Big Data Platforms and Tools for Data Analytics in the Data Science Engineering Curriculum

Yuri Demchenko
University of Amsterdam
Science Park 904, 1098XH Amsterdam
Tel: +31 20 525 7586
E-mail: y.demchenko@uva.nl

ABSTRACT

This paper presents experiences of development and teaching courses on Big Data Infrastructure Technologies for Data Analytics (BDIT4DA) as a part of the general Data Science curricula. The authors built the discussed course 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 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.

CCS Concepts

Computer systems organization -> Architectures -> Distributed architectures -> Cloud computing

Information systems -> Data management systems

Keywords

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

1. INTRODUCTION

Modern Data Science and Business Analytics applications extensively use Big Data infrastructure technologies and tools which are commonly cloud based and are available at all major cloud platform. Knowledge and ability to work with the modern Big Data platforms and tools to effectively develop and operate the data analytics applications is required from the modern Data Science practitioners. Including Big Data Infrastructure topics into the general Data Science curriculum will help the graduates to be easy integrated into the future workplace.

This paper refers to and effectively uses the EDISON Data Science Framework (EDSF), initially developed in the EDISON Project (2015-2017) and currently maintained by the EDISON community [1, 2]. The EDSF provides a general framework for the Data Science education, curriculum design and competences management what has been discussed in the previous author's 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 is focused on the definition of the Data Science Engineering Body of Knowledge and Big Data Infrastructure Technologies for Data Analytics (BDIT4DA) course. The paper provides brief overview of the Big Data infrastructure technologies and existing cloud based platforms and tools for Big Data processing and data analytics which are relevant to the BDIT4DA course. The focus is given on the cloud based Big Data infrastructure and analytics solutions and in particular 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 develop necessary knowledge and strong practical skills. These topics need to be included into both lecture course and hands on practice.

The paper is organised as follows. Section 2 introduces the EDISON Data Science Framework (EDSF), section 3 provides information about the Data Science Engineering Body if Knowledge (DSENG-BoK) and the DSENG Model Curriculum and its main components. Section 4 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; it also includes a brief overview of the Big Data platforms provided by the major cloud providers. Section 5 describes an example of the course taught by the author in different education environments and formats. Conclusion section 6 describes ongoing developments and activities on exchange of best practices in Data Science curriculum development and ongoing education.

2. 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. EDSF can be also used for professional certification and to ensure career transferability.

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

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- 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 both types of skills, those related to basic competences and professional experience and those based on wide range of practical skills including using programming languages, development environment and cloud based platforms.

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.

3. DATA SCIENCE ENGINEERING BOK AND MODEL CURRICULUM

3.1 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.

3.2 DSENG/BDIT - Big Data infrastructure 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 are growingly adopted by the research community and recognised by industry. Data Stewardship is a DMG application domain that combines general and subject domain data management ensuring the FAIR principles are incorporate into the organisational practice.

4. PLATFORMS FOR BIG DATA PROCESSING AND ANALYTICS

This section describes what platforms can be used for teaching the BDIT4DA course and other courses in the Data Science Engineering curricula requiring processing Big Data. The section describes the Hadoop Ecosystem and its main components and functionalities, and provides information about cloud based Big Data Infrastructure and analytics platforms from the major cloud providers.

4.1 Essential Hadoop Ecosystem Components

Hadoop is commonly used as a main platform for Big Data processing, it includes multiple components and applications developed by the Apache Open Source Software community, with rich functionality to support all processes and stages in the data processing workflow/pipeline. Giving general understanding and basic experience with the Hadoop applications and tools is an important part of the practical activity and assignments in the BDIT4DA course. Figure 1 below illustrates the Hadoop main components and few other popular applications for data processing [12, 13].

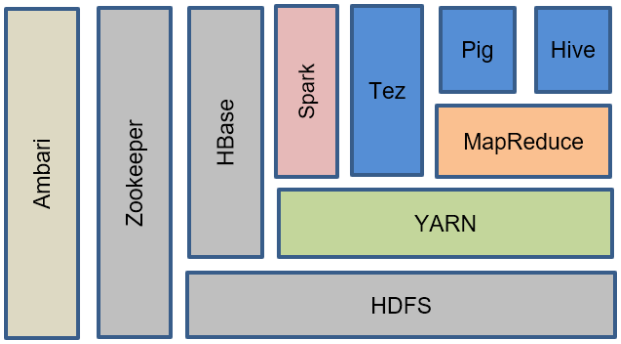

Figure 1. Main components of the Hadoop ecosystem

The following main Hadoop applications constitute the foundation of the Hadoop ecosystem and provide basis for other applications. 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
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.
Mahout: A scalable machine learning and data mining library.
Solr: Open source enterprise search platform that uses lucene as indexing and search engine.
Oozie: Server-based workflow scheduling system to manage Hadoop jobs.
Hue: A user graphical interface providing full functionality for programming Hadoop applications, including dashboard, data upload/download, visualisation.

4.2 Hadoop Programming Languages

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.

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.

4.3 Cloud based Big Data Platforms

Major cloud platforms Amazon Web Services (AWS) [16], Microsoft Azure [17], Google Cloud Platform (GCP) [18] provide rich set of the Big Data services and applications.

AWS Big Data stack includes such services as Elastic MapReduce (EMR) which is a hosted Hadoop platform for Data Analytics, Amazon Kinesis is a managed service for real-time processing of streaming big data, Amazon DynamoDB - highly scalable NoSQL data stores, Amazon Aurora - scalable relational database, and Amazon Redshift - fully-managed petabyte-scale data warehouse. Separately provided is the Machine Learning stack with a number of services. All services and tools are accessible from the AWS Console and can programmed via Command Line Interface (CLI), where the former provide all necessary functionality to program, deploy and operate complex business applications by integrating all necessary components into one data processing pipeline.

Microsoft Azure provides well integrated and supported by development tools the Big Data and Analytics stack that includes such services as HDInsight which is Hortonworks based Hadoop platform, Data Lake Storage and Data Lake Analytics, CosmosDB mufti-format NoSQL database, and other services.

Google Cloud provides general cloud services and a set easy configured Big Data services such as BigQuery column based NoSQL database, Google Spanner Big SQL database, and Machine Learning stack with well defined API that support the whole data analytics

5. EXAMPLE BDIT4DA COURSE STRUCTURE AND CONTENT

This section provides example of the reference Big Data Infrastructure and Technologies for Data Analytics course 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 few guest lectures, to expose the students to external experts and real practices.

5.1 BDIT4DA 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 Stewardship/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.

5.2 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 [19], Hortonworks Data Platform [20], 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.

6. 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 author's experience in teaching cloud computing technologies [21] that provide necessary 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 [22].

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) [23]. Participation and contribution to both the IG-ETHRD and EDSF Community Initiative is open and free.

7. ACKNOWLEDGMENTS

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

8. 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 Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, 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] Amazon Web Services (AWS) [online] https://aws.amazon.com/

[17] Microsoft Azure [online] https://docs.microsoft.com/en-us/azure/architecture/data-guide/

[18] Google Cloud Platform [online] https://cloud.google.com/

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

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

[21] 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.

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

[23] 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