Data Re-archival in IT Application Retirement Scenario
A Case Study

Vidyasagar Uddagiri[1], Amarendra Mohanty[2] and Biswaranjan Jena[2]
[1]*Tata Consultancy Services Limited, Hyderabad, India*
[2]*Tata Consultancy Services Limited, Bhubaneswar, India*

Keywords: Information Lifecycle Management, Application Retirement, Data Archival, Re-archival.

Abstract: The prolific expansion of business operations globally, multi-channel operations and many such newer paradigms are driving voluminous growth for many businesses. This is resulting in tremendous data volume growth within supporting IT systems. Increased need for fact driven decision making and regulatory compliance requirements needs is driving the need for retention and storage of historical data for longer duration. Historical data being referred occasionally does not warrant storage using expensive database systems unlike transactional data. Information Lifecycle Management (ILM) is an emerging discipline within Information Technology that addresses this problem. Data Archival is a concept within ILM used to retain necessary data for reference by live applications, while preserving historical data. The challenges faced during the execution of a Data Archival project by a Fortune 500 organization, the methodology formulated and implemented to work-around the challenges are described in this case.

1 INTRODUCTION

It is a widely known fact that the rate of data growth in business enterprises is increasing at a profound rate. Changing regulatory requirements and new paradigms in business models and ecosystems such as rapid globalization, multi-channel operations, impact of five digital forces are the driving forces to these enterprises for retaining data for a longer period of time as compared to a decade ago. This trend is expected to continue in future. It is also observed that over 80% of such enterprise data is not actively used and is retained for reference purposes only. However, the retention of such data on expensive storage media or databases makes little business sense from the standpoint of cost containment, risk reduction or operational efficiency.

Information Lifecycle Management (ILM) is shaping up as an emerging disciple to address this situation. ILM comprises the policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost effective Information Technology (IT) infrastructure from the time information is conceived to its final disposition. Data Archival, a part of the ILM discipline is a process of preserving and protecting data as inexpensively and effectively as possible for future use.

A case study of a large automobile component manufacturer that embarked on a large IT landscape transformation program that included archival of data for some applications getting retired is the subject of this case. The project success criteria, challenges faced in achieving the success criteria and the methodology evolved to overcome the challenges are described in this case.

Section 2 provides an overview of the literature relevant for the project context. Section 3 describes the detailed project situation, the risks that manifested and the project execution methodology that was evolved as mitigation mechanism. Section 4 describes the details around the design of this methodology, Section 5 describes the results of the methodology applied and reflection on the results and recommendations for future use, while Section 6 provides the conclusion and suggests the future scope for further research.

2 RELATED WORK

Business data is a very critical asset for modern enterprises. The corner stone of Information Lifecycle Management (ILM) revolves around management of this critical asset.

Uddagiri V., Mohanty A. and Jena B..
Data Re-archival in IT Application Retirement Scenario - A Case Study.
DOI: 10.5220/0005109105760581
In *Proceedings of the 9th International Conference on Software Engineering and Applications* (ICSOFT-EA-2014), pages 576-581
ISBN: 978-989-758-036-9

Petrocelli(Petrocelli, 2005) provides a systematic, coherent approach to planning and implementing cost-effective data protection as part of ILM, while, Smith(Smith, 2009) provides a framework for defining and assessing the maturity of Enterprise Information Management.

Ball(Ball, 2012) provides a comparitive analysis of various data management lifecycle models and relevance of data archival within each model and Olson(Olson, 2008) provides a detailed step by step technical procedure for implementing Data Archival solution.

Khadka(Khadka et al., 2014) provides an insight into industry perspective on legacy systems, drivers for modernization and challenges faced in the course of modernization. However the scope does not include methods/approaches to overcome the challenges.

There are some approaches described with respect to flexible approaches in Information Systems Development (Ahituv et al., 1984) that provide foundation elements for devising newer methodologies using a combination of various elements.

Practical insights and best practices into Data Archival are not yet included in the academic / research realm and are confined to the documentation from product/service vendors.

Thus, there exists a gap in literature with respect to the project execution methodologies or best practices to handle risks and challenges not related to technology, with specific reference to data archival projects.

3 PROJECT SITUATION AND SCOPE

A US based fortune 500 automotive components manufacturer Auto-Sup-Co (name changed to protect identity) embarked on a large transformation program to modernize the applications on their legacy Mainframe platform to improve their business flexibility, agility and reduce total cost of operations. The program included replacement of several applications with equivalent modules of an Enterprise Resource Planning (ERP) system (Software Engineering Institute, 2000), while others became obsolete.

For several legacy applications, the modernization involved only replacement of the functionality by equivalent ERP modules, without any data migration. However, there was a regulatory requirement to retain the historical legacy application data for a period of 7 years from the date of its creation. Hence, a Data Archival solution (Ref. Appendix) is necessary. While the application migration work was managed by Auto-Sup-Co with support from their current information systems services partner, they were looking for another information systems services vendor experienced in carrying out the Data Archival project of their transformation program.

Data archival implementation is relevant in the following scenarios across any business enterprise:

- Data Archival for live applications
  - Retain information for an extended period of time for conformance to various legal and regulatory requirements
  - Perform strategic business analysis on historical business data
  - Improve operational costs and application performance by moving off infrequently referred data to a lesser cost storage medium
- Data Archival for Application Retirement
  - When legacy applications are retired some data may have to be retained for compliance reasons even if they are not migrated to newer applications

There are two solution variants for Data Archival namely Bulk Data Archival (BDA) or Change Data Capture (CDC) based on project context as explained below.

- Bulk Data Archival
  - When legacy applications retire, there may be a need for one-time archival.Bulk Data Archival (BDA) is the right choice for this situation.
- Change Data Capture
  - Data archival may be necessary in live applications at pre-determined intervals to improve application performance. CDC is the right choice for this situation. CDC includes a bulk data archival in the first pass, followed by periodic increments.

After a rigorous selection process including evaluation of technical and financial proposals, proof of concept execution results and reference checks, the largest Indian IT Services vendor, TCS was awarded the project.

As part of the project's technical proposal, TCS recommended a commercial data archival product to implement the archival solution as a better alternative over a custom built solution. As a strategic decision, TCS also recommended Auto-Sup-Co to opt for CDC solution variant because of the long term benefits that can be reaped. However Auto-Sup-Co did not see a business case at that point in time for CDC, considering its high implementation costs and complexity and rather opted for BDA variant of the data archival product.

Auto-Sup-Co purchased the product while TCS played the role of System Integrator. The system integration activities for the Data Archival project included configuring, installing and implementing the solution and carrying out the Data Archival process for the business application data in scope.

The scope of the project involved archiving about 5 terabytes (TB) of data related to tens of applications and thousands of data sources. The estimated duration of this project was 8 months starting in mid-October 2012. The major constraint that was considered for the project planning by TCS was the end timeline of the maintenance contract for the legacy application platform as indicated by Auto-Sup-Co. It was also indicated by Auto-Sup-Co that the legacy application platform will be de-commissioned after the end of this maintenance contract. Any delay in the Data Archival project would mean that Auto-Sup-Co had to renew their legacy application platform maintenance contract thereby incurring additional expenses. Thus end timeliness and Total Operational Cost (TCO) of the project were critical success factors for this project.

The end timeline for the Data Archival project was planned to be 5 months prior end date of Auto-Sup-Co's legacy application platform maintenance contract. This left sufficient time to discover any issues after the solution rollout and fix them while the legacy application platform is still available.

The data archiving schedule split for each application, was based on the original schedule of application migration as shared by Auto-Sup-Co. In the application retirement scenario, Data Archival is carried after the respective applications become inactive, lest complexities due to validation of data mutation and impact to service levels hinder the project progress and manifest as risk, which was clearly highlighted as an assumption in TCS technical proposal.

3.1 Risk Manifestation and Remediation Planning

During the course of project execution, while analysis phase was in progress, TCS project team realized that there would be a delay in the application modernization project. This delay in turn impacted the data archival project schedule. A plausible option here for any System Integrator in such a situation was to let the project be kept on hold, till the remaining applications become inactive and become ready for archival.

The consequences of this approach were:

- Cost implications to Auto-Sup-Co as part of contractual obligations for schedule delays
- Loss of momentum on the project leading to additional project restart costs
- Insufficient time for monitoring the project results after full roll-out

These consequences being unacceptable to Auto-Sup-Co, both parties to the project, TCS and Auto-Sup-Co have conducted a brain storming session to discover options for implementing the project without putting on hold while leaving reasonable time after the project end date to ensure any issues that are likely to arise, can be resolved.

4 DESIGN OF RISK MITIGATION METHODOLOGY

However there were some guiding factors laid out for generating options that can fast track the project. The following factors needed to be provided high importance:

- Low TCO
- Ease of execution
- End user satisfaction
- Leveraging the investment in BDA solution variant
- Alignment with the application transformation program at Auto-Sup-Co

The most plausible idea out of the many ideas generated in the session was to detail out the critical path of the data archival process and look at options for fast tracking. This exercise resulted in realizing 6 significant steps in the critical path for carrying out Data Archival. These steps are depicted in Figure 1.

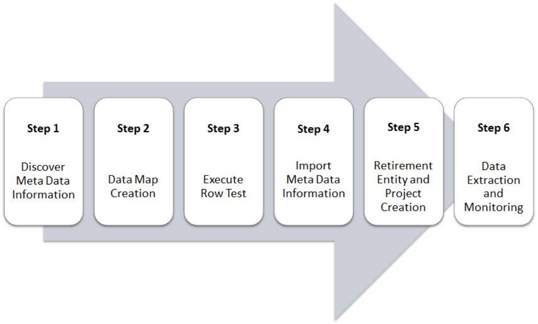
Step 1
Discover Meta Data Information
Step 2
Data Map Creation
Step 3
Execute Row Test
Step 4
Import Meta Data Information
Step 5
Retirement Entity and Project Creation
Step 6
Data Extraction and Monitoring

Figure 1: Critical path steps of Data Archival.

4.1 Outcome of Critical Path Analysis

It was quite evident from the critical path analysis that Re-archival requires lesser duration as compared to the full archival cycle. A strategy of two pass archival process was worked out as a realistic approach.

A cut-off date was mapped for each application to consider the data existing on that date as the baseline for archival purposes. The application data would be archived as per original schedule. This was the first pass of data archival. Later once the applications become inactive, a second pass was decided to be carried out by refreshing the entire data archive with the source application data without recreating the metadata (Step 1 and Step2) since data structures were not expected to change. This two pass technique was believed to help in overcoming the challenge of project delays.

This new methodology was coined as Re-Archival since archival will be attempted on the entire data all over again. This strategy is expected to reduce the critical path duration since two steps will not be executed during Re-Archival.

TCS disinterred their corporate knowledge repository, connected with the Data Archival product vendor and also searched the literature available with respect to prior instances of Re-archival. It was realized that there were no prior recorded instances of implementing the Re-archival methodology anywhere in the world, before this project was executed.

Hence, a pilot was carried out to validate the technical feasibility of the solution and better understand the detailed timeline required for each step of Re-archival. The results of the pilot confirmed that Re-archival was a feasible technical solution.

The detailed planning for the Re-archival methodology was carried out in the last few weeks of the first pass of Data Archival so the results from the initial archival cycle could be leveraged to plan the Re-archival phase of the project as accurately as possible.

4.2 Validating the Decision

The experience and data gathered from the first pass of data archival for all the applications in scope provided good insights that helped in the preparation of an accurate execution plan for Re-archival phase of the project.

It was observed that while the effort and time duration for Step 1 to Step 5 were identical for all types of data sources, they varied for the last step (Step 6) based on the type of data source (namely relational database, networked database, hierarchical database or file type data store). While the effort required for the 6th step was the same across various data sources, the elapsed time required to load the data once it is triggered, varies widely (from minutes to days).

The average effort for each step within the data archival process was determined to understand the timelines that had to be provisioned for the second pass of data archival or Re-archival. The results of these observations are mentioned in Table 1.

Table 1: Effort split among each Data Archival step.

Step No.	Step Description	Percentage of effort split
Step 1	Discover Meta Data Information (of source system)	18%
Step 2	Data Map creation (in Data Archival Tool)	36%
Step 3	Execute Row Test (to ensure correct data source is accessed)	9%
Step 4	Import Meta Data from Source System (into Data Archival Tool)	9%
Step 5	Retirement Entity and Project Creation (in Data Archival Tool)	18%
Step 6	Data Extraction and Monitoring	10%

About 20% of the data sources had exceptions with respect to meta-data creation because of their legacy data structure. The meta-data for exception sources had to be converted into the common structure of the Data Archival system. Consequently, during re-archival, the steps 1 to 6 had to be executed in total for these exception scenarios. Fast tracking by additional resource augmentation seemed a feasible option to handle the exception cases.

One difference between the first pass and second pass of Data Archival is to follow a differential solution implementation strategy for the two different classes of data sources:

1. Data sources with no meta data structure exceptions
2. Data sources with meta data structure exceptions

Based on the data collected during first pass of archival, it was computed that an incremental service cost of 35% was necessary for Re-archival by leveraging BDA itself. This was cost-effective as compared to the alternate option of putting the project on hold and restarting.

5 RESULTS AND REFLECTION

The diligence put in developing the Re-archival methodology enabled evolution of a detailed Work Breakdown Structure (WBS). To support execution of tasks in the WBS and accelerate the execution process, a robust software factory execution model (Nomura et al., 2007) was deployed by TCS.

Re-archival was carried out as a collaborative exercise between TCS and Auto-Sup-Co, leading to the success of the project overcoming the schedule challenges.

There were technical challenges discovered during the course of implementation of the Re-archival process. Since this approach did not have any precedence earlier, everything about Re-archival was learnt in the course of project execution, the hard way.

Though the Re-archival methodology is tried and tested once, it is still not recommended as the option of choice. Better collaboration by including TCS in the overall transformation program planning, thereby providing greater insight into the program schedule, would have resulted in more realistic planning and identification of other Data Archival products or variants to handle possible schedule challenges.

However, TCS experience with Data Archival projects states that CDC solution variant costs at-least 50% more in product licensing costs and 25% more with respect to implementation by System Integrators. A post project TCO analysis also vindicated this decision. The project TCO including Re-archival phase was 15% less a CDC solution.

The re-archival methodology resulted in achieving quicker timelines for overall project completion. The re-archival for the last set of data sources was completed in mid-October 2013 providing a reasonable 10 week time prior to the end date of the current contract for the legacy platform maintenance. Since then, not a single deviation was reported. This proves the robustness of the Re-archival methodology conceptualized and implemented by TCS.

However, if one were to really evaluate Re-archival methodology for future use, the authors suggest that a diligence be carried out to compare all available options as cited in Table 2.

While considering option 1 or 2, due consideration has to be made to quantify the opportunity cost for not considering CDC upfront.This will enable project managers in taking an informed decision in evaluating all the options before making a choice.

Table 2: Applicable cost elements for each option.

Option No.	Option Name	Cost elements to be included
1	Hold and restart project	Cost of BDA and Cost of project restart
2	BDA with Re-archival	Cost of BDA and Cost of Re-archival
3	CDC	Cost of CDC only

6 CONCLUSION

As part of this case a new methodology namely Re-archival was conceptualized and implemented. This was suitable considering the context of the project situation.

However, the authors still recommend that rigorous and collaborative planning by enterprises including their systems integrator (either in-house or outsourced) will help in better planning of risks and their mitigation, lest availability of such a methodology may result into an excuse for not putting enough diligence into planning of Data Archival projects.

Re-archival may turn out to be an effective option after duly evaluating all available options and solutions with due consideration given for data archival scenario, data volume and critical success factors for the project.

The merits of this methodology were bench marked against the data from only one project case. Hence, there exists scope for further research and discussion to include data from larger number of projects before it can be optimized.

ACKNOWLEDGEMENTS

We would like to acknowledge Vasudev Rao, Gaurav Joshi and Smita Prava Mishra for the valuable suggestions and feedback provided during the course of manuscript preparation.

REFERENCES

Ahituv, N., Hadass, M., and Neumann, S. (1984). A flexible approach to information systems development. *MIS Quarterly*, 8:69–78.

Ball, A. (2012). Review of data management lifecycle models. *Opus, University of Bath, UK.*

Khadka, R., Batlajery, B., Saeidi, A., Jansen, S., and Hage, J. (2014). How do professionals perceive legacy systems and software modernization? In *Proceedings of* *the 36th International Conference on Software Engineering*.

Nomura, L., Spinola, M., Tonini, A., and Hikage, O. (2007). A model for defining software factory processes. In *19th International Conference on Production Research*.

Olson, J. (2008). *Database Archiving*. Morgan Kaufmann.

Petrocelli, T. (2005). *Database Protection and Information Lifecyle Management*. Prentice Hall.

Smith, A. (2009). Enterprise information management maturity: Data governance's role. *EIMI Archives*, 3:1.

Software Engineering Institute, C. M. U. (2000). A survey of legacy system modernization approaches. In *Technical Note CMU/SEI-2000-TN-003*.