Brno University of Technology
Faculty of Information Technology

Department of Information Systems (DIFS) Academic year 2021/2022

Master's Thesis Specification

Student: **Šalgovič Marek, Bc.**
Programme: Information Technology and Artificial Intelligence
Specialization: Information Systems and Databases
Title: **Model Driven Development of Spark Tasks by Means of Eclipse Acceleo**
Category: Parallel and Distributed Computing
Assignment:

1. Get familiar with the Apache Spark platform for distributed Big Data processing and with the programming languages it can be used with. Get familiar with Model Driven Development (MDD) and with Eclipse Acceleo. Research the options and existing projects that are used for modeling data processing tasks, especially for the Big Data processing.
2. Design a meta-model for modelling tasks for Big Data processing in Apache Spark and describe its usage with Eclipse Acceleo. Make generation of Spark applications source code from their models possible.
3. After consulting with the supervisor, integrate the designed meta-model and source code generation of Spark applications using their models into Eclipse Acceleo. Verify the functionality by creating models of several Spark tasks for processing Big Data and by generating the corresponding source code.
4. Test the solution, evaluate and discuss the results. Publish the resulting software as open-source.

Recommended literature:

- Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia. *Learning Spark: Lightning-Fast Big Data Analysis*. First edition, 256 pp., O'Reilly Media, 2015. ISBN 978-1-449-35862-4.
- Databricks Spark Reference Applications [online]. 2017 [seen 2021-09-29]. Available at [https://databricks.gitbooks.io/databricks-spark-reference-applications]
- Michele Guerriero, Saeed Tajfar, Damian A. Tamburri, and Elisabetta Di Nitto. Towards a model-driven design tool for big data architectures. In *Proceedings of the 2nd International Workshop on BIG Data Software Engineering (BIGDSE '16)*. Association for Computing Machinery, New York, USA, 2016. ISBN 978-1-4503-4152-3. Available at [http://dx.doi.org/10.1145/2896825.2896835]
- Michele Guerriero, Damian Andrew Tamburri, and Elisabetta Di Nitto. 2021. *StreamGen: Model-driven Development of Distributed Streaming Applications*. ACM Trans. Softw. Eng. Methodol. 30, 1, Article 1 (January 2021), 30 pp. ISSN 1049-331X. Available at [https://doi.org/10.1145/3408895]
- Matúš Bútora. Modelem řízený vývoj Spark úloh. Master's Thesis. Brno University of Technology, Faculty of Information Technology, Brno, 2019. Available at [https://www.fit.vut.cz/study/thesis/21682/]

Requirements for the semestral defence:

- Items 1 and 2 finished and item 3 in progress.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor: **Rychlý Marek, RNDr., Ph.D.**
Head of Department: Kolář Dušan, doc. Dr. Ing.
Beginning of work: November 1, 2021
Submission deadline: May 18, 2022
Approval date: October 21, 2021