PGO Business Analytics & Data Science
Vrije Universiteit Amsterdam, www.acba.nl

Course Manual
Big Data Infrastructures and Technologies

Docent(en)
Prof. Dr. Peter Boncz – Centrum Wiskunde en Informatica en VU
Dr. Hannes Mühleisen – Centrum Wiskunde en Informatica
- de docenten zijn te bereiken voor de cursus via boncz.muehleisen@gmail.com

Communicatie
Communicatie met de docenten gaat via e-mail. De benodigde informatie is te vinden op de pagina: homepages.cwi.nl/~boncz/bads.

Literatuur
Er wordt niet gewerkt met een vast boek; de leerstof is te vinden op bovengenoemde website, die per onderwerp een uitgebreide pagina heeft. Hierop staan de volgende materialen:

- de **slides** – deze stof moet doorgekeken en begrepen worden.
- de **samenvattingen** – deze stof (de tekst onder de kop 'Lecture') moet beheerst worden. Alles wat hier staat kan voorkomen in de test.
- **achtergrond** materiaal – deze papers beschikbaar op het internet over onderwerpen relevant voor de cursus zijn geselecteerd en goed bevonden door de docenten. Het wordt aangeraden om deze op zijn minst door te kijken, of natuurlijk aandachtig te bestuderen als het onderwerp de interesse wekt.
- **extra** materiaal – zie vorig punt. Dit zijn vaak YouTube-videos en slide-decks van anderen.
- **technische literatuur** – dit zijn vaak de wetenschappelijke artikelen die aan de beschreven technologieën ten grondslag liggen. Veelal zullen deze artikelen te technisch zijn voor dit publiek – dit is ter referentie.

Benodigde software
Deelnemers dienen zelf een laptop mee te nemen met werkende internetverbinding en installeerrechten. Het meeste werk gaat via de browser.
Er wordt gewerkt met cloud computing en daarvoor moeten de deelnemers een Amazon Web Services (AWS) account aanmaken op aws.amazon.com/free. Daarvoor is een creditcard vereist. Er zullen AWS vouchers verstrekt worden aan de deelnemers, die in principe ruim genoeg zijn om het practicum te doen. Let op: als het krediet op de vouchers op is, zal de eigen creditcard worden aangesproken, en dit is in principe de eigen verantwoordelijkheid. Houd dus goed het verbruik in de gaten ("niet het licht laten branden als u de kamer verlaat"). Indien er een tekort aan credits dreigt op te treden dient u de docenten aan te spreken, mogelijk kunnen er extra credits verstrekt worden.

Leerdoelen
Deelnemers verkrijgen begrip van de huidige mogelijkheden van openbare cloud computing en software die daarop beschikbaar is voor het analyseren van Big Data, alsmede ervaring in het werken met dit soort systemen.

Vereiste voorkennis

Basiskennis (zowel praktisch als theoretisch) van de informatica is wenselijk. Kunnen programmeren is een plus, alhoewel het niet zoveel uitmaakt in welke programmeertaal men ervaring heeft. Enige ervaring met het werken met computers zonder grafische gebruikersomgeving (GUI) is een plus, hoewel het meeste practicumwerk vanuit de browser gedaan kan worden.

Opzet

Praktische kennis wordt verkregen door de helft van de lestijd beschikbaar te houden voor practicum, en door het uitvoeren van beoordeelde practicumopdrachten buiten de les. De theoretische kennis wordt aangeboden in het hoorcollege en via het achtergrondmateriaal, en richt zich op het begrijpen van zowel de hardware-eigenschappen van deze clouds alsmede de belangrijkste eigenschappen en achterliggende ontwerpprincipes van de gebruikte software.

Wijze van toetsing

Het huiswerk bestaat uit wekelijkse opgaven waarmee gestart wordt tijdens het college en die voor de volgende bijeenkomst ingeleverd moeten zijn in Canvas. Deze vijf opgaven worden beoordeeld met een cijfer. Er is in de zesde bijeenkomst een korte theoretische eindtoets op hoofdlijnen gebaseerd op het materiaal in de slides en de samenvattingen (niet het extra materiaal). Deze toets zal nabesproken worden op de laatste bijeenkomst. De eindbeoordeling wordt bepaald uit de toets-score en het practicum.

Overzicht bijeenkomsten

Datum	Onderwerp
7 sept	*College:* Introduction Cloud Computing *Practicum:* Amazon Web Services 101
14 sept	*College:* MapReduce and Hadoop *Practicum:* Large-scale text analysis with MR
21 sept	*College:* The Spark Framework *Practicum:* Wikipedia topic analysis with Spark
28 sept	*College:* SQL on Big Data *Practicum:* Comparing Athena and Redshift on a TPC-H warehouse
5 okt	*College:* Scalable Machine Learning *Practicum:* Machine learning with Spark MLIB
12 okt	*Tentamen*
19 okt	*College:* Stream Processing *Nabespreking tentamen & evaluatie module*