GHENT UNIVERSITY

Course Specifications

Valid as from the academic year 2024-2025

Big Data Science (C004074)

Course size *(nominal values; actual values may depend on programme)*

Credits 6.0 **Study time 165 h**

Course offerings and teaching methods in academic year 2024-2025

A (semester 2)	Dutch	Gent	seminar	
			peer teaching	0.0h
			group work	0.0h
			lecture	

Lecturers in academic year 2024-2025

Mesuere, Bart	WE02	lecturer-in-charge

Offered in the following programmes in 2024-2025	**crdts**	**offering**
Master of Science in Teaching in Science and Technology(main subject Computer Science)	6	A
Master of Science in Computer Science	6	A

Teaching languages

English, Dutch

Keywords

Horizontal scalability, distributed file systems, MapReduce, Spark, NoSQL databases, time series

Position of the course

This course familiarizes the student with the versatile area of Big Data Science, a scientific discipline that deals with exceptional types of data, including very large data sets, streaming data,..., that cannot be dealt with adequately anymore with classical data mining and computer architectures. The focus of the course is to learn the student a variety of ways to handle such data in a scalable way, and present the technical and architectural frameworks to allow this.
Next to these technical aspects, the course will also pay attention to actual topics related to big data science, such as ethical and privacy aspects.

Contents

Lectures with lab sessions

A selection of relevant big data topics, accompanied by a lab session. Examples of such topics:

- Cassandra
- Time Series Databases
- Recommender Systems
- Hadoop
- Spark

Guest lectures

Around 3 guest lectures that present big data in practice or explain a non-technical aspect of working with big data. Examples are

- A company presenting their data pipeline
- GDPR
- Bias in AI

Student lectures

Students present a technical topic of choice to peers in groups of 2 or 3. Examples of such topics are:

- Dask
- DuckDB
- Scylla
- Kafka
- MongoDB
- OpenNLP
- Arrow en Parquet
- Neo4j

Initial competences

Students are required to have already followed the courses programming, data-structurs and algorithms, parallel computing, machine learning and databases.

Final competences

1 Evaluate simple strategies for executing a distributed query to select the strategy that minimizes the amount of data transfer.
2 Evaluate different methodologies for effective application of data mining.
3 Identify and characterize sources of noise, redundancy, and outliers in presented data.
4 Research a big data technology and present it to peers
5 Decide on an appropriat data storage technology for a given problem
6 Know the pro's and con's of the different types of recommender systems and apply them to build a recommender system.
7 Using nosql databases such as Cassandra in practice
8 Using time series databases such as InfluxDB in practice

Conditions for credit contract

Access to this course unit via a credit contract is determined after successful competences assessment

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Group work, Seminar, Lecture, Peer teaching

Extra information on the teaching methods

Ufora will be used to ensure a smooth organisation and follow-up of the practical assignments.

Study material

Type: Handbook

Name: Mining of Massive Datasets (Jure Leskovec, Anand Rajaraman, Jeff Ullman)
Indicative price: Free or paid by faculty
Optional: yes
Language : English
Author : Jure Leskovec, Anand Rajaraman, Jeff Ullman
Online Available : Yes

Type: Slides

Name: Slides
Indicative price: Free or paid by faculty
Optional: no

References

Mining of Massive Datasets (Jure Leskovec, Anand Rajaraman, Jeff Ullman)

Course content-related study coaching

The exercises and practical assignments are supervised by the lecturer.

Assessment moments

continuous assessment

Examination methods in case of periodic assessment during the first examination period

Examination methods in case of periodic assessment during the second examination period

Examination methods in case of permanent assessment

Skills test, Presentation, Peer and/or self assessment, Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible

Extra information on the examination methods

De studenten maken individueel enkele gequoteerde practica die aansluiten op de hoorcolleges.
De studenten geven per 2-3 een korte les (15-20 minuten) over een technisch onderwerp naar keuze.
De studenten werken per 2-3 aan een project. Ze stellen het eindresultaat voor in een mondeling gesprek op afspraak tijdens de examenperiode.
Only the exam can be retaken in the summer; the points from the project from the first semester will be carried over if the student retakes in the summer.

Calculation of the examination mark

De eindscore van dit vak bestaat uit de gewogen som van de score op de verschillende onderdelen volgens onderstaande weging:

- 4 punten: de practica
- 4 punten: de les die gegeven wordt
- 12 punten: het project