Large Scale Data Management

Instructor: Dr. Lefteris Sidirourgos, lsidir@aueb.gr

Overview

The goal of Large Scale Data Management (LSDM) is to enable modern businesses and sciences to exploit as many machines as possible, to consume as much information as possible, and to process data as fast as possible. In modern LSDM system design, the challenge is how to enable new technologies that transform raw data into useful knowledge. This is a moving target as both the underlying hardware and our methods to extract knowledge and information evolve. In this course, we will examine the building blocks of a database management system (DBMS), the algorithms and the technologies that allow DBMSs to *scale-up* or *-out*, and also we will study specific technologies such as distributed filesystems, column-stores, map/reduce and stream engines, graph systems, noSQL, and solutions for performing machine learning tasks close to where the data naturally resides, the database. We will see how they all rely on the same set of fundamental concepts, which are necessary for a data scientist to understand in order to be efficient on his quest to extract knowledge.

Expected Learning Outcomes

- Learn state-of-the-art research and industry trends in Large Scale Data Management Systems
- Understand the fundamental principles that govern all modern DBMSs
- Be able to make design decisions in deploying large scale data processing applications as well as to identify the bottlenecks of such applications
- Learn how to install and use open source systems and libraries in order to perform meaningful machine learning and text analysis tasks

Requirements and Prerequisites

The course does not assume any prior experience in databases systems.

However, the following skills would be nice to have, but not necessary:

- Undergraduate course in databases
- Some experience with declarative and functional programming
- An understanding on how operating systems and the hardware work
- Knowledge of fundamental computer science concepts
- Some experience with web applications, scripting languages, and installing and using open source software

Reading Material

The main reading material is the slides of the course. All the information that you need in order to succeed in this course is given in the slides and the notes you will take during class.

Most of the fundamental material that you will hear in this course is from the textbook:

Database System Concepts (7th Edition)

Avi Silberschatz, Henry F. Korth, S. Sudarshan

McGraw-Hill

https://www.db-book.com/db7/index.html

Advanced material on fundamental concepts are taken from "the red book":

Readings in Database Systems, 5th Edition

Peter Bailis, Joseph M. Hellerstein, Michael Stonebraker , editors

We will also try to capture during the course the state-of-the-art in large scale data management. This material is taken from conference and journal publications that are mentioned in the course slides and can be used for future reference. You are not required to read these publications in order to succeed in the course.

Software/Computing requirements

You will need to have access to a personal computer or a laptop in order to take part to the programming tasks and successfully finish the course assignments. If you don't have a personal computer or a laptop, please consult with the instructor to find a solution.

All necessary software will be provided to you. We will only use open source and free software.

Grading

Grading will be calculated as follows:

- 70% final exam
- 30% assignments

Participation

You are expected to actively participate in the class. Questions at any time during the lecture are encouraged. The lecture should not be just a lecture but also an active discussion between the participants. Your attendance is mandatory.

You may also ask questions and clarifications by sending an email to the instructor or the teaching assistant. All questions will be then answered in the next class. You may not ask questions during the breaks, use the time to relax your muscles and mind.

No laptops or phones are allowed during the lecture, unless otherwise advised. Using a laptop to take notes distracts both the instructor and you. You may have your mobile on in silent mode for emergencies,

but you are not allowed to use it otherwise or take pictures of the slides. The slides are distributed to you every week.

You may drink or eat during the lecture, provided that it is not against the rules of the room or the building.

All assignments, homework, and exams are assigned individually and not to a group, and you may not copy code or work from other people.

Assignments

You will have to successfully complete three assignments during the semester. Each next assignment requires the previous one to be completed otherwise you will not be able to continue. At the end of the assignment you will know

- how to install an open source DBMS and a number of third-party libraries,
- how to write SQL queries in order to examine and manipulated data,
- how to write User Defined Function in Python, and
- how to perform machine learning tasks inside a modern DBMS

Each assignment amounts for the 10% of the final grade, totaling to 30%. If you fail to submit the assignments, you fail the course.

Course Syllabus

The course comprises ten units of three hours each.

Unit 1: Relational Databases

We start the course by presenting the purpose of a database management system. We then introduce the relational model and get familiar with the relational algebra. We describe in detail important relational operators and we study a primer on SQL.

Unit 2: Files and Storage

The first function of a DBMS is to store the data in a persistent manner. We will study the state of the art on file formats for databases, row- and column- stores, and distributed storage filesystems.

Unit 3: Query processing

The next task of a DBMS is to process queries usually expressed in SQL. We will study all the different query processing models, such as the iterator and the vectorized model. We will concern ourselves with the pros and cons of each model and their applicability in different scenarios.

Unit 4: Scan and Indexes

No database can be efficient without efficient scan operators and sophisticated index structures to locate the relevant data. We will study the classic indexes but also new approaches on the subject.

Unit 5: Joins

We will first study functional dependencies and decomposition in order to understand why joins are fundamental to the operation of a DBMS. We will the continue to present different join algorithms. Finally, we are interested to see how joins and indexes come together in a distribute environment that is required in order to perform large scale data management.

Unit 6: Transactions

Transactions are important for guaranteeing that data remain consistent. We study the ACID properties, serializability, and concurrency control techniques.

Unit 7 and 8: Massively Parallel Databases

We study classic parallel database systems, the map/reduce programming model, and modern key-value stores. We take a close look on the internals and the functionality of important modern large scale systems, such as Spark, Flink, and others.

Unit 9: User Defined Functions and Embedded DBMS

User defined functions is the key to open the functionality of a DBMS to new areas, such as Machine Learning, Text Mining, and Artificial Intelligence. They allow the data scientist to use familiar programming languages such as Python and known libraries such as scikit-learn or tensorflow, to process information without moving the data out of the database. Embedded DBMS can also serve the same purpose, by bringing the database system in the application process.

Unit 10: Practical exercises and Recap

Unit 10 serves as a placeholder lecture for re-iterating over the most important aspects of the course and to have a hands-on experience with the software needed for the assignments. This unit may be divided in three one-hour units distributed over the other 9 units.