WHITE PAPER: DATA DISCOVERY AND PROFILING

Demonstrating the ROI of Data Discovery

OCTOBER 2009

Malcolm Chisholm Ph.D.

INDEPENDENT CONSULTANT

Table of Contents

Copyright © 2009 CA. All rights reserved. All trademarks, trade names, service marks and logos referenced herein belong to their respective companies. This document is for your informational purposes only. To the extent permitted by applicable law, CA provides this document "As Is" without warranty of any kind, including, without limitation, any implied warranties of merchantability or fitness for a particular purpose, or noninfringement. In no event will CA be liable for any loss or damage, direct or indirect, from the use of this document, including, without limitation, lost profits, business interruption, goodwill or lost data, even if CA is expressly advised of such damages.

Executive Summary

Challenge

As data continues to grow in importance to enterprises worldwide, a glaring problem with its quality and accessibility has come to light. Organizations are struggling to identify all viable data sources and leverage it to answer such fundamental questions as “what is the number of current customers?” As senior management, business users and IT grow increasingly frustrated at the inability to fully leverage corporate data, many are looking for a better solution.

Opportunity

Data discovery and profiling is a relatively new data management technique that uses tools to identify all data sources throughout the organization and unearth the meaning of the information itself. As with any new paradigm, today’s budget holders and decision makers must first be educated on data discovery and profiling and then shown the value it can bring to the organization through a clear return on investment (ROI) analysis.

Benefits

The growing need to access and leverage data that is hiding within many disparate systems and applications and across the highly complex, disjointed IT infrastructures common in today’s enterprises, has made data discovery and profiling a necessity. Data discovery and profiling tools help complement many data management technologies in use today and help improve efficiencies and effectiveness, while mitigating risk.

SECTION 1

The Growing Data Crisis

During the past forty years, enterprises have been gradually implementing complex infrastructures for information processing. Originally, the focus was on automating business processes. Today, however, it has shifted to data. While data was once viewed as merely the by-product of automation, it is now recognized as a valuable resource in its own right. It can be used to do things that were unthinkable a few decades ago, such as measure the efficiency of business processes, yield precious marketing insight and show the results of business decisions.

At least that is the promise. In reality, the ad-hoc, organic growth of enterprise IT architectures and lack of meaningful data management have created production data landscapes of almost unimaginable scale — that are almost opaque. In fact, the following general questions cannot always be reliably answered:

- What data does the enterprise manage?
- What does it mean?
- Where it is located?

Worse, many organizations are struggling to ascertain key business data, such as the number of current customers. And, customer details are inevitably distributed over many databases, but few — if any — know exactly which ones. Information may even be hidden in transaction records, since companies put more emphasis on having a new customer complete a transaction than properly capturing all of his or her details in a separate customer record. Even where a specific customer database can be identified, the data content is often plagued with such problems as duplicate records.

The full consequences of this crisis are not yet completely understood, but they are becoming clearer. Business users are increasingly frustrated that the information they need cannot be provided. And IT departments are struggling to integrate enterprise information assets in order to answer business needs — but find that such projects consume a very high level of resources, delivery is delayed and outcomes are often plagued by data quality issues.

One proven way to mitigate these problems is through the use of data discovery and profiling. However, it is a relatively new data management technique. As a result, senior management may be unfamiliar with it and unable to readily understand its value. This paper provides the information needed to clearly demonstrate the return on investment (ROI) for discovery and profiling.

What are Data Discovery and Profiling?

The first steps to explaining the ROI of data discovery and profiling to senior management involve describing what it is, why it is needed now and why it was not needed in the past. Because discovery and profiling are new and may not be intuitive, it's important to ensure that this message is always consistent. This is particularly true for senior executives outside of IT, who inevitably focus on business value. For them, discovery and profiling can be explained as follows:

- Data Discovery. In most organizations, the size and complexity of the production data is so large that no one knows where critical business information is located. Yet, the business demands this information be provided in consolidated and aggregated forms. With thousands of database tables, and hundreds of thousands of database columns in an average mid-sized enterprise, it is virtually impossible to do so manually. Human analysts simply cannot look at data on this scale to find customers, products, sales and so on, to figure out how to bring the data into integrated environments where it can be accessed to satisfy business information needs.

However, a data discovery tool can overcome this problem of scale because it can scan large environments and identify data in a fraction of the time required by a team of human analysts. This tool offers a much greater chance of finding the best sources of critical business data.

- Data Profiling. Even if human analysts were to find a particular source of critical business data in the landscape, they still have to understand it before they can use it for business intelligence purposes. This is remarkably difficult because most data is locked in vendor-supplied packages with proprietary database designs — or home-grown legacy applications whose databases are equally mysterious. Unless the meaning of the data and its relationships in these environments are properly understood, it is impossible to integrate it to produce the kind of actionable reporting the business demands.

  Data profiling tools look at the content of the data itself to find relationships and patterns — making it much easier to unearth the meaning behind the information. Profiling tools also work on a far greater scale than a team of human analysts trying to browse the data to manually infer relationships.

  Additionally, data profiling tools can quickly spot inconsistencies in the source data. Before data profiling, these inconsistencies often appeared unnoticed in data warehouses and marts. The data problems were eventually found when business users discovered anomalies in their reports. All too often, such episodes reduced trust in the reporting environments, but left business users without a reliable alternative.

Source Data Analysis

Defining discovery and profiling is just the first step. Senior management will want to know how these techniques and the tools that implement them can either save — or make — the enterprise money. This is the core of any demonstration of ROI, and one area in which it can easily be done is source data analysis (SDA).

SDA was rarely needed in the early days of IT, because applications were being developed to automate existing business processes, rather than reprocess data that had already been populated in databases. For this reason, it never appeared in such programming-centric methodologies of IT as waterfall or agile. At the time, "analysis" in these activities consisted of looking at documents, or talking to users, to gather requirements and understand processes.

But SDA is completely different. It aims at providing a sufficient understanding of the content of existing data sources that are to be integrated in order to satisfy business requirements. What's more, it simply cannot be done by a few human analysts in the same way that traditional analysis can.

ESTIMATING SOURCE DATA ANALYSIS

It is fairly easy to determine the time required for SDA. Table 1 provides estimates of the size of small, medium and large enterprise database environments. The estimates in the table are taken from the author's experience in working on projects involving limited manual SDA. They can be substituted by estimates for a given enterprise, but because of the normal large scale of data landscapes, the conclusions will be more or less the same.

Table 1 shows that, except for the very smallest enterprises, the work required to perform manual SDA on any integration project is very significant. For example, an integration project that requires 2,000 columns to be analyzed would require 5,000 work hours. At a cost of $100 per analyst work hour this would total $500,000. It would require a team of three analysts working for about 1 year to complete the task.

TABLE 1: VOLUME ESTIMATES FOR MANUAL DATA DISCOVERY AND PROFILING

Volume Metric	Small	Medium	Large
Number of Production Databases	10	100	1,000
Average Tables per Database	20	50	75
Average Columns per Database	8	15	30
Average Rows per table	10,000	20,000	100,000
Total Tables	200	5,000	75,000
Total Columns	1,600	75,000	2,250,000
Total Rows	2,000,000	100,000,000	7,500,000,000
Average Analyst Profiling Effort (Work Hours to Manually Discover and Profile a Single Column)	1.0	2.5	4.0
Work-hours to Manually Discover and Profile All Columns	1,600	187,500	9,000,000
Cost per analyst work-hour	$100	$100	$100
Total cost for Manual Discovery and Profiling (potential savings for automation)	$160,000	$18,750,000	$900,000,000
Average Number of Columns per Integration Project	150	2,000	5,000
Work-hours to Manually Discover and Profile All Columns	150	5,000	20,000
Cost per analyst work-hour	$100	$100	$100
Total cost for Manual Discovery and Profiling for Integration (potential savings for automation)	$15,000	$500,000	$2,000,000

Clearly, this is unacceptable and does not happen on actual projects. Unfortunately, because SDA is so new, it is rarely budgeted for explicitly and facts about costs are difficult to glean. For example, during an integration project for a major financial company that had spent 30% of its budget on manual SDA, this level of effort was unforeseen and caused a major project overrun. It only occurred after serious anomalies were found in the data and the SDA effort was stopped after six months because it was deemed to be holding up the project (which was ultimately unsuccessful). Underscoring the lack of planning and budgeting for SDA, a senior manager in a leading discovery tool company asserted that in his company approximately 30% of all resources on data integration projects are devoted to SDA. He too felt that this was not nearly enough to do the job well if it had to be done manually.

If the task of SDA could be automated in any way, the investment in manual labor could be reduced. Data discovery and profiling tools provide this capability. If the above estimate is correct for an average data integration project in an enterprise with a medium-sized data landscape, then one project would easily pay for the acquisition of such a tool.

It is also worth noting that today most analysis is actually performed on an infrastructure that IT has already implemented. In fact, programmers spend a lot of their time reverse-engineering existing code and databases — estimated to be more than 50% according to a 1996 report by CASE Associates[1]. Management is likely aware that analysts must spend time trying to understand existing applications and databases — which should be taken into consideration during any discussion of ROI for data profiling.

The Mirage of Legacy Data Quality

One issue that sometimes makes it difficult to propose data profiling to CIOs and other senior management is a feeling that data quality may be better than it actually is. All CIOs are aware of the status of their production applications and typically receive daily reports on anything that impacts the service levels of these applications. Such reports usually include few instances of data quality problems that cause service outages. As a result, data profiling is assumed to be unnecessary; the fact that legacy applications are running smoothly is interpreted as confirmation that there must be relatively few data quality problems.

This is the "mirage of legacy data quality." It is caused by the fact that, over the years, program logic and production data content have become bound together in legacy applications. The program code filters, transforms, enhances and relates application data to fit every output that the application produces. The data is not used "as is", but only "as rendered" by the application. Therefore, just because the legacy application works does not mean that data quality problems are absent, only that they are being handled. When data is moved out of this environment, data quality problems are no longer addressed, and they become apparent.

Of course, many CIOs realize this and are only too aware of the need to profile legacy data before it is used in different situations. However, the lack of visibility into legacy data problems makes it impossible to gauge just how bad the problem really is, and may lead to the conclusion that profiling is not really necessary. This can result in a headlong rush into data integration projects in which legacy data is pulled into data warehouses or MDM hubs — or used in messaging in near real time among operational systems. It is much later that the problems inherent in legacy data are properly understood. But by that time, an unacceptable level of resources has been expended.

When the mirage of legacy data quality is a problem, the data analyst must find a way to convincingly demonstrate a need for data profiling before any integration project becomes too advanced. This is a difficult task, because in such situations the analyst is unlikely to have an enterprise-class data profiling tool available — and manual profiling has proven to be too time consuming. However, the analyst does not need to do a full-scale profiling exercise. A sample of data is all that is needed to identify data quality issues. Next, a scaling factor can be used to estimate the full scope of data quality problems and provide an incontestable case for full profiling of the legacy data. Today, there are low-cost data profiling tools available, so the analyst need not necessarily rely on a large, enterprise-class solution. And if the low-cost tools are not available, an analyst with reasonable knowledge of the data may still be able to construct a sample set using SQL queries.

Table 2 shows how results from such a sampling approach might be presented to senior management. In order to estimate the gross number of quality problems likely to occur, the number of errors found is multiplied by the ratio of the data objects sampled to the total.

TABLE 2: PRESENTATION OF RESULTS FROM A SAMPLING APPROACH TO ESTIMATING DATA QUALITY PROBLEMS

Volumes	Sampled	Total
Number of Tables	10	100
Number of Columns	50	2000
Data Quality Test Results	**Number Found in Sample**	**Estimated to Occur**
Overloaded Columns	5	200
Code Columns without Parent Code Tables	7	280
Columns with Orphaned Child Values	3	120

1. David Sharon, "Meeting the Challenge of Software Maintenance," IEEE Software, vol. 13, no. 1, pp. 122-126, January, 1996.

SECTION 2

Enhancing Data Models

An important part of data profiling is building data models that accurately represent the underlying databases. However, in enterprises where data modeling has not been performed in the past, this may not be seen as helpful, making it difficult to get senior management buy-in.

Yet, even in enterprises that have done data modeling, there may still be a need to convince management of the need for profiling. Databases are very expensive pieces of infrastructure. Ideally, logical data models would represent them perfectly and physical data models would specify them exactly. However, they rarely do. Models may be entirely missing, the logical may not match the physical or the physical may not match the database. Profiling can help tremendously in these areas by producing completely reliable data models that are reverse-engineered from production databases.

Still, this may not resonate with management in some enterprises. They may continue to question the need for data modeling, since it was not needed before. However in the past, over a period of decades, each silo was built individually and was self-contained. Thus, it was possible, albeit poor practice, for individual application teams to understand the representation and specification of these databases as they were being built — without retaining or passing on this knowledge.

Such knowledge is usually lost if not contained in data models. What's more, through the years, applications get enhanced, modified, fixed and patched a little bit at a time, resulting in even more drift away from the original design. While all this is happening, personnel come and go, and knowledge that is maintained "tribally" is imperfectly transmitted — if at all. The result is that applications inevitably evolve into "black boxes" — and nobody is sure what is going on inside. Even a small change is risky and involves a disproportionate amount of reverse-engineering. Changes to database structures are particularly feared.

This argument is underscored by the fact that the older an application, the less it is understood. In fact, contrary to most experience, the best understood applications are always the newest. Again, it would not matter if "black box" databases weren't needed for anything beyond themselves, or if there were just one or two of them. But data integration casts a wide net, and data is ultimately needed from all transaction applications. Furthermore, the cumulative set of databases that has been built up over decades needs to be considered. This is completely different from saying that a data modeling tool was not needed when teams built or implemented one or two databases per year.

Trying to understand the structure of a "black box" database, either to enhance it or to use it as a source of data, means developing a data model of the database in question. The least efficient and least effective way of doing this is to rely on human analysts to somehow understand, remember and communicate the structure of the database. No enterprise will have the analyst resources to do this for more than one or two databases at a time. If data models are needed to understand databases, and there is a scarcity of analyst resources, then there is no alternative but to use data profiling tools to automatically generate sharable data models of the databases.

The Role of Technology

In helping senior management understand the ROI for data management, it may also be necessary to overcome the perception that existing technology investments can solve all data-related problems. The fact is that technology has created a great deal of the "data mess" in which most enterprises find themselves. For instance, there are many classes of tools that can quickly and efficiently move data around — and they work very well and have been used extensively. As a result, however, integration environments have been plagued by data quality problems that went unrecognized due to the lack of data profiling.

It can be difficult to make the case that another piece of technology — a profiling tool — is required to make the investment in other tools pay off. The vendors of data movement tools, data warehouse products, MDM hubs, BI tools and the like will not necessarily raise the need for a data profiling tool. In fact, they may even downplay it in order to make a sale. What's more, they may claim that their tools have some features that assure data quality, such as guaranteed message delivery in a middleware product.

The role of such claims needs to be carefully understood when making a case for the ROI of data profiling. Senior management may not mention the claims they have heard from vendors of other products, so it is important to place data profiling tools in a product matrix that shows how they will complement tools currently in use. More to the point, without data profiling, the investment in these other tools may well be wasted. If possible, the engagement with senior management should attempt to flush out any arguments or assumptions about how existing tools can meet data profiling use cases.

SECTION 3

Preparing a Business Case

Gathering the information about ROI for data profiling is one thing, but successfully presenting it as part of a well thought out business case is another. The following three main areas should be covered in any business case that attempts to demonstrate ROI:

- **Efficiency.** Efficiency is the reduction of existing costs, or the act of doing more with the current set of resources. This part of the business case is the easiest to present in quantitative terms, and management expects to see quantification in at least one major area for ROI. Developing the kinds of matrices shown in Table 1 can be helpful here.
- **Effectiveness.** Effectiveness is harder to quantify than efficiency. Basically, it means enabling the enterprise to meet its stated goals. Today, IT is seen as a major problem in adjusting to new business environments, because the architecture that IT has developed over the decades is brittle and inflexible. In this part of the plan, it is necessary to show how data profiling can contribute to agility.
- **Risk Mitigation.** This part of the business case is frequently overlooked. Some quantification may be possible if information is available about losses caused by data-related problems that would have been picked up by profiling. A more direct, but possibly controversial, approach is to look at IT projects that have failed to meet expectations because of lack of data profiling. Obviously, this should be handled tactfully.

Presentation is important because data is full of abstract concepts that are difficult to grasp. The business case should be as easy for the intended audience to read as possible. There is no need to dwell on technical details. Senior management has many demands on its time and attention, so presenting the case and ROI as clearly as possible will be much appreciated.

SECTION 4

Conclusions

While the need for Data Discovery and Profiling tools has become a necessity, the ability to successfully prove its value to senior management remains a challenge. The ROI techniques presented in this paper can be used to help executives understand the need for this new paradigm, appreciate its place within the existing arsenal of IT tools and look forward to reaping the benefits of it.

For more information about how CA ERwin Data Profiler can help your business increase the quality of your data by performing cross system analysis and profiling of both database and legacy systems, visit ca.com/modeling or for local information, go to ca.com/contact/rmdm.

SECTION 5

References

1. David Sharon, "Meeting the Challenge of Software Maintenance," IEEE Software, vol. 13, no. 1, pp. 122-126, January, 1996.

SECTION 6

About the Author

MALCOLM CHISHOLM PH.D.
INDEPENDENT CONSULTANT

Malcolm Chisholm Ph.D. is an independent consultant with over 25 years of experience in IT, and has worked in finance, manufacturing, government, defense, telecommunications, pharmaceuticals and insurance. He is author of the books *Managing Reference Data in Enterprise Databases* (Morgan Kaufmann, 2000) and *How to Build a Business Rules Engine* (Morgan Kaufmann, 2003). Malcolm is a thought leader in the fields of master data management and data governance, and writes and speaks often on these topics. He maintains the websites www.refdataportal.com, www.bizrulesengine.com, and www.dgovernance.com, and can be contacted at DataGovernance@gmail.com.