Last update: 19-07-2019

270221 - BDA - Advanced Databases

Coordinating unit:	270 - FIB - Barcelona School of Informatics
Teaching unit:	747 - ESSI - Department of Service and Information System Engineering
Academic year:	2019
Degree:	BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Teaching unit Compulsory)
ECTS credits:	6
Teaching languages:	Catalan

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.
Ser capable of creating, consulting and manipulating databases with SQL.

Degree competences to which the subject contributes

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.

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.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

Teaching methodology

At the theory hours, the teacher exposes the concepts corresponding to one of the contents. Some of the concepts are not discussed by the teacher, but students must work materials published on the virtual campus. The doubts that may arise when reading these materials are resolved by the teacher or their peers, through cooperative learning activities.

At the laboratory hours, the teacher presents different exercises, which students must solve in pairs and will be solved in class. Apart from that, two projects will also be carried out: one descriptive analysis of data in a data warehouse and the other one for predictive analysis in a Big Data environment.

The course also has an autonomous learning component, given that They will have to work with different data management tools (relational and non-relational). Apart from the support material, they must be able to resolve doubts or use problems of these managers.

Learning objectives of the subject

1.Be able to discuss and justify in detail the bottlenecks of the relational managers in front of alternative storage and processing systems.
2.Be able to analyze the pros and cons of having (or not) a unique model of reference that adapts to all possible storage scenarios.
3.Be able to detect and correct defects in a logical design.
4.Be able to obtain the logical scheme from a conceptual schema expressed in UML taking into account the consequences of the variety and variability of the data.
5.Being able, given certain characteristics of the data (volume, schema / data variability, expected workload), choose the data model and the appropriate physical structures to guarantee a correct balance between maintenance of the database and performance .
6.Be able to explain the operation and calculate the cost of access of the main data structures used by the managers.
7.Be able to obtain the access plan for a consultation based on optimization criteria.
8.Be able to reproduce the execution of algorithms that intervene in a process tree and estimate its cost.
9.Be able to decide the indexes (primary and secondary) that must be defined based on the expected operations.
10.Be able to choose and justify the use of storage based on rows or columns.
11.Be able to explain and use the main mechanisms of parallel processing of queries in distributed environments, and detect bottlenecks.
12.Be able to discuss and justify in detail the architectural principles that share the new non-relational storage systems.
13.Being able, given a specific scenario with user requirements (partial or total), identify what characteristics of relational managers would potentially act as bottleneck and talk about what types of storage managers would be most appropriate.
14.Be able to justify and use distributed functional data processing environments, like MapReduce/Spark.

Study load

Total learning time: 150h	Theory classes:	30h	20.00%
	Laboratory classes:	30h	20.00%
	Guided activities:	0h	0.00%
	Self study:	90h	60.00%

Content

Introduction

Degree competences to which the content contributes:

Description:
Data warehousing and Big Data

Data Warehousing

Degree competences to which the content contributes:

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

Techniques for the improvement of performance of database systems

Degree competences to which the content contributes:

Description:
Materialized views Data structures and indexing techniques (hash, trees and bitmaps). Techniques of compression and columnar storage. Parallelism

Distributed databases

Degree competences to which the content contributes:

Description:
Taxonomy of distributed databases. Architectures. Distributed database design (fragmentation and replication). Measures of scalability. Non-relational Key-Value systems.

Distributed data processing

Degree competences to which the content contributes:

Description:
Importance of parallel sequential access. Synchronization barriers (Bulk Synchronous Parallel model). Distributed processing environments of functional data (MapReduce and Spark). Abstraction of distributed datasets (Resilient Distributed Datasets).

Planning of activities

Introduction

Hours: 2h
Theory classes: 2h
Practical classes: 0h
Laboratory classes: 0h
Guided activities: 0h
Self study: 0h

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

Specific objectives:
1

Study of data warehouses

Hours: 55h
Theory classes: 10h
Practical classes: 0h
Laboratory classes: 12h
Guided activities: 1h
Self study: 32h

Study of techniques for improving the performance of database systems

Hours: 21h
Theory classes: 4h
Practical classes: 0h
Laboratory classes: 4h
Guided activities: 1h
Self study: 12h

Description:
Learning the types of management problems of materialized views and indexing structures, as well as the main associated costs that they have in each case

Specific objectives:
6, 9, 10

Study of distributed databases

Hours: 21h
Theory classes: 4h
Practical classes: 0h
Laboratory classes: 4h
Guided activities: 1h
Self study: 12h

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

Specific objectives:
1, 11, 12, 13

Study of the distributed processing of data	Hours: 45h Theory classes: 10h Practical classes: 0h Laboratory classes: 10h Guided activities: 1h Self study: 24h

Description:
Learning of distributed data processing techniques in functional style environments

Specific objectives:
14

Final exam	Hours: 12h Guided activities: 2h Self study: 10h

Description:
Global examination of the subject

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Qualification 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 E.

Bibliography

Basic:

Garcia-Molina, H.; Ullman, J.D.; Widom, J. Database systems: the complete book [on line]. 2nd ed. Upper Saddle River, NJ: Pearson Education, 2009 [Consultation: 22/07/2019]. Available on: <https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?docID=5174436>. ISBN 9780131873254.

Teorey, T.J. Database modeling and design: logical design. 5th ed. Burlington, MA: Morgan Kaufmann Publishers/Elsevier, 2011. ISBN 9780123820204.

Lightstone, S.; Teorey, T.J.; Nadeau, T. Physical database design: the database professional's guide to exploiting indexes, views, storage, and more [on line]. Amsterdam: Morgan Kaufmann Publishers, 2007 [Consultation: 22/07/2019]. Available on: <https://www.sciencedirect.com/science/book/9780123693891>. ISBN 9780123693891.

Golfarelli, M.; Rizzi, S. Data warehouse design: modern principles and methodologies. New York: McGraw Hill, 2009. ISBN 9780071610391.

Vaisman, A.; Zimányi, E. Data warehouse systems: design and implentation. Berlin: Springer, 2014. ISBN 9783642546549.

Özsu, M. T.; Valduriez, P. Principles of distributed database systems. 3rd ed. New York: Springer, 2011. ISBN 9781441988331.

Sadalage, P.J.; Fowler, M. NoSQL distilled: a brief guide to the emerging world of polygot persistence. Boston, Mass.: Addison-Wesley, 2013. ISBN 9780321826626.

Complementary:

Lewis, J. Cost-based oracle fundamentals. Berkeley, CA: Apress, 2006. ISBN 9781590596364.