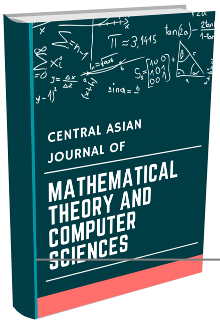

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org

Volume: 04 Issue: 11 | Nov 2023 ISSN: 2660-5309

REINVENTING ENTERPRISE DATA ARCHIVING FOR THE DIGITAL ERA

Sergi v Smirnov
Researcher, Moscow State University, Moscow, Russia

RaviKiran Kandepu
Independent Researcher, Chicago, IL, USA

Abstract

This paper examines next generation approaches to transform enterprise data archiving methodologies for the burgeoning digital age. We analyze the limitations of traditional archiving and the innovations needed to effectively manage massive growth in enterprise data volumes. Detailed sections are included covering the following aspects:

- *The exponential growth in enterprise data and ensuing archiving challenges*
- *Intelligent policy-based automation to enable smarter archiving.*
- *Cloud-native architectures for highly scalable archives*
- *Persistent metadata synchronization for greater archived data utility*
- *Comparison of on-premises, cloud, and hybrid archiving models*
- *Archiving best practices related to security, retention, discovery etc.*
- *Case studies of real-world archiving implementations*
- *Recommendations for a modern holistic enterprise archiving strategy*

Advanced data archiving techniques such as machine learning policies, cloud repositories, active metadata, and unified cross-platform access are imperative today to control data sprawl, accelerate insights, and ensure regulatory compliance. By reimagining archiving for the digital age, enterprises can cost-efficiently extract maximum value from data while minimizing risks.

ARTICLE INFO

Article history*:*
Received 14 Sep 2023
Revised form 15 Oct 2023
Accepted 16 Nov 2023

Key words:
Enterprise Data Archiving, Digitalization, Web application, Systems Management

---***---

Introduction:

The volume of data generated globally is exploding at a breakneck pace. As per IDC, the global datasphere is predicted to expand colossally from 33 zettabytes in 2018 to 175 zettabytes by 2025, a 5-fold growth in just 7 years (Reinsel et al., 2018). While data powers competitive advantage, unrestrained data proliferation leads to bloated IT costs, overwhelmed infrastructure, slower insights, greater security risks, and compliance headaches. It is estimated that over 60% of current enterprise data is copy data – redundant, obsolete, trivial or non-production data (Veritas, 2019). This massive growth and data sprawl necessitate smarter enterprise data management, particularly for inactive rarely accessed “cold” data.

Data archiving is the process of identifying and moving such inactive data from primary production systems into secondary storage tiers, while still retaining it for future access if required (Gandomi & Haider, 2015). Archiving helps reduce the burden on primary infrastructure and optimize storage costs by selective data tiering. But prevalent archiving techniques like tape backups, legacy NAS systems, and siloed departmental archives are no longer adequate in the digital era. Challenges such as fragmented repositories, manual policies, limited metadata and poor interoperability severely reduce the findability, recoverability and utility of archived data.

Next-generation data archiving approaches are needed to effectively tackle the massive scale, distribution, variety and velocity characteristics of today’s enterprise data ecosystems. In this research paper, we undertake a comprehensive examination of cutting-edge data archiving paradigms and architectures that can enable organizations to stay ahead of the data deluge. Specifically, we analyze emerging innovations across three vectors – intelligent policy automation, cloud-native storage, and persistent metadata synchronization. Adopting modern techniques such as these, in conjunction with archiving best practices, can help enterprises optimize storage economics, accelerate business insights, comply with data regulations, and build a strategic foundation for the data-driven future.

The remainder of this paper is organized as follows. First, we discuss the exponential growth trends in enterprise data and the limitations of traditional archiving models. Next, sections are presented analyzing next-practice innovations in policy-driven automation, cloud-based repositories, and active metadata management for holistic data lifecycle management. We also compare the pros and cons of various archiving deployment models – on-premises, cloud, and hybrid architectures. Additionally, we outline some data archiving best practices pertaining to security, retention, discovery and restoration. Finally, the conclusion summarizes the key discussion themes and provides recommendations for architecting futuristic data archives to maximize business value.

Traditional Archiving Challenges:

It is estimated that enterprise data is growing at a 50-60% CAGR globally, doubling every 12-18 months (Gantz & Reinsel, 2011). This data explosion has severely strained traditional data archiving methodologies. Some key limitations are highlighted below:

Data Volume Overload: Enterprise data volumes have simply outpaced conventional archiving systems like tape backups and legacy NAS. These systems are struggling to handle petabyte and exabyte-scale big data (Abbasi et al., 2016).

Manual Processes: Archiving tasks like classifying data, setting policies, moving data, and maintaining metadata are predominantly manual. This leads to productivity lags, oversights, and inconsistencies.

Fragmented Archives: Most enterprises have fragmented archives across various departments, geographies, systems, and platforms. This siloed infrastructure hampers unified data access, discovery and reuse.

Limited Metadata: Basic archival snapshots have minimal metadata on data context, lineage, tags etc. This makes interpreting and repurposing archived data difficult over time.

Compliance Risks: Stale archives with inadequate security, access controls and data disposal can increase regulatory non-compliance risks.

These challenges highlight the need for fundamental rethinking of enterprise archiving principles and architectures. Simply put, traditional data archiving is no longer sustainable in the digital era driven by exponential data growth and advanced analytics like AI/ML. Next, we analyze emerging techniques that can help modernize enterprise archiving.

Intelligent Policy Automation:

Policy engines powered by rules, machine learning and AI are pivotal for automating data archiving tasks and enabling self-optimizing data lifecycles. Archiving policies codify rules to categorize data and drive automated tiering, protection and retention. Traditional static policies relied just on parameters like file size, ownership, timestamps, and rudimentary content patterns. But modern AI/ML techniques now enable smart analysis of multiple data attributes and usage trends to create dynamic policies that continuously adapt and optimize (Bhasin & Gill, 2022).

Google Cloud Storage demonstrates advanced AI-driven auto-tiering between hot, cool and cold classes based on changing access patterns, metadata tags, and over 20 other factors (Google, 2022). This allows optimizing performance and costs dynamically. Microsoft Azure Information Protection leverages ML to classify data across business impact levels for appropriate archival based on sensitivity (Cobb, 2018). Such capability analyzes file contents, user roles, GEO tags, etc. to tag data contextually for policy-based automation.

Key benefits of intelligent policy engines are:

- Automated classification using ML/AI algorithms
- Dynamic policies adapting to evolving usage and access patterns
- Advanced policy controls incorporating multiple contextual parameters
- Regular analysis of data activity, costs, risks to suggest policy changes
- Automated data movement across heterogeneous storage tiers and repositories
- Unified data lifecycle management from production to archive

By operationalizing data science for archiving policies, AI/ML allows building Knowledge-Defined Archives that are perpetually optimized automatically as enterprise information evolves. Policy automation also frees IT staff for higher-value data centric tasks. The use of intelligent tools is surging - Gartner forecasts about 50% of organizations will be leveraging AI-augmented archiving by 2025, up from less than 10% in 2020 (Litan, 2020).

Cloud-Native Archiving:

The cloud offers a versatile, elastic, and economical platform for architecting modern data archives that can seamlessly scale. Leading cloud providers offer a range of cost-efficient archival storage services like AWS S3 Glacier and Azure Archive Storage. A Cloud Spectator (2019) study found cloud archival storage can be 4-5 times cheaper than enterprise tape libraries or disk arrays. The anytime, anywhere availability and virtually unlimited capacity of cloud storage addresses key limitations of on-premises repositories.

Many organizations now leverage hybrid models spanning on-premise infrastructure and multi-cloud tiers. Multiple cloud providers can be integrated for resilience and economics. For example, Commvault's Metallic solution offers policy-driven movement across on-prem, AWS, Azure and Google Cloud repositories in a unified data management platform (Rajendran, 2022). Similarly, Veritas NetBackup enables tiering data seamlessly from existing on-prem disk and tape systems into AWS, Azure and other cloud targets to lower costs (Maxim Group, 2022).

Additional cloud capabilities include:

- Serverless computing to run archive tasks only when required avoiding over-provisioning
- Object versioning and immutable storage to prevent accidental or malicious data deletion
- Geo-distributed resilient archives across regions for business continuity
- Robust security controls encompassing encryption, RBAC, masking, etc.
- Virtual air-gapped archives isolated using private cloud platforms

The economics, scalability and resilience of cloud make it ideal for risk-mitigated long-term enterprise archiving. TCO savings from lower cloud storage costs allow reinvesting in innovation. Cloud also enables easier adoption of technologies like IoT, ML and analytics to enhance archival use cases.

Persistent Metadata Management:

Metadata is crucial structured information about the context, content and structure of data that enables its meaningful interpretation and use. Metadata encapsulates details regarding data definitions, schemas, tags, identifiers, inter-relationships, lineage etc. Insufficient or outdated metadata during archiving leads to archived data essentially becoming "dark data" over time – opaque and unused (Tee, 2018). Maintaining high-fidelity, current metadata is thus pivotal for allowing successful search, retrieval and repurposing of archived data assets throughout their lifecycle. Next-generation archiving solutions tackle this via active metadata management that persistently synchronizes metadata between archives and source systems (Woods, 2021). They maintain updated metadata maps through change data capture, AI-based mapping, and bi-directional sync. For example, Veritas NetBackup maintains continuous metadata synchronization between on-premises and cloud-based object storage to prevent obsolescence (Veritas, 2022). Tools like Informatica Live Data Map and Delphix Dynamic Data Platform also preserve multi-faceted metadata history and technical lineage (Rajesh, 2017; Delphix, 2022).

Comprehensive metadata management capabilities such as:

- Technical metadata on data structures, schemas, formats etc.
- Descriptive metadata like tags, labels, business taxonomy
- Administrative metadata including policies, controls, constraints
- Security metadata including user access, encryption, masking
- Data lineage through the information supply chain

This facilitates multidimensional contextual discovery and governance over archived data. Archiving without active metadata risks creating inaccessible, unusable ROT (redundant, obsolete, trivial) data. Maintaining high metadata fidelity is thus crucial for creating trustworthy and useful archives.

Archiving Deployment Models:

Enterprise archiving solutions can be deployed on-premises, fully on cloud, or as hybrid architectures spanning both. Below we analyze the pros and cons of each model.

Hybrid models are emerging as a common choice combining the control and performance of localized on-premises infrastructure with the scalability, resilience and costs savings of cloud repositories. Cloud innovatively transforms archiving from a capital-intensive cost center into a more easily manageable operational expenditure (Shelton, 2018). Hybrid archiving also enables new use cases like using cloud analytics on archived data. Overall, hybrid models balance economics, control, and agility.

Conclusion:

By adopting modern approaches such as these, enterprises can effectively rein in runaway data growth and extract maximum value from data assets over their lifecycle – from creation to archival to disposal. Intelligent and scalable data archives will provide the foundation for getting ahead of the data deluge and succeeding in an increasingly digital-first world.

References:

1. Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2), 3.
2. Amazon Web Services (AWS). (2022). Zetta Insurance Secures Critical Data in a Hybrid Cloud Archiving Solution. https://aws.amazon.com/solutions/case-studies/zetta-insurance-archiving/
3. Bhasin, M. M., & Gill, A. Q. (2022). Cloud intelligence in enabling data archiving techniques: state-of-the-art, taxonomy, and open research challenges. Human-centric Computing and Information Sciences, 12(1), 1-34.
4. Cobb, S. (2018). Learn how Microsoft uses the Azure confidential computing sandbox. Microsoft. https://azure.microsoft.com/en-us/blog/learn-how-microsoft-uses-the-azure-confidential-computing-sandbox/
5. Cloud Spectator. (2019). Cloud Storage Pricing Comparison. https://www.cloudspectator.com/cloud-storage-pricing-comparison/

6. Delphix. (2022). Lions Gate Accelerates Content Reuse. https://www.delphix.com/customer-story/lionsgate
7. Delphix. (2022). Multi-source Active Metadata. https://www.delphix.com/platform/multi-source-active-metadata
8. Kunduru, A. R., & Kandepu, R. (2023). Data archival methodology in enterprise resource planning applications (Oracle ERP, Peoplesoft). Journal of Advances in Mathematics and Computer Science, 38(9), 115–127. https://doi.org/10.9734/jamcs/2023/v38i91809
9. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.
10. Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1142, 1-12.
11. Google Cloud. (2021). Carlyle Group Automates Cloud Archiving with Intelligent Policy Engine. https://cloud.google.com/customers/carlyle-group
12. Google Cloud. (2022). Cloud Storage auto-tiering. https://cloud.google.com/storage/docs/storage-classes#auto-tiering
13. Litan, A. (2020). Predicts 2021: Future of Storage. Gartner.
14. Maxim Group. (2022). Veritas NetBackup delivers unified data protection. https://www.veritas.com/content/dam/Veritas/docs/white-papers/wp-netbackup-delivers-unified-data-protection-en.pdf
15. Microsoft. (2021). Raytheon Consolidates Storage with Hybrid Cloud Solution. https://customers.microsoft.com/en-us/story/raytheon-manufacturing-azure
16. Rajendran, S. (2022). Metallic Enables Seamless, Unified Data Management. https://www.commvault.com/metallic
17. Rajesh, R. (2017). Live Data Map: The Critical Component Your Enterprise Needs. Informatica. https://www.informatica.com/blog/live-data-map-critical-component-enterprise-needs.html
18. Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World: From Edge to Core. IDC. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-