

Big Data Science (C004074)

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

Studiepunten 6.0 **Studietijd 165 u**

Aanbodsessies en werkvormen in academiejaar 2024-2025

A (semester 2)	Nederlands	Gent	werkcollege	
			peer teaching	0.0u
			groepswerk	0.0u
			hoorcollege	

Lesgevers in academiejaar 2024-2025

Mesuere, Bart WE02 Verantwoordelijk lesgever

Aangeboden in onderstaande opleidingen in 2024-2025

	stptn	aanbodsessie
Educatieve Master of Science in de wetenschappen en technologie (afstudeerrichting informatica)	6	A
Master of Science in de informatica	6	A

Onderwijstalen

Engels, Nederlands

Trefwoorden

Horizontaal scalability, distributed file systems, MapReduce, Spark, NoSQL databases, tijdreeksen

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 adequate manier kan verwerkt worden. De focus van de cursus ligt op het aanbrenge 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

Hoorcollege's met practica

Een selectie van relevante big data onderwerpen, meestal vergezeld van een practicum.

Voorbeelden van onderwerpen die eerder al aan bod kwamen:

- Cassandra
- Tijdreeksdatabanken
- Aanbevelingssystemen
- Hadoop
- Spark

Gastlessen

Er komen een drietal gastsprekers aan bod die inzicht geven over big data in de praktijk of een andere blik werpen op het gebruik van grote hoeveelheden data. Voorbeelden van onderwerpen die eerder al aan bod kwamen:

- Een bedrijf dat hun data pipeline komt voorstellen
- GDPR
- Bias in AI

Lessen door studenten

De studenten geven (in groepjes van 2-3) zelf een korte les over een big data technologie naar keuze. De bedoeling is om aan collega's een technologie te introduceren, kort de werking uit te leggen en de voor- en nadelen te verkennen. Voorbeelden van onderwerpen die eerder al aan bod kwamen:

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

Begincompetenties

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

Eindcompetenties

- 1 Verschillende aanpakken voor het uitvoeren van een gedistribueerde query kunnen evalueren en degene selecteren die voor de minste datatransfer zorgt
- 2 Verschillende aanpakken voor de toepassing van data mining kunnen beoordelen
- 3 Bronnen van ruis, redundantie en outliers kunnen identificeren en karakteriseren in data
- 4 Een big data technologie zelfstandig onderzoeken en presenteren aan peers
- 5 De gepaste opslagtechnologie kunnen bepalen voor een probleem
- 6 De voor- en nadelen van verschillende types aanbevelingssystemen beheersen en kunnen toepassen
- 7 Nosql databanken zoals Cassandra kunnen gebruiken in de praktijk
- 8 Tijdreeksdatabanken zoals InfluxDB kunnen gebruiken in de praktijk

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

Groepswerk, Werkcollege, Hoorcollege, Peer teaching

Toelichtingen bij de didactische werkvormen

Ufora wordt gebruikt voor een vlotte organisatie van de cursus en opvolging van de praktische sessies.

Studiemateriaal

Type: Handboek

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

Richtprijs: Gratis of betaald door opleiding

Optioneel: ja

Taal : Engels

Auteur : Jure Leskovec, Anand Rajaraman, Jeff Ullman

Online beschikbaar : Ja

Type: Slides

Naam: Slides

Richtprijs: Gratis of betaald door opleiding

Optioneel: nee

Referenties

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

Vakinhoudelijke studiebegeleiding

De oefeningen en practica worden begeleid door de lesgever.

Evaluatiemomenten

niet-periodegebonden evaluatie

Evaluatievormen bij periodegebonden evaluatie in de eerste examenperiode

Evaluatievormen bij periodegebonden evaluatie in de tweede examenperiode

Evaluatievormen bij niet-periodegebonden evaluatie

Vaardigheidstest, Presentatie, Peer en/of self assessment, Werkstuk

Tweede examenkans in geval van niet-periodegebonden evaluatie

Examen in de tweede examenperiode is mogelijk

Toelichtingen bij de evaluatievormen

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.

De practica kunnen niet hernomen worden in tweede zitting. De punten voor de practica worden overgenomen van eerste naar tweede zitting.

Eindscoreberekening

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