# 270678 - BDM - Big Data Management

<table>
<thead>
<tr>
<th>Coordinating unit:</th>
<th>270 - FIB - Barcelona School of Informatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching unit:</td>
<td>747 - ESSI - Department of Service and Information System Engineering</td>
</tr>
<tr>
<td>Academic year:</td>
<td>2019</td>
</tr>
<tr>
<td>Degree:</td>
<td>ERASMUS MUNDUS EN BIG DATA MANAGEMENT AND ANALYTICS-BDMA (Syllabus 2017). (Teaching unit Compulsory)</td>
</tr>
<tr>
<td></td>
<td>MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Teaching unit Optional)</td>
</tr>
<tr>
<td>ECTS credits:</td>
<td>6</td>
</tr>
<tr>
<td>Teaching languages:</td>
<td>English</td>
</tr>
</tbody>
</table>

## 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 design; 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

### Basic:

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.

### 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.

CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.

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

### Generical:

CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.

### Transversal:

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.
270678 - BDM - Big Data Management

**Teaching methodology**

The course comprises theory, problems, and lab sessions.

Theory: The theory classes comprise the teacher's explanations and constitute the main part of the course. The students will also have some contents to be read and prepared outside the classroom, and will be asked to participate in cooperative learning activities to solve some problems.

Lab: There will be some lab sessions to introduce some of the technologies in a given execution environment.

Project: At the end of the course, there will be a small project where the student will show all the knowledge acquired during the semester in a proof of concept.

**Learning objectives of the subject**

1. Understand the differences and benefits of in-memory data management.
2. Understand the execution flow of a distributed query.
3. Identify the difficulties of scalability and parallelization.
4. Design a distributed database using NoSQL tools.
5. Produce a functional program to process Big Data in a Cloud environment.
6. Manage and process a Data Stream.
7. Design the architecture of a Big Data management system.

**Study load**

<table>
<thead>
<tr>
<th>Total learning time: 150h</th>
<th>Hours large group: 25h 30m</th>
<th>17.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours medium group: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Hours small group: 25h 30m</td>
<td>17.00%</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 3h</td>
<td>2.00%</td>
</tr>
<tr>
<td></td>
<td>Self study: 96h</td>
<td>64.00%</td>
</tr>
</tbody>
</table>
## Content

### Introduction

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

**Description:**
- Big Data, Cloud Computing, Scalability

### Big Data Design

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

**Description:**
- Polyglot systems; Schemaless databases; Key-value stores; Wide-column stores; Document-stores

### Distributed Data Management

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

**Description:**
- Transparency layers; Distributed file systems; File formats; Fragmentation; Replication and synchronization; Sharding; Consistent hash; LSM-Trees

### In-memory Data Management

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

**Description:**
- NUMA architectures; Columnar storage; Late reconstruction; Light-weight compression

### Distributed Data Processing

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

**Description:**
- Distributed Query Processing; Sequential access; Pipelining; Parallelism; Synchronization barriers; Multitenancy; MapReduce; Resilient Distributed Datasets; Spark

### Stream management and processing

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

**Description:**
- One-pass algorithms; Sliding window; Stream to relation operations; Micro-batching; Sampling; Filtering; Sketching
### Big Data Architectures

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

**Description:**
- Centralized and Distributed functional architectures of relational systems;
- Data Warehousing architectures;
- Service Oriented Architecture;
- Lambda architecture
### Planning of activities

| **Theoretical lectures** | **Hours:** 50h  
Theory classes: 25h  
Practical classes: 0h  
Laboratory classes: 0h  
Guided activities: 0h  
Self study: 25h |
|--------------------------|-------------------|
| **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, 5, 6, 7  |

| **Exam** | **Hours:** 19h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 0h  
Guided activities: 0h  
Self study: 17h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
<td>Written exam of the theoretico-practical concepts introduced along the course.</td>
</tr>
<tr>
<td><strong>Specific objectives:</strong></td>
<td>1, 2, 3, 4, 6, 7</td>
</tr>
</tbody>
</table>

| **Lab** | **Hours:** 81h  
Theory classes: 0h  
Practical classes: 0h  
Laboratory classes: 27h  
Guided activities: 0h  
Self study: 54h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
<td>Students will use different NOSQL tools in a sandbox environment.</td>
</tr>
<tr>
<td><strong>Specific objectives:</strong></td>
<td>2, 3, 4, 5, 6, 7</td>
</tr>
</tbody>
</table>
Final Mark = 50% min (10,1.1 * sum (Ci * Wi) / sum (Wi)) + 10% Pe + 30% E + 10% Pr

Ci = Marks on collaborative activities and laboratories
Wi = Weight equal to 1, 2, 4, 6 or 8 (depending on the relevance and difficulty of the corresponding activity / laboratory)
Pe = Peer evaluation
E = Exam
Pr = Project

Calculation of Pe: Students will have multiple mates in the activities carried out during the semester and those will report on them. Based on these reports, the teacher will assign a mark to each student.

Bibliography

Basic:


Complementary:


Others resources:
Hyperlink

http://it4bi-dc.ulb.ac.be

http://cs.ulb.ac.be/conferences/ebiss.html