Course guide
270952 - DW - Data Warehousing

Last modified: 12/07/2024

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

Degree: MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Compulsory subject).

Academic year: 2024 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer: PETAR JOVANOVIC

Others: Primer quadrimestre:
PETAR JOVANOVIC - 11, 12
XAVIER ORIOL HILARI - 11, 12

PRIOR SKILLS

Basic knowledge on relational databases and SQL.

Specifically, it will be assumed knowledge on:
- UML class diagrams
- Relational algebra
- SQL queries
- Relational views
- B-tree operations (i.e., insertion and splits)
- Basic concepts on physical query optimization

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE2. Apply the fundamentals of data management and processing to a data science problem
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact

Generical:
CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data

Transversal:
CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CT3. TEAMWORK: Ability to work as a member of an interdisciplinary team, as a normal member or performing direction tasks, in order to develop projects with pragmatism and sense of responsibility, making commitments taking into account the available resources.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

Basic:
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
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.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.

TEACHING METHODOLOGY

The course comprises theory, and lab sessions.

Lectures: The teacher presents the topic. Students follow the lesson, take notes, and prepare additional material outside of class. They may also be asked to carry out assessment activities within these sessions.

Laboratory: Some representative tools will be used for the application of theoretical concepts (e.g., PotgreSQL, Oracle, Talend, Tableau). The course includes continuous hands-on through a course project, divided into three logical blocks: data warehouse modelling, data integration and migration (ETL), and descriptive visualisation, in which the students will work in teams. There will be three project deliverables outside the class hours, while in the class the students will be as well individually assessed about the knowledge acquired during each project block.

LEARNING OBJECTIVES OF THE SUBJECT

1.Be able to model multidimensional data warehouses and visually analyze their data
2.Be able to apply specific physical design techniques for decisional systems
3.Be able to design and implement data migration processes (i.e., ETL)

STUDY LOAD

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

Total learning time: 150 h

CONTENTS

Introduction

Description:
Comparison of operational and decisional systems; Metadata

Data warehousing architectures

Description:
Corporate Information Factory; DW 2.0

Multidimensional modeling, OLAP tools

Description:
Structure; Integrity constraints; Operations; Advanced concepts

Database physical desing for analytical queries

Description:
Star-join and join indexes; Bitmaps; Materialized views; Spatio-temporal data

Extraction, Transformation and Load

Description:
Data quality; Schema and Data Integration; ETL management

Visualization and descriptive analytics

Description:
Key Performance Indicators; Dashboarding

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

Related competencies :
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
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.
CE2. Apply the fundamentals of data management and processing to a data science problem
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CT3. TEAMWORK: Ability to work as a member of an interdisciplinary team, as a normal member or performing direction tasks, in order to develop projects with pragmatism and sense of responsibility, making commitments taking into account the available resources.

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

Hands-on sessions

Description:
The student will be asked to practice the different concepts introduced in the theoretical lectures. This includes problem solving either on the computer or on paper.

Specific objectives:
1, 2, 3

Related competencies :
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
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.
CE2. Apply the fundamentals of data management and processing to a data science problem
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CT3. TEAMWORK: Ability to work as a member of an interdisciplinary team, as a normal member or performing direction tasks, in order to develop projects with pragmatism and sense of responsibility, making commitments taking into account the available resources.

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

Exam

Description:
Written exam of the theoretical concepts introduced along the course.

Specific objectives:
1, 2, 3

Related competencies :
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
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.
CE2. Apply the fundamentals of data management and processing to a data science problem
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CT3. TEAMWORK: Ability to work as a member of an interdisciplinary team, as a normal member or performing direction tasks, in order to develop projects with pragmatism and sense of responsibility, making commitments taking into account the available resources.

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

GRADING SYSTEM

Final grade = max(20%EP+40%EF ; 60% EF) + 40% P

EP = partial (mid term) exam mark
EF = final exam mark
P = Weighted average of the marks of the project deliverables

BIBLIOGRAPHY

Basic:
- Golfarelli, M.; Rizzi, S. Data warehouse design: modern principles and methodologies. New York [etc.]: McGraw Hill, 2009. ISBN 9780071610391.
- Jensen, C.S.; Pedersen, T.B.; Thomsen, C.W. Multidimensional databases and data warehousing. San Rafael, Calif: Morgan & Claypool, 2010. ISBN 9781608455379.
- Inmon, William H; Imhoff, Claudia; Sousa, Ryan. Corporate information factory. 2nd ed. New York: John Wiley, 2001. ISBN 0471399612.
- Kimball, R. [et al.]. The data warehouse lifecycle toolkit. 2nd ed. Indianapolis: Wiley publishing, 2008. ISBN 9780470149775.
- Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database systems: the complete book [on line]. Second edition, Pearson new international edition. Harlow, Essex: Pearson Education Limited, 2014 [Consultation: 14/03/2025]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5174436. ISBN 9781292024479.
- Vaisman, A.; Zimanyi, E. Data warehouse systems: design and implentation. Berlin: Springer, 2014. ISBN 9783642546549.
- Taniar, D.; Rahayu, W. Data warehousing and analytics: fueling the data engine. Cham: Springer, 2021. ISBN 9783030819781.

Complementary:
- Teorey, Toby ... [et al.]. Database modeling and design: logical design. 5th ed. Burlington, MA: Morgan Kaufmann Publishers/Elsevier, 2011. ISBN 9780123820204.
- Lightstone, S.; Teorey, T.J.; Nadeau, T. Physical database design: the database professional's guide to exploiting indexes, views, storage, and more. Amsterdam [etc.]: Morgan Kaufmann Publishers, 2007. ISBN 9780123693891.
- Ramakrishnan, R.; Gehrke, J. Database management systems. 3rd ed. Boston [etc.]: McGraw-Hill, 2003. ISBN 0071151109.
- Lewis, J. Cost-based oracle fundamentals. Berkeley, CA: Apress, 2006. ISBN 9781590596364.

RESOURCES

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