



POSTACADEMISCHE OPLEIDING

# BIG DATA HANDS-ON

24 april 2019 – 19 juni 2019



UNIVERSITEIT  
GENT



De huidige IT biedt ons vandaag almaar meer informatie, wat gepaard gaat met nieuwe uitdagingen voor gegevensbeheer, -verwerking en -analyse. Deze opleiding is opgebouwd uit een aantal praktische hands-on rond het omgaan met Big Data waarmee u zelf aan de slag kan. Naast praktisch inzicht verwerven, wordt ook aandacht besteed aan valkuilen en good practices. Dit levert u essentiële praktische know-how die u kan gebruiken voor uw Big Data projecten.

Deze opleiding vraagt programmeerervaring en een goede vertrouwdheid met informatica en bouwt verder op de basisopleiding 'Big Data' (deze opleiding gevolgd hebben is een meerwaarde, maar geen vereiste).

Door het aantal deelnemers te beperken, beogen we een goede individuele begeleiding.

### DOELPUBLIEK

De lessen zijn bedoeld voor iedereen die een goede professionele vertrouwdheid met informatica heeft en die graag praktisch, via hands-on, aan de slag wil met Big Data. Deelnemers hebben een hogere opleiding in de informatica gevolgd of hebben een gelijkwaardige ervaring opgebouwd. Deelnemers hebben programmeerervaring met Python of een aanverwante programmeertaal.

Het aantal deelnemers is beperkt tot 28.

Na deelname ontvangt u een aanwezigheidsattest.

### WETENSCHAPPELIJKE COÖRDINATIE

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

### LESGEVERS

- **Giannis Bekoulis**, Vakgroep Informatietechnologie, Universiteit Gent
- **Toon Boeckling**, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Sander Borny**, Vakgroep Informatietechnologie, Universiteit Gent
- **Cedric De Boom**, Mediaaan - De Persgroep Publishing
- **Dries Decap**, Vakgroep Informatietechnologie, Universiteit Gent
- **Thomas Demeester**, Vakgroep Informatietechnologie, Universiteit Gent
- **Robin De Mol**, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Jan Fostier**, Vakgroep Informatietechnologie, Universiteit Gent
- **Merlijn Sebrechts**, Vakgroep Informatietechnologie, Universiteit Gent
- **Lucas Sterckx**, Vakgroep Informatietechnologie, Universiteit Gent
- **Gregory Van Seghbroeck**, Vakgroep Informatietechnologie, Universiteit Gent
- **Bruno Volckaert**, Vakgroep Informatietechnologie, Universiteit Gent

## MEER INFO EN INSCHRIJVEN

[WWW.UGAIN.UGENT.BE/BIGDATAHANDSON](http://WWW.UGAIN.UGENT.BE/BIGDATAHANDSON)

## 1. GEGEVENSBEHEER

- |          |                                                                                                                                                                                                                                           |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 24 april | <b>NoSQL (Volume, Velocity)</b><br>Toon Boeckling en Robin De Mol<br><b>Vereisten (laptop):</b> De oefeningen kunnen gemaakt worden op uw eigen laptop. Er is geen connectiviteit nodig met externe machines.                             |
| 8 mei    | <b>Datakwaliteit en data-integratie (Variety, Veracity)</b><br>Toon Boeckling en Robin De Mol<br><b>Vereisten (laptop):</b> De oefeningen kunnen gemaakt worden op uw eigen laptop. Er is geen connectiviteit nodig met externe machines. |
| 15 mei   | <b>Document classificatie: inleiding tot Python</b><br>Cedric De Boom, Thomas Demeester en Lucas Sterckx<br><b>Vereiste voorkennis:</b> kennis van basisconcepten in programmeren                                                         |
| 22 mei   | <b>Document classificatie</b><br>Giannis Bekoulis, Thomas Demeester en Lucas Sterckx                                                                                                                                                      |

## 2. GEGEVENSANALYSE

- |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 29 mei & 5 juni | <b>Het MapReduce-programmeermodel</b><br>Jan Fostier en Dries Decap<br><b>Vaardigheden:</b> Basiskennis van Linux commando's voor bestandsmanipulatie (cd / mv / rm / mkdir / rmdir / ls / cat / grep / head / tail / wc / etc). Basiskennis van programmeren in Python is ten eerste aangeraden. In principe zal het programmeerwerk zeer beperkt zijn (aanpassen scripts, lopen scripts, etc.).<br><b>Vereisten (laptop):</b> Eigen laptop met een recente Linux distributie en root access (voor Docker installatie). Het downloaden en testen van de Docker container gebeurt best op voorhand aan de hand van een tutorial.                                                                  |
| 12 juni         | <b>De Kappa-architectuur</b><br>Sander Borny, Merlijn Sebrechts, Gregory Van Seghbroeck en Bruno Volckaert<br><b>Vaardigheden:</b> Kennis van een programmeertaal of een scripttaal, zoals bijvoorbeeld Java, Scala, Python of Javascript, is vereist. Tijdens de sessie wordt Python als scripttaal gebruikt, maar alle nodige Big Data en Python specifieke concepten worden zeker tijdens de sessie behandeld, waardoor het voldoende is om te kunnen programmeren, los van taal en omgeving.<br><b>Vereisten (laptop):</b> Jupyter Notebooks wordt gebruikt waardoor niets effectief op de laptop van de deelnemers moet geïnstalleerd worden. Alles kan gebeuren binnen de browser omgeving. |
| 19 juni         | <b>Deep learning</b><br>Giannis Bekoulis, Thomas Demeester en Lucas Sterckx                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

## PROGRAMMA

### MODULE 1: GEGEVENSBEHEER

#### NoSQL (Volume, Velocity)

Tijdens deze lesavond leert u hoe u grote datavolumes en snelle data-invoerstromen kunt beheeren met NoSQL databanken. We leren u 'high velocity streams' weg te schrijven naar verschillende 'data stores' en confronteren u met de verschillen tussen NoSQL databanksystemen, relationele databanksystemen en systemen voor het beheer van tijdseries. Verder besteden we aandacht aan de manier waarop databanksystemen omgaan met het gelijktijdige, 'high load' operaties. Tenslotte laten we u ervaren welke impact verschillende parameterinstellingen en benaderingen hebben op het wegschrijven van data in een NoSQL databank.

#### Datakwaliteit en data-integratie (Variety, Veracity)

De bedoeling van deze lesavond is om in de praktijk te leren omgaan met de twee andere hoofdkenmerken van Big Data, nl. kwaliteit en variëteit. We leggen hierbij de focus op technieken om de kwaliteit van data te meten. Er wordt onder andere gekeken naar volledigheid, courantheid en consistentie tussen metingen. Er wordt ook getoond hoe een kostenmodel kan gebruikt worden om kwaliteitsniveaus tastbaar te maken.

#### Document classificatie: inleiding tot Python

Tijdens deze lesavond leert u een aantal basisvaardigheden rond programmeren met de programmeertaal Python die vereist zijn voor een aantal van de verdere praktische lessen binnen de opleiding. Bijzondere aandacht gaat naar het werken met Python packages, in het bijzonder packages voor het bouwen van machine learning toepassingen.

#### Document classificatie

Op deze lesavond wordt er een praktisch overzicht gegeven van technieken om tekst-gebaseerde data voor te bereiden voor het trainen en evalueren van machine learning toepassingen. Daarnaast worden een aantal basisconcepten rond machine learning aangebracht, die ook voor de vervolglessen belangrijk blijven. Er wordt gewerkt rond een toepassing die bestaat uit het automatisch classificeren van volledige documenten in specifieke categorieën via klassieke machine learning technieken. Voor de toepassing worden de Python packages uit de vorige lesavond gebruikt.

### MODULE 2: GEGEVENSANALYSE

#### Het MapReduce-programmeermodel

In de deze les leert u werken met gedistribueerde gegevensopslag en -verwerking. Eerst besteden we aandacht aan het gebruik van het Hadoop Distributed File System (HDFS). Daarbij leert u de basiscommando's en illustreren we hoe u data nodes en name nodes kunt inspecteren en hoe u bestanden kunt opsplitsen in partities voor gedistribueerde dataopslag. Vervolgens brengen we u, aan de hand van MapReduce streaming, basiskennis bij over Hadoop MapReduce. We starten met enkele eenvoudige voorbeelden zoals word counting en gaan dan verder met iets moeilijker oefeningen waarbij 2 MapReduce jobs noodzakelijk zijn. Tijdens de oefeningen leert u tevens hoe u MapReduce streaming applicaties kan debuggen en hoe u de 'partitioner', 'combiner' en 'sort' modules in Hadoop Streaming kan gebruiken. Ter illustratie voeren we een performantietest uit op de UGent HPC. Tenslotte leert u werken met Spark. We bestuderen de basisfunctionaliteit van Spark aan de hand van PySpark Shell / Notebooks, we leren werken met de Resilient Distributed Dataset (RDD) en leren omgaan met acties en transformaties, Stand-alone batch processing en interactive, schaalbare dataexploratie via Notebooks.

#### De Kappa-architectuur

Tijdens deze lesavond kunt u proeven van enkele technologieën die gebruikt kunnen worden in een Kappa-architectuur. Deze architectuur is een courant gebruikte architectuur voor gedistribueerde gegevensverwerking en bestaat uit een samenstelling van de volgende technologieën:

- Kafka: een gedistribueerd message systeem
- Spark: een schaalbaar platform voor het analyseren van Big Data, dat zowel geschikt is voor stream als batch-analyses
- HDFS: een gedistribueerd en shared file systeem
- ArangoDB: een nieuwe populaire multimodale data store

Tijdens deze hands-on leren we u hoe u rechtstreeks op Kafka datastromen kunt invoeren en hoe u deze kunt analyseren. We leren u ook hoe u een eenvoudige data clean-up kunt doen met Spark en de resultaten hiervan kunt opslaan in HDFS. Vervolgens voeren we batch-analyses uit op de data in HDFS en sturen we de resultaten naar ArangoDB. Tenslotte leren we u hoe uw bevindingen kunt presenteren via Notebooks.

#### Deep learning

Het doel van deze lesavond is een praktische kennismaking met deep learning, via opnieuw een toepassing met tekst-gebaseerde data maar deze keer op zinsniveau. We maken opnieuw oefeningen in Python. Na een korte inleiding in courante deep learning pakketten via python worden gradueel complexere modellen aangebracht. De reeds aangebrachte basisconcepten in machine learning worden uitgebreid met een aantal frequent gebruikte bouwstenen van diepe neurale netwerken. Daarnaast leren we u hoe training via gradient-gebaseerde optimalisatietechnieken werkt. De nadruk zal liggen op de praktische aspecten, eerder dan op de achterliggende wiskunde.

## PRAKTISCH

### Prijs

Deelnameprijs omvat lesgeld, hand-outs, frisdranken, koffie en broodjes.

Betaling geschiedt na ontvangst van de factuur.

Alle facturen zijn betaalbaar dertig dagen na dagtekening.

Alle vermelde bedragen zijn vrij van BTW.

**Volledige opleiding**

**€ 1.600-**

### 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 en geassocieerde hogescholen.
- Kortingen zijn niet cumuleerbaar.

### Annulering

Raadpleeg onze annulatievoorwaarden op [www.ugain.ugent.be/annulatievoorwaarden](http://www.ugain.ugent.be/annulatievoorwaarden)

### KMO-portefeuille

Universiteit Gent aanvaardt betalingen via de KMO-portefeuille ([www.kmo-portefeuille.be](http://www.kmo-portefeuille.be); gebruik autorisatiecode DV.0103194).

### Tijdstip en locatie

- De lessen worden gegeven van 18u tot 21u30, in 2 delen, gescheiden door een broodjesmaaltijd en vinden plaats aan de Universiteit Gent, UGent Academie voor Ingenieurs, Technologiepark 60, 9052 Zwijnaarde.
- Data onder voorbehoud van wijzigingen om onvoorziene omstandigheden.

### Laptop

Eigen laptop is vereist. Per lesavond staan de vereisten voor uw laptop vermeld.

### Organisatie

#### Universiteit Gent

UGain (UGent Academie voor Ingenieurs)

Technologiepark 60

9052 Zwijnaarde

09 264 55 82

[ugain@ugent.be](mailto:ugain@ugent.be) - [www.ugain.ugent.be](http://www.ugain.ugent.be)

**MEER INFO EN INSCHRIJVEN**

**[WWW.UGAIN.UGENT.BE/BIGDATAHANDSON](http://WWW.UGAIN.UGENT.BE/BIGDATAHANDSON)**