Course guides
270956 - SDM - Semantic Data Management

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: 2021 ECTS Credits: 6.0 Languages: English

LECTURER
Coordinating lecturer: OSCAR ROMERO MORAL
Others: Segon quadrimestre:
BESIM BILALLI - 11
ANNA QUERALT CALAFAT - 11
MD ATAUR RAHMAN - 11
OSCAR ROMERO MORAL - 11

PRIOR SKILLS
The student must be familiar with basics on databases and data modeling. Advanced programming skills are mandatory.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
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
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science

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
CG3. Define, design and implement complex systems that cover all phases in data science projects

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 spans theory and lab sessions.

Theory: These lectures are based on teacher’s explanations and constitute the main part of the course. The students will also have some contents to read and prepare outside the classroom and will be asked to participate in cooperative learning activities during the lectures.

Laboratory: Mainly, the lab sessions will be dedicated to hands-on of the concepts introduced in the theory lectures. Specific and relevant tools will be introduced in these sessions. Small-sized projects will be conducted using these tools.

Project: The course contents are applied in a realistic problem in the course project.

LEARNING OBJECTIVES OF THE SUBJECT

1. Learn, understand and apply the fundamentals of property graphs
2. Learn, understand and apply the fundamentals of knowledge graphs
3. Perform graph data processing both in centralized and distributed environments
4. Integrate, combine and refine semi-structured or non-structured data using graph formalisms
5. Determine how to apply graph formalisms to solve the Variety challenge (data integration)
6. Apply property or knowledge graphs to solve realistic problems such as data integration, graph-based data analysis, etc.

STUDY LOAD

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

Total learning time: 150 h
**CONTENTS**

<table>
<thead>
<tr>
<th>Introduction and formalisation of Variety in Big Data and its management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Definition of data management tasks: from a database perspective and knowledge representation.</td>
</tr>
<tr>
<td>Definition of Variety and Big Data. Syntactic and Semantic heterogeneities. Impact of heterogeneities in the identified data management tasks.</td>
</tr>
<tr>
<td>Data integration. Theoretical framework for the management and integration of heterogeneous data sources.</td>
</tr>
<tr>
<td>Main components of an integration system: data sources, global schema and mappings.</td>
</tr>
<tr>
<td>The concept of canonical model for data integration. Definition of data model. Main characteristics of a canonical data model.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property graphs management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Data structures. Integrity constraints.</td>
</tr>
<tr>
<td>Basic operations. Based on topology, content and hybrid.</td>
</tr>
<tr>
<td>Graph query languages: GraphQL.</td>
</tr>
<tr>
<td>Graph database concept: tool heterogeneity when implementing the graph structures. Impact of such decisions in the main operations.</td>
</tr>
<tr>
<td>Distributed graph databases. Need and difficulties. Thinking like a vertex paradigm as standard de facto in distributed graph processing.</td>
</tr>
<tr>
<td>Main distributed graph algorithms.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge graph management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Data structure. RDF. Origin and relationship with Linked Open Data. Integrity constraints.</td>
</tr>
<tr>
<td>Data structure: RDFS and OWL. Relationship with first order logic. Foundations in Description Logics. Integrity constraints. Reasoning.</td>
</tr>
<tr>
<td>Basic operations and query language. SPARQL and underlying algebra. Entailment regimes (reasoning).</td>
</tr>
<tr>
<td>Triplestores. Differences with graph databases. Native implementations. Implementations based on the relational data model. Impact of such decisions on the basic operations.</td>
</tr>
<tr>
<td>Distributed triplestore. Needs and difficulties. Graph Engine 1.0 as paradigm of distributed triplestore.</td>
</tr>
<tr>
<td>Main distributed algorithms.</td>
</tr>
</tbody>
</table>
Graphs as solution to the Variety challenge

Description:
Graphs as the best canonical model for data integration.
Graph data models main features. Differences with other data models (specially the relational data model).
Data and metadata concepts and their formalization in graph models.
Use cases (highlighting topological benefits): fraud detection, bioinformatics, traffic and logistics, social networks, etc.
Introduction to the main graph models: property graph and knowledge graphs.

Property and knowledge graphs comparison. Use cases

Description:
Recap about both models. Commonalities and differences. Concepts to borrow between both paradigms.
Main use cases. Metadata management: Data Lake semantification and data governance.
Main use cases. Exploitation of their topological features: recommenders on graphs and data mining.
Visualization: by means of a GUI (Gephi) or programmatically (D3.js or GraphLab).
ACTIVITIES

Lectures

Description:
During lectures the main concepts will be discussed. Lectures will combine master lectures and active / cooperative learning activities. The student is meant to have a pro-active attitude during active / cooperative learning activities. During master lectures, the student is meant to listen, take notes and ask questions.

Specific objectives:
1, 2, 3, 5

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
CG3. Define, design and implement complex systems that cover all phases in data science projects
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE5. Model, design, and implement complex data systems, including data visualization
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
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.

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.
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.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
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.
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.

Full-or-part-time: 53h
Theory classes: 25h
Self study: 28h
**Hands-on Session**

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

**Specific objectives:**
4, 5, 6

**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
- CG3. Define, design and implement complex systems that cover all phases in data science projects
- CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
- CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
- CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
- CE5. Model, design, and implement complex data systems, including data visualization
- CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
- 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.

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.

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.

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.

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.

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.

**Full-or-part-time:** 87h
Laboratory classes: 27h
Self study: 60h
Final Exam

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

Specific objectives:
1, 2, 3, 4, 5

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
CG3. Define, design and implement complex systems that cover all phases in data science projects
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE5. Model, design, and implement complex data systems, including data visualization
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
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.

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.
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.
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.
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.
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.

Full-or-part-time: 10h
Theory classes: 2h
Self study: 8h

GRADING SYSTEM

Final mark = 40% EX + 50% LAB + 10% P

EX = Final exam mark
LAB = Weighted mark of the labs
P = Project
BIBLIOGRAPHY

Basic: