270653 - OD - Open Data

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 747 - ESSI - Department of Service and Information System Engineering
Academic year: 2018
Degree: MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Teaching unit Optional)
ECTS credits: 6 Teaching languages: Catalan

Prior skills

The student must be familiar with basics on databases and data modeling. Programming skills are also mandatory.

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.
- CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.

Generical:
- CG4. Capacity for general and technical management of research, development and innovation projects, in companies and technology centers in the field of Informatics Engineering.
- CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.

Transversal:
- CTR1. ENTREPRENEURSHIP AND INNOVATION: Capacity for knowing and understanding a business organization and the science that rules its activity, capability to understand the labour rules and the relationships between planning, industrial and commercial strategies, quality and profit. Capacity for developing creativity, entrepreneurship and innovation trend.
- 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.

Teaching methodology

The course comprises theory and lab sessions.

Theory: These lectures comprise the 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.

Laboratory: Mainly, the lab sessions will be dedicated to the practice (with and without computer) 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.

Learning objectives of the subject

1. Determine how to apply graph formalisms to solve the Variety challenge (data integration)
2. Master the main semantic-aware formalisms to enable semantic modeling
3. Determine how to apply graph formalisms to solve the Variety challenge (data integration)
4. Reinforce teamwork capabilities in order to develop innovative solutions by means of complementing the organization data with external data.

5. Perform graph data processing both in centralized and distributed environments.

### Study load

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total learning time:</strong></td>
<td>150h</td>
<td>36h</td>
<td>24.00%</td>
</tr>
<tr>
<td>Theory classes:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practical classes:</td>
<td>0h</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>Laboratory classes:</td>
<td>18h</td>
<td>12.00%</td>
<td></td>
</tr>
<tr>
<td>Guided activities:</td>
<td>0h</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>Self study:</td>
<td>96h</td>
<td>64.00%</td>
<td></td>
</tr>
</tbody>
</table>
Content

**Introduction and formalisation of Variety in Big Data and its management**

**Degree competences to which the content contributes:**

**Description:**
Definition of data management tasks: from a database perspective and knowledge representation.

Definition of Variety and Big Data. Syntactic and Semantic heterogeneities. Impact of heterogeneities in the identified data management tasks.

Data integration. Theoretical framework for the management and integration of heterogeneous data sources.

Main components of an integration system: data sources, global schema and mappings.

The concept of canonical model for data integration. Definition of data model. Main characteristics of a canonical data model.

**Graphs as solution to the Variety challenge**

**Degree competences to which the content contributes:**

**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 graphs management**

**Degree competences to which the content contributes:**
Description:
Data structures. Integrity constraints.

Basic operations. Based on topology, content and hybrid.

Graph query languages: GraphQL.

Graph database concept: tool heterogeneity when implementing the graph structures. Impact of such decisions in the main operations.

Distributed graph databases. Need and difficulties. Thinking like a vertex paradigm as standard de facto in distributed graph processing.

Main distributed graph algorithms.

Knowledge graph management

Degree competences to which the content contributes:

Description:
Data structure: RDF. Origin and relationship with Linked Open Data. Integrity constraints.

Data structure: RDFS and OWL. Relationship with first order logic. Foundations in Description Logics. Integrity constraints. Reasoning.

Basic operations and query language. SPARQL and underlying algebra. Entailment regimes (reasoning).

Triplestores. Differences with graph databases. Native implementations. Implementations based on the relational data model. Impact of such decisions on the basic operations.

Distributed triplestore. Needs and difficulties. Graph Engine 1.0 as paradigm of distributed triplestore.

Main distributed algorithms.

Property and knowledge graphs comparison. Use cases

Degree competences to which the content contributes:

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).
270653 - OD - Open Data

**Qualification system**

Final mark = 10% EC + 40% EX + 40% LAB + 10% C

- **EX** = Final exam mark
- **LAB** = Weighted mark of the labs
- **EC** = Mark from the activities in the theoretical sessions
- **C** = Peer evaluation

EC = In some theory sessions some activities will be proposed. The students need to solve it and hand it out to the lecturer before the end of the session.

LAB: There are three lab sessions, each one with a potential different weight. LABs will be performed in groups assigned by the lecturer.

C: During the course practice each student will interact with other students. Since the practice is meant to entail several working hours each student will peer-mark his / her teammates. The lecturer will assign a mark to each student according to the peer-marking received by his / her teammates.

**Bibliography**

**Basic:**


Sahu, Siddhartha; Mhedhbi, Amine; Salihoglu, Semih; Lin, Jimmy; Özsu, M. Tamer. The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing. Cornell University Library, 2017.