The 2019 High Performance Computing & Simulation International Conference
(HPCS 2019)
Dublin, Ireland
July 15 - 19, 2019

TUTORIAL

Cloud Based Big Data Infrastructure and Tools for Data Analytics and Data Management

Yuri Demchenko
University of Amsterdam

TUTORIAL DESCRIPTION

This tutorial will provide comprehensive overview and introduction into the Big Data Infrastructure technologies and existing cloud based platforms and tools for Big Data processing and data analytics. 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. The tutorial will also overview industry best practices and models for enterprise data architectures to ensure effective data management and governance.

The listeners will learn the core functionality of the major Big Data Infrastructure components and how they integrate to form a coherent solution with business benefit. 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 HiveQL.

This will be supported by demonstrating the services and tools available from the major cloud providers Amazon Web Services (AWS) and Microsoft Azure, as well as the Cloudera Hadoop Cluster deployed on cloud (can be also installed as a starter edition on the laptop) and its component applications.

The course intends to provide a basis for further self-study and practical use of the Big Data technologies and competent evaluation and implementation of practical projects in the listeners' organisations. Links to online hands-on exercises will be provided.

Tutorial OUTLINE

The tutorial will cover the following topics:

- Overview the basic concepts of Big Data and related technologies, and their application to data analysis and organisational needs
- Overview the Big Data Infrastructure services from the major Cloud Service Providers (such as AWS Elastic Map Reduce, Azure HDInsight, Azure Data Lakes, others) and their use for enterprise data management and analysis
- SQL and NoSQL databases, their properties and cloud based implementation
- Functionality and programming models of the main Hadoop ecosystem components MapReduce, Spark, HBase, Hive, Pig, Kafka, others; examples programming simple tasks using scripting or programming languages such as Hive SQL, Pig Latin, Java.
- Enterprise Data Governance Architecture and corresponding organisational roles; Data Management Plan (DMP) and Data Quality assessment framework.

REFERENCES

Data Science Competence Framework (CF-DS), EDISON Data Science Framework (EDSF), EDISON Community Initiative [online]
https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
https://github.com/EDISONcommunity/EDSF/
https://github.com/EDISONcommunity/EDSF/tree/master/data-science-competence-framework

(BD-RA, 2015) NIST Special Publication NIST SP 1500: NIST Big Data Interoperability Framework (NBDIF). Volume 6, Reference Architecture [online]
http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf

(NIST-BD-Req, 2015) NIST Special Publication NIST SP 1500: NIST Big Data Interoperability Framework (NBDIF). Volume 3, Use cases and General Requirements [online]
http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-3.pdf

REQUIREMENTS AND TARGET AUDIENCE

The attendees are advised to apply for Amazon Educate account (https://aws.amazon.com/education/awseducate/) and/or install Cloud Hadoop Starter edition on their laptops (https://www.cloudera.com/documentation/enterprise/5-7-x/topics/cloudera_quickstart_vm.html)
The expected target audience is wide but primarily oriented on Data Science practitioners and those who want to understand how to move their existing business applications to cloud
There is no specific knowledge of Data Science or Big Data required.

TUTORIAL DURATION

The tutorial material will be presented in one 2 hours session.

A/V AND EQUIPEMNT

Standard presentation facilities, no AV required.

INSTRUCTOR BIOGRAPHY AND PHOTO

Yuri Demchenko is a Senior Researcher at the System and Network Engineering of the University of Amsterdam. He is graduated from the National Technical University of Ukraine "Kiev Polytechnic Institute" where he also received his PhD (Cand. of Science) degree. His main research areas include Data Science and Data Management, Big Data and Infrastructure and Technologies for Data Analytics, DevOps and cloud based software development, general security architectures and distributed access control infrastructure for cloud based services and data centric applications. He is currently involved in the European projects GEANT4, MATES, FAIRsFAIR where he develops different elements of cloud based infrastructures for scientific research, and issues related to Data Science and digital skills development. Yuri has coordinated the EU funded EDISON project (2015-2017) which has developed the EDISON Data Science Framework (EDSF) that provides a conceptual foundation and practical basis for building the Data Science profession. His recent research are also extending into data economics and open data market models.
He is actively contributing to the standardisation activity at RDA, OGF, IETF, NIST, CEN on defining Big Data Architecture Framework, Data Science competences, and data properties as economic goods.