Course guide
270952 - DW - Data Warehousing

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: 2022 ECTS Credits: 6.0 Languages: English

LECTURER
Coordinating lecturer: ALBERTO ABELLO GAMAZO
Others: Primer quadrimestre:
ALBERTO ABELLO GAMAZO - 11, 12
PETAR JOVANOVIC - 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.

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

Laboratory: Some representative tools will be used for the application of theoretical concepts (e.g., Indyco Builder, Postgresql, Oracle, Pentaho Data Integration, 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

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guided activities</td>
<td>6,0</td>
<td>4.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>24,0</td>
<td>16.00</td>
</tr>
<tr>
<td>Self study</td>
<td>96,0</td>
<td>64.00</td>
</tr>
<tr>
<td>Hours small group</td>
<td>24,0</td>
<td>16.00</td>
</tr>
</tbody>
</table>

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
<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multidimensional modeling, OLAP tools</td>
<td>Structure; Integrity constraints; Operations; Advanced concepts</td>
</tr>
<tr>
<td>Database physical design for analytical queries</td>
<td>Star-join and join indexes; Bitmaps; Materialized views; Spatio-temporal data</td>
</tr>
<tr>
<td>Extraction, Transformation and Load</td>
<td>Data quality; Schema and Data Integration; ETL management</td>
</tr>
<tr>
<td>Visualization and descriptive analytics</td>
<td>Key Performance Indicators; Dashboarding</td>
</tr>
</tbody>
</table>
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:**
- CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data
- 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
- 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.

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.

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.

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.

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.

**Full-or-part-time: 50h**
Theory classes: 25h
Self study: 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:
CG1. Identify and apply the most appropriate data management methods and processes to manage the data life cycle, considering both structured and unstructured data
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

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.

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.

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.

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.

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.

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

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

**Specific objectives:**
1, 2, 3

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

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.
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.
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.
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.

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

**GRADING SYSTEM**
Final grade = min(10 ; max(20%EP+40%EF ; 60% EF) + 40% P + 10% C)

EP = partial (mid term) exam mark
EF = final exam mark
P = Weighted average of the marks of the project deliverables
C = participation in the class
BIBLIOGRAPHY

Basic:

Complementary:

RESOURCES

Hyperlink: