DSI DATA SCIENCE AND INFORMETRICS
VOL.3, NO.3, AUG. 2023

RESEARCH ARTICLE

Large–scale data archiving: At the interface of archive science and computer science

Chaolemen Borjigin*, Qingwen Jin

a. Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China

b. School of Information Resource Management, Renmin University of China, Beijing, China

ABSTRACT

Both computer science and archival science are concerned with archiving large-scale data, but they have different focuses. Large-scale data archiving in computer science focuses on technical aspects that can reduce the cost of data storage and improve the reliability and efficiency of Big Data management. Its weaknesses lie in inadequate and non-standardized management. Archiving in archival science focuses on the management aspects and neglects the necessary technical considerations, resulting in high storage and retention costs and poor ability to manage Big Data. Therefore, the integration of large-scale data archiving and archival theory can balance the existing research limitations of the two fields and propose two research topics for related research - archival management of Big Data and large-scale management of archived Big Data.

KEYWORDS

Data archiving; Archive Science; Computer Science; Large-scale data; Data storage

Introduction

Integration and innovation of disciplines have become one of the new trends in the development of traditional disciplines. Archival science has always made important contributions to social memory, cultural heritage, national governance, and talent training (Sabiescu, 2020; White & Gilliland, 2010; Elsayed & Ammar, 2020; Ringel & Ribak, 2021). In the era of Big Data, archival science research must also appropriately utilize and incorporate the research findings of related disciplines, especially related problems in the field of computer science. On the one hand, the application of computer technology has given historians access to more archival materials, and history has received more social attention as a result of technological development (Pessanha & Salah, 2022). On the other hand, the change in the technological environment leads to the development and transformation of archival materials from the analog state to the digital state and to the data state. At the same time, experts and scholars in the field of archival science are beginning to pay attention to the interface between archival science and computer science, proposing the concept of Computational

* Corresponding Author: chaolemen@ruc.edu.cn

Archival Science (CAS), which they position as a new development in archival science (Proctor & Marciano, 2021).

Archiving is a pervasive and universal behavior, and various disciplines in almost all fields engage in archiving and conduct archival activities. Therefore, it is likely that archival activities will break through at the interface between archival science and other disciplines. In addition to archival science, research on archival problems is also being conducted in computer science, and there are more profound theoretical breakthroughs and mature solutions. Therefore, this paper discusses the theory, technology, and tools for big data from the perspective of computer science to relate to related research in archival science, especially in the area of electronic records management.

The rest of this paper is organized as follows: In the second part, based on the realistic background and theoretical foundations of data archiving, the main characteristics of big data archiving in computer science and the difference between data archiving and data preservation are analyzed; the third part focuses on the comparative analysis of big data archiving and archival science archiving, highlighting the differences in archival management and complementarities in functional positioning; the fourth part mainly discusses the technology related to archival science archiving activities in large-scale data archiving, including: object storage, hierarchical storage, security control and access use of large-scale data; the fifth part highlights the integration and innovation of large-scale data archiving and archival science archiving from two aspects: the application of archival science in large-scale data archiving and the application of large-scale data archiving technology in archival science. Finally, the research results are summarized and three suggestions for the integration of computer science and archival science are proposed.

1 Data archiving in Computer Science

From a computer science perspective, the most important theoretical basis of data archiving is the Data Tiered Storage (DTS) theory. The main background of this theory is that the read speed of different data storage technologies strongly depends on their cost. In general, the higher the read speed, the higher the cost. Therefore, in the era of exponential growth of data volume, it is necessary to introduce the DTS theory, divide the storage strategy into hot storage tiers, warm storage tiers, and cold storage tiers, and classify data into different storage types according to their access frequency, as shown in Table 1. Using Oracle's tiered storage model shown in Figure 1 (Moore, 2015), data is divided into four tiers (Tier 0~Tier 3) based on lifetime and reuse probability, which provides differentiated storage solutions (T0~T3) for the data on the different tiers, including high-performance flash storage, primary disks, secondary disks, tapes, and archive storage. Through SAM-QFS (The Sun Storage Archive Manager, Quick File System), Oracle's integrated storage stack is combined to build a scalable, tiered storage architecture solution.

Table1 The DTS theory in the field of computer science

	Hot Tiers	**Warm Tiers**	**Cold Tiers**
Usage frequency	High	Less	Very few
Cost of storage carrier	High	Low	Low
Archiving strategy	Not archive (backup)	Short–term archiving	Long–term archiving
Retrieval time requirements	High	Medium	Low
Recovery time requirements	High	Medium	Low

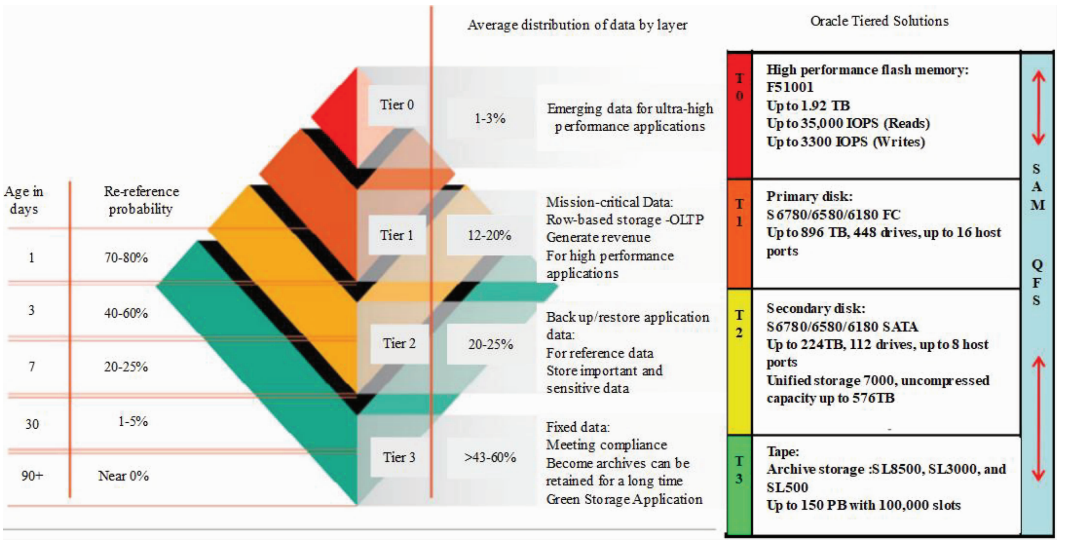

Figure 1 Oracle Tiered Storage (Moore, 2015)

1.1 What is data archiving

In computer science, data archiving is a phase of the data lifecycle (Stodden, 2020) that specifically refers to moving data that is no longer frequently accessed in production systems to low-cost long-term storage devices (MSP360, 2014). Large-scale data archiving in computing has the following five basic characteristics.

(1) prerequisites and aims. data that is no longer frequently accessed. Fred Moore (2022) suggested that "data are archived when they exist for 90 to 120 days and their access probability falls below 0.5%" (Mellor, 2022). Over time, the volume of data increases rapidly, but the probability of accessing the data decreases rapidly (MarkLogic, 2022), as shown in Figure 2. When the volume of data is very large, the probability of access is very low, which increases storage costs. Consequently, Big Data that is accessed infrequently must be archived on lower-cost secondary storage devices and managed using various technologies such as HDFS, NAS, and Amazon S3.

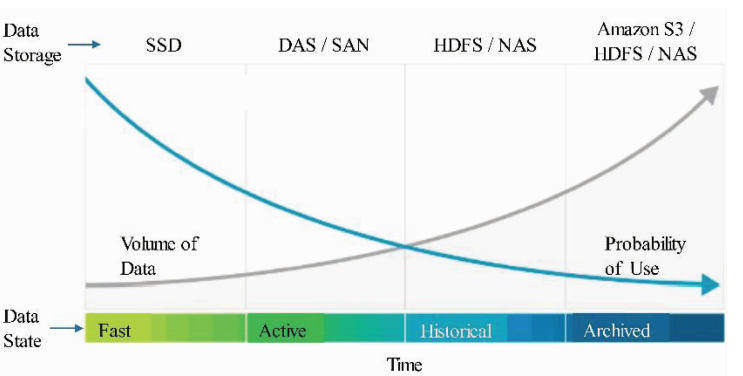

Figure 2 Relationship between data life cycle and data archiving (Adapted from Marklogic, 2022)

(2) Main motivations. In data processing, the main motivation for data archiving is to reduce the cost of data storage. Storage devices used in production systems often have low latency and high throughput, but their cost is too high. When there is a large amount of "sleep data" (big data with low accessibility) in the production system, the usage efficiency of the storage devices in the production system decreases, and the storage cost increases significantly. Therefore, the original motivation of data archiving is to move this "sleep data" to secondary storage devices with low access efficiency and low cost. When re-access is required, we can read at the cost of some access delay and efficiency. Based on the goal of cost reduction, data archiving in the information technology field naturally brings some additional functions or benefits, such as avoiding the risk of data loss, supporting legal and business archiving requirements, eliminating redundant data, and reducing the amount of data in the production system.

(3) Storage location. According to the theory of staged data storage, data archiving is the process of transferring data from hot storage to warm or cold storage. Therefore, in computer science, the storage location of archived data is called the warm or cold storage layer.

(4) Usage requirements. After data archiving, accessing the data is not as efficient as accessing the data in the production system, nor is it completely impossible for the original data subject to access and use the data again. Data archiving in the field of information technology lowers the cost of data storage at the expense of the convenience of data collection. Therefore, Big Data archiving must balance the contradiction between the low cost of data storage and the inconvenience of data use (Calhoun et al., 2019). However, data processing supports the use of archived data by the original data subject at the cost of access performance. In general, the data archiving system must support at least the unified indexing and retrieval function for archived data. Here, unified indexing and retrieval of archived data refers to unified indexing and retrieval of archived data stored in cold storage tier or warm storage tier and current data stored in hot storage tier production systems.

(5) Object Hierarchy. The object hierarchy is mainly concerned with the archiving of hard disk files at the physical level of the storage device, not with the data sets from the business aspect. In computer science, data archiving objects are usually file objects that are no longer frequently accessed. Of course, there are some data archiving systems in computer science that have begun to support encapsulation functions and archiving requirements at the logical level, such as SAP Data Archiving.

1.2 Data archiving vs. data backup

Table 2 shows the main differences between Data Archiving and Data Backup.

Table 2 Differences between data archiving and data backup

	Data archiving	Data backup
Primary purposes	To reduce storage costs	To recover lost data, load balance, or (and) improve reliability
Types of activity	Transferring data	Replicating data
Implementation strategies	Simple	Complex
Exploitation methods	Unified indexing and retrieval	Disaster recovery and emergency treatment, separation of data flow and control flow, load balancing scheduling

(1) Primary purposes. The purpose of data archiving is to reduce the cost of data storage in the production system and to transfer data that is no longer frequently used to a storage medium with lower storage costs. The purpose of data backup is to quickly recover lost data, load balance, or avoid a single point of failure to improve production system reliability. However, it is not directly related to reducing the cost of data storage in the production system.

(2) Types of activity. Data archiving is essentially a transmission behavior. After data archiving, the production system no longer has the data. Data backup is essentially replication. After data backup, the production system still has the data.

(3) Implementation strategies. The data archiving strategy is relatively simple. In general, the selected append method does not directly overwrite or replace the archived data. There are many strategies for data backup (Lenhard, 2022), such as full backup, bulk backup, real-time backup, and mirroring. The operation strategy is not only a single form of attachment, but also full coverage and partial replacement.

(4) Exploitation methods. The data in data archiving generally use a unified index and retrieval method and are retrieved together with the current data of the production system. There are many ways to utilize backup data or replicas, such as data recovery, separation of data and control flow, load balancing planning and disaster recovery, but they are not uniformly indexed and retrieved.

2 Comparation of data archiving in Computer Science and Archival Science

Large-scale data archiving and archival science are two distinct but related fields. The main relationship between the two is reflected in the complementarity of their functional positioning.

2.1 The main differences

In archival science, "archiving" refers to the act of transferring completed records by records-processing departments and business units to archives (Wisniewska-Drewniak, 2021); in electronic records management, "archiving" refers to the process of transferring processed and systematically organised electronic records worthy of evidence, research, and preservation, as well as their metadata management, to the archives department (Specifications for Electronic Records and Electronic Records Management, 2016). Currently, a distinction is made between "archiving" in computer science and "archiving" in archival science, as shown in Table 3.

Table 3 The difference between archiving in Computer Science and Archival Science

	Archiving in Computer Science	Archiving in Archival Science
Theoretical basis	Data Tiering or Tiered Storage Theory	Archival Science and Electronic Records Management
Main motivation	To reduce storage costs	To ensure compliance with laws, regulations and standards
Prerequisite	Infrequency of data access	Completion of file related business
Perspective	Technological perspective	Management perspective
Location	Lower cost storage devices	Archives or Archives Department
Focus	Technical realization	Management specifications

(1) Theoretical basis. The main theoretical basis for data archiving in computer science is the theory of data tiering or tiered storage (DTS), while the main theoretical basis for file archiving in archival science is the theory of archival science and electronic records management.

(2) The main motivation for archiving activities. The main motivation for data archiving in computer science is to reduce the cost of data storage and to store data that is no longer frequently accessed on a less expensive storage medium. However, the main motivation for archiving files in archival science is the need to comply with laws, regulations, and standards.

(3) Prerequisites for archiving activities. The identification of data archiving activities in computer science is based on the frequency of data access, while the prerequisite for file archiving activities in archival science is that the relevant transactions have been performed on files. Thus, "archiving" in computer science does not require that the corresponding transaction has been completed.

(4) Research perspective of archiving objects. The aspect of data archiving discussed in computer science is more fundamental than that in archival science, where issues of data archiving are mainly discussed from a technical point of view, such as file storage, block storage, and object storage. However, archiving in archival science is mainly concerned with the aspect of management, i.e., records generated in the business system that have value as a document, storage and retention.

(5) Location of data archiving. The goal of data archiving discussed in computer science is a cost-effective storage medium. The administration's ownership of the archived files remains unchanged. The goal of archiving activities in archival science is the archive department. The management's ownership of the files before and after archiving changes, and the archival department is responsible for the management.

(6) Research focus. Research on data archiving in computer science is mainly discussed from the technical aspect and focuses on the technical realization of archiving activities. In contrast, research on archiving in archival science is mainly discussed from the management perspective and focuses on ensuring the management of archiving activities.

2.2 The complementary features

The complementarity of large-scale data archiving in computer science and data archiving

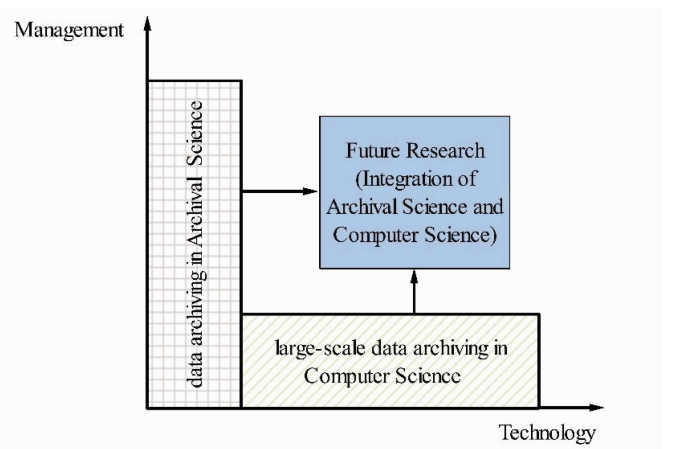

Figure 3 The complementarity of large-scale data archiving in computer science and data archiving in archival science

in archival science is reflected primarily in the complementarity of technology and management. On the one hand, large-scale data archiving based on the theory of data layers has achieved a new breakthrough in technical terms, but lacks the reference and introduction of ideas of archival management. On the other hand, archival theory based on archival science has gained rich experience in the management dimension, but there is an urgent need for innovation in the technical dimension to reduce the cost of storing and maintaining large-scale archives (see Figure 3).

(1) The breakthrough of large-scale data archiving in the technical dimension and the lack of archive management. Unlike archival science, large-scale data archiving focuses on technological innovation and ignores the reference and introduction of archival management ideas and methods. Due to the lack of archival management, large-scale data archiving has resulted in insufficient evidence and investigation. Large-scale data archiving mainly relies on the theory of data layers. The focus of technical innovation is on DTS and storage cost reduction. If archival science is introduced to Big Data archiving, it can not only enhance the value and significance of Big Data archiving, but also provide a new direction for the development of large-scale data archiving.

(2) Archival science takes the lead in the management dimension and compromises with the traditional technical conditions. On the one hand, archival science has accumulated many research results in the field of archival management, which have promoted the sustainable development of archival science. The theory in the field of archival science is not only of great significance to archival management, but also of great reference value to archival management, the management of non-archival data according to the ideas and technologies of archival management. On the other hand, the emergence and development of archival science were synchronous with the technical conditions of the time. Influenced by the lack of traditional technical conditions, the phenomenon occurred that the cost of archiving and management was high, and the level of development and use was low. Thus, traditional archiving and research is still mainly carried out manually, there is a lack of research into algorithms and models, and there is no thorough indexing and rapid review of archival collections. The theory of archival and research management is influenced to a certain extent by the technical conditions of the time. Therefore, the introduction of Big Data archiving into archival science can not only reduce the cost of archiving and management, but also enable the synchronous development of archival management technology and Big Data storage technology.

It can be seen that the integration of Big Data archiving technology in computer science and archive management methods in archival science can promote the joint development of these two fields and become a new growth point for the mutual integration of computer science and archival science.

3 Key technologies for large–scale data archiving

The study and analysis of the common Big Data archiving platform shows that there are four related technologies in the field of Big Data archiving that can be applied to archiving science: Object Encapsulation, Tiered Storage, Security Control, and Access Usage. Table 4 shows the comparative analysis of three popular Big Data archiving platforms: Microsoft Azure Blob Archive, Amazon S3 Glacier Deep Archive, and Google Cloud Coldline.

Table 4 Typical platforms for large-scale data archiving

Data Archiving Solutions	Azure Blob Archive	S3 Glacier Deep Archive	Cloud Coldline
Provider	Microsoft	Amazon	Google
Main features	Binary large object level tiered storage	Tape gateway and virtual tape	Low latency
Data tiered storage	Hot tier, cool tier, Archive tier	S3 Intelligent, S3 Standard, S3 Standard–IA, S3 One Zone–IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, S3 Outposts 和 S3 Glacier Deep Archive	Standard Storage, Near-line Storage, Coldline Storage, Archive Storage
Encapsulation of archived data	Binary Large Object, Blob	Vault	Bucket
Security of archived data	AES–256	AES–256	HMAC
Access for archived data	Azure storage data movement library and the Azure import and export service	AWS Command Line Interface, AWS CLI or REST–based web services	Google Cloud Console, gsutil, Cloud Storage Client libraries and REST API

3.1 Object encapsulation

In general, Big Data archiving systems need to transfer the data to the archival storage device after data collection and encapsulation in the production system, and should perform cross-domain storage, retrieval, and use based on the encapsulation object. In Big Data archiving systems, there are two common archiving technologies.

(1) Block storage

Block storage is a common storage technology in cloud computing, in which data is divided into multiple blocks with a fixed size. The advantages of block storage are that each block has the same size, access is fast, secure, reliable, and easy to update. The disadvantage is the lack of semantic metadata, which is not conducive to supporting query functions. The main feature of Azure Blob Archive is the data encapsulation strategy, the archiving scheme based on Binary Large Object (Blob). A large binary object refers to a large binary data set stored in a single object (Kaur et al., 2021). Binary Large Object technology can support unified storage of complex and mutable data (such as text, image, and video). With the introduction of Binary Large Object technology, Microsoft Group has the key technical means to archive large data and solves the following archiving functions (Microsoft, 2022): Uploading images or documents via browsers, distributed access to storage files, streaming video and audio, writing log files, storing archived data, and supporting local or Azure hosting services for analytics. Figure 4 shows the relationship between binary large object storage resources and related terms from Azure Blob Archive, an archiving platform from Microsoft. Users set up one or more containers under their own storage accounts and archive files in each container in the unit of Binary Large Objects. Containers are collections of Binary Large Object (Blob).

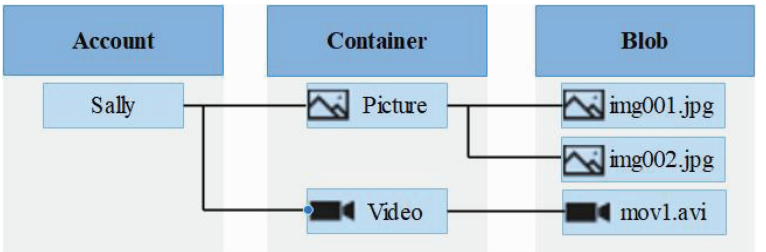

Figure 4 Microsoft Azure Blob Archive (Source: microsoft.com)

For the various application scenarios of binary large objects, Microsoft Group offers three solutions: Block Blocks, Append Blocks and Page Blocks. They are used to transfer new large data, append data to existing data, and transfer page data. When creating binary large objects, users must specify certain types and assign unique identifiers. At the same time, Microsoft has provided a solution for transferring a variety of data to binary big object storage, such as the AzCopy command line tool, Azure Storage Data Movement Library application programming interface, and Azure Import/Export service.

(2) Object Storage

Object storage technology is a logic-level encapsulation technology. It encapsulates data in objects with self-describing metadata, but there is no hierarchical relationship between all objects. Object storage mainly consists of three elements: Data, metadata, and universal unique identifiers (Patil et al., 2020). The advantages of object storage are that it can encapsulate large amounts of data, has the required metadata description and strong retrieval capability, and supports elastic storage of objects of different sizes and quantities as needed. The disadvantages of object storage technology are low access performance, difficult access control, and lack of support for local updates of objects.

Amazon Simple Storage Service (S3) is an object storage integration service from Amazon that includes CloudFront, RDS, Glacier, EBS, EMR, and VPC. S3 Glacier is a storage service for archiving large amounts of data. In terms of data encapsulation, Amazon S3 uses bucket-based encapsulation technology. In Amazon S3 storage, a bucket is a container of objects, and an object is a disk file with metadata. After the data is archived in S3 Glacier, "Bucket" and "Object" are converted to "Vault" and "Archive" respectively, and named in different areas of the "Vault" unit, which are accessed, retrieved, and stored through the API. In S3 Glacier, "Vault" is the container for archived files. As for the security of archived data, S3 Glacier Deep Archive mainly uses AES -256 encryption technology. Besides AES -256 encryption technology, Tape Gateway's Write Once Read Many (WORM) also provides better protection against data loss. As for data archiving tools, S3 Glacier Deep Archive does not support real-time access to archived data. After restoring archived data, users can retrieve the data via the AWS command line interface (AWS CLI) or REST -based Web services.

In Google Cloud Coldline, the most basic container for storing data is called a bucket. Figure 5 shows the object structure of Google Cloud Storage. Google Cloud Storage uses a layered structure that includes four main tiers: Organization (such as Example Inc.), Project (such as messaging app), Bucket (such as photos), and Object (such as puppet.png). Buckets are the units of data management and control in Google's cloud storage. A bucket can store multiple data objects and their metadata, but no other bucket can be nested within it. Each bucket has its own unique identifier and can be stored in multiple geographic locations. In terms of geographic location, Google Cloud Storage offers users three different storage so-

lutions: Single Region, Dual Region and Multi Region. Dual Region and Multi Region are part of the geographically redundant data storage strategy, where the same bucket data is stored in two or more geographic locations that are at least 100 miles apart.

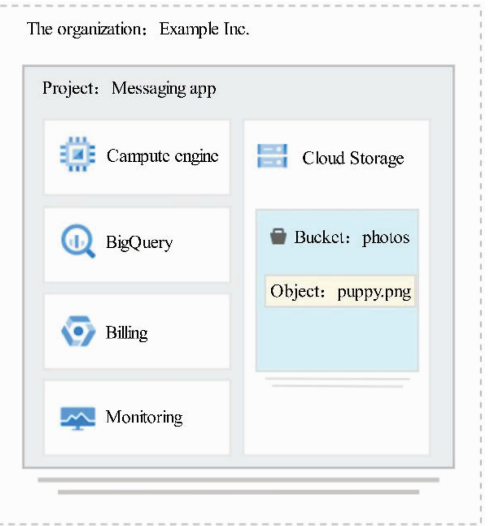

Figure 5 Google object storage (source: Google.com)

3.2 Tiered storage

In addition to cost considerations, Big Data archiving requirements, especially storage volume and read performance requirements, are dynamically changing. Therefore, Big Data archiving requires storage technologies that are highly scalable and have low storage costs.

(1) Multi–tier storage

Azure Blob Archive is a Big Data archiving solution provided by Microsoft Corporation. From the perspective of multi-tier data storage theory, the platform divides data into three tiers: Hot Tier, Cold Tier and Archive Tier. Although the designations are different, they are essentially the same as the Hot Tier, Warm Tier, and Cold Tier designations in the theory of tiered data storage. The Archive Tier represents data with very low access probability and focuses on minimizing storage costs. The unique feature of the Archive Tier of Microsoft Azure Blob Archive is that once the data enters the Archive Tier, it cannot be read or modified. If you want to read and access the data in the Archive Tier, you must first transfer it to the Hot Tier or Cold Tier. From the perspective of CAP theorem, Azure Blob Archive follows the CP strategy, which means it sacrifices the availability of archived data to ensure its quality and stability.

Google Cloud Storage is divided into four storage categories: Standard Storage, Nearline Storage, Coldline Storage and Archive Storage. Standard storage is used to store current data. Nearline storage is used to store backups and multimedia content with a long runtime. Coldline storage is mainly used to store data for disaster recovery. Archive storage has the lowest cost and is used to store archived data (Google, 2022). Archival storage comes at the cost of availability and is still a CP strategy from a theoretical perspective (CAP).

The most salient feature of Google Cloud Coldline in data archiving is its low latency. Unlike Azure Blob Archive and S3 Glacier Deep Archive, the archive storage in Google Cloud Storage has higher availability, and its low latency is not much different from the other three

storage categories. It no longer takes days or hours for archived data to change to the available state, but only milliseconds.

(2) Lake warehouse integration

Azure Blob Archive also supports another Microsoft technology, Azure Data Lake Storage Gen2. Compared with the archiving function of Azure Blob Archive, Azure Data Lake Storage Gen2 supports the data lake function with higher availability and data quality. The compatibility of the above two systems enables the integration of data archiving and data management systems and represents a new trend in large-scale data archiving.

3.3 Data security

Both the archiving process and the management of archived data require data security technology. There are three main data security technologies commonly used in large-scale data archiving activities.

(1) Encryption. Encryption technology used in data archiving can be divided into two types: Encryption of data in the "transmission state" of the data archiving process and encryption of data in the "storage state" after data archiving. For encryption of data in the "transfer state", HTTP-based security technologies such as HTTPs are usually used. For the encryption of data in the "storage state", asymmetric encryption technology is usually used, with AES -256 being the most popular encryption technology in the data archiving field.

(2) Access control. It is usually necessary to implement access control for archived data to control which users have which operating rights for archived data. In information security, there are many access control technologies (Langmead, 2022), such as Mandatory Access Control (MAC), Role-Based Access Control (RBAC), Rule-Based Access Control (RBAC), and other technologies. In a data archiving system, you can choose a specific access control technology depending on your business needs.

From the security perspective, Google Cloud Storage mainly uses hash-based Message Authentication Code (HMAC) keys and V4 signature-based authentication technology. As for archiving tools, Google provides a variety of interactive solutions, including command-line tools (such as gsutil), client libraries (such as Cloud Storage Client Libraries), and API programming interfaces (such as REST APIs).

(3) Identity authentication and digital signature. The data archiving system typically needs to authenticate the people who create and transfer the archives and access the archived files. Therefore, digital signatures and authentication technologies from CA are widely used in data archiving platforms.

3.4 User interface

Typically, archived data is not stored on local storage devices that are directly connected to the production system. Therefore, for data archiving, the data access interface between the client and the archiving platform is also an important technology. Currently, there are three access technologies for data archiving systems.

(1) The graphical interface is a special graphical interface tool provided by the data archiving platform for users. Some platforms support submitting and accessing archived data through the browser. Compared to command-line tools and API interfaces, graphical interface tools are more user-friendly for most non-experts, but their functionality and flexibility are limited.

(2) The command–line tool is the most flexible and powerful tool interface for accessing

archived data. In general, graphical interface tools implement only some of the common functions of command line tools. Compared to graphical interface tools, command line tools provide more flexible command selection and parameter setting capabilities.

(3) API mainly includes HTTP protocol-based REST API interfaces, client-side programming packages, and Soap protocol-based web services. The advantage of API interface is that it can be easily integrated into newly developed applications.

4 Integrating Computer Science and Archive Science for the large–scale data achieving purposes

From the above analysis and discussion, it is clear that Big Data archiving in computer science and archival theory in archival science focus on the technical aspect and the management aspect, respectively. As for the technical aspect, large-scale data archiving reduces the cost of data storage and improves the reliability and efficiency of Big Data management. However, the management dimensions, especially the archival management, are not considered, resulting in inadequate data recording and management irregularities. However, research on archiving in archival science focuses on the management aspect and the lack of technical discussion, which leads to the problems of high storage and curation costs and low scalability of archive management. Therefore, the integration of Big Data archiving and archival theory can balance the existing research limitations of the two fields and bring new research topics to related research, as shown in Figure 6.

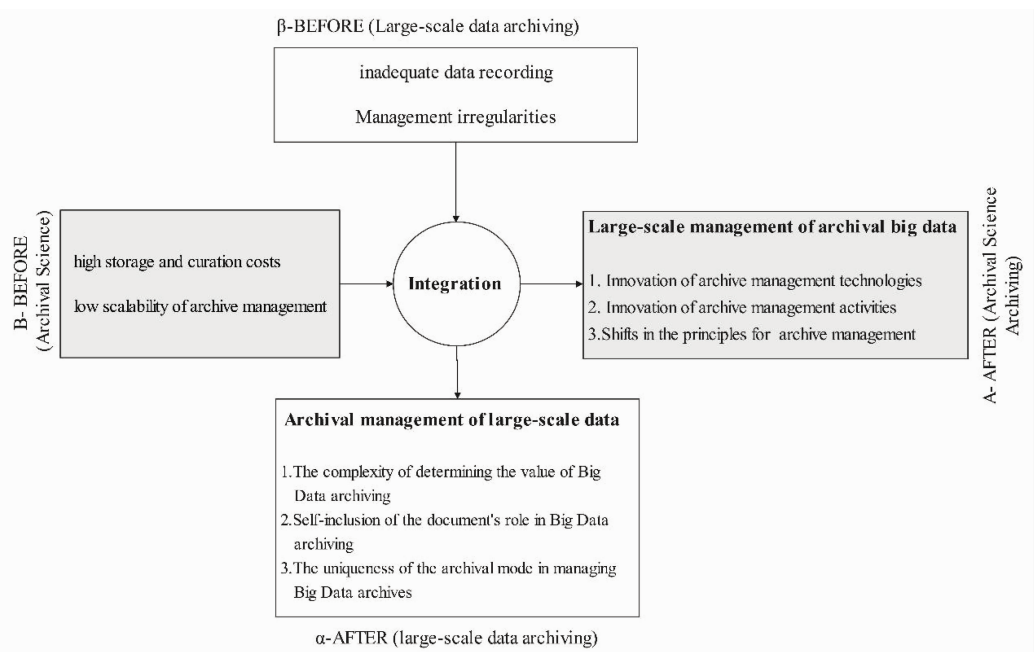

Figure 6 Integrating computer science and archival science for data archiving purposes

4.1 Introducing archival science into computer science

Archiving Big Data refers to the introduction of archival science ideas into the management of Big Data, improving the value of documents, the query function, and the storage ca-

pability of Big Data. However, archiving Big Data does not mean turning Big Data into archives and handing it over to the archives department. From the perspective of archival science, not all objects can become archives after archival management (Cook, 2014). The archiving of Big Data not only needs to introduce the functions of archival order, archival identification, archival recording and research, and archival management and protection into the system of Big Data archiving, but also needs to adopt the ideological model of archival data management, such as the source principle, the inventory principle, the principle of organic connection and the principle of simplification and convenient use, etc., as the new theoretical basis of the system of Big Data archiving. The archiving of Big Data differs from traditional archival science in three main aspects.

(1) The complexity of determining the value of Big Data archiving. Unlike individual archives or small archives discussed in archival science, the value of Big Data tends to be more complex, especially in two ways: first, the value curve of Big Data and its member data is not uniform. In general, the value of archival member data in Big Data changes over time in a "high-low-high" manner, while the overall value of Big Data changes over time differently than member data, and its change can be described as "low-high-low," as shown in Figure 7. It can be seen that the identification of Big Data from the perspective of archives cannot directly copy the archival identification theory in archival science, and the value of Big Data and the value of member data need to be weighed. On the other hand, the characteristics of the value of Big Data become clear. Sometimes, individual data have little value, but when some of the worthless data are combined into Big Data, they become very valuable. For example, the archival value of a person's emotional state at a particular time is not high, but the archival value becomes very high when the records of the emotional state of all the people in a group or a person at all times are combined into Big Data. They can be used as important supporting data for analyzing records of personal growth or social collective memory.

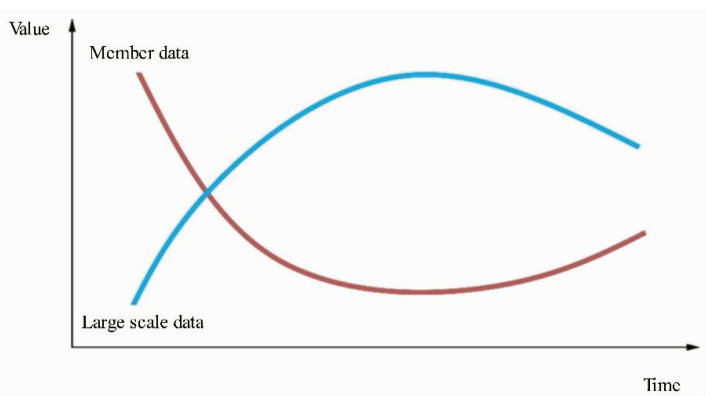

Figure 7 Value curve of large-scale data and its member data

(2) Self–inclusion of the document´s role in Big Data archiving. In the archiving of Big Data, the data is not necessarily managed by the archiving department. Therefore, the document value of Big Data is not guaranteed by the authority of the archive department and the standardization of archive management. However, compared to small data, the document value of Big Data is relatively easy to verify and evaluate. The main reason is that Big Data contains self-describing information, including data and its metadata, mappings and citation relation-

ships between data, and technical information to verify data integrity with a hash value. Therefore, the document value of Big Data is inherent, which can be demonstrated and evaluated by analyzing the Big Data itself. In addition, archival science emphasizes the "principle of organic coherence" - archives are an organic whole that should not be separated at will when managing archival units (Eastwood, 2010). It can be seen that the "principle of organic coherence" of archival science has some guiding significance for the construction, maintenance, and identification of the Big Data dataset.

(3) The uniqueness of the archival mode in managing Big Data archives. Unlike archival science, the archival object in Big Data archiving is not data in the state of "relevant transaction completed", so the management mode must change. One of them is the sequence relationship between sorting and identification behaviors. In archival science, it is common to first perform value identification, then establish a retention period, and perform archival sorting operations such as distinguishing holdings, filing records, and creating the record catalog. However, this method of identification and subsequent archiving is not suitable for large amounts of data, especially for Big Data management scenarios where the relevant business is ongoing, the amount of data is growing rapidly, and the data content is constantly changing. In Big Data archiving, a different strategy is usually followed - sort first, then identify, i.e., the fast-growing and dynamically changing data is first managed with reference to the archival organization methodology of archival science, and then its archival value is identified and evaluated according to the self-identification characteristics of the document value in the Big Data when accessed. On the other hand, the dominant relationship between retention and destruction. In the traditional archiving of data, which is relatively little accessed, the purpose and motivation of data archiving is often preservation, and the archives with evidential, testimonial, and reference value are kept for a long time. However, with very large amounts of data and a significant reduction in the cost of data storage, the motivation and goal of archiving in large-scale data management will change. Destruction will be the main motivation and goal, which means that the data that needs to be destroyed in large-scale data management will be destroyed before it is transferred to the archive department.

4.2 Introducing computer science into archival science

Large-scale Big Data archive management and protection refers to the application of Big Data archiving technologies in the field of archival science, thereby reducing the cost of managing Big Data archives and improving their manageability. The key to large-scale Big Data archive management lies in the division of labor and collaboration between experts in the field of archival science and experts in the field of computer science. It is necessary to integrate the existing digital archives of the archives department and the growing archival data in the future, to use archival management in the technological environment for Big Data management, and to achieve large-scale management and protection of archived Big Data. The key to large-scale Big Data management in archives lies in the following three aspects.

(1) Innovation of Archive management technologies. Large-scale Big Data archive management requires technological innovation, especially the adoption of Big Data archival management technology. The first aspect is the innovation of technology to encapsulate archive objects. From a technical perspective, archival science can introduce object encapsulation technology, which is commonly used in large-scale data archiving, including block storage technology and object storage technology. The second aspect is the introduction of lake ware-

house integration technology. Lake warehouse integration is the main trend of Big Data management in the future and is also a new technology emerging in large-scale data archiving. The adoption of lake warehouse integration technologies such as Azure Data Lake Storage Gen2 and Databrick's Data Lakehouse can not only effectively support the seamless connection between business information systems and archive information systems, but also improve the flexibility and performance of archive data management, especially the standardized management of archive Big Data. The third point is to improve security control technology. Big Data archive management can adopt Big Data archive security control technologies, including HMAC keys and identity authentication technology based on V4 signature, to improve Big Data archive security and access control effectiveness.

(2) Innovation of archival management activities. The introduction of technology-based archival management in archival science is not only an innovation at the technical level, but also brings forth some new research topics at the theoretical level, mainly reflected in the following points: First, large-scale archival management units. Unlike the traditional management of single archives or archives based on holdings, the Big Data management of archives needs to introduce the management unit of data mode, rather than the mode of managing holdings and archives in the traditional document mode. Data archives include the unified management of structured data, unstructured data, and semi-structured data, as well as the management of Big Data replicas. The second aspect is to explore the theory of post-assessment of Big Data in archives. Unlike archival science, the management and protection of Big Data in archives are based on the principle of "sorting before identifying", and the identification of the value of archives is usually postponed to the phase of access and use of archives. The large-scale identification of Big Data in archives requires the extraction and analysis of self-contained data to realize the identification and evaluation of the value of Big Archives. Third, it involves rethinking archival and research methodology. The "manual processing as the main task and automatic processing as the supplement" mode discussed in traditional archival science is not suitable for the acquisition and research of Big Archives. Automatic or fully automatic processing must be used in the acquisition and research of Big Archives, and the algorithms and technical frameworks for the acquisition and research of archives must be innovated.

(3) Shifts in the principles for big scale Archives management. The CAP theorem is one of the central theories of Big Data management and also has a certain reference significance for Big Data archive management. The CAP theorem states that a distributed system cannot satisfy the three requirements of consistency, availability, and partition tolerance at the same time, but at most two (Brewer, 2012). For Big Data archive management, the theory of CAP has changed the traditional pursuit of perfectionism in archive management and preservation and provided a new idea for Big Data archiving, which is to make tradeoffs between the quality, availability, and reliability of archived Big Data data according to the requirements of the actual production system. Unlike traditional archive management, the management of large Big Data archives needs to strike a balance between the quality, availability, and stability of Big Data archives. CP should implement a strategy (consistency and partition tolerance) to ensure the quality and reliability without compromising the availability of the data. The Big Data archive management proposal based on the theory of CAP will change the idea of perfectionism in traditional archive management, unite Big Data archive management and Big Data data archiving from a technical point of view, and achieve the integration and innovation of both.

5 Conclusions

Storage costs and access performance for large data sets are a dichotomy. Assuming that archival requirements are fixed, high storage costs can lead to high access performance. However, if the data is stored in a low-cost storage device, access performance is limited. For this reason, computer science has established the theory of multi-tier data storage, and data archiving usually belongs to the cold storage tier. Based on this, a solution has been found to reduce storage costs at the expense of access performance. In the era of Big Data, the integration of the theory of multilevel data storage and the theory of CAP has become the two main theoretical foundations for Big Data archiving in the field of computer science. In computer science, many breakthroughs have been made in the technical implementation of Big Data archiving, providing the technical foundation for data archiving and more mature solutions such as Microsoft Azure Blob Archive, Amazon S3 Glacier Deep Archive, and Google Cloud Coldline.

Archiving is a widely studied research topic in the fields of archival science and computer science. However, the research levels and perspectives are different, each focusing on the management aspect of the methodology and the technology aspect of the solutions. The theoretical research and technological development of Big Data archiving in the field of computer science have the following impact on archival science: first, research in the field of archival science has attracted more and more attention in many fields. Archival management has become one of the fundamental contents of human data management, and most disciplines urgently need theoretical support from the field of archival science. In the field of archival science, we should emphasize "getting out there" to enable relevant research in other disciplines, including computer science. Second, archival science should also emphasize "inviting" by stimulating research growth in various disciplines on archives and related topics and promoting the sustainable development of archival science itself. Third, we should emphasize not only interdisciplinary and cross-disciplinary research, but also research in archives with multidisciplinary integration. Archival science needs to involve experts and scholars from different disciplines to facilitate in-depth, problem-oriented collaboration. At the same time, archival science needs to train more talents with genuine collaborative character, promote the deep integration of archival science and informatics, and realize the rapid development of new research directions, such as computer-aided archival science.

Biographical Note

Chaolemen Borjigin is an associate professor in School of Information Resources Management, Renmin University of China. With a background in computer science, his research focuses on data archiving, data science, and semantic ontology recognition. He has published several papers on electronic document management, document management system, data continuity and other related topics in the Journal of Archival Science.

Qingwen Jin is a PhD candidate at the School of Information Resources Management, Renmin University of China. His research focuses on interpretable machine learning, information analysis.

Author contributions

C.B. contributed to the study conception, design, data collection, analysis, and final manuscript, and Q.J. contributed to data collection and investigation.

Funding

This work is supported by the National Natural Science Foundation of China (grant number 72074214).

Reference

Brewer, E. (2012). CAP twelve years later: How the "rules" have changed. *Computer, 45*, 23–29. https://doi.org/10.1109/MC.2012.37

Calhoun, S. P., Akin, D., Zimmerman, B., & Neeman, H. (2019) Large scale research data archiving: Training for an inconvenient technology. *Journal of Computational Science, 36*, 100523 –100530. https://doi.org/10.1016/j.jocs.2016.07.005

Cook, M. (2014). *The management of information from archives.* Routledge, London.

Eastwood, T. (2010). A contested realm: The nature of archives and the orientation of archival science. In H. MacNeil & T. Eastwood (eds.), *Currents of archival thinking* (pp.3–21). Libraries Unlimited.

Elsayed, N., & Ammar, S. (2020). Sustainability governance and legitimisation processes: Gulf of Mexico oil spill. Sustainability Accounting, *Management and Policy Journal, 11*, 253–278. https://doi.org/10.1108/SAM–PJ–09–2018–0242

Google. (2022). *Cloud Storage.* https://cloud.google.com/storage.

Kaur, P., Pannu, H. S., Malhi, A. K. (2021). Comparative analysis on cross–modal information retrieval: A review. *Computer Science Review, 39*, 100336. https://doi.org/10.1016/j.cosrev.2020.100336

Langmead, P. (2022). Comparative evaluation of access control models [PhD Thesis, Hood College Computer Science and Information Technology]. MDSOAR.

Lenhard, T. H. (2022). *Data Backup and Restore* (pp. 61–64). Wiesbaden: Springer.

MarkLogic. (2022). *Administrator' s Guide — Chapter 18*. https://docs.marklogic.com/guide/admin/tiered–storage.

Mellor, C. (2022). *Zettabyte era brings archiving front and center*. https://blocksandfiles.com/2022/07/11/zettabyte–era–brings–archiving–front–and–center/.

Microsoft. (2022). *Introduction to Azure Blob storage*. https://docs.microsoft.com/en–us/azure/storage/blobs/storage–blobs–introduction.

Moore, F. (2015). *Tiered Storage Takes Center Stage.* https://www.oracle.com/assets/oracle–tiered–storage–takes–center–194075.pdf.

MSP360. (2014). *Data archiving 10*1: Methods and storage to use. http://cdn.ttgtmedia.com/CascadingTargetedDownloads/downloads/data–archiving%20definition.pdf.

Patil, A., Rangarao, D., Seipp, H., Lasota, M., Santos, R. M., Markovic, R., ... Medlin, T. (2020). *Cloud Object Storage as a Service: IBM Cloud Object Storage from Theory to Practice–For developers, IT architects and IT specialists.* IBM Redbooks, US.

Pessanha, F., & Salah, A. A. (2022). A computational look at oral history archives. *Journal on Computing and Cultural Heritage, 15*, 1–16. https://doi.org/10.1145/3477605

Proctor, J., & Marciano, R. (2021). An AI–assisted framework for rapid conversion of descriptive photo metadata into linked data. *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, Orlando, FL, USA,2255–2261. https://doi.org/10.1109/BigData52589.2021.9671715

Ringel, S., & Ribak, R. (2021). ´Place a book and walk away´: Archival digitization as a socio–technical practice. *Information, Communication & Society, 24*, 2293–2306. https://doi.org/10.1080/1369118X.2020.1766534

Sabiescu, A. G. (2020). Living archives and the social transmission of memory. *Curator: The Museum Journal, 63*, 497–510. https://doi.org/10.1111/cura.12384

ICS 01.140.20. (2016). Specifications for Electronic records archiving and electronic archives management. Part 3: Terms and Definitions.

Stodden, V. (2020). The data science life cycle: A disciplined approach to advancing data science as a science. *Communications of the ACM, 63*, 58–66. https://doi.org/10.1145/3360646

White, K. L., & Gilliland, A. J. (2010). Promoting reflexivity and inclusivity in archival education, research, and practice. *The Library Quarterly: Information, Community, Policy, 80*, 231–248. https://doi.org/10.1086/652874

Wisniewska–Drewniak, M. (2021). Archival description in Polish community archives: Three examples from a multiple case study. *Education for Information, 37,* 121–145. https://doi.org/10.3233/EFI–190361