Distributed and Parallel Programming Curriculum Definition/Design using EDISON Data Science Framework Methodology

Yuri Demchenko, Adam Belloum, Zhiming Zhao

University of Amsterdam, the Netherlands
{y.demchenko | a.s.z.belloum | z.zhao}@uva.nl

Abstract— This paper provides a practical example of how the EDISON Data Science Framework (EDSF) methodology can define a consistent Distributed and Parallel Programming (DPP) curriculum with customization elements for different hosting programs. The EDSF includes four main components for customized curriculum definition: Competence Framework, Body of Knowledge, Model Curriculum, and Data Science Professional Profiles. The EDSF methodology allows for defining curriculum structure for the specific set of competencies defined by a selected professional profile (by the hosting program, e.g., Computer Science or Information Systems). First, competencies are mapped to a group of Knowledge Areas (KA) or Knowledge Units (KU) defined in the Body of Knowledge. KAs and KUs can then be used directly for curriculum definition by the teacher, consulted with the Model Curriculum that contains examples of curricula for different scientific or technology domains defined in compliance with the ACM/IEEE Classification Computer Science (CCS2012). The paper describes the Distributed and Parallel Programming course taught/offered at the University of Amsterdam.

Keywords—Parallel and Distributed Computing, Curriculum Design, EDISON Data Science Framework (EDSF), Competence Framework, Body of Knowledge, Bloom's Taxonomy.

I. Introduction

Parallel and Distributed Computing (PDC) is an essential technology in modern Cloud Computing, Big Data platforms, and large-scale computation and applications. PDC knowledge and competencies are important components of many bachelor's and Master's programs in Computer Science, Information Systems, Data Science, and Artificial Intelligence. Modern data-driven and AI-enabled research and industry created strong demand for new types of specialists that can support all stages of the digital products and services lifecycle from managing sensor networks, data production, and input to (continuous) data processing, Machine Learning (ML) model building, deployment, and operation, as well and technological processes control and automation. Therefore, practical PDC courses and education are important to create a solid foundational knowledge of students in relation to all Computer Science programs.

This paper presents the experience of extending/redesigning/advancing the current PDC course and ensuring its continuous evolution based on science and industry demand. The presented development of the PDC curriculum design/definition is based on the EDSF Release 4 (2022), which is extended with new technologies development and summarizes various experiences in using EDSF for different curricula and course development [1, 2].

II. EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework (EDSF) [1, 2], which is the product of the EDISON Project, provides a basis for Data Science education and training, curriculum design, and competencies management that can be customized for specific organizational roles or individual needs. EDSF can also be used for professional certification and career transferability.

The main EDSF components include (specified in separate documents Part 1 – Part 4):

- 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

The EDSF Part 5 "Use cases and Applications" [3] includes examples of using EDSF for different Data Science curricula and course definitions and practical teaching; it is intended to provide practical guidance for universities, training organizations, data science teams, practitioners to define their Data Science curricula, as well as guide competences and knowledge assessment.

The CF-DS provides the overall basis for the whole framework. The following core CF-DS competence and skills groups are defined:

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, and others) (DSDA)
- Data Science Engineering (including Software Engineering, Cloud Computing, PDC, Big Data Infrastructure, and Tools) (DSENG)
- Data Management and Governance (including data stewardship, quality assurance, and metadata) (DSDM)
- Research Methods and Project Management (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

The DS-BoK defines the Knowledge Areas (KA) and Knowledge Units (KU) for building Data Science curricula required to support identified Data Science competencies. DS-BoK is organized by Knowledge Area Groups (KAG) corresponding to the CF-DS competence groups. The DS-BoK is based on the ACM/IEEE Classification Computer Science (CCS2012) [4] and incorporates best practices in defining domain-specific BoK's. DS-BoK includes KAs related to Computer Science, including Parallel and Distributed Computing.

The MC-DS is built based on DS-BoK and linked to CF-DS, where Learning Outcomes are defined based on CF-DS competencies, and Learning Units are mapped to Knowledge Units in DS-BoK. In addition, three mastery (or proficiency) levels are defined for each Learning Outcome (in alliance with Bloom's Taxonomy [5]) to allow for flexible curricula development and profiling for different Data Science professional profiles.

The EDSF methodology allows for defining curriculum structure for the specific set of competencies defined by a selected professional profile or required set of competencies (for example, by the hosting program like Computer Science or Information Systems) [6, 7]. First, a collection of competencies is mapped to a group of Knowledge Areas (KA) or Knowledge Units (KU) defined in the Body of Knowledge. KAs and KUs can then be used directly for curriculum definition by the teacher or consulted with the Model Curriculum, which contains examples of curricula for different scientific or technology domains defined in compliance with the CCS 2012.

III. Experience in defining DPP curricula

The Distributed and Parallel Programming (DPP) course at the University of Amsterdam was initially developed in 2012 and currently is taught by the authors since 2019 and is currently well-established with a structure that responds to the need of the local job market. However, the enrollment includes many international students. The course is taught as a part of the Computer Science bachelor's and counts for six credits. It runs for two months and has eight contact hours per week, of which 4 hours are lectures and 4 hours are practice and labs.

This course teaches concepts and technologies of parallel and distributed computing, including concurrency, multi-threading, message-passing interface, GPU, cloud, and decentralized applications. In addition, the course gathers the recent development in cloud-based services and applications development, deployment, and operations based on the best industry practices. Students have the opportunity to apply these concepts to understand how they can be best implemented to automate development, test, and release practices.

The course includes 5 group assignments, four programming and one literature study assignment.

Practical skills are defined based on the skills defined for the required PDC competencies and used to define practical and labs/hands-on activities. The lab assignments include the basic hands-on tasks to learn the cloud platforms services The students are required to submit reports on practice, which contribute to the final grade on the course.

IV. Conclusion and Further Developments

Presented experience and using EDSF methodology provide a basis for efficient management of the PDC course. Understanding the importance of teaching PDC foundations, the adopted practice allows for agile/continuous course adjustment to new technologies and developments. Future development will include the adoption of the recommendations by the NSF/TCPP curriculum initiative on Parallel and Distributed Computing [8], which will contribute to the DS-BoK extensions.

Acknowledgement

The work is partially supported by the EU H2020 program through ENVRI-FAIR (824068) and LifeWatch ERIC.

References

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

[2] The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7

[3] EDISON Data Science Framework: Part 5. EDSF Use Cases and Applications (EDSF-UCA) Release 4, 31 December 2022, 111 pp. [online] https://zenodo.org/record/7538447

[4] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012

[5] Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

[6] 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. The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.

[7] 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.

[8] NSF/IEEE-TCPP Curriculum Initiative on Parallel and Distributed Computing - Core Topics for Undergraduates Version 2.0 [online] https://tcpp.cs.gsu.edu/curriculum/?q=system/files/TCPP%20PDC%20Curriculum%20V2.0beta-Nov12.2020.pdf