GHENT UNIVERSITY

Course Specifications

Valid in the academic year 2023-2024

Distributed data processing (E761040)

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

Credits 3.0 **Study time 90 h**

Course offerings and teaching methods in academic year 2023-2024

B (semester 2)	Dutch	Gent	lecture seminar

Lecturers in academic year 2023-2024

Volckaert, Bruno	TW05	lecturer-in-charge

Offered in the following programmes in 2023-2024	crdts	offering
Bachelor of Science in Engineering Technology(main subject Information Engineering Technology)	3	B
Exchange Programme Information Engineering Technology	3	B
Linking Course Master of Science in Information Engineering Technology	3	B
Preparatory Course Master of Science in Information Engineering Technology	3	B

Teaching languages

Dutch

Keywords

Distributed data storage, parallel data processing, data processing models, Big Data, Business Intelligence, Computer Science, Informatics, Computer technology

Position of the course

In this course the challenges and techniques are studied as to how to retrieve business intelligence from large volumes of data and/or a multitude of data streams. To achieve this a data pipeline is constructed: collecting and extracting data from internal and external sources, storing this raw data, transforming it to qualitative and useable data and finally analysing and visualising the data. Students will gain insights into distributed data processing, both from a computational-processing point of view (how current data processing platforms work) as from a programming-algorithmic point of view.

Contents

Computational-processing concepts which will be tackled are
- Collecting data from different sources (files, SQL, NoSQL, streaming data, scraping)
- Distributed filesystems at the basis of a data lake
- Programming models and archtectures for processing data
- Lambda, Kappa and Zeta architecture
- Batch-processing, mini-batch, streaming
- MapReduce and modern offerings

Underlying mechanisms for parallel data processing (Bringing compute to the data) on a distributed filesystem
- Redundancy
- Sort and join
- Parallel processing
- Identification of potential bottlenecks

Programming concepts which will be explored are
- Transforming raw data to qualitative data (data cleaning)
- Noise, normalisation, transforming to proper data format
- Explorative data-analysis

- Organisation of datasets
- (Re)-sampling

Data warehouse (vs data lake) with structured data for business intelligence

- Role of data management: metadata, data discovery, data ownership, data privacy
- Datawarehouse vs datamart vs operational data store
- Extract-Transform-Load
- Facts, dimensions, star schemas, schema evolution ...
- Online Analytical Data Processing databases optimised for BI-workloads: small number of users, loads of data

Potential technologies that can exemplify above concepts are

- Extraction of data form sources: Sqoop, Flume, Hadoop Connector
- Explorative data-analysis: Python Pandas
- Data lake: Hadoop DFS and Yarn
- Configuring data workflows and basic algorithms: Apache Spark Streaming, Apache Hive, Apache Pig, Apache Beam, Apache Flink, etc.
- Notebooks and visualisation of data: Tableau, Plotly, Jupyter notebook, Apache Zeppelin

Initial competences

Object oriented programming and software development, Python

Final competences

1 Students can analyse large and varied datasets by means of distributed data processing
2 Students know programming models and platforms for processing large and varied datasets
3 Students know the techniques and visualisations of explorative data-analysis

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

Extra information on the teaching methods

Hoorcollege, werkcollege: oefeningen, groepswerk (project)

Learning materials and price

Slides on the electronic learning platform

References

Course content-related study coaching

Interactive support via the electronic learning environment; assistant-guided labs; contact with professor and assistants through mailing list and personally by means of an appointment.

Assessment moments

end-of-term and continuous assessment

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

Written assessment with open-ended questions

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

Written assessment with open-ended questions

Examination methods in case of permanent assessment

Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible

Extra information on the examination methods

First term:

- PE1: written exam with open questions

- NPE1: evaluation of result project based on report / code

Second term:

- PE2: written exam with open questions
- NPE2: evaluation of result project based on report / code

Calculation of the examination mark

- 50% of the final grade is determined by the answers to the written exam
- 50% of the final grade is determined by evaluation of the result of the project

To pass, a student needs to receive at least 9/20 for the PE. If this not the case and the calculated result is 10 or more, the final grade will be changed and the student receives 9/20.