2EL1560 – Models and Systems for Big Data management

Instructors: Nacera Seghouani
Department: DÉPARTEMENT INFORMATIQUE
Language of instruction: FRANCAIS
Campus: CAMPUS DE PARIS - SACLAY
Workload (HEE): 60
On-site hours (HPE): 35,00
Elective Category : Fundamental Sciences
Advanced level : Yes

Description

Existing data management technologies continue to evolve and adapt to challenges related to the heterogeneity, the volume and the velocity of data. NoSQL (Not only SQL) databases are a family of DataBase Management Systems (DBMS) which differs from the traditional relational SGBD paradigm. The purpose of such DBMSs is the simplicity of design, the horizontal scaling, the real-time access, the availability and performance in a distributed infrastructure.

The objective of *Models and Systems for Big Data Management* course is to study the theoretical foundations, the conceptual models and technologies for storing, monitoring and querying big data. From SQL to NoSQL, NoSQL the data representation models, the data querying and analytics, the performance measures in a distributed environnement are the main aspects addressed in this course.

Quarter number

SG6

Prerequisites (in terms of CS courses)

Basic knowledge of relational databases and SQL query language.
Desired basic knowledge of client/server architecture.

Syllabus

1. Conceptual models for data representation: SQL (relational) and NoSQL (document, key-value, column, graph).
2. Data querying and analytics languages.
3. Transaction concepts, ACID properties, CAP theorem.
4. Main concepts related to data distribution in a cluster.

5. Practical works on data modeling, querying real data (social network data, Wikipedia, ...).
6. Softwares : Oracle/postgres, Cassandra, Neo4J, Giraph, ElasticSearch.

Class components (lecture, labs, etc.)
The course is organized as follows:

- 12 slots of 1h30 of lecturer,
- 5 slots of 3h practical/lab works, groups of 25 students (maximum).

Grading
Continuous assessment during practical/lab work classes, project, written exam (2h):
50% written exam,
30% project (last lab),
and 20% continuous assessment (quizzes related to labs, 2 best marks).

Resources
Slides, practical exercices/works and solutions, QCMs, bibliography references.
Use of different data management system softwares available on MyDocker.

Learning outcomes covered on the course
At the end of this course, students will be able to:

- choose a data management model which fits with the application and the data nature.
- define, deploy, manipulate a SQL and NoSQL database.
- acquire the fundamental concepts underlying distributed data.

Description of the skills acquired at the end of the course
C1.2 Develop and use appropriate models, choosing the correct modelling scale and simplifying assumptions when addressing a problem
C6.4 Solve problems through mastery of computational thinking skills.
C6.5 Operate all types of data, structured or unstructured, including big data.