Course guide
270678 - BDM - Big Data Management

Last modified: 02/02/2024

Unit in charge: Barcelona School of Informatics
Teaching unit: 747 - ESSI - Department of Service and Information System Engineering.

Degree: MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Optional subject). ERASMUS MUNDUS MASTER'S DEGREE IN BIG DATA MANAGEMENT AND ANALYTICS (BDMA) (Syllabus 2021). (Compulsory subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer: ALBERTO ABELLO GAMAZO

Others: Segon quadrimestre:
ALBERTO ABELLO GAMAZO - 11, 12
BESIM BILALLI - 11, 12
SERGI NADAL FRANCESCH - 11, 12

PRIOR SKILLS

Being Big Data Management the evolution of Data Warehousing, such knowledge is assumed in this course. Thus, general knowledge is expected on: Relational database desing; Database management system architecture; ETL and OLAP

Specifically, knowledge is expected on:
- Multidimensional modeling (i.e, star schemas)
- Querying relational databases
- Physical design of relational tables (i.e., partitioning)
- Hash and B-tree indexing
- External sorting algorithms (i.e., merge-sort)
- ACID transactions

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.

Generical:
CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.

Transversal:
CTR3. TEAMWORK: Capacity of being able to work as a team member, either as a regular member or performing directive activities, in order to help the development of projects in a pragmatic manner and with sense of responsibility; capability to take into account the available resources.

Basic:
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

TEACHING METHODOLOGY

The course comprises theory, and lab sessions.

Theory: Inverted class techniques will be used, which require that the student works on the provided multimedia materials before the class. Then, theory lectures comprise the teacher's complementary explanations and problem solving.

Lab: The course contents are applied in a realistic problem in the
course project, done in teams, where students will put into practice the kinds of tools studied during the course. Since this course is part of the BDMA Erasmus Mundus
master syllabus, this project is conducted jointly with the Viability of
Business Projects (VBP), Semantic Data Management (SDM) and Debates on Ethics
of Big Data (DEBD) courses.

LEARNING OBJECTIVES OF THE SUBJECT

1.Understand the main advanced methods of data management and design and implement non-relational database managers, with special emphasis on distributed systems.
2.Understand, design, explain and carry out parallel information processing in massively distributed systems.
3.Manage and process a continuous flow of data.
4.Design, implement and maintain system architectures that manage the data life cycle in analytical environments.

STUDY LOAD

Type	Hours	Percentage
Hours large group	27,0	18.00
Hours small group	27,0	18.00
Self study	96,0	64.00

Total learning time: 150 h

CONTENTS

Introduction

Description:
Big Data, Cloud Computing, Scalability

Big Data Design

Description:
Polyglot systems; Schemaless databases; Key-value stores; Wide-column stores; Document-stores

Distributed Data Management

Description:
Transparency layers; Distributed file systems; File formats; Fragmentation; Replication and synchronization; Sharding; Distributed hash; LSM-Trees

In-memory Data Management

Description:
NUMA architectures; Columnar storage; Late reconstruction; Light-weight compression

Distributed Data Processing

Description:
Distributed Query Processing; Sequential access; Pipelining; Parallelism; Synchronization barriers; Multitenancy; MapReduce; Resilient Distributed Datasets; Spark

Stream management and processing

Description:
One-pass algorithms; Sliding window; Stream to relation operations; Micro-batching; Sampling; Filtering; Sketching

Big Data Architectures

Description:
Centralized and Distributed functional architectures of relational systems; Lambda architecture

ACTIVITIES

Theoretical lectures

Description:
In these activities, the lecturer will introduce the main theoretical concepts of the subject. Besides lecturing, cooperative learning techniques will be used. These demand the active participation of the students, and consequently will be evaluated.

Specific objectives:
1, 2, 3, 4

Related competencies :
CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR3. TEAMWORK: Capacity of being able to work as a team member, either as a regular member or performing directive activities, in order to help the development of projects in a pragmatic manner and with sense of responsibility; capability to take into account the available resources.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

Full-or-part-time: 50h
Theory classes: 25h
Self study: 25h

Exam

Description:
Written exam of the theoretico-practical concepts introduced along the course.

Specific objectives:
1, 2, 3, 4

Related competencies :
CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR3. TEAMWORK: Capacity of being able to work as a team member, either as a regular member or performing directive activities, in order to help the development of projects in a pragmatic manner and with sense of responsibility; capability to take into account the available resources.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

Full-or-part-time: 19h
Theory classes: 2h
Self study: 17h

Lab

Description:
Students will use different NOSQL tools in a sandbox environment.

Specific objectives:
1, 2, 3, 4

Related competencies :
CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR3. TEAMWORK: Capacity of being able to work as a team member, either as a regular member or performing directive activities, in order to help the development of projects in a pragmatic manner and with sense of responsibility; capability to take into account the available resources.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

Full-or-part-time: 81h
Laboratory classes: 27h
Self study: 54h

GRADING SYSTEM

Final Mark = min(10 ; 60%E + 40%L + 10%P)

L = Weighted average of the marks of the lab deliverables and tests
E = Final exam
P = Participation in the class

BIBLIOGRAPHY

Basic:
- Özsu, M.T.; Valduriez, P. Principles of distributed database systems. 4th ed. New York: Springer, 2020. ISBN 9783030262525.
- Liu, L.; Özsu, M.T. Encyclopedia of database systems. New York ; London: Springer, 2009. ISBN 9780387399409.
- Sadalage, P.J.; Fowler, M. NoSQL distilled: a brief guide to the emerging world of polygot persistence. Boston, Mass. ; London: Addison-Wesley, 2013. ISBN 9780321826626.
- Plattner, H.; Zeier, A. In-memory data management. 2nd ed. Berlin: Springer, 2012. ISBN 9783642295744.
- Zaharia, M. An architecture for fast and general data processing on large clusters. ACM Books, 2016. ISBN 9781970001563.
- Leskovec, J.; Rajaraman, A.; Ullman, J.D. Mining of massive datasets. 3rd ed. Cambridge: Cambridge University Press, 2020. ISBN 9781108476348.
- Aggarwal, C.C. (ed.). Data streams: models and algorithms. New York: Springer, 2007. ISBN 9780387287591.

Complementary:
- Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database system: the complete book. 2nd ed. Harlow, Essex: Pearson Education Limited, 2013. ISBN 9781292037301.
- Loshin, D. Master data management. Amsterdam ; Boston: Morgan Kaufmann/Elsevier, 2009. ISBN 9781282285507.

RESOURCES

Hyperlink:
- http://cs.ulb.ac.be/conferences/ebiss.html- https://deds.ulb.ac.be