Database Archival
A Bird's Eye view on Design & Execution Approach for efficient Archival

Rajalakshmi Thiruthuraipondi Natarajan

rajalan11@gmail.com

Abstract

Data Archival is an activity or a process, that almost every organization performs to isolate and retain inactive and historical data outside of the primary storage device, for reference and reporting purposes. Though archived data might not be actively used in daily operations, it is imperative that proper planning and execution needs to be put in place. There are several out-of-the box tools available in the market for easy and effective archiving, however, it is the onus of each organization to understand the reason for archiving, the scope, who is it for and the end game to provide customized archival solution.

This following article provide some of the key steps and focus points that every archival team needs to consider (if not all), for successful data archival along with examples using Oracle ERP Applications as reference, wherever possible, for better understanding of the topic

Keywords: Database Archival, Archival Strategy, Data Isolation, Application Retirement prep

Introduction

Database Archival is a collaborative activity in an organization between business and IT for isolating the sets of data that is rarely or no longer used as part of normal operations. While there can be various reason for archiving, such as, storage recovery, performance, end of life etc., it is critical that the archival team understand the need, scope, and target users for proper archiving, without which, there is a risk of the whole process being ineffective, unusable, and potentially counterproductive, resulting in loss of time and money.

Why are we archiving?

A clear understanding of the need and potential benefit for the organization is vital for any project. Though archiving does not directly impact any business, it is important to know the reason for this activity and company's vision which will ensure that optimal planning and execution. One needs to remember that the data archival is an expense for the company over the data that does not help in day-to-day operations and is seldom used and hence the cost needs to be as minimal as possible, hence any gap in this step might result in significant deviation from the goal and the end product might not be effective or even unusable. Following are some of the reasons for archiving the data and it is highly likely that the company might have one among the following reasons.

Application Retirement

This is the most common reason for data archival. With advent of new technologies flooding the market, organizations are continuously reevaluating their existing applications for efficiency, capabilities and

productivity and replace or consolidate with a newer or better fit solution for their operations and retire the original application big-bang or in a phased manner. During such transition, more often than not, only active data and / or data created within the past few years will be carried over to the new application. All other data, that the soon-to-be retired application accumulated during its lifetime, which does not fit these criteria, the company might decide to archive them in accordance to company norms or legal requirement.

For example, if an organization decides to consolidate several Payable applications into a centralized solution, it might plan to convert only invoices that are unpaid. Any invoices that were already paid and accounted in prior period is of no significance to the organization. However, they might want to retain such invoices for reference or tax reporting or auditing purpose.

Reclaim Storage

From start-ups to multinational conglomerates or a simple app on a mobile phone to federal tax authority, everyone collects a wide range of data to analyze and steer the business or country in the right direction. These days companies collect, retain, and analyze a plethora of data to stay competitive in the market. Though the cost of storage devices has steadily decreased over the years, it is still not an economical to continuously increase the storage capacity to accommodate such high volume.

Additionally, the organizations need to consider the other costs that comes with increase in storage, such as licensing, maintenance and cost they might have to incur to maintain the efficiency, such as increase in CPU or memory or even go for more powerful server, etc. A study shows that around 80% of the data that is older than 90 days are inactive and not used except for auditing and reporting. Retaining such large volume of data in a powerful primary storage serves no benefit to the company, on contrary is a overhead.

With more and more businesses moving to cloud, the organizations are billed based on the amount of storage consumed. Hence, it is a direct financial impact to keep data in the main infra and pay the cloud provider for data that is seldom used.

Usually, Security and Protection companies retain the footages from the recordings from CCTV or other devices anywhere between a week to six months or more, depending on the commitment with the customer. However, most users do not look-back beyond a week. Since view files are much bigger in size, retaining all recordings in the main server might not be the most effective option, instead, an optimal duration can be identified and retained for quick retrieval and the rest can be archived.

Improve Efficiency

This is in part corelated to the prior section. In a fixed server capacity Data volume is directly proportional to turn-around time, i.e., as the amount of data collected increases, longer the time taken for the application to parse and return the result, in other words decreased performance. The parser does not know if the data is active or not and even though almost all requests pertain to just 20 – 25 % of the stored data, it needs to scan through the entire table to find the result. Agreed that the database providers are aware of these stats and are continuously trying to fine tune their search engines and algorithms accordingly, yet it is a hard fact that there is only so much that can be done and most product companies trying play catch-up to the exponential increase in data volume.

Every year, a substantial amount of the IT budget is allocated for upgrading the infrastructure to maintain or improve the efficiency and small and medium scale industries, which have limited cash reserve might find itself in a very difficult situation.

In an organization, which retains all their financial transactions in the main server, in the first year, to retrieve the Journal balances for the current Fiscal Year, the parser needs to run through one year of data. However, in the tenth year, the same needs to be done for ten years' worth of data. While most databases provide indexes for faster retrieval, any query outside the indexed columns will display a substantial decrease in turnaround time.

P.S. The one factor that is common across all the above reasons (or any other reason that may be) is the cost. Like any business initiative, this is a one of the most important deciding factors. Every organization looks at the return on investment and closely weigh in on the benefits, either in terms of savings or business development or to avoid penalties.

What data are we archiving?

Every database underlying an application will have a wide verity of data ranging from master data to transactional data to set-up data to reporting data. One need to know which set of data needs to archive and what to do with unarchived data. This largely depends on the company policy, business requirement and of course, government mandate.

Master data are the core of any business. Sone examples of master data are Inventory items, suppliers, customers, etc. There are almost never obsolete since they need to be available as a whole in the primary storage for business to run smoothly or fully converted into the replacing application prior to retiring the existing application. Also, they take up very little storage and rarely contribute to efficient issues due to the relatively low volume.

Set-up data is key for the application to function properly, and their significance can be described as all or nothing, i.e., this type of data need to remain as a whole in the primary server for the application to function correctly (if work at all) and can be fully discarded once application sun set.

Foe example, inventory items are master data type and should avoid being archived since they can be activated anytime and used by the business. Similarly, Item categories are set-up data and can never be archived as long as the application is alive and can be discarded once the application retires, since the new application might have its own native categories.

Predominantly, the reporting data and transactional data are almost always considered for archival since these take up most of the storage. An example of reporting data is the GL balances, which is the final consolidated data by accounts of all the journals for any given period. These types of data are frequently requested for auditing and justification, even after years and the organizations are obligated to provide these per company or government laws.

Transactional data are underlying records which to bundle up to the reporting data. These are generated by daily business activities and make up around 95% of the total storage. Transactional data are extremely important when they are active, but when closed, they are mostly considered obsolete, (though certain transaction are used in planning and research). These transactions have different levels of granularity, where one header record might have multiple lines and each line have multiple distributions and henceforth.

An example of transactional data is Purchase Orders. These records are extremely important to the business, however, once they are received and invoiced, they are virtually of no use to the business. These purchase orders and invoices, when accounted would record a consolidated financial entry in General Ledger, which is nothing but reporting data. it is very common for external audit and tax authorities to request for data from general ledger, however, there might also be additional request to show transactional

evidence for these Journals, in which case, the organization need to provide purchase order details, but not obligated to provide line or distribution level detail, which can influence the archival decision.

Who are the target users?

The next logical step is to understand who, or which team might use the archived data. While it is natural that all team would prefer to have access to historical data, it is practically not a viable option. Since archival is see an overhead, the amount of data, infra needed, maintenance and access need to be as concise as possible. The request from each team needs to be weighed in and determine if the access is absolutely necessary. For instance, both purchasing, and production team might need access for determining the buying trend and plant performance, however, while the former is key for business for planning and negotiation, the later might really not be important as the historical efficiency factors might not be the same as the current situation.

Additionally, the target users are dictate the kind of access and data availability. While majority of data retrieval from archive is IT centric, there might be certain team, who would want front end screens and / or reports that might be needed, such as the audit or tax team might need a quicker turnaround and rely on IT completely, as they might have more frequent requests compared to other teams, where as payables team might have a rare request to check for payments made in a certain period, which does not justify building an user interface.

How are we archiving?

Now that we arein the actual archival phase, there certain steps, validations, and precautions that need to be taken for successful archival as follows –

One-Time or Continuous Archival

Data archival, can be a one-time activity or a continuous process adopted by the organization for periodic maintenance or better manage the volume. Each of this option have their own sets of pros and cons and might need different solution.

One-Time Archival is usually done when the application retires or if there is an issue with the storage, whose solution involves cutting down the data volume. In this scenario, there possibly is no predefined template or process defined and would need a long discovery phase to determine the archival strategy. But this provides a much clearer visibility and an opportunity of design a tailor-made solution that would suit the business.

On the other hand, the continuous archival is a much-appreciated approach, since the business has adopted a process to periodically isolate the unused data. In this approach, the requirements are pre-defined, and the entire process is likely automated. This also means the in each iteration, we are only dealing with delta records and the archival process completes rather quickly. However, an automated process means that less visibility and supervision, which might deliver undesired results. these cycles need to be carefully planned to avoid any active data being archived. Also, any data anomaly or deviation might not be considered in the entire process and inadvertently cause data corruption or loss.

Data Fluidity

Data Fluidity is arguably the most critical step in the archival process. It is quite common that an application is retired in a phased manner and the decision is made to archive similarly. The archival team needs to make absolutely sure that the data that meet the criteria is static prior to cut-over and does not change during or

post cut-over. Since archived data is non transactional and used for reference and reporting, there is little to no room for updation. Any change in data after it is archived defeats the purpose and makes the data unreliable.

Data Archival and Accessibility

This is the section, where the team needs to determine how to carry over the data from the primary storage to archival system. For smaller databases, usually, the entire database before a certain date will be moved to archival system. The volume of data might be relatively insignificant, that the entire database can be archived. For much larger databases, there needs to be balanced approach where the future need for the data needs to be determined and archived accordingly. For example, while archiving General Ledger data for tax and audit purpose, one must also keep in mind the possibility that the audit team might ask for additional transactional evidence, in which case the team needs to identify the level of commitment and add the records to archival.

Also, there might be several intermediate tables / structures that merely add as connectors or additional information which are not needed, and the company is not obligated to retain and eliminate or merely retain only the important fields. To site an example, in Oracle XLA tables are significantly large tables merely act as a link between transactions and GL table and serves no purpose of their own. hence while archival, the team can design a solution only to retain the links as part of attributes or reference tables, there by eliminating a significant amount of data.

Data Security

Data security standards that are applicable to the primary storage, applies to archived data too and by extension, data compliance. Any Personal Identifiable Information (PII) must be properly secured or encrypted based on the business requirements. The organization security defined as part of separation of Duties should be enforced to archived data too. Every data set that is getting archived need to be reviewed to find the right set of users who can view the data and restrictions placed on the same. For instance, the overall financial books report is a sensitive information that should only be viewed finance controller team. Similarly, price negotiations details with the suppliers are confidential, and should be for contract team's eyes only. Any breach or violation of security even after archival, would result in legal action affecting the organization reputation and financially.

Archival Date

It goes without saying that the archival team needs to have a clear idea as to when the archival activity starts and how far before are we archiving. As part of the discovery phase, there will be a fair idea on what data will be archived, however, unlike data type, archival date will not be defined in the system. Identifying the archival date, provides the team ability to analyze the data volume, plan the migration strategy and the implementation time. Quite often, the date range requirement would be ambiguous, such as x years of data will be retained, x1 years of data will be archived. However, the actual date when the archival starts might make a difference in the volume and ability of archival.

Also, its preferable that the archival data's start at the end of a period, quarter, and even better fiscal year. It gets far easier for the reporting teams to generate reports when all of the requested data in in a single system, rather than read through multiple systems. Similarly, its is highly not recommended that the dates be mid-period for the same reason mention above.

End-of-Life Planning

This is an often-neglected step, and usually several assumptions are made resulting incoherency between teams. While decision to archive is made, there is also a need to determine how long the data stays in archival, before it is deleted for good. For instance, a HR team might want to retain the historical payroll indefinitely, however, the legal team would have emphasis on retaining only seven years of data as per the government mandate and to eliminate any additional obligation and IT team might raise concern over utilization of the allocated archival space.

There needs to a coordinated discussion among operations, legal and IT team (or any other team who need be) to come up with a definitive timeline for archived data retention.

Disaster Recovery

This is yet another overlooked step in any project. Post archival, its is likely the data is removed from the primary storage, rendering the archival system being the only source of the data. The archival team needs to ensure that disaster recovery is amounted as part of the entire project, tested and backed up so that in the unfortunate event of data loss, either accidentally or deliberately, the organization is able to restore the necessary data.

Conclusion

Though deemed insignificant, archival is and integral part of an application life cycle and needs to the performed with utmost precision, since in most cases, data once archived, it serves as the only place where the data is available and any oversight might result in invalid or unavailable data resulting in a complex and expensive recovery, if at all possible. Thought this article covers the key points and checklists for successful archival, one need to always read the landscape in detail to design appropriate archival solution.

References

1. What Are the Benefits of Data Archiving with Oracle?https://www.oracle.com/assets/why-archive-with-oracle-article-seo-2608143.pdf
2. Overview of Archive Storage
3. https://docs.oracle.com/en-us/iaas/Content/Archive/Concepts/archivestorageoverview.htm
4. Understanding Archiving Strategyhttps://docs.oracle.com/cd/F44200_01/pt859pbr2/eng/pt/tadm/concept_UnderstandingArchivingStrategy-077a9e.html