Course guide
270221 - BDA - Advanced Databases

Last modified: 11/07/2024

Unit in charge: Barcelona School of Informatics
Teaching unit: 747 - ESSI - Department of Service and Information System Engineering.

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Compulsory subject).

Academic year: 2024 **ECTS Credits:** 6.0 **Languages:** Catalan, English

LECTURER

Coordinating lecturer: ALBERTO ABELLO GAMAZO

Others: Primer quadrimestre:
ALBERTO ABELLO GAMAZO - 11, 12, 13
BESIM BILALLI - 11, 12, 13

PRIOR SKILLS

Be able to read and understand materials in English.
Be able to list the stages that make up the software engineering process.
Be able to understand conceptual schemas in UML.
Be able to create, query and manipulate databases with SQL.
Be able to program using functional programming like Spark.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

Generical:
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

Transversal:
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Basic:
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

TEACHING METHODOLOGY

The course consists of theory and laboratory sessions.

Theory: Reverse class techniques will be used that require the student to work on multimedia materials before class. Theory classes consist of complementary teacher explanations and problem solving.

Laboratory: Representative tools will be used for the application of theoretical concepts (for example, PotgreSQL, Talend, HDFS, MongoDB). There will also be two projects, in which students will work in teams: one on descriptive data analysis in a data warehouse and the other on predictive analysis in a Big Data environment. Consequently, there will be two deliverables outside of class hours, but students will also be assessed individually in the classroom on the knowledge gained during each of the projects.

The course has an autonomous learning component, as the students will have to work with different data management and processing tools. Apart from the support material, students should be able to resolve doubts or problems using these tools.

LEARNING OBJECTIVES OF THE SUBJECT

1.Be able to discuss and justify in detail architectural principles and the bottlenecks of the relational managers in front of alternative storage and processing systems.
2.Be able to obtain the logical scheme of a data warehouse from a conceptual schema expressed in UML, detect and correct defects in it.
3.Be able to explain and use the main mechanisms of parallel processing of queries in distributed environments, and detect bottlenecks.
4.Be able to justify and use NOSQL storage systems.

STUDY LOAD

Type	Hours	Percentage
Hours small group	30,0	20.00
Self study	90,0	60.00
Hours large group	30,0	20.00

Total learning time: 150 h

CONTENTS

Introduction

Description:
Data warehousing and Big Data

Data Warehousing

Description:
Data warehousing. ETL data flows. Data integration. OLAP tools.

Distributed databases

Description:
Taxonomy of distributed databases. Architectures. Distributed database design (fragmentation and replication). Parallelism. Measures of scalability. Distirbuted file systems.

Distributed data processing

Description:
Importance of parallel sequential access. Synchronization barriers (Bulk Synchronous Parallel model). Big Data architectures and NOSQL systems.

ACTIVITIES

Introduction

Description:
Introduction of the subject, motivation and overview of existing data management tools, their advantages and disadvantages

Specific objectives:
1

Related competencies :
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 2h
Theory classes: 2h

Study of data warehouses

Specific objectives:
2

Related competencies :
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 62h
Self study: 38h
Theory classes: 10h
Laboratory classes: 14h

Study of distributed databases

Description:
Learning the principles of distributed databases and their application in NOSQL systems

Specific objectives:
1, 3

Related competencies :
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 14h
Self study: 4h
Theory classes: 6h
Laboratory classes: 4h

Study of the distributed processing of data and NOSQL systems

Description:
Learning of distributed data processing techniques and NOSQL systems

Specific objectives:
1, 3, 4

Related competencies :
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 60h
Self study: 38h
Theory classes: 10h
Laboratory classes: 12h

Final exam

Description:
Global examination of the subject

Specific objectives:
1, 2, 3, 4

Related competencies :
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 12h
Self study: 10h
Guided activities: 2h

GRADING SYSTEM

Final grade = max(20%EP+40%EF ; 60% EF) + 40% P

EP = partial (mid term) exam mark
EF = final exam mark
P = project mark, as a weighted average of the course projects

For students who may take the resit session, the reassessment examination mark will replace EF.

BIBLIOGRAPHY

Basic:
- Garcia-Molina, Hector; Ullman, Jeffrey D; Widom, Jennifer. Database systems : the complete book [on line]. Second edition, Pearson new international edition. Essex: Pearson Education, 2013 [Consultation: 14/03/2025]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5174436. ISBN 9781292024479.
- Database Technologies and Information Management. Slides on Advanced Databases course.
- Golfarelli, M.; Rizzi, S. Data warehouse design: modern principles and methodologies. New York [etc.]: McGraw Hill, 2009. ISBN 9780071610391.
- Vaisman, A.; Zimányi, E. Data warehouse systems: design and implentation. Second edition. Berlin: Springer, 2022. ISBN 9783662651667.
- Özsu, M.T.; Valduriez, P. Principles of distributed database systems. 4th ed. New York: Springer, 2020. ISBN 9783030262525.
- Sadalage, P.J.; Fowler, M. NoSQL distilled: a brief guide to the emerging world of polygot persistence. Addison-Wesley, 2013. ISBN 9780321826626.
- Badia, Antonio. SQL for data science : data cleaning, wrangling and analytics with relational databases. Springer, 2020. ISBN 9783030575915.
- Abelló, Alberto; Jovanovic, Petar. Data Warehousing and OLAP.
- Abelló, Albero; Nadal, Sergi. Big Data Management.

Complementary:
- Exercises Big Data Management.
- Exercises Data Warehousing.

RESOURCES

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