Big Data Science Architecture for Continuous Technology Transfer from Research to Industry Operations

Richard A. E. Leibrandt

WidasConcepts Unternehmensberatung GmbH,
Maybachstrae 2, 71299 Wimsheim, Deutschland
richard.leibrandt@widas.de
http://www.widas.de

Abstract. Big Data without analysis is hardly anything but dead weight. But how to analyse it? Finding algorithms to do so is one of the Data Scientist's jobs. However, we would like to not only explore our data, but also automatise the process by building systems that analyse our data for us. A solution should enable research, meet industry demands and enable continuous delivery of technology transfer.

For this we need a Big Data Science Architecture. Why? Because in Big Data Science (BDS) projects, Big Data (BD) and Data Science (DS) – influencing each other – can't be handled separately. Thus, their complexities (and gain) multiply: $BDS \neq BD + DS$, $BDS = BD \cdot DS$.

This complexity boost increases further by the clash of the two different worlds of scientific research programming (DS) and enterprise software engineering (BD). The former thrives on explorative experiments which are often messy, ad hoc and uncertain in their findings. The later requires code quality and fail-safe operation, achieved by well defined processes with access control and automated testing and deployment.

We present a blue print for a Big Data Science Architecture. It includes data cleaning, feature derivation and machine learning, using Batch and Real-time engines. It spans the entire lifecycle with three environments: Experiments, close-to-life-tests, life-operations, enabling creativity while ensuring fail-safe operation. It takes the needs of data scientist, software engineers and operation administrators into account.

Data can be creatively explored in the experimental environment. Thanks to strict read governance no critical systems are endangered. After algorithms are developed, a technology transfer to the test environment takes place, which is build the same as the life-operations environment. There the algorithm is adapted to run in automated operations and tested thoroughly. On acceptance the algorithms are deployed to life-operations.

Keywords: Big Data, Data Science, Architecture, Industrial Challenges, Technology Transfer, Continuous Delivery, Batch- and Real-Time-Processing

In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org