© Sergey Nivens | Adobe Stock

From Data to Company Values

AUTHORS

Hendrik Bohlen
is Technical Director at Werum Software & Systems AG in Lüneburg (Germany).

Dr. Stefan Unterschütz
is Head of Program Test Process Management at Werum Software & Systems AG in Lüneburg (Germany).

The test process in vehicle development must ensure high quality at maximum cost efficiency while meeting compliance requirements. However, data from simulation and testing all too often ends up in data silos. With a suitable data analytics tool chain, the different types of data can be linked with added value being generated.

TOOL CHAIN INTEGRATES ALL TEST DATA

Today's vehicle development makes high demands on the test process. High quality must be ensured at maximum cost efficiency – while meeting compliance requirements.

Valuable data on the vehicle are generated in simulations and tests as part of the test process. Among them are tests on endurance, engineering strength, safety and performance. With the use of smart devices, increasing amounts of data are expected to come directly from the field in the future.

Usually, data are stored at the place of their origin, in so-called data silos. Only rarely, they are available across the whole enterprise. That is why it is not possible to derive the added value from them as it could be done by linking these different types of data.

State-of-the-art IT offers solutions for the digital transformation of test processes to tap the data's maximum potential. This article outlines the requirements imposed on a data analytics tool chain and a feasible implementation. **FIGURE 1** shows a simplified depiction of the tool chain process. It illustrates the process steps

data go through on their way from their source to the interested user or continuative systems.

The solution can be set up independently of existing IT systems, and the different data suppliers like testbeds, IIoT devices or applications can be integrated gradually.

CHALLENGES IN DATA PROCESSING

A state-of-the-art tool chain for the test process has to deliver many functional and non-functional requirements. Once acquired, the test data need to be pre-processed. The data records need to be cleansed, the data are qualified and enriched with additional information or initial statistical evaluations are necessary. In this context, the data may also undergo verification, that is, they are checked for completeness and formal correctness. It may also be sensible to harmonize and standardize the data as part of the pre-processing as it facilitates their future use.

Test data require audit-proof storage for the applicable retention periods. Here, accompanying information, so-called metadata, should be filed next to the actual test data. Short access times are desirable to achieve fast processing. Rapidly increasing amounts of data in today's test process require a shift toward scalability.

Traceability, transparency and auditing are important elements in dealing with data. Therefore it must be logged how, on which testbed, and by the use of which sensor the test data were generated. Any data modification whatsoever must be documented.

Most diverse methods may be necessary to analyze the test data. They range from aggregation and compression, transformation or calculation of statistical characteristics to highly complex machine-learning algorithms. The number of available analysis tools also varies, which calls for a high degree of flexibility in integration.

The tool chain needs to enable access to the data for users and for external applications, such as a PLM tool. This also entails the need for searches across all data and their visualization. Powerful access control, too, is required – frequently in combination with release processes. Next to access control based on roles and rights, further mechanisms may be necessary to classify and protect the data.

The number of data-related compliance requirements enterprises have to meet are increasing, for example requirements on information protection according to GDPR, ISO/IEC 27001, TISAX or proving the required product characteristics, for example WLTP or homologation. In addition to these standards, verifying the test laboratories' competence according to DIN EN ISO 17025 is gaining in importance.

There may also be other requirements to fulfill: fail safety, availability, reliability, fault tolerance, and distributability are among the substantial demands made on all tool chain services apart from performance requirements. Finally, the requirements on system operation need to be defined. Essential key words in this context are orchestration and monitoring.

Enterprises need to assess all the requirements according to their relevance, which has a significant impact on the tool chain design. Currently, out-of-the-box solutions are not available. The enterprises' requirements are far too different, applicable regulations and laws change far too quickly, and there are far too many tools for data analysis.

In the following, it is explained how a state-of-the-art tool chain can be set up. To the extent deemed practical and useful, the tools stated are open source tools. Where necessary, commercial equivalents are used to cover certain functional aspects.

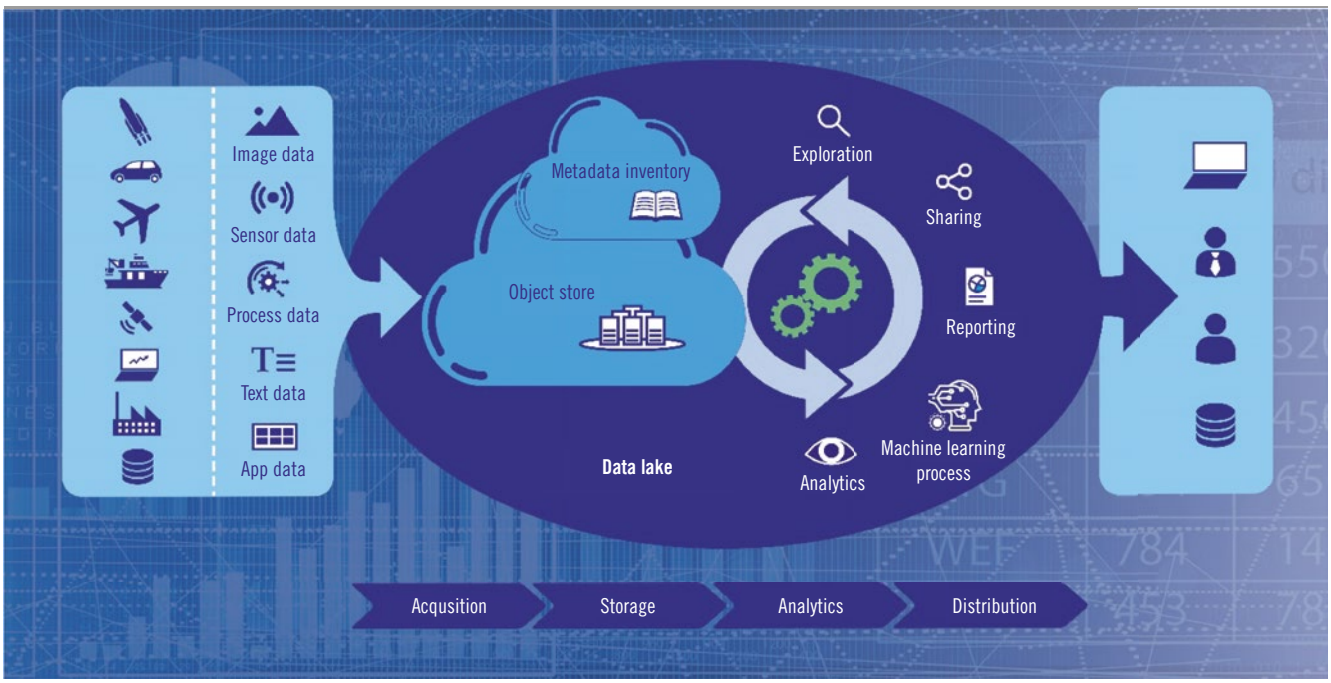

FIGURE 1 Process of the data analytics toolchain (© Werum)

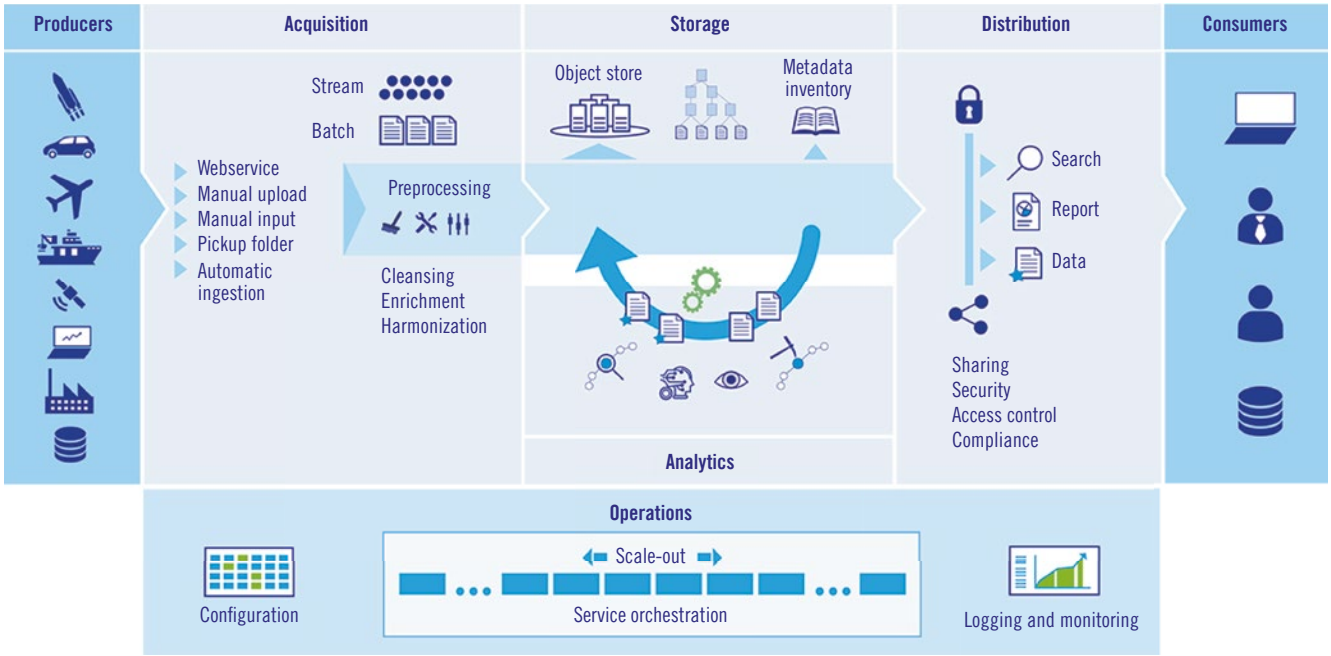

FIGURE 2 Function-oriented structure of a toolchain (© Werum)

OF DATA AND METADATA

The most suitable analogy to illustrate the processing of data certainly is the processing of raw materials. Raw materials are refined and cultivated in complex process steps to manufacture high-quality products until the desired final good is achieved. A tool chain basically applies the same mechanism. It transforms raw data into higher-end data products, which again can be transformed into other data products.

Besides data transformation and analysis, enriching the data with metadata also plays an important part. Metadata are pieces of information about the data. Without metadata, measured data are worthless for any valid further use by the company. Metadata can be divided into different categories:

- descriptive metadata which identify data, make sure they are findable and enrich them with additional context, for example the instruments that were used for the measurement, the place and the time when it was done and who did it
- structural metadata which provide information on the data's type and internal structure
- administrative metadata which enable the management of the data (for example access rights)
- statistical metadata which provide statistical information on the data (for example KPIs).

Metadata are of particular importance when it comes to linking and processing data in a tool chain. Existing applications producing data may need enhancements in order to make the required metadata available.

DATA ANALYTICS TOOL CHAIN STRUCTURE

FIGURE 2 shows the function-oriented structure of a tool chain. A tool chain resides between the producers and the consumers of data. Typical producers are testbeds or measurement facilities or, in the future, any kind of IIoT device. Data may also be manually acquired parameters or photos that were taken as part of the diagnosis. Apart from the actual measured data, the descriptions of test order and test as well as potential comments and documented events are added as metadata.

The next section describes the functionality implemented by adequate services and applications.

FROM PRODUCER TO CONSUMER

The user manually uploads the data, for example by copying files, using a web frontend, by entering single values, or in an automated process using for example a REST API or a transfer directory. In the next step, the data are accepted and pre-processed. A batch or stream process can be used for implementation, where the latter enables timely data analysis.

The essential functions of pre-processing are cleansing and enrichment of the data as well as their harmonization and standardization. Moreover, the metadata are treated in suitable manner to enable searches and further processing. At the end of pre-processing, the data and metadata are filed in a store.

An object store is the ideal medium for this purpose. It is optimized for the storage of large quantities of unstructured data and has prevailed over the last years against file-based storage or Apache HDFS. Object stores offer almost unlimited scalability while costs are low and access times are short.

Object stores have the capability to store the metadata in addition to the actual data. In real life, however, employing a relational database solution has proven to be more advantageous as it stores the metadata in structured manner plus a reference to the actual data, a so-called

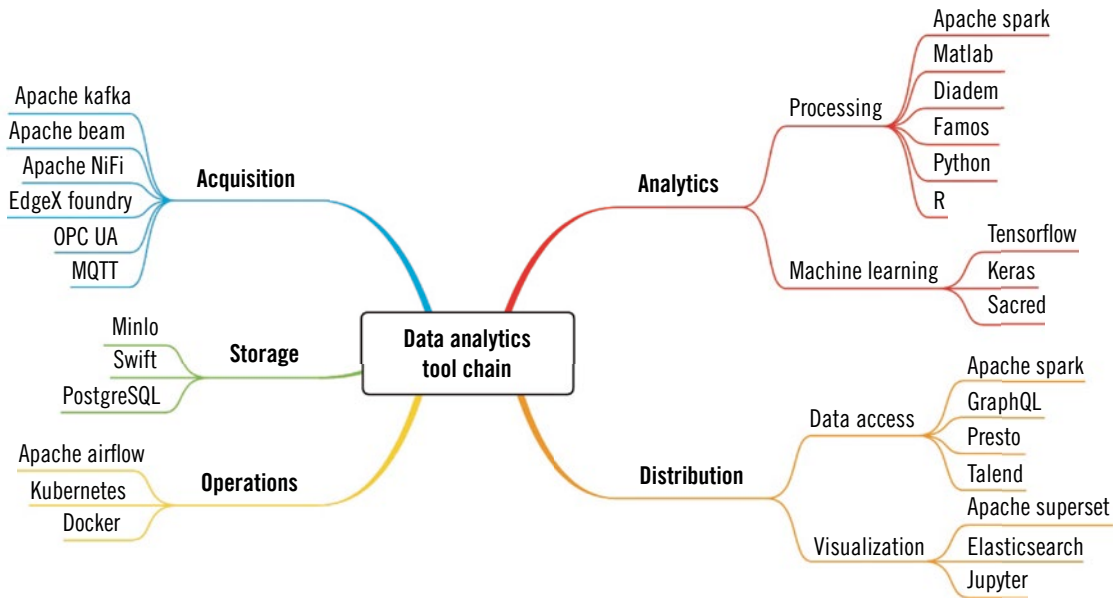

FIGURE 3 Possible tool-chain tools (© Werum)

(meta-)data inventory. It makes it much easier to adjust metadata, define expressive search queries and to implement data processing based on metadata.

The underlying IT infrastructure must be capable of safeguarding distribution of the data to make sure consumers can access the data and metadata. The issues of access and access control should be considered in a multi-faceted approach as, next to the role-based model, access to data may also be controlled by metadata, such as location or release status. Usually, specialized solutions need to be involved for this highly complex kind of requirements.

The processing or transformation of data can be performed as part of the data distribution so that required data formats or entire reports can be provided.

ANALYSIS AND EVALUATION

To have the previous functionality of a tool chain in place is the fundamental prerequisite for data analysis and their processing into higher-end products.

Explorative analysis and evaluation of single data sets using adequate tools is possible just like applying standardized evaluation routines. So, an order workflow could be established that accepts test orders from a PLM system. The test data are then assigned to the test orders and can be processed driven by specific processes. The final result can be a report for the customer, which the tester has to release beforehand. The benefit of a tool chain is that all generated data are available to the tools using a standardized interface while ensuring access control at the same time.

Spark offers scalable and paralelized processing of large amounts of data. It meanwhile has become quasi-standard and has outclassed early approaches like Apache Hadoop.

The data can be utilized further using machine learning methods, which open up new insights and fields of application. Approaches like predictive maintenance can be applied in endurance and field tests, for example, to identify anomalies or potential component failures at an early stage. Machine learning can also be used to point out “suspicious” data records to the tester. In view of the increasing number of data expected to be generated in tests in the future, this kind of machine-supported data assessment is becoming indispensable.

OPERATIONS AND TOOLS

A service-oriented architecture is well-suited for a tool chain. Additional tools are needed to orchestrate, configure and monitor all the services. A workflow engine can be used to manage the processing steps described above and to have them executed automatically.

The decision whether it should be operated on-premise, as a corporate cloud or as a public cloud, strongly depends on the individual enterprise’s requirements. Apart from cost aspects, IT and compliance requirements need to be considered. **FIGURE 3** shows suitable tools to build up a tool chain.

CONCLUSION

A data analytics tool chain is of central importance to tap the full potential of data and to open up data silos. The overview of existing tools and their interaction shows that today’s requirements on the test process can be met by a suitable tool chain. Using it can accelerate the digital transformation in the test process area.

The available tools still need to be integrated into the enterprise’s environment, as there are no out-of-the-box tool chains that would meet an enterprise’s requirements.