Realtime publishers

Tips and Tricks Guide™ To

Windows Administration

Don Jones and Dan Sullivan

Copyright Statement

Tip, Trick, Technique 20: Understanding the Sources of Growing Volumes of Data

Systems administrators are the last people who need to be told the volumes of data are growing at staggering rates and will probably continue on the same trajectory. As they are the ones responsible for keeping up with this growth, it's worth taking a look at where all this data is coming from. After all, computers are nothing new; they have been running business applications since the 1950s. What is it about today's use of information technology that is generating such high growth rates? The answer is that there is no single culprit; rather a confluence of technical and organizational issues drives this growth. Some of the most important drivers contributing to this phenomenon are:

- Data-intensive applications
- The growing importance of unstructured data
- Compliance

Systems administrators will have influence on some of these drivers, such as new applications, while other areas, such as compliance, have more rigid requirements that may not leave much room for optimization.

Data-Intensive Applications

The days of business running a limited number of back office applications are over. Of course, pretty much any business will be running financial packages that track revenues and expenditures along with whatever form of sales they may have—that is, products or services. Any but the smallest will likely have some kind of customer relationship management, human resources, and inventory management package as well. These will often generate a fairly constant rate of data or grow in proportion to the business activity. These kinds of applications do not generate significant growth in data—that comes from other data-intensive applications.

Data-intensive applications come in many forms, including those that capture detailed interactions with customers, instrumentation, data analysis applications, and content management systems. We will consider each of these in turn.

Customer Interaction Data

More and more interactions with customers are being tracked. In the past, we could track customer interactions at a point of sale. For example, when we shop at a national retailer, the business captures their first pieces of data about us at the point of sale system. At that point, we are done shopping, we have a full cart, and we are ready to pay. The retailer can capture information about:

- Items purchased
- Type of payment used
- Amount of sale
- If provided, additional tracking data such as postal code or phone number

That is a relatively small amount of data compared with what could be gathered through an online catalog. If we were to shop at the same retailer's Web site, the list of data elements that could be tracked would grow to include:

- Items purchased
- Type of payment used
- Amount of sale
- List of products viewed
- Types of pages viewed, such as product description, customer reviews, ratings, and so on
- Contents of abandoned carts
- Navigation paths through the Web site
- Search terms entered
- Time spent at Web site
- Additional demographic data provided by third-party tracking and Web analytics services

If we were to multiply the size of the new data by the additional number of customers that come to Web sites over retail stores, we would start to get a sense of how much additional data can be generated.

Tracking customer interaction in detail is only useful if we do something with that data, and that is where business intelligence and analytics come in.

Business Intelligence and Analytics

Business intelligence and analytics are applications designed for internal use. Customers reviewing product offerings, checking the status of orders, or making purchases work with online transaction processing systems. These are optimized for rapid response to high volumes of concurrent users. Business intelligence systems are a different breed of application.

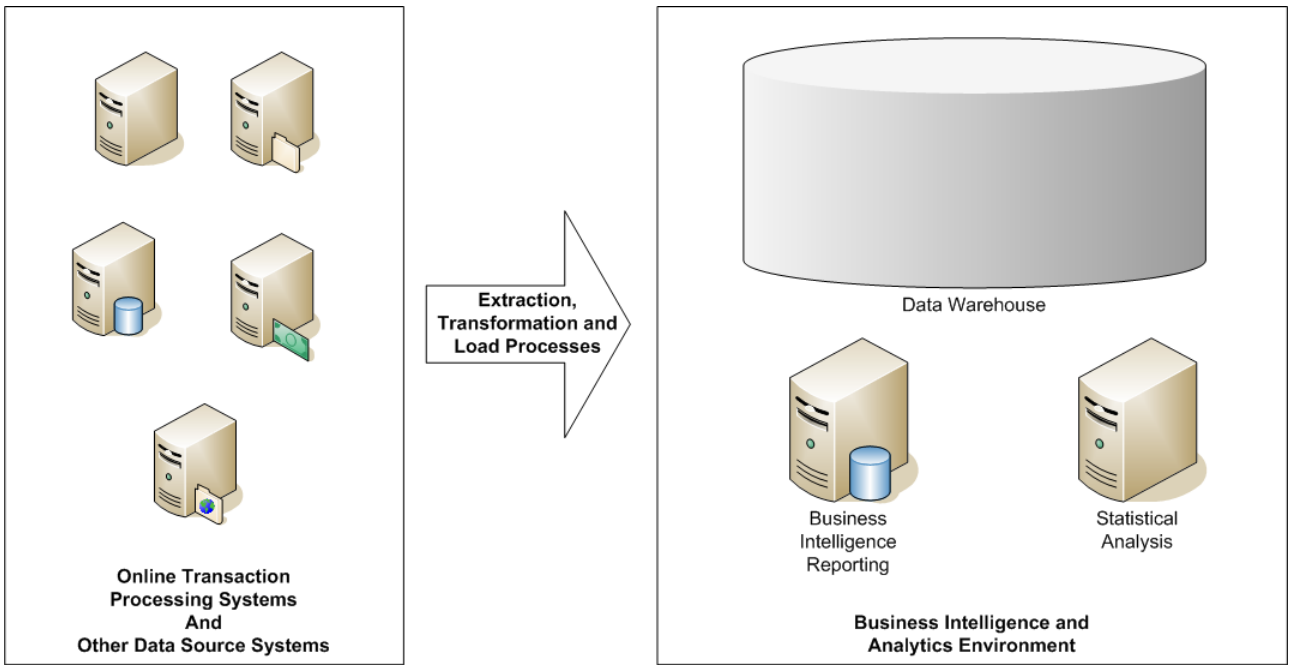

Figure 39: Business intelligence environments duplicate data found in transaction processing systems.

Business intelligence systems are designed for managers, analysts, and others who need to delve into data and make comparisons across time, products lines, sales regions, and so on. For example, if you want to know how sales in the Southeast sales region are doing this quarter compared with the same time last quarter, you would use a business intelligence system. Similarly, if you wanted to find branch offices with the poorest revenue to expense ratio, you would use a business intelligence system. The problem from a data storage perspective is that the business intelligence systems duplicate the data found in transaction processing systems using operations known as extraction, transformation, and load (ETL) processes.

Why duplicate data? After all, if is already in the transaction processing system, why not use that? A complete answer is beyond the scope of this topic, but the quick answer to that question is:

- Online transaction processing systems are designed to rapidly retrieve, update, and delete individual records; they are not designed to rapidly return aggregate data, such as the sum of all sales in a given quarter.
- Online transaction processing systems may not keep sufficient history to answer business intelligence questions; keeping large volumes of historical data could slow the response time of interactive queries.
- Business intelligence systems use data warehouses that are designed to store large volumes of historical data organized in a form that allows for rapid querying of aggregate data. For example, a business analyst can quickly move from looking at annual summaries to quarterly to monthly to weekly data.

- Many business intelligence reporting tools are designed to take advantage of the data models used in data warehouses but not online transaction processing systems.
- Database schemas of online transaction processing are often difficult for non-database professionals to navigate; data warehouse schemas are simpler and more intuitive allowing for more *ad hoc* reporting and exploratory analysis.

In addition to traditional business intelligence reporting, there is a growing use of statistical analysis and data mining techniques known collectively as business analytics. These applications are consumers of data—and the more data, the better in some cases. Like data warehousing, they require data in a particular format that does not usually correspond to the way online transaction processing systems structure data. As a result, data is copied from source systems and reformatted into a format more amenable to analysis.

Business intelligence and analytics are formalized processes that duplicate data; information practices contribute to data duplication as well. Microsoft Excel and other spreadsheets are something of a double edge sword for data warehouse designers. On the one hand, it is convenient to have the option of exporting data from the data warehouse into a spreadsheet so that users can take advantage of the additional features of the application. On the other hand, users take advantage of this. Data that was originally in an online transaction processing system is now in a data warehouse and some unknown number of spreadsheets in user directories. Of course, some of these will be emailed to multiple recipients who may potentially save their own version.

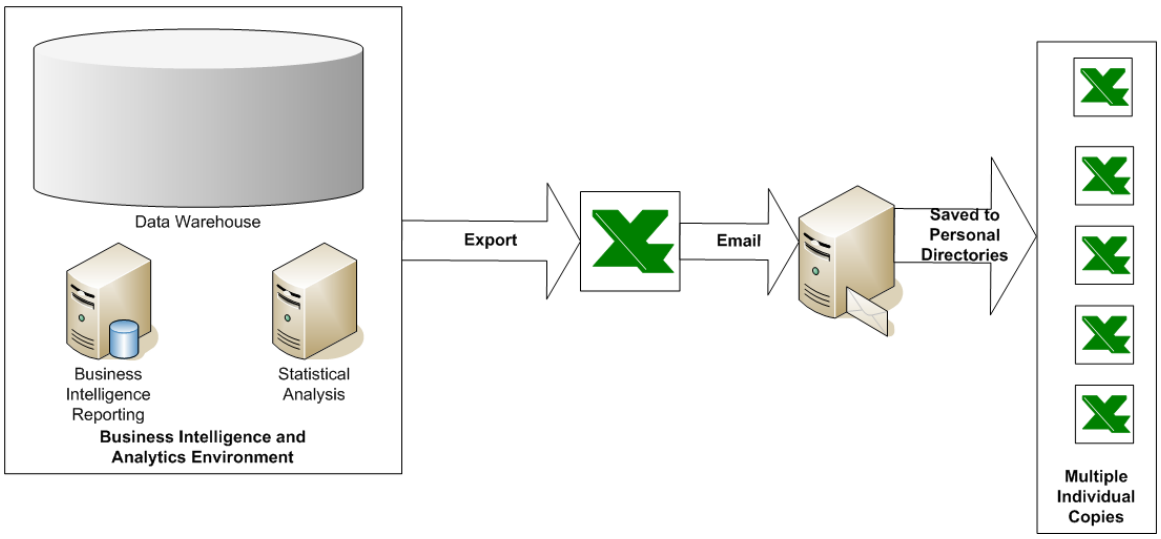

Figure 40: Useful data warehousing features, such as the ability to export to spreadsheets, can quickly become a means to duplicate data many times over.

Data-intensive applications are significant drivers behind the growth in data volumes due to both improved methods for collecting data and the need to duplicate data to meet multiple needs. These examples focus on what is typically known as structured data. They tell only part of the story.

Growing Importance of Unstructured Data

Unstructured data is data that does not fit into well-defined data structures, such as database tables or spreadsheets. Free-form text, audio, and video are all examples of unstructured data. Unstructured data is ubiquitous in today's organizations with sources including:

- Email messages
- Instant messaging, texting, and other short communications
- Word processing documents and presentations
- Web pages, including content management systems, Wikis, and so on

In addition to the fact that many of us generate unstructured data on a daily basis, we are constantly duplicating it. When we reply to an email message and embedded the original text in our response, we create more unstructured data. When we save attachments as personal copies, we add to the growing volume of unstructured data. The Web makes it easy to bring in additional data from outside the organization as well. Find an especially useful article? You might save a local copy so that you do not have to search again or risk having the site remove the content. The rate at which we create and duplicate unstructured data is yet another driver behind the growth in data volumes.

The "Unstructured" Misnomer

Calling text unstructured is something of a misnomer. Linguists study the structure of natural languages and can describe their complex structures in detail. If anything, natural language is highly structured. For most IT needs, though, we can safely ignore the structure of natural language. Instead, we treat the entire text as a single object and do not delve into the structure within.

To appreciate the importance of unstructured data, we only have to consider how our organizations would function without email, shared directories, or SharePoint servers. Applications such as these can be just as business critical as application servers and databases. Both structured and unstructured data can be subject to yet another factor in data growth: compliance.

Compliance and Data Generation

Regulatory compliance and other legal drivers, such as e-discovery, are shaping the way we generate, store, and archive data. Regulations such as the Sarbanes-Oxley Act (SOX), the Health Insurance Portability and Accountability Act (HIPAA), and others define certain requirements with regard to how businesses report on their financial status and protect customer privacy. A common aspect of many regulations requires businesses to not only comply with the regulation but also be able to prove that they are in compliance. To do so, they must document, in detail, with policies, procedures, documentation and audit trails.

At first glimpse, this documentation requirement may sound simple and not terribly data intensive; however, in many cases, the level of detail required can result in significant data generation. For example, consider the type of events that may have to be logged:

- User logins and logouts
- Changes to access controls on files
- Changes to privileges on database tables
- Granting of privileges to user accounts
- Addition of users to administrator groups
- Updates to sensitive information, such as corporate financials
- Transmission of protected information, such as patient records

A related driver is known as e-discovery. During legal proceedings, a company may be required to produce electronic documents, such as emails and word processing documents, relevant to the case. In the past, companies that have been unable to produce those documents have been subject to severe fines. In 2008, Qualcomm was fined $8.5 million dollars for e-discovery violations (Source: Kristine L. Roberts, "Qualcomm Fined for 'Monumental' E-Discovery Violations—Possible Sanctions Against Counsel Remain Pending" at http://www.abanet.org/litigation/litigationnews/2008/may/0508_article_qualcomm.html). It does not take many such examples to motivate businesses to retain and catalog electronic communications.

Data-intensive applications, both those that are designed to capture and generate data as well as those designed to analyze it, are significant contributors to data volume growth. Unstructured data is easily created and duplicated, further contributing to that growth. If these factors were not enough, compliance and e-discovery concerns are prompting businesses to preserve data and to maintain it longer than they might otherwise.

Tip, Trick, Technique 21: Understanding Systems Administrator's Responsibilities for Growing Volumes of Data

Who is responsible for managing the growing volumes of data? It is a shared responsibility of the business owners, who are responsible for setting policies and procedures governing the generation, use, and destruction of data; application managers, who are responsible for maintaining their applications and ensuring they function as required; and systems administrators. In many ways, it is the systems administrator who is on the front line of managing data in an organization.

Some of the key responsibilities of systems administrators, with respect to growing volumes of data, is keeping up with

- Backup and recovery
- Security
- Infrastructure management
- Planning and architecture

These responsibilities range from the mundane but essential, such as setting and verifying access controls on files, to making recommendations on the use of emerging technologies, such as cloud computing, to accommodate even more data.

Backup and Recovery

Backup and recovery procedures are standard operating tasks for systems administrators, but these tasks become more difficult with growing volumes of data. In particular, systems administrators have to grapple with:

- How to perform backups in the time windows allotted for them
- How to restore fast enough to meet recovery time objectives (RTOs)
- How to capture changed data frequently enough to meet recovery point objectives (RPOs)
- How to detect trends in data growth before current procedure breakdown because of insufficient time or storage space to perform necessary operations

Part of the solution is to understand what has to be backed up and how frequently; a related part is to understand how long different types of data have to be kept. Information life cycle management practices can help here; they are discussed in more detail in Tip, Trick, Technique 22.

Most of the difficulties previously listed can be at least mitigated with deduplication technologies. Backup vendors are incorporating deduplication technologies in their software packages to combat the problem of growing data volumes. The basic idea behind deduplication is that data is often duplicated and rather than storing multiple copies of identical data blocks, a backup can be constructed using a single copy of such data blocks and references or pointers back to that copy.

Security

Security concerns can be distilled to three words: confidentiality, integrity and availability. In the case of confidentiality, security comes down to the question: How do we ensure that private and sensitive data is accessed only by those with legitimate reasons to have it? To maintain the integrity of data, we have to ensure that only processes that follow established protocols can change data. This translates into a question of How do we ensure that no one tampers with data, for example, making an unauthorized change to a revenue statement or delete entries from a system log file? Availability is a bit different from the other two security fundamentals. In one sense, systems administration is all about ensuring availability. In a security context, though, it has more to do with preventing adverse events from disrupting services (think Denial of Service—DoS—attacks) but also includes recovering from adverse events (recovering from backups made prior to a malware infection).

The challenge of growing volumes of data has not introduced security responsibilities; it has only made them more difficult. Let's just consider some of the ways that increasing volumes of data can tax policies and procedures.

Challenges to Maintaining Confidentiality and Integrity

One of the ways we protect confidentiality and integrity of data is with the use of access controls. These consist of three parts:

- Identity management elements that are used to authenticate individuals or processes and assign to them privileges
- Resources that are assigned access controls, such as devices, directories, and files
- Privileges, which are rights granted to an identity to perform some operation

Figure 41 shows the basic security dialog on the Windows file system.

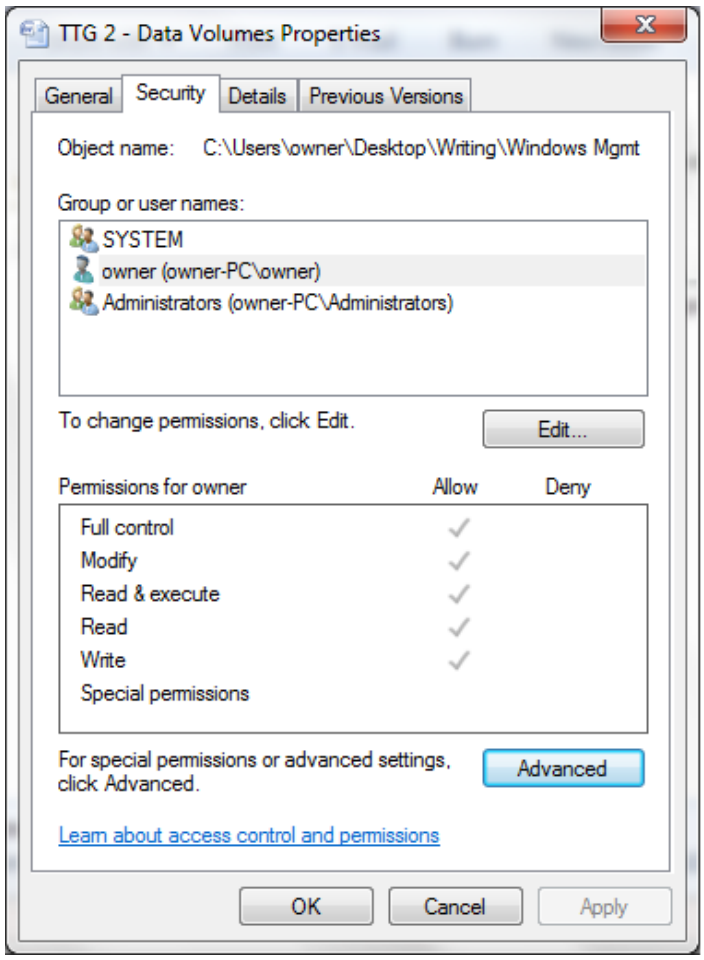

Figure 41: Windows supports several types of privileges on files, which are used to preserve confidentiality and integrity of file data.

With growing volumes of data, these simple building blocks for protecting confidentiality and integrity become more difficult to apply, track, and monitor. Some of the reasons for this include:

- Data is stored in different forms, such as files; content management systems, such as SharePoint; relational databases, such as SQL Server; and specialized applications that may use different sets of access controls.
- Individuals may have multiple identities, such as one user account for a domain and another shared identity for working with a database application. This problem can arise when application administrators decide it is easier to have a small number of shared access accounts rather than maintain individual accounts for each user.
- Applications and servers, such as databases and content management systems may have different default configurations and access settings. As departments and individual users start to adapt these applications to help manage growing volumes of data, they may not understand the security implications of default configurations, which could leave some data vulnerable to viewing or tampering.

Ironically, with growing volumes of data come growing challenges to protecting the data in the first place.

Challenges to Maintaining Availability

With more data comes more servers, more storage, and potentially more applications and this ultimately leads to more potential points of failure. Recovery management practices, such as backups and disaster recovery planning, can mitigate the risk of losing data to a hardware failure, human error, natural disaster, and in some cases to malicious attack or other security breach. Another area that should be considered is application vulnerabilities (this applies to protecting confidentiality and integrity as well).

As noted earlier, one of the reasons for growing data volumes is new applications, both customer-facing applications and internally-oriented systems such as decision support applications. Each of these new applications increases the options available to malicious attackers looking to either steal private and confidential data or disrupt services. This is called "increasing the attack surface" in security parlance and it essentially means the more applications and the more complexity, the more opportunity for vulnerabilities.

One thing we should understand is that attackers do not need detailed knowledge of our applications (although that helps). Automated vulnerability scanning tools can be used to detect vulnerabilities to well-established attack methods such as cross-site scripting attacks, which exploit weaknesses in Web applications to compromise them. Another class of tools, known as fuzzers, probe application programming interfaces looking for exploitable errors. Fuzzers, for example, can generate random input of varying sizes to detect unhandled errors in applications accepting user input.

How Real Is the Risk to Cyber Security?

It's real. The days of cyber-vandalism look benign in retrospect. Identity theft and credit card fraud are real threats, but a larger, more costly threat is to businesses, government agencies, and other organizations with valuable sensitive information. Recent Congressional hearings on threats to cyber security summarized the situation as "computer-based network attacks are slowly bleeding US businesses of revenue and market advantage" (Source: Elinor Mills, "Experts Warn of Catastrophic Cyberattack" at http://news.cnet.com/8301-27080_3-10458759-245.html). Businesses are now facing the kind of sophisticated, long-term attacks once limited to governments; see "Report Details Hacks Targeting Google and Others" at http://www.wired.com/threatlevel/2010/02/apt-hacks/ for a glimpse into the world of advanced persistent threats.

The growing volumes of data increase the amount of data to be protected while likely be accompanied by an increasing number of applications for manipulating that data. For systems administrators, this means two things:

- More data has to be protected with access controls available in the data management systems in use—these include file systems, database management systems, and content management and related portal systems, such as SharePoint. Access controls can minimize the damage caused by malware, application vulnerabilities, social engineering, and insider abuse.
- More applications have to be assessed for vulnerabilities that could potentially expose data.

Systems administrators not only have to protect the growing volumes of data but also help control that growth by utilizing information life cycle management practices.

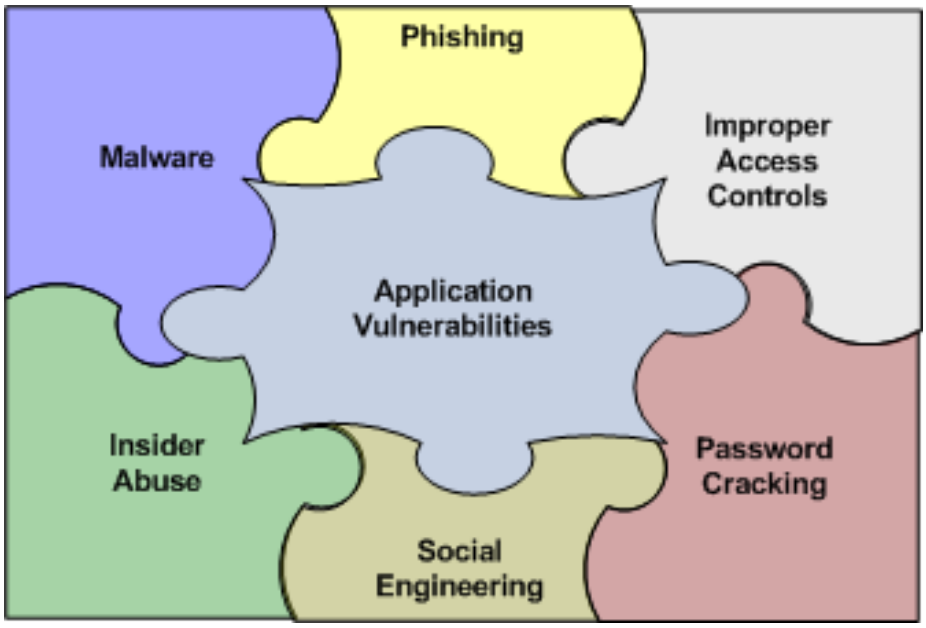

Figure 42: Increasing amounts of data combined with increasing numbers of applications expand the opportunities for exploiting existing threats to compromise confidentiality, integrity, and availability of data.

Tip, Trick, Technique 22: Getting Control of Data Growth with Information Life Cycle Management

With the growing volumes of data naturally comes a response from those who are tasked with managing it. The term "information life cycle management" is used to describe an array of management practices designed to rationalize the process of creating, collecting, storing, and in some cases destroying data. Information life cycle management makes use of tools and technologies, such as backup and recovery software and information classification systems, but it is primarily a business practice.

In its most basic form, information life cycle management answers key questions about data within an organization:

- What is the value of this data to the organization?
- How long must this data be kept?
- What are the retrieval requirements for different types of data?
- What security protections are appropriate for different types of data?
- Under what conditions can data be destroyed?

To answer these questions, we have to look at the business case for keeping data. There are several important drivers:

- Customer expectations—For example, customers may expect to look up account history for the past 3 years. If the customer self-service application provides a briefer history, then a key business requirement may not be met.
- Compliances—Government and industry regulations can specify the kinds of data that must be kept about business operations and the types of protections that must be applied to data, such as private information about patients.
- Non-compliance legal requirements—We mentioned e-discovery briefly. The focus was on the need to be able to produce documents relevant to legal proceedings.
- Maintaining institutional knowledge—Documents, emails, wikis, and other unstructured data sources can capture valuable intellectual property and institutional knowledge. This issue is especially important in industries in which the aging workforce brings the potential to lose institutional knowledge when significant numbers of employees retire.
- Mitigating risk—Mission-critical applications are backed up frequently and in some cases their data is replicated to enable rapid failover and recovery.

These drivers show the range of reasons we generate, store, and maintain data. Now let's turn our attention to implementing an information life cycle management practice.

Step 1: Classifying Data

The first step in information life cycle management is differentiating data so that we can treat data according to its value to the organization. For example, confidential engineering diagrams may need to be encrypted when they are backed up; whereas, the contents of the public Web site do not. Classifying data should include determining

- Where data should be stored
- How it should be backed up and archived
- Whether it should be replicated for automatic failover
- Whether it should be deleted

Classifying data would be a tedious and costly operation without automation. Fortunately, with Windows Server 2008 R2, additional capabilities in File Sever Resource Manager provide the tools we need to classify data efficiently. Windows Server 2008 R2 includes the File Classification Infrastructure (FCI), which can be used to classify data file attributes such as:

- File name
- Data type
- Location
- Content

These properties can be used to execute particular commands based on the classification. For example, if a file is on a high-performance disk array but has not been accessed for more than 2 years, it may be moved to a slower, less expensive disk archive. In addition, the FCI provides reports along with the ability to apply policies according to classifications.

> **Cross Reference**
>
> See Tip, Trick, Technique 22: Classifying Files in R2 for more details on how to use the FCI.

Step 2: Determining Access Requirements for Categories of Data

This step is also part of a security management process. The goal here is to ensure that only users with legitimate business need for data have access to that data. This information can be used to

- Determine whether directory protections are proper for the type of data in a directory
- Determine what files need to be encrypted during backup to protect confidentiality
- Identify all users with access privileges to sensitive information

Step 3: Defining Recovery Requirements for Data

The next step is defining recovery requirements for data. This includes defining Recovery Point Objectives (RPOs) and Recovery Time Objectives (RTOs). The RPO specifies the point in time in which data should be recoverable, such as any day in the past week or any week in the past month. RTOs define how long the recovery operation should take. Critical systems may have short recovery windows, such as a few minutes, while other relatively stale systems may not be needed for days.

Step 4: Defining Explicit Policies for Destroying Data

With so much emphasis on protecting data from tampering and keeping multiple backup copies so that we can restore, it is easy to forget about destroying data. Let's face it, not every email we write, spreadsheet we put together, or database we compile is worthy of study by some future archeologist. It often is not worth keeping after a few years. Some examples of data that can be purged include:

- Draft versions of documents preserved elsewhere in final form
- Documents in user directories that have not been accessed in 2 years but are archived
- Copies of data in analytic databases that are more than 3 years old and no longer used for reporting and analysis

The first four steps address the organizational factors of information life cycle management. The fifth step focuses on implementing those policies.

Step 5: Implementing Information Life Cycle Policies

The FCI mentioned in Step 1 for classifying data can also be used to enforce policies. (Policies are essentially rules that fit into a pattern of "IF a certain set of conditions are met, then execute this script.") Policies themselves have to be managed, so once they are defined in the FCI, be sure to:

- Determine who will monitor the process to ensure the policies are applied as expected
- Set a schedule to review and revise policies
- Occasionally audit file systems to ensure polices are applied properly and comprehensively

When dealing with unstructured data, policies may be applied in unintentional ways. For example, a policy may specify that any document with word "confidential" in the title be categorized as proprietary information. This rule would apply to a document entitled "Introduction to Data Classification Policies for Documents Ranging from Public to Confidential."

Limits of Information Life Cycle Management

Information life cycle management is no panacea for the challenges of data growth. At best, these practices will help:

- Protect our data by guiding security decisions
- Ensure proper levels of backups and archives are made
- Control the cost of storing data by optimizing the distribution of data based on access requirements and use high cost/high performance storage only where needed

A couple of problems will continue to challenge information life cycle management. First, automated classification techniques are not foolproof and their results should be reviewed. Second, data is easily copied both informally to employees' directories and workstations and formally to backups and failover servers. Destroying old data may never be 100% successful; copies may linger for years in unexpected places.