Archiving the insurance data warehouse

Surajit Basu, IBEXI Solutions

Surajit Basu is an IT consultant with over 20 years of experience in design, development and implementation of enterprise-wide solutions in different industries, especially insurance.
Surajit Basu received his B. Tech in Computer Science from IIT Kanpur, and his MBA from IIM Calcutta.

He can be reached at surajit_basu@ibexi.com

Executive Summary

Over time, as the insurance data warehouse grows larger and larger, the much-postponed archival becomes essential.

At this stage, many questions come up – from “What's in the insurance warehouse?” to “What should we keep?' to “How should one archive?” and finally to “Who will pay for the project”.

We seek the answers.

Table of Contents

To archive – or not to archive?

That is the question. The first, at least.

Data in a warehouse grows like the things in a large family home. Initially, there is a lot of space. Over the years, items accumulate in every part of the house. Some of these items are valuable, some are not. After many years, we have no reliable information of what is available. It is difficult to find some items that we know were available and used a long time ago. We know some of the items "must be" junk, but we don't know which items and where they are. We also know that we should not throw away the family heirlooms with the junk. We are worried that as soon as we throw some things away, someone will want one of those things. We know we cannot take the same approach to cleaning the basement, attic and the living rooms.

It seems a lot of effort to catalogue the items and separate the useful from the useless, to decide what to keep – and throw away, to decide how to clean up the various parts of the house, and to do the spring-cleaning. So we postpone it to the next spring. Until, we run out of space, or it becomes such a mess that we can't find the useful items.

Data warehouses have a similar story. Data warehouses start big and get bigger. In a "Data Warehousing Satisfaction Survey" 80 percent of data warehouses surveyed were 1TB or larger [1] . Over the years, data accumulates, and much of the data - 85% according to some estimates [2] – becomes irrelevant, unusable, or rarely used. The sheer size of the data causes slowness in searches, leading to longer response times. There is need for greater database space, more and faster processors, more administration staff, design changes. It takes more and more time and money.

But there are risks of archiving critical data, and of retrieving them if required. No easy categorisation of useful and useless data is available. Often, more than adequate space is allocated initially, but it runs out faster than expected. And only then, do the questions come up:

- What's in the insurance warehouse?
- Is there a difference in archiving the different layers?
- What if the archived data is needed later?
- How do we decide what to keep – and archive?
- What should we do now to prepare for future archival analysis?
- What guidelines can we create for future archival ?
- How should one archive?
- Who will pay for data archival and why?

What's in the insurance warehouse?

A typical insurance warehouse accumulates data from many sources. The variety of these sources represents the typical insurance software landscape. In it, there are usually many business applications, and in some cases, multiple major modules within one business application. These range from the multiple core systems – for managing policies, claims, reinsurance, channels and incentives, valuation – to the generic – for managing tasks and workflow, human resources, finance. It is also common to find miscellaneous yet critical information recorded in spreadsheets. In addition, external data – related to the market and competition – are also available.

These are re-organised in the data warehouse in appropriate business-driven sections. Business rules are available in the Master Data. All individual and organisations for different roles - policyholder, insured, member, payor, beneficiary, assignee, trustee – and maintained in one section. Similarly, information on policies, the endorsements, policy servicing transactions, claims, reinsurance, the insurance collections & payments, channel and incentives – are collated across all source systems. All insurance related accounting is usually similarly collated. Non-insurance accounting recorded in the General Ledger is collated with this. True earnings or valuation information derived from all these sources are stored. Informations on tasks, human resources are also collated.

In addition, in order to simplify reporting and usage, information is re-organised and summarised for the use of various stakeholders – departments, management teams, and regulators.

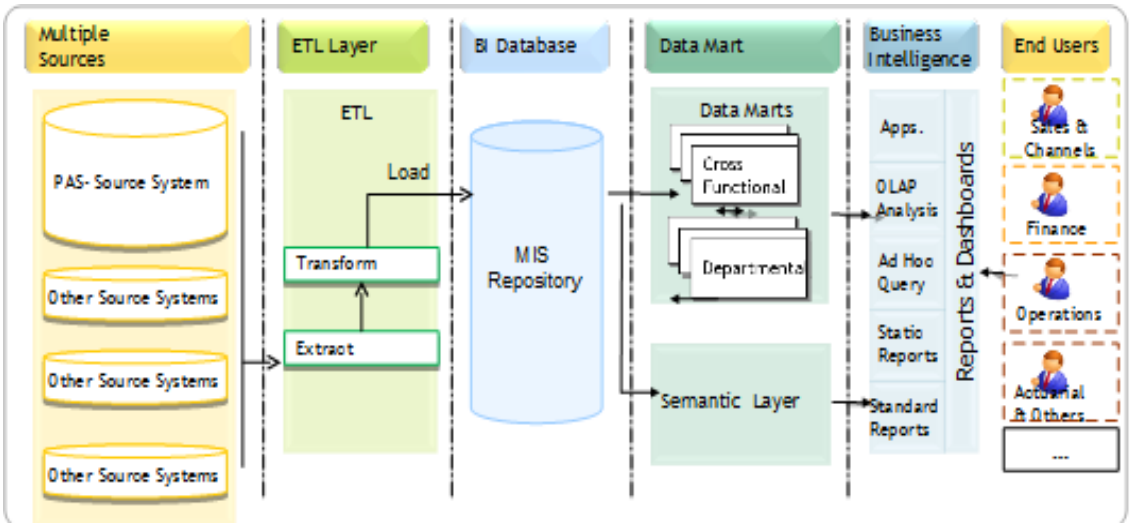


Figure: A standard data warehouse (Biz-I, IBEXI Solutions)

Typically, the three parts of the data warehouse are organised differently and store different types and forms of data. A typical view would be:

Source Systems	The base of the warehouse	Usage summaries in the warehouse
Policy Administration Claims Reinsurance Channel management Valuation Workflow Human Resources General Ledger Excel for miscellaneous information External data – market, competition	Masters/ Business rules Client – with history Policies Policy Transactions Insurance Collections, Disbursements Claims – with history Reinsurance Channel structure, hierarchy and history Commission Insurance Accounting Journals Earnings / Valuation Workflow Human Resources Accounting Journal (General Ledger) Insurance Market and Competition	New Business cubes Marketing cubes Operations cubes Claims cubes Sales cubes Actuarial cubes HR cubes Finance cubes Management cubes Regulatory cubes

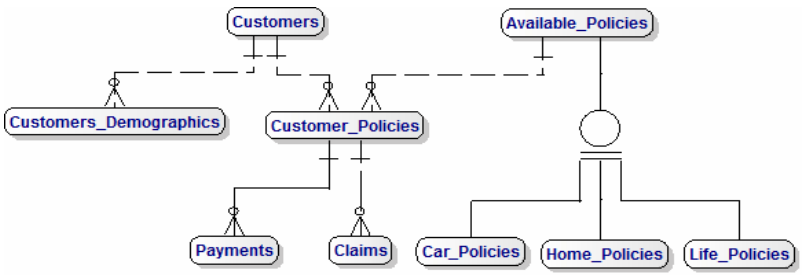
The different layers of the data warehouse – transactional and informational – are typically structured differently. Transactional data is stored similar to operational systems using 3rd Normal Form (3NF) design with a focus on referential integrity, consistency and speed for real-time usage. Informational data – oriented towards usage, typically with summaries – uses a dimensional Star Schema design.

Is there a difference in archiving the different layers?

Since the design approaches for the different layers of the data warehouse are different, the issues for data archiving are also very different.

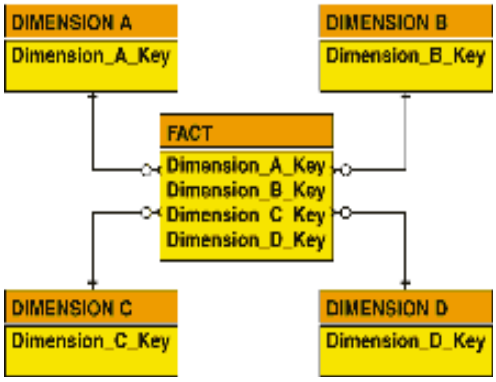
Transactional data uses 3rd Normal Form (3NF) design with a focus on referential integrity for consistency. Thus, for archival, maintaining referential integrity within the database is essential.

For each candidate table where archival is planned, child tables need to be examined to ensure that related rows are also be archived. If the rows of the child tables cannot be archived, then it should be checked if the absence of the relevant parent records will make the child tables



inaccessible. This may be so, depending on the primary keys of the tables and the access methods. If accessing the rows of the child tables requires the rows of the parent table planned for archival, then design changes may be required to implement the archival.

On the other hand, the informational data – oriented towards usage, typically with summaries – uses a dimensional Star Schema design, where archival can usually be achieved much more easily with less effort. The fact tables should be the largest tables and therefore typical archival targets. Usually, it is easy to identify the archival candidate rows on the fact tables – typically based on a date key. There is no need to archive related dimension rows.



What if the archived data is needed later?

It is difficult to say when the archived data will be required for restoration. At that time, there will probably be an urgency for restoration. So, the process for restoring archive data needs to be finalised – and tested - during archival.

For the transactional 3rd Normal Form (3NF) design, for restoration to work, referential integrity is essential. Thus, when candidate rows have been identified for restoration, related parent table rows – if archived - also need to be restored. Restoring archived fact table data is much simpler. Just the specific fact rows/partitions need to be restored.

In either layer, there can be two possible restoration target models. In the more common approach, the data is restored to the same tables from which the data was archived. Existing models and queries will automatically include this data. It is also possible to restore the data into a separate archive table or set of archive tables . Existing models and queries will not include this data. Either these have to be modified, or new models and queries have to be designed.

If the restored data is needed in multiple queries and usage shall be similar to the existing data, e.g. for interactive drill-downs, then the first method is probably the better fit. Also, if the data is likely to be used repeatedly, over a longer period, this is probably the better fit. Data space management for the restored rows needs to be considered, as the table may become very large, leading to performance issues with current model, loading, queries, usage. Note that re-archival of the data will have to go through the original archival process – including identifying related rows in other tables.

On the other hand, if the restored data is needed in a few one-off queries and with different usage, e.g. for arriving at a temporary summary for the past, then the second method is probably the better fit. Note that re-archival of the data will have to be simple; just the deletion of this table will suffice. Data space management for the archive table is also easier, as it is independent, and will have less impact on the performance of the existing model, loading, queries, usage.

How do we decide what to keep – and archive?

After deciding on the need for archival and the archival strategy, the difficult questions of what to keep and what to archive come up. What are candidates for archival? The main considerations are size and usage.

Firstly, the large tables need to be identified, and monitored. In the insurance warehouse, the largest tables in the base data are usually in Insurance Accounting Journals, Earnings and Valuations, and Workflow.

There are also likely candidates for each line of business. In health insurance, especially for Outpatient coverage, with the high claim frequency, Claims transactions can be candidates. In life insurance, with long-term policies, Policy Transactions especially monthly deductions and fund deductions for unit-linked policies can be candidates. For P&C policies, Unearned Premiums.

In addition, for fact tables, the allocated facts – where 1 original transaction value creates many facts – are likely candidates. For instance, a likely candidate is an Accounting cube with a grain of many sub-ledgers (e.g. product, branch, department) and transactions allocated to these sub-ledgers (e.g. head office rent allocated to each sub-ledger combination) for analysis.

A table similar to this for the largest 50 tables helps in the analysis:

Layer	Table name & Description	Space allocated currently	Space allocated at start of previous year	Growth %	No. of rows currently	No. of rows at start of previous year	Growth%

Secondly, consider usage. It is important to monitor report usage and data usage to determine what summaries/cubes and base data may could be archived. Many data warehouse projects do not track report usage; over the years, many reports are added but none are deleted/ archived. Often, many scheduled reports go on using resources, adding to time and money for the enterprise. It is important to track these early, using a table like this:

Report name	Report Description	Report Owner - Department	Report Owner - Person	Ad-hoc/ Scheduled	Ad-hoc report: last run date

For ad-hoc reports, the "last run date" is usually indicative of usage. We need to be careful about analysis using this. There may be important reports which are used only once a year, or when regulators ask for analysis. For all reports, and especially scheduled reports, it is important to track the report owner, who should be able to explain if it is still needed.

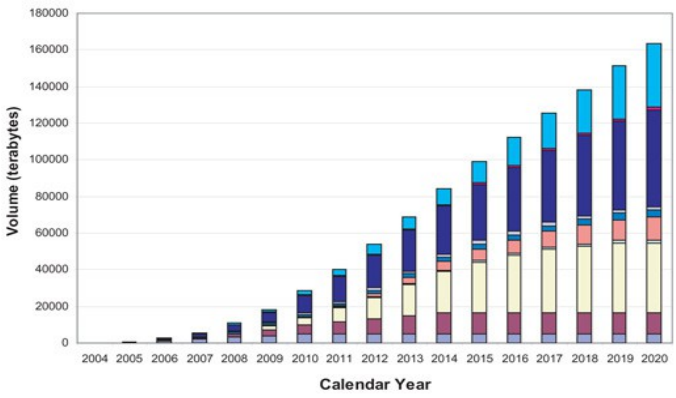
Similarly, data usage history is critical for analysis of how long to keep the data. Database can track usage and provide information on utilisation of tables, and partitions. It is important to partition using the relevant factors – e.g. policy year, accounting year – to simplify archival.

For the transaction layer, old transactions can usually be archived. Detailed Insurance Accounting Journals, Earnings and Valuations, and Workflow tables can typically be archived for rows which are more than 3 years old. In health insurance, Claims transactions greater than 5 years old can be archived. In life insurance, monthly deductions and fund deductions for unit-linked policies greater than 5 years old can be archived.. For P&C policies, Unearned Premiums can typically be archived for rows which are more than 3 years old. In addition, policies which cannot be revived (E.g. death, cancellation, surrender, lapsed for 3 years) can be archived after 5 years. Related data – reinsurance, claims, accounts – can be archived with the policy.

For fact tables in the reporting layer, the ones with the lowest grain can be archived. It may be required to implement design changes to keep aggregates of the archived data.

What should we do <u>now</u> to prepare for future archival?

In order to prepare for future archival, it is important to collect information over time to enable reliable analysis. A monthly collection of data can be simple, and very useful. Such data can be used to estimate future growth, plan for sizing of infrastructure, and for savings for archival projects.



A table such as this can be used to analyse database growth.

Year & Month	Layer	Table name & Description	Space allocated currently	No. of rows currently

To assess the performance degradation with increase data over time, the memory and CPU utilisation should also be tracked.

Year & Month	RAM utilisation online - peak	RAM utilisation batch - peak	CPU utilisation online - peak	CPU utilisation batch - peak	Batch time required

For report usage, ensure that the BI environment can easily provide usage information such as "last run date" for each report. For all reports, and especially scheduled reports, it is important to track the report owner, which may change over time. So, this should be tracked from the project start.

Report name	Report Description	Report Owner - Department	Report Owner - Person

What guidelines can we create for insurance archival ?

After the first archival analysis, it is important to create the general guidelines for the archival. These may be summarised like this:

Layer	Table name & Description	Archival strategy	Archival condition	Comments
Transaction	Masters/ Business rules	No archival		Low volume, needed for past
Transaction	Client – with history	No archival		Needed for ad-hoc CRM analysis
Transaction	Policies	Archive	Last revival date > 5 years.	
Transaction	Policy Transactions	Archive UL History, and Transactions for specific policies	UL History date > 5 years Transactions for policies with last revival date > 5 years.	
Transaction	Insurance Collections, Disbursements	Archive	Accounting date > 5 years	
Transaction	Claims – with history	Archive	Closed date > 5 years ago	
Transaction	Reinsurance			
Transaction	Channel structure, hierarchy and history	No archival		Low volume, needed for past
Transaction	Commission	Archive	Accounting date > 10 years	

Layer	Table name & Description	Archival strategy	Archival condition	Comments
Transaction	Insurance Accounting Journals	Archive	Accounting date > 5 years	
Transaction	Earnings / Valuation	Archive	Valuation date > 5 years	
Transaction	Workflow	Archive	Task date > 3 years	
Transaction	Human Resources	No archival		Low volume, needed for past
Transaction	Accounting Journal (General Ledger)	Archive	Accounting date > 10 years	
Transaction	Insurance Market and Competition	No archival		Low volume, needed for past
Reporting	Finance cubes	Archive	Remove grain of Allocated Sub-ledger for Accounting date > 5 years; keep aggregates	

However, it is important to ensure that the archival guideline is not blindly followed, but reviewed periodically using the data collected since the last archival.

How should one archive?

Specific programs need to be developed to archive the data based on your specific archival guidelines. Target formats need to be decided first. Since the archive data does not have to organised for regular access, it is usually adequate to keep it in formats which minimise the storage capacity. Thus, for archives, it is adequate to store as text files rather than database tables. The text file can be compressed and stored in an external storage backup. Multiple copies need to be kept to prevent loss of archive data due to storage device corruption.

In some cases, the archival is simple e.g. for the accounting fact tables– where the archival is based on 1 dimension: accounting date, or the insurance accounting journal tables – where the archival is based on 1 field: accounting date. In such cases, a simple archival program - which selects the relevant rows, saves as text in a defined folder, and deletes the rows – is sufficient.

In cases where the archival is complex, it is preferable to create shadow tables for each table which is to be archived. The programs should copy data in an integrated way from existing tables to the shadow tables. A detailed reconciliation process should be conducted. After this, the copied records in the existing tables should be deleted. At this stage, before archival, it is safe to allow regular reporting process for some time, before the removal of the shadow tables. This ensures that – in case there is a need, the relevant rows can be restored easily. After this “parallel run”, a simple archival program - which selects the entire shadow table, saves as text in a defined folder, and deletes the table – is sufficient.

Who will pay for data archival and why?

The reality is that most businesses pay for not doing data archival! Data warehouses start big and get bigger. Over the years, data accumulates, and much of the data - 85% according to some estimates [2] – becomes irrelevant, unusable, or rarely used. The sheer size of the data causes slowness in searches, leading to longer response times. There is need for greater database space, more and faster processors, more administration staff, design changes. It takes more and more time and money. Archiving is a good way to improve the performance of the data warehouse, without investing more in IT infrastructure and maintenance.

The business case for archival requires an analysis of the costs and benefits, and identifying a return on investment (ROI). Based on the project scope, it is usually not difficult to quantify the costs. The challenge usually lies in quantifying the benefits in savings in storage, processing capacity, backup capacity, deferment of system upgrades, administrative costs, and performance improvements of the data warehouse.

One way to estimate the savings is to identify the costs if the current performance has to be maintained with the additional data. Based on the current size and the data growth over the last 3 years, it is possible to estimate the growth over the next 5 years. This provides the additional disk space requirements. To maintain current performance, increases in RAM and CPU are required. Based on the current memory and CPU utilisation and the increase in utilisation over the last 3 years, it is possible to estimate the needs over the next 5 years. It is important to note the requirement of memory and CPU sometimes increases faster than linearly. Also, increases in memory and CPU are not always possible on the same machine. Hence, a machine upgrade may be required in order to provide the additional memory and CPU.

Also, there are multiple environments – production, disaster recovery, user integration testing environments; the increases of all the environments need to be planned. In addition, equivalent increases in back-up storage capacity are required. To ensure back-up within a defined batch window, the back-up devices and software may also need upgrade.

There are additional administrative costs for maintenance of the increased database, backup maintenance. These can be estimated using a pro-rata model. In practice, though, costs are often higher than proportional, as much more expertise is needed for managing extremely high volumes.

The deferment of infrastructure upgrades, and the maintenance costs are major benefits derived from archiving. Also, there are benefits of performance improvements of the data warehouse, and the ability to derive complex analysis faster.

A summary of costs may be presented like this:

Savings components	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Disk space	Additional for 5 years			Move to new machine		
Memory	Additional for 5 years			Move to new machine		
CPU	Additional for 5 years			Move to new machine		
Disaster recovery	Equivalent increase					
UAT environment	Equivalent increase					
Back-up storage	Additional for 5 years					
Back-up device/ software				Upgrade back-up		
Administrative overheads	Additional costs	Additional costs	Additional costs	Additional costs	Additional costs	Additional costs
Other savings						
Total savings	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**

On the other hand, there is a one-time cost for the data archiving project, and the possible restoration of archived data. The probability of restoration and the quantum of data to be restored need to be assessed. Finally, the present value of the total annual savings needs to be compared with the estimated present value of the data archiving project.

Summary: Costs, Savings	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Total savings	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**	**Annual savings**
Total costs	**Data Archival project cost**					

We hope this will help you derive an adequate Return of Investment for the archival for your data warehouse.

References

1. Lou Agosta and Kevin Modreski. “The Data Warehousing Satisfaction Survey” DM Review Special Report, October 2007.
2. Noel Yuhanna. “Database Archiving Remains an Important Part of Enterprise DBMS Strategy.” Forrester.com, August 13, 2007.
3. Data Warehousing Meets Data Archiving in Information Lifecycle Management
4. Keep it Simple When Archiving Dimensional Data