

erodriguezl@faculty.ie.edu

SUBJECT DESCRIPTION

In an era where data storage and management technologies are advancing at an unprecedented rate, the "Advanced Data Structures and Storage" course offers a timely and insightful exploration into this rapidly evolving field. Mirroring the exponential growth trajectory outlined by a Moore's Law analog for storage capacity, this course delves into the intricate world of scalable storage systems, addressing the burgeoning challenges of data volume, scalability, security, and privacy.

This comprehensive 30-session course is a natural progression from the foundational "Algorithms and Data Structures" course. It aims to equip students with a nuanced understanding of advanced enterprise-level data structures and storage technologies.

The first part of the course covers topics such as tensors (critical in high-performance computing and AI applications), data compression algorithms, error detection mechanisms, and enterprise storage solutions, including cloud storage.

The second part of the course places a significant emphasis on understanding and utilizing NoSQL databases, including column-family, graph, and document-based systems, which are pivotal in managing the complexity of modern data sets.

Furthermore, students will gain insights into the latest trends and future directions in data management.

LEARNING OBJECTIVES

Understanding of Multidimensional Arrays and Vectors: Gain a comprehensive understanding of multidimensional arrays and vectors, their significance in high-performance computing and machine learning, and practical applications in Python.
Fundamentals of Data Management: Develop a solid foundation in data management principles, focusing on memory and disk storage techniques, including an understanding of volatile and non-volatile memory and storage technologies.
Data Compression Techniques: Acquire knowledge and skills in data compression and transmission, understanding both lossy and lossless compression methods, and their applications in various formats and codecs.
Error Detection and Correction Strategies: Learn about error detection and correction methods crucial for reliable data storage and transmission, including hash functions, parity bit coding, and error control coding.
Mastery of NoSQL Databases: Become proficient in non-relational (NoSQL) database systems, including key-value, document, column-family, and graph databases, and their use in managing large-scale, unstructured data.
Vector Database Utilization: Understand the role and implementation of vector databases in data storage and retrieval, with hands-on experience in Python libraries like Milvus and Pinecone for applications like image similarity search.
Insights into Data Center and Cloud Environment Management: Delve into data center architectures and cloud computing environments, covering aspects like cloud storage models, virtualization, and containerization with tools like Docker.
Advanced Data Structure Topics: Explore advanced topics such as distributed and blockchain storage, scalable databases, differential privacy, and the impact of disruptive technologies like quantum computing on data management.
Critical Analysis and Decision-Making in Data Structures: Develop critical thinking skills to assess the pros and cons of various data structures and storage technologies, enhancing decision-making in complex data management scenarios.
Practical Application and Hands-On Experience: Through labs and project work, apply theoretical knowledge to real-world scenarios, gaining practical experience in managing and manipulating multidimensional vectors, databases, and advanced data structures.

With these learning objectives, we aim to cultivate your critical thinking, decision-making skills, and practical expertise, preparing you for the dynamic challenges and opportunities in the field of data science and storage

TEACHING METHODOLOGY

IE University teaching method is defined by its collaborative, active, and applied nature. Students actively participate in the whole process to build their knowledge and sharpen their skills. Professor’s main role is to lead and guide students to achieve the learning objectives of the course. This is done by engaging in a diverse range of teaching techniques and different types of learning activities such as the following:

Learning Activity	**Weighting**	**Estimated time a student should dedicate to prepare for and participate in**
Lectures	26.67 %	40.0 hours

Discussions	13.33 %	20.0 hours
Exercises in class, Asynchronous sessions, Field Work	20.0 %	30.0 hours
Group work	20.0 %	30.0 hours
Individual studying	20.0 %	30.0 hours
TOTAL	100.0 %	150.0 hours

PROGRAM

SESSION 1 (LIVE IN-PERSON)

WELCOME/COURSE ORGANIZATION

Introduction to the course: logistics, evaluation system, and content at-a-glance.

SESSION 2 (LIVE IN-PERSON)

MULTIDIMENSIONAL ARRAYS

The curse of dimensionality. Memory ordering; endianness; strides. Common formats. Trends: the role of multidimensional arrays in high-performance computing and machine learning; tensor processing units. Multidimensional arrays in Python.

SESSION 3 (LIVE IN-PERSON)

MULTIDIMENSIONAL ARRAYS

Sparse Matrices: Coordinate list, CSR, generation functions.

Numpy Lab:

- Basic Attributes of the ndarray Class
- Basic Numerical Data Types Available in NumPy
- Summary of NumPy Functions for Generating
- Arrays Created from Lists and Other Array-Like Objects
- Arrays Filled with Constant Values
- Arrays Filled with Incremental Sequences
- Arrays Filled with Logarithmic Sequences
- Meshgrid Arrays: Multidimensional coordinate grids
- Creating Arrays with Properties of Other Arrays
- Creating Matrix Arrays
- Indexing and Slicing
- Multidimensional Arrays
- Views
- Fancy Indexing
- Boolean-Valued Indexing
- Visual summary of indexing methods for NumPy arrays
- Reshaping and Resizing
- Vertical and Horizontal Stacking

- Vectorized Expressions
- Operators for Elementwise Arithmetic Operation on NumPy Arrays
- Elementwise Functions
- Boolean Arrays and Conditional Expressions
- Set Operations
- Operations on Arrays
- Matrix and Vector Operations

SESSION 4 (LIVE IN-PERSON)

DATA COMPRESSION

Lossy versus lossless compression. Run-length and entropy encoding. Important compression formats and multimedia codecs. Signal Processing principles.

SESSION 5 (LIVE IN-PERSON)

DATA COMPRESSION

Applications of Information Theory. Source Code Theorem. Source Code Theorem. Number Systems. Review of most popular Compression Algorithms

SESSION 6 (LIVE IN-PERSON)

ERROR DETECTION AND CORRECTION

The role of hash functions for check summing. Error detection and error correction strategies for data storage and transmission over unreliable channels such as the Internet. Watermarking and steganography.

SESSION 7 (LIVE IN-PERSON)

ERROR DETECTION AND CORRECTION

Error Control Coding. Cyclic Redundcy Check (CRC) and Forward Error Correction (FEC) codes. Parity Bit Coding. Hamming Distance.

SESSION 8 (LIVE IN-PERSON)

VECTOR DATABASES

Introduction to Vector Databases

The Role of Vectors in Data Storage and Retrieval

Python Libraries for Vector Database Management: Milvus, Pinecone,...

Hands-On Demonstration: Image Similarity Search with Milvus

SESSION 9 (LIVE IN-PERSON)

DISK AND MEMORY MANAGEMENT

Volatile vs Non-volatile memory. Storage technologies: HDD, SSD, Magnetic, Optical. Filesystems. RAID technology:striping, mirroring, and parity. Business Continuity. Information Availability concept. Data Serialization.

SESSION 10 (LIVE IN-PERSON)

DISK AND MEMORY MANAGEMENT

- The concepts of Average Access Time and Response Time
- Common Metrics for Storage Devices
- Basic Storage Performance Metrics: IOPS, Bandwidth (a.k.a throughput), Latency (a.k.a Response Time)
- Enterprise Storage Solutions: DAS, NAS and SAN

SESSION 11 (LIVE IN-PERSON)

DATA CENTER AND CLOUD ENVIRONMENT

- Data Center Environment
- NIST Model of Cloud Computing
- Cloud Deployment Models (IaaS, PaaS, SaaS)
- Cloud Computing Service Models (Public, Private, Hybrid)
- Cloud Storage Basis: Block vs File vs Object Storage
- Backup as a Service
- Cloud Storage Solutions: Azure and AWS.

Azure Storage Practice Lab

SESSION 12 (LIVE IN-PERSON)

VIRTUALIZATION, DATA CENTERS AND CLOUDS

Virtualization and Containers. VMWare and Docker.

Docker Lab: Setting up a containerized Jupyter Server.

SESSION 13 (LIVE IN-PERSON)

IN-CLASS QUIZ AND RECAP

Multiple-choice/answers quiz.

SESSION 14 (LIVE IN-PERSON)

RECAP, Q&A

Review session ahead of Mid-term Exam.

SESSION 15 (LIVE IN-PERSON)

MIDTERM EXAM

SESSION 16 (LIVE IN-PERSON)

INTRODUCTION TO NOSQL

SQL, its context, and disadvantages. Principles of NoSQL and non-relational databases as a response to SQL's limitations. Types of NoSQL: key-value, document, column-family, and graph databases. Normalized vs. denormalized data.

SESSION 17 (LIVE IN-PERSON)

KEY-VALUE DATABASES

Key-value databases: Putting hash tables to use. Design principles. Key-value architecture and Use Cases. Limitations. Python sets, dictionaries, shelve and pickle. HDF5. Redis.

SESSION 18 (LIVE IN-PERSON)

REDIS LAB

REDIS Setup. Basic Commands. Data Types. Transactions. Geospatial Features. HyperLogLog. Python SDK.

SESSION 19 (LIVE IN-PERSON)

DOCUMENT DATABASES

Operations on document collections. Partitioning and sharding. Query processing. Basics of MongoDB and PyMongo.

SESSION 20 (LIVE IN-PERSON)

MONGODB LAB

MongoDB Setup. Simple Queries. Running Javascript. Aggregations.

SESSION 21 (LIVE IN-PERSON)

GRAPH TECHNOLOGY

Graphs basics and definitions. Algorithms. Graph Analytics packages. Graph Database Technology.

SESSION 22 (LIVE IN-PERSON)

GRAPH ANALYTICS LAB

Using NetworkX to model graphs and networks.

SESSION 23 (LIVE IN-PERSON)

NEO4J LAB

Neo4j Setup. Graph Data Modelling. Cypher Query Language. Querying data. Creating Nodes and Relationships. Aggregations and Graph Functions. Date and TCalculations.

SESSION 24 (LIVE IN-PERSON)

NEO4J LAB

Neo4j Aggregations and Functions. Knowledge Graph modelling.

SESSION 25 (LIVE IN-PERSON)

COLUMNAR FAMILY DATABASES. PROBABILISTIC DATA STRUCTURES

Interpretation and components. The advantages of denormalization. Processes and protocols for maintaining a column family database. Differences with relational databases. Probabilistic Data Structures: bloom filters.

SESSION 26 (LIVE IN-PERSON)

CASSANDRA LAB.

Cassandra Data Model. Cassandra Distributed Architecture: Nodes, Rings, Tokens, Replication. CQL (Cassadra Query Language). Data Types and Collections. Secondary Indexes. Materialized Views.

SESSION 27 (LIVE IN-PERSON)

TEAM DEMOS I

Team project demo about an agreed course topic.

SESSION 28 (LIVE IN-PERSON)

TEAM DEMOS II

Team project demo about an agreed course topic.

SESSION 29 (LIVE IN-PERSON)

IN-CLASS QUIZ AND RECAP

Multiple-choice/answers quiz

SESSION 30 (LIVE IN-PERSON)

FINAL EXAM

EVALUATION CRITERIA

criteria	**percentage**	**Learning Objectives**	**Comments**
Class Participation	15 %		
Group Project	20 %		
Individual Work	15 %		
Midterm Exam	25 %		
Final Exam	25 %		

RE-SIT / RE-TAKE POLICY

Class Participation

This includes quizzes, discussion board (forum) activity, and participation in in-class activities.

Students will have to individually solve two in-class quizzes in sessions 13 and 29, each weighing 5% of the final grade.

Throughout the semester, readings will be shared by the Professor on the discussion board, about current topics (for example blockchain storage, scalable databases, differential privacy, cloud storage, predictive data management, ...). Students are expected to read and contribute to the forum with the own views and resources.

Individual Assignments

Multiple compulsory individual assignments will have to be completed for each of the course blocks, including computational thinking & algorithmic problems, coding exercises and system setup scenarios. All assignments will be published on Blackboard.

Group Projects

Students will work in groups on a research activity involving one course topic and a practical implementation and demo.

Midterm Exam

The midterm exam will comprise all materials given in class, including readings, up to session 14.

Final Exam

The final exam will include all the materials given in class, including readings, from session 16 and it will be held during session 30.

Minimal Marks

A minimum passing grade in the midterm and final exam (3.5) is required to pass the subject. If a student scores lower than this minimum, he will have to go to June retake, irrespective of their overall course grade. The overall passing course grade is 5.0. All the presentations/videos/exams will be submitted via Blackboard. No other option will be accepted.

BIBLIOGRAPHY

Recommended

- Gnanasundaram & Shrivastava. (2012). *Information Storage and Management.* Wiley & Sons. ISBN 9788126537501 (Printed)

- Sullivan. (2015). *NoSQL for Mere Mortals.* Addison-Wesley. ISBN 0134023218 (Digital)

- Colton McAnlis, Aleks Haecky. (2016). *Understanding Compression.* O'Reilly Media, Inc.. ISBN 9781491961537 (Digital)

- Harrison. (2015). *Next Generation Databases.* Apress. ISBN 9781484240243 (Printed)

BEHAVIOR RULES

Please, check the University's Code of Conduct here. The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy here. The Program Director may provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code here. The Program Director may provide further indications.