Making Everything Easier!™

Informatica Special Edition

Data Integration FOR DUMMIES®

A Wiley Brand

Learn:

- What data integration is and why you should care
- How data integration can help your business become more agile
- Common data integration challenges and benefits
- What to consider when looking for data integration tools

Compliments of

informatica

Put potential to work.™

Brian Underdahl

Data Integration

FOR DUMMIES®

A Wiley Brand

Informatica Special Edition

by Brian Underdahl

FOR DUMMIES®

A Wiley Brand

Data Integration For Dummies®, Informatica Special Edition
Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-118-89658-7 (pbk); ISBN 978-1-118-89692-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Jennifer Bingham

Acquisitions Editor: Connie Santisteban

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Custom Publishing Project Specialist: Michael Sullivan

Project Coordinator: Melissa Cossell

Special Help from Informatica: Andrew Taylor, Dominic Sartorio, Todd Goldman, and David Lyle

Table of Contents

Introduction

Data integration refers to an industry category and a technology designed to make it easy for you to combine different sources of data together to produce useful business information. Those sources may include legacy mainframe databases, modern relational databases, desktop applications, social media comments, blog postings, machine sensor data, and so on.

But modern data integration tools shouldn't be limited to today's data sources because the best of these tools are also flexible enough to encompass whatever new technology becomes the hot item tomorrow.

About This Book

You may be familiar with data integration but are still using hand-coding approaches to do it. Or perhaps you're trying to figure out exactly what data integration is and whether it should be part of your data process. If so, this book is designed to help.

This book shows you what data integration is, how it works, and how you can use the technology to become more competitive as you combine your existing data sources with new data sources to provide the business intelligence your organization needs to compete effectively.

This book was created with and for Informatica.

How This Book Is Organized

This book is divided into five chapters.

Chapter 1 – More Data, Better Tools, More Opportunity

This chapter looks at the rise of data and discusses how both data and data infrastructure have changed over the years. The chapter also provides a look at some examples that show how data integration has been used to solve problems in several different industries.

Chapter 2 – Data Integration 101

This chapter introduces you to some common data integration terminology and offers a basic understanding of how data integration works.

Chapter 3 – Understanding Data Integration Challenges

This chapter examines the challenges that organizations face in implementing data integration projects.

Chapter 4 – Understanding the Benefits of Data Integration Tools

This chapter shows you how modern data integration tools address the challenges and provide benefits such as increased efficiency, better decision making, and faster development of the solutions you need.

Chapter 5 – Top Ten Things to Look For in a Data Integration Tool

Finally, this chapter provides some tips on what to look for in your data integration solution.

Conventions and Icons Used in This Book

This book uses several standard *For Dummies* conventions:

- Defined terms are *italicized*.
- **Boldface** type indicates the keyword in a bulleted list.
- Web addresses are in a distinctive `monofont typeface`.

  Some web addresses may have needed to break across two lines of text. If that happened, no extra characters (such as hyphens) were included to indicate the break.

In addition to the standard *For Dummies* conventions, this book makes use of some standard *For Dummies* icons — those little illustrations in the margins of the book meant to draw your attention to the text next to them.

The information marked by this icon is important. This way, you can easily spot noteworthy information when you refer to the book later.

This icon points out extra-helpful information.

Paragraphs marked with the Warning icon call attention to common pitfalls that you may encounter.

Beyond the Book

You can find additional information about data integration (and about Informatica's approach to it) by visiting the following websites:

- **Data Integration for the Enterprise:** `www.informatica.com/us/products/data-integration`
- **Data Integration for the Department:** `www.informatica.com/us/products/data-integration/department`

- **Informatica Cloud:** www.informaticacloud.com/cloud-integration
- **Data Quality:** www.informatica.com/us/products/data-quality
- **Informatica:** www.informatica.com/us

Chapter 1

More Data, Better Tools, More Opportunity

In This Chapter

- Looking at data's evolution
- Examining data and IT infrastructure over time
- Understanding the impact of data

Business thrives on data. That's because it can be turned into useful information and insight that provides a competitive business advantage. Today, many additional data sources are available beyond the traditional relational database and legacy mainframe. Despite what many people would say, legacy data is still very important, but now you need to also integrate data from customer spreadsheets, SaaS CRM systems, social media dialogues, log records, and even sensor data. You have to accumulate and sort through data that's created by businesses, governments, social media sites, and intelligent devices. The more relevant data you can capture, process, and provide to the business, the more competitive advantage and differentiation you can create.

This chapter looks at how the world of data has changed and also goes over how integrating legacy and modern data sources can help you become more business agile and competitive.

The Rise of Data

Before computers, most data consisted of handwritten ledgers tallied from paper records such as sales receipts or company accounting information. In that era, creating useful business information was cumbersome, time-consuming, and prone to human error.

When business organizations began using computers, the handwritten ledgers were replaced by mainframe databases where the data was entered manually on terminals. This manual entry process meant that data collection was still fairly slow and expensive, as well as limited in scope. Over the years, new and distributed systems have been developed that enable far more efficient collection and use of data.

Today's computer systems exchange data with systems and devices both inside and outside the organization. Data is collected from a growing number of sources and turned into information that you simply wouldn't have had in the past. You're able to integrate and use publicly available data collected and managed by others to further enhance your organization's knowledge and database.

To better understand how this flood of data from new sources can have a major effect on business competitiveness, consider what has happened to securities trading firms. Their ability to collect data on price differences for the securities they trade from various markets around the world allows them to quickly buy and sell those securities at a profit. By collecting more data from more sources, they're able to analyze the data and apply their unique models to gain a competitive advantage. Even slight advantages in obtaining and analyzing data for faster processing can be hugely profitable.

The ability to collect and use more data brings both opportunity and risk. For example, data about the human genome has raised concerns regarding individual rights to privacy while at the same time providing new opportunities for improved health care.

Much of today's new data comes from sources such as machine sensors or smart devices like Internet-connected refrigerators, smart thermostats, or even your cell phone. These new data sources have come to be known as the *Internet of Things,* because the "things" automatically communicate on their own without human intervention.

For example, did you know that your car creates and consumes data and can probably transmit that data without you even knowing about it? In fact, most modern electronic devices have this capacity. The Internet of Things consists of millions of devices, sensors, and other components that

can be uniquely identified and connected to the Internet. This system is a massive source of fragmented data which, when combined with your traditional data sources and used in the right way, can allow you to reduce uncertainties and unknowns and provide you a tremendous advantage over your competition — but only if you can turn the raw material into high quality, useful information.

Looking at Data and IT Infrastructures Since the 1960s

The past 50 years have seen dramatic changes in data and IT infrastructures. In the beginning, the mainframe computer was the single repository for data. If you needed to generate business information from your data, you got that information from your central mainframe.

Now, however, as Figure 1-1 shows, the number of data sources has exploded. Instead of a single source of data, you now must contend with multiple, diverse sources.

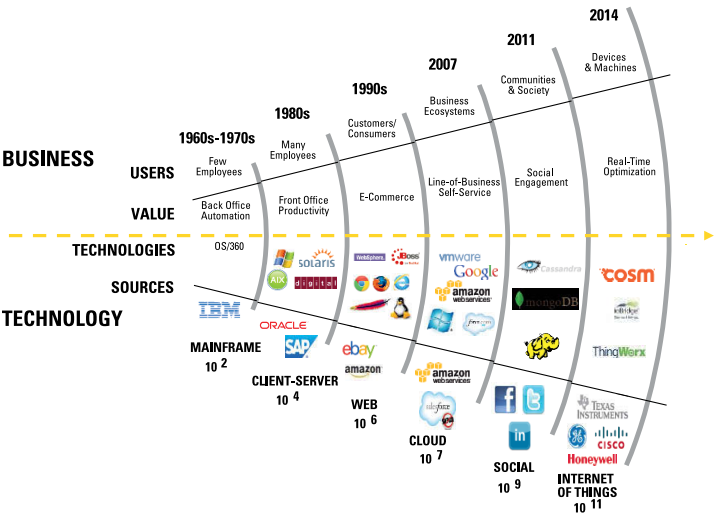

Figure 1-1: Historical view of how data and IT infrastructures have grown.

At the same time, the legacy data didn't go away — you now need to integrate it with the new data that's available from many different sources so you can produce more useful and differentiating business information.

Here's a quick look at how data and IT infrastructure have changed over the past 50 years:

- **1960s and 1970s:** This was the era of Big Iron. Mainframes and a limited number of employees skilled in managing and programming in that environment were the only option business and government had. Basically, there was very little choice of how to automate the business beyond what a limited number of vendors offered. Many of the applications were hand coded and maintained.
- **1980s and 1990s:** Client-server architectures created more choice and loosened the control of the back-end data center. Organizations gained more flexibility over their computing processes and the value they derived from the new technology. More vendors meant more opportunities. And along came more data that had to be managed and stored.
- **1990s and 2000s:** Organizations begin to see more distributed technologies that brought in more data — not just from within the organization, but from outside of it, too. The new technologies started to open up the world and data sharing became more common among organizations and individuals. The old ways of creating and managing data began to drastically change, and the technologies supported that change in a big way. Consider, for example, how many things have moved to the web. Not only has e-commerce made a big dent in traditional brick and mortar, but social media has changed the way people interact and share their opinions.
- **Today:** Businesses and government entities now have the technologies to take advantage of data on the web, in the cloud, from social media, mobile devices, and from machine sensors. The amount of data and the number of sources of information organizations can use has exploded. Valuable data is growing and being stored at phenomenal rates. This data must now be managed and integrated from numerous sources for businesses to get an accurate and clear view into its meaning. Without that analysis, businesses can't differentiate themselves in the way they compete.

Examining Some Data Integration Examples

Data integration may sound interesting on its own, but there's nothing like a good real-world example to show its true value. This section gives you a look at how several types of organizations can use data integration to improve their services and bottom line along with a use case for each.

Healthcare

By integrating legacy data with social media data, health organizations can make better predictions about the spread of contagious diseases — and thus make more informed decisions to protect public health. Even a day or two saved in getting vaccines to the right location or implementing a quarantine can make a huge difference in limiting the extent of a health crisis.

Very recently, researchers at a renowned East-Coast university experimented with a new approach aimed at providing more immediate information about the current status of flu infections. Rather than relying on traditional data sources, the researchers used advanced algorithms to look at unstructured data from Twitter feeds. The researchers analyzed hundreds of thousands of tweets to determine locations where people had the flu. This social media data provided an up-to-the-minute and predictive look at where the flu was spreading at a particular point in time. Without integrating the new types of data from social media with the existing legacy data, it simply wouldn't be possible to gain this quick insight.

In the past, the data that's been available for tracking flu outbreaks has been traditional, legacy type data that only tells you what happened after the fact — and after the data has been collected and analyzed. In other words, you can see what has happened in various parts of the country, but because the information is at least several weeks old, it's not very useful for predicting where you'll need to send additional supplies of flu vaccine or where you'll need to ramp up staffing levels at clinics and hospitals in order to stem the epidemic.

Financial

Financial institutions are also finding that they need the power of data integration to better compete in today's market. They need to understand who their customers are and how to deliver services that fit their specific needs.

A large bank wanted to offer better, more customer-oriented services that required it to rapidly access and integrate existing customer, product, and activity data across multiple business applications and legacy transactional systems. The bank also wanted to be able to find and fix data quality problems, such as incorrect customer address data, duplicates, misspellings, and inconsistent values.

To meet these goals, the bank embarked on an enterprise-wide data integration and data quality strategy to reorient around the customer. By integrating data from different disconnected sources, the bank was able to understand not only what each customer valued but also the value the customer represented to the bank's bottom line. The result of using data integration was better business intelligence and customer knowledge to segment customer audiences, tailor business streams, deliver value to customers, and target the customers that would deliver the bank more revenue.

Retail/B2B

Retail organizations need to adapt to an ever more competitive marketplace. Quite simply, customers have more options, so stores need to be able to provide better service to remain competitive. And they need their systems to be organized and up to date so internal teams aren't working at odds or hunting for information that should be easy to find.

A large business product sales group faced the challenge of creating a unified sales order management system. The sales professionals responsible for business solutions relied on a CRM system in the cloud for their selling strategy. However, all the other crucial sales data, including prospects, sales orders, and devices deployed out in the field resided in the company's on-premise enterprise resource planning (ERP) platform.

This created a lack of business agility that was clearly highlighted when sales had to close a deal. First, they had to switch from the CRM system to the ERP system to find existing contract information. Then they emailed the order to the order administration team, which in turn manually entered the information into the ERP system. Talk about slow, repetitive, and error prone!

The company used data integration tools to provide bidirectional connectivity between the CRM and the ERP systems. This hybrid solution provides everyone in the company with a timely, complete view of all data associated with every business deal, including customer information, pricing, products, devices in the field, order data, and stock control. The unified integration process is completely automated, ensuring that no sales data is more than one hour old.

Government

Governmental organizations aren't immune to the need to leverage data in new ways, either. Tight budgets mean that more efficiency is vital to providing services with the limited available resources. Data integration makes it possible for government departments to make the best use of both data and funding.

Consider a state transportation department with thousands of miles of roads and bridges. Its transportation system is recognized as one of the most accessible, efficient, and safe systems in the United States; however, the same could not be said of the IT infrastructure that supported the state department of transportation's (DOT's) financial, construction, maintenance, and traffic safety programs. Most of the agency's information systems — some dating to the late 1970s — were based on a legacy mainframe and a rather old-fashioned database, with a highly indexed, hierarchical data structure at odds with modern relational data systems. Using that legacy data was very difficult, and trying to combine it with modern data efficiently was almost impossible.

Streamlining the data integration process was vital in enabling the DOT to rapidly realize its primary objectives. Modern data integration tools made it possible to quickly combine the legacy data and the newer data sources into valuable information that would make sound decision making possible.

Chapter 2

Data Integration 101

In This Chapter

- Looking at data integration
- Getting an understanding of the terminology
- Understanding agile development with data integration
- Considering how IT and business fit into the picture

So if the number of data sources has increased almost beyond recognition (see Chapter 1 for more on that), how can you get your arms around them and use the information most effectively and efficiently? I'm glad you asked. Data integration is the answer.

To make the best use of data integration technology, you need to understand not only what data integration is, but also the terminology that you'll see in discussions of data integration. This chapter provides that background, shows how data integration can make your business more agile, and provides insight into how IT and business can cooperate in making data integration work for you.

Defining Data Integration

Data integration allows you to move or visualize data from different sources, consolidate that data with more data, and apply business transformations to that data so it conforms to your organization's needs. It then provides standardized data in a format and place such as a data warehouse where it is consumable by the business. The data integration process addresses the problem of data coming from and being stored in many different, fragmented places, and formats (see Chapter 1 for more on why this is a problem). The goal is to

provide an end-to-end view of the business to make it more efficient, identify new opportunities, and improve decision making and future planning.

It's estimated that 80 percent of the work and expense of using data today comes in the data preparation process rather than the data analysis process. In the late 1990s and early 2000s, the figure was closer to 70 percent. Although neither figure is scientific, it's interesting that, while data integration technology has significantly improved over the years, at the same time, the complexity and variety of the data has gotten more complex at an even faster rate! Clearly, without the proper data integration tools, using data can simply be cost prohibitive in many cases.

Using modern data integration tools

Once you start using modern data integration tools, you'll be able to see what's happening with your information and where it's coming from.

Modern data integration tools handle the data collecting, data transformation, and data provisioning functions in a transparent and highly adaptable manner. They automate the documentation process and provide a visual representation that shows the data sources, what processing is done to the data, and how the resulting data is delivered.

In other words, data integration tools provide the ability to move data from many different sources, aggregate and transform that data to support how the business wants to consume it, and then make it available to the users who can make decisions based on that data.

Because of this greater visibility into information, one of the major benefits of data integration is enabling better decision making through more insightful data rather than seat-of-the-pants guesses. When decision makers know what's really going on instead of relying on assumptions, they can base their decisions on facts. For example, comments on social media sites may help you understand where best to expend time and money on product variations. A product manager may discover that what people really want is a green and purple polka dot smartphone case rather than the expected (but to some, boring) white one.

By providing a visual representation of the entire process, these tools help you determine the complete *lineage* of all data (so you can see where the data originated and where it is being used) and enable effective auditing of the process. This visual representation makes it far easier for people on the business side to understand and trust the data without relying completely on the IT department.

Why you should avoid hand coding

Before modern data integration tools were developed, companies knew they needed to consolidate and process data from numerous sources. In the past, however, integrating multiple data sources typically meant a lot of ad-hoc hand coding between different data sets, which resulted in great expense and extremely difficult maintenance.

Consider, for example, that old-school data integration projects might involve hand coding in SQL, C++, or even JavaScript to produce the desired results. Unfortunately, most one-off, hand-coded systems were either poorly documented or not documented at all, making upgrades or other maintenance a nightmare. If the programmer who developed a system left the company, someone else might have to spend considerable time trying to understand exactly what the application was doing in order to make any changes. A single small tweak, such as changing the data type of a field in one of the source tables, could easily break the entire system.

Today, larger companies recognize the cost of maintaining code much more clearly than small companies. But regardless of the size of your company, you simply can't afford to compete while trying to use yesterday's hand-coding techniques, especially when your competitors are integrating data with modern tools at a faster pace. The efficiency gains provided by the modern tools are enormous. Rather than hand coding, wasting time troubleshooting, and then redoing everything when requirements change, you can concentrate on getting results quickly using visual tools that enable you to see where data originates, how it is processed, and where it ends up. Plus, good automated tools let you reuse your development over and over again.

With a hand-coded system, you generally don't know what impact a change in the source code will have on the end result. But modern data integration tools make it easy for you to understand the impact of any proposed changes. For example, say you have a legacy system that uses Boolean values in a gender field (like a 1 to represent a female and a 0 to represent a male) to indicate that an employee must be labeled female if the male option isn't selected. Changing the data type for that field to a text or numeric field with additional options (like whether the employee qualifies for maternity benefits) could break applications that depend on the value of the field. With a modern visual data integration tool, it's easy to see where that field's data is used and therefore quite easy to see how making such a change will impact applications.

Understanding Data Integration Terminology

Just like any other complex topic, data integration has its own lingo. You may be familiar with some of it, and you can skip this stuff if so. In the following sections, I take you through some data integration terminology.

Understanding data

The term *data* can mean many different things. But ultimately, data is the raw information that you process and manipulate to produce business intelligence. For the purposes of data integration, you need to consider several different types of data:

- **Structured data:** This term refers to data that conforms to clearly defined rules regarding the format and content of the data. A typical business database consists of structured data with fields that contain specific data types, like dates, numerals, text, and so on. The database application enforces the rules so that users can't enter unexpected types of data within those fields: text in a numeric field, for instance. Structured data is typically the easiest

type of data to manipulate programmatically because you generally know what to expect from it. But keep in mind that just because data is structured, that doesn't mean it's simple. There are hundreds of structured systems, each with thousands of tables and hundreds of thousands of data structures. And the complexity of data goes beyond the relatively simple structure and format of the data to the semantics of what each data value actually means and how it relates to other values.

- **Semistructured data:** This term refers to data that is formatted according to certain accepted rules, but that can vary in structure depending on exactly what data is being provided. Examples of semistructured data include industry standard XML formats like HIPAA (Health Insurance Portability and Accountability Act), ACORD (Association for Cooperative Operations Research and Development), and SWIFT (Society for Worldwide Interbank Financial Telecommunication) files. Semistructured data typically defines the data structure and its meaning within the file, and requires that the application understand how to exchange data among different systems.
- **Unstructured data:** This term defines data that doesn't really follow a specified format. For example, emails, blog postings, comments on social media sites, video files, audio files, PDF-formatted purchase orders, machine sensor logs, and RFID data are all examples of unstructured data. Although it's true that something like an email contains specific information such as the sender, the recipient, a subject, and body text, the body information is generally more difficult to extract in an ordered fashion. Until fairly recently, processing unstructured data wasn't cost-effective due to the difficulties involved. Now, with *natural language processing* (computer understanding of normal human language), data integration technology can understand the sentiment or other ideas contained within the body of emails or social media communications.
- **Sensor data:** This term refers to data that is generated by various machines. For example, the sensors on jet engines that provide information about operating conditions and failures.

Understanding metadata

Metadata is information about data. Basically, metadata tells you what types of information are contained in any piece of data. For example, in a typical relational database, the field and table definitions are metadata. The metadata enables you to determine that a particular field contains date information, numeric currency values, or is a text field, as well as what kind of text is stored in this field (address, last name, product brand, and so on).

Many different types of data include metadata. When you download an MP3 file, the header of the MP3 file includes metadata that describes the type of file and information about its origin such as the label, the album title, and the artist.

You may be generating metadata and not even know it. For example, digital cameras typically add metadata to every digital photo that identifies the type of camera used as well as the date, the time, and the resolution of the image. Some smartphone cameras even include geotagging information in the metadata using the phone's GPS sensor — so you can determine exactly where a photo was taken. This information is part of the data but not the part you would typically view.

Looking at big data

One data integration expert has defined *big data* as being an amount of data larger than your current systems and processes will allow you to handle. Effectively, big data refers to the storing and analyzing of data that used to be thrown away because there wasn't an efficient way to process it or extract information from it.

One interesting application of big data comes from the aviation industry. Large airliners are fitted with hundreds of sensors that can produce over half a terabyte of data on a single long-range flight. By properly integrating this data with other data regarding the expected performance of various components, it's possible to do jet engine predictive analysis and thereby schedule maintenance before a failure occurs, greatly reducing the costs associated with maintenance.

In discussions of big data, you'll hear about things like *data lakes* and *Hadoop*. A data lake is a place to store practically unlimited amounts of data of any format, schema, and type. Data lakes are relatively inexpensive and massively scalable. *Hadoop* has all the attributes of a data lake. Hadoop is an open source software framework for storing and processing data on clusters of commodity hardware.

Location, location, location

Real estate professionals are well known for saying that the three most important factors in their field are location, location, and location. With data integration, the fragmentation of data across various locations can also be a big factor in the success of the project.

Traditionally, all data existed within the walls of the organization. You had your mainframe computer system in your data center, and that's where all your data was stored. This well-defined structure meant that it was easy to control access to that data, but it also meant that you were limited to using only the data that was available in your system.

In the past few years, much data storage has been moved to the cloud. And in many large organizations, applications have moved to the cloud, too. These days, justifying onsite development or data storage is sometimes more difficult than justifying cloud only. But it all depends on the nature of your business and its level of maturity. Some organizations prefer to have their data onsite for security purposes. Others use a hybrid model in which some of their data is in the cloud and some is on premises. Regardless, on-premise, cloud, and hybrid data strategies are factors to consider when integrating data.

Although you want a data integration tool that supports hybrid environments with cloud and on-premise, choose one that's also flexible enough to support new technologies as they come along. Remember, only a few years ago, no one had heard of the cloud or of technologies such as Hadoop. You can bet that new technologies no one has thought of will be knocking on your door, and you want to make sure that your data integration tool won't become obsolete — which may well happen if you focus only on what's hot right now.

One reason the cloud has become so important is that the people who need to use data and applications aren't always located in the same physical location. Data and applications accessible through the cloud can be used by people anywhere they can get a secure Internet connection.

In a data integration process, you can use the cloud to integrate data or, more likely, you can use a hybrid system where some data is in the cloud and some is stored locally. Also, the user interface can be in the cloud and accessed through a web browser or similar device. Even the actual data integration execution can be done in the cloud. Always consider the right solution for your business and business users but make sure the solution is flexible enough to support all expected scenarios seamlessly.

A good data integration tool should allow you to reuse the skills of your developers.

Mapping, business glossaries, and data quality

The data integration world uses some additional terms that are important to understand. These include the following:

- **Mapping:** Mapping is the process of defining the source and destination of data as well as the transformations to perform as you move the data. You might want to use data from different sources, such as two types of databases. In a good data integration tool, this mapping is depicted visually so that anyone can easily follow the path of the data from beginning to end.
- **Business glossary:** This glossary helps to ensure that everyone is on the same page regarding business terms and synonyms. For example, the glossary might define common data values that are important to the business process and the common terminology of reference to be used.

  One reason for needing this precise definition is that different groups within the company may have different functional definitions for the same things. For example, the sales team defines a *customer* as the person or

company who buys the company's products or services. But to IT, the customer is the business group that consumes the data and uses the applications.

- **Data quality:** The quality of your data is vital. Data quality problems can easily break a business process and have employees making inaccurate and misinformed decisions. For example, consider how poor-quality data can affect a delivery carrier such as FedEx. Obviously, it's important that the carrier knows when a business will be open so that it doesn't attempt to make deliveries outside of normal business hours. Data quality problems can include missing, incomplete, or incorrect data, and any of these issues will have adverse effects on your data integration process. The time to deal with data quality issues is as early in your process as possible.

Data integration doesn't necessarily include data quality functionality. These are two separate tasks and can be deployed separately. But you should make data quality a priority from day one — moving bad data just means bad data is provided to the business user. Often a company ignores the data quality problem initially and just focuses on getting the data to the users.

Seeing How the Agile Development Process Fits In

In today's competitive business environment, organizations need to be agile in order to succeed. You need to cut out the fat and move faster than your competitors to win. Conversely, if your competitors move faster than you, you might lose.

Agile methods can be compared to how Toyota changed car assembly processes in order to improve quality. In the old days, car companies considered quality control an afterthought. As vehicles went down the production line, there wasn't much thought given to defective parts or poor assembly. Defects would be dealt with after the vehicle came off the line. In the 1960s, Toyota introduced a new way of building cars that dealt with problems immediately. If a problem was discovered, the

entire assembly line was stopped and the problem corrected immediately. Not only did this method reduce manufacturing costs, but it also delivered higher quality products to customers. As a result, Toyota became known for its exceptional products and is now one of the world's largest auto manufacturers.

Around ten years ago, 17 software developers met at the Snowbird ski resort in Utah and set guidelines in place to improve the software development process. The group produced the *Agile Manifesto*, which defined the agile development process. Several of the developers at the conference later formed the Agile Alliance, a nonprofit organization dedicated to the principals discussed in the manifesto. This agile process recommends that development teams work closely with their customers, producing frequent small updates, and using the resulting feedback to improve the end product far more quickly.

The improvements you can reap from more agile processes and tools are truly remarkable, but they don't come for free. They require people to alter the way they work. A good data integration tool supports this new way of working.

Data has meaning and relationships that change over time, so you need data integration tools that not only move data but enable you to adapt quickly as needs change.

Chapter 3

Understanding Data Integration Challenges

In This Chapter

- Seeing the technical issues affecting data integration
- Understanding the process issues in implementing data integration
- Looking at how politics and organization play a part

Any large project presents challenges, and data integration is no exception. The challenges may be technical, such as making sure you can effectively handle the variety, volume, and velocity of data. Or you may need to spend some time figuring out how to streamline processes so you can repeat your successes. You may even need to convince certain groups that "their" data needs to be shared. Fortunately, understanding challenges ahead of time can help prepare you for them — and guide stakeholders through ups and downs. Knowing about issues before they arise can help ensure success for your project.

Understanding the Technical Challenges to Data Integration

The most obvious challenges to implementing a data integration project are probably technical. But fortunately for you, those technical challenges are usually the easiest to resolve — largely because technical issues are understandable, measurable, and have been addressed before on many other projects. Plus, modern data integration tools can help you benefit from the lessons already learned — why learn them again?

In any discussion of data integration, you're likely to hear numerous mentions of big data (see Chapter 2 for more information on that). In reality, what many would call big data is no more of a challenge than any other data. Sure, there's more volume, but the right data integration tools make handling that larger volume just about as easy as handling any data. Once everyone gets used to the fact that data went big and is going to stay big, the terminology will likely change from *big data* to *data.*

More important than whether or not the data is big data is the question of why you want to move or aggregate your data. There can be many answers to this question, such as a desire to improve marketing efforts or the need to comply with new regulations, but fundamentally, data integration consolidates data from a number of different sources and formats in order to produce useful business information that better illustrates the bigger picture.

The technical challenges to data integration are often lumped into the three Vs: variety, volume, and velocity. A number of people talk about veracity as a fourth V. A few people are starting to talk about value as the fifth V. All five are important and should be vital to your planning process. The following sections discuss the specifics of these five categories.

Variety

The first technical challenge to data integration is the *variety* of data. Today, many kinds of data exist, ranging from highly structured data found in legacy mainframe databases, to semi-structured data that follows various industry standards such as XML, to relatively unstructured data such as web content and social media comments. (This is discussed in more detail in Chapter 2.)

A relatively new type of data that you need to consider is machine sensor data. For instance, the smart meters that utility companies use to remotely measure energy consumption collect sensor data. Or the hundreds of sensors incorporated into modern automobiles that monitor how your car is running as well as logging data immediately before an accident.

The challenge provided by data variety is that your data integration tool must be able to map and transform these various types of data sources into something you can use. Most data integration tools are able to handle structured data quite well, but often lack the ability to easily incorporate newer data types. A key issue to address when evaluating data integration solutions is extensibility and whether the tool will be able to work with new data types as they come down the pike.

Don't assume that today's data structures won't change over time. Even structured data has new columns added to tables or fields that are changed to a different data type to suit new needs.

Volume

There's always more data, never less. The constantly growing *volume* of data is at the heart of many data integration challenges. For example, you need to consider how to consume and move large quantities of data in a short time. In legacy systems, you might only need to consider something like filling your data warehouse. Now, you still need to fill the data warehouse, but you have many more customers, many more transactions, and many more sources of data to consider. Make sure you have the right hardware and systems that can handle this greatly increased volume of data.

Partitioning is a data integration term that refers to taking large quantities of data that you break up into smaller pieces and process in parallel to improve data processing performance.

Some companies are turning to Hadoop — a technology for consuming and storing vast quantities of (typically unstructured) data in parallel and on clusters of low-cost hardware. Hadoop is open source technology and is designed to automatically handle hardware failures. But Hadoop isn't a data integration tool per se.

Make sure that the data integration tools you choose can easily scale to handle increased data volumes across different platforms and that those tools have enough headroom for future expansion.

Velocity

The third technical challenge to consider is *velocity*, or how quickly data is coming at you and your ability to consume and move it. Twenty years ago, moving the data overnight, in batch mode, was usually considered fast enough, which meant that your business analysts were using yesterday's data. Everyone knew they were using yesterday's data, and this was fine.

But things move much more quickly now, and working with yesterday's data won't do. It simply puts you at a great competitive disadvantage. Real-time data is vital for processes that use predictive analytics, such as pinpointing credit card fraud. It doesn't do much good to determine that someone used a stolen credit card yesterday considering how quickly identity thieves can move to purchase thousands of dollars of goods and services online. Such processing is really only effective when done in real time or near real time. Other examples where real-time processing is vital include counterterrorism work or a power company detecting that part of the power grid has gone down. In these examples, waiting until the next day to process the data is unacceptable and dangerous.

Latency can be a real challenge for data integration tools. Latency is the amount of delay in processing data. It can cover a broad spectrum of values from the slow end where you're processing yesterday's data to the fast end where you're processing real-time data in nanoseconds. Most enterprises have data integration requirements and policies that cover this whole spectrum. You need to make sure that your tools can monitor processing for the required amount of latency to make sure that data is being delivered to the business within the desired timeframe.

Veracity

Veracity refers to the quality or cleanliness of data and how certain you can be that the data you're seeing is indeed accurate. Veracity addresses the question: Can this data be trusted? Things that cause trustworthiness of data to be called into question include: ambiguities, duplicates, latency, spam, inconsistencies and model approximations, among other things.

The biggest positive impact you can make on your data integration project is to ensure data quality. One expert maintains that at least 20 percent of all raw data is incorrect, and that this inaccuracy is the largest challenge faced by any data integration project. Inaccurate data leads users to question the veracity of any information the system provides.

Value

Last, but certainly not least is *value*. It is probably the most important but least mentioned of the Vs. Unless you can turn your data into something of value, it's pretty much useless. Simply collecting a lot of data without making a good business case around the project is meaningless. You need to consider what all of that data means to your bottom line both in terms of the costs and of the benefits.

It's quite easy for executives to be caught up in buzz words and insist on pushing forward with the latest trend without understanding the big picture. Yes, big data does have a lot to offer in many cases, but don't fall into the trap of thinking that "everyone else is doing this so we have to as well." Make certain that your big data project will contribute more to your bottom line than simply some extra costs. Make sure that your project actually returns some real value so that you can feel confident in your recommendations.

Looking at Process Issues

Technical challenges aren't the only hurdles you face in integrating your data — you also need to consider process issues. Process challenges can be categorized as either repeatability or collaboration — for instance, you need to make sure you can repeat your data integration processes more than once while also streamlining those processes and encouraging groups to work together.

Repeatability

In the old days, you probably had a process to load your data warehouse, and that was the end of what you needed to do. You may have had a batch file that went through the various

steps of loading the data warehouse and after you set that up, you probably didn't need to make any changes for a considerable time. Loading your database was the last step in what is called ETL. ETL refers to a process in which data is *extracted* from data sources, *transformed* to fit or support business needs, and then *loaded* into the final location (like a data warehouse).

Now, with so many new and ever-changing sources of data available and desired by the business, you probably need to consume and move that data to many different places so that it can be used by different business groups. For example, you may need to combine data in different ways to serve the needs of order fulfillment, marketing, and new product development.

Data integration isn't about just one project; more typically it's a bunch of projects that happen over time. Different business, groups may ask for different data sets to support their projects. As a result of these different needs, you'll have to develop new mappings or incorporate new data sources in order to produce the desired results. But you don't want to start from scratch each time: Your data integration tool should allow you to design repeatability into your system so that you can reuse many of the same pieces in new ways.

Fortunately, people's skills improve over time. Developers become familiar with the tools, business managers gain better understanding of the process, and organizations generally get more efficient at data integration projects with experience.

Try to develop a core group of people who will be involved in your data integration project, so that their skill sets will improve and make your future projects easier. This core group should develop best practices that answer the questions of how to improve the processes and data quality over time.

Newer or smaller companies sometimes overlook the need for repeatability, treating data integration as a one-time project in which a developer simply loads the data warehouse and walks away. Make sure that you consider the need for repeatability. A data integration tool helps retain the process and knowledge within the organization and allows new users to pick up from where someone may have left off.

Collaboration

Next, consider the collaboration that must happen between the business groups and IT for a data integration process to be successful. Generally, each business group has different reasons for moving data, specific requirements about how that data should appear, and a need to receive the data while it is still useful and timely.

The person who specifies the precise requirements of the business group is typically called a *business analyst*. The business analyst is the person who communicates the requirements to IT, and data integration tools should help support and accelerate collaboration between business and IT.

Often, business analysts have business expertise, but little in the way of technical skills. Likewise, IT may have great technical skills, but little business experience. The result of these two differing skill sets is that it's sometimes difficult for the business analyst and IT to understand each other.

Such a lack of understanding can easily lead to extended project deadlines, where what IT develops simply doesn't meet the needs that the business analysts thought they were clearly expressing. This back-and-forth process between business and IT is fraught with human error and miscommunication, all of which a data integration tool can help address. Better data integration and data quality tools facilitate collaboration between business and IT, reducing the lag time between requirements definition and delivery — and ensuring a better end-result.

Specifically, better data integration tools provide the means for business analysts to define exactly what they need. Then IT can use the same tool to understand what needs to be done to fulfill those needs. The two groups should also use this same tool for several rounds of reviews to make sure that everyone understands the process in the same way. After the first round of reviews, developers can do a quick prototype to validate that they understand what the business analysts want and need. See Chapter 2 for more information about how your company can use the agile development process to make frequent reports on small steps in the process. This can help keep projects on track, going in the right direction, and meeting the business's deadlines.

Considering Political and Organizational Roadblocks

Unfortunately, political and organizational roadblocks can challenge your data integration project. You need to do more than provide great tools and communicate effectively about them. For those larger projects you have to motivate others in your company, people with influence, to see the benefits of what you have to offer. Once you have that, you can work to ensure you have the proper budget to build a team and the influence to encourage data sharing. After all, others need to share data in order for you to integrate it.

Finding a sponsor

Find an executive sponsor who understands the value of data to the business, and can make sure that your project has the proper budget and that the right people are working together to ensure success. Without an executive sponsor, you'll have different people doing different things that advance their priorities without considering what other groups need.

In a large organization, your data integration group should be a separate group with its own manager and its own budget. Smaller companies may not be able to swing this, but may want to investigate products designed for smaller organizations that can scale up as needs grow.

Budgeting

A successful data integration project requires the right tools and resources. This includes having the right people assigned as well as doing the proper planning and obtaining the correct tools. Most of these things aren't free or even cheap, and it's important to budget properly as well as choose tools that will scale. If you need to start small, be sure to use a tool that can do what you need today and grow with your needs in the future.

Look for tools that meet a broad set of needs, from companies just dipping their toes into data integration to large enterprises with the most complex data integration needs. If you start small, make sure you can grow with the same tools so you don't have to throw everything away and start over. The business department can experiment with ownership and self-service tools before spending any money.

Encouraging data sharing and data quality

Different groups often feel they own their data, and therefore they become very protective of it. Rather than allowing other groups to access their data, they sometimes don't want to share or insist upon sharing on their terms only.

Data quality (or perceptions about data quality) can be part of what makes people territorial. For example, say one business group (Group A) is consuming data from another group (Group B). Group A feels the data it's getting from Group B has quality issues because it doesn't quite fit its needs. Group B feels the data is just fine for its needs and doesn't even want to share with Group A any longer. But data quality isn't a "them versus us" issue — it's important to the whole company. You need collaboration where everyone understands how important data quality really is. For example, a customer database needs to be up-to-date with accurate spellings as well as correct customer contact information.

You may want to have a chief data officer who owns data quality and ensures that the data can be trusted (or at least someone who takes ownership if you don't have the budget for another executive). This person is responsible for policies ensuring best practices are followed and collaboration occurs on data quality issues. Such a person can sit down with Group A and Group B and come up with a solution to make both groups happy. If Group A needs Group B to harvest additional data or double check accuracy, maybe Group A can dedicate some of their budget to the data harvesting process. See "Finding a sponsor" earlier in this section for more thoughts on this subject.

Chapter 4

Understanding the Benefits of Data Integration Tools

In This Chapter

- Seeing how data integration tools can help
- Considering agility
- Making the most of the people you have
- Looking at how you can scale to fit future needs

The right tools can make quite a difference in the success of any project. You won't see a watchmaker using the same tools as an auto mechanic, because although the auto mechanic's tools are useful for working on your car, they're not very efficient for the fine work of repairing a watch. And just as the watchmaker's tools enable the watchmaker to be far more efficient at watch repair, specialized data integration tools greatly improve the process of implementing data integration solutions, especially when they do away with hand coding.

This chapter examines how specialized data integration tools can offer great benefits in efficiency and productivity to your data integration projects.

Considering the Benefits of Data Integration Tools

Modern data integration tools allow you to consolidate data from many different sources to create a clearer perspective of the business for differentiated decision making. These

specialized tools improve the data integration process by providing a highly visual user interface that enables a clear view of data in motion. You can see exactly what's going on. For example, data mappings that are presented visually allow everyone to understand precisely where each piece of data originates, how the data is processed or transformed as it passes through the system, and exactly where the transformed data is going. Figure 4-1 shows an example of this type of tool.

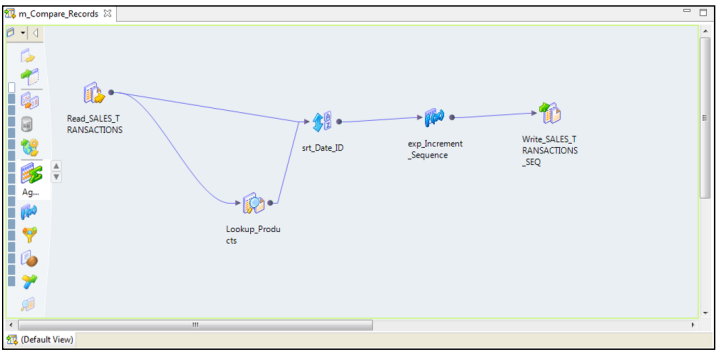

Figure 4-1: Visual Data Integration tools show you exactly what's going on.

For instance, with a modern data integration tool, business or data analysts can open a screen to see where developers are getting information and how they're going to process it. While it's being processed, analysts can see what's going on — they're using the same tool and are able to view the same screen.

The old, traditional, and time consuming way of collaboration has no true real-time or visual representation. Business analysts explain what they want (often in a document, spreadsheet, or email) and then some time later — weeks or months — developers respond with something that may not represent what the analyst had in mind. Using these methods, the whole process is error-prone, time consuming, and frustrating. Figure 4-2 shows an example of the way traditional hand-coding methods made it very difficult for business analysts to understand what developers were doing.

```
            AND l.lgcy_typ_cd <> '32' THEN 'PRODTYPE XTND'
        WHEN l.lgcy_product_cd
             || l.lgcy_typ_cd IN ( '0605', '0932' ) THEN
        'PRODTYPE RURAL'
        WHEN l.lgcy_product_cd = '02'
             AND l.lgcy_typ_cd <> '06' THEN 'PRODTYPE LGL'
        WHEN l.lgcy_product_cd
             || l.lgcy_typ_cd = '0206' THEN 'PRODTYPE GEN'
        WHEN l.lgcy_product_cd = '08' THEN 'PRODTYPE 08'
        ELSE 'OTHER'
    END                 AS product,
    SUM(f.aa_pol_cnt) AS value_current,
    0                   AS value_pl2,
    0                   AS value_pm
FROM   edw.edw_std_coll a,
    edw.d_time_mo d,
    insya_dm.f_actv_ratn f,
    insya_dm.d_lgcy_product_typ_view l
WHERE  f.dt_key = d.dt_key
    AND f.prty_id = a.person_key
    AND f.lgcy_product_typ_key = l.lgcy_product_typ_key
    AND l.lgcy_product_cd NOT IN ( '05', '07' )
    AND d.period_st_dt BETWEEN (SELECT
        Add_to_date(d.period_st_dt, 'MM', -23)
                                FROM   edw.d_time_mo d
                                WHERE  d.curr_mo_ind = 'Y')
                               AND
                               (SELECT d.period_st_dt
                                FROM   edw.d_time_mo d
                                WHERE
                                   d.curr_mo_ind = 'Y')
GROUP  BY a.person_key,
```

Figure 4-2: Traditional hand-coded techniques were hard to understand.

Unless business analysts are highly technical, about all they can do is hope the developers understand exactly what they want. As a result, analysts probably won't be able to figure out if developers got it right until after all the coding is finished. Often, the only way to recover from these mistakes is to start the process all over again.

But with modern data integration tools, the business analyst and developer can collaborate in real time to figure out problems with how the data is being processed. Validation of the specification happens early in the process against real data. The analyst can say, "Look at this part right here, I'm not getting what I want, can we try this other data source instead?"

Visually oriented data integration tools enable you to easily follow the lineage of data, making change management easier because you can see exactly how any changes in the source data will affect the outcome. Without this type of visual representation, you'll end up doing a significant amount of rework due to unintended consequences from seemingly simple changes such as changing a data type.

Another benefit of visual data integration tools comes in the area of governance. If you happen to be in a highly regulated industry such as banking or healthcare, you need to demonstrate exactly where data originates, how it has been handled, and how it is stored securely. Robust data integration tools can provide the documentation you need to comply with these types of regulations.

One of the big dangers of trying to use spreadsheets and SQL or other scripting languages for data integration projects is that hand coding lacks visual clues and makes future modifications or reuse extremely difficult, if not impossible. Hand-coded projects tend to be very poorly documented (if they're documented at all), and even the person who did the original coding usually finds that maintenance can be a nightmare.

Understanding How Data Integration Can Make You Agile

Agility is important. You want your company to be able to perform quickly — without necessarily making a big deal out of it. For example, if you're moving your CRM system to something like Salesforce in the cloud, you need to be able to load your customer data into that new system. That's a data integration function, and a good visual data integration tool can make the process easier and faster.

Integrating data from many functions, departments, partners, ecosystems, and so on allows you to innovate and differentiate your business because you have a clearer perspective of what's happening. The innovation also applies to the tools you use to achieve the movement of the data and its flexibility.

But just as important as making the original process quick and easy is that a good data integration tool makes it easier for you to reuse what you've already done and adapt it to new purposes. For example, you might want to use that same customer data along with data from social media to develop

a rewards program for customers who spread the word about your company on social media sites. Developing this type of application by hand would be difficult, slow, time consuming, and probably not very cost-effective. On the other hand, good data integration tools could help you consolidate, integrate, and cleanse the data so you can quickly get to the analysis phase and speedily create the results you want.

Most data integration problems you encounter aren't unique. Developing your own data integration solution from scratch means that you're reinventing the wheel: relearning information, re-creating mistakes others have made — and repeating these processes slowly rather than quickly. Using good data integration tools means that you can reuse the knowledge other people have gained over the years — and you'll be quite agile and able to respond quickly as your organization's needs change.

Vendors with modern data integration tools may facilitate communities and online stores where complimentary resources and mappings are available and shared. A robust community and online store are great resources to check out before you start any development, because they can save you a great deal of time if an existing resource (such as a mapping) could be leveraged as-is or fine-tuned to meet your specific requirements.

Using the People and Skills You Have

Different people have different skill sets, of course. Hiring different teams of developers who are experts in different technical areas can be very expensive. Ideally, you want to make the best use of the people and skills you already have, rather than spending large sums to support new technologies.

One way to leverage the people you have is to use data integration tools that work visually and hide the complexities of the underlying technologies. Rather than hiring expensive developers with Hadoop and other cloud technology experience, select data integration tools that function the same across all platforms as shown in Figure 4-3.

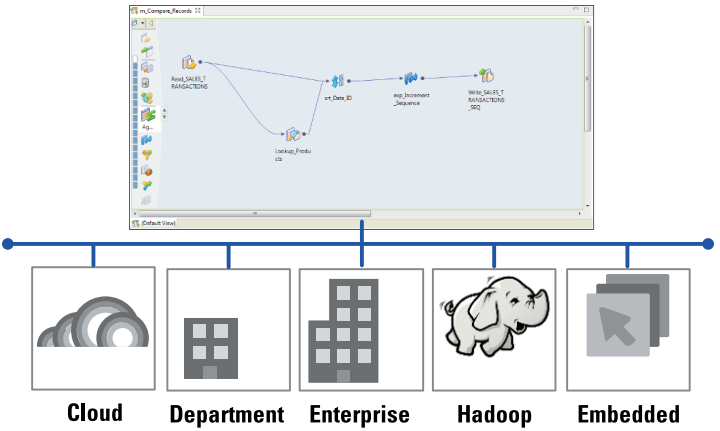

Figure 4-3: Choose tools that work across many platforms and technologies.

New technologies come down the pike all the time. If you have data integration tools that are designed to be future proof, you'll be able to support and use those new technologies without either hiring new developers or sending your existing developers for expensive training.

Hiring new developers with expensive skill sets like Hadoop can easily eat up your entire budget. A better alternative may be to choose data integration tools that hide the complexity of the underlying process so that your existing developers can do the job — no matter what platform you use.

So the bulk of the job is simply integration: how you get the data together, how you get it clean, and how you get it into a consistent format. This is where visual data integration tools can help because they make the data integration process easier and faster.

Scaling With Changing Needs

Good data doesn't happen by accident, it happens by design — by planning for good data. Consider, for example, how more and more sensors are being used to record what's happening with machines. A few years ago, you probably didn't use much machine sensor data. Now, however,

something like a jet engine has as many as 3,000 sensors. All these sensors generate huge amounts of data. In fact, a typical modern airliner can log anywhere from 100 gigabytes up to a half a terabyte on a flight. This is an awful lot of new data to process.

Unfortunately, a lot of that new data isn't very useful. For example, it's estimated that 98 percent of sensor data alerts recorded for that airline flight represents false positives. That means there was an alert to a potential engine problem that wasn't actually a problem. You need to have a system that can filter out the extraneous data so that you can focus on the meaningful data. Otherwise, there will be too much time spent checking engine problems that aren't actually problems and the cost of maintenance would go up, and as a result, so would the cost of an airline ticket! It's pretty clear that having someone go through half a terabyte of data manually for each airline flight wouldn't be reasonable, economical, or very efficient. Rather, you need an automated system that can quickly analyze the data and ignore the garbage. This is where a data integration system comes into play. Manual processes simply can't scale fast enough to meet these changing needs.

In addition to machine sensor technology, data integration tools can help you scale up as your needs embrace future technologies. For example, many organizations are moving to cloud-based applications and storage. Good data integration tools support this type of move in a transparent and seamless manner. Essentially, your business analysts and IT developers can use the same data integration tools for in-house projects, for projects deployed using Hadoop, and for cloud-based projects.

Another very important consideration is that with the right tools, you can grow from small projects to enterprise-level projects without having to move to something new. For example, a good tool would be appropriate for entry-level data integration projects, those smaller discrete projects, but have the capability to move up and scale to support projects that grow and become business critical. This consistency across a broad range of capabilities means you don't need to relearn new tools as you grow. Rather, you can leverage the knowledge you've gained without having to start from scratch.

A big factor in getting data integration tools that can scale to suit your future needs is remembering that the pace of technological innovation doesn't show any sign of slowing anytime soon. Consider that, for example, virtually everything has some connection to the Internet today, but just 20 years ago the Internet was primarily a private playground for college students and government researchers. Back then, even Bill Gates missed how important the Internet would become. Ten years ago, if you mentioned the cloud, pretty much everyone would have assumed that you were talking about the weather. Today, the Internet and the cloud have both become integral parts of everyday life. Who knows what exciting new technology is just around the corner? Fortunately, good data integration tools will enable you to take advantage of the next big thing — and innovate and stay ahead of your competition — without going back to square one.

Chapter 5

Top Ten Things to Look for in a Data Integration Tool

In This Chapter

- Making the right connections
- Encouraging agility
- Ensuring data quality

Buying a data integration solution, whatever the size of your project, shouldn't be undertaken lightly. It will affect how your business will and can function for years to come — especially if you buy the wrong product and need to revisit the whole process next year. So what do you need when you're out on the open market looking to buy a data integration tool?

In this chapter I discuss ten very specific features that you want to make sure your data integration solution includes.

The Right Connections

Data integration enables you to combine data from many different and rich sources to produce new business information you couldn't get from a single source.

Make sure your data integration tools are able to connect to any data source (both current and legacy) including RDBMS, NOSQL, mainframe, text, applications, and so on — and not just the data sources you consume today. It's this universal set of connections that makes it possible to bring all that data together.

Not all connectors are built the same. So just because a vendor can "tick the box" for having a connector, that doesn't mean that their connector is any good. This is especially true for connectors to applications like ERP or CRM. A good connector will hide the complexity of those applications. So don't just take a vendor's word for it — try it out or ask for references.

If a vendor provides a connector that translates the data poorly, you won't get the results you want.

The Right Data Types

Just as data integration draws data from many different sources, it also must be able to consume various and multiple data types, including structured, semistructured, and unstructured data sources in batch and real-time modes. You need a tool that is flexible enough to work with any type of data you encounter.

Rapid Development

You need data integration tools that can accelerate development with rapid prototyping and conversion of prototypes to production without recoding.

You can't afford to waste time hand-coding solutions or redoing your work by using a tool that can't turn your prototype into a real functioning system without having to start from scratch. In addition, vendors should have robust communities and online stores where you can access complimentary resources to kick-start your own project development.

Lean and Agile

You can't afford to create and execute projects using traditional, isolated development methods anymore. Your data integration tools need to support lean and agile integration processes that enable business and IT collaboration so that development happens quickly and interactively. Your competition certainly isn't going to wait for you to catch up!

Proactive Alerts

Mistakes and errors happen. Your data integration tools need the ability to monitor integration tasks and proactively alert administrators when there are exceptions so that corrections can be made quickly.

Waiting until someone complains about the results wastes time and delays the usefulness of your systems.

Automated Testing

Data integration projects are complex, and, like every development process, you need to test to make sure you've made the right assumptions, find the bugs in your code, and eliminate them. Most people still develop their data integration tests by hand. Look for tools that can automate development testing and verify that everything functions as planned.

Ability to Scale

Companies grow, and so do the sizes of their projects. You don't want to be locked into tools that are only appropriate for today's projects. Rather, you want tools that have the ability to scale, grow, and move projects from small departmental innovative exercises to large enterprise mission-critical environments, or vice versa.

Ideally, choose tools that provide this capability no matter how large your projects are today or might become, because scalable tools will help you leverage the skill sets developed on your smaller projects.

Cloud/On-Premise Hybrid Support

Who knows what technologies will come along tomorrow? You want tools that can support a hybrid business environment with a mix of on-premise, departmental, and cloud

development and execution options, as well as whatever comes along next. Such support is vital in protecting your investment. Things can change; new technologies come along and can dominate quickly. So you want to make sure your tool is flexible and future proof.

Data Profiling

With so many different sources of data involved, you need to have a means to make sure that your data is what you expect. It's important that your tools allow a level of data profiling so that you can verify the data going into and out of your system, and ensure that you'll end up with the desired results.

Data Quality

Finally, you need to remember that poor data quality can sink any project. It's absolutely essential that your data integration tools enable you to embed data quality into the data integration process. After all, you know what to expect at the output if garbage was the source!

Why you can't ignore data integration and how it gives your organization an edge

Are you trying to figure out how to use data from a variety of sources, such as legacy mainframes, modern relational systems, social media, and even sensor data? Are you still using hand-coding methods to integrate data? If so, this is the book for you! This book shows you how data integration tools can make it easy to use data whether you're a small business or a huge enterprise with either small or big data.

- ***Data integration 101** — why there's so much data today, what your business can do with it, and how data integration helps you use it*
- ***Data integration challenges** — the issues you face when trying to combine data from different sources*
- ***Data integration benefits** — how the right data integration tools can help you easily consolidate data sources and give your business the agility it needs to be competitive*

Open the book and find:

- **The difference between data, metadata, structured data, unstructured data, and more**
- **Insights into data integration pain points**
- **How your business can benefit from data integration**
- **A list of ten things to look for in a data integration solution**

Go to Dummies.com®
for videos, step-by-step examples, how-to articles, or to shop!

Brian Underdahl is the author of over 100 books and has been active in the tech industry for over 30 years helping businesses large and small.

FOR DUMMIES®
A Wiley Brand

978-1-118-89658-7
Not for resale