Course: Big Data Infrastructure and Technologies for Big Data Analytics (BDIT4DA)

Lectures and practice with the Hadoop tools

Lecturer(s):
Yuri Demchenko, email: y.demchenko@uva.nl

Objectives:

This course provides students with a clear understanding of the Big Data Infrastructure technologies and existing cloud based platforms and tools for Big Data processing and data analytics. This will be supported by practical knowledge of the Cloudera Hadoop Cluster and its component applications. The course will provide basis for further self-study and practical use of Big Data technologies and competent evaluation and implementation of practical projects in the students' organisations.

Upon completing this course, students will be able to:

- Outline the basic concepts of Big Data and related technologies, and apply them to analyse general use cases and those related to their organisations
- Compare and select the Big Data Infrastructure services from the major Cloud Service Providers (such as AWS Elastic Map Reduce, Azure HDInsight, Azure Data Lakes, others) to use them for enterprise data management and analysis
- Describe main properties of the SQL and NoSQL databases, select appropriate database type depending on used data and analysis
- Outline the major components and processes of the enterprise Data Governance Architecture and corresponding organisational roles; develop the company's Data Management Plan (DMP) and corresponding implementation plan.
- Become acquainted with the functionality and programming models of the main Hadoop ecosystem components MapReduce, Spark, HBase, Hive, Pig, Kafka, others; program simple tasks using scripting or programming languages such as Hive SQL, Pig Latin, Java.
- Outline the main security and privacy challenges in using Big Data technologies; apply industry best practices and existing applications to protect companies' data and customers personal data.

Contents:

This course provides comprehensive overview and introduction into Big Data infrastructure technologies and tools. It establishes working knowledge of the concepts, techniques and products associated with the Big Data infrastructure and corresponding cloud based services. The focus is given on the cloud based Big Data infrastructure and analytics solutions and how cloud based services can be integrated into company's IT and data infrastructure.

Students will learn the core functionality of the major Big Data Infrastructure components and how they integrate to form a coherent solution with business benefit. Hands-on exercises aim to provide insight into how the cloud based services and tools can simplify processing of Big Data by using cloud based services for Hadoop, Machine Learning and general data analytics. Specific attention will be given to understanding and using the major Big Data platform Apache Hadoop ecosystem, its main functional components MapReduce, Spark, HBase, Hive, Pig, and supported programming languages Pig Latin and Hive.

The course describes industry best practices and models for enterprise data architectures to ensure effective data management and governance. The course also provides sufficient insight into Big Data security and compliance issues including those that are related to EU General Data Protection Regulation (GDPR).

Format:

8 sessions of 3 hours including lectures, discussions, and hands-on exercises, practical assignments, and the group project.

Study materials:

Study material will be made available on the selected learning platform in a form of Lecture slides, lecture notes, also set of recommended articles.

Schedule:

Lecture/Session 1	**Lecture 1 Introduction into course. Cloud Computing foundation.** Cloud service models, cloud resources, cloud services operation, multitenancy. Virtual cloud datacenter and outsourcing enterprise IT infrastructure to cloud. Cloud use cases and scenarios for enterprise. Cloud economics and pricing model.
Read	Lecture Notes/slides for session 1
Practice	Getting started with Amazon Web Services cloud; cloud services overview EC2, S3, VM instance deployment and access.
Assignment:	No assignment for session 1
Lecture/Session 2	**Lecture 2 Big Data architecture framework, cloud based Big Data services** Big Data Architecture and services. Overview major cloud based Big Data platform: AWS, Microsoft Azure, Google Cloud Platform (GCP). MapReduce scalable computation model. Overview Hadoop ecosystem and components.
Read	Lecture Notes/slides for session 2
Practice	Continue with Amazon Web Services cloud; VM instance deployment and access.
Assignment 1	Wordcount example using MapReduce algorithm (manually or with Java MapReduce library)
Lecture/Session 3	**Lecture 3 Hadoop platform components for Big Data analytics** Hadoop ecosystem components: HDFS, HBase, MapReduce, Kafka, YARN, Pig, Hive, others.
Read	Slides and recommended links for session 3
Practice	Work with Cloudera Hadoop cluster: Hue interface, uploading, downloading data.
Lecture/Session 4	**Lecture 4 SQL and NoSQL Databases** SQL basics and popular RDBMS. Overview NoSQL databases types. Column based databases and their use (e.g. HBase). Modern large scale databases AWS Aurora, Azure CosmosDB, Google Spanner.
Read	Slides and recommended links for session 4
Practice	Running simple SQL scripts. Using Pig scripting language for programming Big Data workflows.
Assignment 2	Processing Big Data of the educational database using Pig scripting language
Lecture/Session 5	**Lecture 5 Data Streams and Streaming Analytics**

	Data streams and stream analytics. Spark architecture and components. Popular Spark platforms, DataBricks. Spark programming and tools, SparkML library for Machine Learning.
Read	Slides and recommended links for session 5
Practice	Run simple Hive script for data processing
Lecture/Session 6	**Lecture 6 Data Management and Governance.** Enterprise Big Data Architecture and large scale data management. Data Governance and Data Management. FAIR Principles in data management.
Read	Slides and recommended links for session 6
Practice	Continue with Hive for data processing
Assignment 3	Processing Big Data sets using Hive scripting language (HQL)
Lecture/Session 7	**Lecture 7 Big Data Security and Compliance.** Big Data Security challenges, Data protection. cloud security models. Cloud compliance standards and cloud provider services assessment.
Read	Lecture Notes and slides for session 7
Practice	Cloud compliance practicum
Lecture/Session 8	**Course reflection and project progress presentations** Discussion and feedback. Project progress presentation by groups.

Assignment and assessment

This course doesn’t include formal exam.

The students are required to submit 3 assignments by the end of course and 1 practice report on the AWS cloud work experience.

Assignment 1 is a manual use of the MapReduce algorithm for wordcount. Assignments 2 and 3 working with Hadoop Hive and Pig will be assisted with the course teaching staff.

Practice report should provide the results of the practical work with AWS cloud following AWS practice guidelines (including steps, screenshots and summary)

Grading scheme to be defined based on the master programme requirements.

Remarks:

Current Syllabus and course content must be adjusted to the enclosing Master programme and available practice/lab base (in particular available cloud platform and services).

Recommended Prior Knowledge:

Familiarity with at least one of programming language, scripting language such as SQL, however detailed guidelines will be provided.

Software Installations:

Modern web browser: Chrome, Edge, FireFox, Safari, Opera
SSH client
VirtualBox or VMware desk virtualisation platform and Cloudera Quickstart Cluster (optional)

Remote Computing:

This course will require Internet access to services of some of the Cloud Service Providers and related Big Data platforms and to some of the major cloud service providers: Practice is planned with Amazon Web Services (AWS) educational class and account.