

POSTACADEMISCHE OPLEIDING

BIG DATA

(via e-learning)

18 maart 2021 – 24 juni 2021



UNIVERSITEIT
GENT

Deze opleiding is breed opgevat en behandelt de belangrijkste aspecten van datavergaring, -beheer, -analyse en -presentatie. Zo krijg je een totaalinzicht dat je beter in staat zal stellen om jouw Big Data efficiënt aan te wenden ten voordele van jouw onderneming. De aangeleerde technische inzichten worden aangevuld met een degelijk basisinzicht in de juridische uitdagingen rond Big Data projecten zoals privacy en gegevensbescherming, discriminatie en intellectuele eigendomsrechten.

Big Data kunnen worden omschreven als gegevenscollecties die niet efficiënt met traditionele gegevensbeheer en -verwerkingstechnieken kunnen worden behandeld. Bepalende factoren daarbij zijn de grotere datavolumes, de grotere snelheden waarmee de data worden aangeboden en de grotere variëteit aan dataformaten en de kwaliteit van de data. De tendens naar Big Data wordt gevoed door de almaar groeiende beschikbaarheid van digitale informatie uit nieuwsbronnen, multimedia, sensors, ... en gaat gepaard met nieuwe uitdagingen om deze data efficiënt te kunnen verzamelen, opslaan, beheren, analyseren en presenteren.

Het inzetten van geavanceerde technologieën die specifiek zijn afgestemd op het verwerken van zeer grote hoeveelheden data, kan bedrijven helpen om beter tegemoet te komen aan de steeds groter wordende informatienoden die vaak vereist zijn om gegevensanalyse nog beter te kunnen onderbouwen. Een beter inzicht in de beschikbare data en een optimale exploitatie ervan levert de beste garantie om met meer kennis van zaken belangrijke beslissingen te onderbouwen en daar dan ook een concurrentieel voordeel mee te behalen.

DOELPUBLIEK

U krijgt inzicht in de problematiek die gepaard gaat met Big Data en in de beschikbare ICT-oplossingen die momenteel voorhanden zijn. Er wordt aangetoond hoe de aangereikte oplossingen werken, wat hun beperkingen en voordelen zijn en waar en wanneer ze het beste kunnen worden ingezet.

Voor de lessen wordt bewust gekozen voor een sterke academische aanpak waarbij de hoofdaccenten liggen op het verwerven van kennis in de breedte zonder daarbij productgebonden te zijn. Daarnaast wordt ruim aandacht besteed aan biomedische en businessstoepassingen.

De opleiding is dusdanig opgevat dat deze toegankelijk is voor iedereen die ietwat vertrouwd is met informatica. Er wordt gewerkt rond hoorcolleges die handelen rond vier thema's: gegevensbeheer, gegevensanalyse, gebruikaspecten en juridische aspecten.

GETUIGSCHRIFT

U ontvangt een getuigschrift, indien u deelneemt aan de volledige opleiding (minstens 80% aanwezigheid vereist tijdens de live online sessies) en slaagt voor het bijbehorende examen (8 juli 2021 om 14u of 9 september 2021 om 16u).

WETENSCHAPPELIJKE COÖRDINATIE

Prof. dr. Guy De Tré, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent

LESGEVERS

- **Michael Brands**, Dynactionize NV
- **Antoon Bronselaer**, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Thomas Demeester**, Vakgroep Informatietechnologie, Universiteit Gent
- **Guy De Tré**, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Jan Fostier**, Vakgroep Informatietechnologie, Universiteit Gent
- **Simon Geiregat**, Vakgroep Metajuridica, Privaat- en Ondernemingsrecht, Universiteit Gent
- **Peter Lambert**, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- **Eva Lievens**, Vakgroep Metajuridica, Privaat- en Ondernemingsrecht, Universiteit Gent
- **Filip Pattyn**, Ontoforce
- **Dirk Van den Poel**, Vakgroep Marketing, Universiteit Gent
- **Katrien Verbert**, Departement Computerwetenschappen, KU Leuven
- **Ruben Verborgh**, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- **Bruno Volckaert**, Vakgroep Informatietechnologie, Universiteit Gent

MEER INFO EN INSCHRIJVEN

WWW.UGAIN.UGENT.BE/BIGDATA

PROGRAMMA

1. GEGEVENSBEHEER

Inleiding en NoSQL (3 uren)

In de introductie wordt aandacht besteed aan de oorsprong van de term Big Data. Aspecten zoals de interpretatie, het belang, de problematiek en de kritiek op Big Data worden besproken.

Daarna komen de verschillende vormen en karakteristieken (Volume, Variety, Velocity en Veracity) van Big Data aan bod. Er wordt gekeken naar de tekortkomingen en beperkingen van traditionele databanksystemen en er wordt dieper ingegaan op mogelijke oplossingen. Vervolgens worden de belangrijkste NoSQL databankoplossingen ('Not only' SQL) gesitueerd en bestudeerd. Zowel key/value stores, documentdatabanken, column stores als graafdatabanken worden daarbij behandeld.

Datakwaliteit (3 uren)

In deze les wordt een overzicht gegeven van de verschillende technieken waarmee men datakwaliteit kan meten. Er wordt vervolgens uitgelegd hoe de meetresultaten van de verschillende technieken geïnterpreteerd moeten worden en hoe ze verder kunnen worden gebruikt in bijvoorbeeld rapportering en strategische analyses. Nadien worden ook enkele technieken uitgelegd voor de verbetering van kwaliteit. Alle methoden worden toegelicht aan de hand van cases uit de praktijk.

Information retrieval en document classificatie (3 uren)

Eerst wordt er een inleiding gegeven op 'Information Retrieval' (IR), de technologie die toelaat om grote digitale collecties van ongestructureerde teksten efficiënt te doorzoeken. De bedoeling is eerder om een overzicht te geven van de basisconcepten zoals indexering en retrieval modellen, dan om hier heel technisch op in te gaan. Nadien worden de mogelijkheden bekeken om in dergelijke collecties structuur aan te brengen door de documenten op specifieke manieren te gaan groeperen. Er worden enkele basisconcepten aangebracht rond machinaal leren in het algemeen, en rond classificatie en clustering. Vervolgens worden een paar belangrijke algoritmen besproken, en toegelicht via enkele applicaties.

Linked Data (3 uren)

Wat komt er na Big Data, als we gegevens niet meer in één plek bij elkaar kunnen brengen omwille van praktische, legale of andere redenen? In deze les bekijken we Linked Data, een ander manier om met gegevens om te gaan waarin data inherent verspreid zit over een netwerk. We behandelen technologieën uit het semantisch web met het oog op het machine-leesbaar maken van data en informatie. We bestuderen de noodzaak van semantiek om die data aan elkaar te koppelen. Daarnaast komen ook de principes van Open Data aan bod, met als doel het uniform beschikbaar stellen van data via het web.

2. GEGEVENSANALYSE

Het MapReduce-programmeermodel (3 uren)

MapReduce is een generiek raamwerk dat bijzonder geschikt is om op eenvoudige wijze analyses uit te voeren op enorme hoeveelheden data, gebruik makend van een parallel computersysteem. De basisconcepten rond 'Mappers' en 'Reducers' komen aan bod, alsook enkele veelgebruikte ontwerp patronen. Deze laatste worden gestaafd aan de hand van eenvoudig te begrijpen voorbeelden. De Hadoop-implementatie van MapReduce wordt besproken, alsook het verwante Hadoop Distributed File System (HDFS). De doelstelling is dat deelnemers na deze sessie in staat zijn Hadoop MapReduce toe te passen op eigen problemen.

Gedistribueerde gegevensverwerking (3 uren)

Voor de efficiënte verwerking van Big Data is men grotendeels naar volledige gedistribueerde vormen van gegevensverwerking overgestapt. Tijdens deze lesavond wordt u wegwijs gemaakt in enkele van de belangrijkste architecturen voor gedistribueerde gegevensverwerking (stream-gebaseerde dataverwerking, Lambda architectuur, Kappa architectuur, Microservices architectuur, Zeta architectuur, etc.). De opbouw en werking van deze architecturen worden besproken, hoe deze geheel of gedeeltelijk mappen op bestaande technologieën / implementaties (Apache Storm, Apache Samza, Apache Spark, Apache Kafka, etc.) en wat hun belangrijkste voor- en nadelen zijn. Dit geheel wordt aangevuld met voorbeelden van gedistribueerde architecturen die technologie reuzen zoals LinkedIn, Netflix, etc. geadopteerd hebben om hen om te laten gaan met de enorme hoeveelheid data die ze dagelijks moeten verwerken.

Deep learning (3 uren)

Kunstmatige neurale netwerken zijn in staat om het menselijk leerproces na te bootsen door het veranderen van de sterkte van gesimuleerde neurale verbindingen, een eigenschap die ervoor zorgt dat deze netwerken uiterst effectief zijn in het automatisch terugvinden van patronen in grote hoeveelheden data (deep learning). Dit heeft onlangs geleid tot een aantal doorbraken op het vlak van taalverwerking en audiovisuele analyse. Voortbouwend op een aantal basisconcepten uit het domein van machinaal leren, wordt er in deze les bijzondere aandacht besteed aan het gebruik van meerlagige neurale netwerken, alsook aan de technieken die de inzet van deze netwerkarchitecturen praktisch haalbaar hebben gemaakt. Vervolgens wordt er stilgestaan bij verschillende toepassingen op het vlak van taalverwerking en audiovisuele analyse, illustrerend hoe meerlagige neurale netwerken kunnen aangewend worden om kennis te extraheren uit grote hoeveelheden ruizige data. Tot slot wordt er eveneens een overzicht gegeven van toekomstige uitdagingen op het vlak van onderzoek en ontwikkeling in het domein van deep learning.

3. GEBRUIKSASPECTEN

In deze module wordt Big Data benaderd vanuit het standpunt van de gebruiker. Daarbij lichten we twee belangrijke toepassingsdomeinen nader toe, nl. business-toepassingen en biomedische toepassingen en staan we stil bij de nieuwste technologie om Big Data te visualiseren en tekstuele data semantisch te interpreteren en te verwerken.

Biomedische data-analyse (3 uren)

In deze lessen wordt uitgelegd en gedemonstreerd hoe biomedische Big Data aan elkaar kunnen worden gelinkt en doorzocht. De gepresenteerde aanpak illustreert tevens hoe Big Data kunnen worden aangewend om te komen tot beter doordachte, data-gedreven beslissingen, wat op zijn beurt bijdraagt tot betere inzichten en versneld biomedisch onderzoek. In de les wordt aandacht besteed aan de gevolgde aanpak, de mogelijke valkuilen en aandachtspunten voor de ontwikkelaar en gebruiker.

Data-analyse voor business (3 uren)

Deze lessen lichten het analyseren van Big Data in business context toe. Er wordt gestart met een inleiding tot 'Analytics' met aandacht voor 'descriptive', 'predictive' en 'prescriptive analytics'. Daarna wordt aandacht besteed aan het gebruik van de Spark (Streaming) open source software stack Berkeley Data Analytics Stack (BDAS). Vervolgens wordt het gebruik van MLib (machine learning library) binnen Spark voor 'Analytics' toegelicht en worden enkele gebruikerscases besproken. Tenslotte wordt een hands-on demo gegeven van de UGent-implementatie van Berkeley open source software stack.

Visualisatie (3 uren)

Een uitgelezen manier om mensen te helpen om Big Data te exploreren en te begrijpen, is het visualiseren van de data: we zijn immers vaak erg goed in staat om patronen, tendensen, uitschieters, ... te begrijpen met behulp van visualisaties. Mede aan de hand van een groot aantal concrete voorbeelden wordt uiteengezet hoe een goede interactieve visualisatie kan worden opgebouwd, wat de typische misvattingen zijn, hoe visualisaties kunnen misbruikt worden, ... Er wordt ook een overzicht gepresenteerd van een aantal typische technieken en hulpmiddelen voor interactieve informatie-visualisatie.

Tekst en natuurlijke taal (3 uren)

Algemeen wordt aangenomen dat 80% van alle beschikbare informatie vervat zit in tekstuele documenten. Tekstuele data adequaat semantisch kunnen interpreteren en koppelen aan elkaar is één van de vereisten en tegelijkertijd grote uitdaging voor veel Big Data projecten. In deze les wordt uitgelegd hoe tekstuele data semantisch kan worden geanalyseerd en beheerd met een NoSQL databankbeheersysteem. Daarnaast wordt gedemonstreerd hoe deze data efficiënt kunnen worden doorzocht via interactieve 'dashboard'-toepassingen. Bovendien worden enkele reële casussen besproken.

4. JURIDISCHE ASPECTEN (9 UREN)

U wordt wegwijs gemaakt in de juridische uitdagingen rond Big Data projecten zoals privacy en gegevensbescherming, discriminatie, intellectuele eigendomsrechten en andere relevante topics. Aan de hand van voorbeelden worden een paar veel voorkomende juridische problemen uit de praktijk en hun mogelijke aanpak nader toegelicht. U krijgt tevens een aantal tools en checklists aangereikt omtrent het op een juridisch correcte manier aanvatten en uitvoeren van Big Data projecten.

PRAKTISCH

Lesmateriaal

Het lesmateriaal wordt enkel digitaal aangeboden.

Prijs

Betaling geschiedt na ontvangst van de factuur.

Alle facturen zijn betaalbaar dertig dagen na dagtekening.

Alle vermelde bedragen zijn vrij van BTW.

Module 1 GEGEVENSBEHEER	€ 660,-
Module 2 GEGEVENSANALYSE	€ 495,-
Module 3 GEBRUIKSASPECTEN	€ 660,-
Module 4 JURIDISCHE ASPECTEN	€ 495,-
Volledige opleiding	€ 2.079,-

Voor iedere module kan er afzonderlijk ingeschreven worden.

Korting

- Indien minstens één deelnemer van een bedrijf inschrijft voor de volledige opleiding wordt voor alle bijkomende gelijktijdige inschrijvingen van hetzelfde bedrijf een korting van 20% verleend. Facturatie geschiedt dan d.m.v. een gezamenlijke factuur.
- Aangepaste prijzen voor personeel van UGent
- Kortingen zijn niet cumuleerbaar.

Annulering

Raadpleeg onze annulatievoorwaarden op www.ugain.ugent.be/annulatievoorwaarden

KMO-portefeuille

Universiteit Gent aanvaardt betalingen via de KMO-portefeuille: www.kmo-portefeuille.be; gebruik autorisatiecode DV.0103194.



WWW.UGAIN.UGENT.BE/BIGDATA



Tijdstip en locatie

- Deze opleiding wordt **volledig online** aangeboden.
- De **theoretische lessen worden via opnames** aangeboden op een digitaal leerplatform en na ieder lesonderdeel wordt er een live online sessie georganiseerd (via Teams). Tijdens deze live sessie gaat de lesgever dieper in op de inhoud van de les en wordt uw kennis getoetst. Er is ook mogelijkheid om vragen te stellen aan de lesgever.
- De opnames zullen telkens 14 dagen voor de online meeting beschikbaar worden gesteld.
- Aanwezigheid tijdens (minstens 80% van) de live sessies is verplicht. De live sessies starten telkens om 19u.
- Data onder voorbehoud van wijzigingen om onvoorziene omstandigheden.

Overzicht online live sessies

1. GEGEVENSBEHEER

18 maart 2021	Inleiding en NoSQL Guy De Tré
25 maart 2021	Datakwaliteit Antoon Bronselaer
1 april 2021	Information retrieval en document classificatie Thomas Demeester
22 april 2021	Linked Data Peter Lambert en Ruben Verborgh

2. GEGEVENSANALYSE

29 april 2021	Het MapReduce-programmeermodel Jan Fostier
12 mei 2021	Gedistribueerde gegevensverwerking (Bruno Volckaert) Deep learning (Thomas Demeester)

3. GEBRUIKSASPECTEN

20 mei 2021	Biomedische data-analyse Filip Pattyn
27 mei 2021	Data-analyse voor business Dirk Van den Poel
3 juni 2021	Visualisatie Katrien Verbert
10 juni 2021	Tekst en natuurlijke taal Michael Brands

4. JURIDISCHE ASPECTEN

24 juni 2021	Gegevensbescherming, discriminatie en informatieveiligheid (Eva Lievens) & Intellectuele rechten (Simon Geiregat)
--------------	--

Organisatie

Universiteit Gent

UGain (UGent Academie voor Ingenieurs)
Technologiepark 60
9052 Zwijnaarde
09 264 55 82

ugain@ugent.be - www.ugain.ugent.be