DataOps & MLOps: using data more intelligently

_AI-generated image. Prompts: electronic devices falling into a giant funnel, digital art.

Expert article

_Heiko Faller, Managing Director Enterprise Solutions and Member of the Executive Board

_Dr. Wilhelm Kleiminger, Head of Data Science

Ergon Informatik

Published in SMART insights 2023 magazine

ergon smart people – smart software®

Big data has turned out to be a flop for many companies. Hunting for buried data treasure is costly, the added value rather small. That’s where MLOps and DataOps come in, with efficient processes to turn raw data into a genuine competitive advantage.

Many companies have amassed large volumes of data in recent years. What most of them have done with it is less than impressive, however. Whether in an SME, the healthcare sector, or the finance industry, often this supposed gold mine of data causes high costs with little real return. It doesn’t have to be that way. Deployed properly, the data-centric approach can generate clear added value in everyday business, and shorten time to market for new products. The magic formula is ‘DataOps plus MLOps’. It uses dedicated pipelines and high-quality datasets to produce ongoing competitive advantages.

DataOps and MLOps encompass processes, best practices and technologies that allow data and machine-learning models to be provided faster. They are structured, automated and reuseable. DataOps alone can increase a company’s productivity by 10 per cent and cut time to market by as much as 30 per cent, according to a study by McKinsey. Instead of the time and expense of digging through huge mountains of data, a DataOps–MLOps pipeline produces the output in house. The right processes turn raw data into real value.

How companies profit from machine learning
Machine learning (ML) allows companies to find solutions to problems where traditional software development falls short. Unlike conventional software, an ML model is not fully programmed. Instead it ‘learns’ its functionality with exemplary input and output data. That is particularly helpful in areas in which functionality cannot be described within the program. Automatic speech recognition is one example. The many characteristics of human speech cannot be captured exhaustively in source code. Today, although we have numerous ML models in the widest range of fields that can be adapted at little cost to a specific use case, it is still a great leap from technical prototype to productive use. This makes data preparation, and the selection and training of the ML model, all the more important.

Dedicated pipelines for a structured approach
DataOps and MLOps are based on the concept of pipelines. At one end of the pipeline, the system records the raw data, and on the other it produces finished reports, dashboards and machine-learning models. Within the pipeline itself, well-defined

transformation steps make the desired results verifiable and replicable. In practice, this means efficient, transparent and faster data processing.

Let's take an example: a company develops its own voice-user interface to operate an application using voice commands. To do this, the company uses an MLOps pipeline to adapt an existing language model to its own use case with domain-specific data. First, it inputs training data. Depending on requirements, it transforms this data, from time domain to frequency domain, for example. It then enriches this data further, by masking certain frequencies, for example, to remove distracting background noise.

This freshly prepared data enables an existing language model to be modified to suit a new use case. This type of transfer learning can produce good results even with little training material, thereby producing marked efficiency gains.

Quality not quantity
Gathering application-specific data is a laborious and expensive task. That is why data quality generally has a higher priority than quantity. Metainformation is particularly important because it permanently records the conditions surrounding data capture. These can be factored in to training a model. For a language model, for example, variables such as the weather, the device used and the user's gender can all affect the model's precision.

DataOps offers the necessary strategies and tools to collate and prepare data from a range of sources. A DataOps pipeline can register these datasets, filter them and convert them into a standard format. It also allows various statistics to be visualised, such as the frequency with which words are used, or the gender composition of the body of speakers. In this way, a DataOps pipeline always helps to keep an overview, even at the early stage when collecting proprietary datasets.

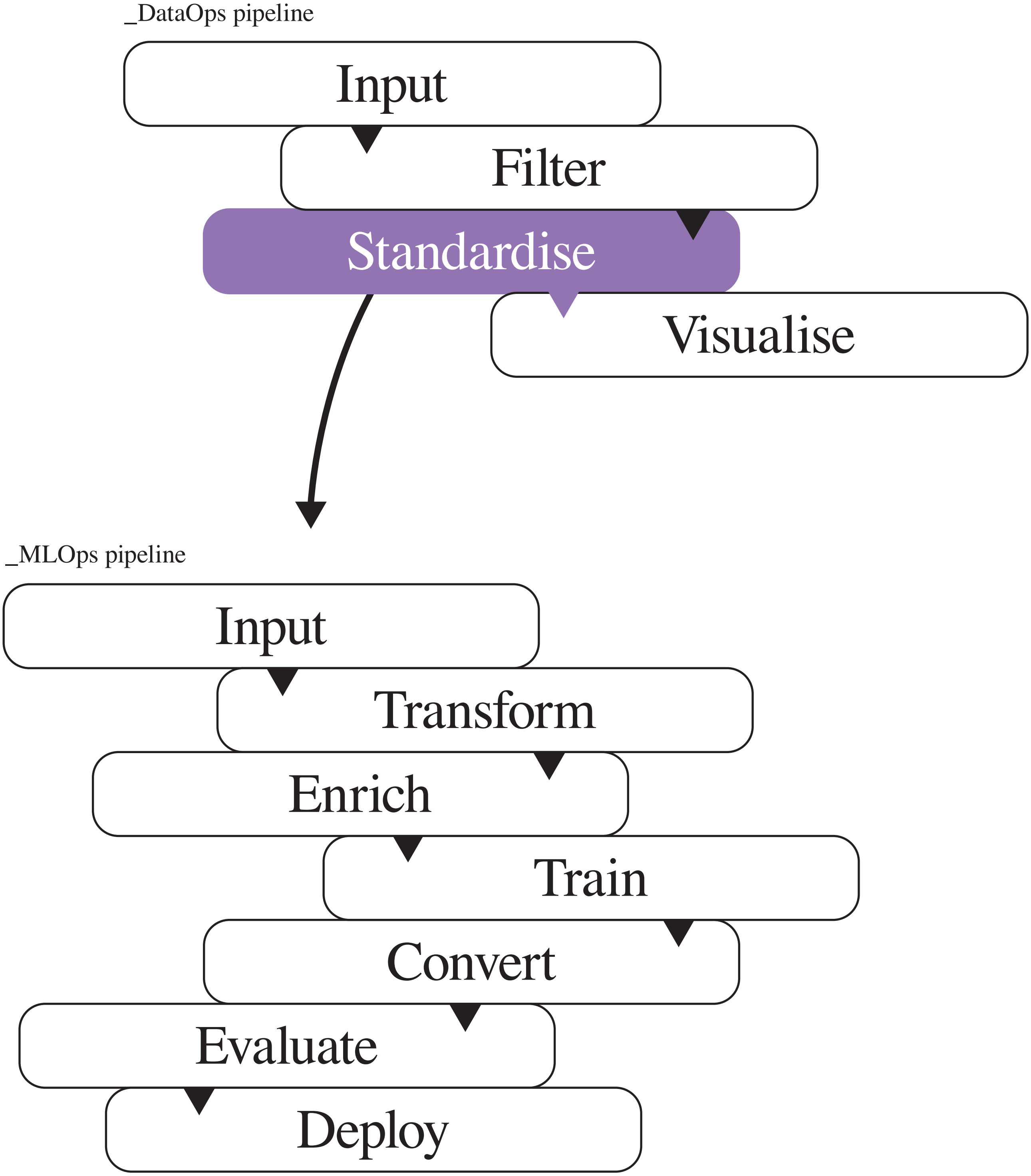
Data and models require management
Robert C. Martin, author of 'Clean Code', said that the truth in conventional software development can only be found in the code. In machine learning models, though, this 'truth' is distributed between the chosen ML algorithm, the training data used, and any configuration parameters. To understand and reproduce results, the individual input and output artifacts must be neatly versioned and managed. Which datasets have been used for training, and which not, must always be clear. That is the only way that the model can respond to new requirements, such as additional vocabulary or new data from ambient factors such as background noise.

Permanently integrated into productive environments
Further modifications may have to be made to integrate a model into the productive environment. For example, if the language model is to be run not in the cloud but on a mobile device, an additional conversion step will have to be included in the MLOps pipeline. This converted model will also have to be tested again, to ensure that the conversion has not significantly impaired performance.

Experience of operation in a real-life setting may then result in a further round of changes to the model. In many cases, a model has to be in productive use before it becomes clear which marginal conditions have not been given enough attention. To keep improving a model, the productive data has to be useable. Furthermore, it is worth having a separate release cycle for ML models that is independent of the release cycle of the app itself. That means you can respond in good time if it is time for the data to be updated.

Data projects: start small and agile
Companies that want to use DataOps/MLOps are well advised to start with a clear, straightforward business case in a small team. This team works

_Dedicated pipelines for a structured approach



_DataOps pipeline

Input

Filter

Standardise

Visualise

_MLOps pipeline

Input

Transform

Enrich

Train

Convert

Evaluate

Deploy

“DataOps and MLOps make companies competitive in the age of AI.”
_Heiko Faller, MD Enterprise Solutions and Member of the Executive Board, Ergon

“DataOps and MLOps are processes that have to be improved all the time.”
_Dr. Wilhelm Kleiminger, Head of Data Science, Ergon

out which steps are actually needed to develop a productive solution. Without a clear strategy, having a company-wide location to store structured and unstructured data can be costly, and is often the wrong way to go about things. Full-stack data scientists can help keep the focus on what counts. They support the necessary steps; from requirements engineering to putting the solution to productive use. As is the case with agile software engineering, the task is to get an initial productive version into operation fast to gather valuable experience.

This kind of approach has the advantage of being easy to implement, and it proves its worth directly with every further successful project. If development teams are working in different data silos around the company, it is important that they stay in dialogue. This builds a DataOps/MLOps culture with a long-term impact. Here, it is not about the chosen tools or a one-size-fits-all methodology, but about clear processes and clear business cases.

Particular attention must be paid to the data needed to cover the use case. Its quality must be impeccable. With structured DataOps/MLOps pipelines, all steps of the process, from data gathering to the productive integration of the ML model are carried out consistently and transparently. It is a convergent approach, and therefore offers greater prospect of success than hunting for buried, big-data treasure. It costs a lot less, too. />

DataOps and MLOps in brief

DataOps makes data pipelines more efficient. DataOps tools and methods oversee data quality and make data processing systems scaleable. MLOps, also known as DevOps for machine learning (ML), focuses on developing, providing and overseeing ML models. MLOps methods implement, monitor and optimise ML models in productive systems. Both approaches require data scientists and IT teams to work together.