Rethinking Big Data Scale: Balancing Accuracy, Collaboration, and Storage

Panos K. Chrysanthis
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, USA
panos@cs.pitt.edu

Constantinos Costa
Rinnoco Ltd
Limassol, Cyprus
costa.c@rinnoco.com

***Abstract*—The scale of modern big data is increasingly shaped by collaboration as a result of the constant exchange and sharing of information between organizations, users, and systems. As data collaboration grows, the cost of moving, storing and synchronizing shared datasets has become a major challenge. Existing systems focus on capacity and compression within isolated environments, overlooking the broader need to reduce redundant data movement while keeping results accurate and consistent.**

In this vision paper, we argue that balancing *accuracy, collaboration, and storage* should be at the core of rethinking the big data scale. We propose an approach where systems learn when and what to share, reducing unnecessary transfers through collaborative compression, shared summaries, and intelligent reconstruction. Rather than sending entire datasets, collaborating sites can exchange compact representations that preserve essential information, while AI methods fill in or refine details when needed. By aligning accuracy with collaboration efficiency, future data systems can scale not by storing or moving more data, but by *sharing smarter*. We outline key design ideas for such collaboration-aware systems and discuss open challenges in building intelligent, efficient, and sustainable data infrastructures.

I. Introduction

Today, this growth is driven not only by how much data we create, but by how we *use* and *share* it. Organizations, research groups, sensor networks, and online services constantly exchange and combine information to support joint analytics, artificial intelligence (AI), and decision-making [1], [2]. The rise of the Internet of Things (IoT) has amplified this trend with billions of sensors continuously generating and transmitting data that can fuel AI models, automation, and real-time monitoring [3], [4]. As AI systems demand larger and more diverse datasets, the cost and complexity of storing, moving, and synchronizing shared data across distributed environments have increased dramatically.

As collaboration and network activity increases, the effort required to move, store, and update shared data has become a major challenge. Many copies of the same information are stored across systems and transferred repeatedly between partners or devices. Existing storage and cloud platforms focus on capacity and compression within individual systems, without considering how the same data circulates across multiple systems. This leads to wasted time, cost, and energy [5], and limits the scalability of AI-driven and collaborative environments.

Several research directions have begun to address these challenges from different perspectives. Approaches in distributed and federated learning aim to reduce dependence on centralized data movement by enabling localized training and selective information sharing [6]. Decentralized and blockchain-based architectures explore how data can evolve or decay over time to control growth in collaborative environments [7]. Complementary efforts in adaptive compression and accuracy management [8], [9], [10] demonstrate how systems can balance fidelity and efficiency through dynamic data-aware optimization. Together, these directions signal a broader shift toward intelligent infrastructures that adapt to both workload dynamics and shared data ecosystems. These advances improve efficiency, but they act locally within each system and do not coordinate across collaborators. There is still no unified way for systems to decide *what to share*, *when to share it*, or *how much detail is needed*.

In this vision paper, we argue that the future of big data lies in balancing *accuracy, collaboration,* and *storage*. We propose a new way of thinking about scale, where systems learn to share data more intelligently—exchanging information in a compact manner instead of full datasets and reconstructing details when needed. By focusing on collaboration efficiency rather than replication, we can build data infrastructures that scale by *sharing smarter*, not by storing more.

The remainder of this paper outlines our vision for such systems. We first review existing storage solutions and their limitations (Section II). We then introduce our reimagined paradigm (Section III) and discuss its key design dimensions (Section IV). To ground the vision, we highlight specific mechanisms such as signature-based compression and data postdiction (Section V). We conclude by identifying open research challenges (Section VI) and discussing the broader impact of collaboration-aware and AI-driven storage on the future of big data and IoT systems (Section VII), before closing with a call to action (Section VIII).

II. Background

Big data systems have long relied on distributed frameworks such as MapReduce [11], the Hadoop Distributed File System

(HDFS) [12], and Spark [13]. These platforms introduced scalability and fault tolerance by spreading data across many machines and supporting parallel computation. However, they were designed mainly for batch analytics within single organizations, where data movement and replication were treated as internal processes rather than network-wide concerns. Modern ecosystems are increasingly collaborative and data-driven, where datasets are continuously exchanged across users, cloud services, and connected devices.

A. *Collaboration and Data Exchange*

The rise of data-sharing frameworks, such as data meshes [1] and federated analytics [2], [6] reflects a renewed shift toward decentralized collaboration [14]. In these environments, multiple parties work with different fragments of data that must be combined without excessive replication or loss of control. Federated learning, for instance, allows models to be trained locally and only share updates, reducing central storage needs while still requiring coordination and consistency [6]. Recent studies have also explored practical mechanisms for distributed collaboration, including marketplaces for IoT sensor sharing [15], [16], AI-assisted network management for cooperative operations [17], and federated learning strategies for multi-party environments such as UAV networks [18]. Together, these works highlight the growing need for systems that not only process data efficiently but also share it intelligently and securely across diverse environments, while minimizing redundant movement.

B. *IoT, Telco, and Sensor Networks*

The Internet of Things (IoT), telecommunications infrastructures, and large-scale sensor networks further amplify this challenge. Billions of connected sensors and network elements continuously produce streams of data that feed AI systems for automation, optimization, and monitoring [3], [4]. Telecommunication providers, in particular, face massive collaborative data workloads across edge sites and data centers. Previous research introduced the concepts of *postdiction* and *data decay* for telco analytics, enabling storage systems to reduce volume while preserving predictive value [9], [10]. These studies demonstrated that it is possible to trade strict fidelity for efficiency in continuous, network-intensive environments, which is a core idea of this vision paper. Subsequent work on decay-aware storage for IoT data [7] further confirmed the potential of controlled information loss to maintain scalability as workloads grow, though most current solutions remain static and rule-based rather than adaptive.

C. *Compression, Accuracy, and Adaptation*

Early research on adaptive compression explored how systems could automatically synthesize compression techniques for heterogeneous files, laying the groundwork for adaptive selection strategies [19]. Subsequent work demonstrated that evolutionary optimization could intelligently combine multiple codecs, achieving higher fidelity with reduced redundancy [20]. More recently, machine learning–based predictors have extended this vision by using data characteristics to forecast the most effective compression algorithm dynamically [21]. Compression and approximation techniques remain essential for efficient data management. Systems such as COMPASS [22] apply signature-based *lossless* compression to reduce storage while preserving exact accuracy, whereas postdiction pipelines [23] reconstruct or estimate decayed information when full precision is unnecessary. Recent research has further extended these principles with adaptive, AI-driven data management methods that optimize resource allocation and balance accuracy with performance. Together, these approaches demonstrate that accuracy and efficiency can coexist, yet most operate within specific organizations or domains rather than across federated, multi-party systems. To fully support collaboration, compression and reconstruction must be combined into a coordinated, learning-based process guided by shared policies and workload awareness.

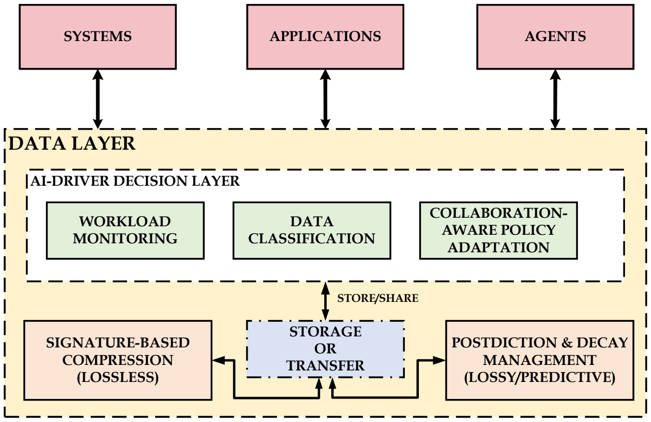


Fig. 1: Conceptual view of the reimagined adaptive collaboration paradigm. The AI-driven decision layer integrates workload monitoring, data classification, and adaptive policy enforcement to balance accuracy, cost, and collaboration.

III. A Reimagined Data Collaboration Paradigm

We envision a new data paradigm in which infrastructures are no longer static repositories, but *adaptive, intelligent, and collaboration-aware systems*. Instead of relying on simple heuristics such as last-access time or fixed replication factors, our approach integrates **AI-driven decision making**, **data classification**, and **collaboration-aware policy adaptation** directly within the data layer, as shown in Figure 1.

This paradigm is built on the observation that the data is not uniform and is rarely used in isolation. Some datasets are *fresh* and intensively explored, others are *shared* across teams, while large volumes become *archived* yet remain valuable for AI retraining or long-term analytics. Existing systems often treat these categories independently, whereas our reimagined system explicitly accounts for **data characteristics** (such as freshness, accuracy requirements, and collaboration frequency) to guide storage, movement, and compression policies.

An AI-driven data layer continuously monitors *workload dynamics*, *data characteristics*, and *collaboration behavior*. Thus, it can recommend and apply strategies such as:

- Keeping frequently co-accessed or evolving data in situ with lightweight compression for interactive, low-latency access.
- Migrating less-shared or stable data into distributed tiers using signature-based *lossless* compression (e.g., COMPASS [22]), achieving significant space savings without sacrificing fidelity.
- Applying decay-aware and postdiction-based reconstruction methods [9], [10], [23] to balance analytical accuracy with long-term storage cost.

This paradigm emphasizes **holistic optimization**, aiming to balance accuracy, cost, and collaboration efficiency rather than treating them as separate concerns. It unifies three key design principles:

1) *Data-driven classification:* understanding workloads and data movement driven by collaboration to inform storage placement and representation.
2) *AI-guided policy selection:* leveraging prediction and postdiction to align data management decisions with workload and sharing patterns.
3) *Continuous adaptation:* dynamically tuning compression, storage, and replication as collaboration and usage evolve.

By embedding these principles directly into the data layer, we move toward systems that no longer treat storage as a passive repository, but as a **smart, dynamic mechanism for data collaboration**. Such infrastructures are cost-efficient, accuracy-aware, and self-adaptive—qualities that are essential for sustainable growth in data-driven ecosystems.

The size of stored data plays a decisive role in collaboration efficiency. When large datasets are replicated or transferred without coordination, network bandwidth becomes a limiting factor, slowing synchronization, and increasing both energy and financial costs. An adaptive collaboration paradigm must therefore manage not only accuracy and access patterns, but also data growth itself, ensuring that shared information remains lightweight, relevant, and network-aware.

IV. Key Design Dimensions

Realizing an adaptive, collaboration-aware data paradigm requires understanding the multiple, interdependent dimensions that shape both system design and operation. In this section, we outline the key aspects that define such systems and guide their evolution toward intelligent, scalable, and efficient collaboration.

A. Data Classification and Workload Awareness

Data does not contribute uniformly to collaboration or decision-making. Datasets differ in freshness, importance, and frequency of use. By dynamically classifying data into categories such as *fresh* (hot), *shared*, *archival* (cold), or *derived*, a system can apply tailored strategies rather than uniform policies. This classification enables more informed decisions about storage, compression, and accessibility, aligning data management with both workload dynamics and collaboration patterns.

B. Accuracy, Fidelity, and Usability

Data accuracy and storage efficiency are inherently conflicting objectives. Exact replication guarantees fidelity but increases cost, whereas aggressive lossy compression reduces storage space by trading off data accuracy. An adaptive system must continuously balance these trade-offs according to context and workload. Techniques such as signature-based *lossless* compression (e.g., COMPASS [22]) and postdiction-based reduction [9], [10], [23] demonstrate that systems can preserve analytical value while managing storage dynamically. Future architectures should expose accuracy as a tunable parameter, enabling collaborative environments that adapt precision to the needs of each user or task.

C. Data Volume and Network Awareness

The volume of stored and exchanged data directly influences the efficiency of collaboration. Large, redundant datasets consume bandwidth, slow synchronization, and increase both energy and financial costs. To remain sustainable, data systems must become network-aware, monitoring transfer patterns, identifying unnecessary movement, and minimizing replication across collaborators. AI-driven prediction models can forecast when data movement is likely to exceed network thresholds and recommend actions such as local summarization, selective caching, or delayed synchronization. By regulating storage growth and network usage together, the system enhances both collaboration responsiveness and environmental efficiency.

D. Collaboration Context and Policy Adaptation

Collaboration patterns evolve continuously as teams, projects, and applications change over time. Some groups may require real-time shared access, while others depend on periodic updates or derived summaries (e.g., Oracle OBISEE, SQL SSRS, Crystal Report, Tableau). Adaptive systems should capture these differences through collaboration metadata, tracking who accesses what, how frequently, and with what latency or accuracy requirements. This contextual information supports policy adaptation mechanisms that can prioritize critical collaborations, defer less urgent ones, or dynamically allocate resources to balance fairness and performance.

E. AI-Driven Prediction and Continuous Learning

At the core of this paradigm lies continuous learning. Machine learning and AI models can classify workloads, predict usage spikes, detect redundant operations, and optimize compression or data placement policies. Instead of relying on static heuristics, the system evolves through feedback, learning from observed collaboration patterns and data flows. Over time, these learning loops enable infrastructures that self-tune, anticipating collaboration needs and proactively balancing accuracy, cost, and resource utilization.

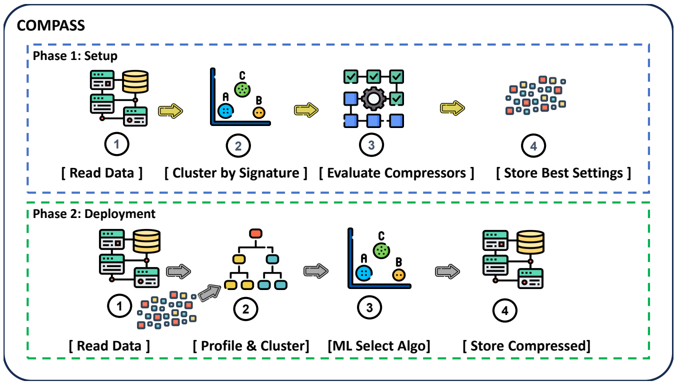


Fig. 2: Illustration of signature-based *lossless* compression (COMPASS). Data columns are represented by signatures, which are generated by data characteristics.

F. Transparency, Fairness, and Trust

As AI becomes integrated in storage and collaboration decisions, transparency, trust, and fairness become essential. Stakeholders must be able to understand how policies are derived, how shared data is prioritized, and how resource allocation impacts participants. Future collaboration-aware systems must therefore incorporate explainable models and auditing mechanisms to ensure that adaptive optimization aligns with human and organizational expectations.

Summary: *Together, these dimensions outline a roadmap for developing intelligent, adaptive, and sustainable data infrastructures. By linking data classification, accuracy management, and network awareness with AI-driven learning and fair collaboration policies, we can advance toward systems that not only store and process information, but also understand how to share it efficiently.*

V. Example Mechanisms

To illustrate how an adaptive and collaboration-aware data paradigm can be realized, we highlight two representative mechanisms: *signature-based compression* and *postdiction-driven decay management*. These techniques demonstrate how intelligence within the storage layer can directly reduce data volume, network load, and coordination cost while maintaining analytical fidelity.

A. Signature-Based Compression (COMPASS)

Our COMPASS framework [22] decomposes relational data into columns and applies the most suitable compression scheme to individual columns or column groups. COMPASS employs K-Means clustering to group *similar* columns based on data values or entropy and then applies the most effective compression technique to each cluster, as shown in Figure 2. The optimal scheme for each column or group is determined empirically by testing multiple combinations.

Empirical Evaluation: Through a series of controlled experiments, we gained practical insights into how adaptive compression performs under real workloads. We used readily available algorithms (i.e., LZMA, BZIP2, and DEFLATE) from the *zipfile* library, previously used in SIBACO. K-Means clustering was applied to group columns after scaling with the *StandardScaler* and encoding string columns with the *OrdinalEncoder* from the *scikit-learn 1.4.2* library.

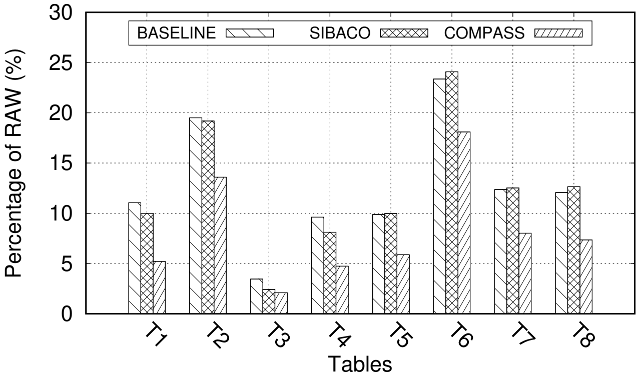


Fig. 3: Disk space usage for the eight largest tables in the Envmon database.

Compared Techniques: Our first experimental series compares four techniques:

BASELINE: Compresses each table using the best single compression algorithm from the *zipfile* library, without considering data characteristics.

SIBACO: Our previous multi-scheme technique [8], which partitions columns into two groups based on entropy and applies the best compression scheme to each group.

COMPASS: The proposed technique, which applies K-Means clustering based on column entropy, achieving lower computational complexity by focusing on one characteristic.

Envmon Dataset: This real-world dataset originates from the Environmental Monitoring Platform, developed through the STEAM project (Sea Traffic Management in the Eastern Mediterranean)[1]. The database contains environmental, meteorological, and oceanographic measurements collected over three years, with a total size of approximately 1 GB.

Experimental Results: Across all tables, COMPASS consistently achieves the highest storage efficiency, reducing disk usage to between 2% and 18.2% of the original RAW size. In contrast, the BASELINE and SIBACO methods reduce disk space requirements to 4.6–23.4% and 4–23.6%, respectively. COMPASS outperforms BASELINE and SIBACO by more than 22% in the worst case and up to 56% (approximately 2×) in the best case, as shown in Figure 3.

B. Data Postdiction and Decay-Aware Reconstruction

The concept of **data postdiction** [9], [10], [23] extends predictive analytics to historical and decayed data. Instead of retaining all original records, systems maintain models that can infer or reconstruct values that have been compressed or discarded.

[1] STEAM: https://steam.cut.ac.cy

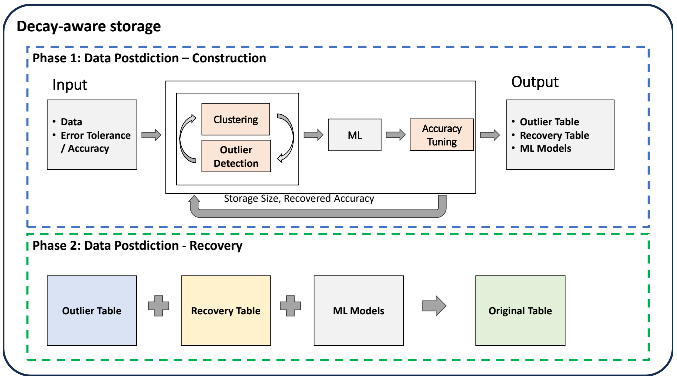


Fig. 4: Pipeline for decay-aware storage using postdiction. As raw data decays, predictive models preserve analytical value by reconstructing or approximating missing information.

Particularly, postdiction operates through a pipeline, as shown in Figure 4, that integrates clustering, outlier detection, machine learning, and accuracy tuning [23]. The process groups similar data instances, isolates and stores *outliers* separately, and trains regression or neural *models* to predict the deleted or decayed attributes. Accuracy tuning enforces a user-defined error threshold, ensuring that reconstructed values remain within acceptable bounds. This approach enables selective, error-bounded recovery using a *recovery table*, eliminating the need to retain the complete *original* dataset.

Empirical Evaluation: The dataset contains one million power consumption records collected from residential environments, with twelve features (six raw and six derived from timestamps). The size of each column is approximately 6 MB. The task involved predicting a single target column representing household power consumption.

Experimental Results: The experiment compared several machine learning techniques for reconstructing decayed data. Table I summarizes the results, including normalized root mean square error (NRMSE), symmetric mean absolute percentage error (sMAPE), model size, and compression ratio. Feedforward Neural Networks (FNN) achieved the best overall balance between accuracy and compactness, while Kernel Attention Networks (KAN) delivered the highest accuracy at a significantly larger size. Decision Trees and Gradient Boosting provided efficient and interpretable midpoints.

These results highlight that lightweight models such as FNNs or Decision Trees can serve as effective postdiction models, preserving analytical quality while drastically reducing data footprint. In additional empirical evaluations conducted on diverse datasets, the results indicate that linear regression models achieve a more favorable trade-off between predictive accuracy and model size reduction (see below). These findings suggest that the selection of an optimal model is inherently data-dependent and can be established through empirical assessment.

TABLE I: Performance of postdiction models on 1M sensor records (12 features). Model size is compared against the raw data size (5641.59 KB).

Model	**NRMSE**	**sMAPE**	**Size (KB)**	**% RAW**
Linear Regression	0.292	31.7%	0.64	0.011
Polynomial Regr.	0.295	31.4%	1.25	0.022
Decision Tree	0.046	5.2%	10.05	0.178
Random Forest	0.045	5.1%	19.74	0.350
Gradient Boosting	0.034	4.2%	102.84	1.823
SVR	0.211	38.0%	36.29	0.643
FNN	0.037	4.9%	8.40	0.149
LSTM	0.221	11.0%	18.16	0.322
KAN	0.032	3.9%	1023.38	18.14
Transformer	0.040	5.2%	4263.74	75.58

Overall, postdiction transforms data storage from a passive capacity problem into an active process of *semantic compression* and intelligent reconstruction, aligning storage growth with collaboration efficiency and sustainability goals.

C. Synergy Between Mechanisms

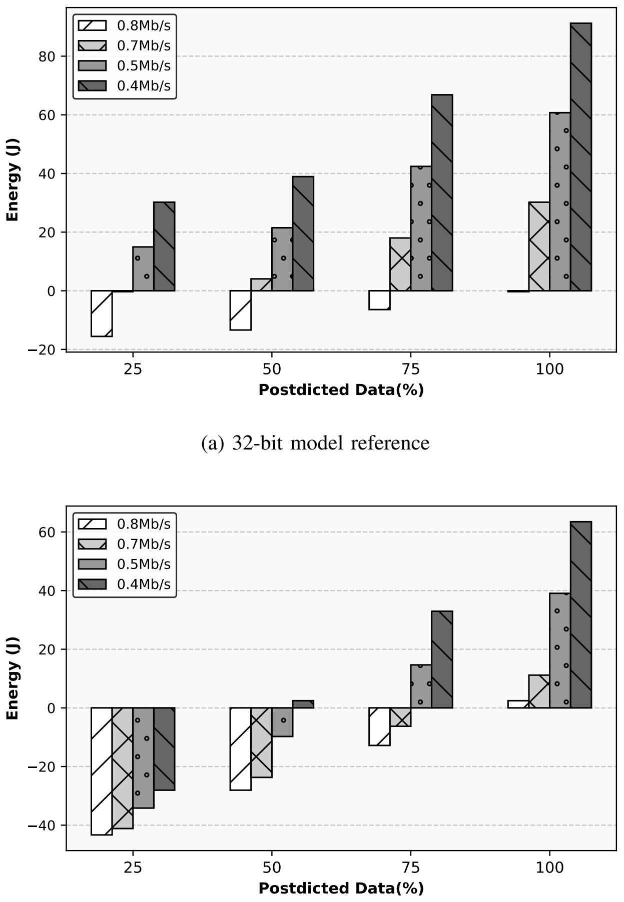
Both mechanisms address complementary aspects of collaboration efficiency. Signature-based compression preserves fidelity through compact representations, while postdiction manages data decay and inference when approximation is acceptable. Together, they establish a foundation for adaptive systems that can continuously balance accuracy and efficiency according to workload, network conditions, and collaborative context.

When guided by AI-driven decision layers, these mechanisms enable infrastructures capable of adaptively minimizing data replication and transfer, optimizing the trade-off between accuracy and cost, and sustaining performance in the face of growing collaborative scale and complexity.

In large-scale, data-intensive collaborative environments, three predominant bottlenecks are storage capacity, bandwidth utilization, and communication latency. Within a smart-sharing paradigm, collaborating parties dynamically adapt their data exchange strategies by selecting among raw datasets, compressed datasets, or compact models, contingent upon contextual and resource constraints.

Empirical Evaluation: The following experimental evaluation illustrates the influence of resource constraints on data exchange dynamics in a sensor-based collaborative framework. In particular, we investigate the network conditions that determine whether raw sensor readings or postdicted sensor readings are transmitted to reduce energy cost.

Setup: We implemented a C++ postdiction pipeline that applies linear regression models to synthetic sensor data. In this setup, sensor readings are replaced with compact model references (4–7 bits), each representing one of several pretrained regression models. This approach significantly reduces transmission size compared to sending 32-bit raw values. The system quantifies both computational energy consumption (associated with regression inference) and transmission energy consumption (associated with wireless data upload via a Raspberry Pi 5).



(a) 32-bit model reference

(b) 64-bit model reference

Fig. 5: Energy efficiency for regression models with different postdiction accuracy. Postdiction achieves net energy savings when network bandwidth is below approximately 0.8 Mb/s for 32-bit models and 0.4 Mb/s for 64-bit models.

At full CPU utilization, the Raspberry Pi 5 consumes approximately 8 W, while the Wi-Fi module draws about 0.8 W. The energy trade-off is beneficial when the time saved in transmission exceeds the computation overhead.

For a dataset containing one million records and four features, postdiction achieved substantial storage savings; however, these savings do not necessarily translate into energy efficiency when the data is shared externally rather than stored locally.

Experimental Results: Experimental results (Figure 5) indicate that the energy efficiency of postdiction is strongly influenced by dataset size and network throughput. In scenarios where 32-bit models achieved a 50% prediction accuracy, energy savings were realized when the upload rate was below 0.8 Mb/s; for 64-bit models, the corresponding threshold declined to 0.4 Mb/s. IIn both scenarios, energy savings are substantial when the models achieve 100% prediction accuracy, regardless of the available bandwidths. Together, compression and postdiction techniques offer complementary strategies for adaptive, collaboration-aware data management.

1) *compressed aggressively (lossless)* when fidelity is required,
2) *decayed gracefully (lossy) when approximation is acceptable*, and
3) *reconstructed intelligently*, either through selective decompression or AI-guided postdiction.

The synergy of these two techniques demonstrates how storage intelligence can be integrated across both compression and inference, enabling a dynamic trade-off among fidelity, cost, and communication efficiency.

This balance transforms storage from a passive repository into an active component of collaboration, capable of reasoning about accuracy, cost, and connectivity in real time.

VI. Research Challenges and Open Questions

While signature-based compression and data postdiction illustrate the potential of adaptive, collaboration-aware storage, several open challenges remain before such systems can be deployed at scale. Real-world data ecosystems are highly dynamic, heterogeneous, and decentralized, making it difficult to maintain consistency, efficiency, and fairness across collaborators. This section highlights the key technical and organizational questions that must be addressed to advance the vision of intelligent, sustainable data infrastructures.

A. Modeling and Predicting Collaborative Workloads

Adaptive systems rely on accurate workload models to perform effectively. However, collaborative data environments combine diverse activities, ranging from interactive exploration to automated pipelines and continuous streaming, making it difficult to maintain predictive accuracy. Anticipating when and where data will be produced, accessed, or shared remains an open problem. Future research should focus on developing hybrid models that integrate statistical analysis, AI-based prediction, and feedback from real usage patterns to better anticipate collaboration dynamics and guide adaptive resource allocation.

B. Balancing Fairness, Transparency, and Adaptivity

Embedding AI-driven decision layers into storage systems introduces new concerns related to fairness, transparency, and accountability. Collaborators often have differing priorities, such as cost, latency, or data ownership, that must be respected even as the system self-optimizes. Future research should focus on developing explainable policies, auditable adaptation mechanisms, and cooperative governance frameworks to ensure that adaptive storage systems remain trustworthy, equitable, and aligned with human intent.

C. Integration Across Heterogeneous Infrastructures

Collaborative data spans increasingly multiple cloud providers, edge devices, and IoT networks. Each environment presents distinct cost, latency, and reliability profiles. Achieving end-to-end adaptivity requires interoperable standards for data classification, model exchange, and compression. Open challenges include designing protocols that maintain efficiency without sacrificing portability or security.

D. Energy Efficiency and Environmental Sustainability

The sustainability impact of large-scale data collaboration is often underestimated. As shown in our experiments, energy consumption depends not only on computation and storage but also on network transfer costs. Future systems must incorporate energy-awareness into their optimization logic, balancing accuracy, responsiveness, and environmental impact simultaneously.

E. Toward Fully Adaptive Data Ecosystems

Finally, a critical research direction involves extending adaptivity beyond the storage layer to the full data lifecycle, integrating query processing, caching, visualization, and AI model retraining. A unified framework for adaptive data ecosystems would allow storage, computation, and learning components to co-evolve, driven by continuous feedback and shared optimization goals across collaborators.

VII. Vision and Impact

The vision outlined in this paper redefines the foundation of data scalability. Rather than treating growth as a matter of accumulating storage or expanding computation, we argue that the future of big data depends on how intelligently systems can *share, adapt, and collaborate*. By aligning accuracy with collaboration efficiency, we can build infrastructures that scale not by storing more data, but by *sharing smarter*.

Our approach brings together several key ideas: adaptive storage guided by AI agents, collaboration-aware compression and reconstruction, and policies that evolve with data lifecycles. Together, these components form a unified framework for balancing accuracy, cost, and responsiveness across distributed participants. This shift changes the role of storage from a passive, static layer to an active, learning participant in the data ecosystem.

The impact of this paradigm extends across domains. In large-scale scientific collaborations, it can reduce redundant transfers and accelerate the sharing of results. In telecommunications and IoT, it can regulate network pressure and energy consumption by minimizing unnecessary data movement. For AI-driven analytics, it can maintain model accuracy while reducing the cost and delay of feeding distributed training pipelines. Across all these domains, intelligent sharing mechanisms enable a new level of sustainability—reducing the environmental and economic footprint of data growth.

Beyond immediate technical benefits, this vision also repositions collaboration as a first-class concept in data system design. Future infrastructures will not simply store information but understand its *value in context*: who needs it, how it evolves, and what level of precision is necessary at any moment. Embedding such intelligence into the data layer, whether this is a database, data warehouse, or data repository or lake, will allow systems to self-regulate, anticipating workload shifts, and optimizing collaboration without manual intervention.

Ultimately, realizing this vision requires interdisciplinary effort—connecting advances in databases, AI, networking, and systems architecture. But its potential is transformative: infrastructures that are leaner, smarter, and more equitable in how they share knowledge. By balancing accuracy, collaboration, and storage, we move toward a data future that is not only larger and faster, but also more sustainable, adaptive, and intelligent.

VIII. Conclusion

In this paper, we presented a vision for rethinking big data scalability through adaptive, collaboration-aware storage. We argued that future infrastructures must move beyond static, capacity-driven designs toward systems that learn, adapt, and coordinate across distributed participants. By embedding intelligence within the storage layer, data systems can reason about accuracy, cost, and workload dynamics, balancing these factors in real time to support efficient collaboration.

We introduced two complementary mechanisms, signature-based *lossless* compression and postdiction-driven decay, that illustrate how accuracy and efficiency can coexist. These mechanisms, combined with AI-driven decision layers, demonstrate the feasibility of infrastructures that optimize data sharing and storage holistically, reducing redundancy and enabling sustainable collaboration at scale.

Realizing this vision will require bridging advances across data management, artificial intelligence, and network systems. Doing so can redefine how organizations exchange, process, and preserve knowledge—building infrastructures that scale not by storing more, but by *sharing smarter*. Through this shift, big data systems can become not only larger and faster, but also more sustainable, adaptive, and equitable in how they serve global collaboration.

Acknowledgment

The authors would like to thank the students of the ADMT Lab of the University of Pittsburgh, specifically Brian T. Nixon, Kartik Hans and Vasilis E. Sarris, and the interns at Rinnoco Ltd, namely Spyros Sachmpazidis and Stelios Mavris, both students at the University of Cyprus. All have contributed to the discussions, experiments, and prototypes that supported this work.

This research was supported in part by the National Institutes of Health (NIH) under NIH award R01HL159805, and under the programme of social cohesion “THALIA 2021-2027” co-funded by the European Union through the Research and Innovation Foundation (RIF). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the supporting organizations.

References

[1] A. Goedegebuure, I. Kumara, S. Driessen, W.-J. Van Den Heuvel, G. Monsieur, D. A. Tamburri, and D. D. Nucci, "Data mesh: A systematic gray literature review," *ACM Comput. Surv.*, vol. 57, no. 1, Oct. 2024. [Online]. Available: https://doi.org/10.1145/3687301

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Nitin Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, *Advances and Open Problems in Federated Learning*. Hanover, MA, USA: Now Publishers Inc., Jun. 2021, vol. 14, no. 1–2. [Online]. Available: https://doi.org/10.1561/2200000083

[3] W. Khan, M. Rehman, H. Zangoti, M. Afzal, N. Armi, and K. Salah, "Industrial internet of things: Recent advances, enabling technologies and open challenges," *Computers & Electrical Engineering*, vol. 81, p. 106522, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790618329550

[4] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013, including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services & Cloud Computing and Scientific Applications — Big Data, Scalable Analytics, and Beyond. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X13000241

[5] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012, special Section: Energy efficiency in large-scale distributed systems. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X11000689

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3298981

[7] P. Drakatos, C. Costa, A. Konstantinidis, P. K. Chrysanthis, and D. Zeinalipour-Yazti, "A blockchain datastore for scalable iot workloads using data decaying," *Distrib. Parallel Databases*, vol. 42, no. 3, p. 403–445, May 2024. [Online]. Available: https://doi.org/10.1007/s10619-024-07441-9

[8] C. Costa, P. K. Chrysanthis, M. Costa, E. Stavrakis, and N. Nicolaou, "Towards a signature based compression technique for big data storage," in *39th IEEE International Conference on Data Engineering, ICDE 2023 - Workshops, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2023, pp. 100–104. [Online]. Available: https://doi.org/10.1109/ICDEW58674.2023.00022

[9] C. Costa, A. Charalampous, A. Konstantinidis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Decaying telco big data with data postdiction," in *19th IEEE International Conference on Mobile Data Management, MDM 2018, Aalborg, Denmark, June 25-28, 2018*. IEEE Computer Society, 2018, pp. 106–115. [Online]. Available: https://doi.org/10.1109/MDM.2018.00027

[10] C. Costa, A. Konstantinidis, A. Charalampous, D. Zeinalipour-Yazti, and M. F. Mokbel, "Continuous decaying of telco big data with data postdiction," *GeoInformatica*, vol. 23, no. 4, pp. 533–557, 2019. [Online]. Available: https://doi.org/10.1007/s10707-019-00364-z

[11] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, p. 107–113, Jan. 2008. [Online]. Available: https://doi.org/10.1145/1327452.1327492

[12] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 2010, pp. 1–10.

[13] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. USA: USENIX Association, 2010, p. 10.

[14] A. Bonifati, P. K. Chrysanthis, A. M. Ouksel, and K.-U. Sattler, "Distributed databases and peer-to-peer databases: past and present," *SIGMOD Rec.*, vol. 37, no. 1, p. 5–11, Mar. 2008. [Online]. Available: https://doi.org/10.1145/1374780.1374781

[15] D. Georgakopoulos and A. Dawod, "Ubiquitous iot sensor sharing via an open self-manged marketplace," in *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, 2024, pp. 29–34.

[16] A. Dawod, D. Georgakopoulos, P. P. Jayaraman, J. K. Milovac, K. Liao, and P. K. Chrysanthis, "Demo: Senshamart - A sensor sharing marketplace for iot," in *43rd IEEE International Conference on Distributed Computing Systems, ICDCS 2023, Hong Kong, July 18-21, 2023*. IEEE, 2023, pp. 1025–1028. [Online]. Available: https://doi.org/10.1109/ICDCS57875.2023.00122

[17] T. Sulthana, A. S. Jourabchi, S. Song, and B.-Y. Choi, "Sinema: Semantics-driven intelligent network management using ai assistance," in *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, 2024, pp. 35–43.

[18] Q. Zeng, S. Olatunde-Salawu, and F. Nait-Abdesselam, "Fga-ids: A federated learning and gan-augmented intrusion detection system for uav networks," in *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, 2024, pp. 50–59.

[19] W. H. Hsu and A. E. Zwarico, "Automatic synthesis of compression techniques for heterogeneous files," *Softw. Pract. Exper.*, vol. 25, no. 10, p. 1097–1116, Oct. 1995. [Online]. Available: https://doi.org/10.1002/spe.4380251003

[20] A. Kattan and R. Poli, "Evolutionary lossless compression with gp-zip*," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1211–1218. [Online]. Available: https://doi.org/10.1145/1389095.1389333

[21] B. A. Burtchell and M. Burtscher, "Using machine learning to predict effective compression algorithms for heterogeneous datasets," in *Data Compression Conference, DCC 2024, Snowbird, UT, USA, March 19-22, 2024*, A. Bilgin, J. E. Fowler, J. Serra-Sagristà, Y. Ye, and J. A. Storer, Eds. IEEE, 2024, pp. 183–192. [Online]. Available: https://doi.org/10.1109/DCC58796.2024.00026

[22] C. Costa, P. Chrysanthis, H. Herodotou, M. Costa, E. Stavrakis, and N. Nicolaou, "A multiple compression approach using attribute-based signatures [version 1; peer review: 2 approved with reservations]," *Open Research Europe*, vol. 5, no. 49, 2025.

[23] A. Baskin, S. Heyman, B. T. Nixon, C. Costa, and P. K. Chrysanthis, "Remembering the forgotten: Clustering, outlier detection, and accuracy tuning in a postdiction pipeline," in *New Trends in Database and Information Systems - ADBIS 2023 Short Papers, Doctoral Consortium and Workshops: AIDMA, DOING, K-Gals, MADEISD, PeRS, Barcelona, Spain, September 4-7, 2023, Proceedings*, ser. Communications in Computer and Information Science, A. Abelló, P. Vassiliadis, O. Romero, R. Wrembel, F. Bugiotti, J. Gamper, G. Vargas-Solar, and E. Zumpano, Eds., vol. 1850. Springer, 2023, pp. 46–55. [Online]. Available: https://doi.org/10.1007/978-3-031-42941-5_5