Course: Big Data Infrastructure and Technologies for Data Analytics

Lectures and practice with the Hadoop tools

Instructor:
Yuri Demchenko, email: y.demchenko at uva.nl

Objectives:

This course provides students with understanding of the Big Data Infrastructure technologies and existing cloud based platforms and tools for Big Data handling and Data Analytics. The course includes lectures and practical work with the Hadoop tools using cloud Hadoop cluster installation and/or quickstart installation on own laptop/notebook. The course will provide basis for further self-study and practical use of Hadoop tools and other Big Data tools.

The following main topics will be covered:

- Cloud services architecture, use cases and scenarios. Example cloud services by major Cloud Service Providers (CSP): Amazon Web Services (AWS), Microsoft Azure, Open Source cloud management platforms.
- Big Data architecture framework and cloud based Big Data services. Big Data services from the major cloud providers: AWS, Azure, Google Cloud Platform (GCP).
- Hadoop as a platform for Big Data storage, processing and Data Analytics. Major Hadoop components: HDFS, MapReduce, YARN, Tez, HBase, Hive, Pig, Hue, Spark, Kafka, Solr.
- NoSQL databases: properties, characteristics, types (HBase, Cassandra, MongoDB, Accumulo, others). Modern SQL databases: Amazon Aurora, Google Spanner, Azure CosmosDB. CAP theorem for distributed systems and SQL/NoSQL databases.

Acknowledgement:
Cloudera Hadoop Cluster cloud installation is provided by the Institute of Theoretical Physics of the Academy of Science of Ukraine (http://horst-7.bitp.kiev.ua, http://bitp.kiev.ua/en/welcome)

Programme:

Day 1 9:00 – 11:00	**Lecture 1 Introduction into course. Cloud Computing foundation** Cloud Computing architecture and service models, cloud resources, cloud services operation, multitenancy. Cloud use cases and scenarios, cloud economics and pricing model.
11:30 – 13:30	**Lecture 2** Big Data architecture framework and cloud based Big Data services. MapReduce and Hadoop overview. Overview major cloud

	based Big Data platform: AWS, Microsoft Azure, Google Cloud Platform (GCP)
13:30 – 14:00	**Discussion**
Practice 15:00 – 19:00	**Practice 1: Getting started with AWS cloud:** Cloud account creation, installation user side tools, accessing cloud services, S3, EC2, Lambda cloud services deployment, users and groups management with AIM.
	Getting started with Cloudera Hadoop cluster, Cloudera Manager and components access and navigation. Configuring access to the cloud based Cloudera Hadoop cluster.
Day 2 9:00 – 11:00	**Lecture 3 Detailed look into Hadoop ecosystem** Hadoop as a platform for Big Data storage, processing and Data Analytics. Major Hadoop components and programming models: HDFS, MapReduce, HBase, Hive, Pig, Solr, Spark, Kafka, Hue, others.
11:00 – 13:00	**Lecture 4 NoSQL databases for Big Data** NoSQL databases: properties, characteristics, types (HBase, Cassandra, MongoDB, Accumulo, CosmosDB, others). Modern SQL databases: Amazon Aurora, Google Spanner, Azure CosmosDB. CAP theorem for distributed systems and SQL/NoSQL databases.
13:30 - 14:00	**Discussion**
Practice 15:00 – 19:00	Practice 2: Working with Hadoop ecosystem tools. Hadoop Distributed File System (HDFS) - Basic commands and data transfer. Running simple MapReduce tasks
	Working with HBase, example handling large datasets
	Programing Hive – Hadoop SQL database.
	Using Pig Latin scripting language for programming simple workflows
	Dynamic search dashboards with Solr, Hue.