Towards Quality-Aware Development of Big Data Applications with DICE

Giuliano Casale, Elisabetta Di Nitto, Ilias Spais*

Abstract. Model-driven engineering (MDE) has been extended in recent years to account for reliability and performance requirements since the early design stages of an application. While this quality-aware MDE exists for both enterprise and cloud applications, it does not exist yet for Big Data systems. DICE is a novel Horizon2020 project that aims at filling this gap by defining the first quality-driven MDE methodology for Big Data applications. Concrete outputs of the project will include a data-aware UML profile capable of describing Big Data technologies and architecture styles, data-aware quality prediction methods, and continuous delivery tools.
Keywords: quality-driven development, Big Data, UML

1 Overview

Big Data systems [4] are rapidly emerging and their popularity on the ICT market calls for novel software engineering methods to support their development. In particular, independent software vendors (ISVs) need to create novel data-intensive products, but this is complicated by the lack of expertise in technologies such as NoSQL databases, MapReduce/Hadoop analytics, or real-time processing. Pressure to hit the market first can therefore shift the development focus primarily on functional aspects, at the expense of non-functional properties such as reliability, performance or safety of the resulting applications.

The goal of the DICE project is to deliver a methodology and a toolchain to help ISVs develop Big Data applications without compromising on quality. DICE proposes innovations concerning both functional and non-functional properties of data-intensive software systems. For what concerns functional properties, DICE wants to extend model-driven engineering approaches based on UML with a novel profile to annotate properties of data such as volume, velocity, location or data transformations. The traditional ecosystem of models used in MDE, which encompasses model ranging from platform-independent to technology-specific, will also consider techologies and architecture styles that are specific to Big data, such as the lambda architecture[1]. The main challenge of this generalization is to develop the model annotations, a consistent methodology, and the underpinning

* G. Casale is with Imperial College London, UK; E. Di Nitto with Politecnico di Milano, Italy; I. Spais with ATC Technologies, Greece. Contact: g.casale@imperial.ac.uk . This paper has been supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No. 644869.

[1] https://en.wikipedia.org/wiki/Lambda_architecture

model-to-model transformations. Furthermore, DICE aims at translating such high-level design models into a concrete deployment plan and execute it.

On the non-functional side, the extended UML models will be annotated with performance and reliability requirements using specific annotations, such as the UML MARTE and UML DAM profiles [3, 1], but also with novel annotations that describe the data used by the application. Then, tools will be developed to predict the fulfillment of these requirements before and during application development. In particular, DICE envisions the co-existence of multiple simulation, verification, and testing tools that can guide the developer through the quality assessment of early prototypes of the Big Data application. For example, a developer could initially describe the application architecture, an expected user behaviour, and the technologies to be used; based on this specification, he could then explore the forecasted response times under increasing volumes or rates of data intakes. This information can be helpful to assess if a given architecture design is appropriate to meet customer requirements. The novelty is the explicit accounting for the data volumes or rates in the predictions.

Application Domains. The DICE development environment will offer a general methodology, that can be useful in a number of application domains. In particular, the project plans to develop demonstrators in the areas of *News & Media*, *e-Government*, and *Maritime Operations*. In News & Media, streaming solutions that connect to social platforms will need to be modelled, together with Hadoop/MapReduce processing of the acquired social data. The case of e-Government provides a test scenario for the DICE methodology to apply in an environment with legacy data systems, where decision-making related to the best Big data technologies to adopt is complex. Lastly, Maritime Operations is a sector where streaming data related to vessel movement needs to be processed and analyzed in real-time to guarantee safe and correct port operations.

Future work. The DICE project has started in February 2015 and a first public release of the DICE MDE tools is scheduled for Spring 2016. News and updates on the project are available at http://www.dice-h2020.eu and a detailed project vision can be found in [2].

References

1. S. Bernardi, J. Merseguer, and D. C. Petriu. Dependability modeling and analysis of software systems specified with UML. *ACM Comput. Surv.*, 45(1):2, 2012.
2. G. Casale and *et al.* DICE: Quality-driven development of data-intensive cloud applications. In *Proceedings of the 7th International Workshop on Modeling in Software Engineering (MiSE 2015)*, 2015.
3. R. Soley and *et al. Modeling and Analysis of Real-Time and Embedded Systems with UML and MARTE.* MK/OMG Press, 2013.
4. P. Zikopoulos and *at al. Understanding big data: Analytics for enterprise class hadoop and streaming data.* McGraw-Hill Osborne Media, 2011.