Data Science Model Curriculum Implementation for Various Types of Big Data Infrastructure Courses

Tomasz Wiktorski
University of Stavanger, Norway
tomasz.wiktorski@uis.no

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl

Oleg Chertov
National Technical University of Ukraine, Ukraine
oleg.r.chertov@gmail.com

***Abstract*— This paper presents experiences of development and teaching three different types of Big Data Infrastructure courses as a part of the general Data Science curricula. The authors built the discussed courses based on the EDISON Data Science Framework (EDSF), in particular, Data Science Body of Knowledge (DS-BoK) related to Data Science Engineering knowledge area group (KAG-DSENG). The paper provides overview of the sandboxes, Cloud-based platforms and tools for Big Data Analytics and stresses importance of including into curriculum the practical work with Clouds for future graduates or specialists workplace adaptability. The paper discusses a relationship between the DSENG BoK and Big Data technologies and platforms, in particular Hadoop-based applications and tools for data analytics that should be promoted through all course activities: lectures, practical activities and self-study.**

Keywords- Education, EDISON Data Science Framework (EDSF), Data Science Body of Knowledge (DS-BoK), Data Science Engineering, Big Data Infrastructure Technologies, Hadoop ecosystem, Cloud Computing

I. Introduction

Knowledge and ability to work with the modern Big Data platforms and tools to effectively develop and operate the data analytics applications is one of important competences required from the modern Data Science specialists, and need to be included into the general Data Science curriculum.

The EDISON Data Science Framework (EDSF), initially developed in the EDISON Project (2015-2017) and currently maintained by the EDISON community [1, 2], provides a general framework for Data Science education, curriculum design and competences management what has been discussed in the previous authors' works [3, 4, 5]. Big Data Infrastructure Technologies (BDIT) is a part of the defined in EDSF the Data Science Engineering Body of Knowledge (DSENG-BoK) and Model Curriculum (MC-DSENG) described in details below.

This paper provides comprehensive overview and introduction into the Big Data infrastructure technologies and existing cloud based platforms and tools for Big Data processing and data analytics that are relevant to the Big Data Infrastructure Technologies for Data Analytics (BDIT4DA) courses. The focus is given on the cloud based Big Data infrastructure and analytics solutions and how Cloud-based services can be integrated into the company's IT and data infrastructure. Specific attention should be given to understanding and using the Apache Hadoop ecosystem as the major Big Data platform, its main functional components MapReduce, Spark, HBase, Hive, Pig, and supported programming languages Pig Latin and HiveQL.

Knowledge and basic experience with the major cloud service providers (e.g., Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform GCP) as well as the Cloudera Hadoop Cluster or Hortonworks Data Platform are important to form strong practical knowledge and skills; they need to be included into both lecture course and hands on practice.

The paper is organised as follows. Section II introduces the EDISON Data Science Framework (EDSF), section III provides information about the Data Science Engineering Body if Knowledge (DSENG-BoK) and the DSENG Model Curriculum and its main components. Section IV describes the Hadoop ecosystem used as a main platform for the Big Data applications, including core components and other important applications, used programming and query languages. Section V provides example courses taught by the authors in different education environments and formats. Conclusion section VI describes ongoing developments and activities on exchange of best practices in Data Science curriculum development and ongoing education.

II. EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework (EDSF), that is the product of the EDISON Project, provides a basis for Data Science education and training, curriculum design and competences management that can be customised for specific organisational roles or individual needs [24]. EDSF can be also used for professional certification and to ensure career transferability.

The following are the main EDSF components that are specified in the corresponding separate documents [2]:

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge

978-1-7281-2451-3/19/$31.00 ©2019 IEEE
DOI 10.1109/eScience.2019.00074

- MC-DS – Data Science Model Curriculum
- DSPP - Data Science Professional profiles and occupations taxonomy
- Data Science Taxonomy and Scientific Disciplines Classification

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path.

The following core CF-DS competence and skills groups have been identified (refer to CF-DS specification for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science competences must be supported by knowledge that are defined primarily by education and training, and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills (refer to CF-DS specification for full definition of the identified knowledge and skills groups):

- Skills Type A, which are related to the professional experience and major competences, and
- Skills Type B,which are related to wide range of practical computational skills including using programming languages, development environment and cloud based platforms.

CF-DS defines workplace skills, also referred to as “soft” skills or professional attitude skills, which are becoming increasing important in modern data driven and future Industry 4.0 economy. This includes two groups of skills that are increasingly demanded by employers: Data Science Professional skills (Thinking and acting like Data Scientist), and so called the 21st Century skills that comprises a set of workplace skills that include critical thinking, communication, collaboration, organizational awareness, ethics, and others.

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [6], incorporates best practices in defining domain specific BoK’s and provides reference to existing related BoK’s. It also includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS) competences (specifically skills type A), and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. Practical curriculum should be supported by corresponding educational environment for hands on labs and educational projects development.

The formal DS-BoK and MC-DS definition creates a basis for Data Science educational and training programmes compatibility and consequently Data Science related competences and skills transferability.

III. Data Science Engineering BoK and Model Curriculum

A. *DSENG Model Curriculum Components*

Data Science Engineering Knowledge Group builds the ability to use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management. It includes Knowledge Areas that cover: software and infrastructure engineering, manipulating and analysing complex, high- volume, high-dimensionality data, structured and unstructured data, cloud based data storage and data management.

Data Science Engineering includes software development, infrastructure operations, and algorithms design with the goal to support Big Data and Data Science applications in and outside the cloud. The following are commonly defined Data Science Engineering Knowledge Areas (as part of KAG02-DSENG):

- KA02.01 (DSENG/BDI) Big Data infrastructure and technologies, including NOSQL databased, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approached to big data analysis and common MapReduce algorithms;
- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with focus on data warehouse and data centers.

The DS-BoK provides mapping of the DS-BoK to existing classifications and BoKs: ACM Computer Science BoK (CS-BoK) selected KAs [7], Software Engineering BoK (SWEBOK) [8], and related scientific subjects from CCS2012 [6]: Computer systems organization, Information systems, Software and its engineering.

B. *DSENG/BDIT - Big Data infrastructure and technologies course content*

Big Data infrastructures and technologies shape many of the Data Science applications. Systems and platforms behind Big Data differ significantly from traditional ones due to specific challenges of volume, velocity, and variety of data that need to be supported by data storage and transformation. Data Lakes and SQL/NoSQL databases must be included in the DSENG curriculum

Deployment of Data Science applications is usually tied to one of most common platforms, such as Hadoop or Spark, hosted either on private or public clouds. The applications workflow must be linked to a whole data processing pipeline including ingestion and storage for variety of data types and source. Data Scientists should have a general understanding of data and application security aspects in order to properly plan and execute data-driven processing in the organization. This module should provide an overview of the most important security aspects, including accountability, compliance and certification.

Data Management and Governance (DMG) [9, 10], although belonging to different KAG4-DSDM, must accompany the DSENG courses and short overview of the DMG common practices must be included into the BDIT curriculum. This should also include the introduction of the FAIR data principles (data must be Findable, Accessible, Interoperable, Reusable) [11] that is growingly adopted by the research community and recognised by industry.

Curriculum components have some minor overlaps with Cloud Computing curricula [23].

IV. ESSENTIAL HADOOP ECOSYSTEM COMPONENTS

Hadoop became the main platform for Big Data processing with multiple applications and rich functionality developed by the Apache Open Source Software community. Giving general vision and basic experience with the Hadoop applications and tools is a key part of the practical activity and assignments in the BDIT4DA course. Figure 1 below illustrates the Hadoop main components and other popular applications for data processing [12, 13].

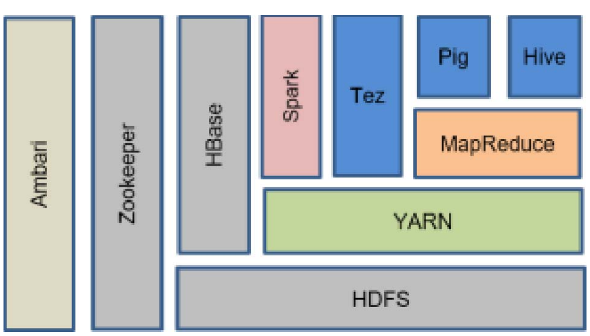

Figure 1. Main components of the Hadoop ecosystem

A. *Hadoop applications and projects*

The following main Hadoop applications constitute the foundation of the Hadoop ecosystem and provide basis for other applications

Apache Hadoop software stack includes the following main modules:

Hadoop Common: The common utilities that support the other Hadoop modules and includes utilities and drivers to support different computer cluster and language platforms.

HDFS: Hadoop Distributed File System optimized for large scale storage and processing of data on commodity hardware

MapReduce: A YARN-based system for parallel processing of large data sets.

YARN: A framework for job scheduling and cluster resource management.

Tez: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.

Other Hadoop-related projects at Apache that provide rich set of functionalities for data processing during the whole data lifecycle:

Hive: A data warehouse system that provides data aggregation and querying.

Pig: A high-level data-flow language and execution framework for parallel computation.

HBase: A distributed column oriented database that supports structured data storage for large tables

ZooKeeper: A scalable coordination service for distributed applications.

Spark: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

Kafka: Fast, scalable, durable, and fault-tolerant publish-subscribe messaging system

Mahout: A scalable machine learning and data mining library.

Sqoop: command-line interface application for transferring data between relational databases and Hadoop file system

Solr: Open source enterprise search platform that uses lucene as indexing and search engine.

Oozie: Server-based workflow scheduling system to manage Hadoop jobs.

Avro: A data serialization system that supports rich data structures

Ambari: A web-based tool for provisioning, managing, and monitoring YARN jobs and Apache Hadoop clusters

Hue: A user graphical interface providing full functionality for programming Hadoop applications, including dashboard, data upload/download, visualisation.

Few other recently added Hadoop based application and packages that should be considered for BDIT4DA courses:

Flume: Distributed, reliable, and available software for efficiently collecting, aggregating, and moving large amounts of log data.

Flink: Distributed streaming data-flow engine written in Java and Scala that executes arbitrary dataflow programs in a data-parallel and pipelined manner.

B. Programming Hadoop

Introducing multiple Hadoop programming options is essential to allow future integration of the Hadoop platform and tools into research and business applications. Hadoop is natively programmed in Java, with current support for Scala by many applications. There is also support for Hadoop API calls from many popular programming and data analytics IDE and tools for R, Python, C, .NET. Specific for Hadoop are query languages to work with HBase, Hive, Pig as shown in Figure 2.

While Hadoop, and to some extent Spark, are Java based framework, Python becomes a *de facto* main language for Data Scientists. Both Hadoop and Spark provide good support for programming in Python. In Hadoop that can be achieved through Hadoop Streaming library and third-party libraries such as MRJob. Streaming is available natively, but is rather inconvenient to use for more complex jobs. MRJob provides a an easy-to-use overlay over Streaming library. Not only does it make creating complex jobs easy, it also has native support for Amazon Elastic MapReduce service and in-line testing without a Hadoop cluster.

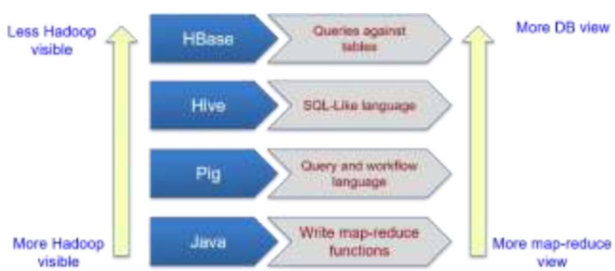

Figure 2. Query languages for Hadoop

Hive Query Language (HiveQL or HQL) [14]: Provides higher-level data processing language, used for Data Warehousing applications in Hadoop. Query language is HiveQL, variant of SQL, tables are stored on HDFS as flat files. HiveQL facilitates large-data processing that compiles down to Hadoop jobs.

Pig Latin [15] is a scripting language used for large-scale data processing system to describe a data processing flow. In fact, Pig Latin has similarity to HiveQL query commands with additional flow control commands. Similar to HiveQL, it compiles down to Hadoop jobs and relies on MapReduce or Tez for execution.

C. Functional Abstraction

While students become familiar with new technologies and tools, the course should emphasize the underlying principles. big bata processing both in Hadoop and Spark, relies on MapReduce type algorithms. MapReduce is not specific to big data. The concept originates from functional programming and become adapted to big data first by Google [21].

The algorithms development process gravitates with time from low-level MapReduce programming to higher-level tools such as Hive or SparkSQL, but the underlying data processing still relies on functional principles in MapReduce.

Some newer textbook, e.g. Data-intensive System [ref to book] introduce functional abstraction for big data processing as an independent concept that aids in understanding further algorithms design choices and performance optimization strategies.

V. EXAMPLE BDIT4DA COURSES

This section provides example of three Big Data Infrastructure and Technologies for Data Analytics courses that can be adjusted to different academic or training programmes. BDIT4DA includes lectures, practice/hands on labs, projects and such engaging activities as literature study and seminars. The course should beneficially include a few guest lectures, to expose the students to external experts and real practices.

A. BDIT4DA Course for Big Data Engineering Masters

1) Lectures

Lectures must provide a foundation for understanding the whole BDIT4DA technology domain, available platforms, tools and link other course activities. However, form and technical level must be adjusted to the incumbent programme, for example distinguishing Computer Science and MBA programs. The same should be related to the selection of practical assignments and used tools and programming environment.

The following example is the set of lectures has been developed and taught by the authors' (presented in a form of sessions that actually can combine lectures, practice, interactive activities):

Lecture 1 Cloud Computing foundation and economics.
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.

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.

Lecture 3 Hadoop platform for Big Data analytics
Hadoop ecosystem components: HDFS, HBase, MapReduce, YARN, Pig, Hive, Kafka, others.

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.

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.

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.

Lecture 7 Big Data Security and Compliance.

Big Data Security challenges, Data protection. cloud security models. Cloud compliance standards and cloud provider services assessment. CSA Consensus Assessment Initiative Questionnaire (CAIQ) and PCI DSS cloud security compliance.

2) Practice and project development

Recommended practice includes working with the main Hadoop applications and programming simple data processing tasks. Different Hadoop platforms can be used for running practical assignments using either dedicated Hadoop cluster installations (e.g. Cloudera Hadoop Cluster [16], Hortonworks Data Platform [17], or cloud based AWS Elastic MapReduce (EMR), or Azure HDInsight platform). Students can be also recommended to install personal single host Hadoop cluster using either Cloudera Starter edition or Hortonworks Sandbox that are available for both VirtualBox and for VMware.

The following are example topics for practice and hands on assignments.

Practice 1: Getting started with the selected cloud platform. For example, Amazon Web Services cloud; cloud services overview EC2, S3, VM instance deployment and access.

Practice 2: Understanding MapReduce, Pregel, other massive data processing algorithms. Wordcount example using MapReduce algorithm (run manually and with Java MapReduce library).

Practice 3. Getting started with the selected Hadoop platform. Command line and visual graphical interface (e.g. Hue), uploading, downloading data. Running simple Java MapReduce tasks.

Practice 4. Working with Pig: using simple Pig Latin scripts and tasks. Develop Pig script for programming Big Data workflows. This can be also done as a part of practical assignment on Pig.

Practice 5. Working with Hive: Run simple Hive script for querying Hive data base. Import external SQL database into Hive. Develop Hive script for processing large datasets. This can be also a part of practical assignment on Hive.

Practice 6: Streaming data processing with Spark, Kafka, Storm. Using simple task to program Spark jobs and using Kafka message processing The option for this practice can also use Databricks platforms that provides a good tutorial website.

Practice 7: Creating dashboard and data visualisation. Using tools available from the selected Hadoop platform to visualise data, in particular using results from Practice 5 or 6 that is dealing with large datasets where dashboard is necessary

Practice 8. Cloud compliance practicum. This practice is important for the students to understand the complex compliance issues for applications run on cloud. Using Consensus Assessment Initiative Questionnaire (CAIQ) tools.

B. Course for Data Science Masters

In contrast to the Big Data Engineer example, a course for Data Scientists spends more time on algorithm design aspect. All basic tool and concepts are introduced, but less time is spent on topics related to security and governance.

The course follows a Problem-based Learning approach [22]. First five lectures have corresponding laboratory sessions. Afterwords, students begin on working on group projects on datasets of their choice, applying concepts, technologies, and tools from lectures. Progress in projects is presented at plenary presentations sessions, in the middle and at the end of the course.

To further motivate and guide students 1 or 2 guest lectures with practitioners from industry are organized. The can be scheduled any time after Lecture 5. In some case, it might also be scheduled together with Lecture 1.

The project work goes through several stages, starting from developing a scenario, identifying relevant data sources, data acquisition and preprocessing [25], and finally performance optimization [26].

1) Lectures

Lecture 1: Introduction to data intensive systems and use cases. Data as 4th paradigm of science. Increasing focus on data collection, data architectures, data centers. Use cases in search, commerce, healthcare, energy.

Lecture 2: Hadoop 101 and Functional abstraction. Introductory, but fully functioning MapReduce program in Python with execution from command line.

Lecture 3: MapReduce. Detailed description of file splitting, mapping, combining, shuffling, reducing, and storage of results.

Lecture 4: Hadoop Architecture. Resource management, permanent and temporary storage, batch processing, real-time processing, higher-level tools.

Lecture 5: MapReduce algorithms and patterns. Counting, summing, and averaging. Processing multiline input. Random sampling. Search Assist. Inverted index.

Lecture 6: HBase and other NoSQL databases. Alternative permanent storage for big data. CAP/PACELC theorems. Interaction between Hadoop/MapReduce and NOSQL databases.

Lecture 7: First project presentation. Focused on choice of dataset, data preprocessing, identification of interesting problems.

Lecture 8: Spark (RDD based). Data model. Programming model, actions, transformations, other operations. Architecture.

Lecture 9: Spark (SQL/other structures/MLlib). Alternative programming models, advantages and drawbacks. Incorporating existing libraries in the programming workflow.

Bonus: 1-2 industrial guest lectures. Usually focusing on data quality and data workflow in industry.

Lecture 10: Final project presentations. Focused on MapReduce implementation of identified problems. Performance tuning.

2) Practice and project development

Lab 1: refresh Bash knowledge, setup Docker and Hortonworks Sandbox. Ensures that students have a working test environment on their laptops.

Lab 2: recreate steps from lecture (system setup, file coping, running ready MRJob and Hive examples). Ensures that students can correctly execute examples in the book/lecture.

Lab 3: introduce modifications to MRJob based program on the Sandbox. Ensures that students understood basic concepts related to MapReduce programming.

Lab 4: setup Hadoop from scratch on a VM (not Sandbox). Ensures that students understood Hadoop architecture.

Lab 5: in-depth analysis of typical algorithms and patterns in groups. Ensures that students understood details of MapReduce programming

After five laboratory sessions students work on **group projects**. They are still encouraged to come on a regular basis to laboratory sessions where the can discuss and get support with any technical problems they meet.

C. Big Data Infrastructure Technologies (BDIT) Course for MBA in Big Data

Big Data and Data Analytics tools is important part of the business supporting infrastructure and services which are growingly cloud based. The specifics of the MBA Data Science groups is the diverse background of the students from the economics and business to Computers Science and engineering. The main goal of the BDIT course is to provide knowledge sufficient for the future business managers to make assessment and advice development of necessary services in their future organisations. The practical work is entirely based on using cloud based applications and tools. The course includes also project where the students working in groups need to deliver the design of the cloud based Big Data infrastructure supporting the business processes of their hypothetic company.

1) Lectures

BDIT lectures include subset of topics outlined in the section V.A but enriched with examples and closely linked to practices and labs.

Lecture 1 Cloud Computing foundation and cloud economics: Provides basic for understanding and working with clouds.
Lecture 2 Big Data architecture framework, cloud based Big Data services: Overview of cloud based Big Data platforms and tools, including AWS, Azure and Google Cloud Platform.
Lecture 3 MapReduce and Hadoop platform: Introduce the Hadoop ecosystem and main components; example of use
Lecture 4 Spark and Streaming Analytics: Including data structure, programming with Scala.
Lecture 5 SQL and NoSQL Databases: Database classification and types, Cloud based big data bases, Hadoop based HBase, Hive
Lecture 6 Data Management and Governance: based on DAMA DMBOK, extended with FAIR and QA methods
Lecture 7 Big Data Security and Compliance: Cloud data security services, access control, CSA Compliance framework.

2) Practice and Project Development

Practice covers major aspects of working with two main cloud platforms AWS and Microsoft Azure, starting with AWS as presenting more generic cloud services model, and following with Microsoft Azure as providing better aligning with Business Analytics processes. The following topics were included in the course:
Practice 1: Getting started with Amazon Web Services cloud
Practice 2: AWS services EC2, S3 deployment and access.
Practice 3: Amazon Elastic MapReduce (EMR). Running MapReduce wordcount example manually and using EMR.
Practice 4. AWS Aurora scalable SQL database, deployment and simple queries exercises.
Practice 4. Getting started with Microsoft Azure cloud, Storage and Compute services, instances deployment
Practice 5. Azure HDInsight business analytics platform, deployment and Hadoop cluster visual interface. Running simple Spark examples.
Practice 6: Cloud compliance practicum using CSA Consensus Assessment Initiative Questionnaire (CAIQ) tools.

VI. CONCLUSION

The presented in this paper the general approach and practical experience in teaching the Big Data Infrastructure Technologies for Data Analytics is based in the EDISON Data Science Framework, which is widely used by universities, professional training organisations and certification organisations, providing valuable feedback for further framework development and continuous courses evolution. The presented work is also based on long authors' experience in teaching cloud computing technologies [18] that are providing basis for Big Data technologies.

The academic education or professional training must provide strong basis for graduates and trainees to continue their further self-study and professional development in conditions of the fast developing technologies and agile business environment adopted by majority of modern companies. To achieve this, the Data Science curriculum needs to be supported by the professional skills development courses such as to develop the general 21st Century skills and specific Data Science workplace skills. One of general skills for data workers is considered the Research Data Management and Stewardship adopting FAIR data principles, which is part of the FAIRsFAIR project [19].

The EDSF maintenance and continuous development as well as collection of the best practices in Data Science education and training is supported and coordinated by the EDISON community, in cooperation with national and EU projects as well as supported by the Research Data Alliance (RDA) Interest Group on Education and Training on Handling Research Data (IG-ETHRD) [20]. Participation and contribution to both IG-ETHRD and EDSF is open and free.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Horizon2020 projects FAIRsFAIR (grant number 831558) and EDISON (grant n. 675419).

References

[1] EDISON Community wiki. [online] https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome

[2] EDISON Data Science Framework (EDSF). [online] Available at https://github.com/EDISONcommunity/EDSF

[3] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Daya and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kigndom, ISBN978-1-4503-6186-6/19/03

[4] Demchenko, Yuri, et all, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, Proc. The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 Dec 2016, Luxembourg.

[5] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.

[6] The 2012 ACM Computing Classification System [online] http://www.acm.org/about/class/class/2012

[7] ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] http://dx.doi.org/10.1145/2534860

[8] Software Engineering Body of Knowledge (SWEBOK) [online] https://www.computer.org/web/swebok/v3

[9] Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

[10] Data Maturity Model (DMM), CMMI Institute, 2018 [online] https://cmminstitute.com/data-management-maturity

[11] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] https://www.nature.com/articles/sdata201618

[12] Apache Hadoop [online] https://hadoop.apache.org/

[13] Hadoop Ecosystem and Their Components – A Complete Tutorial [online] https://data-flair.training/blogs/hadoop-ecosystem-components/

[14] Apache Hive Tutorial [online] https://cwiki.apache.org/confluence/display/Hive/Tutorial

[15] Apache Pig Tutorial [online] https://data-flair.training/blogs/hadoop-pig-tutorial/

[16] Cloudera Hadoop Cluster (CDH) [online] https://www.cloudera.com/documentation/other/reference-architecture.html

[17] Hortonworks Data Platform [online] https://hortonworks.com/products/data-platforms/hdp/

[18] Demchenko, Yuri, David Bernstein, Adam Belloum, Ana Oprescu, Tomasz W. Wlodarczyk, Cees de Laat, New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering. Proc. The 5th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2013), 2-5 December 2013, Bristol, UK.

[19] FAIRsFAIR Project [online] https://www.fairsfair.eu/

[20] Research Data Alliance (RDA) Education and Training on Handling of Research Data Interest Group (IG-ETHRD) [online] https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html

[21] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." In OSDP04: Proceedings of the 6th conference on Symposium on Opearting Systems Design, vol. 6, no. 5, pp. 10-10.

[22] T. W. Wlodarczyk and T. J. Hacker, "Problem-Based Learning Approach to a Course in Data Intensive Systems," 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, Singapore, 2014, pp. 942-948.

[23] Y. Demchenko, D. Bernstein, A. Belloum, A. Oprescu, T. W. Wlodarczyk and C. d. Laat, "New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, 2013, pp. 112-119.

[24] Y. Demchenko et al., "EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry," 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Luxembourg City, 2016, pp. 620-626.

[25] B. Jia, T. W. Wlodarczyk and C. Rong, "Performance Considerations of Data Acquisition in Hadoop System," 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, IN, 2010, pp. 545-549.

[26] T. W. Wlodarczyk, Y. Han and C. Rong, "Performance Analysis of Hadoop for Query Processing," 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications, Singapore, 2011, pp. 507-513.