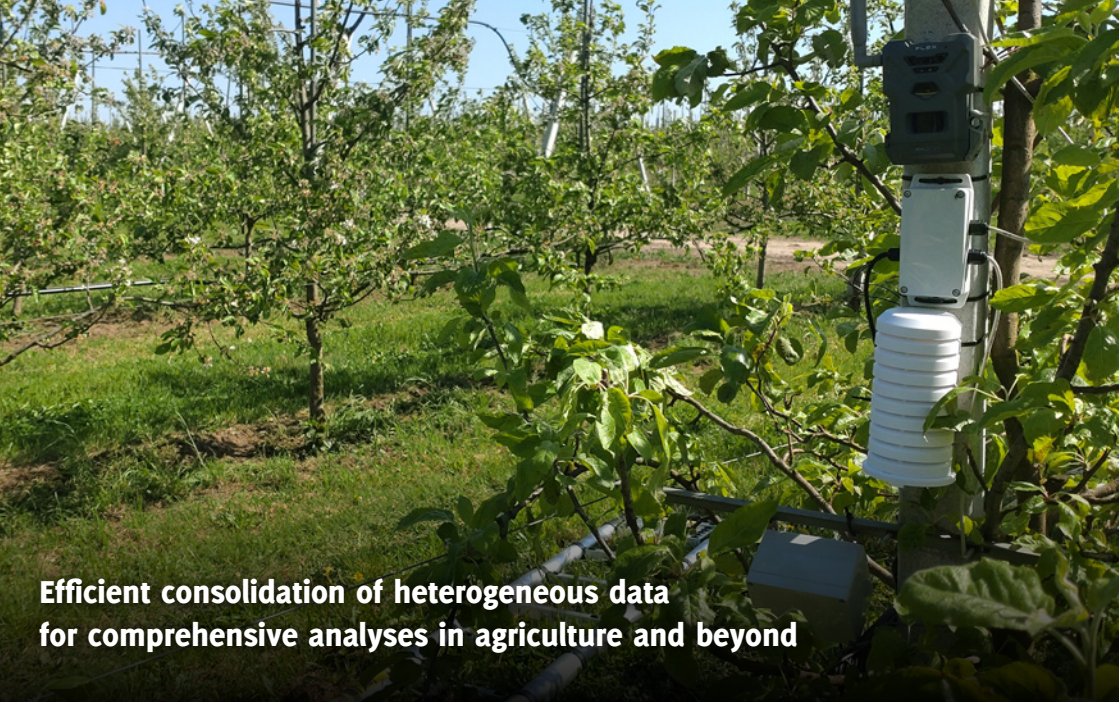


Efficient consolidation of heterogeneous data for comprehensive analyses in agriculture and beyond

In MIRO, data is collected in the form of time series and used, among other things, for research on the digital twin of regional fruit varieties. IMMS has developed a concept for the efficient merging of heterogeneous data for comprehensive analyses in agriculture, which can also be used for other applications. Photograph: IMMS.

Motivation and overview

www.imms.de/miro

In the "Mitteldeutsche Innovationsregion Obstbau" ("Central German Innovation Region for Orcharding", MIRO) project, IMMS is working on digitalisation solutions with the aim of strengthening the future security of the entire fruit value chain from cultivation and processing to marketing in the region of Central Germany and thus addressing issues such as a rapidly changing climate and a challenging skilled labour and competitive situation in the agricultural economy, especially in fruit growing. For example, problems in processing could be avoided by adapting cultivation methods or finding locally suitable fruit varieties depending on the specific environmental and soil conditions, even in times of climate change.

The exchange and consolidation of data from different systems and actors is central to digitalisation. For example, as in many other disciplines, a wide variety of data is required in agriculture to ensure efficient processes on and across farms along the value chain. This ranges from producers and suppliers to logistics with refrigerated warehouses and processors such as wineries or canning manufacturers to finished

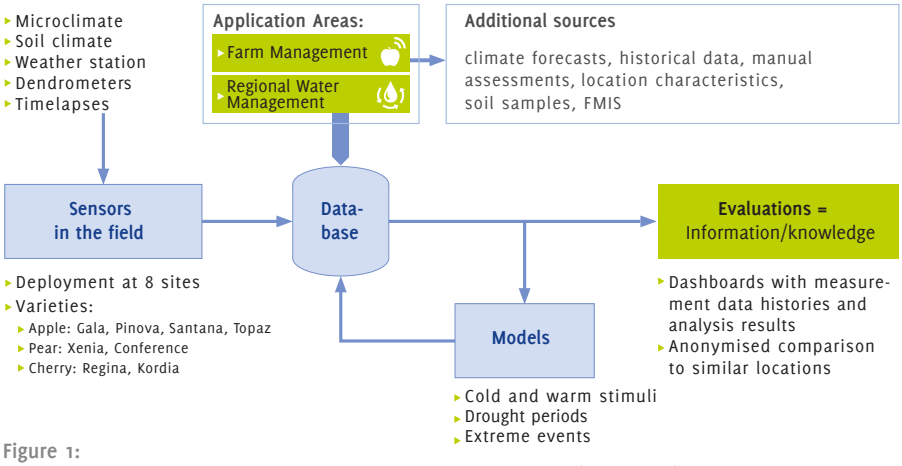


Figure 1:
Data requirements and processing for the digital twin in MIRO (schematic). Graphic: IMMS.

products such as juice, wine or apple sauce for end customers in retail. However, other interest groups such as plant breeders, tree nurseries or authorities also increasingly expect data to be exchanged with the farms. This is made more difficult by numerous different systems, unsuitable interfaces (e.g. manufacturer-specific) and separate initiatives. As a result, data has to be processed multiple times, sometimes with increased manual effort if things have to be documented or data from different systems have to be exported and merged manually. Such activities are not among the core tasks of farmers.

Against this background, IMMS has worked on two use cases in MIRO: the data exchange between the players in the region just briefly outlined and a digital twin for feedback on variety characteristics at different locations, along the value chain.

www.imms.de/miro

IMMS has analysed the data exchange use case with regard to both agricultural aspects in general and the project partners involved. First of all, the needs of both target groups were analysed and the efficient, ideally automated merging of data and its storage in a form suitable for further use for various purposes were identified as objectives.

The solution approach based on currently available platforms should implement a standardised, central repository for all types of structured and unstructured data, which facilitates the retrieval and storage of data as well as subsequent retrieval or analysis using data from multiple data. This approach should also not be specifically

limited to the processing of data in MIRO, but rather also benefit other data-intensive applications, e.g. in AI applications.

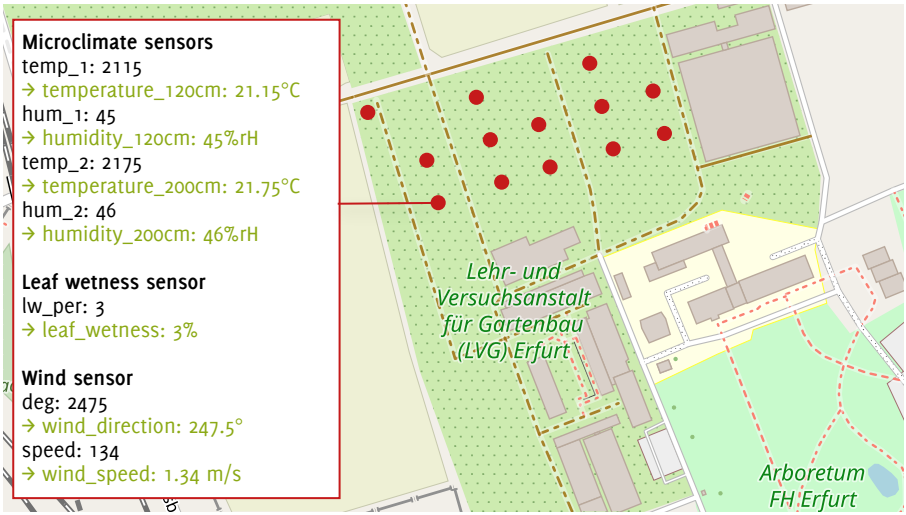
Data collection of time series using the example of digital twins of regional fruit varieties

At IMMS, extensive data is collected not only in MIRO, but also in other projects, such as EXPRESS, specifically time series data from various types of sensors in agricultural fields, weather data and image data. Data sources are sensor installations operated by IMMS (such as wireless sensor networks, weather stations or wildlife cameras for plant observation), partners or also public and commercial data providers on the Internet, such as Deutscher Wetterdienst (DWD, German weather service) or sensor platforms from which data can be obtained via various interfaces (APIs).

Although time series data in these projects are already systematically entered into time series databases at the institute, they are written by gateways in the field without further processing. Aspects such as the assignment of measuring nodes to measuring points (which may vary over time) or the assignment of index-referenced sensors to specific measuring depths or heights can only be associated with the data in subsequent steps. In practical terms, this means that this meta-information

www.imms.de/express

Figure 2:
Raw vs. semantically processed data for a sensor node. The processing includes the assignment of semantic references such as installation height/depth and conversions. Map created with MapOSMatic/OCitysMap on 21 July 2025. Map styles: Baumkarte by Oliver Rudzick; Allotments overlay; Data source: Map data ©2025 OpenStreetMap.org and contributors (see http://osm.org/copyright).



must be recombined with the data for any further use, which is prone to errors if not done automatically. Furthermore, time series databases, especially InfluxDB, frequently used at IMMS, typically cannot be queried using the standard SQL database query language, which makes it difficult to use for analyses.

The work in MIRO on the digital twin use case requires a more extensive evaluation of data and has shown the necessary starting points for the company's own databases. Here, sensor-based and manually collected data, such as that from manual assessments, should allow conclusions to be drawn about the suitability of varieties against the background of climate change. On the other hand, other projects dealing with entirely different research topics but also with the aggregation and evaluation of large amounts of data have equally shown a need for and the potential of new approaches.

www.imms.de/miro

Modern approaches for large, heterogeneous databases

Work in the MIRO use case of data exchange therefore began by analysing the state of the art for data storage in general. Conceptually, after the older concepts of data warehouses for the efficient storage of standardised, structured data and data lakes for the efficient storage of heterogeneous (including unstructured) data, the state of the art has arrived at the so-called (data) lakehouse, which strives to combine the advantages of both approaches.

There are various open source solutions for lakehouses that are complex to set up and use and rely on object stores as data storage. An object store saves files as objects in so-called buckets. In these object collections, they are stored by name, an optional path and possibly other attributes. Examples of object stores are cloud storage services such as Amazon AWS S3 or the open source solution MinIO with a compatible interface. Prominent lakehouse solutions in the open source space are Apache Iceberg or Apache Hudi, or, commercially, Databricks or Snowflake. Iceberg and Hudi were analysed in more detail as they can also be operated on-premise.

When looking at full-blown lakehouse solutions, it quickly became clear that they require powerful hardware for operation on company infrastructure, such as clusters for Spark SQL as the query engine. On the other hand, they also require considerable expertise for operation and use. The latter represents a significant hurdle for an in-

stitute or other operator in terms of IT resources and training of specialists without in-depth database knowledge. However, some aspects of the complex lakehouse solutions are motivated by big data applications and Fortune 500 companies, as it is the backend for the entirety of their corporate data and thus business. In contrast, this exceeds the requirements for the storage of measurement data and analyses (WORM, write-once read-many) by a significant margin.

Looking at current technologies and alternative approaches has led us to a viable alternate approach that realises a lakehouse by storing Apache Parquet files directly in a suitable hierarchical file system structure in the object store. Parquet is a binary file format for tabular data that is now being widely used. In contrast to CSV, Parquet files are much more efficient in terms of storage space and read performance and also avoid typical problems when working with CSV files (such as unclear data types of columns, handling of missing values, etc.). Furthermore, they allow for the integration of metadata. Parquet files in MinIO/S3 can be queried via Apache Spark SQL or via the in-process database DuckDB via SQL, much the same as regular databases. This approach avoids the complexities of full-blown lakehouse table formats in use and operation; restrictions compared to these are insignificant for the utilisation scenarios under consideration.

Goodbye CSV, hello Parquet – More flexible yet lightweight data storage for MIRO

Instead of writing data from the source, especially sensor installations in the field, directly to the lakehouse, it has advantages to retain a time series database such as the InfluxDB already being used in these projects. This is easier and more lightweight to access from resource-constrained devices in the field such as IoT gateways and can also be located somewhere other than centrally on a server at IMMS, if required. The flexibility of InfluxDB compared to SQL databases has proven itself in that no rigid schema needs to be defined and maintained. And since it does not make sense to maintain all the other information necessary for data transformation and metadata augmentation on the gateway, this step is better realised downstream and separately.

The ingress into the lakehouse is carried out periodically using a set of flexible custom tools. The data is processed (filtered, converted) and enriched with metadata (physical quantities, units, locations etc.). This process results in Parquet files pertaining to individual sensor installations and specific time ranges, which are stored in a well-thought-out organisational structure in MinIO. Data can then be retrieved from them and processed further.

Data storage using InfluxDB or other databases

Figure 3 shows an overview of the concept with two tiers of data storage: The first tier in this case consists of an InfluxDB instance (own preference: other databases would also be conceivable here if the approach was to be adopted by others) for sensor data collection and other existing or externally-provided sources, e.g. other databases or simply CSV files. A set of dedicated tools is used to "lift" data from this first tier into the lakehouse. This involves some initial processing and metadata annotation. The lakehouse can be used for further processing, visualisation and as a basis for interfaces or data exports, as is planned for the digital twin of regional fruit varieties.

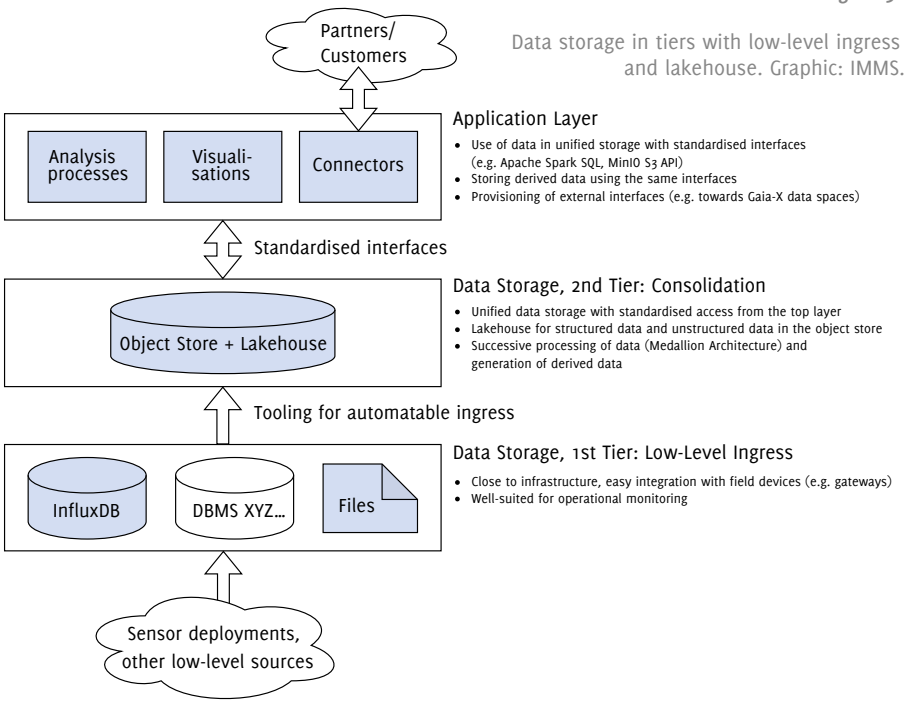


Figure 3:
Data storage in tiers with low-level ingress and lakehouse. Graphic: IMMS.

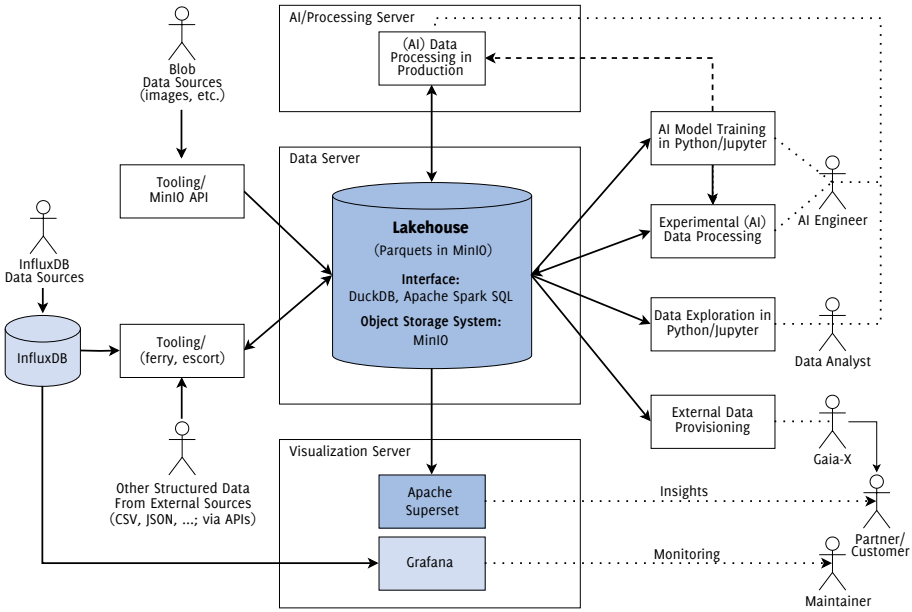


Figure 4: Sustainable data storage via a lakehouse based on Parquets in MinIO. Graphic: IMMS.

Figure 4 further details the concept: At the heart of the data storage is the lakehouse based on Parquet files in a MinIO object store. In the MinIO/lakehouse, there are individual buckets for each project or sensor deployment. In each of these buckets, there are two directories at the top level: “warehouse” for structured data, further organised according to the levels of the so-called Medallion Architecture – with levels Bronze, Silver and Gold for different degrees of processing – and “blobs” for differently or unstructured data such as images, videos or AI models.

Automatic and continuous data transfer

The ingress, i.e. the transfer of structured data into the lakehouse from InfluxDB instances and other sources, is automated and continuous using two software components implemented for this purpose. These are flexibly configurable and generate time slice Parquets (each for one year, one month, one day or one hour). By writing data in individual files for shorter time periods and merging them into larger time periods as soon as those are over, inefficient constant rewriting of all available data is avoided. When the Parquets are created, they are enriched with metadata based on a JSON structure conceived by us. For each time series, the metadata is also stored in a separate JSON file with a related name that can be accessed even more easily.

Figure 5: Monitoring dashboard, here for the commissioning of a Mira sensor network, based on Grafana and raw data including technical metadata in InfluxDB. Source: IMMS.

InfluxDB remains the preferred source for monitoring sensor installations as it is easier to understand problems at the level of raw data and other technical metadata, which are, on the other hand, not relevant for the actual data analysis and therefore not transferred into the lakehouse. Here, Grafana has proven itself for plotting data in dashboards. In future, Apache Superset will be used as the primary means of visualising processed data in the lakehouse. In contrast to Grafana, Superset can also visualise data other than time series and individual values and offers interaction options.

Data processing

Data processing on the lakehouse can be carried out using various approaches: Python and R make it equally easy to access Parquets in the lakehouse via DuckDB or Spark SQL. This applies to algorithmic processing approaches as well as AI-based ones. When developing processing algorithms, it is also easy to initially work with fixed exemplary Parquets and then switch to “live data” in the lakehouse by adjusting the path or URL during production, without having to make any adjustments to the code. Processing results can also easily be stored as Parquets in the lakehouse, but can also be written and validated locally during development for cross-checking.

In MIRO, a MinIO production instance has been set up on servers at IMMS. The setup of an Apache Spark SQL cluster was dispensed with in favour of using DuckDB to query Parquet files. In addition to the MinIO instance, productive instances of the tools created for automated data ingress have been configured and are running in production.

On this basis, all time series data previously and still being stored in InfluxDB instances for the digital twin in MIRO could be fed into an automated continuous ingress into the lakehouse, which preprocesses the data, semantically assigns it and annotates it with metadata. Further automation also continuously performs interval normalisations of the data typically recorded asynchronously, and also script-based processing using Python and R. This means that data from all sensor installations is continuously being preprocessed and made available in the lakehouse with low la-

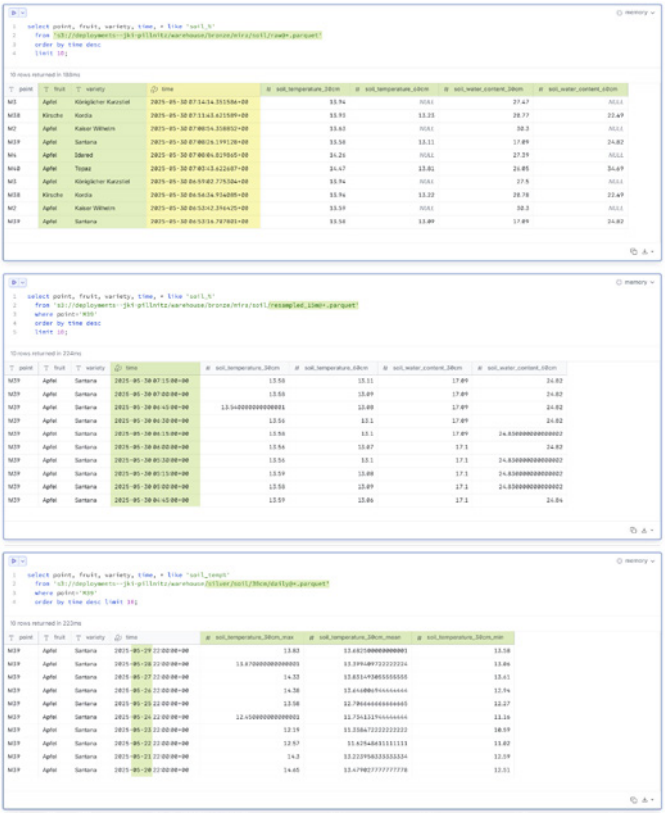


Figure 6:

Illustration of data preparation using sample data from SQL queries on the lakehouse via DuckDB.

From top to bottom:

1. Raw data from different measurement points written on Bronze, with non-uniform/asynchronous timestamps and metadata annotation.

2. Interval-normalised data of a measuring point written on Bronze.

3. Daily aggregates for a measuring point on Silver. The URLs in the queries also show the organisational structure used in the lakehouse/MinIO.

Graphic: IMMS.

tency. The data generated at the Medallion Architecture's Bronze level is periodically further processed into Silver (e.g. through aggregations). In addition to structured data, unstructured data, such as images or even firmware images as backups, are also stored in the lakehouse, where it is centrally accessible.

In MIRO, image data is also stored in the lakehouse, which can be easily accessed with MinIO/S3 client libraries and thus be incorporated in data processing. The intermediate results obtained from working with the data in the form of derived data can in turn be stored in the warehouse part of the lakehouse.

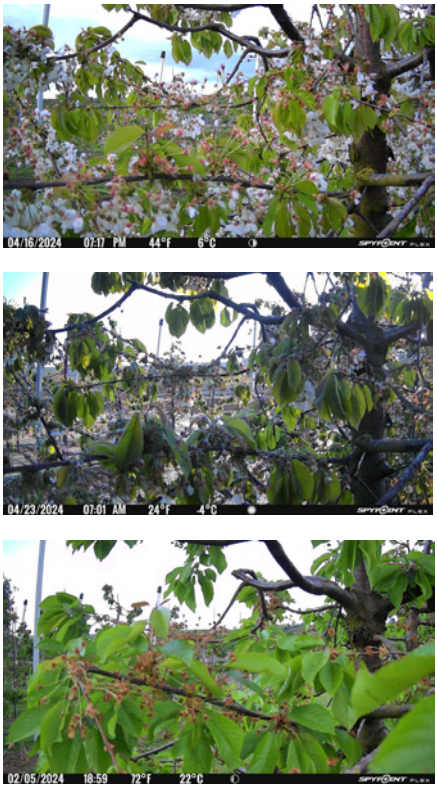


Figure 7: Frost damage to cherry blossoms at LVG Erfurt, documented by wildlife camera photos in April 2024: blossoms before frost, during frost, after frost. Photosgraphs: IMMS.

In addition to data collected by us, various data from external services are also integrated and stored as Parquet files. In particular, data from the DWD, the Helmholtz Centre for Environmental Research (UFZ), but also data from various weather station providers are integrated, depending on which sensor technology was already installed at the respective partners' or synergistically deployed by partners in the project.

An initial site analysis has already been carried out with the aim of visualising the effects of climate change. In doing so, data on historical climatic development provided by the DWD, climate projections provided by project partner UFZ and automated aggregations of monthly temperature and precipitation data were used. Figure 8 shows an exemplary analysis in Superset.

www.imms.de/miro

Temperature (monthly average)

Metric	Average temperature [°C]												
month	1	2	3	4	5	6	7	8	9	10	11	12	Total (Average)
slice													
200x	1.8	2.0	4.4	8.5	12.4	16.0	18.3	18.0	13.6	9.3	4.1	2.2	9.2
201x	1.9	3.2	5.2	8.5	12.8	15.8	18.1	18.1	14.4	9.3	4.4	2.0	9.5
202x	2.7	2.8	5.5	8.8	13.3	16.0	18.1	18.2	14.0	9.6	4.6	2.5	9.7
203x	2.4	3.1	5.8	9.0	13.1	16.3	18.6	18.4	14.5	10.0	4.8	3.0	9.9
204x	2.7	3.5	5.1	9.2	13.8	16.8	19.2	19.3	14.8	10.0	4.7	3.0	10.2
205x	2.7	3.6	5.8	9.5	13.3	16.7	19.1	18.9	14.4	10.3	5.2	2.7	10.2
206x	3.2	3.4	6.2	9.5	13.9	17.3	19.9	19.3	15.1	10.7	5.4	3.4	10.6
207x	3.0	3.9	6.0	9.6	14.5	16.7	19.6	19.5	15.3	10.9	5.9	3.6	10.7
208x	3.0	3.5	5.8	10.1	14.4	17.5	19.6	19.0	15.2	10.5	5.5	3.5	10.7
209x	3.1	3.9	6.1	9.7	14.3	17.3	20.0	19.9	16.0	10.9	6.1	3.6	10.9
Total (Average)	2.7	3.3	5.6	9.3	13.6	16.6	19.1	18.9	14.7	10.1	5.1	2.9	10.2

Precipitation (monthly sums average)

Metric	Average precipitation sum [mm]												
month	1	2	3	4	5	6	7	8	9	10	11	12	Total (Average)
slice													
200x	32.1	38.9	44.9	43.9	62.8	53.9	67.5	57.3	48.8	35.6	50.6	50.7	48.9
201x	43.2	38.4	49.0	44.5	55.1	49.3	72.4	56.3	55.6	38.7	64.9	52.8	51.7
202x	39.6	40.1	49.7	41.9	58.6	54.5	84.5	52.8	52.1	40.0	60.1	53.7	52.3
203x	42.3	39.4	45.9	41.0	58.7	53.5	78.6	50.5	59.5	39.1	57.9	61.2	52.3
204x	40.7	41.8	49.8	40.8	55.4	54.7	67.5	55.5	54.0	37.9	61.2	61.5	51.7
205x	46.1	41.7	49.3	41.4	52.5	50.6	77.4	46.1	54.2	47.0	61.6	51.1	51.6
206x	46.6	39.4	42.2	48.5	53.2	54.1	72.2	52.7	51.6	37.1	63.2	56.7	51.5
207x	50.9	37.1	53.7	40.8	57.9	58.4	76.1	53.0	57.6	43.4	63.8	58.1	54.2
208x	45.0	41.7	51.0	40.5	61.5	50.2	78.3	58.7	56.4	36.6	57.0	59.7	53.0
209x	44.3	43.9	55.5	41.7	58.9	55.5	74.5	45.7	45.2	39.3	63.6	60.4	52.4
Total (Average)	43.1	40.2	49.1	42.5	57.5	53.5	74.9	52.9	53.5	39.5	60.4	56.6	52.0

Figure 8:

Forecasts for annual temperature and precipitation for the LVG Erfurt site using the model RCP4.5 with a global warming of 2.6 K by 2100.

Data: UFZ, visualisation: IMMS via Apache Superset.

Outlook in MIRO

In MIRO, a data management concept was developed as a basis for data collection and analyses, which can be used in future to derive well-founded decisions for the cultivation of locally suitable fruit varieties that can cope with climate change, based on climate and soil conditions, among other things.

Based on the Silver data currently available, further evaluations beyond the current site comparison are to be added in the future. To this end, further assessment data is to be automatically written to the Lakehouse, thus enabling images to be analysed (e.g. to verify identified flowering times) and make it possible to correlate yield data. Further sensor systems are also planned to support farms in recording data and to automate the acquisition of additional data as much as possible. The architecture presented enables customised, automated data exchange, even between different actors, through targeted configurations of ingresses and export mechanisms if the corresponding interfaces are known.

The presented lakehouse architecture is not limited to the context of MIRO. By standardising the approach presented, data retrieval and storage can be easily implemented using generic, reusable libraries for common analysis tools and languages (e.g. Python, R). As an S3-compatible object store, MinIO also allows the same tools and libraries to be used for other S3-compatible object stores, meaning that the solution can also be used with cloud storage, for example, without the need for any other adjustments.

Contact persons:
Dipl.-Inf. Marco Götze, marco.goetze@imms.de
Dr.-Ing. Silvia Krug, silvia.krug@imms.de

With support from

Federal Ministry of Food and Agriculture

by decision of the German Bundestag

Project manager

Federal Office for Agriculture and Food

In 2024, the MIRO project was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany. The Federal Office for Agriculture and Food (BLE) provided coordinating support for future farms and future regions as funding organisation, grant number 2822ZR0005.

www.imms.de/miro

Effizientes Zusammenführen heterogener Daten für übergreifende Analysen in der Landwirtschaft und darüber hinaus

In MIRO werden Daten in Form von Zeitreihen erhoben und u.a. für die Forschung am digitalen Zwilling regionaler Obstsorten genutzt. Das IMMS hat hierfür ein Konzept für ein effizientes Zusammenführen heterogener Daten für übergreifende Analysen in der Landwirtschaft entwickelt, das sich auch für andere Anwendungen nutzen lässt. Foto: IMMS.

Motivation und Überblick

www.imms.de/miro

Das IMMS arbeitet im Projekt „Mitteldeutsche Innovationsregion Obstbau“ (MIRO) an Digitalisierungslösungen mit dem Ziel, die Zukunftssicherheit der gesamten Wertschöpfungskette Obst von Anbau, Weiterverarbeitung bis Vermarktung in der Region Mitteldeutschland zu stärken und damit Themen wie ein sich rasch wandelndes Klima sowie eine herausfordernde Fachkräfte- und Wettbewerbssituation in der Landwirtschaft, v.a. im Obstbau, zu adressieren. So ließen sich beispielsweise Probleme bei der Verarbeitung durch angepasste Anbaumethoden vermeiden oder lokal geeignete Obstsorten je nach konkreten Umwelt- und Bodenbedingungen auch in Zeiten des Klimawandels finden.

Zentral für die Digitalisierung ist der Austausch bzw. das Zusammenführen von Daten aus unterschiedlichen Systemen bzw. von unterschiedlichen Akteuren. So werden, wie in vielen anderen Disziplinen auch, in der Landwirtschaft unterschiedlichste Daten benötigt, um effiziente Abläufe im Betrieb und über Betriebe hinweg entlang der Wertschöpfungskette zu gewährleisten. Diese reicht von Erzeugern und Zulieferern

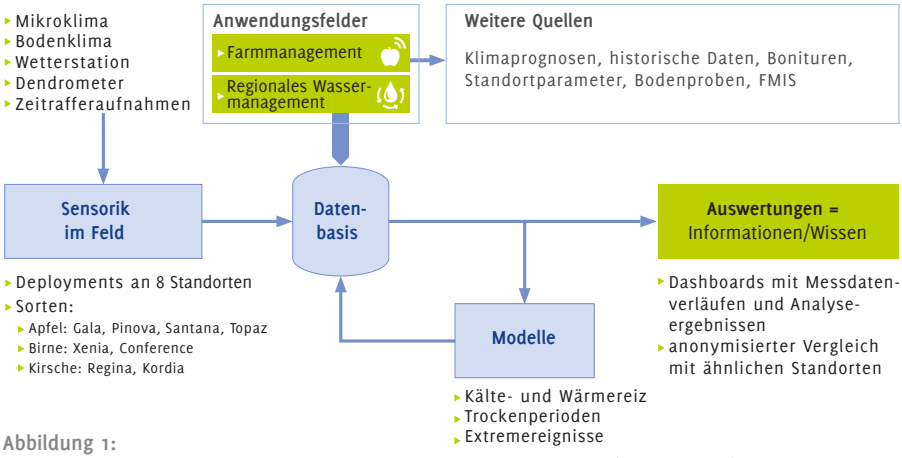


Abbildung 1:
Datenbedarfe und -verarbeitung für den digitalen Zwilling in MIRO (schematisch). Grafik: IMMS.

über Logistik mit Kühllagern sowie Verarbeitern wie Keltereien oder Konservenherstellern hin zu fertigen Produkten wie Saft, Wein oder Apfelmus für Endkunden im Handel. Aber auch andere Interessengruppen wie Pflanzenzüchtung, Baumschulen oder Behörden erwarten zunehmend einen Datenaustausch mit den Betrieben. Dieser wird durch zahlreiche unterschiedliche Systeme, nicht passende, z.B. herstellerabhängige, Schnittstellen und separate Initiativen erschwert. Im Ergebnis müssen Daten mehrfach aufbereitet werden, und das zum Teil mit erhöhtem manuellen Aufwand, wenn Dinge von Hand dokumentiert und ins System übertragen oder Daten aus unterschiedlichen Systemen manuell exportiert und zusammengeführt werden müssen. Solche Aktivitäten sind nicht Kernaufgabe der Landwirte.

Vor diesem Hintergrund hat das IMMS in MIRO zwei Anwendungsfälle bearbeitet – den eben kurz umrissenen Datenaustausch zwischen den Akteuren in der Region und den digitalen Zwilling für Feedback zu Sorteneigenschaften an unterschiedlichen Standorten entlang der Wertschöpfungskette.

www.imms.de/miro

Den Anwendungsfall Datenaustausch hat das IMMS sowohl mit Blick auf die Landwirtschaft als auch auf die beteiligten Projektpartner betrachtet. Zunächst wurden Bedarfe beider Zielgruppen analysiert und dabei das effiziente, idealerweise automatisierte Zusammenführen der Daten und deren Ablage in einer für eine Weiternutzung für verschiedene Zwecke geeigneten Form als Zielstellung identifiziert.

Der Lösungsansatz auf Basis aktuell verfügbarer Plattformen sollte eine einheitliche, zentrale Ablage für alle Arten von strukturierten wie unstrukturierten Daten realisie-

www.imms.de/express

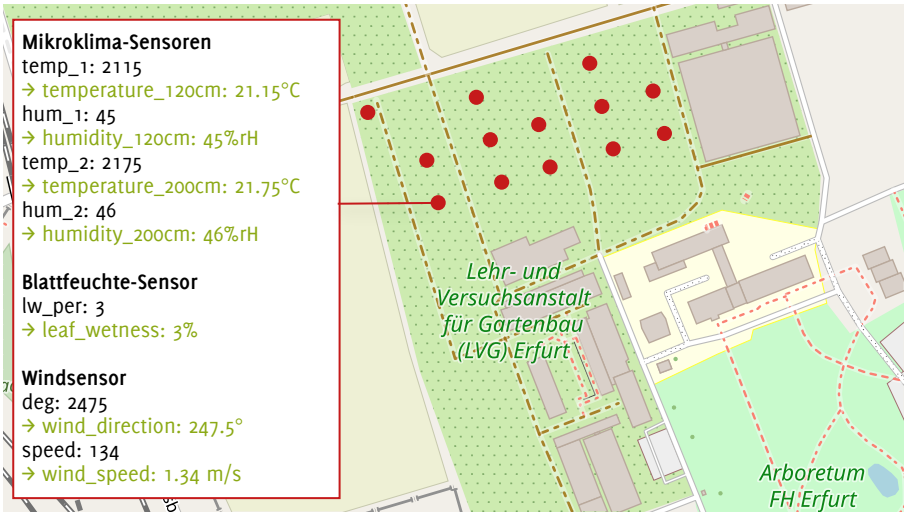
ren, die die Abfrage und Ablage von Daten ebenso erleichtert wie das spätere Auffinden oder Analysieren anhand mehrerer Datenquellen. Dieser Ansatz sollte auch nicht spezifisch auf die Verarbeitung der Daten in MIRO begrenzt sein, sondern vielmehr auch anderen Daten-intensiven Anwendungen, z.B. im KI-Bereich, zugutekommen.

Datenerhebung von Zeitreihen am Beispiel digitaler Zwilling regionaler Obstsorten

Am IMMS werden nicht nur in MIRO, sondern in weiteren Projekten wie z.B. EXPRESS, umfangreiche Daten erhoben, konkret Zeitreihendaten aus verschiedenartigen Sensoren in landwirtschaftlichen Schlägen, Wetterdaten oder auch Bilddaten. Datenquellen sind eigene Sensorikinstallationen wie drahtlose Sensornetze, Wetterstationen oder Wildkameras zur Pflanzenbeobachtung, Partner oder auch öffentliche und kommerzielle Datenbereitsteller im Netz wie der Deutsche Wetterdienst (DWD) oder Sensorikplattformen, von denen Daten über verschiedene Schnittstellen (APIs) bezogen werden können.

Zwar laufen Zeitreihendaten in diesen Projekten bereits systematisch in Zeitreihendatenbanken am Institut ein, jedoch werden sie von Gateways im Feld ohne weitere Aufbereitung geschrieben. Aspekte wie die zeitlich variierende Zuordnung von Messknoten zu Messstellen oder die Zuordnung von Index-referenzierten Messfühlern zu

Abbildung 2: Roh- vs. semantisch aufbereitete Daten für einen Sensorknoten, Verortung der Knoten nur illustrativ). Bei der Aufbereitung erfolgen u.a. die Zuordnung semantischer Bezüge wie der Einbauhöhe/-tiefe und Umrechnungen. Karte erstellt mit MapOSMatic/OCitysMap am 21. Juli 2025. Map styles: Baumkarte by Oliver Rudzick; Allotments overlay; Data source: Kartendaten ©2025 OpenStreetMap.org und Mitwirkende (siehe http://osm.org/copyright).



bestimmten Messtiefen oder -höhen sind erst bei der weiteren Arbeit mit den Daten in Zusammenhang zu bringen. Praktisch heißt das, dass diese Metainformation mit den Daten bei jeglicher Weiternutzung erneut kombiniert werden müssen, was fehleranfällig ist. Außerdem lassen sich Zeitreihendatenbanken, vor allem die genutzte InfluxDB, nicht über die übliche Datenbank-Abfragesprache SQL abfragen, was die Nutzung für Auswertungen erschwert.

Die Arbeiten in MIRO im Anwendungsfall Digitaler Zwilling erfordern eine intensivere Auswertung von Daten und zeigten die notwendigen Ansatzpunkte für die eigenen Datenbestände auf. Hier sollen sensorisch und manuell erhobene Daten wie die von Bonituren Rückschlüsse zur Sorteneignung vor dem Hintergrund des Klimawandels zulassen. Aber auch für fachlich gänzlich anders gelagerte Projekte, in denen Daten zu erfassen und intensiver zu verarbeiten sind, wurde das Potenzial für neue Ansätze deutlich.

www.imms.de/miro

Zeitgemäße Ansätze für große, heterogene Datenbestände

Daher wurde im MIRO-Anwendungsfall Datenaustausch zunächst der Stand der Technik für Datenhaltung im Allgemeinen eruiert. Konzeptuell ist dieser nach den älteren Konzepten Data Warehouse zur effizienten Ablage vereinheitlichter strukturierter Daten und Data Lake zur effizienten Ablage heterogener, einschließlich unstrukturierter Daten beim sog. Lakehouse angekommen, das die Vorteile beider Ansätze zu vereinen versucht.

Für Lakehouses gibt es verschiedene Open-Source-Lösungen, die komplex zu realisieren und nutzen sind und auf Object Stores als Datenspeicher zurückgreifen. Ein Object Store speichert Dateien als Objekte in sog. Buckets. In diesen Objektsammlungen sind sie mit Namen, optionalem Pfad und ggf. weiteren Attributen zu finden. Vertreter von Object Stores sind Cloud-Speicher wie Amazon AWS S3 oder die Open-Source-Lösung MinIO mit dazu kompatibler Schnittstelle. Verbreitete Lakehouse-Lösungen sind im Open-Source-Bereich Apache Iceberg oder Apache Hudi, im kommerziellen Bereich Databricks oder Snowflake. Iceberg und Hudi wurden näher betrachtet, da diese auch in eigenen Instanzen lokal betrieben werden können.

Bei den Betrachtungen ausgewachsener Lakehouse-Lösungen wurde schnell klar, dass diese zum einen leistungsfähige Hardware für den Betrieb auf eigener Infra-

struktur benötigen, wie z.B. Cluster für Spark SQL als Abfrage-Engine. Zum anderen erfordern sie erhebliches Know-how für den Betrieb und in der Nutzung. Letzteres stellt für ein Institut oder andere Betreiber bezüglich der IT-Ressourcen und Schulung von Fachkräften ohne tiefgehendere Datenbankkenntnisse durchaus eine nennenswerte Hürde dar. Allerdings sind Teile der komplexen Lakehouse-Lösungen durch Big-Data-Anwendungen und Fortune-500-Unternehmen motiviert, da deren gesamte Datenbasis darauf abgebildet wird. Die Anforderungen für die Ablage von Messdaten und Auswertungen (WORM, write-once read-many) werden aber bei weitem überstiegen.

Als tragfähige Alternative stellte sich bei Betrachtung aktueller Technologien ein Ansatz heraus, der ein Lakehouse dadurch realisiert, dass Dateien im Apache-Parquet-Format in geeigneter hierarchischer Dateisystemstruktur direkt im Object Store abgelegt werden. Bei Parquet handelt es sich um ein mittlerweile verbreitetes binäres Dateiformat für tabellarische Daten, das speicherplatzeffizient und performant einlesbar ist, typische Probleme bei der Arbeit mit CSV-Dateien (wie unklare Datentypen von Spalten, Varianten der Darstellung von Fehlwerten u.a. vermeidet) und darüber hinaus die Integration von Metadaten erlaubt. Parquet-Dateien in MinIO/S3 können per Apache Spark SQL oder über die In-Prozess-Datenbank DuckDB performant und gleich einer Datenbank per SQL abgefragt werden. Dieser Ansatz vermeidet die Komplexitäten „ausgewachsener“ Lakehouse-Tabellenformate in Nutzung und Betrieb; Einschränkungen diesen gegenüber sind für die betrachteten Nutzungsszenarien unerheblich.

Tschüss CSV, hallo Parquet – Datenhaltung für MIRO leichter und flexibler

Statt Daten von der Quelle, insbesondere Sensorikinstallationen im Feld, direkt ins Lakehouse zu schreiben, hat es Vorteile, eine Zeitreihendatenbank wie die in diesen Projekten bereits genutzte InfluxDB beizubehalten. Diese ist aus dem Feld von ressourcenbeschränkten Geräten wie IoT-Gateways einfacher und leichtgewichtiger anzusprechen und kann bei Bedarf auch anderswo als zentral auf einem Server am IMMS liegen. Die Flexibilität der InfluxDB gegenüber SQL-Datenbanken hat sich dahingehend bewährt, dass kein starres Schema definiert und gepflegt werden muss. Und da es nicht sinnvoll ist, am Gateway bereits alle weiteren Informationen konfigurativ vorzuhalten, die für die angestrebte Datenaufbereitung und Metadatenannotation notwendig sind, ist dieser Schritt nachgelagert und separat besser zu realisieren.

Der Ingress in das Lakehouse geschieht durch geschaffene flexible Werkzeuge nachgelagert periodisch. Dabei werden die Daten aufbereitet (grundlegend gefiltert, umgerechnet) und mit Metadaten (Größen, Einheiten, Standorte etc.) angereichert. Das Ergebnis dieses Prozesses sind Parquet-Dateien mit Bezug zu einzelnen Sensorikinstallationen und bestimmten Zeitbereichen, die in einer durchdachten Organisationsstruktur im MinIO abgelegt und anschließend von dort abgefragt und weiter verarbeitet werden können.

Datenhaltung mit InfluxDB oder anderen Datenbanken

Abbildung 3 zeigt das Konzept im Überblick mit zwei Ebenen der Datenhaltung: Die erste besteht in diesem Fall aus einer InfluxDB-Instanz (eigene Präferenz, andere Lösungen wären hier bei Adaptierung des Ansatzes durch Andere ebenso denkbar) für sensorische Datenerhebungen und weiteren vorhandenen oder extern bereitgestellten Quellen, z.B. anderen Datenbanken oder auch einfach CSV-Dateien. Ein geeigneter Satz an Werkzeugen wird verwendet, um Daten aus dieser ersten Ebene ins Lakehouse zu „heben“. Dabei erfolgt eine erste Aufbereitung und Metadatenannotation. Auf dem Lakehouse können weitere Verarbeitungen erfolgen sowie Visualisierungen

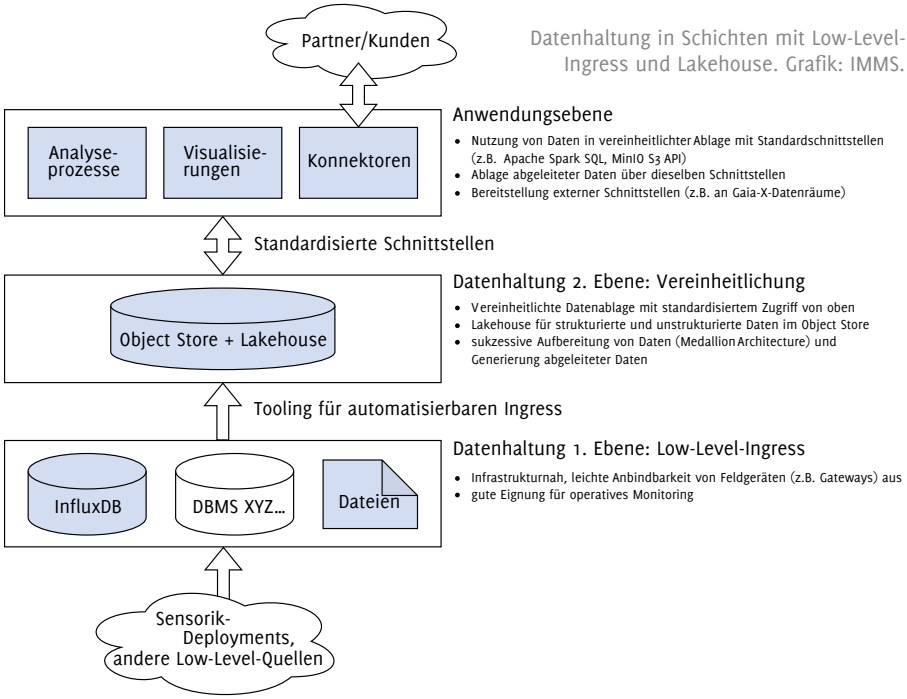


Abbildung 3:

Datenhaltung in Schichten mit Low-Level-Ingress und Lakehouse. Grafik: IMMS.

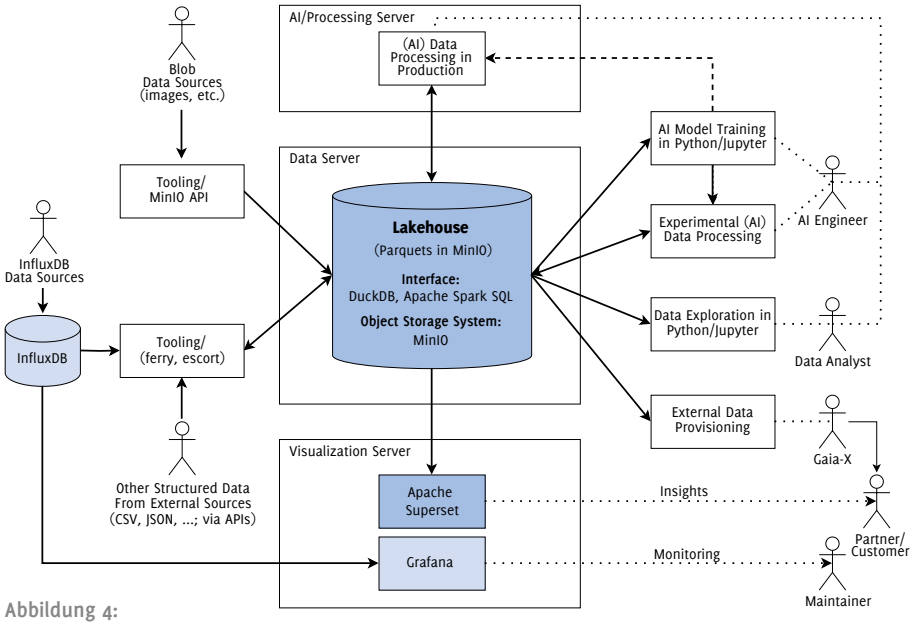


Abbildung 4:

Nachhaltige Datenhaltung per Lakehouse auf Basis von Parquets in MinIO. Grafik: IMMS.

und Schnittstellen oder Datenexporte aufsetzen, wie das für den digitalen Zwilling des regionalen Obstes geplant ist.

Abbildung 4 detailliert das Konzept weiter: Im Kern der Datenhaltung steht das Lakehouse auf Basis von Parquet-Dateien im Object Store MinIO. Im MinIO/Lakehouse gibt es einzelne Buckets je Projekt oder Sensorikfeldinstallation (Deployment). In jedem dieser Buckets gibt es auf der obersten Ebene zwei Verzeichnisse: „warehouse“ für strukturierte Daten, weiter strukturiert nach den Ebenen der sog. Medallion-Architektur mit den Ebenen Bronze, Silver und Gold für unterschiedliche Grade der Verarbeitung, und „blobs“ für anders oder unstrukturierte Daten wie Bilder, Videos oder KI-Modelle.

Automatisierter und kontinuierlicher Datentransfer

Der Ingress, also der Transfer strukturierter Daten in das Lakehouse aus InfluxDB-Instanzen und anderen Quellen, erfolgt automatisiert und kontinuierlich mittels zweier dafür implementierter Software-Komponenten. Diese sind flexibel konfigurierbar und erzeugen Zeitscheiben-Parquets (jeweils für ein Jahr, einen Monat, einen Tag oder eine Stunde). Das Schreiben von Daten in Einzeldateien für kürzere Zeiträu-

Abbildung 5: Monitoring-Dashboard, hier für die Inbetriebnahme eines Mira-Sensornetzes, auf Basis von Grafana und Roh- incl. technischer Metadaten in einer InfluxDB. Quelle: IMMS.

me und Zusammenführen dieser zu größeren Zeiträumen, sobald diese passé sind, vermeidet ein ineffizientes ständiges erneutes Schreiben aller vorliegenden Daten. Beim Erzeugen der Parquets werden diese mit Metadaten auf Basis einer eigenen JSON-Struktur versehen; pro Zeitreihe werden die Metadaten zudem in einer namentlich zugeordneten, separaten JSON-Datei abgelegt, auf die noch leichter zugegriffen werden kann.

Die InfluxDB bleibt die bevorzugte Quelle für ein Monitoring von Sensorinstallationen, da man hier (auf der Ebene von Rohdaten und weiterer technischer Metadaten, die für die eigentliche Datenauswertung uninteressant sind und daher nicht ins Lakehouse überführt werden) Probleme besser nachvollziehen kann. Hierbei hat sich Grafana zur Anzeige in Dashboards bewährt. Für die Visualisierung aufbereiteter Daten im Lakehouse soll zukünftig primär Apache Superset genutzt werden, das im Unterschied zu Grafana auch andere Daten als Zeitreihendaten und Einzelwerte visualisieren und dabei Interaktionsmöglichkeiten bieten kann.

Datenverarbeitung

Die Datenverarbeitung auf dem Lakehouse kann mit verschiedenen Ansätzen erfolgen, Python und R sind gleichermaßen leicht nutzbar, mit Zugriff auf Parquets im Lakehouse via DuckDB oder Spark SQL. Dies gilt für algorithmische Verarbeitungsansätze ebenso wie für KI-basierte. Zudem kann bei der Entwicklung von Verarbeitungen zunächst leicht mit festen beispielhaften Parquets gearbeitet und dann

www.imms.de/miro

durch Anpassungen des Pfads bzw. URLs bei der Produktivsetzung auf „Live-Daten“ umgeschwenkt werden, ohne dass Anpassungen am Code nötig sind. Verarbeitungsergebnisse können ebenfalls einfach als Parquets im Lakehouse abgelegt, während der Entwicklung zur Gegenprüfung aber auch zunächst lokal erzeugt und validiert werden.

Umsetzung für den digitalen Zwilling des regionalen Obstes

In MIRO wurde eine MinIO-Produktivinstanz auf Servern am IMMS aufgesetzt. Auf die Einrichtung eines Apache-Spark-SQL-Clusters wurde verzichtet, stattdessen wird auf DuckDB gesetzt. Neben der MinIO-Instanz laufen Produktivinstanzen der geschaffenen Werkzeuge für automatisierte Daten-Ingresse und -verarbeitungen.

Auf dieser Basis konnten und können für den digitalen Zwilling in MIRO sämtliche bislang und weiterhin in InfluxDB-Instanzen eingehenden Zeitreihendaten einem au-

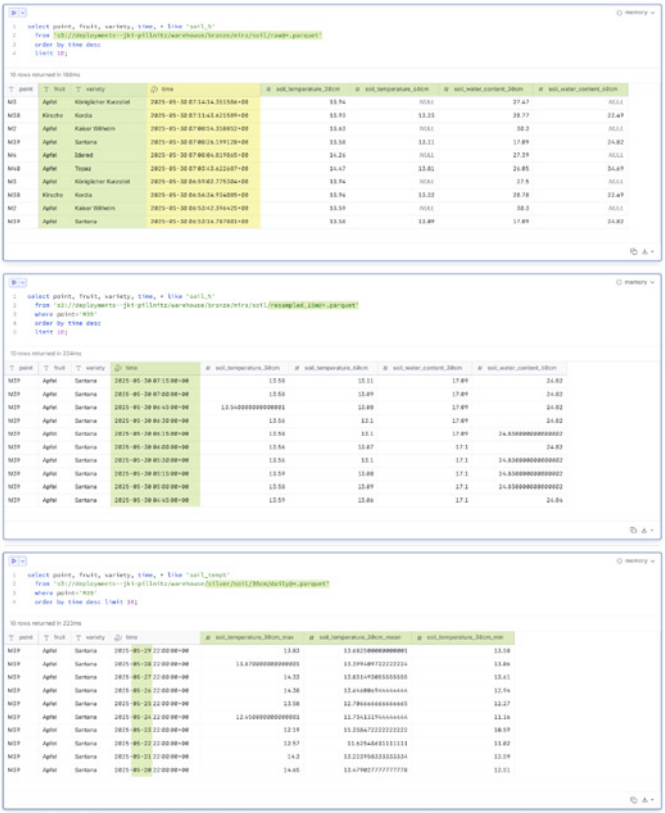


Abbildung 6:

Illustration der Datenaufbereitung anhand von Beispieldaten aus SQL-Abfragen auf dem Lakehouse via DuckDB. Von oben nach unten:

1. Auf Bronze geschriebene, bereits metadatenannotierte Rohdaten verschiedener Messpunkte mit uneinheitlichen/asynchronen Zeitstempeln.

2. Auf Bronze intervallnormierte Daten eines Messpunkts.

3. Auf Silver zu Tagesaggregaten zusammengefasste Werte eines Messpunkts. Zudem ist anhand der URLs in den Abfragen die verwendete Organisationsstruktur im Lakehouse/MinIO ersichtlich.

Grafik: IMMS.

tomatisierten kontinuierlichen Ingress in das Lakehouse zugeführt werden, der die Daten aufbereitet, dabei semantisch zuordnet und mit Metadaten annotiert. Weitere Automatisierungen führen ebenfalls kontinuierlich Intervallnormierungen der typischerweise asynchron erfassten Daten und auch Script-basierte Verarbeitungen per Python und R durch. Damit werden mit geringer Latenz kontinuierlich Daten aller Sensorikinstallationen im Lakehouse bereitgestellt und dort aufgearbeitet. Periodisch erfolgen Verarbeitungen dieser auf der Medallion-Schicht Bronze erzeugten Daten in die Schicht Silver (z.B. durch Aggregationen). Neben strukturierten Daten werden im Lakehouse auch unstrukturierte Daten, wie etwa Bilder oder Firmware-Images als Sicherungen, zentral zugänglich abgelegt.

Abbildung 7: Frostschäden an Kirschblüten am LVG Erfurt, dokumentiert durch Wildkamera-Fotos im April 2024: Blüten vor dem Frost, bei Frost, nach dem Frost. Fotos: IMMS.

In MIRO werden zusätzlich Bilddaten im Lakehouse abgelegt, auf die mit MinIO-/S3-Client-Bibliotheken leicht zugegriffen werden kann und die so bei Datenverarbeitungen einbezogen werden können. Aus der Arbeit mit den Daten resultierende Zwischenergebnisse in Form abgeleiteter Daten können wiederum im Warehouse-Teil des Lakehouses abgelegt werden.

www.imms.de/miro

Neben selbsterfassten werden zusätzlich verschiedene Daten externer Dienste eingebunden und ebenfalls als Parquet-Dateien abgelegt. Dabei werden vor allem Daten vom DWD, dem Helmholtz-Zentrum für Umweltforschung (UFZ), aber auch Daten von verschiedenen Anbietern von Wetterstationen eingebunden, je nachdem, welche Sensorik bei den Partnern bereits vorhanden war oder synergetisch durch Partner im Projekt ausgebracht wurde.

Temperatur (Monatsmittel)

Metric	Durchschnittstemperatur [°C]												
month	1	2	3	4	5	6	7	8	9	10	11	12	Total (Average)
slice													
200x	1.8	2.0	4.4	8.5	12.4	16.0	18.3	18.0	13.6	9.3	4.1	2.2	9.2
201x	1.9	3.2	5.2	8.5	12.8	15.8	18.1	18.1	14.4	9.3	4.4	2.0	9.5
202x	2.7	2.8	5.5	8.8	13.3	16.0	18.1	18.2	14.0	9.6	4.6	2.5	9.7
203x	2.4	3.1	5.8	9.0	13.1	16.3	18.6	18.4	14.5	10.0	4.8	3.0	9.9
204x	2.7	3.5	5.1	9.2	13.8	16.8	19.2	19.3	14.8	10.0	4.7	3.0	10.2
205x	2.7	3.6	5.8	9.5	13.3	16.7	19.1	18.9	14.4	10.3	5.2	2.7	10.2
206x	3.2	3.4	6.2	9.5	13.9	17.3	19.9	19.3	15.1	10.7	5.4	3.4	10.6
207x	3.0	3.9	6.0	9.6	14.5	16.7	19.6	19.5	15.3	10.9	5.9	3.6	10.7
208x	3.0	3.5	5.8	10.1	14.4	17.5	19.6	19.0	15.2	10.5	5.5	3.5	10.7
209x	3.1	3.9	6.1	9.7	14.3	17.3	20.0	19.9	16.0	10.9	6.1	3.6	10.9
Total (Average)	2.7	3.3	5.6	9.3	13.6	16.6	19.1	18.9	14.7	10.1	5.1	2.9	10.2

Niederschlag (Monatssummen-Mittel)

Metric	Durchschnittliche Niederschlagssumme [mm]												
month	1	2	3	4	5	6	7	8	9	10	11	12	Total (Average)
slice													
200x	32.1	38.9	44.9	43.9	62.8	53.9	67.5	57.3	48.8	35.6	50.6	50.7	48.9
201x	43.2	38.4	49.0	44.5	55.1	49.3	72.4	56.3	55.6	38.7	64.9	52.8	51.7
202x	39.6	40.1	49.7	41.9	58.6	54.5	84.5	52.8	52.1	40.0	60.1	53.7	52.3
203x	42.3	39.4	45.9	41.0	58.7	53.5	78.6	50.5	59.5	39.1	57.9	61.2	52.3
204x	40.7	41.8	49.8	40.8	55.4	54.7	67.5	55.5	54.0	37.9	61.2	61.5	51.7
205x	46.1	41.7	49.3	41.4	52.5	50.6	77.4	46.1	54.2	47.0	61.6	51.1	51.6
206x	46.6	39.4	42.2	48.5	53.2	54.1	72.2	52.7	51.6	37.1	63.2	56.7	51.5
207x	50.9	37.1	53.7	40.8	57.9	58.4	76.1	53.0	57.6	43.4	63.8	58.1	54.2
208x	45.0	41.7	51.0	40.5	61.5	50.2	78.3	58.7	56.4	36.6	57.0	59.7	53.0
209x	44.3	43.9	55.5	41.7	58.9	55.5	74.5	45.7	45.2	39.3	63.6	60.4	52.4
Total (Average)	43.1	40.2	49.1	42.5	57.5	53.5	74.9	52.9	53.5	39.5	60.4	56.6	52.0

Abbildung 8:

Prognosen für Jahrestemperatur- und -niederschlagsverlauf für den Standort LVG Erfurt anhand des Modells RCP4.5 mit Erderwärmung um 2,6 K bis 2100.

Daten: UFZ, Visualisierung: IMMS via Apache Superset.

Es wurde bereits eine erste Standortanalyse durchgeführt mit dem Ziel, die Auswirkungen des Klimawandels sichtbar zu machen. Genutzt wurden hierfür bereits Daten des DWD zur historischen klimatischen Entwicklung, Klimaprognosen des Projektpartners UFZ sowie automatisierte Aggregationen von monatlichen Temperatur- und Niederschlagsdaten. Abbildung 8 zeigt eine exemplarische Analyse in Superset.

Ausblick in MIRO

In MIRO wurde ein Datenhaltungskonzept als Grundlage für Datenerhebungen und -analysen erarbeitet, mit denen künftig u.a. auf Basis von Klima- und Bodenbedingungen fundierte Entscheidungen für den Anbau von lokal geeigneten Obstsorten abgeleitet werden können, die dem Klimawandel gewachsen sind.

Auf Basis der aktuell vorhandenen Silver-Daten sollen perspektivisch weitere Auswertungen über den aktuell begonnenen Standortvergleich hinaus ergänzt werden. Dazu sollen auch weitere Bonitur-Daten automatisiert in das Lakehouse geschrieben werden und so die Auswertung der Bilder (z.B. zur Verifikation der identifizierten

Blühzeitpunkte) wie auch die Korrelation von Ertragsdaten ermöglichen. Hier sind zudem weitere Sensorsysteme geplant, um Betriebe bei der Erfassung zu unterstützen und diese möglichst zu automatisieren. Die vorgestellte Architektur ermöglicht durch gezielte Konfigurationen von Ingressen und Exportfunktionen einen angepassten, automatisierten Datenaustausch auch zwischen verschiedenen Akteuren, wenn die entsprechenden Schnittstellen bekannt sind.

Die vorgestellte Lakehouse-Architektur ist dabei nicht nur auf den Kontext von MIRO beschränkt. Durch die Vereinheitlichung des vorgestellten Ansatzes können Datenabruf und -ablage durch generische, wiederverwendbare Bibliotheken für die gängigen Auswertungswerkzeuge und -sprachen (z.B. Python, R) auf einfache Weise realisiert werden. MinIO als S3-kompatibler Object Store erlaubt die Nutzung derselben Tools und Bibliotheken auch für andere S3-kompatible Object Stores und damit ggf. auch eine Nutzung der Lösung auf Basis z.B. eines Cloud-Anbieters, ohne dass sonstige Anpassungen notwendig wären.

Kontakt:
Dipl.-Inf. Marco Götze, marco.goetze@imms.de
Dr.-Ing. Silvia Krug, silvia.krug@imms.de

Gefördert durch
Bundesministerium für Ernährung und Landwirtschaft
aufgrund eines Beschlusses des Deutschen Bundestages

Projektträger
Bundesanstalt für Landwirtschaft und Ernährung

Die Förderung des Vorhabens MIRO erfolgte 2024 aus Mitteln des Bundesministeriums für Ernährung und Landwirtschaft (BMEL) aufgrund eines Beschlusses des deutschen Bundestages. Die Projektträgerschaft erfolgte über die Bundesanstalt für Landwirtschaft und Ernährung (BLE) im Rahmen der Bekanntmachung über die Förderung der Einrichtung von Experimentierfeldern als Zukunftsbetriebe und Zukunftsregionen der Digitalisierung in der Landwirtschaft sowie in vor- und nachgelagerten Wertschöpfungsketten mit dem Förderkennzeichen 2822ZR0005.

www.imms.de/miro