UNIVERSITEIT GENT

# Studiefiche

Vanaf academiejaar 2022-2023

## Big Data Science (C004074)

**Wegens Covid19 kan mogelijk afgeweken worden van de onderwijs- en evaluatievormen. Dergelijke afwijkingen zullen via Ufora worden gecommuniceerd.**

**Cursusomvang** *(nominale waarden; effectieve waarden kunnen verschillen per opleiding)*

**Studiepunten 6.0** **Studietijd 165 u** **Contacturen** 62.5 u

**Aanbodsessies en werkvormen in academiejaar 2022-2023**

| | | | | |
|---|---|---|---|---|
| A (semester 2) | Nederlands | Gent | werkcollege: PC-klasoefeningen | 15.0 u |
| | | | zelfstandig werk | 10.0 u |
| | | | hoorcollege | 22.5 u |
| | | | begeleide zelfstudie | 15.0 u |

**Lesgevers in academiejaar 2022-2023**

| | | |
|---|---|---|
| Mesuere, Bart | WE02 | Verantwoordelijk lesgever |

| **Aangeboden in onderstaande opleidingen in 2022-2023** | **stptn** | **aanbodsessie** |
|---|---|---|
| Educatieve Master of Science in de wetenschappen en technologie (afstudeerrichting informatica) | 6 | A |
| Master of Science in de informatica | 6 | A |

**Onderwijstalen**

Nederlands, Engels

**Trefwoorden**

Horizontal scalability, distributed file systems, MapReduce, NoSQL databases, large-scale machine learning

**Situering**

Deze cursus maakt de student vertrouwd met het veelzijdige gebied van Big Data science, een vakgebied dat zich bezig houdt met uitzonderlijke aspecten van data, zoals zeer grote data, streaming data,... die niet meer met de klassieke data mining technieken en architecturen op een adekwate manier kan verwerkt worden. De focus van de cursus ligt op het aanbrengen van schaalbare technieken om zeer grote hoeveelheden data te kunnen verwerken, en de technische en architecturale frameworks die hiervoor gebruikt worden.
Voorts wordt in de cursus ook aandacht besteed aan actuele onderwerpen die aan bod komen bij Big Data science, zoals ethische en privacy-aspecten.

**Inhoud**

**Inleiding tot Big data**
- Geschiedenis van Big Data science
- 5V definitie van Big Data
- Toepassingen van Big Data science

**Distributed file systems**
- Google File System
- Hadoop File System (HDFS)
- Graph computation systems

**Map-Reduce en SPARK**
- MapReduce framework
- Toepassingen en limitaties van MapReduce
- SPARK

**Big Data Databases**
- NoSQL databases

**Datastructuren voor Big Data**
- Matrixrepresentaties
- Locality-sensitive hashing
- Frequent itemset mining

**Case studies in Big Data**
- Omgaan met streaming data
- Werken met grote grafen.
- Visualizatie van grote datasets

**Ethische aspecten van Big Data**
- Privacy-aspecten en GDPR regulatie
- Privacy-preserving machine learning technieken

**Begincompetenties**

De student wordt verwacht reeds cursussen programmeren, data-structuren en algoritmen, parallel computing, machine learning en databases gevolgd te hebben.

**Eindcompetenties**

1 Explain the techniques used for data fragmentation, replication, and allocation during the distributed database design process. [Familiarity]
2 Evaluate simple strategies for executing a distributed query to select the strategy that minimizes the amount of data transfer. [Assessment]
3 Explain how the two-phase commit protocol is used to deal with committing a transaction that accesses databases stored on multiple nodes. [Familiarity]
4 Describe distributed concurrency control based on the distinguished copy techniques and the voting method. [Familiarity]
5 Describe the three levels of software in the client-server model. [Familiarity]
6 Compare and contrast different uses of data mining as evidenced in both research and application. [Assessment]
7 Explain the value of finding associations in market basket data. [Familiarity]
8 Characterize the kinds of patterns that can be discovered by association rule mining. [Assessment]
9 Describe how to extend a relational system to find patterns using association rules. [Familiarity]
10 Evaluate different methodologies for effective application of data mining. [Assessment]
11 Identify and characterize sources of noise, redundancy, and outliers in presented data. [Assessment]
12 Identify mechanisms (on-line aggregation, anytime behavior, interactive visualization) to close the loop in the data mining process. [Familiarity]
13 Describe why the various close-the-loop processes improve the effectiveness of data mining. [Familiarity]

**Creditcontractvoorwaarde**

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk mits gunstige beoordeling van de competenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

**Didactische werkvormen**

Begeleide zelfstudie, hoorcollege, zelfstandig werk, werkcollege: PC-klasoefeningen

**Toelichtingen bij de didactische werkvormen**

Ufora wordt gebruikt voor een vlotte organizatie van de cursus en opvolging van de praktische sessies.
Omwille van COVID19 kunnen gewijzigde werkvormen uitgerold worden indien dit noodzakelijk blijkt

**Leermateriaal**

Slides zijn beschikbaar, en een handboek waarvan de PDF gratis kan verkregen worden wordt gevolgd.

**Referenties**

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

**Vakinhoudelijke studiebegeleiding**

De oefeningen en practica worden begeleid door de lesgever en de assistenten.

**Evaluatiemomenten**

periodegebonden en niet-periodegebonden evaluatie

**Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode**

Mondeling examen, werkstuk

**Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode**

Mondeling examen, werkstuk

**Evaluatievormen bij niet-periodegebonden evaluatie**

Vaardigheidstest, verslag

**Tweede examenkans in geval van niet-periodegebonden evaluatie**

Examen in de tweede examenperiode is mogelijk

**Eindscoreberekening**

De eindscoreberekening wordt in begin van het semester door de lesgever meegedeeld via UFORA.

**Cursusomvang** *(nominale waarden; effectieve waarden kunnen verschillen per opleiding)*

| | | | |
|---|---|---|---|
| **Studiepunten** | **6.0** | **Studietijd** | **165 u** |
| **Contacturen** | 62.5 u | | |

**Aanbodsessies en werkvormen in academiejaar 2022-2023**

| | |
|---|---|
| werkcollege: PC-klasoefeningen | 15.0 u |
| zelfstandig werk | 10.0 u |
| hoorcollege | 22.5 u |
| begeleide zelfstudie | 15.0 u |

**Lesgevers in academiejaar 2022-2023**

| | | |
|---|---|---|
| Mesuere, Bart | WEO2 | Verantwoordelijk lesgever |

| **Aangeboden in onderstaande opleidingen in 2022-2023** | **stptn** | **aanbodsessie** |
|---|---|---|
| Educatieve Master of Science in de wetenschappen en technologie (afstudeerrichting informatica) | 6 | A |
| Master of Science in de informatica | 6 | A |

**Onderwijstalen**

Nederlands, Engels

**Trefwoorden**

Horizontal scalability, distributed file systems, MapReduce, NoSQL databases, large-scale machine learning

**Situering**

Deze cursus maakt de student vertrouwd met het veelzijdige gebied van Big Data science, een vakgebied dat zich bezig houdt met uitzonderlijke aspecten van data, zoals zeer grote data, streaming data,... die niet meer met de klassieke data mining technieken en architecturen op een adekwate manier kan verwerkt worden. De focus van de cursus ligt op het aanbrengen van schaalbare technieken om zeer grote hoeveelheden data te kunnen verwerken, en de technische en architecturale frameworks die hiervoor gebruikt worden.
Voorts wordt in de cursus ook aandacht besteed aan actuele onderwerpen die aan bod komen bij Big Data science, zoals ethische en privacy-aspecten.

**Inhoud**

**Inleiding tot Big data**

- Geschiedenis van Big Data science
- 5V definitie van Big Data
- Toepassingen van Big Data science

**Distributed file systems**
- Google File System
- Hadoop File System (HDFS)
- Graph computation systems

**Map-Reduce en SPARK**
- MapReduce framework
- Toepassingen en limitaties van MapReduce
- SPARK

**Big Data Databases**
- NoSQL databases

**Datastructuren voor Big Data**
- Matrixrepresentaties
- Locality-sensitive hashing
- Frequent itemset mining

**Case studies in Big Data**
- Omgaan met streaming data
- Werken met grote grafen.
- Visualizatie van grote datasets

**Ethische aspecten van Big Data**
- Privacy-aspecten en GDPR regulatie
- Privacy-preserving machine learning technieken

**Begincompetenties**

De student wordt verwacht reeds cursussen programmeren, data-structuren en algoritmen, parallel computing, machine learning en databases gevolgd te hebben.

**Eindcompetenties**

1 Explain the techniques used for data fragmentation, replication, and allocation during the distributed database design process. [Familiarity]
2 Evaluate simple strategies for executing a distributed query to select the strategy that minimizes the amount of data transfer. [Assessment]
3 Explain how the two-phase commit protocol is used to deal with committing a transaction that accesses databases stored on multiple nodes. [Familiarity]
4 Describe distributed concurrency control based on the distinguished copy techniques and the voting method. [Familiarity]
5 Describe the three levels of software in the client-server model. [Familiarity]
6 Compare and contrast different uses of data mining as evidenced in both research and application. [Assessment]
7 Explain the value of finding associations in market basket data. [Familiarity]
8 Characterize the kinds of patterns that can be discovered by association rule mining. [Assessment]
9 Describe how to extend a relational system to find patterns using association rules. [Familiarity]
10 Evaluate different methodologies for effective application of data mining. [Assessment]
11 Identify and characterize sources of noise, redundancy, and outliers in presented data. [Assessment]
12 Identify mechanisms (on-line aggregation, anytime behavior, interactive visualization) to close the loop in the data mining process. [Familiarity]
13 Describe why the various close-the-loop processes improve the effectiveness of data mining. [Familiarity]

**Creditcontractvoorwaarde**

Toelating tot dit opleidingsonderdeel via creditcontract is mogelijk mits gunstige beoordeling van de competenties

Examencontractvoorwaarde

Dit opleidingsonderdeel kan niet via examencontract gevolgd worden

**Didactische werkvormen**

Begeleide zelfstudie, hoorcollege, zelfstandig werk, werkcollege: PC-klasoefeningen

**Toelichtingen bij de didactische werkvormen**

Ufora wordt gebruikt voor een vlotte organizatie van de cursus en opvolging van de praktische sessies.
Omwille van COVID19 kunnen gewijzigde werkvormen uitgerold worden indien dit noodzakelijk blijkt

**Leermateriaal**

Slides zijn beschikbaar, en een handboek waarvan de PDF gratis kan verkregen worden wordt gevolgd.

**Referenties**

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

**Vakinhoudelijke studiebegeleiding**

De oefeningen en practica worden begeleid door de lesgever en de assistenten.

**Evaluatiemomenten**

periodegebonden en niet-periodegebonden evaluatie

**Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode**

Mondeling examen, werkstuk

**Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode**

Mondeling examen, werkstuk

**Evaluatievormen bij niet-periodegebonden evaluatie**

Vaardigheidstest, verslag

**Tweede examenkans in geval van niet-periodegebonden evaluatie**

Examen in de tweede examenperiode is mogelijk

**Eindscoreberekening**

De eindscoreberekening wordt in begin van het semester door de lesgever meegedeeld via UFORA.