Course guides
270955 - BDM - Big 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: ALBERTO ABELLO GAMAZO
Others: Segon quadrimestre:
ALBERTO ABELLO GAMAZO - 11
SERGI NADAL FRANCESCH - 11

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:
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
CE2. Apply the fundamentals of data management and processing to a data science problem
CE4. Apply scalable storage and parallel data processing methods, including data streams, once the most appropriate methods for a data science problem have been identified
CE5. Model, design, and implement complex data systems, including data visualization

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 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: There will be a project done in teams where students will put into practice the kinds of tools studied during the course. This will be evaluated in two deliverables and individual tests.

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

<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>Self study</td>
<td>96,0</td>
<td>64.00</td>
</tr>
<tr>
<td>Guided activities</td>
<td>6,0</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Total learning time: 150 h
## CONTENTS

<table>
<thead>
<tr>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Big Data, Cloud Computing, Scalability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data Design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Polyglot systems; Schemaless databases; Key-value stores; Wide-column stores; Document-stores</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distributed Data Management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Transparency layers; Distributed file systems; File formats; Fragmentation; Replication and synchronization; Sharding; Distributed hash; LSM-Trees</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In-memory Data Management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>NUMA architectures; Columnar storage; Late reconstruction; Light-weight compression</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distributed Data Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Distributed Query Processing; Sequential access; Pipelining; Parallelism; Synchronization barriers; Multitenancy; MapReduce; Resilient Distributed Datasets; Spark</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stream management and processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>One-pass algorithms; Sliding window; Stream to relation operations; Micro-batching; Sampling; Filtering; Sketching</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Centralized and Distributed functional architectures of relational systems; Lambda architecture</td>
</tr>
</tbody>
</table>
### Theoretical lectures

**Description:**
In these activities, the lecturer will introduce the main theoretical concepts of the subject. The active participation of the students will be required.

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

**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
- CE2. Apply the fundamentals of data management and processing to a data science problem
- CE13. Define, design and implement complex systems that cover all phases in data science projects (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
- 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
- CE5. Model, design, and implement complex data systems, including data visualization
- CE4. Apply scalable storage and parallel data processing methods, including data streams, once the most appropriate methods for a data science problem have been identified
- 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.

**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.

**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.

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

**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:**
- 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
- CE2. Apply the fundamentals of data management and processing to a data science problem
- CE3. Define, design and implement complex systems that cover all phases in data science projects
- 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
- 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
- CE5. Model, design, and implement complex data systems, including data visualization
- CE4. Apply scalable storage and parallel data processing methods, including data streams, once the most appropriate methods for a data science problem have been identified
- 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.

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.

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.

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

**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:
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
CE2. Apply the fundamentals of data management and processing to a data science problem
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
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
CE5. Model, design, and implement complex data systems, including data visualization
CE4. Apply scalable storage and parallel data processing methods, including data streams, once the most appropriate methods for a data science problem have been identified
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.
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.
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.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.

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:

Complementary:

RESOURCES

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