POSTACADEMISCHE OPLEIDING

BIG DATA HANDS-ON

25 oktober 2021 - 20 december 2021

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 sessies 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 sessies, 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 (zie VEREISTE VAARDIGHEDEN). Er wordt gewerkt met eigen laptop. Deze moet krachtig genoeg zijn (minimum 8GB Ram) en deelnemers moeten administratierechten hebben voor het installeren van de nodige programma's.

Het aantal deelnemers is beperkt tot 28.

Na deelname ontvangt u een aanwezigheidsattest.

MEER INFO EN INSCHRIJVEN
WWW.UGAIN.UGENT.BE/BIGDATAHANDSON

VEREISTE VAARDIGHEDEN

Introductie tot NLP met Python
- Basisvaardigheden in programmeren in Python - lijsten, woordenboeken, klassen en objecten, ... - worden hier niet expliciet aangeleerd. Deze moet u reeds onder de knie hebben.

Het MapReduce-programmeermodel
- Basiskennis van Linux commando's voor bestandsmanipulatie (cd / mv / rm / mkdir / rmdir / ls / cat / grep / head / tail / wc / etc).

De Kappa-architectuur
- Basiskennis van programmeren in Python is ten zeerste aangeraden. In principe zal het programmeerwerk zeer beperkt zijn (aanpassen scripts, lopen scripts, etc.).
- Kennis van een programmeertaal of een scripttaal, zoals bijvoorbeeld Java, Scala, Python of Javascript, is vereist.

WETENSCHAPPELIJKE COÖRDINATIE

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

LESGEVERS

- **Toon Boeckling**, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Sander Borny,** Vakgroep Informatietechnologie, Universiteit Gent
- **Cedric De Boom,** Vakgroep Informatietechnologie, Universiteit Gent
- **Dries Decap,** Vakgroep Informatietechnologie, Universiteit Gent
- **Thomas Demeester,** Vakgroep Informatietechnologie, Universiteit Gent
- **Jan Fostier,** Vakgroep Informatietechnologie, Universiteit Gent
- **Joachim Peeters,** Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Merlijn Sebrechts,** Vakgroep Informatietechnologie, Universiteit Gent
- **Lucas Sterckx,** Vakgroep Informatietechnologie, Universiteit Gent
- **Yoram Timmerman,** Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- **Bruno Volckaert,** Vakgroep Informatietechnologie, Universiteit Gent

PROGRAMMA

MODULE 1: GEGEVENSBEHEER

Introductie tot NLP met Python

Tijdens deze lesavond maken we kennis met het domein van Natural Language Processing (NLP) en brengen we de basisconcepten en basisblokken aan die onontbeerlijk zijn in datatoepassingen die tekstuele gegevens verwerken. We zullen de hele keten doorlopen, vertrekkende van het inlezen van datasets, het opschonen van de ingelezen data, tot het berekenen van features en het implementeren van een eenvoudige AI-toepassing. Hiervoor maken we gebruik van de programmeertaal Python.
Let op! Basisvaardigheden in programmeren in Python - lijsten, woordenboeken, klassen en objecten, ... - worden hier niet expliciet aangeleerd. Deze moet u reeds onder de knie hebben.

NoSQL (Volume, Velocity)

Tijdens deze lesavond leert u hoe u grote datavolumes en snelle data-invoerstromen kunt beheren 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 hoofdkarakteristieken 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

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 vervolgsessies 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.

MODULE 2: GEGEVENSANALYSE

Het MapReduce-programmeermodel

In 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 moeilijkere 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.

Vaardigheden: 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 zeerste aangeraden. In principe zal het programmeerwerk zeer beperkt zijn (aanpassen scripts, lopen scripts, etc.).

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 ArrangoDB. Tenslotte leren we u hoe uw bevindingen kunt presenteren via Notebooks.

Vaardigheden: 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.

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 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
- 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 authorisatiecode 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.**
- De lessen vinden plaats op maandagavond.
- Data onder voorbehoud van wijzigingen om onvoorziene omstandigheden.

Laptop

Er wordt gewerkt met eigen laptop. Deze moet krachtig genoeg zijn (minimum 8GB Ram) en deelnemers moeten administratierechten hebben voor het installeren van de nodige programma's. Zo moet uw laptop een recente Linux distributie en root access (voor Docker installatie) hebben. Men krijgt de nodige instructies op voorhand zodat men alles reeds thuis kan installeren en proberen.

1. GEGEVENSBEHEER

25 oktober 2021	**Introductie tot NLP met Python** Cedric De Boom, Thomas Demeester en Lucas Sterckx
8 november 2021	**NoSQL (Volume, Velocity)** Toon Boeckling, Joachim Peeters en Yoram Timmerman
15 november 2021	**Datakwaliteit en data-integratie (Variety, Veracity)** Toon Boeckling, Joachim Peeters en Yoram Timmerman
22 november 2021	**Document classificatie** Cedric De Boom, Thomas Demeester en Lucas Sterckx

2. GEGEVENSANALYSE

29 november en 6 december 2021	**Het MapReduce-programmeermodel** Dries Decap en Jan Fostier
13 december 2021	**De Kappa-architectuur** Sander Borny, Merlijn Sebrechts, en Bruno Volckaert
20 december 2021	**Deep learning** Cedric De Boom, Thomas Demeester en Lucas Sterckx

Organisatie

Universiteit Gent
UGain (UGent Academie voor Ingenieurs)
Technologiepark 60
9052 Zwijnaarde
09 264 55 82
ugain@ugent.be - www.ugain.ugent.be

MEER INFO EN INSCHRIJVEN

WWW.UGAIN.UGENT.BE/BIGDATAHANDSON

DIENSTVERLENER VOOR DE
KMO-PORTEFEUILLE

UNIVERSITEIT GENT

FACULTEIT INGENIEURSWETENSCHAPPEN EN ARCHITECTUUR

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN