`mlwhatif`: Data-Centric What-If Analysis for Native Machine Learning Pipelines

Stefan Grafberger, Paul Groth*, Sebastian Schelter
AIRLab and INDELab, University of Amsterdam

1 Data-Centric What-if Analysis

An important task of data scientists is to understand the sensitivity of their models to changes in the data that the models are trained and tested upon. What if my testing data has more errors than my training data? What if there are high amounts of missing data? Currently, these *data-centric what-if analyses* require significant and costly manual development and testing with the corresponding chance to introduce bugs. Such analyses are becoming increasingly relevant with the current push towards *data-centric AI* and are closely related to recently proposed benchmark tasks in this direction. Examples of such data-centric what-if analyses include the investigation of the *robustness of ML models against data errors* in train and test data, *data cleaning on a budget*, or understanding the *impact of different data preprocessing operations on the fairness* of a ML model.

2 Automation & Optimisation Potential

Instead of forcing the data scientist to spend a long time manually implementing these what-if analyses, our vision is to allow data scientists to declaratively specify a given what-if scenario, and to automatically generate, optimise, and execute the required workloads based on their existing pipeline. In particular, we focus on *natively written* ML pipelines, e.g., pipelines that use code from established libraries from the data science ecosystem, such as pandas, scikit-learn or keras, to enable data scientists to easily apply our techniques without having to manually change their code. The automation and optimisation of data-centric what-if analyses impacts not just the efficiency of the execution but also the coverage of potential problems that may have substantial business and social impact, as well as better reuse and sharing of these experiments.

3 Current State & Ongoing Work

In previous work [2, 3], we developed machinery to instrument natively written ML pipelines and extract an abstract dataflow representation. Recently, at the DEEM workshop at SIGMOD, we proposed directions on how to optimise the execution of what-if analyses by modeling pipelines as dataflow computations and reusing trained models and intermediate results [1]. In ongoing work, we implement our ideas in a new system `mlwhatif`, based on our machinery for native ML pipelines.

This work was supported by Ahold Delhaize. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their employers.

References

[1] S. Grafberger, P. Groth, and S. Schelter. Towards data-centric what-if analysis for native machine learning pipelines. *DEEM workshop @ SIGMOD*, 2022.

[2] S. Grafberger, P. Groth, J. Stoyanovich, and S. Schelter. Data distribution debugging in machine learning pipelines. *VLDBJ*, 2022.

[3] S. Grafberger, S. Guha, J. Stoyanovich, and S. Schelter. MLINSPECT: A data distribution debugger for machine learning pipelines. *SIGMOD*, 2021.

*Prospective Speaker