Chapter 13

Where's the Data?

In This Chapter

- Ridding yourself of data silos
- Making information into a service
- Scouting out the metadata repository
- Making sure you can trust your data

If you've decided to dive into the world of service oriented architecture — thereby reaping the benefits of sharing critical business services across the organization — you need to consider how to maximize the trust and confidence you have in your company's data. You may decide to begin your SOA journey by eliminating some of the redundancies in businesses technology systems and software, but you can't stop there. The next step is to ensure that all the company's data is both *consistent* and *accurate*. This chapter shows you how to achieve both those goals.

When Good Data Goes Bad

Service oriented architecture represents a new way of thinking about everything in a company's IT structure, including how one thinks about data. It begins with the goal of achieving consistency between data sources. In order to achieve data consistency, you begin by separating your data from its tight dependency on the business applications that created it and update it.

Data is one of the organization's most precious assets, but these critical data stores are typically segregated by business function in *data silos*. Traditionally, business data has been managed in a way that tightly associates specific data definitions to specific business applications, such as finance, human resources, or operations. Take, for example, a sales force automation application that manages all your sales force data. This is likely to be a *silo* of data because the data structures will have been designed to satisfy the particular needs of the sales force automation application. Your customer relationship management system will probably sit over another silo of data — one that overlaps the sales

force automation silo. Some organizations might have hundreds of such data silos scattered all over their enterprise. This may sound like an exaggeration, but it is not. If you have hundreds of applications, you likely have hundreds of data silos.

The problem is that an organization's data resources were not designed for global use by all applications. They were designed to suit one specific business application or, at best, several applications. The data was designed for a specific context. For example, the sales order processing system might record a person's name, address, date of birth and sex, but not their marital status, or what they do for a living. This is fine for taking orders and delivering goods, but not for the customer relationship management system, which needs much richer information about the customer.

When separate systems gather their own data, simple errors in entering data make it difficult, and sometimes impossible, to aggregate the data that has been collected about a customer (or any other entity). No matter how effectively data is gathered, corruption creeps in. The rate of error can sometimes be as high as 15 percent and it is never zero.

The siloed approach to working with data may provide great information to a particular business unit, but it creates some startling inconsistencies in data when viewed at an enterprise level. This happens because data is often defined to fit the precise view of a single business unit. So, one department thinks of the customer as the manager of a department that procured a service from the company. Another department defines a customer as the company itself. A third department defines a customer as the local office of that same company. How can you trust the information your business uses to make strategic decisions if poor-quality data keeps you from having a complete view of your customers or products? For example, you might underestimate the importance of one of your key customers if that customer makes purchases from several business subsidiaries, using slight variations of the company name and your system doesn't recognize them as a single purchaser.

Other inconsistencies in data are based on *semantic* differences. *Semantics* here refers to the rules that govern how one talks about data, just like the semantics of English govern how one conveys meaning when speaking English. A semantics of data is used to ensure that everyone in the business has a common understanding of the business information and rules — that everyone speaks the same "language," as it were. For example, one business unit might consider the word *customer* as referring to the local office of a particular company, whereas another unit might use *customer* to mean the entire corporate entity. These semantic differences in the use of basic business terms like *customer, partner, department,* and the like lead many organizations to devote significant resources to interpreting and reconciling differences in reports from various divisions or subsidiaries. (For a more technical take on data semantics, check out the "Data semantics" sidebar.)

Data semantics

Data semantics is the *meaning* of data. It's a meaning that goes deeper than definitions to include an understanding of the data in context with business products, people, or events. *Semantic interoperability* is an architectural quality that measures how people and technology understand data and how this level of understanding impacts the exchange of information. You have to understand data in context in order to make accurate and appropriate decisions.

Semantic interoperability means that both business managers (humans) and software applications (machines) can understand the subtler meaning in data. Humans and machines need to know something about how the data is defined or calculated and where it came from in order to determine whether this is the right data. Humans from different business entities need to agree on certain rules, definitions, and policies for critical data. However, there are times when a human may be able to make manual adjustments to account for slight nuances in meaning, but a machine cannot.

Much of the transfer of data in a service oriented architecture is done from machine to machine, without human intervention, so that you need the highest level of semantic accuracy or interoperability to really achieve trusted information. In the following example, the terms "balance" and "remainder" may mean the same thing, but the machine requires specific instructions to account for the semantic difference in the two terms:

- A billing application needs a customer balance. The application calls the data it needs *balance.*
- An accounting application supplies a customer balance. This application calls the data it supplies *remainder.*
- In order for the accounting application to automatically supply the billing application with the correct customer balance, an adjustment or mapping must be made between *balance* and *remainder.*

Reconciling inconsistent data can take a lot of time — and can result in missed business opportunities. In addition, imagine the confusion that often occurs when one organization buys or merges with another and tries to integrate the data. It's imperative to determine whether the soon-to-be-merged companies in fact share a certain subset of customers so that the customer can be appropriately served.

One of the main objectives of service oriented architecture is to make sense out of business chaos. This includes providing accurate information about the business — in the right form and at the right time — to everyone involved in the business. One critical step for making this happen is to ensure that each component of data can be used independently from its current implementation. With service oriented architecture, you need to begin to think of data as a reusable resource. We call this new concept *information as a service*.

Dastardly Data Silos

A database (or other data store) is called a *data silo* if it is tightly dependent on (and designed to manage data for) a specific application or a specific region or functional area of the company. The tight dependency between a siloed data store and the application makes it almost impossible to get a complete and consistent view of data across the enterprise.

Silos of data are a natural extension of how business applications have been designed for decades. For example, each department in a very large (fictitious) commercial bank (which we're calling Big Global Bank) has its own applications — such as the personal account system, the human resources system, and the mortgage origination system. Each of these applications uses and creates lots of data. Likewise, partners have their own sets of data about the products they sell and the customers they serve. In addition to all this data, Big Global Bank has recently acquired several other banks and must contend with overlapping sets of systems and related data, hampering business interaction and decision making. The mainframe systems that store the customer and product data from different departments and subsidiary banks cannot easily connect with each other or with externally located data stores belonging to Big Global Bank's business partners. This siloed approach to storing and managing data inhibits the flow of critical information at Big Global Bank.

Is it information, or is it data?

Often, the terms *data* and *information* are used interchangeably, and in general that's okay. However, people who spend a lot of time with data usually give these two terms slightly different meanings:

- *Data* generally refers to facts, like temperature and humidity.
- *Information*, on the other hand, is the collection of these facts in a specific context from which conclusions can be drawn.

So, if the temperature is high and the humidity is high little can be deduced, but if the temperature is high and the humidity is high in a given place for a given length of time, the conclusion could be that it will be uncomfortable for anyone in that place during that time. The facts — the actual temperature and humidity statistics — are the data. The conclusion that's drawn (uncomfortable weather) is the information.

Businesses use the term *data* to refer to words and numbers that represent what a business needs to know. For example, a data element like "Peter Jones" is an instance of a customer name, or "24601" is an instance of a style number. These pieces of data need to be placed in context, along with other pieces of data, in order to be used for analysis and decision making. Because companies use data to derive information to make decisions, and this data must be qualified and consistent, we think the term *information* sounds more appropriate. Hence, we use the phrase *information as a service* rather than *data as a service.*

Individual departments and bank subsidiaries have each defined data items for their own purposes — not taking into account the rest of the company's needs. So, when Global Bank needs to bring together data from one department with a dozen other departments in order to make new business decisions, they have problems. Definitions of everything from what a customer is to the names of products are different. The same customer may have different types of accounts, and Global Bank may not be able to associate all the accounts with this customer. Therefore, management simply cannot trust the data to be consistent and accurate when viewed at an enterprise level. When organizations like Global Bank discover this problem, they typically come up with ways to work around the problem on a case-by-case basis. Not only is this time consuming, but each situation also requires development teams to start from scratch — there is no reuse of expensive development efforts. They really need SOA.

Trust Me

Businesses like our fictitious Big Global Bank have become increasingly complex, resulting in many situations in which trust in data, and therefore trust among entities reliant on the data, has been compromised. Company mergers and acquisitions, electronic commerce, and economic globalization have all contributed to the increased level of complexity in organizational data. Government regulations like Sarbanes-Oxley and Basel II require organizations to make significant changes to the way data is managed to ensure accuracy, reliability, and auditability. But, in addition to responding to regulations, it makes good business sense for organizations to ensure the integrity and security of corporate data assets. A higher level of trust among companies, their partners, and their customers leads to more efficient business because transactions can be done more quickly and cost effectively.

In order to make data more reliable, consistent, and trusted, enterprises link data sources between departments or regions of their organization by using various data integration processes. Service oriented architecture is changing both the philosophy and the architectural framework for deploying the data integration software tools that manage the integration process. Some of the key processes required to bring the data together in a meaningful way include locating and accessing data from a data store (*data extraction*), changing the structure or format of the data so it can be used by the business application (*data transformation*), and sending the data to the business application (*data load*). Software programs that automate these processes are often grouped together as *Extract-Transform-Load* (ETL) tools.

Integrating data across business entities was previously done by creating a system of tight linkages or connections that were fixed in place and could not be easily changed. In many cases, they didn't (and still don't) provide for a two-way flow of information. Implementing a SOA approach enables the business to access, manipulate, and share data across the organization in a repeatable and consistent way. This approach provides the business with more useful information to help make sound business decisions. For example, the business knows more about John Parker Jones as a customer after the purchases of J. Jones, Mr. Jones, J.P. Jones, and John Parker Jones are aggregated. Service oriented architecture ensures that this aggregation can be done quickly and efficiently, and that the system is flexible enough to adapt to changes required by the addition of a new product line or subsidiary.

Service oriented architecture enables the business to put the priorities of the business first instead of holding the information hostage to the restrictions based on the structure of the IT system. The ETL and software tools for other data integration processes (data cleansing, profiling, data transformation, and auditing, for example) all work on different aspects of the data to ensure that the data will be deemed trustworthy. The following sections show how that's done.

Data profiling

Data profiling tools help you understand the content and structure of your data by first collecting the necessary information on the characteristics of the data in a database or other data store — a crucial first step when it comes to turning the data into a more trusted form. The tools then analyze the data to identify errors and inconsistencies so they can make the necessary adjustments and corrections. The tools check for acceptable values, patterns, and ranges and help identify overlapping data. The data profiling process, for example, checks to see if the data is expected to be alphabetical or numeric. The tools also check for dependencies or to see how these data relate to data from other databases.

Data quality

High-quality data is essential if a company is to make sound business decisions. The quality of data refers to characteristics about the data, such as consistency, accuracy, reliability, completeness, timeliness, reasonableness, and validity. Data-quality software makes sure that data elements are represented in the same way across different data stores or systems in order to increase the consistency of the data.

For example, one data store may use two lines for a customer's address, and another data store may use only one line. This difference in the way the data is represented can result in inaccurate information about customers, such as one customer being identified as two different customers. A corporation might use dozens of variations of the company name when they buy products. Data-quality software can be used to identify all the variations of the company name in your different data stores and ensure that you know everything that a particular customer purchases from your business. This process is called *providing a single view of customer or product.* Data-quality software matches up data across different systems and cleans up or removes redundant data. The data-quality process provides the business with information that is easier to use, interpret, and understand.

Data transformation

Data transformation is the process of changing the format of data so it can be used by different applications. This may mean a change from the format the data is stored in into the format needed by the application that will use the data. This process also includes *mapping* instructions so that applications are told how to get the data they need to process.

The process of data transformation is made far more complex by the staggering growth in the amount of unstructured data. A business application, such as a customer relationship management or sales management system, typically has specific requirements for how the data it needs should be stored. These data are likely to be *structured* in the organized rows and columns of a relational database. Data is *semi-structured* or *unstructured* if it doesn't follow these very rigid format requirements. (The information contained in an e-mail message is considered *unstructured*, for example.)

Some of a company's most important information is in unstructured and semi-structured forms, including things as ubiquitous as documents, e-mail messages, customer support interactions, transactions, and information coming from packaged applications like ERP and CRM. Many data transformation tools don't handle unstructured data very well, and if you need to incorporate the information into your integration strategy, a significant amount of manual coding may be involved.

Data governance and auditing

The primary role of establishing SOA data governance and auditing services is to enable and manage the enforcement of business and security policy as it is applied to data. The need for this technology is particularly urgent because

of regulations like Sarbanes-Oxley. Data governance provides a level of accountability that is equally critical for business customers, suppliers, partners, auditors, shareholders, and regulatory agencies. This technology includes security services such as data encryption, digital certificate management, and authentication. It also includes processes for managing user privileges regarding data access control that determine who can see data as well as who can change the data.

As information becomes more loosely coupled (independent of any specific application), data auditing ensures that an organization can manage and adhere to requirements imposed by regulatory agencies and that access to data is kept confidential. It also helps the enterprise answer questions like these:

- Who has access to sensitive data?
- When was it accessed and by whom?
- How can I track data that may have been deleted?

Providing Information As a Service

When organizations begin to apply SOA principles to managing their data assets, they move from fixing problems on the fly to delivering information as a service. Information as a service is an architectural approach that loosens the tight connections between data and applications so that data can be controlled and shared across the enterprise. This approach allows businesses to reach a consistent view of enterprise-wide information that has previously been very hard to achieve.

By applying the principles of service oriented architecture, such as loose coupling of data to applications, businesses can increase the consistency and accuracy of their data. And they can do this in an efficient, cost-effective way without actually moving or redesigning the data stores that exist in their business today. In essence, you achieve the goal of getting at the data you need without performing major surgery. Although creating one massive centralized data store would help control data management and synchronize data definitions, it would be impossible to manage.

Data control

Control of data is a controversial issue in many companies. If you are responsible for a departmental budget or are responsible for meeting specific business objectives, you really don't want someone from another department

manipulating your data. Businesses need to find the balance that lets managers retain enough control over the data that matters most to their department and also allows for a single, consistent view of customers or products at the enterprise level.

What the business needs is data that can be trusted and understood at all levels of the organization. The information as a service approach is designed to ensure that business services are able to use and deliver the data they need in a trusted, controlled, consistent, and flexible way across the enterprise regardless of the requirements specific to individual systems or applications.

As with so much of the SOA approach, the ability to provide information as a service is a work in progress. While no specific method will work for all enterprises, in order for information to be delivered as a service, the data must meet the following three requirements:

- **Consistent data definitions:** The meaning of data needs to be unambiguous so it can be interpreted and processed appropriately both by businesspeople and by machines.
- **Ensured quality of data:** Businesses should use whatever tools they need to ensure that data from many different sources can be trusted to be accurate and consistent no matter how the data ends up being used.
- **Data independence:** If the data is loosely coupled from its original sources, those data elements can be more easily brought together in different ways to meet many different business needs.

The following sections address each of these concerns in turn.

Consistent data and the metadata repository

To provide information as a service to everyone in the business — from sales to operations to finance and senior management — all the data the businesspeople need must be treated consistently across the enterprise. Consistent definitions and rules for data must be based on the way the business *as a whole* needs to understand sales, customers, products, and profit. If you have a right to view, change, and report on data, you should be able to get the quality data you need when you need it.

Information delivered as a service has been effectively certified by the enterprise as trusted data. This means you can trust that you and your counterparts across the enterprise are basing decisions on data that is secure, clean,

and structured correctly. Everyone in the business is working with consistent rules about how the data is structured, accessed, and used. This common understanding of the data must extend across business units and regions to include information provided to partners and customers.

The definitions, mappings, and other characteristics used to describe how to find, access, and use the company's data are called *metadata*. Business services need to be able to access metadata in order to consume and deliver the data they need. Metadata is stored in the *metadata repository* — a container of consistent definitions of business data and rules for mapping data to their actual physical location in the system. This repository resides in a technical layer between the actual data stores and the business services.

The metadata repository is often referred to as a *metadata layer* because of its position in the information infrastructure. A more complete technical term for the metadata repository is a *metadata abstraction layer.* This is because it includes the rules, definitions, and mapping instructions about the data that are either replicated or separated from the data stores. The process of abstracting the data rules and definitions adds flexibility to the data infrastructure, which provides programmers with a way to loosen the tight connections between data stores and specific business applications.

The purpose of the metadata repository is to help you bring together all the components of your business in an orderly way without requiring you to replace your existing data stores. The abstraction of rules, definitions, and other instructions from the data stores provides your business with a way to achieve consistent data while still maintaining your extensive investments in data management assets. By abstracting data from the context in which it is held and used, you are better able to work with it in a variety of situations. The metadata repository ensures that the data is of the right structure and quality before it's consumed by a business service. The metadata repository also ensures that data from different sources can be linked together correctly. The semantics and rules that apply to all your company's data can be organized, tracked, and managed through the metadata repository. For more on metadata repositories see Chapter 15.

Know Your Data

You should be as careful and curious about your business data as you are when meeting a new person at a party. The data definitions, data lineage, and other characteristics about the data in the metadata repository provide details about your data in the way that you might put together background details about a person you have just met. For example, last week, Elizabeth

met Bob at a party. Initially, Elizabeth wasn't sure whether she wanted to spend the time to get to know Bob. She had no context in which to judge whether he was honest or thoughtful and would be a good person to get to know. Elizabeth began to collect data about Bob in a very simple way. She asked him a lot of questions. She found out his age, his hometown, where he lives now, and where he works. These are some of the descriptive characteristics about Bob.

She also found out that he went to college with one of her close friends, that he plays basketball with someone who works in her office, and that one of his co-workers is married to her cousin. This is the *lineage* — history about where Bob has been and some of the connections between Bob and other people. Now she knows enough to conclude that she would like to know Bob well, and she knows how to locate him to find out even more.

Think about data the same way. The metadata repository allows you to ask and get answers to the following types of questions:

- How is the data structured?
- What does the data look like?
- What rules apply to the data?
- How is the data used?
- How do you find the data?
- What does the data mean?
- Where does the data come from?
- Who has the rights to access or change the data?
- What is the context for the data?
- What impact will changing the data definitions create?

The answers to these questions provide context for data and enable applications to use data properly. The lineage or background history provided on the data answers the type of questions that Elizabeth wanted to know about Bob at the party. You need to know where the data has been and how it has been accessed, changed, and used to be able to achieve data consistency. (Safe dating, safe data.) A metadata repository helps you to understand the impact of changes to data. You need to be able to follow the history of changes to data and make the connections to the business services that use the data. This requires a link between the SOA registry and the metadata repository (detailed in Chapters 8 and 15).

Data services

Data services are all the technical processes that *qualify* the business data to ensure that it is trustworthy. These processes include the data integration technologies that we talk about earlier in this chapter (data profiling, data quality, data transformation, and data governance and auditing). Although businesses have successfully used software tools for data integration without applying SOA principles, using the data services approach gives you a more comprehensive — and more business-focused — view of your data. Data services bring all the modular data integration components together to deliver trusted information to the enterprise consistently and as needed. In the past, a data profiling tool or data-quality tool may have been applied to specific data stores on a case-by-case basis. The data services approach applies all these technical processes as required to the data requested by the businessperson. It is the automatic and integrated nature of this approach that ensures that all the data the business needs is accessible, accurate, consistent, timely, and complete.

The metadata repository is a critical part of the infrastructure that all data services need in order to work effectively. If sales, finance, and operations all need to get data about customer John Parker Jones for different business processes, the data services for the business ensure that everyone is working with consistent and accurate information. The data profiling and data-quality services, for example, look to the metadata repository to find and correct the different variations of the customer's name. The metadata repository provides data on the linkages between John Parker Jones and his various accounts. It also provides the security and access level so you can get this information only if you are entitled to do so. The metadata repository provides the data service with all the contextual information about the data to provide a complete picture of the customer.

Loose coupling

The third key requirement for delivering information as a service is to ensure that the data is available as a reusable resource. Loosening the dependencies between data and the applications where the data originated provides the infrastructure flexibility that supports reusability. Using a federated approach ensures that the data can stay in its original location. (More about federation in Chapter 5.) Federating data sources provides consistent rules and definitions so that data from various types of data stores can all work together. This means that the business can avoid changing the data and its location but can still combine data from a variety of sources depending on the requirements of a business application.

Service oriented architecture has the potential to allow businesses to grab the elusive brass ring of business achievement through flexibility and innovation. Achieving trust in data needs to be an integral part of any business's SOA. Even the most efficiently created, easy-to-use business services for payment, invoicing, or other business processes will not provide long-term value to your business if the data is misunderstood or of poor quality. Business services exist only to read, monitor, calculate, analyze, report, and otherwise manage the business data.

Implementing information as a service leads to increased business flexibility, business trust in data, and reduced costs. The integrity of the data is strengthened because when a business service receives or consumes data that is delivered as a service, the data has been effectively certified by the enterprise as trusted data. The ultimate goal of this approach is to provide a seamless way for the business user to access data that is both trusted *and* consistent with company rules and polices.