SCIENCE CHINA
Information Sciences

• PERSPECTIVE •

May 2026, Vol. 69, Iss. 5, 156101:1–156101:3
https://doi.org/10.1007/s11432-026-4800-x

Entity-centric data management for the ubiquitous computing era

Yanyan SHEN[1], X. Sean WANG[2], Xiaoyong DU[3], Beng Chin OOI[4] & Hong MEI[5*]

[1]*School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China*
[3]*School of Information, Renmin University of China, Beijing 100872, China*
[4]*School of Software Technology, Zhejiang University, Hangzhou 310027, China*
[5]*School of Computer Science, Peking University, Beijing 100871, China*

Received 12 January 2026/Accepted 9 February 2026/Published online 7 April 2026

Citation Shen Y Y, Wang X, Du X Y, et al. Entity-centric data management for the ubiquitous computing era. Sci China Inf Sci, 2026, 69(5): 156101, https://doi.org/10.1007/s11432-026-4800-x

The ubiquitous computing era has arrived, realizing a long-envisioned world where computing weaves itself into the fabric of everyday life and becomes indistinguishable from the reality [1]. At its core, ubiquitous computing represents a deep integration of the human society, the cyberspace, and the physical world into a unified socio-cyber-physical environment. While the major wave of digitalization over the past three decades has focused on connecting people, fundamentally reshaping how individuals communicate and collaborate, the process is rapidly extending toward a universal digitalization of the physical world. This expansion encompasses a massive number of real-world entities, including natural entities (e.g., individuals, devices, and buildings) and social entities (e.g., organizations and communities).

The universal digitalization, featuring an explosive growth in data continuously gathered from the real-world entities, catalyzes two far-reaching visions. The first is the emerging vision of digital twins[1)], which seeks to create digital mirrors that faithfully represent not only the static properties and dynamic states of the entities but also their complex associations and interactions across time and space. The second is data valorization, which aims to fully unlock the value of these rapidly expanding data resources to sustain scientific discovery, industrial innovation, and societal development. Central to realizing both visions is effective data management that establishes a unified and well-organized data infrastructure capable of empowering ubiquitous computing. Specifically, this imposes three fundamentally new requirements on data management, comprising data completeness to ensure faithful representations of entities, long-term usability to sustain data utility over time, and clear data ownership to facilitate trusted usage and equitable value distribution.

For decades, data management technologies have evolved in a requirement-driven manner, ranging from relational databases optimized for online transaction processing to data warehouses and data lakes designed for analytical decision support. As Internet services expanded rapidly, new requirements emerged for handling Big Data, driving the development of NoSQL and NewSQL systems. Regardless of these technological advances, data is typically organized and managed to support specific organizational and application needs. That is, data schema, life-cycles, and access mechanisms are largely dictated by the application logic and business operations. While this model has proven effective for operational needs, it cannot be easily extended for the three following issues.

First, data silos inherently preclude data completeness. Being enterprise-centric, the data of an entity is captured across multiple disconnected organizations. Each organization manages a particular subset of attributes and relationships needed for its specific purpose, with no individual application holding a complete view of the entity. While federated learning and data sharing with zero-knowledge proof have emerged in recent years, they are not well adopted due to data ownership and breach concerns.

Second, data is meaningful only with respect to the intended applications. To a great extent, data is self-contained within the applications for which it is captured, but it lacks a "life" that extends beyond these applications. Its persistence, evolution, and semantics are all constrained by the applications it is tied to. Its existence is highly dependent on the applications, and it vanishes when the applications are no longer supported. However, the entity from which the data is collected may still exist in some form across the databases of many other applications.

Third, data ownership rests with the enterprise, which pays for the creation and maintenance for its business operation. Entities do not own data, despite the fact that the data is created because of their actions.

Let us use the healthcare domain to illustrate the above mentioned problems. Hospitals own the data even when the data is about a patient entity's medical checkups and treatment. Data of a particular patient may be captured in different hospitals and clinics, yet none of which may contain the full history of the patient's health condition. In terms of data ownership, to comply with the data privacy regulations, one-off patients' consent is sometimes

* Corresponding author (email: meih@pku.edu.cn)

1) Digital twins remain primarily visionary without established technological solutions, focusing on creating comprehensive digital representations that faithfully reflect real-world entities.

sought for in order to use their health data on non-medical care purposes such as research. Oftentimes, data is technically owned by the individual care unit. In a general social setting, data created by the users is owned by the business entities that capture the data instead of by the users, and is used to train large models without the consent of creators. The ownership of the data created due to user participation remains unresolved.

With the development of technologies, entity data can be made immutable, consistent across organizations and applications, verifiable, secure, and accessible. In this study, we propose a new entity-centric paradigm to meet the future data management requirements, which aims to organize and manage data with respect to real-world entities. The new paradigm restructures the relationship among data, applications, and real-world entities. Essentially, we propose to anchor the data lifecycle to the entities rather than to the applications. The evolution of data is triggered by the entity, as opposed to the applications. To this end, data ownership has to explicitly bind to the entities, enabling them to establish ownership rights over their own data. Accordingly, applications such as AI-powered analytics and agentic recommendation are classified as "consumers" that must obtain permission from the entities for the use of data.

Rationales and challenges. The entity-centric data management paradigm is driven by the following three key rationales.

• Persistent abstraction. Application logic is subject to frequent modifications driven by shifting business needs and data may therefore exist ephemerally if it is tied to the application. In contrast, real-world entities have significantly longer lifetime. Organizing data around entities instead of applications should lead to a maintainable and resilient architecture that ensures data persistency and completeness, transcending the applications and hence time. This is in line with the historical transition from procedural to object-oriented programming, where organizing operations around stable data abstractions rather than transient procedures has significantly improved software development.

• Sustainable data engineering. By managing data around the entities that generate it rather than around applications, the new paradigm ensures that data persists independently as applications evolve or are replaced, thereby preserving long-term data usability and reducing the cumulative cost of data engineering.

• Digital asset ownership. When data is organized around the associated entities and governed at the entity-level rather than controlled by applications, ownership can be restored to the entities themselves. This enables data owners to determine the use of data and realize the economic potential of their data.

Realizing the entity-centric data management paradigm requires addressing several research challenges regarding data organization, usage, and ownership.

(1) Data organization: How to develop a high-fidelity, evolvable, and self-describing data model for real-world entities?

The primary challenge is to develop a data model that functions as faithful representations of real-world entities rather than disordered collections of data fragments. In practice, data associated with the same entity is generated by multiple systems with different formats. The data model should logically integrate these heterogeneous fragments into a coherent representation, while accommodating diverse information types including static properties, temporal dynamics, event histories, relational structures, and multimedia artifacts. Furthermore, as real-world entities continuously change, the data model should support both state evolution and structural evolution. Such evolutions should be non-destructive, enabling any past state or structure to be reconstructed when needed. Finally, the data model should be self-describing, embedding semantic information such as the meaning and constraints of data items, definitions of relationships, and provenance metadata. Such information enables the entity data to exist independently of specific applications while remaining interpretable over time.

(2) Data usage: How to enable secure, privacy-preserving, and effective utilization of entity data by heterogeneous applications?

Applications interact with entities under varying task objectives, perspectives, and abstraction levels, creating mismatches between the unified entity representations and the diverse application requirements in semantics, granularity, and structure. This requires mechanisms that interpret application data needs in terms of task semantics and desired results, and translate them into precise specifications over the entity data model. These specifications may encompass structured queries, derived features, statistical aggregates, or task-specific computations. Moreover, although an application may access only partial information about an entity, exposing raw data may violate privacy or regulatory constraints. Entity-centric data usage should therefore incorporate controlled access mechanisms that enable applications to obtain task-relevant results without unrestricted access to underlying data. To achieve this, the system should expose data through well-defined methods or operations analogous to object-oriented interfaces. Applications invoke these methods to query, compute over, or learn from the data, while the entity retains control over how the data is accessed, transformed, or revealed.

(3) Data ownership: How to achieve ownership determination, auditable usage guarantees, and flexible governance of data?

The enforcement of data ownership is challenging in three aspects. The first is ownership determination. While data generated by an entity naturally belongs to the entity, determining the ownership of data arising from multi-entity interactions remains complex. The determination should be grounded in verifiable entity identities, ensuring that ownership claims are authentic and enforceable. Such identities have to be unique and persistent, and possibly borderless to handle entity mobility. The second aspect is ownership auditability. Effective enforcement of ownership rights requires the ability to reliably track how data is accessed, processed, and shared over time. This demands trusted provenance and accountability mechanisms that can provide verifiable and tamper-evident records of data usage. The third aspect is ownership governance, which concerns how data owners maintain authority over their data under dynamic and context-dependent access demands. Conventional database authorization mechanisms rely on static and coarse-grained permission assignments to predefined identities. In contrast, entity-centric data ownership requires fine-grained and context-aware governance mechanisms that allow owners to specify usage constraints, access conditions, and delegation policies, enforcing ownership rights while still enabling flexible and legitimate data access by diverse data consumers.

Opportunities. The idea of organizing data around the represented entities is not entirely new. Several research streams have explored this entity-centric or broader data-centric perspective. In earlier times, object-oriented database systems [2] were developed for complex engineering data management, introducing key constructs including object identifiers, encapsulation, and inheritance. Memex [3] and MyLifeBits [4] focused on creating personal data repositories capable of capturing the digital footprints of individuals. Linked data and the semantic web [5] proposed interoperable data objects with the RDF data model, aiming to facilitate machine-readable knowledge representation at scale. The Solid project [6] emphasized decentralized data hosting to restore user ownership via access control lists. While prior efforts share a strong conceptual alignment with our perspective, none have evolved into mainstream due to various practical constraints, e.g., rigid data models that hindered entity evolution, insufficient semantic mapping for diverse application contexts, and absence of scalable trusted infrastructure for ownership governance.

Recent technological developments have significantly altered the situation, making the realization of the entity-centric data management paradigm increasingly feasible.

(1) The data organization challenge involves three aspects, each of which can benefit from existing techniques. The first is flexible methods [7] that avoid rigid schemas. Schema-on-read techniques allow entity data to be stored in its original form and interpreted at query time, enabling evolution as entity attributes and relationships expand or transform. In-situ processing enables computation directly on data in its native format, eliminating costly transformations. Modern systems like MongoDB demonstrate that such flexibility does not necessarily compromise query efficiency.

The second aspect concerns how to preserve complete evolutionary histories. Immutable database technologies [8] treat data as append-only records, enabling entities to maintain complete state change histories and supporting temporal queries that reconstruct entity states at any point in time. They naturally support tracking both state and structure evolution.

The third aspect is about embedding essential semantic information within the data model to ensure entity data remains interpretable. Machine learning approaches offer promising capabilities for automating semantic enrichment, including extracting relationships from raw data, suggesting schema evolutions based on usage patterns, and generating descriptive metadata. Vector embeddings provide an additional layer that supports interoperability by encoding semantic information in a standardized format.

(2) The data usage challenge comprises two aspects that stem from the gap between a unified entity-centric organization and diverse application data requirements. The first aspect involves interpreting application data needs expressed in terms of task semantics, expected outputs, and preferred structures, then mapping them to entity-level representations. Agentic AI systems demonstrate potential in interpreting natural language or semi-structured data specifications, inferring semantic relationships between application concepts and entity attributes, and utilizing task contexts alongside historical usage patterns to produce data access plans.

The second aspect concerns maintaining application-specific data views that remain consistent with the evolving entity states under security and privacy constraints. Traditional view management approaches struggle to handle continuous entity evolution and heterogeneous application data needs, as numerous parameters related to efficiency, scalability, and consistency should be carefully configured. Self-driving data management techniques [9] may address this by learning from recurring data access and transformation patterns to optimize view maintenance strategies adaptively. When combined with privacy-preserving computation [10], they enable secure use of entity data by applications.

(3) The data ownership challenge concerns technical support for tracking data owners, auditing ownership actions, and enforcing ownership-based access controls. The first aspect requires mechanisms to uniquely and verifiably identify entities across systems and domains. Decentralized identity technologies [11] enable entities to maintain persistent and verifiable identities that are not controlled by any single service provider. In multi-party scenarios, ownership may need to be distributed rather than exclusive, necessitating mechanisms that support joint ownership.

For auditability, data provenance and lineage tracking technologies [12] provide systematic recording of data origins, transformations, and access events. When combined with immutable storage mechanisms, these records provide tamper-evident audit trails that support accountability and dispute resolution. Cryptographic techniques such as Merkle tree-based hashing enable efficient verification of provenance and access records, allowing data owners to verify whether their data has been accessed or redistributed in accordance with declared policies. Watermarking techniques can further embed ownership information directly into data.

Ownership governance requires authorization mechanisms that go beyond static identity-based permissions. Attribute-based access control methods [13] are well-suited in this setting, as they enable authorization decisions to incorporate attributes of the requesters, the data objects, the access purposes, and the execution contexts, without transferring the ownership rights. As access demands become increasingly dynamic, learning-based approaches can leverage historical access logs and authorization decisions to automatically refine policies and identify inconsistencies, allowing governance frameworks to evolve with data.

Discussion. While the building blocks discussed above address individual challenges, integrating them into a practical system requires careful consideration of several system-level issues. These include tradeoffs among cost, reliability, and availability in deploying entity data across personal devices or hosted data stores, architectural extensions to existing relational or NoSQL systems for persistent identifiers and temporal tracking, and conflicts between components such as immutable storage versus schema-on-read flexibility, or privacy-preserving computation versus real-time access. Addressing these issues may require middleware layers, standardized protocols, and careful evaluation of trade-offs for specific entity types and deployment contexts.

Summary. The pervasive digitalization demands fundamental research on data management, particularly in terms of data completeness, long-term usability, and ownership. The proposed entity-centric data management paradigm offers a conceptual foundation, where data becomes a first-class asset bound to the real-world entities it represents. While recent technological advances are promising, realizing this paradigm requires moving from isolated capabilities to coherent infrastructure, presenting significant research and engineering opportunities. This study presents a research vision rather than a complete solution. Many technical details require further investigation, and detailed implementation will reveal many challenges we have yet to anticipate. We therefore invite the research community to explore the theoretical and practical dimensions of this paradigm, advancing toward a new horizon for data management in the ubiquitous computing era.

References

1 Weiser M. The computer for the 21st Century. Sci Am, 1991, 265: 94–104
2 Maier D, Jacob S. Development and implementation of an object-oriented DBMS. In: Research Directions in Object-Oriented Database Systems. Cambridge: MIT Press, 1987. 355–392
3 Bush V. As we may think. Atl Mon, 1945, 176: 101–108
4 Gemmell J, Bell G, Lueder R. MyLifeBits: a personal database for everything. Commun ACM, 2006, 49: 88–95
5 Wood D, Zaidman M, Ruth L, et al. Linked Data. Shelter Island: Manning Publications Co., 2014
6 Mansour E, Sambra A V, Hawke S, et al. A demonstration of the solid platform for social web applications. In: Proceedings of the 25th International Conference Companion on World Wide Web, 2016. 223–226
7 Armbrust M, Ghodsi A, Xin R, et al. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the 11th Annual Conference on Innovative Data Systems Research, 2021
8 Yue C, Xie Z, Zhang M, et al. Analysis of indexing structures for immutable data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2020. 925–935
9 Huang S Y, Qin Y Z, Zhang X Y, et al. Survey on performance optimization for database systems. Sci China Inf Sci, 2023, 66: 121102
10 Zhu Y, Wu Y, Luo Z, et al. Secure and verifiable data collaboration with low-cost zero-knowledge proofs. Proc VLDB Endow, 2024, 17: 2321–2334
11 W3C. Decentralized Identifiers (DIDs) v1.0. W3C Recommendation, 2022. https://www.w3.org/TR/did-1.0/
12 Pan B, Stakhanova N, Ray S. Data provenance in security and privacy. ACM Comput Surv, 2023, 55: 1–35
13 Servos D, Osborn S L. Current research and open problems in attribute-based access control. ACM Comput Surv, 2017, 49: 1–45