WHITE PAPER

DATA ARCHIVING: STRATEGIES FOR ACCELERATING DIGITAL TRANSFORMATION

Infosys®
Navigate your next

Overview

Data archiving is the process of long-term storage of enterprise's structured and unstructured data (also referred as Enterprise Information Archiving). It can be defined as moving the data from the live business application to a dedicated archival solution or storage for long-term storage and management (data search/ retrieval, retention,purge etc.). Data archiving is typically performed to better organize the business data based on data value and compliance requirements, by keeping the high value and/or live business data in live application and offloading the non-core data to a different and independent setup (an on-premises or cloud based archive solution), thus improving the solution performance, reducing associated operational, as well as storage costs and ensuring compliance with relevant regulations and policies.

Archiving enables organizations to manage and preserve their data for the long term, however there is no one-size-fits-all solution available and there are several factors, design considerations and execution approaches that affect the success of an enterprise's archival strategy more effectively. From standards perspective, OAIS (Open Archival Information System) reference model (ISO 14721) provides a framework for efficient archival practices and serves as an international standard for digital archives management. It covers the key components and processes involved in the archival of digital information, including preservation, access, and long-term digital information management

This whitepaper investigates core archiving aspects, provides a view of data archival needs, factors to consider, best practices as well as execution considerations. Please note that this whitepaper provides a data oriented viewpoint of archival exercise and archival product selection/ setup/tweaking, or other functional/non-functional aspects are not discussed in this document.

Data Archiving and Digital Transformation :

Data archival is a crucial component of an enterprise's digital transformation journey. Modern archival solutions (e.g. InfoArchive) deploy digital features to fundamentally change the way business data is archived, interacted with, regulated, and retained/purged. Compared to legacy approaches, modern data archival solutions lean more towards seamless data access experiences, delivered real-time via APIs to various data consumption channels, e.g., search UIs, reporting applications. Analytics capabilities of modern archival solutions aid the digital transformation process further, with detailed insights on archived data, ready-to-use reports to determine data distribution patterns (e.g., data spread under a given retention policy) etc., enabling the enterprise to make faster decisions on data tiering, cost optimizations & retention/purge strategies.

Dynamic storage integration capabilities of modern archival solutions, coupled with cloud options, enhance the digital transformation quotient further, providing flexibility & scalability, along with rule based data movement options between storage media. This allows automated & efficient balancing of storage requirements, access needs and cost impacts and helps reduce the overall cost footprint greatly (especially for large data volumes).

A good level of built in resilience capabilities, augmented by cloud storage help enterprise hedge the risk of data loss due to hardware failure/human actions (automated backups, disaster recovery setups)

To summarize, a good archival strategy can help ensure enterprise can protect its data, comply with regulations, facilitate better data management, and reduce costs, ultimately supporting digital transformation journey.

Data Archiving: What and Why

Data archiving refers to the movement of end of life/inactive/infrequently used data from the primary application to an archival solution or storage. It involves moving the enterprise data (e.g., documents, images, emails, tabular data) from the primary data repository to the archival solution which may be located on-premises/cloud (or hybrid, a mix of both) and may be managed by the enterprise itself or hosted by a 3rd party vendor. In few cases, another variant, termed as live archival (characterized by real-time reads) is also deployed. This consists of pushing live business data to a WORM (Write-Once-Read-Many) setup in real time/near real time & retaining it in read-only mode there. This variant is usually geared more towards exploiting cost & performance advantages of WORM setups, rather than archiving and compliance aspects.

Archiving process consists of ETL, i.e., Extract (data extraction from source), Transform (data conversion/mapping to archival format), and Load (data ingest in archival solution) stages, done using off-the-shelf software or custom utilities (often a mix of both) designed to move data from the source repository to data archive, while maintaining the data structure and integrity. The process may be a one-time activity (e.g., application decommissioning) or a recurring task (data offloading from source in batches), or a mix of both in some cases (initial bulk data archival, followed by recurring batches)

Below diagram provides an overview of a typical archival cycle for business data –

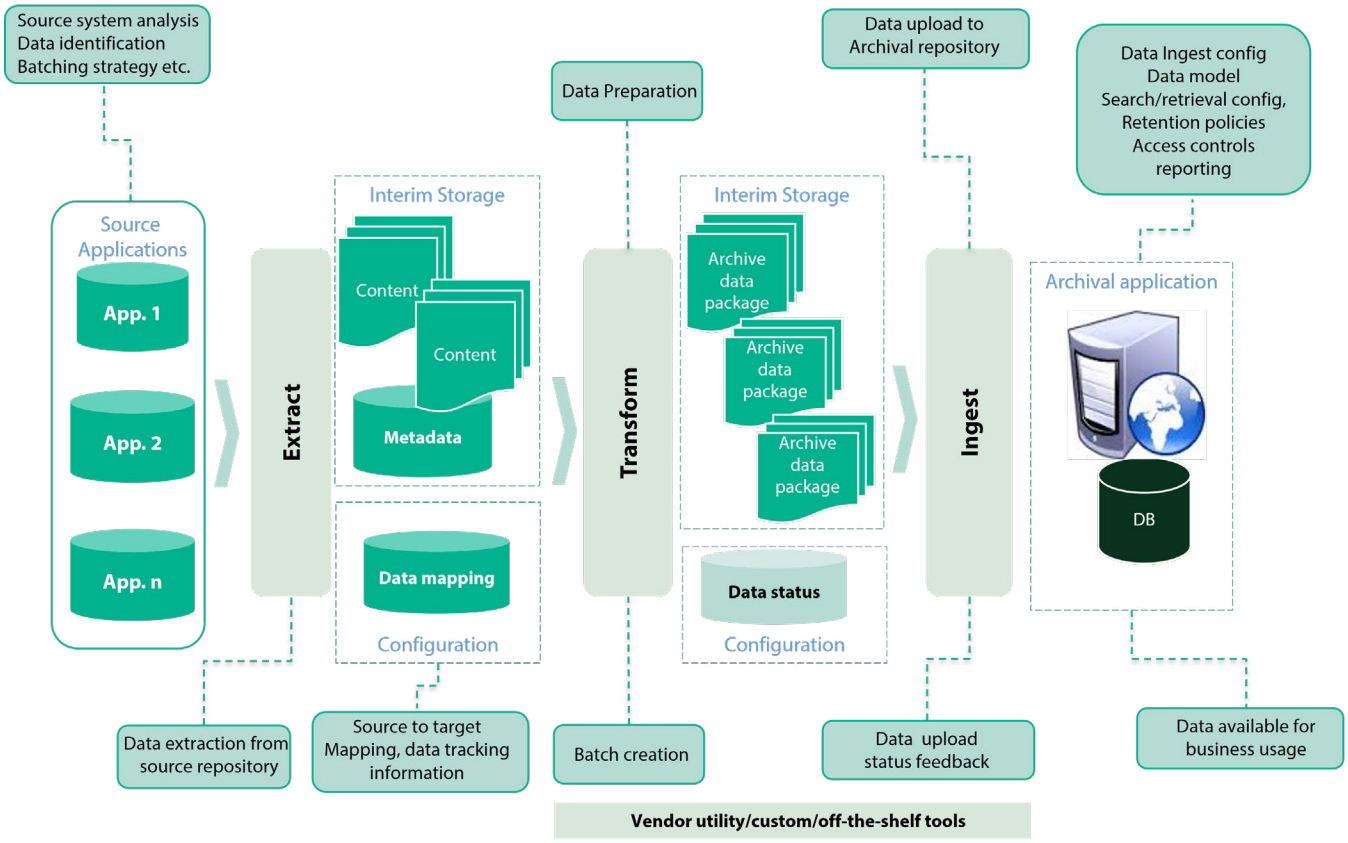

Figure 1: Data Archival: A typical archival flow

In enterprise landscape , following factors usually necessitate undertaking the data archival activity –

Compliance - Many industries have regulatory requirements for retention & disposal of electronic records, e.g., –

- *Record retention laws, mandating how long certain types of records must be kept*
- *Data protection regulations, such as the General Data Protection Regulation (GDPR) in the EU and the California Consumer Privacy Act (CCPA) in the US, outlining requirements for personal data retention and protection*
- *Financial regulations dictating how long financial records must be kept, e.g., Sarbanes-Oxley Act (SOX) in the US*

Furthermore, often legal requirements also necessitate data archival & retention. For example, in case of a legal dispute or investigation, organizations may need to provide certain documents as evidence, and may need to retain them for a specific duration.

Archiving the data to a dedicated archival solution helps ensure that the organization stays in compliance with applicable laws & regulations. While some of the business applications may provide some level of retention management, a dedicated archival solution offers a greater range of archival capabilities, thus helping improve the compliance & adherence to applicable standards, while also facilitating easier tracking of individual document/large data sets' compliance status & compliance metrics

Costs – Storing large volumes of data in live business applications can be expensive, especially as the amount of data grows over time. In addition to storage and compute costs increase, it may also attract additional product license costs. On the operational front too, solution maintenance costs increase due to system size and complexity

Hence freeing up space on the primary business applications can provide considerable cost advantages, by offloading data from costly storage systems to cheaper storages on archival solutions (tiered storage can bring in further cost advantages). It may also allow product license/maintenance cost benefits due to capacity/numbers reduction on primary application (lesser nodes to be run). There is also scope for operational efficiencies, as a centralized archival solution can help streamline data maintenance activities

A specific use case here is application decommissioning, where the business application is no longer required & needs to be retired. In such cases, archiving the data is the most logical choice, to get rid of license & maintenance costs, while retaining the data in archive for compliance purposes, in a cost effective manner

System performance/availability - With massive data sets, business applications often suffer performance issues, e.g., search/retrieval latency. Also, it needs to run larger number of nodes, storage racks & other supporting hardware components, which become difficult to manage, resulting in increased downtime/maintenance overhead, further hampering the system performance & availability

Maintaining a smaller, active business data set on live business application can significantly improve system performance, especially for data search/retrieval. Archiving also brings in operational efficiencies for business application, due to a leaner shape (lesser compute & storage requirements) post data movement, resulting in system maintenance reduction benefits. It may also help bring in faster backup and recovery cycles (due to smaller data sets at business application)

Data management – Managing data based on data value requires different data destinations and opens further avenues for business application landscape optimization. Moving end of life/inactive/infrequently used business data to an archive allows live application to focus more on critical business data, tasks & activities. Also, having only the active business data in live application eases strategic & operational decisions (e.g., retiring search UIs built for legacy data). Similarly, a dedicated archival solution can help refine enterprise data strategies, e.g., centralized retention management/purge/discovery/reporting etc., instead of dealing with data in multiple different applications.

Data Archival: Analysis and Planning

A good data archival strategy is key to a successful archival outcome. It involves initial analysis (data boundaries firm-up, impact analysis), planning the extract, transform, and load activities, as well as overall design including data migration, archival data model & security among others.

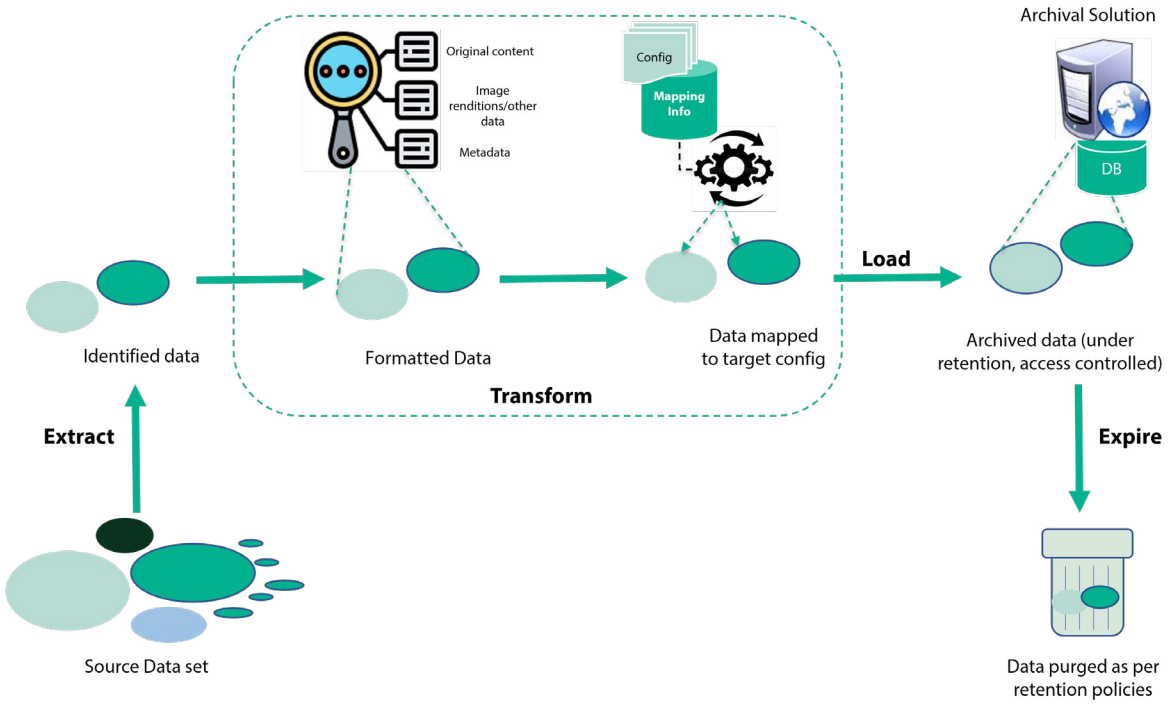

Figure 2: Data Archival to an Archival Solution: Data journey

Below is an overview of archival analysis & design activities and factors involved –

1. Data Analysis

Data analysis involves evaluating and examining the source data in scope, determine its relevance, accuracy, and completeness. It also involves analysis of source application(s) to determine other factors that impact the data and may influence the archival design (e.g., integrations, retention policies).

a. Data identification and categorization: A thorough set of data identification and categorization rules can help focus the efforts on optimal archival strategy, ensuring data is properly archived, protected and accessible. It also helps eliminate the need for archiving unnecessary data. The rules need to include the criteria for

- *Candidate data selection (age/type etc.)*
- *Archival fitment assessment (data sensitivity evaluation, legal/ regulatory requirements impact, fitment for archival/purge etc.)*
- *Data categorization & prioritization (based on business value/data state/metadata to be archived etc.)*
- *Data access requirements (real-time data access/ infrequent but SLA bound access etc.)*

b. Retention: It is important to have clear retention policies and procedures in place for archival of data. These policies & guidelines should be well-documented and communicated to relevant stakeholders to ensure compliance with relevant regulations and data accessibility. The retention information required includes:

- *Policies for retention periods*
- *Type of retention required (e.g., scheduled/event based/mixed)*
- *Related capabilities (e.g., legal hold requirements)*

Below are few examples of data regulations, describing data retention compliance in different parts of the world:

- General Data Protection Regulation (GDPR): EU
- Health Insurance Portability and Accountability Act (HIPAA): USA
- Sarbanes-Oxley Act (SOX): USA
- National Archives and Records Administration (NARA): USA

c. Dependencies: While data destined for an archival solution is often inactive and hence usually not part of any active business processes, it may be referenced by other applications in read-only mode, for their internal processing or for reporting purposes. For example, at times, integrating applications have static query string URLs, containing document identifiers from business application. Moving this data to archival solution will break all such references, so impact on integrated applications due to data archival should be thoroughly assessed

2. Archival Design

Design for data archival needs to consider data characteristics (format, structure), accessibility requirements, as well as data archival objectives. The design also includes consideration of security and privacy requirements, in addition to hardware requirements firm up (compute, storage, future growth as applicable).

Below are the core considerations, contributing to design phase:

a. **Formats:** While archival process usually involves movement of data in as-is formats, the long term view of data usage should be given due consideration, for proper formatting and structuring of data. For example, PDF/A format should be considered for long term data archival (particularly for non-image formats), ensuring the document will remain readable and retain its formatting over time (PDF/A is a standardized format, optimized for archiving and long-term preservation of electronic documents)

Data formats acquire even greater importance, when the source application stores data in a proprietary, non-human-readable format (e.g., AFP, MSAR data). In such cases, format conversion approach requires evaluation of various parameters & after thorough analysis, an approach should be arrived at. Below whitepaper provides some insights on AFP format handling –

https://www.infosys.com/services/digital-marketing/documents/content-archiving-infoarchive.pdf

On a related note, if image renditions (pdf versions) are to be generated for the given document set for long term viability, additional software/hardware requirements and storage assessment as well as cost impact should also be part of overall planning

b. **Access:** Post archival data access requirements & mechanisms should be ironed out and documented in detail. An issue often observed with large data set archival is the lack of data access planning, resulting in sub optimal archival management.

Few of the factors that need to be emphasized upon, during access planning exercise –

- **Business Roles** – *what are roles required to effectively manage the archived data from business perspective, e.g., retention manager, business owner, eDiscovery admin, end user*
- **Mechanisms** – *how the data will be accessed, e.g., access anywhere/desktop applications only? Does data need to be accessible to other applications as well (e.g., used by enterprise reporting tools)?*

c. **Search and Retrieval:** Data search/retrieval capability is a vital aspect for archived data, making it crucial to thoroughly analyze the requirements. Few points to consider while defining search & retrieval design –

- **Search type** – Determine whether real time searches are required or given data set will be accessed with batch queries only. Accordingly, background searches/real-time searches (or both) need to be configured
- **Search forms** – Determine the search requirements, including
  - *Number of searches*
  - *Search criteria (e.g., Type, Date, other business criteria), operators (like, equals etc.) & combinations (multiple search criteria on a given search form)*
  - *Search results, i.e., what columns to display in results*
  - *Result(s) export, i.e., zip & export, export as original/transformed export (convert to image formats) etc.*

Below are few best practices, that should be considered while designing the archival searches –

- *Consider the business value of data as an input to search design. Often the ask is to replicate searches as-is (from source business application to archival solution), but consider the data value & usage frequency to determine the design, e.g., can a given search requirement be satisfied using a support ticket based process or a real-time search is required for it*
- *Avoid complex rule creation (e.g., validations/complex joins/ cascading searches/other custom logic) for searches. An archive is a data preservation destination, not business processing engine*
- *Similarly, thoroughly evaluate any full-text search requirements, as it'll incur additional costs (consider adding metadata during transformation, where feasible)*
- *Treat archived data set as a flat structure, instead of business hierarchies/dependencies (as present at source business application end), use it to apply rules/search data (this is driven by data organization during the archival, discussed next)*

d. **Data organization for archival:** Some data sets in business application may have complex structures, such as nested data or data with multiple levels of hierarchy/relationships (e.g., document versions, annotations, compound documents). Archiving such data could be challenging due to expectation from business for structure preservation, however efforts should be made to flatten out the data structure, i.e., treat each data item at same level (e.g., no parent/child relationship). An option could be to include structure information as metadata, which can help determine hierarchy/relationship in the archived data.

Annotations present an interesting use case here, as these are often vital to image data stored. Burning annotations into the image during transformation phase can be a good option. Depending upon use case, exporting as metadata may also be explored

Another factor to be ironed out during data organization is data export format from source (e.g., Zip/Csv) & data load format (driven by target application/migration tool and loading mechanism, i.e., API/batch based)

e. Data security: To ensure archived data is protected from unauthorized access, modification, or destruction, security measures need to be designed & deployed (encryption, access controls etc.). Note that these mechanisms are usually set up as part of overall archival solution architecture & may not be customizable for specific archival cycles

f. **Integrations:** In certain scenarios, accessing archived data from other applications within an organization may be required, e.g. searching for data from another application or using it for reporting purposes. These scenarios primarily relate to capabilities of archival solution in place, however, may influence the archive design to some extent (e.g., fields to be searchable) hence should be kept in consideration for archival design

g. **Storage Choices:** While overall archival solution architecture is established during the initial setup & may not be modifiable for individual data archival cycles, modern archival solutions (e.g., InfoArchive) provide a lot of configuration/customization options, to suit specific archival requirements of a given data set.

One of such configurable options is storage where a given data set can choose from a mix of storage media to best serve its storage needs. The storage choices can include use of on-premises or cloud-based storage (or both) for archived data storage. The optimal mix of storage can help balance the costs & functionality aspects and hence it's an important factor, in addition to determining the storage capacity & future growth.

Storage selection for data archiving is often influenced by an organization's data storage, data management, performance, and cost requirements. Although a single storage option may suffice in some cases, a hybrid approach combining on-premises & cloud-based storage may prove advantageous in other scenarios.

Table below lists some important factors associated with on-premises & cloud storage choices –

Parameter	On-premises storage	Cloud storage
Cost	On-premises storage may be more expensive upfront, due to the purchase and maintenance of physical hardware	Cloud-based storage, is typically subscription-based and usually offers more flexible pricing options
Data Security	On-premises storage may offer greater control over data access and security, as an organization can physically secure its servers and implement its own security measures	Cloud-based storage providers too, offer robust security measures, but organizations may have less control over how their data is protected (hence for sensitive customer data, often, organizations are reluctant to use public cloud storage offerings)
Scalability	Less flexibility & longer provisioning timings. Demands need to be forecast well in advance, to keep up with storage expansion requirements	More scalable than on-premises solutions, allows to quickly add/remove storage capacity as needed. This can be especially useful for organizations with rapidly changing storage needs or those that expect their data to grow significantly over time
Storage Tiering	Limited options compared to cloud, complications around feasibility and setup depending upon volume	Available with all leading cloud vendors, allowing data to be stored on different tiers based on access patterns, can help optimize storage costs greatly
Data Accessibility	On-premises storage may require more complex integration setup, to make the data accessible to other applications (e.g., building an API layer over data/rely on archival product's API offerings)	Cloud-based storage usually offer more flexibility in terms of integration with other systems, as data can be accessed via cloud APIs or other cloud tools. It can be further augmented by API capabilities of the archival product in use

An important aspect of storage choices is storage tiering, which can influence the cost factor a lot, along with performance choices. This is prevalent for the cloud storage models, i.e. providing a method of organizing data on different storage media based on the access frequency requirements. This can help organizations optimize their storage costs and improve system performance by storing data in the most appropriate tier based on its access patterns. In addition, leading cloud storage vendors also offer automated, rule based movement of data from faster, costly storage to less fast & cheaper storage based on access frequency reduction.

There are broadly three main tiers of storage under this model:

Below are the core considerations, contributing to design phase:

1. **Hot storage:** The highest-performance storage tier is termed as hot storage and is used for frequently accessed data. This may include data used by live applications/business processes, or data accessed frequently by users. Hot storage typically offers high availability and low latency, however its more expensive compared to other tiers. AWS S3 is considered a hot storage choice (in archival context).
2. **Warm storage:** Warm storage is a medium-performance storage tier, used for data set with less frequent access (compared to hot storage), but with relatively fast access requirement (i.e., occasionally used, but with stringent SLAs). Warm storage typically offers lower performance and higher latency, but is more cost effective usually, compared to hot storage. AWS S3-IA is an example of warm storage.
3. **Cold storage:** Cold storage is the lowest-performance tier and is used for infrequently accessed data/data not needed for immediate use. This may include data retained for archival or compliance purposes only, or data rarely accessed. Cold storage is characterized by the lowest performance and highest latency of all the tiers, but it's usually the most cost-effective option for storing large amounts of inactive data with no immediate use/ relaxed retrieval SLAs. AWS Glacier is a cold storage offering example from AWS.

Note that cloud vendors offer further variations of above storage categories, however logically, most will align with above mentioned categorization.

h. **Reporting:** While driven by target archival solution capabilities, reporting on archived data should be considered a vital design element as it can help organizations maintain good data governance practices and ensure that they are not retaining unnecessary data.

- *Reporting on archived data can help organizations monitor the compliance with retention requirements and ensure that data is retained in line with desired policies.*
- *Reporting can help ease archived data governance, by helping organizations track and manage their data, aiding data review process (for example, allowing to see if any data set can be purged due to changes in regulations), bringing in cost & governance efficiencies.*

i. **Notification:** Configuring notifications or alerts on archived data for any access/modification/deletion enables organizations to track and monitor the usage of archived data. It also helps monitor that their data is being used in accordance with relevant regulations and policies. Ideally, organizations should have clear policies describing types of events that should trigger notifications, and the roles or groups for receiving notifications. If not present, the rules & guidelines should be clearly defined & followed for notification design.

3. **Data Movement Design**

Archival design defines the desired target state for data, while data movement design defines plan and execution strategy to move data to target state. It involves strategies for overall ETL process, determining the hardware & software requirements for process execution as well as reconciliation & security approach finalization. A well designed data movement approach can ensure a smooth transition and minimize disruption to business operations. Few important points to keep in view while designing the data movement framework -

a. **Tool considerations** – Migration tool is the component responsible for performing the overall ETL (or a subset of it) activities. There could be couple of options that can be considered for tool selection –

- *Often, there are archival product connectors available, with OOB capabilities to integrate with many source applications, to perform the ETL process (e.g., Documentum connector for InfoArchive)*
- *Alternatively custom utilities can be built to perform the data movement.*
- *In some cases, source applications have extract tools available, allowing data extraction from source, requiring only the transformation & load to be taken care of by the migration utilities/ scripts.*
- *Extraction strategy (e.g., API based, source application utilities or data copy in migration environment) also plays a part in tool selection.*

For large data sets, usually dedicated migration tools provide better results. There are many off-the-shelf tools available, which can help this process (though these may add significant costs to program). Custom tool development is also an option, however efforts required for development & tweaking (e.g., performance optimization) should be factored into. Overall, a decision on archival tool should be taken keeping in view the tool capabilities, performance, costs & fitment into overall archival plan.

b. **Infrastructure setup** – Consider setting up a dedicated intermediate migration environment (compute & storage) between source & target (especially for large data sets). The dedicated setup will ease tracking & fixing any issued during ETL process and simplify tracking individual data object's overall journey & status and help facilitate any remedial steps as required. A dedicated setup typically also provides higher performance and helps avoid impacting the source/target compute as well as simplifying the transformation process.

c. **Hardware considerations** – This includes arriving at number of environments, compute capacity & storage required for archival process, in migration environment as well as any additional hardware requirements in target archival solution. It also includes assessing source capability to support extraction.

d. **Operational considerations** – This involves planning the runtime factors including batching strategy, error handling & reconciliation mechanism.

e. **Data movement approach** – Data can be moved from source application to archival solution using a Big-bang or Recurring approach. Big-bang approach involves moving all the eligible data, to the archive in one go, as a one-time, large-scale migration, while recurring approach divides the data into smaller batches, and moves it to archival application on regular, scheduled basis.

There are pros and cons associated with both approaches, and the best approach will depend on several factors including data volume, requirements, resources, and the impact on source system performance and availability during the migration process. The advantages and disadvantages of both approaches are as listed below –

Approach	Advantages	Disadvantages
Big bang	**Shorter timeline:** moves entire data set to the archive in a single batch, allowing the overall process to be completed quicker than recurring approach	**Planning and resource intensive:** requires significant upfront planning and resources, including the development of a detailed archival plan and design, acquisition of necessary hardware and software and allocation of sufficient personnel and resources to carry out the overall activity
	Potential cost advantages: can be more cost-effective for large data volumes, as it avoids the need for ongoing data migration efforts & associated resources	**Operational Disruptions:** may introduce disruptions to live business application operations, due to longer downtime requirements or source application performance/stability impact due to heavy resource consumption during archival execution window
Recurring	**Manageability:** For large data volumes, its more manageable to split data in smaller batches and archive it on a recurring basis, compared to big-bang approach as it is less resource intensive and can be tweaked anytime to include any changes as required. It can also help to incorporate new data more easily into archive, as it is created/modified	**Overall duration:** A recurring approach takes longer than a big bang migration, as it involves moving data to the archive on a regular, ongoing basis rather than all at once
	Lesser Operational interferences: can help minimize disruption to ongoing operations during archival, as it allows organizations to move data to the archive in an incremental and controlled manner	**Costs:** A recurring approach may turn out to be overall more expensive, as it involves ongoing data migration efforts rather than a one-time migration

There could be another option as well, with an initial bulk data load, followed by recurring batches.

Ultimately, the best approach to data archiving will depend on the specific needs and resources of the organization. It is important for organizations to carefully consider their options and choose the approach best meeting their needs.

Below whitepaper can be referred to, for a more detailed view of best practices associated with data migrations –

https://www.infosys.com/services/digital-experience/insights/documents/docs-open-filenet-journey.pdf

Data Archival: Execution

Analysis & design phase is followed by the execution phase, where the data set is extracted, transformed & eventually loaded into target archival solution. Sections below provide a view of factors and considerations involved in each of the execution stages -

1. Extraction

As part of data extraction activity, identified data set is extracted from source business application. Few points to keep in consideration while performing the extraction –

- *Any data fixups (e.g., missing/junk data remediation) if required, should be done in source application before data is extracted (many off-the-shelf business applications provide data consistency checker jobs)*
- *A unique identifier should be maintained by the migration tool database to keep track of data extracted & its status (or a custom table in case of standalone utilities/scripts)*
- *Apart from count & sample based validations, hash based content integrity checks may also be used (where feasible).*

2. Transformation

Data transformation deals with mapping & conversion of the source data set as per the target archival solution requirements (e.g., xml, zip formats) as well as format conversions required (e.g., image version creation, annotation conversion to text/another format). Points to consider –

- *Data mapping rules must be solidified in design phase, to align source data with archive data model.*
- *Validation checks should be done to ensure data quality and integrity standards (may include count reconciliation, data state & integrity checks)*

3. Load

Loading data into the archival solution can be done using migration tool/utility (usually API based) or archival product's load capabilities. Below are few factors that can influence selection of a suitable approach:

Volume: For large data sets, batch-based approach may be more suitable (usually more efficient for large volumes)

Complexity: Usually for complex data set with significant transformation/mapping requirements, an API-based loading approach may provide better results. This approach allows transformation and loading of the data on an individual record level, providing more flexibility for complex data sets

Availability: If the data set needs to be available immediately in archive, post movement from source, an API-based approach may be more suitable. This approach allows loading data in real-time as it is generated (as output of extraction & transformation phases). For infrequently accessed data set, a batch-based approach may be more suitable

Post data load, the load status is updated in migration database (or custom table). Daily reports should be generated & reconciled with source to ensure completeness. Loaded data should be quality checked for any integrity issues/errors.

Data Archival: Post Execution

a. **Data deletion from source:** Data that is successfully archived in the target archival solution, needs to be deleted from the source business application. While each organization may have its own procedures & protocols around data destruction which need to be adhered to, below are the few considerations to be factored in –

   1. Data identification: Use the archived data set reports to identify the deletion candidates
   2. Impact notification: Data changes may potentially impact dependent systems/applications (determined during the design phase). Deletion phase should ideally notify the affected groups of the deletion exercise
   3. Deletion method: This may involve automated data deletion scripts, manually data deletion using admin tools, or data destruction tools usage, to permanently erase the data. Ensure adherence to organization guidelines if any present, for method selection
   4. Data deletion: Delete data from source business application using the deletion method selected. Organizational procedures and protocols should be followed during data deletion, to ensure data is destroyed and unrecoverable (in line with organization guidelines). Verification is important to confirm data deletion success and that data is no longer available. This may involve reviewing reports/logs as well as conducting a manual review of the data in the business application

b. **Data review:** Data retention requirements can change over time. Data archive should be regularly reviewed to ensure that it is current and relevant, and to remove or destroy any data that is no longer needed.

Archival: Some unique scenarios

At times, some unique scenarios emerge in organization, which may complicate the archival design/execution & may require developing a strategy to address them. Listed below are few such examples –

1. **Source data under retention:** Several off-the-shelf business applications provide retention capabilities, and it may be possible that source business data set identified for archival, may already have retention policies applied to it. Archiving such a data set would complicate the compliance aspect, hence for such scenarios, there are several factors that organizations need to consider, to ensure compliance is maintained:

- ***Data fitment re-assessment*** *– the data already under retention should be re-assessed for a cost/benefit analysis, e.g.*
- *Can such data be kept in source itself, allowing it to expire & be purged in source business application itself? How long till the documents expire?*
- *Any significant benefits accrued by data movement, keeping in view the complexity involved?*
- ***Review & update:*** *Review the retention policy to understand specific requirements for retaining the documents, analyze whether the retention period can be shortened. This may involve updating the source retention schedule (to a shorter duration), to retain & eventually expire such data in source itself or creating a new retention policy for the document (for application in target), post compliance aspect evaluation.*
- *Re-application: If data transfer is deemed feasible, transfer documents to archival solution, apply the new retention policy.*

2. ***Archiving sensitive/confidential data:*** Some data sets may contain sensitive or confidential information that requires special handling when it is archived. This may involve implementing additional security measures or taking steps to ensure that the data is not accessed by unauthorized individuals in-flight & at-rest.

Summary

It is important for organizations to implement a well-planned and structured archival strategy to ensure that their data is properly managed and preserved for the long term. This may involve the development of clear policies and procedures for the selection, retention, and destruction of data, as well as the use of appropriate software and systems to manage and protect the data in the archive. Overall, data archiving is an important aspect of organizations' compliance duties and can help them manage, preserve, and access their data over the long term more effectively, while also helping to reduce costs and ensure compliance with relevant regulations and policies.

Author

Ravi Kumar
Senior Technology Architect;
Digital Marketing Professional

Mentor

Girish Pande
Principal Technology
Architect

For more information, contact askus@infosys.com

Navigate your next