

# moduli: A Disaggregated Data Management Architecture for Data-Intensive Workflows

Paolo Ceravolo

Università degli Studi di Milano

and

Tiziana Catarci AND Marco Console

Sapienza University of Rome

and

Philippe Cudré-Marroux

Université de Fribourg

and

Sven Groppe

University of Lübeck

and

Katja Hose

TU Wien and Aalborg University

and

Jaroslav Pokorný

Charles University

and

Oscar Romero

Universitat Politècnica de Catalunya

and

Robert Wrembel

Poznan University of Technology and Artificial Intelligence and Cybersecurity Center

---

As companies store, process, and analyse bigger and bigger volumes of highly heterogeneous data, novel research and technological challenges are emerging. Traditional and rigid data integration and processing techniques become inadequate for a new class of data-intensive applications. There is a need for new architectural, software, and hardware solutions that are capable of providing dynamic data integration, assuring high data quality, and offering safety and security mechanisms, while facilitating online data analysis. In this context, we propose *moduli*, a novel disaggregated data management reference architecture for data-intensive applications that organizes data processing in various *zones*. Working on *moduli* allowed us also to identify open research and technological challenges.

---

## 1. INTRODUCTION

A data-intensive workflow is a process, composed of a series of tasks, in which the primary focus is on handling and manipulating large volumes of multi-modal data (a.k.a. heterogeneous data), like structured relational records, semi-structured xml/json, graphs, web pages, short and long texts. Such workflows typically involve the acquisition, storage, processing, analysis, and management of substantial amounts of data. Today, machine learning (ML) plays a crucial role in data-intensive workflows by providing tools and techniques to extract valuable insights, patterns, and predictions from large and complex datasets. This impact spans various fields, including: (1) data analysis (utilizing machine learning for classification, regression, clustering, and anomaly detection), (2) performance optimization of tasks and processes, (3) automation of repetitive tasks, (4) enhancement of decision-making processes through data-driven insights, and (5) the personalization of user experiences by recommending

products, content, or services, based on user behavior and preferences.

The most common business areas of machine learning applications include: Recommender Systems, Natural Language Processing, web search engines, E-commerce, User Behaviour Prediction, and Social Media Analysis.

- In **Recommendation Systems** ML is used to analyze current and past user behavior, preferences, and historical data to recommend articles, products, or videos, tailored to individual users.
- Natural Language Processing (NLP)** techniques like Large Language Models (LLMs) empower **chatbots and virtual assistants** for customer support to answer user queries and provide recommendations as well as for well being of elderly people.
- For **Search Engine Optimization (SEO)**, high-ranking keywords need to be identified; moreover, web content needs to be optimized accordingly, i.e., its quality rated and improved to increase search engine visibility.
- In **E-commerce** applications, ML may help to adjust product prices in real-time, based on factors like demand, competition, and historical sales data.
- Tasks of the **User Behavior Prediction** include the prediction of the time various users spent on a given website and the likelihood of a user clicking on an ad or link to optimize ad placements.
- Social Media Analysis** is concerned, e.g., with the identification of social media influencers who are relevant to a brand or product, and with monitoring social media trends and sentiment to inform marketing strategies.

Despite the different purposes, all these applications are data driven, i.e., require data processing workflows to extract, ingest, store, clean, homogenize, integrate, and analyse multi-modal data to generate insights, typically, in the form of dashboards or mathematical models. A typical data science approach to running data processing workflows is based on Python or R scripts run in notebooks [Psallidas et al. 2022]. However, such an approach has a number of clearly identifiable drawbacks, especially for data-intensive workflows, including: poor performance, lack of data safety and security, and no central (master) repository of code and data. Yet, we observe that current practices largely ignore the relevance of data engineering solutions such as database technologies, data integration architectures, data curation techniques, and the optimisation of data processing pipelines [Romero and Wrembel 2020].

In this paper, we highlight the relevance of data management for data-intensive workflows and argue that it must be considered as a first-class citizen to unleash the full potential of any kind of data-driven technologies. We put forward a database management system (DBMS)-centric approach, where data are **systematically managed** and where the main data engineering tasks are **operationalised** to promote good practices among analysts. Combining the concept of a DBMS, i.e., data engineering technologies, with the needed functionalities of data-intensive workflows is however not trivial. As a result, we discuss the need to extend the traditional architecture of a DBMS to accommodate such needs in what we call a **disaggregated DBMS architecture**. Furthermore, we identify the main research and technological challenges to facilitate its adoption by data engineers and data scientists.

We argue that a systematic approach, following a DB-centric approach, is required in order to successfully manage and analyse data regardless of the final objective.

This paper results from:

- a panel discussion entitled *Future Trends in Databases: from Data Science through Artificial Intelligence to Quantum Computing*<sup>1,2</sup>, organized within the 60th anniversary of the *International Federation for Information Processing (IFIP)*<sup>3</sup>
- the research carried out by the IFIP Working Group 2.6 - Databases<sup>4</sup>.

## 2. DISAGGREGATED ARCHITECTURE

A disaggregated DBMS architecture consists of a set of system modules, repositories, and data flows that inter-operate to provide, as a whole, not only the functionality of a DBMS, but also providing means to collect, store, transform, format, homogenize, clean, and integrate highly heterogeneous data. These data will be made available for analytical tasks. Even though architectures of this type have been previously proposed [Amer-Yahia et al.

<sup>1</sup>[https://www.youtube.com/watch?v=Ge6T0Tx\\_4nQ](https://www.youtube.com/watch?v=Ge6T0Tx_4nQ)

<sup>2</sup><https://www.ifipnews.org/ifip60-upcoming-future-information-processing-events/>

<sup>3</sup><https://ifip.org/>

<sup>4</sup><https://www.cs.put.poznan.pl/ifip-wg26/>

2021; Ghosh et al. 2022; Hai et al. 2021], they focus on specific applications. Therefore, a general and comprehensive reference architecture, with a description of its core modules, is still missing. For this reason, we start by introducing a reference architecture, named *moduli* that will help us drive the discussion.

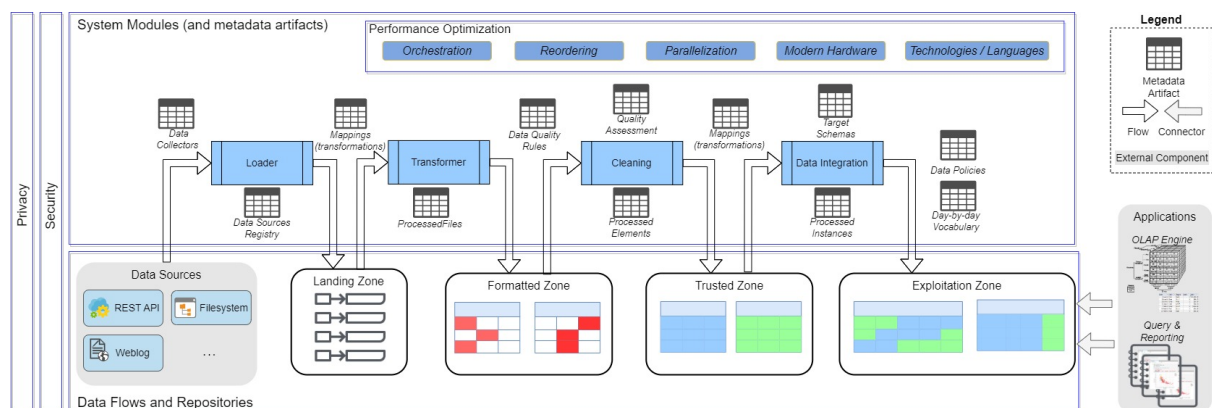


Fig. 1. *moduli*: the data backbone and modules

Figure 1 rethinks the concept of a DBMS within *moduli*. *moduli* is disaggregated by nature (as opposed to single-product DBMS solutions) and spans different types of tools (e.g., scripting, notebooks, NoSQL databases, distributed processing such as Spark) to flexibly accommodate the needs of data-intensive workflows and to serve as a master data repository for organisations.

First, the *loader* module ingests data from data sources via a pull (e.g., API-based) or push (e.g., streaming) mechanism. Each dataset (or its version) is stored as is in the *landing zone*. A key (e.g., data source name plus a timestamp) must be defined to find and fetch a given dataset. The *transformer* module transforms the raw datasets into new datasets following a common data model (the system’s canonical data model) and moves them into the *formatted zone*. In this zone, datasets are homogenized w.r.t. a common data model and format. The *cleaning* module reads the formatted zone, applies data quality rules and moves the data into the *trusted zone*. Since data are common for all organisations, quality rules applied by this module must be general and not specific to a given data analysis. Finally, the *data integration* module integrates the individually processed datasets into single views for the analysts. These views, which reside in the *exploitation zone* are exposed to the analysts and are enriched with the day-by-day vocabulary in the organisation, thus facilitating their understanding and usage.

The separation of concerns is essential in *moduli* to facilitate maintenance and incremental data management, while accommodating specific solutions tackling specific needs of the organisation. Furthermore, metadata must be generated and stored to be later exploited (e.g., for provenance or model explainability purposes).

In addition, throughout the data backbone, optimisation processes are run. They are traversal and dependent on the chosen architecture but generic in nature. Optimisation tasks in *moduli* are, in essence, are the same as conducted in a DBMS but they are challenged by the disaggregated architecture and heterogeneous hardware. Traditional descriptive data analysis techniques (e.g., OLAP) connect with the *exploitation zone*. However, more advanced analytical pipelines require further data management specific to the analysis at hand (as shown in Figure 2). For example, ML algorithm require specific data pre-processing, including the so-called feature engineering [Boeschoten et al. 2023] and vectorization [Villamizar et al. 2023].

Every analytical project must define an analysis backbone. It starts with the *data discovery* module that helps data analysts find relevant data views in the *exploitation zone* for their targeted analysis. Relevant datasets are made available in *sandboxes* where analytical transformations happen. First, the *feature engineering* module generates features (based on available data) and labels the instances, if needed. The generated features are stored in the *feature store*, which is a repository handling the required metadata to understand how a feature has been generated. The *data preparation* module prepares data for the learning phase by applying preparation rules (e.g., specific data imputation or discretisation) and as a result, it generates training and validation datasets. These datasets are used by the *model training* module to learn and validate a model given a set of hyperparameters. Performance metrics are generated and stored as metadata to guarantee traceability.

The analysis backbone must largely automate the data flows and automation of experiments, based on the specific analytical objectives. Also, it must provide out-of-the-box support to run most of these steps using current good

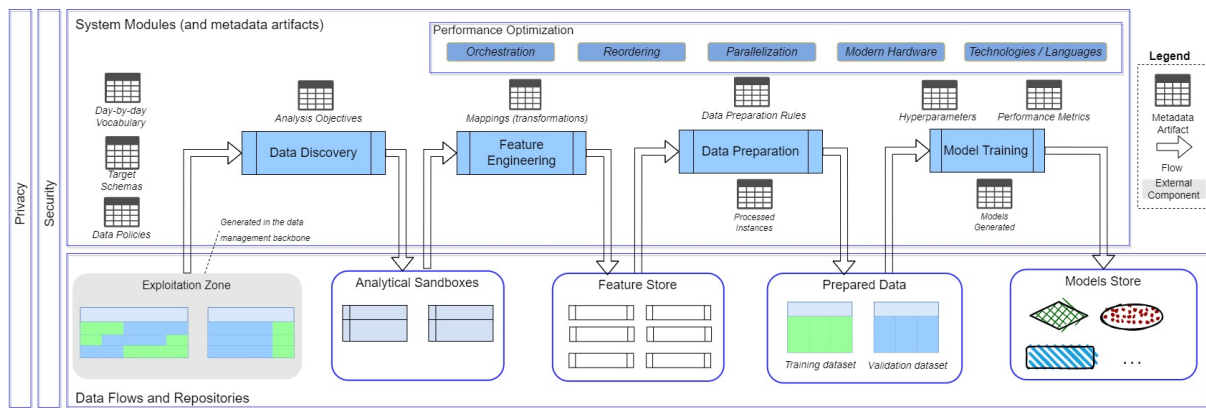


Fig. 2. *moduli*: the analysis backbone and modules

practices (e.g., Python / R notebooks), similar to in-DB-processing currently offered for languages such as Java. Last but not least, the backbone must provide *validation* modules allowing to assess the accuracy of the discovered models using methods that are specific to the studied problem. It must also verify the application of non-biased decisions and the reproducibility of the results.

The presented new backbones open up new research opportunities and expose a number of challenges around *moduli*, which are discussed in the following sections of the paper.

### 3. METADATA MANAGEMENT AND PROVENANCE

*moduli* identifies a set of key metadata artifacts, which are an essential part of the architecture. Most of them are meant to enable first-class citizen data governance. Whereas most research has been focusing on providing *data provenance* for answers of structured queries evaluated over datasets [Glavic et al. 2013; Herschel et al. 2017], provenance can also be used to describe the datasets themselves, independent of the analysis that they are used for.

As we target a trustworthy data management and processing platform, it is important to manage metadata about datasets as well, i.e., data describing the data. The basic information to capture include: (1) when a dataset was uploaded, (2) by whom, (3) in which format, (4) a dataset content. If we also capture information about data ownership and policies about who should have access to which dataset, then we lay the foundations of data governance, which aims at supporting data availability, usability, integrity and security [Stedman 2022].

Driven by new regulations such as GDPR and HIPAA, this point becomes increasingly important. Yet, it remains an open challenge how to support this at scale, especially in the presence of heterogeneous data and mostly undocumented processes going through legacy data workflows.

Often, datasets that are being stored and used are the result of a process, where the original data were extracted (maybe even from a natural language text), cleansed and maybe integrated with other datasets. Capturing metadata about such provenance is vital to enable trustworthiness and explainability. This can, for instance, include who pre-processed the data, which version of an original dataset was used, which version of a particular software, etc.

Such datasets then typically become part of data science pipelines and complex processes to run specific analyses. Data scientists typically select only relevant subsets, integrate them with other datasets – maybe also external data, build and train ML models, and make predictions.

Keeping metadata and provenance about such pipelines is yet another important aspect that forms the foundation of repeatability and explainability; explaining the ML model is one aspect attracting much attention in research at the moment [Dwivedi et al. 2023] but it remains incomplete if we cannot explain what data the model was trained on, how it was pre-processed, and where it came from.

Open challenges stretch from the lack of sufficient standardization, scalability, privacy-compliance, versioning of datasets and software to interactive support of meta-data based search and data exploration. Another line of research tries to learn from available metadata and workflows to help data scientists across different platforms exchange information and learn from each other [Mansour et al. 2021], which eventually leads to better support of AutoML [He et al. 2021], where data scientists are assisted in formulating pipelines.

Yet another unexplored challenge is *structure lineage*, which allows to track dependencies between data structures

in a database (e.g., tables, attributes, views, procedures, functions) along the whole data processing pipeline. Such dependencies are needed to be able to discover which objects are affected by changes made to other objects. The difficulty in this context results from (1) temporal objects (e.g., temporal tables), which cause lineage breaks, (2) a large number of big heterogeneous data sources (e.g., banks integrate hundreds of data sources in one system) to be tracked, (3) dozens or hundreds of thousands of dependencies to be legibly visualized. In standard data science processing pipelines, which are based on files as storage and Python/R as data processing tools, structure lineage is practically impossible.

#### 4. DATA INTEGRATION

Data integration is the last step in *moduli*'s data backbone (see Figure 1) and the module responsible for this step produces the datasets available in the exploitation zone. Classic data integration was mostly considered and designed for relational database systems [Golshan et al. 2017]. The basic principle is to distinguish between a global schema that integrates the data from all the source datasets and the local schemas of the sources themselves, and then define mappings that convert the data and queries (e.g., the GAV, LAV, GLAV approaches). This principle can be applied to Big Data [Dong and Srivastava 2015] with data warehouses and ETL processes, including advanced topics, such as data fusion and duplicate detection [Christophides et al. 2021], as well as to distributed and loosely coupled mediator-wrapper architectures [Wiederhold 1992] in distributed environments. A new data integration architecture has been proposed recently - it is called a *data mesh*. It uses virtual integration, like mediator-wrapper, to build a federation of independent data sources (see for example [Wrembel 2023] for an overview of various data integration architectures).

However, especially in recent years, the requirements have significantly changed. For example, whereas in the past it was often possible to design a global schema (almost) manually, the amount and heterogeneity of data that is being collected in today's use cases go far beyond the scalability of traditional approaches. It is still an open challenge how to integrate the vast amounts of multi-modal data [Stonebraker and Ilyas 2018] that come in different formats, quality, and versions. In general, it is not always possible to design a single global schema that fits all possible applications. Instead, we have witnessed an on-demand style of data integration, where data are integrated when needed and for specific analyses. Although this development goes hand-in-hand with Data Lakes (DL) [Hai et al. 2021; Nargesian et al. 2019a] or Data Lakehouses [Harby and Zulkernine 2022; Tagliabue et al. 2023], which become the standard way of collecting and storing large amounts of heterogeneous data, there are no efficient and advanced solutions yet (going significantly beyond keyword search) that help data scientists find relevant datasets for a specific information need or analysis.

Nevertheless, we argue that the *data integration module* in *moduli* needs to be flexible, as in modern DLs, and it needs to assist data scientists to find and integrate tables and other data objects relevant for a given task. To achieve this flexibility and to efficiently accommodate changes and evolution of the underlying datasets, graph technologies, in particular knowledge graphs (KGs), have become popular – in some cases, entire analytical frameworks can be built upon this paradigm, with graphs being used as the common data format [Cudré-Mauroux 2020; Gu et al. 2022; Nadal et al. 2023; Nath et al. 2022; Noy et al. 2019; Sakr et al. 2021], or as a compact representation of heterogeneous raw data [Mavlyutov et al. 2017]. KGs can also be used to introduce semantics that assists in solving challenges, such as: (1) identifying semantically relevant and related datasets – beyond straightforward string similarity of keywords, and (2) integrating them in a meaningful way by exploiting semantic links between entities across heterogeneous datasets. Still, scaling this principle to automatically integrate thousands of heterogeneous datasets in a general-purpose framework remains an open challenge.

#### 5. DATA QUALITY

Successful data-intensive workflows require high-quality data, as the results they produce strictly depend on the input data [Eppler and Helfert 2004; Haug et al. 2011]. Data quality is still an issue in standard relational data repositories owned by corporations. External data sources being integrated in data lakes (e.g., open data repositories, foras, web portals, social media) provide data of much lower quality (e.g., missing, misspelled, erroneous, contradictory, differently formatted).

Experts agree in identifying high-quality data as *data fit for its intended purpose*. This definition, akin to the notion of quality for business products [Juran and Godfrey 1999] is well-accepted in several fields of application [Kohavi et al. 2004; Karkouch et al. 2016].

Data quality techniques are even more important in data-intensive workflows, which process large volumes of data from multiple and heterogeneous sources [Dong and Srivastava 2013]. The heterogeneity and the lack of a global

structure, render low-quality data even harder to process [Nargesian et al. 2019b]. Whenever possible, data quality issues should be addressed at ingestion time, i.e., before data enters the *loading zone* (see, e.g., [Ceravolo et al. 2018]). Since this is typically not possible or guaranteed, the *data cleaning* module must guarantee the right fit for the intended purpose. Specific algorithms can make some data-intensive workflows *noise robust* [Liu et al. 2021; Mehrabi et al. 2022; Natarajan et al. 2013; Ratner et al. 2016].

To reflect the need for high quality data, *moduli* imposes that data must reach the *trusted zone* only after their quality has been assessed (and possibly improved) by the *cleaning module*. Cleaning and assessment processes must be traced with the corresponding metadata artefacts (see Figure 1).

The *cleaning module* may consist of any combination of data quality techniques that are fit for the data arriving at the *formatted zone*. Usually, these techniques are tailored to work in specific domains [Batini et al. 2009] and may tackle only specific data issues. This extreme specificity makes their results hard to compare [Wang et al. 1995] and generally insufficient for our needs because the data in the *formatted zone* may belong to multiple domains and their quality may depend on cross-domain information.

In this scenario, domain-specific data quality techniques should be used alongside *domain-independent* ones that are capable of describing the bigger picture. Accordingly, multiple efforts have produced frameworks that can accommodate such techniques and make them work in practice [Mezzanzanica et al. 2015; Sessions and Valtorta 2009]. These efforts have been further fostered by the introduction of the ISO 25012 standard that prescribes a unified framework for data quality. Despite the broad interest, the problem of defining data quality techniques that can work effectively across different domains remains still open.

In this context, we foresee the importance of a framework that can provide a unified view of the quality of the data in the *formatted zone*. This framework should be able to reconcile the results of different domain-specific techniques without losing the specificity of their results and foster appropriate cleaning techniques.

To define quality characteristics that remain relevant across different domains, the notion of *data quality dimensions* (from now on simply dimensions) [Batini and Scannapieco 2016] may be used. Intuitively, a dimension is an aspect of the overall quality that can be interpreted, assessed, and possibly improved individually [Ceravolo and Bellini 2019; Sattler 2009]. Dimensions usually come with a corresponding *assessing technique* that defines how the specific characteristics can be measured on data.

In *moduli*, we foresee the use of dimensions as a conceptual tool to define the *data quality rules* applied by the *cleaning module*. Accordingly, an important challenge is to define dimensions and corresponding assessing techniques that are suitable for use in specific data-intensive workflows, while still remaining relevant across different domains. The definition of the set of dimensions relevant to data analytics pipelines is still preliminary [Paggi et al. 2021].

In the context of data-intensive workflows, traditional data quality approaches face additional challenges related to scalability, since data quality techniques are time and resource consuming, and therefore unfit for fast-paced production and consumption. Techniques addressing data quality issues in real-time are necessary, but yet missing. To this end, together with known techniques based on random sampling, we foresee the importance of metadata and semantic data annotations (i.e., providing context) [Führung and Naumann 2007] for data reaching the *formatted zone*.

Finally, as data-intensive workflows are becoming ubiquitous in decision-making processes, data quality techniques should also aim at avoiding or mitigating biases. We thus foresee the importance of data quality techniques that can provide *support for human interaction*. These techniques should augment the conventional automatic data analysis process by providing proactive support throughout the phases of analysis and refinement of the models (i.e., in the analysis backbone), and, if required, propagating data quality issues and solutions backward to the data backbone. All these tasks should be performed iteratively and foster the creation of the *Human in the Loop* model of interaction. To the best of our knowledge, there is no approach supporting these needs yet.

## 6. IN-DATABASE-ML

Reducing the time to market for products is one of the keys to business success. Following practices of continuous software engineering, ML Operations (MLOps) aim to efficiently and reliably deploy and maintain ML models in production [Kreuzberger et al. 2023]. MLOps map to *moduli* analysis backbone (see Figure 2). MLOps refer to the entire lifecycle: from continuous training and evaluation, continuous development and deployment of models, workflow orchestration of ML pipelines to continuous performance monitoring, model improvement, and experiments reproducibility. According to a recent survey [Makinen et al. 2021], companies are on their way to applying

the entire ML lifecycle, resulting in a continuously increasing importance of MLOps.

For many years, research on data management systems offering ML functionalities [Günemann 2017; Park et al. 2022] has been popular, aiming at implementing MLOps in databases. An obvious benefit of combining ML technologies with those of data engineering is to leverage the advanced memory management, data access optimization, safety, security, parallelism, and fault tolerance solutions of databases for ML tasks. For processing ML tasks, data do not need to leave a database if using *in-database-ML*, thus avoiding time-consuming model mappings, data types conversions, and data transfers [Xu et al. 2022], failures of networks as well as privacy and security compliance concerns. In comparison to classical data processing tasks, ML poses additional requirements like: (1) data pre-processing specific to ML algorithms, (2) specific storage of pre-processed data, e.g., vector databases, (3) scalable training phase on large-scale datasets, (4) managing models and their parameters for fast and durable access, (5) out-of-the-box script and query language support for ML tasks as well as (6) automatic selection and tuning ML models.

## 7. PERFORMANCE

Modern hardware solutions are key to improve not only standard analytical processing but also ML-based systems performance. *moduli* acknowledges it by including an optimisation layer in its analysis backbone that should automatically decide when it is appropriate to use different hardware solutions. To the best of our knowledge, such an optimiser has not been considered yet.

Common ML tasks like training and inference heavily use matrix operations and convolution operators [Park et al. 2022]. Hence, most hardware accelerators for ML, such as General Purpose Computation on Graphics Processing Units (GPGPUs) [Park et al. 2022; Nurvitadhi et al. 2016], Tensor Processing Units (TPUs) [Jouppi et al. 2018], Application-Specific Integrated Circuits (ASICs) [Nurvitadhi et al. 2016] and Field-Programmable Gate Arrays (FPGAs) [Groppe 2020; Nurvitadhi et al. 2016; Backasch et al. 2014] specialize on speeding up these kinds of operations. Beyond these widely-used accelerators, new classes of hardware are adding computational capacities to historically passive components. In this context, new generations of smart SSDs [Lee et al. 2022], switches [Lerner et al. 2019], or NICs [Lerner et al. 2022] have recently been leveraged to significantly accelerate computational tasks in modern data ecosystems. Modern data processing architectures, like *moduli* must be able to take advantage of this new hardware plugged into the architecture. This calls for not only new query optimizers, which would utilize this hardware, but also optimizers of the whole processing pipelines.

Utilizing quantum computers for ML tasks is an upcoming trend also for the data management community [Winker et al. 2023; Çalıklarıılmaz et al. 2023]. Quantum ML promises exponential speedups [Rebentrost et al. 2014] and learning on fewer data points than classical methods [Caro et al. 2022]. By looking at the properties of datasets, these datasets that have a potential quantum advantage in learning tasks may be identified [Huang et al. 2021]. Recently, approaches have been studied to optimize join orders in database queries via quantum machine learning [Winker et al. 2023] and quantum annealing [Nayak et al. 2023]. The potentials and challenges of combining quantum computing and data management (called *quantum data management*) inclusive of the integration of quantum ML in *in-database-ML* have been discussed in [Groppe et al. 2022].

Whenever large-scale data needs to be handled by ML tasks (see Figure 1), the memory demands of ML approaches may outpace the system's main memory. Using distributed ML approaches can overcome the limitations inherent in ML approaches that are typically designed for optimal in-memory performance.

For some applications, e.g., healthcare [Ng et al. 2021], data shuffling is prohibited because of data privacy and data security. For these applications, the concept of federated learning [McMahan and Ramage 2017] has been developed, where raw data remain on users' devices for collaboratively training a model. Besides being hands-off and non-invasive, federated learning has further advantages like yielding real-time predictions due to local copies of the model and inherently supporting heterogeneous devices and access to heterogeneous data.

## 8. SUMMARY

In this paper, we discussed the need for a non-silo reference architecture for data-intensive workflows. Such workflows process structured (e.g., row-oriented) and semi-structured (e.g., xml/json, html) data, graphs as well as short and long texts. These workflows typically involve the acquisition of data from legacy internal company systems and from external repositories, like open data repositories and web portals with static data and streamed data. Systems implementing data-intensive workflows are inherently complex and most challenges identified here are due to the lack of a (system-wide) big picture.

To address this missing picture, we proposed a disaggregated architecture and explored the main research gaps and challenges identified in this context, namely: metadata management, multi-modal data integration, techniques for assuring data quality, DBMSs extensions for ML and new software as well as hardware-based techniques to improve performance. Metadata management is the key aspect to provide this system perspective, enriched with appropriate contexts. Inspired by the traditional idea of a DBMS, *moduliis* is a first step towards a reference architecture that deals with end-to-end data management aspects for data-intensive heterogeneous workflows.

## REFERENCES

- AHMADOV, A., THIELE, M., EBERIUS, J., LEHNER, W., AND WREMBEL, R. 2015. Towards a hybrid imputation approach using web tables. In *IEEE/ACM Int. Symposium on Big Data Computing (BDC)*. IEEE, 21–30.
- ALI, S. M. F. AND WREMBEL, R. 2017. From conceptual design to performance optimization of ETL workflows: current state of research and open problems. *The VLDB Journal* 26, 6, 777–801.
- AMER-YAHIA, S., KOUTRIKA, G., BRASCHLER, M., CALVANESE, D., LANTI, D., LÜCKE-TIEKE, H., MOSCA, A., DE FARIAS, T. M., PAPAPOPOULOS, D., PATIL, Y., RULL, G., SMITH, E., SKOUTAS, D., SUBRAMANIAN, S., AND STOCKINGER, K. 2021. INODE: building an end-to-end data exploration system in practice. *SIGMOD Record* 50, 4, 23–29.
- BACKASCH, R., HEMPEL, G., WERNER, S., GROPPE, S., AND PIONTECK, T. 2014. Identifying homogenous reconfigurable regions in heterogeneous fpgas for module relocation. In *International Conference on ReConfigurable Computing and FPGAs (ReConFig)*, Cancun, Mexico.
- BARRENO, M., NELSON, B., JOSEPH, A. D., AND TYGAR, J. D. 2010. The security of machine learning. *Mach. Learn.* 81, 2, 121–148.
- BATINI, C., CAPIELLO, C., FRANCALANCI, C., AND MAURINO, A. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3, 16:1–16:52.
- BATINI, C. AND SCANNAPIECO, M. 2016. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer.
- BLEIHOLDER, J. AND NAUMANN, F. 2009. Data Fusion. *ACM Comput. Surv.* 41, 1.
- BOESCHOTEN, S., CATAL, C., TEKINERDOGAN, B., LOMMEN, A., AND BLOKLAND, M. 2023. The automation of the development of classification models and improvement of model quality using feature engineering techniques. *Expert Systems with Applications* 213, Part, 118912.
- CARO, M. C., HUANG, H.-Y., CERESO, M., SHARMA, K., SORNBORGER, A., CINCIO, L., AND COLES, P. J. 2022. Generalization in quantum machine learning from few training data. *Nat. Commun.* 13, 1.
- CERAVOLO, P., AZZINI, A., ANGELINI, M., CATARCI, T., CUDRÉ-MAUROUX, P., DAMIANI, E., MAZAK, A., VAN KEULEN, M., JARRAR, M., SANTUCCI, G., SATTLER, K., SCANNAPIECO, M., WIMMER, M., WREMBEL, R., AND ZARAKET, F. A. 2018. Big data semantics. *J. Data Semant.* 7, 2, 65–85.
- CERAVOLO, P. AND BELLINI, E. 2019. Towards configurable composite data quality assessment. In *IEEE Conf. on Business Informatics (CBI)*. IEEE, 249–257.
- CHRISTOPHIDES, V., EFTHYMIU, V., PALPANAS, T., PAPADAKIS, G., AND STEFANIDIS, K. 2021. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys* 53, 6, 127:1–127:42.
- CHU, X., ILYAS, I. F., KRISHNAN, S., AND WANG, J. 2016. Data cleaning: Overview and emerging challenges. In *Int. Conf. on Management of Data (SIGMOD)*, F. Özcan, G. Koutrika, and S. Madden, Eds. ACM, 2201–2206.
- CODD, E. F. 1970. A relational model of data for large shared data banks. *Commun. ACM* 13, 6, 377–387.
- CONSOLE, M. AND LENZERINI, M. 2014. Data quality in ontology-based data access: The case of consistency. In *AAAI Conf. on Artificial Intelligence*. AAAI Press, 1020–1026.
- CUDRÉ-MAUROUX, P. 2020. Leveraging knowledge graphs for big data integration: the XI pipeline. *Semantic Web* 11, 1, 13–17.
- DASGUPTA, D. 2021. Delta lake: New hybrid between data lake & data warehouse.
- DONG, X. L. AND SRIVASTAVA, D. 2013. Big data integration. *VLDB Endow.* 6, 11, 1188–1189.
- DONG, X. L. AND SRIVASTAVA, D. 2015. *Big Data Integration*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.
- DURNER, D., CHANDRAMOULI, B., AND LI, Y. 2021. Crystal: A unified cache storage system for analytical databases. *VLDB Endow.* 14, 11, 2432–2444.
- DWIVEDI, R., DAVE, D., NAIK, H., SINGHAL, S., RANA, O. F., PATEL, P., QIAN, B., WEN, Z., SHAH, T., MORGAN, G., AND RANJAN, R. 2023. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.* 55, 9, 194:1–194:33.
- EPPLER, M. AND HELFERT, M. 2004. A classification and analysis of data quality costs. In *Int. Conf. on Information Quality*. MIT, 311–325.
- FARID, M. H., ROATIS, A., ILYAS, I. F., HOFFMANN, H., AND CHU, X. 2016. CLAMS: bringing quality to data lakes. In *Int. Conf. on Management of Data (SIGMOD)*, F. Özcan, G. Koutrika, and S. Madden, Eds. ACM, 2089–2092.
- FRÉNAV, B. AND VERLEYSSEN, M. 2014. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.* 25, 5, 845–869.
- FÜHRING, P. AND NAUMANN, F. 2007. Emergent data quality annotation and visualization. In *Int. Conf. on Information Quality*. MIT, 424–430.
- GHOSH, D., GUPTA, P., MEHROTRA, S., AND SHARMA, S. 2022. A case for enrichment in data management systems. *SIGMOD Rec.* 51, 2, 38–43.
- GLAVIC, B., SIDDIQUE, J., ANDRITSOS, P., AND MILLER, R. J. 2013. Provenance for Data Mining. In *Worksh. on the Theory and Practice of Provenance (TaPP)*.
- GOLSHAN, B., HALEVY, A., MIHAILA, G., AND TAN, W.-C. 2017. Data integration: After the teenage years. In *ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems (PODS)*. 101–106.
- GROPPE, S. 2020. Emergent models, frameworks, and hardware technologies for big data analytics. *The Journal of Supercomputing* 76, 3, 1800–1827.

- GROPPE, S., GROPPE, J., ÇALIKYILMAZ, U., WINKER, T., AND GRUENWALD, L. 2022. Quantum data management and quantum machine learning for data management: State-of-the-art and open challenges. In *EAI Int. Conf. on Intelligent Systems and Machine Learning (EAI ICISML)*.
- GU, Z., LANTI, D., MOSCA, A., XIAO, G., XIONG, J., AND CALVANESE, D. 2022. Ontology-based data federation. In *Int. Joint Conference on Knowledge Graphs (IJCKG)*. ACM, 10–19.
- GÜNNEMANN, S. 2017. Machine learning meets databases. *Datenbank-Spektrum* 17, 1, 77–83.
- HAI, R., QUIX, C., AND JARKE, M. 2021. Data lake concept and systems: a survey. *CoRR abs/2106.09592*.
- HARBY, A. A. AND ZULKERNINE, F. H. 2022. From data warehouse to lakehouse: A comparative review. In *IEEE Int. Conf. on Big Data*. IEEE, 389–395.
- HAUG, A., ZACHARIASSEN, F., AND VAN LIEMPD, D. 2011. The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)* 4, 2, 168–193.
- HE, X., ZHAO, K., AND CHU, X. 2021. AutoML: A survey of the state-of-the-art. *Knowl. Based Syst.* 212, 106622.
- HERSCHEL, M., DIESTELKÄMPER, R., AND BEN LAHMAR, H. 2017. A survey on provenance: What for? What form? What from? *VLDB J.* 26, 6, 881–906.
- HUANG, H.-Y., BROUGHTON, M., MOHSENI, M., BABBUSH, R., BOIXO, S., NEVEN, H., AND MCCLEAN, J. R. 2021. Power of data in quantum machine learning. *Nat. Commun.* 12, 1.
- HUANG, L., JOSEPH, A. D., NELSON, B., RUBINSTEIN, B. I. P., AND TYGAR, J. D. 2011. In *ACM Worksh. on Security and Artificial Intelligence (AISec)*. ACM, 43–58.
- ILYAS, I. F. AND REKATSINAS, T. 2022. Machine learning and data cleaning: Which serves the other? *ACM J. Data Inf. Qual.* 14, 3, 13:1–13:11.
- JOUPPI, N., YOUNG, C., PATIL, N., AND PATTERSON, D. 2018. Motivation for and evaluation of the first tensor processing unit. *IEEE Micro* 38, 3, 10–19.
- JURAN, J. AND GODFREY, A. 1999. *Juran's Quality Handbook*. McGraw Hill.
- KARKOUCH, A., MOUSANNIF, H., MOATASSIME, H. A., AND NOËL, T. 2016. Data quality in internet of things: A state-of-the-art survey. *J. Netw. Comput. Appl.* 73, 57–81.
- KOHAZI, R., MASON, L., PAREKH, R., AND ZHENG, Z. 2004. Lessons and challenges from mining retail e-commerce data. *Mach. Learn.* 57, 1-2, 83–113.
- KREUZBERGER, D., KÜHL, N., AND HIRSCHL, S. 2023. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access* 11, 31866–31879.
- LEE, S., LERNER, A., RYSER, A., PARK, K., JEON, C., PARK, J., SONG, Y. H., AND CUDRÉ-MAUROUX, P. 2022. X-SSD: A storage system with native support for database logging and replication. In *Int. Conf. on Management of Data (SIGMOD)*. ACM, 988–1002.
- LERNER, A., HUSSEIN, R., AND CUDRÉ-MAUROUX, P. 2019. The case for network accelerated query processing. In *Biennial Conf. on Innovative Data Systems Research (CIDR)*. [www.cidrdb.org](http://www.cidrdb.org).
- LERNER, A., JASNY, M., JEPSEN, T., BINNIG, C., AND CUDRÉ-MAUROUX, P. 2022. DBMS annihilator: A high-performance database workload generator in action. *VLDB Endow.* 15, 12, 3682–3685.
- LI, Z., SHARAF, M. A., SITBON, L., SADIQ, S. W., INDULSKA, M., AND ZHOU, X. 2014. A web-based approach to data imputation. *World Wide Web* 17, 5, 873–897.
- LIU, Z., PARK, J., REKATSINAS, T., AND TZAMOS, C. 2021. On robust mean estimation under coordinate-level corruption. In *Int. Conf. on Machine Learning ICML*, M. Meila and T. Zhang, Eds. Vol. 139. PMLR, 6914–6924.
- MAKINEN, S., SKOGSTROM, H., LAAKSONEN, E., AND MIKKONEN, T. 2021. Who needs MLOps: What data scientists seek to accomplish and how can MLOps help? In *IEEE/ACM Worksh. on AI Engineering - Software Engineering for AI (WAIN)*. IEEE.
- MANSOUR, E., SRINIVAS, K., AND HOSE, K. 2021. Federated Data Science to Break Down Silos. *SIGMOD Rec.* 50, 4, 16–22.
- MAURI, L. AND DAMIANI, E. 2022. Estimating degradation of machine learning data assets. *ACM J. Data Inf. Qual.* 14, 2, 9:1–9:15.
- MAVLYUTOV, R., CURINO, C., ASIPOV, B., AND CUDRÉ-MAUROUX, P. 2017. Dependency-driven analytics: A compass for uncharted data oceans. In *Biennial Conf. on Innovative Data Systems Research (CIDR)*. [www.cidrdb.org](http://www.cidrdb.org).
- MCMAHAN, B. AND RAMAGE, D. 2017. Federated learning: Collaborative machine learning without centralized training data.
- MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. 2022. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6, 115:1–115:35.
- MEZZANZANICA, M., BOSELLI, R., CESARINI, M., AND MERCORIO, F. 2015. A model-based approach for developing data cleansing solutions. *ACM J. Data Inf. Qual.* 5, 4, 13:1–13:28.
- MIAO, X., GAO, Y., GUO, S., AND LIU, W. 2018. Incomplete data management: a survey. *Frontiers Comput. Sci.* 12, 1, 4–25.
- NADAL, S., ABELLÓ, A., ROMERO, O., VANSUMMEREN, S., AND VASSILIADIS, P. 2023. Graph-driven federated data management. *IEEE Trans. Knowl. Data Eng.* 35, 1, 509–520.
- NARGESIAN, F., ZHU, E., MILLER, R. J., PU, K. Q., AND AROCENA, P. C. 2019a. Data lake management: Challenges and opportunities. *VLDB Endow.* 12, 12, 1986–1989.
- NARGESIAN, F., ZHU, E., MILLER, R. J., PU, K. Q., AND AROCENA, P. C. 2019b. Data lake management: Challenges and opportunities. *VLDB Endow.* 12, 12, 1986–1989.
- NATARAJAN, N., DHILLON, I. S., RAVIKUMAR, P., AND TEWARI, A. 2013. Learning with noisy labels. In *Annual Conf. on Neural Information Processing Systems (NIPS)*. 1196–1204.
- NATH, R. P. D., ROMERO, O., PEDERSEN, T. B., AND HOSE, K. 2022. High-level ETL for semantic data warehouses. *Semantic Web* 13, 1, 85–132.
- NAYAK, N., REHFELD, J., WINKER, T., WARNKE, B., ÇALIKYILMAZ, U., AND GROPPE, S. 2023. Constructing optimal bushy join trees by solving qubo problems on quantum hardware and simulators. In *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments (BiDEDE)*, Seattle, WA, USA.
- NG, D., LAN, X., YAO, M. M.-S., CHAN, W. P., AND FENG, M. 2021. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quantitative Imaging in Medicine and Surgery* 11, 2, 852–857.

- NORTHCUTT, C. G., ATHALYE, A., AND MUELLER, J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Neural Information Processing Systems Track on Datasets and Benchmarks 1*, J. Vanschoren and S. Yeung, Eds.
- NOY, N. F., GAO, Y., JAIN, A., NARAYANAN, A., PATTERSON, A., AND TAYLOR, J. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8, 36–43.
- NURVITADHI, E., SIM, J., SHEFFIELD, D., MISHRA, A., KRISHNAN, S., AND MARR, D. 2016. Accelerating recurrent neural networks in analytics servers: Comparison of FPGA, CPU, GPU, and ASIC. In *Int. Conf. on Field Programmable Logic and Applications (FPL)*. IEEE, 1–4.
- PAGGI, H., SORIANO, J., LARA, J. A., AND DAMIANI, E. 2021. Towards the definition of an information quality metric for information fusion models. *Comput. Electr. Eng.* 89.
- PARK, K., SAUR, K., BANDA, D., SEN, R., INTERLANDI, M., AND KARANASOS, K. 2022. End-to-end optimization of machine learning prediction queries. In *Int. Conf. on Management of Data (SIGMOD)*. ACM.
- PSALLIDAS, F., ZHU, Y., KARLAS, B., HENKEL, J., INTERLANDI, M., KRISHNAN, S., KROTH, B., EMANI, K. V., WU, W., ZHANG, C., WEIMER, M., FLORATOU, A., CURINO, C., AND KARANASOS, K. 2022. Data science through the looking glass: Analysis of millions of github notebooks and ML.NET pipelines. *SIGMOD Rec.* 51, 2, 30–37.
- RATNER, A. J., SA, C. D., WU, S., SELSAM, D., AND RÉ, C. 2016. Data programming: Creating large training sets, quickly. In *Annual Conf. on Neural Information Processing Systems (NIPS)*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds. 3567–3575.
- REBENTROST, P., MOHSENI, M., AND LLOYD, S. 2014. Quantum support vector machine for big data classification. *Phys. Rev. Lett.* 113, 130503.
- ROMERO, O. AND WREMBEL, R. 2020. Data engineering for data science: Two sides of the same coin. In *Int. Conf. Big Data Analytics and Knowledge Discovery (DAWAK)*. LNCS, vol. 12393. Springer, 157–166.
- SAKR, S., BONIFATI, A., VOIGT, H., IOSUP, A., AMMAR, K., ANGLES, R., AREF, W. G., ARENAS, M., BESTA, M., BONCZ, P. A., DAUDJEE, K., VALLE, E. D., DUMBRAVA, S., HARTIG, O., HASLHOFER, B., HEGEMAN, T., HIDDERS, J., HOSE, K., IAMNITCHI, A., KALAVRI, V., KAPP, H., MARTENS, W., ÖZSU, M. T., PEUKERT, E., PLANTIKOW, S., RAGAB, M., RIPEANU, M., SALIHOGLU, S., SCHULZ, C., SELMER, P., SEQUEDA, J. F., SHINAVIER, J., SZÁRNYAS, G., TOMMASINI, R., TUMEO, A., UTA, A., VARBANESCU, A. L., WU, H., YAKOVETS, N., YAN, D., AND YONEKI, E. 2021. The future is big graphs: a community view on graph processing systems. *Commun. ACM* 64, 9, 62–71.
- SATTLER, K.-U. 2009. *Data Quality Dimensions*. Springer, 612–615.
- SESSIONS, V. AND VALTORTA, M. 2009. Towards a method for data accuracy assessment utilizing a bayesian network learning algorithm. *ACM J. Data Inf. Qual.* 1, 3, 14:1–14:34.
- STEDMAN, C. 2022. What is data governance and why does it matter? <https://www.techtarget.com/searchdatamanagement/definition/data-governance>.
- STEIN, D. 2022. Open sourcing feathr – linkedin’s feature store for productive machine learning.
- STONEBRAKER, M. AND ILYAS, I. F. 2018. Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.* 41, 2, 3–9.
- SURIARACHCHI, I. AND PLALE, B. 2016. Provenance as essential infrastructure for data lakes. In *Int. Provenance and Annotation Worksh. (IPAW)*. LNCS, vol. 9672. Springer, 178–182.
- TAGLIABUE, J., GRECO, C., AND BIGON, L. 2023. Building a serverless data lakehouse from spare parts. In *Workshops at the Int. Conf. on Very Large Data Bases VLDB*. CEUR Workshop Proceedings, vol. 3462. CEUR-WS.org.
- TERRIZZANO, I. G., SCHWARZ, P. M., ROTH, M., AND COLINO, J. E. 2015. Data wrangling: The challenging journey from the wild to the lake. In *Biennial Conf. on Innovative Data Systems Research (CIDR)*.
- VILLAMIZAR, N., WAHRMAN, J., AND VILLASANA, M. 2023. Comparing vectorization techniques, supervised and unsupervised classification methods for scientific publication categorization in the UNESCO taxonomy. In *IFIP WG 12.5 Int. Conf. Artificial Intelligence Applications and Innovations AIAI*. IFIP Advances in Information and Communication Technology, vol. 675. Springer, 356–368.
- WAND, Y. AND WANG, R. Y. 1996. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* 39, 11, 86–95.
- WANG, R. Y., STOREY, V. C., AND FIRTH, C. P. 1995. A framework for analysis of data quality research. *IEEE Trans. Knowl. Data Eng.* 7, 4, 623–640.
- WANG, R. Y. AND STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* 12, 4, 5–33.
- WARD, J. S. AND BARKER, A. 2013. Undefined by data: A survey of big data definitions. *CoRR abs/1309.5821*.
- WIEDERHOLD, G. 1992. Mediators in the architecture of future information systems. *Computer* 25, 3, 38–49.
- WINKER, T., GROPPE, S., UOTILA, V., YAN, Z., LU, J., FRANZ, M., AND MAUERER, W. 2023. Quantum machine learning: Foundation, new techniques, and opportunities for database research. In *Int. Conf. on Management of Data (SIGMOD)*.
- WINKER, T., ÇALIKYILMAZ, U., GRUENWALD, L., AND GROPPE, S. 2023. Quantum machine learning for join order optimization using variational quantum circuits. In *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments (BiDEDE)*, Seattle, WA, USA.
- WREMBEL, R. 2023. Data integration revitalized: From data warehouse through data lake to data mesh. In *Int. Conf. Database and Expert Systems Applications DEXA*. Lecture Notes in Computer Science, vol. 14146. Springer, 3–18.
- XU, L., QIU, S., YUAN, B., JIANG, J., RENGGLI, C., GAN, S., KARA, K., LI, G., LIU, J., WU, W., YE, J., AND ZHANG, C. 2022. In-database machine learning with CorgiPile: Stochastic gradient descent without full data shuffle. In *Int. Conf. on Management of Data (SIGMOD)*.
- ÇALIKYILMAZ, U., GROPPE, S., GROPPE, J., WINKER, T., PRESTEL, S., SHAGIEVA, F., ARYA, D., PREIS, F., AND GRUENWALD, L. 2023. Opportunities for quantum acceleration of databases: Optimization of queries and transaction schedules. *Proc. VLDB Endow.* 16, 9, 2344–2353.

topics, he has published more than 300 papers in international peer-reviewed journals and conferences. Since 2014 he has been chair and subsequently vice-chair of the IFIP 2.6 Working Group on Database. He is a member or chair of several program committees at international conferences. As a data scientist, he was involved in several internationally funded research projects and innovative startups.

**Tiziana Catarci** is full professor in Computer Science and Engineering at Sapienza University of Roma, director of the Department of Computer, Control, and Management Engineering “Antonio Ruberti”. Moreover, she is the Editor-in-Chief of the ACM Journal of Data and Information Quality. Tiziana Catarci’s main research interests are in the hci and database areas and in the intersection between the two. She is recently working on AI ethics as a founder of SIPEIA, the Italian Scientific Society for Ethics in Artificial Intelligence. On these topics she has published over 200 papers in international journals and conferences. In 2020 she has been included in the list of World’s Top 2% Scientists compiled by the Stanford University (<https://data.mendeley.com/datasets/btchxktzyw/2>). In her career she has received many honors, just to cite the last ones: in 2016 she has been included among the “100 Women for Science” project - <http://www.100esperte.it/>. In 2017 she received the Levi Montalcini Association award for the “diffusion of scientific culture among younger generations”. In 2018 she has been included among the “InspiringFifty”, <https://italy.inspiringfifty.org/>, the most influential women in the tech world. In 2021 she received the Women&Tech international award “Le Tecnovisionarie”. Finally, she is very active in combating gender disparities and promoting the STEM disciplines among female students.

**Marco Console** is a tenured assistant professor at Sapienza, University of Rome in the Department of Computer, Control, and Management Engineering “Antonio Ruberti”. His research interests span over several aspects of data management with a special attention to the logical foundations of semantic information integration. His papers are published in several top-ranked conferences and journals of his reference areas; his work on the logical foundations of incomplete SQL databases won the best paper award in KR2018 and was subsequently invited to the Artificial Intelligence Journal. Recently, he served as guest editor for the Special Issue on Quality Aspects of Data Preparation of the ACM Journal of Data and Information Quality.

**Philippe Cudré-Marroux** Full Professor and the Director of the eXascale Infolab at the University of Fribourg in Switzerland. He received his Ph.D. from the Swiss Federal Institute of Technology EPFL, where he won both the Doctorate Award and the EPFL Press Mention in 2007. Before joining the University of Fribourg, he worked on information management infrastructures at IBM Watson (NY), Microsoft Research Asia and Silicon Valley, and MIT. He recently won the Verisign Internet Infrastructures Award, a Swiss National Center in Research award, a Google Faculty Research Award, as well as a 2 million Euro grant from the European Research Council. His research interests are in next-generation, Big Data management infrastructures for non-relational data and AI. Webpage: <https://exascale.info/phil>.

**Sven Groppe** is a Professor at the University of Lübeck, Germany. He received 7 project grants from DFG, BMBF, and BMWi in the area of data management. He is the project coordinator of the BMBF-funded QC4DB project about accelerating relational database management systems via quantum computing. He published more than 150 journal, conference, and workshop papers at top-ranked publication venues including SIGMOD, VLDB, and ICPP with over 155 co-authors from 20 countries worldwide. He is a member of over 110 program committees of various conferences and workshops and a reviewer of over 35 journals. He is a workshop chair of SBD@SIGMOD (2016-2020), BiDEDE@SIGMOD (2021-2023), VLIoT@VLDB (2017-2022) and QDSM@VLDB 2023. He is a general chair of the International Semantic Intelligence Conference (ISIC) (2021-2022), International Health Informatics Conference (IHIC) (2022- 2023), and the International Conference on Applied Machine Learning and Data Analytics (AMLDA) in 2023. More information is available on <https://www.ifis.uni-luebeck.de/~groppe/>.

**Katja Hose** is a full professor of Data Management at TU Wien, databases and AI research unit. Her research is rooted in data and knowledge engineering and spans theory, algorithms, and applications of Data Science including graph databases, knowledge graphs, querying, analytics, and machine learning. As a full professor in the Department of Computer Science at Aalborg University, she has been leading the Data, Knowledge, and Web Engineering group. Prior to joining Aalborg University, she was a postdoc in Gerhard Weikum’s Databases and Information Systems group at the Max Planck Institute for Informatics in Saarbrücken, Germany, and received her PhD in Computer Science from Ilmenau University of Technology, Germany. She has co-authored more than 100 peer-reviewed scientific publications and regularly serves as a reviewer for a broad range of conferences and journals, including TheWebConf/WWW, VLDB, SIGMOD, etc. She has served in many different organizational roles for international conferences, such as VLDB, SIGMOD, EDBT, TheWebConf/WWW, and ISWC, incl. program co-chair roles for EDBT 2023 and ESWC 2021.

**Jaroslav Pokorný** received his Ph.D. degree in theoretical cybernetics from Charles University, Prague, Czechoslovakia, in 1984. He is a full professor in the Faculty of Mathematics and Physics, Charles University in Prague. He has published more than 350 papers and books on data modeling, relational databases, query languages, XML technologies, and data organization. His current research interests include semi-structured data, Web technologies, database architectures, Big Data, and social networks. Jaroslav is involved as a chair and co-chair in the organization of conferences, e.g. ADBIS-DASFAA, EDB, ISD, IDEAS, ADBIS, etc. He is a member of ACM and IEEE. He works also as the representative of Czech Republic in IFIP.

**Oscar Romero** is an associate lecturer at Universitat Politècnica de Catalunya (UPC). He is a member of the Database Technologies and Information Management (DTIM) and his research mainly focuses on complex information systems that automate the data management lifecycle. He has participated in several technology transfer projects with relevant companies or organisations such as the World Health Organisation (WHO), SAP, HP Labs, Siemens, Atos and Zurich Insurance among others. He has published more than 100 papers in peer-reviewed conference and journals.

**Robert Wrembel** (PhD, Dr. Habil.) is an associate professor in the Faculty of Computing and Telecommunications, at Poznan University of Technology (Poland). In 2008 he received a post-doctoral degree in computer science (habilitation), specializing in database systems and data warehouses. He has been a deputy dean of the Faculty of Computing and Management (2008-2012) and the Faculty of Computing (2012-2016). Since Jan 2023 he is the chair of the Data Processing Technologies group at Poznan University of Technology; since May 2023 he is the leader of the Artificial Intelligence and Cybersecurity Center in Poznań. In years 2020-2023 he was leading a project on data quality for the biggest Polish bank. For his work, in 2010 he received the IBM Faculty Award for highly competitive research, in 2011 he was awarded the Medal of the Committee of National Education (from the Minister of National Education), in 2016 - the Silver Medal for Long-lasting Service (from the President of the Republic of Poland), in 2019 - IBM Shared University Research Award, and in 2019 - International Federation for Information Processing (IFIP) Service Award. He is a senior ACM member, a country representative in the IFIP Technical Committee TC 2 - Software: Theory and Practice, and a chair of the IFIP Working Group 2.6 (Database).