UNIVERSITY of NICOSIA

Course Syllabus

Course Code	Course Title	ECTS Credits
COMP-548DL	Big Data Management and Processing	10
Prerequisites	**Department**	**Semester**
COMP-543DL	Computer Science	Fall
Type of Course	**Field**	**Language of Instruction**
Elective	Data Science	English
Level of Course	**Lecturer(s)**	**Year of Study**
2nd Cycle	Dr. D. Trihinas	1st
Mode of Delivery	**Work Placement**	**Corequisites**
Distance Learning	N/A	None

Course Objectives:

The main objectives of the course are to:

- Provide a comprehensive overview of the data evolution landscape and why traditional data solutions are inadequate for the specific requirements of modern scalable, reliable and fault-tolerant applications.
- Introduce the principles, concepts and modelling abstractions for Big Data and data-intensive computing at scale.
- Present the fundamental principles for popular Big Data programming models (e.g., MapReduce, Dataflow, etc.) and discuss how performance and robustness are significantly improved compared to traditional models.
- Provide a comprehensive overview on distributed storage models for big data and elaborate on the challenges that arise for both data availability and consistency.
- Introduce advanced programming design patterns used extensively for processing Big Data (e.g., counting, sorting, relational algebra, matrix multiplication, etc.).
- Provide a comprehensive overview of how to manage and process graph data that is too large to fit in a single host.
- Describe alternative programming models and algorithms for data that must be processed online and also features low-latency requirements (e.g., streaming data).
- Demonstrate basic techniques towards architecting big-data solutions.
- Introduce various popular and open-source tools for big-data storage, processing and analytic insight extraction.

Learning Outcomes:

After completion of the course students are expected to be able to:

- Describe the multiple dimensions and challenges involved in storing, processing and modelling Big Data.
- Comprehend the contexts in which Big Data principles models are applied, while also recognizing potential implications and trade-offs depending on the context.
- Conceptually understand the capabilities and pitfalls of Big Data storage models towards relational storage models when applied on structured and unstructured data.
- Evaluate data analysis problems to determine whether and how Big Data algorithms, programming models and techniques can be applied.
- Understand the underlying principles and concepts of key Big Data programming models (e.g., MapReduce, Dataflow, etc.) and present the ability to design applications adopting these principles.
- Acknowledge how to model, adapt and extend data analysis techniques to process streaming data with low-latency requirements.
- Realize how different tools fit in the frame of Big Data analytics stacks.
- Demonstrate the ability to use open-source technologies to design components of Big Data solutions for data storage, processing and analytics extraction.

Course Content:

1. Scalable Data-Intensive Computing
    a. The Big Data Dimensions (3V's, 5V's)
    b. Thinking at Scale
    c. Parallel and Distributed Computing for Data-Intensive Applications
2. Big Data Advanced Models I
    a. Data Replication
    b. Data Rebalancing
3. Big Data Advanced Models II
    a. Data Partitioning
    b. Schema Sharding
    c. Secondary Indexes
4. Distributed Databases
    a. Relational vs Non-Relational Data
    b. Query Routing – Data Lookups
    c. Distributed Indexing
    d. NoSQL Databases (e.g., column stores, document stores)
5. Failures and Errors in Big Data Systems
    a. Network Interruptions
    b. Transactional Consistency

c. Conflict Resolution
d. ACID properties and the CAP Theorem
e. NewSQL Models and Databases

6. The MapReduce Programming Model
   a. Data Parallel Problems
   b. The MapReduce Programming Model
   c. MapReduce Algorithmic Design Patterns
7. Batch Processing Systems
   a. Data in Batches
   b. Distributed Processing Engines
   c. MapReduce as an Execution Framework (e.g, Hadoop)
   d. Local Aggregation and Latency Improvements
8. Graph Management and Processing
   a. Graph and Network Datasets
   b. Graph Storage and Data Lookups
   c. Iterative MapReduce Model
   d. Graph Processing
9. Data Warehousing
   a. Data Warehousing and Big Data
   b. Structuring Unstructured Data
   c. ETL Operations
   d. Relational Queries over Distributed Processing Engines
10. The Dataflow Programming Model
   a. MapReduce Limitations
   b. The Dataflow Programming Model
   c. Dataflow Algorithmic Design Patterns
11. Data Streams
   a. The Data Stream Model
   b. Stream vs Batch Data
   c. Scalable Streaming Algorithms
12. Stream Processing Systems
   a. Resilient Distributed Datasets
   b. Idempotence and Microbatching
   c. Dataflow as an Execution Framework (e.g., Spark, Flink)

Learning Activities and Teaching Methods:

Lectures, Exercises, Software Tool Tutorials, Case-Study Presentations, Discussions.

Assessment Methods:

Exam, Homework, Lab Reports, Semester Project.

Required Textbooks / Readings:

Title	Author(s)	Publisher	Year	ISBN
Designing Data-Intensive Applications	Martin Kleppmann	O'Reilly	2017	978-1-449-37332-0
Data-Intensive Text Processing with MapReduce*	Jimmy Lin and Chris Dyer	Morgan and Claypool	2010	978-1-608-45342-9
Mining Massive Datasets (2nd edition) **	Jure Leskovec, Anand Rajaraman and Jeff Ullman	Cambridge University Press	2014	978-1-107-07723-2

* Made freely available online by the authors: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
** Made freely available online by the authors: http://www.mmds.org/#book

Recommended Textbooks / Readings:

Title	Author(s)	Publisher	Year	ISBN
Hadoop: The Definitive Guide (4th edition)	Tom White	O'Reilly	2015	978-1-491-90168-7
Big Data Fundamentals	Thomas Erl and Wajid Khattak and and Paul Buhler	Prentice Hall	2016	978-0-134-29107-9
Big Data	Nathan Marz and James Warren	Manning	2015	978-1-617-29034-3
Spark: The Definitive Guide	Matei Zaharia and Bill Chambers	O'Reilly	2018	978-1-49191-221-8

Big Data: Principles and Paradigms	Rajkumar Buyya and Rodrigo N. Calheiros and Amir Vahid Dastjerdi	Morgan Kaufmann	2016	978-0-128-05394-2