› White Paper

Data Evolution

Why a Comprehensive Data Management Platform Supersedes the Data Integration Toolbox

sas
THE POWER TO KNOW.

Contents

Content for *Data Evolution: Why a Comprehensive Data Management Platform Supersedes the Data Integration Toolbox* was provided by David Barkaway and Mark Torr, Director of the SAS Technology Practice, Heidelberg, Germany.

Executive Summary

In this era of big data, today's organizations have incredible amounts of information to be managed, and in many cases it is quickly spiraling out of control. An increased number of channels and routes to market; the rise in machine sensor data known as the "Internet of Things"; globalization of businesses; expansion from traditional data repositories such as databases to unstructured data such as emails, blogs and networking sites; and a sharper focus on regulatory compliance have all contributed to the exponential increase in the amount of data that is captured, processed, analyzed and archived. It is a "perfect data storm."

To address the emerging issues around managing, governing and utilizing data, organizations have been acquiring quite a toolbox of data integration tools and technologies over the past five to 10 years. One of the core drivers for these data integration tools and technologies, and the subsequently assembled data integration toolbox, has been the ever-evolving world of the data warehouse.

Today's data integration toolbox to support data warehouses usually contains the following components:

- **ETL tools and technologies** that support the core processes of extraction, transformation and loading (ETL) typically are associated with data warehousing. ETL tools extract data from a chosen source or sources, transform it into new formats according to business rules and then load it into target data structures. ETL tools and technologies provide comprehensive data connectivity, powerful transformations, batch scheduling or real-time capabilities, and the ability to execute data processing at high volumes in the server or push down processing to the database. Most recently, an extract, load and transform (ELT) process has begun to emerge that allows transformation to occur in a database as opposed to a specialized engine, thereby avoiding data duplication and using the extra capacity on the database platform.

- **Data quality tools and technologies** support the process of cleansing the data so that it is "fit for purpose." Data quality is improved by parsing, standardizing, integrating/matching, de-duplicating and enhancing the data. The tools and technology will often use complex algorithms to match the data, and enhancements may include geocoding and address verification.

- **Data profiling tools and technologies** support a process that examines the data and metadata in a data repository to collect statistics and information about that data. The statistics identify whether the existing data can be used for other purposes. If not, the scope of work to rectify the situation is determined. Profiling technologies are intrinsically linked to data quality, and evolved out of the need to identify source-system data issues prior to embarking on data integration initiatives.

- **Data federation tools and technologies** are used to dynamically aggregate data from multiple sources into a single virtual view of the data and expose that data through SQL queries or a service. These technologies are popular for low-latency queries to data repositories and came into their element when organizations needed the ability to link historical data warehouses with a real-time view of the operational systems.

- **Data exploration tools and technologies** support the process of identifying where data resides within the organization, followed by the categorization and documentation of that data. The tools that automate this process provide for more accurate and focused data profiling.

- **Metadata management tools and technologies** provide the capabilities to store, search, report on, link to, categorize and govern information about data. As the data and information management landscape becomes more complex, the need to understand where data resides, and how or when it is moved or transformed, is critical to the understanding and management of an enterprise information environment.

- **Master data management (MDM) tools and technologies** support the process of defining and maintaining consistent definitions of a business' reference or master data; storing and sharing that data across IT systems and groups throughout the enterprise; and ensuring the master data files remain in a standardized, controlled format as they are accessed and updated.

To build their required data integration toolbox, organizations have had to acquire the tools and technologies from different sources. This means that today's organizations generally have a multitude of tools running on a diverse set of underpinning infrastructure that is not shared. And the tools are not well integrated – even if they are under the same brand following a wave of acquisitions. As projects increasingly require the deployment of many tools and technologies from the data integration toolbox, a significant proportion of the overall time on the project is spent resolving technology integration issues that could instead be applied to data integration requirements to support the business.

Sounding the Death Knell for Data Integration Toolboxes

Data-related projects continue to grow in importance, and organizations are shifting from a focus on integrating data mainly for decision support to how they can manage all of their organizational data. This area is known as data management because it refers to the management and governance of all data in the organization. As the area of data management has emerged, new staff has not been added at the same rate as new projects.

Organizations need tools and technologies that can address new requirements and enable employees to focus on the job at hand instead of spending their time constantly integrating disparate technologies in the toolbox. The need for a single, integrated data management platform that can address all aspects of data integration, data quality and master data management could be sounding the death knell for the data integration toolbox. These key areas will be underpinned by adapters and a federation capability, and will share technical and business metadata that aids in collaboration. Ultimately, a single user interface should surface all the capabilities of this platform rather than a disparate set of user interfaces.

Having a single platform and single interface provides the additional benefit of being able to apply a consistent methodology that spans the different aspects of the data management life cycle, an approach that proves difficult or impossible with a toolbox approach. With a single platform for all data management initiatives and a consistent interface, a methodology can be incorporated into the platform and used to guide the business analyst or developer through the project phases and the tasks that need to be completed. This approach not only reduces the learning curve for a user, but also reduces risks and accelerates project delivery.

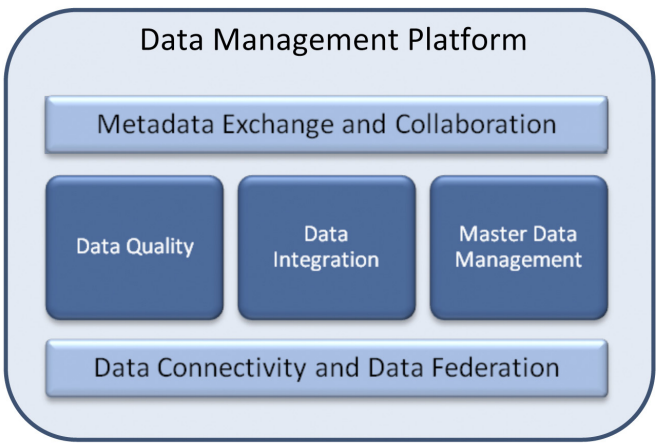

Figure 1: A comprehensive data management platform includes data quality, data integration and master data management – underpinned by data federation and connectivity technologies and overlaid with shared metadata.

Today, SAS offers market-leading data integration technologies and a market-leading portfolio of data quality technologies through SAS® Data Management, including its master data management solution. For decades, SAS has been providing an integrated data integration framework more advanced than any other software vendor. We are now moving to better support the data management needs of organizations by accelerating the delivery of a single data management platform to address any type of data integration, data quality or master data management initiative with a shared architecture and, in the future, a single role-based user interface.

If your organization is not yet looking at a data management platform, you should do so soon. The benefits of better governance, reduced risk, higher quality, more accessible data and shorter project cycles could be of enormous importance to your organization over the next five to 10 years.

The Evolution of the Data Integration Toolbox

The following sections of this paper briefly describe the evolution of data integration and the organizational requirements that drove the need for the tools and technologies in the toolbox.

The Emergence of ETL Tools and Technologies

Historically, as organizations developed more than one data repository, it became necessary to integrate heterogeneous data sources. The rapid adoption of databases in the 1970s naturally led to the need to share or merge existing data repositories. Initially, scripts were written to transfer data from one system to another. Soon, there were multiple data repositories with numerous scripts being executed to transfer data between repositories. We then saw the first generation of tools that automated the generation of these scripts. (These tools automated the generation of COBOL or C programs that extracted, split, merged and loaded data.) The majority of these tools initially generated code to execute on the mainframe, and these processes would be run as a batch process on a regular basis. This approach, now known as ETL, has since migrated to UNIX platforms and PCs as they have become available.

Figure 2 illustrates a single ETL process. Sales on a daily basis are captured in the Sales Data repository. The data is then split and sent to other systems such as Accounts, Human Resources, Customer Service, Warehousing and Shipping so the item can be delivered. All data from these systems would then be merged and loaded back to the Sales repository. This example shows only one batch interaction, although hundreds if not thousands of these processes are executed on a daily and monthly basis.

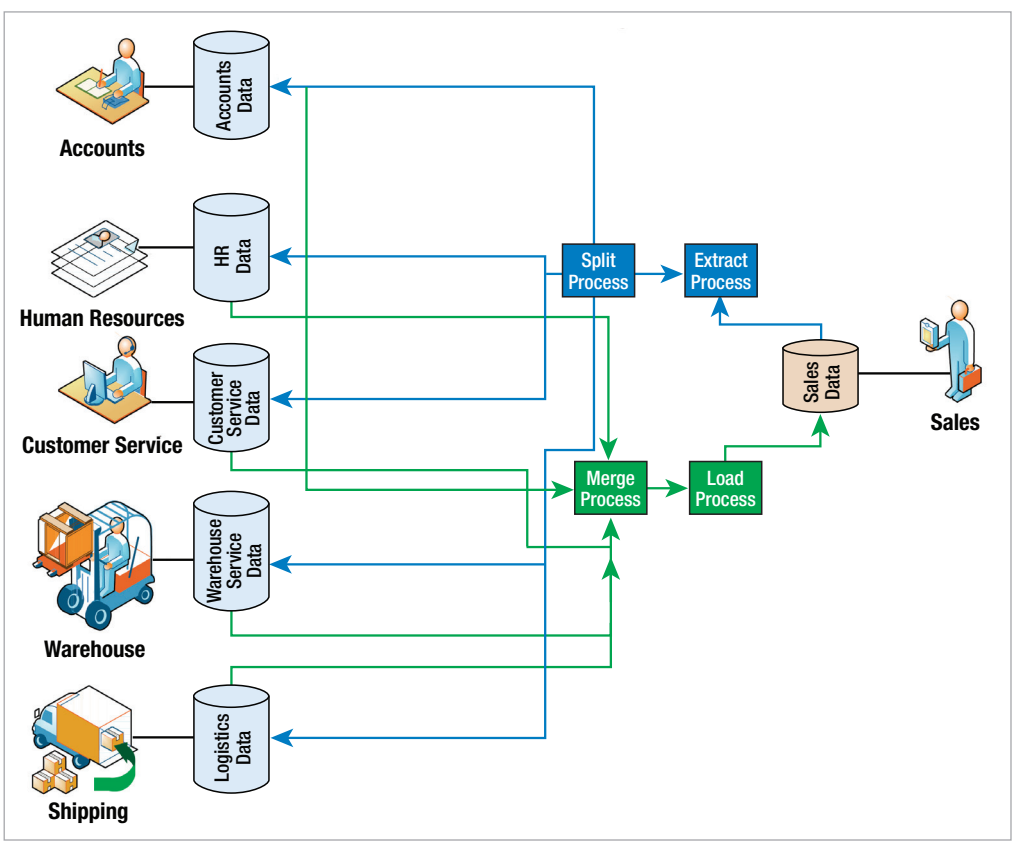

Figure 2: Shown here, a single ETL process. In most organizations, hundreds, if not thousands, of single ETL processes are executed on a daily and monthly basis.

Growing Data Warehouses Drive the Adoption of ETL Tools and Technologies

Although integration between operational databases was a key driver in the emergence of the ETL tools market, the most important driver for the general adoption was the evolution of the data warehouse. Data warehouses became a distinct type of computer database during the late 1980s and early 1990s. They were developed to meet the growing demand by management for information and analysis that could not be met by a single operational system. Operational systems were unable to meet the need for several reasons, including the fact that the processing loads for reporting reduced operational response time for users. Management reports also generally required information from multiple data repositories.

As a result, separate computer databases were specifically designed and built to support management information and analysis purposes. These data warehouses were able to bring in data from a range of different sources – such as mainframe computers, minicomputers, personal computers and spreadsheets – and then integrate that information into a single location.

The Evolving Data Warehouse Propels the Next Generation of ETL Tools

Data warehouses quickly evolved from ancillary reporting repositories to core systems that are active 24/7 – with varying data-latency links to all operational systems. At the same time, there was a parallel evolution in what tools were needed in the data integration toolbox, expanding beyond the basic ETL (extract, transform and load) tools where this all started.

The effect of data warehousing on the ETL tools and technologies market was quite dramatic. ETL tools began providing specific transformations to support the requirements of the data warehouse and became the hub in front of the data warehouse that extracted source data, transformed it and loaded the data into the data warehouse. New transformations were continuously added to ETL tools to support the emerging requirements of the data warehouse. Slowly-changing-dimensions transformations and data-pivoting or transposition transformations were added, to name a few. Native adapters were also built to provide access to the multitude of data source types, and the data integration tool interacted with these adapters to extract and load the data in the most efficient manner.

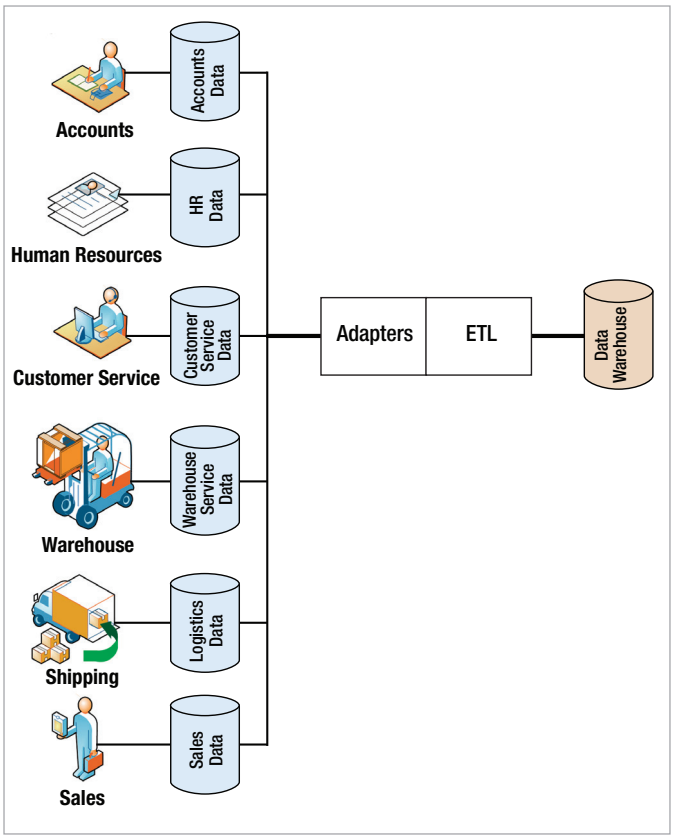

Figure 3: ETL tools and adapters became the hub in front of data warehouses.

Data Quality's Arrival Heralds the Beginning of the Data Integration Toolbox

In addition to the need for propagating data between repositories, there was an increasing demand to ensure that the data being propagated was correct. Joseph Muran, a 20th century evangelist for quality management, probably described data quality best. He noted that data sources are of high quality if they are fit for their intended uses in operations, decision making and planning.

As organizations have evolved to be more informational, it has been recognized that true value exists in the data held – not just the physical assets of the machinery or infrastructure. Improving the critical raw materials needed for success with information required a technology to identify issues in the data, correct these issues and enhance the data where needed.

Incorrect data has a negative effect, because decisions based on the data are likely to be imperfect (at best) and just plain wrong (at worst). With data warehouses as the first point where information was merged from multiple sources and then used for business intelligence and analytics to improve decision making, it was the first place where organizations began to identify the poor quality of their data. Although technically the teams building the data warehouse were not the cause of the problem, they became the focus of the problem and were thereby tasked to fix it. Data quality tools and technologies were soon added to ETL capabilities to solve the data quality issues of the data warehouse, and the data integration toolbox was born. Initially, the integration between data quality tools and ETL tools was through a standard file format or database structure. Over time, more sophisticated methods have been introduced through the use of web services.

It was estimated as far back as 2004 that poor data quality costs the typical company at least 10 percent of its revenue.[1] Today that estimate is thought to be much higher. There has been a considerable increase in organizations adopting data quality technology in conjunction with ETL. In our experience, approximately 50 percent of sales that include ETL tools also included data quality tools and technologies. Some of these investments can be attributed to the desire for an improved bottom line; but others are driven by stricter regulatory compliance requirements. Many organizations still conclude that poor data quality is a fact of life, and will only change their outlook on data quality when a small amount of bad data leads to a disaster of epic proportions.

> Joseph Muran noted that data sources are of high quality if they are fit for their intended uses in operations, decision making and planning.

Knowing the Problems Before It's Too Late with Data Profiling

A large proportion of initial data warehousing projects failed in the 1990s and early 21st century. The majority of these failures can be attributed to poor data quality – data that could not be merged or was not fit for purpose once loaded. Data migration projects failed for the same reason. The scope of the problem and the fitness of the data for purpose were not identified until **after** the initial loading of data. In September 2007, Bloor Research published a paper, *Data Migration in the Global 2000*, which identified that:

- Only 16 percent of data migration projects were delivered on time and on budget.
- More than half of respondents that overran their budgets and two-thirds of respondents that overran time blamed a failure to properly scope the project in advance as a primary cause.

Accurately and proactively scoping the ETL and data quality tasks, rather than uncovering issues after building and deploying the solution, required technology that could examine the source systems for their current state to enable better planning. Specific tools and technologies to profile data emerged in the mid-to-late '90s to support this need. The tools were soon used in the early stages of ETL and data quality projects. Data profiling tools and technologies enable three types of analysis:

- **Structure discovery**. Does your data match the corresponding metadata? Do the patterns of the data match expected patterns? Does the data adhere to appropriate uniqueness and null value rules?
- **Content discovery**. Is the data complete? Is it accurate? Does it contain information that is easily understood and not ambiguous?
- **Relationship discovery**. Does the data adhere to specified and required key relationships across columns and tables? Are there inferred relationships across columns, tables or databases? Is there redundant data?

About half of the customers who purchase an ETL tool also eventually purchase data profiling tools to support their projects.

[1] Thomas C. Redman, "Data: An Unfolding Quality Disaster," *DM Review*, August 2004.

To Copy or View: Data Federation and Data Virtualization Make Their Entrance

While data federation is fundamentally different from ETL, it is usually perceived as complementary. With ETL, you physically copy or move the data. With data federation, you query the data but don't physically move it. Data federation technology has been available as long as ETL tools. You could query multiple databases with gateways in the '80s. But, it wasn't until the late '90s that we saw the emergence of enterprise information integration (EII) vendors providing data federation technology that matched and competed with the ETL vendors.

Although data federation had been around for two decades, there hadn't been a serious demand for this solution until the late '90s. Suddenly, organizations were faced with two issues that were better addressed with data federation tools and technologies than with ETL. First, organizations in the late '90s were under pressure to have a "web presence." Data federation provided a simple method to query the existing systems in real time and provide a single environment to interface with the front-end system. The initial versions of this technology were limited to a unidirectional approach. For example, the federation system could respond to questions such as "What is the stock availability of the items in my basket?" or "What is the balance of my accounts?" Later releases provided a read/write capability and a web service interface to complement the initial SQL-type queries, thus supplying a data services abstraction layer over the physical data.

The second key driver for this technology was the sudden change in demand from data warehouse users. With new channels to market, such as the Internet, the volumes of data increased. Organizations were also switching from eight hours a day, five days a week to 24/7, so the available windows for batch processing were shrinking as the volumes were increasing. If that wasn't enough of an issue for IT, organizations now wanted data refreshed more frequently. Rather than an overnight refresh, they wanted it hourly if not more frequently. This was not a simple solution for traditional ETL systems. Running batch processes every few seconds (combined with the added load of the new online customers) would bring operational systems to a standstill. Data federation provided an elegant solution to this conundrum. If you ran your ETL batch processes overnight to update the warehouse, you could place the data federation technology over your data warehouse and operational systems. The business intelligence technology queries the data federation technology – and this accesses the data warehouse for historical data – "firing" low-impact queries at the operational system for a real-time view of the business.

In the early 2010s, there was an increase in use of data federation technology to create logical data warehouses that combine traditional data warehouses with emerging big data technologies such as Hadoop. Hadoop is an open-source technology that takes advantage of low-cost commodity blades to store large amounts of data and process them in parallel across a grid of servers.

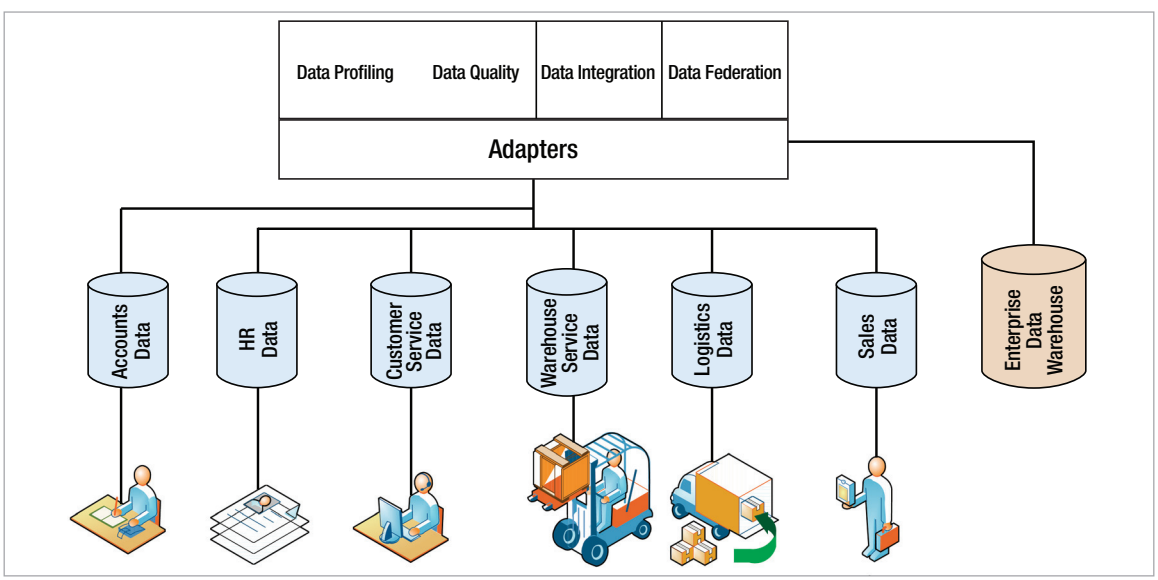

Figure 4: Operational systems are profiled to identify data quality issues. The data is loaded in batch to a data warehouse and cleansed in the process. Data federation technology provides an abstracted view of the data warehouse and operational systems for a variety of consumers, such as business intelligence applications, dashboards, MDM and mashups.

The badge of EII for data federation has now been replaced by data virtualization, capitalizing on the popularity of hardware virtualization in corporate data centers. This new badge reinforces the idea that data federation tools abstract physical data repositories behind a common interface, either SQL or business/data services.

Over the last 10 years, we've seen the data federation tools and technologies become an essential item in the data integration toolbox. They provide a complementary technology to traditional ETL. Organizations use the technology in a variety of situations that require unified access to heterogeneous data sources for data warehousing, reporting, dashboards, mashups, and more recently, master data management (MDM). As organizations realized the importance of data federation, independent vendors of this technology were quickly acquired by the business intelligence, ETL and SOA software houses to add to the data integration toolboxes they were selling.

Data, Data Everywhere – What Does It All Mean and How Is It All Connected? Metadata Management Bursts onto the Scene

Data integration teams now faced another challenge. They had a useful toolbox of data integration technologies to deliver enterprise integration solutions, but the working environment was getting more complex. There were more data sources coming online, and the data warehouse was constantly changing to meet the demands of the business. It became a struggle to keep track of all of the integration processes that were running, and to understand what impact a data repository change had across all of the systems. All of the separate tools that software vendors had acquired and placed into their toolbox with minimal integration had separate metadata repositories. To identify the effect of change in a source or target environment meant trawling through all of the processes in each of the tools.

This complexity affected the business user as well, who might look at a report and ask the following questions:

- "Which system did the data come from?"
- "How was the metric calculated?"
- "How fresh was the data for the report?"
- "Did the data for the report pass all of the data quality, business rules and audit processes so it could be trusted?"

To provide the business user with answers to the above questions required an unbroken chain of metadata to be propagated from the data integration layer to the presentation layer.

Software vendors faced the problem of multiple metadata repositories across all of the tools. To solve this problem, they created a metadata layer that spanned the data integration tools and provided the metadata visualization and propagation. This concept evolved further with search metapedias and business glossary-type capabilities so the user can link, search, govern and categorize metadata, as well as report on it.

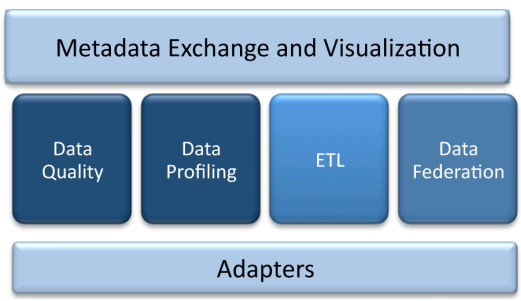

Figure 5: To successfully track data from its source to target, a metadata layer that spans across all data integration tools is needed.

"This move towards being a data quality/integration/master data stack provider is an increasingly common story in this space. However, integration between the different elements of this stack is a big issue. Some major vendors are so far away from a coherent story about integration that they do not really merit the description of a stack supplier, even if they have all of the relevant components. So the fact that SAS now has a genuinely integrated suite should give it a significant advantage over its competitors."

Bloor Research

Data Exploration: What Do I Have and Where Is It?

When an organization has multiple data sources, it is possible there is a limited understanding of what data resides within the data repositories – not an unusual position if you have acquired a company or are managing an outsourced system. Data exploration tools and technologies emerged to support the automation of data identification before data profiling to provide an understanding of what data resided within each repository. This helps to quickly and easily accumulate detailed knowledge of thousands of data sources across many systems and platforms. It provides the ability to analyze metadata and sample the data from existing data sources and furnishes the foundation for more precise data profiling and data quality efforts. In addition, the results of the exploration can be the foundation for a valuable business data glossary.

Some exploration technologies and tools focus on the identification and categorization of data, where it is located and how it is linked across systems. The methods used for data exploration range from intelligently capturing relationships and determining applied transformations and business rules, to using vocabularies and matching technology to identify and categorize data and build up a glossary. Either mechanism is beneficial. The idea is to target profiling and accelerate the build phase. Data exploration tools therefore became the late addition to the now rapidly growing data integration toolbox.

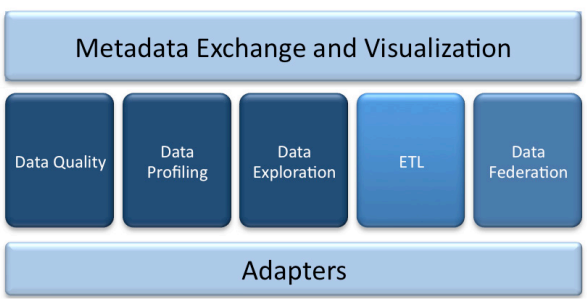

Figure 6: Data exploration tools emerged to automate data identification before data profiling to provide an understanding of what data resides within each repository

The Latest Toolbox Addition: Master Data Management

The final set of tools and technologies that have been added most recently to the data integration toolbox are related to the area of master data management (MDM). Wikipedia refers to MDM as comprising "a set of processes and tools that consistently defines and manages the non-transactional data entities of an organization (which may include reference data). MDM provides processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization to ensure consistency and control in the ongoing maintenance and application use of this information."

At a basic level, master data management ensures that an organization does not use potentially inconsistent versions of the same master data in different parts of its operations, which can easily occur in large organizations.

Master data management, being a requirement for data integration and synchronization and supported with data federation and data quality, is a logical fit to the data integration toolbox to deliver consistent and trusted data across the enterprise.

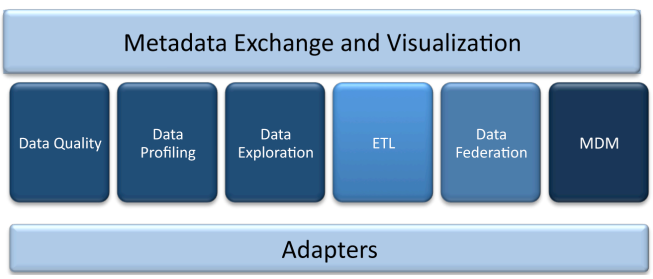

Figure 7: Master data management (MDM) technologies and solutions emerged as an additional tool to deliver consistent and trusted data.

Are Other Tools and Technologies Needed in the Toolbox?

The tools and technologies mentioned above could be seen as the minimum set needed for today's enterprise data integration requirements, whether this is to run projects to build data warehouses, carry out data migrations and improve data quality or run comprehensive master data management programs.

However, some organizations have incorporated other technologies into their data integration toolboxes, such as change data capture. This paper has focused mainly on the tools that support data warehousing initiatives. It has not included categories of tools supporting activities such as enterprise application integration (EAI) because they don't really support the creation and maintenance of data warehouses or data marts for business intelligence and analytics. Neither has the paper delved into complex event processing or event-stream processing because these areas, although supportive of business intelligence, analytics and data synchronization for MDM, are not perceived by organizations to be part of the default data integration toolbox at this point. They could and probably will emerge as core components in the not-too-distant future.

Problems with Data Integration Toolboxes

Having built a data integration toolbox, the problem most organizations face is that they are left with a disparate set of technologies that at best are loosely integrated, even if most appear to come from one vendor. The fact is that almost no major vendor has organically built their set of technologies – instead favoring a bolt-on approach to ensure they have all the pieces needed to look like a complete toolbox, with most technologies and tools being added through aggressive acquisition strategies. Different software vendors have had different levels of success with this approach, but one thing is clear – using diverse, loosely integrated tools and technologies affects the business in numerous ways.

- **Metadata**. If each tool has its own metadata repository as you switch between phases of a project, you need to propagate the metadata from one tool to the next to gain the benefit of the previous phase. If the tools are not integrated properly, you lose that captured metadata. If metadata isn't shared, you also have a metadata lineage issue. An unbroken chain of metadata is required from the source system to the consumer of the data to achieve accurate impact analysis – and for the business user to understand the provenance of the data and trust the data.
- **Administration**. Each tool provides its own administration, scheduling, deployment and management services. The overhead of managing and maintaining each of these separate tool environments is significant. Like juggling, the overhead becomes exponentially more difficult as you bolt on different tools to achieve an end-to-end process.
- **Methodology**. Having different tools means that having a defined set of standards and methodology applied across the tools themselves is difficult.
- **Consistency and training**. As a user, learning the intricacies of different tools is difficult. If you are using only one tool, it is probably not a problem; but a key contributor to a project might be switching between profiling, ETL, data federation and data quality tools, each of which requires a certain expertise.
- **Performance and efficiency**. Some tools consume the services of other tools. For example, an ETL tool could consume the services of a data quality or federation tool. There is the inherent performance overhead of transferring data between one tool and another, service instantiation, data transformations between tools, etc. Having the same types of services, such as data transformation and data parsing services, within each tool is also inefficient. And, there is the risk that the same type of service in each separate tool may be inconsistent.
- **Teamwork, collaboration and reuse**. Data integration has evolved from one person coding a simple process to global initiatives involving numerous people from business and technical services. Having multiple tools makes it difficult to propagate and coordinate valuable information captured in each tool. As an example, profiling information in one tool should be immediately available to a person defining an ETL task on the same data. The metadata captured in the data exploration, the notes added by the business user, and the profile task executed on the data should be immediately available to a person working on that data. Having multiple tools – each with its own separate silo of information – makes this collaboration and sharing of information highly restrictive.

Conclusion: Moving from a Data Integration Toolbox to a Comprehensive Data Management Platform

This data integration toolbox approach has not been an issue for SAS or its customers. SAS has been delivering many of the benefits that a single, integrated data management platform promises for many years, though not with a single user interface.

From a data integration perspective, SAS provides the SAS Data Management solution, which lies at the foundation of the platform for SAS Business Analytics. Data quality is delivered seamlessly through SAS functions or as web services.

That said, and considering the issues with multiple users and multiple tools, there is a fundamental change occurring in how software vendors need to deliver data integration tools and technology to organizations. To support organizational data and information needs of both an operational and decision support

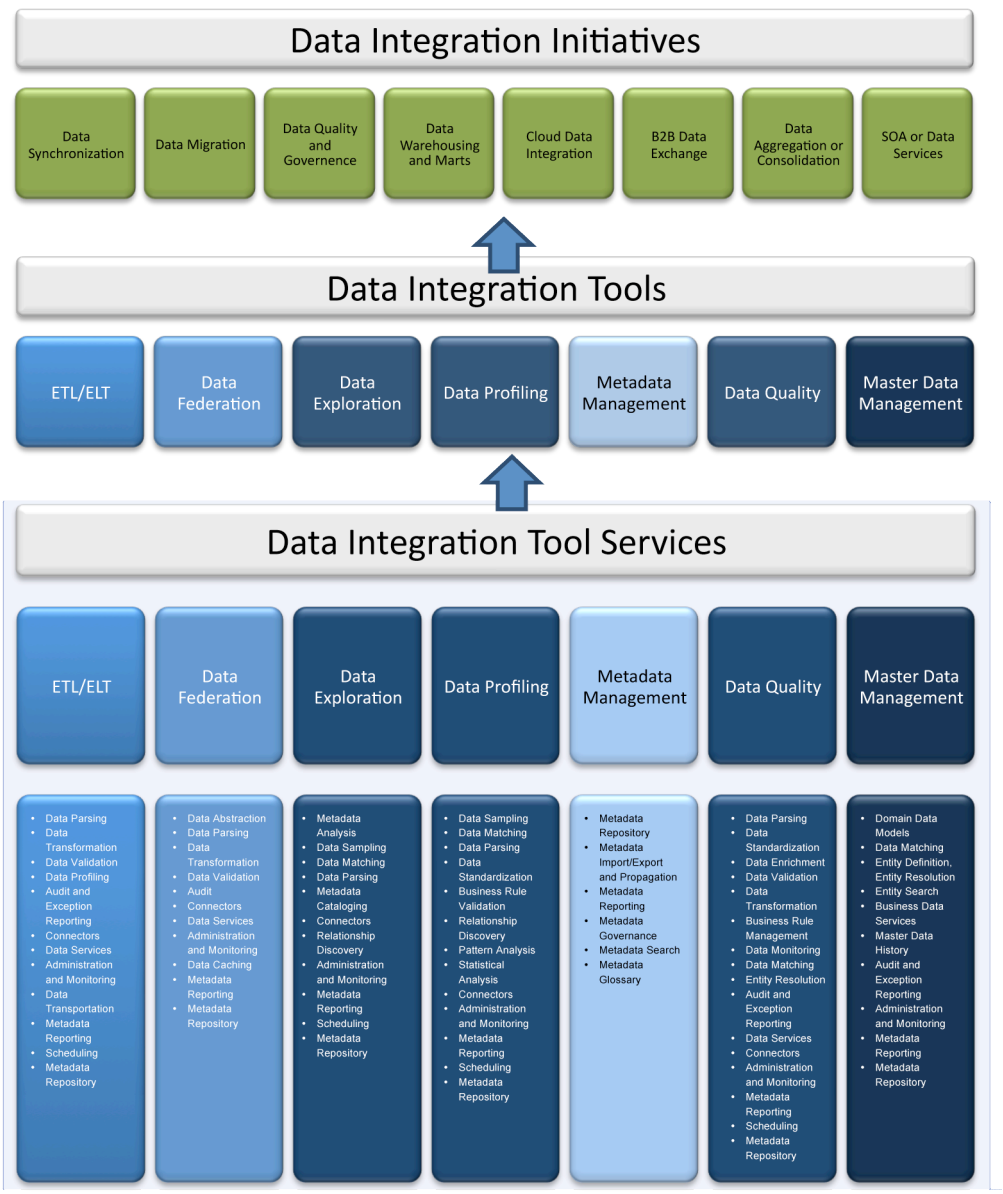

Figure 8: A single-vendor data management stack helps eliminate integration issues.

nature, vendors must move from supplying a toolset of loosely (and sometimes poorly) integrated tools to delivering a single data management platform with three key elements that deliver an enterprise solution.

The three key data integration tools and technologies coming together are data integration, data quality and master data management. When bringing them together, data quality encompasses all aspects of data exploration, profiling, cleansing, monitoring and business rule validation. Data integration covers ETL/ELT, and master data management shares functionality from data quality but provides all of the key services around business data services, entity definition and management, etc. These three key areas must be underpinned by adapters and a federation capability and have metadata sharing, which provides propagation of technical and business metadata and also aids in collaboration.

To be successful, your data management platform of the future will need to be based on a single, shared architecture. Depending on your role, users of the platform will launch modules within a single user interface to provide data exploration, profiling, business rules, ETL/ELT and other capabilities. As a business user, you might be defining business rules or looking at a profile report to identify where data in a source system doesn't comply with the business requirements defined. On top of a single, shared architecture, anything captured or developed by one user is available to be consumed by another user through remediation, workflow or other mechanism, enhancing collaborative efforts.

Once people realize and experience the benefits of a unified data management platform, it is easy to draw the conclusion that the data integration toolbox of the past will no longer be viable for the future.

Learn More

SAS is a recognized leader in data management and business analytics software and services. For more information, please visit sas.com/data.

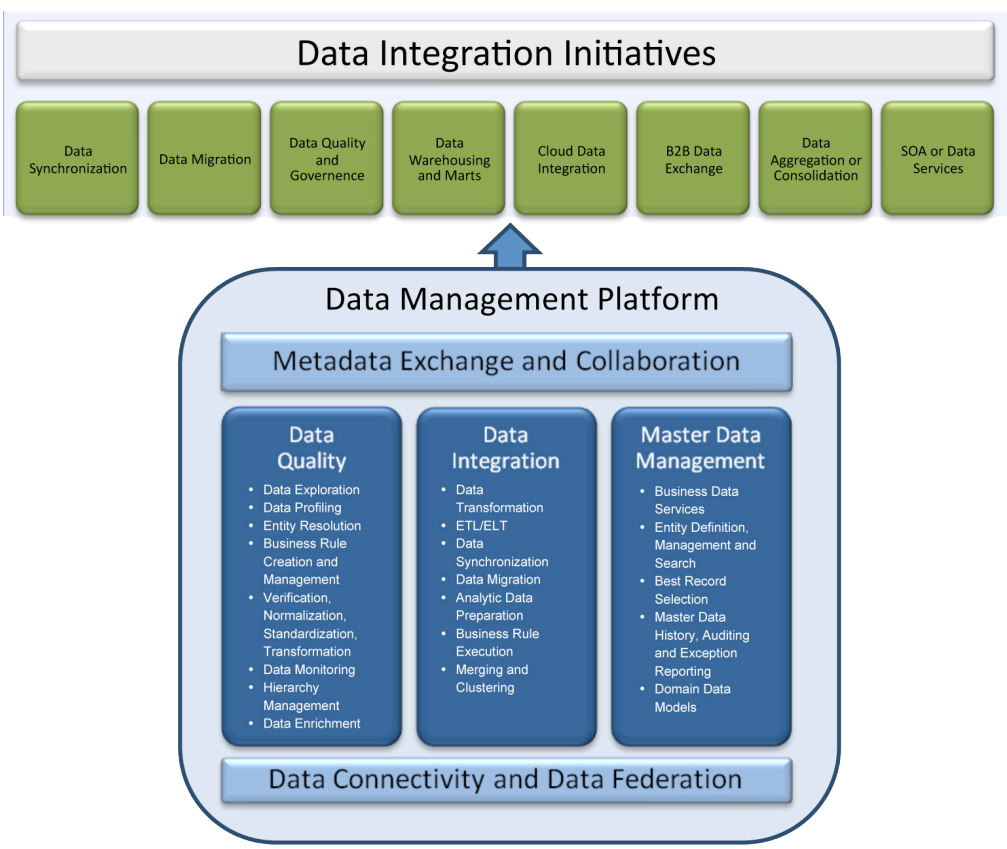

Figure 9: A single data management platform with integrated tools enables organizations to better support their operational and decision support information needs.

To contact your local SAS office, please visit: sas.com/offices

104819_S118250.0714

THE POWER TO KNOW.