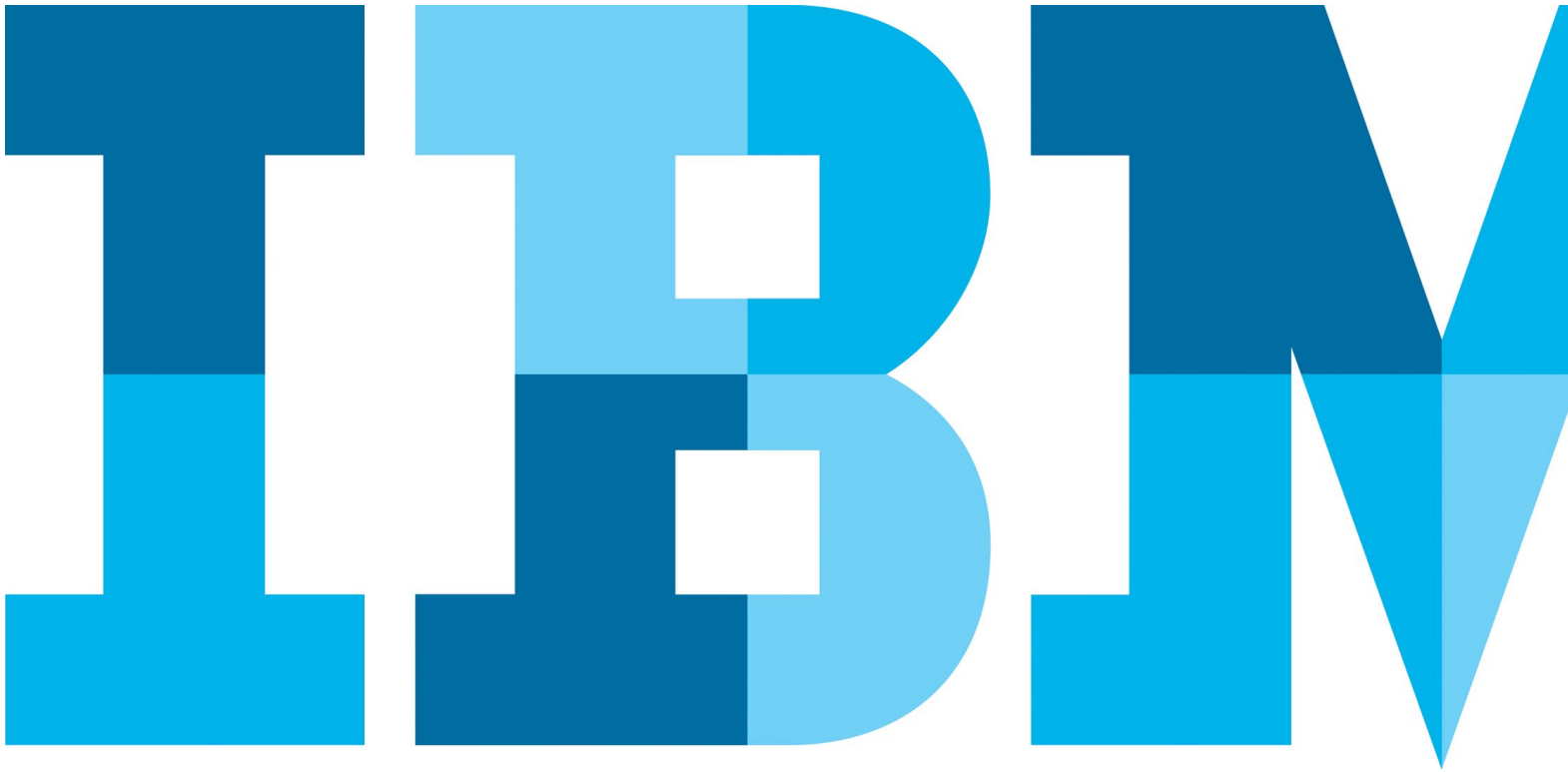
The truth about information governance and the cloud



ENTER »

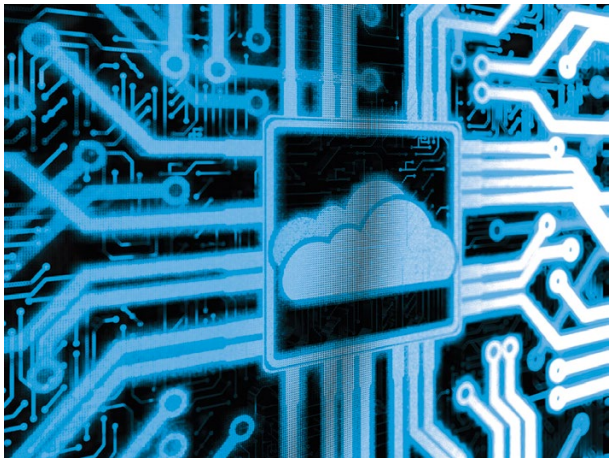
Table of contents

The truth about cloud

Chatter about the cloud is everywhere. You can't turn on your TV, look at your smartphone, open a magazine or browse websites without being inundated with messages about the cloud. Proponents tell you the cloud will save you time, give you a place to store data, create a way to manage hard drive space on your phone and much more. Detractors will scare you with stories of hackers gaining access to personal photos and bank account numbers.

In the realm of business information technology, the conflicting stories are much the same: Will the cloud save you money on your balance sheet, or will the cloud expose your sensitive data to unwanted prying eyes? Some analysts claim that over 50 percent of businesses will do a majority of their computing on the cloud before the end of the decade. Others say the number is more like 10 percent.

As with most innovations in business information technology, the ultimate truth about cloud lies somewhere in between. There is little doubt that cloud-based infrastructures offer an immediate opportunity for smaller organizations to avoid the costly investment needed for a robust on-premises computing environment. Data can be found, processed and managed on the cloud without investing in any local hardware. Large organizations with mature on-premises computing infrastructures are looking to Hadoop platforms to help them benefit from the vast array of structured and unstructured data from cloud-based sources. Organizations have feet in both cloud and on-premises worlds. In fact, one could easily argue that we already live in a "hybrid" world.



The emergence of the hybrid environment

Hybrid simply means a mixture of public cloud and on-premises data sources and computing in support of business operations (see Figure 1).

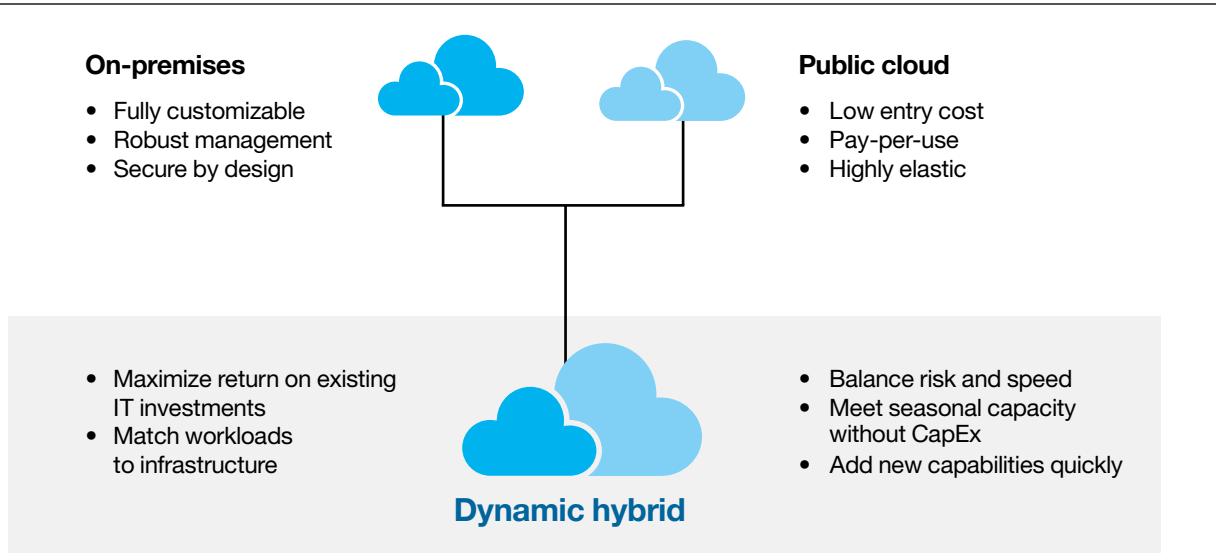


Figure 1. A hybrid environment blends data and computing from both public cloud sources and on-premises systems.

But here's a secret: no one sets out to plan a hybrid environment; it just happens. Consider this scenario:

A marketing executive wants to understand whether a new campaign is influencing customer sentiment. She is in the midst of a large media buy, and needs to know if the radio and TV ads are having any impact. With limited resources, pulling the plug on a bad campaign could save a few million dollars—money she could reinvest in other marketing activities.

So, she tasks her analysts with figuring out a way to gauge customer sentiment before actual sales numbers arrive (see Figure 2). Her team goes to third-party sources like Twitter and raw retail scanner data that they feed into their own analytical tools to create a model of customer sentiment impact. All of this research, data collection, integration, processing and analysis takes place without IT involvement.

Assuming the executive gets a satisfactory answer, regardless of whether the news is good or bad, she will be pleased by her team's speed in assembling data and producing information without IT. This is an approach she will return to again in the future.

This organization is in hybrid mode: IT is hosting on-premises sales information that is reported on a monthly basis, and marketing is curating data from the cloud to spot customer sentiment trends in near-real time. The setup works very well for the marketing executive; however, IT will worry about security and scalability. What if Sales and Operations want to emulate Marketing's approach? The business units want speed and flexibility, while IT wants scalability and security. Can these competing interests be unified?

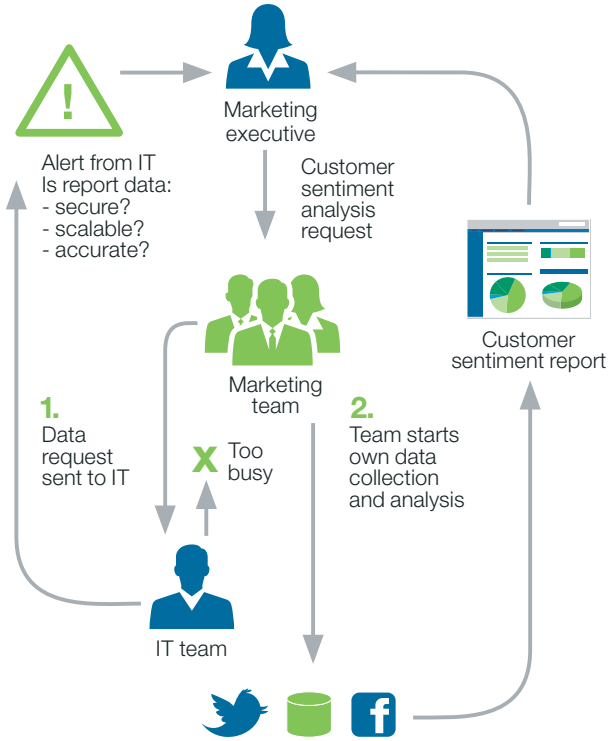


Figure 2. When teams bypass IT to build their own on-the-fly analysis reports, it may create concerns about scalability and accuracy.

Ownership of strategic information

Adopting a hybrid environment does not imply you must have your IT strategy completely worked out. In fact, cloud-based aspects of the environment will evolve rapidly in response to business priorities.

However, even if only a small percentage of data is flowing in from cloud-based sources, IT does need a plan for data integration and security. IT needs to help the organization ensure it "owns" the information created from all data and processing, no matter where it is located.

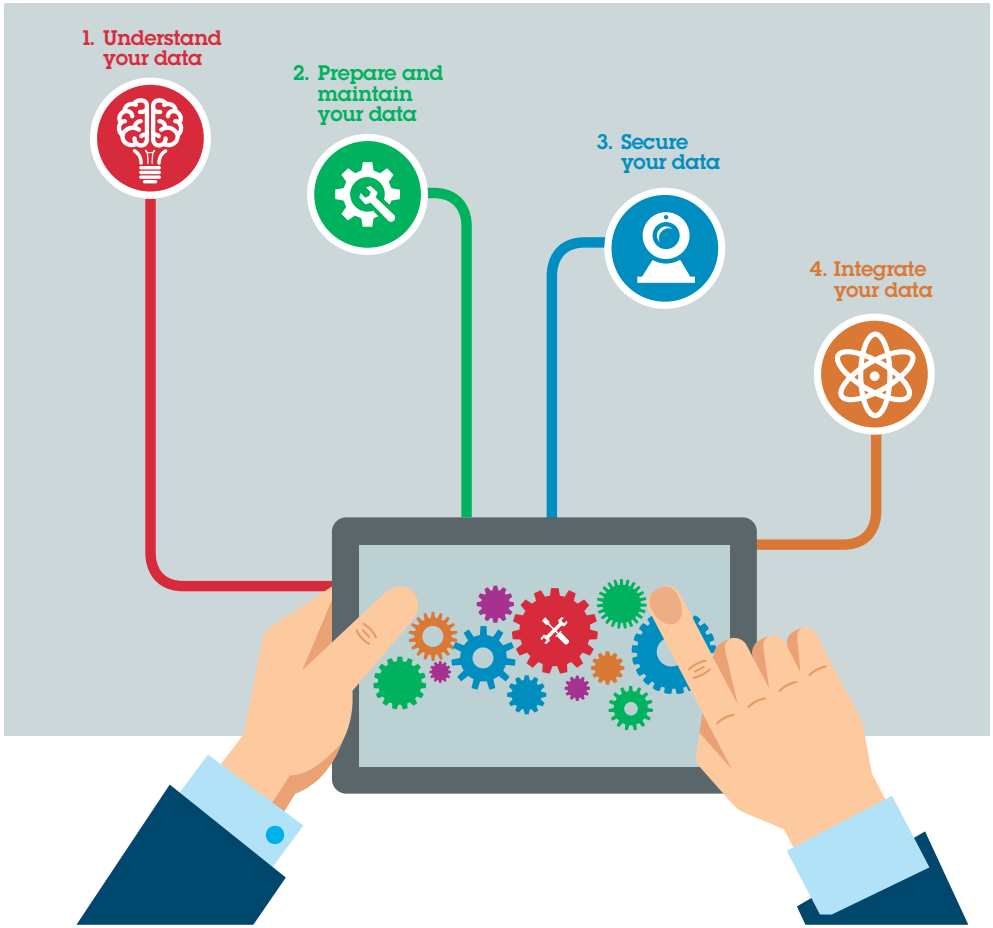
The hybrid infrastructure and decentralized computing are merely means to the ultimate end of creating strategic information assets. Embracing this fundamental notion lends clarity to what IT should be concerned with, and importantly, how IT can more effectively partner with the business users.

How can organizations realize the financial benefits of the cloud while ensuring information culled from cloud sources is secure and trustworthy? The answer is governance. Good hybrid information governance implies several priorities for IT and the business:

1) **Broad agreement on what information means**, including metadata on common policies and plain-language rules for the information the business needs and how it will be handled.
2) **Clear agreement on how owned information assets will be maintained and monitored**, for example, operational data quality rules to master data management in on-premises systems.
3) **Enterprise- and departmental-standard practices for securing and protecting strategic information assets**, such as articulating role-based access to information, creating rules governing how information is shared, and protecting sensitive information from third parties.
4) **Enterprise data integration strategy that includes lifecycle management**, architecting how data will flow and be assembled into strategic information, and also understanding how that data/information will be maintained over time.

These are the foundations of information governance in a hybrid environment. In each case, a blend of process, organizational and technical enablers are needed to make it work. **With these pillars in place, organizations will have the flexibility to move with speed and confidence. Let's look at what each pillar involves**.

Roll over the icons below for more on the top priorities for good hybrid information governance.



Pillar #1: What information means: Terminology and metadata management

The most important business requirement for organizations today is the ability to completely understand and trust their information. For instance, financial analysts looking at loan applications for prospective borrowers need to know exactly what a "mortgage risk" entails: the approved definition of "risk," acceptable levels of risk, the business rules that are associated with "high-risk" and "low-risk" clients and where a specific prospect falls on that scale. If the analysts understand this information and trust that the data in their applications is aligned with the same business rules and definitions, they can make better approval decisions, more quickly and confidently.

Unfortunately, in most organizations, business analysts and data analysts often have multiple definitions for a given business term. Even worse, all of those definitions may be correct, depending on context and usage. For example, "mortgage risk" for existing customers and "mortgage risk" for new prospects are two completely different concepts, definitions and related business processes.

This problem is amplified when third-party cloud-based data sources are added to the mix. Without an understanding of the business context of third-party data, it could negatively impact analysis.

The dilemma of business definition ambiguity and inconsistency is often attributed to the absence of an enterprise-wide business glossary and stewardship program, which is often part of a larger metadata strategy plan. A metadata strategy comprises two elements: *technical metadata* and *business metadata*.

Technical metadata describes the shape, size and format of data, content, business processes, services, business rules and policies. Business metadata describes the business context for those assets. Linking business metadata to technical metadata through a common metadata repository facilitates collaboration and better communication between business and technical users.

Business metadata: What it does

- **Provides business context and meaning around IT assets to promote shared understanding of data, business processes, analytics and key performance indicators**
- **Increases trust and confidence in information for faster decision making in response to market changes and new opportunities**
- **Fosters collaboration, knowledge sharing and information reuse**
- **Supports data governance initiatives that require common business vocabulary definitions and data stewardship**

A critical path to information governance involves putting in place correct data definitions that the organization as a whole can use to better understand information (see Figure 3). While a full understanding of business context and meaning resolves ambiguity and leads to more accurate decisions, users often require more detail behind their data. Understanding where the data is coming from—and who modified it and when—significantly affects its value, authenticity and accuracy.

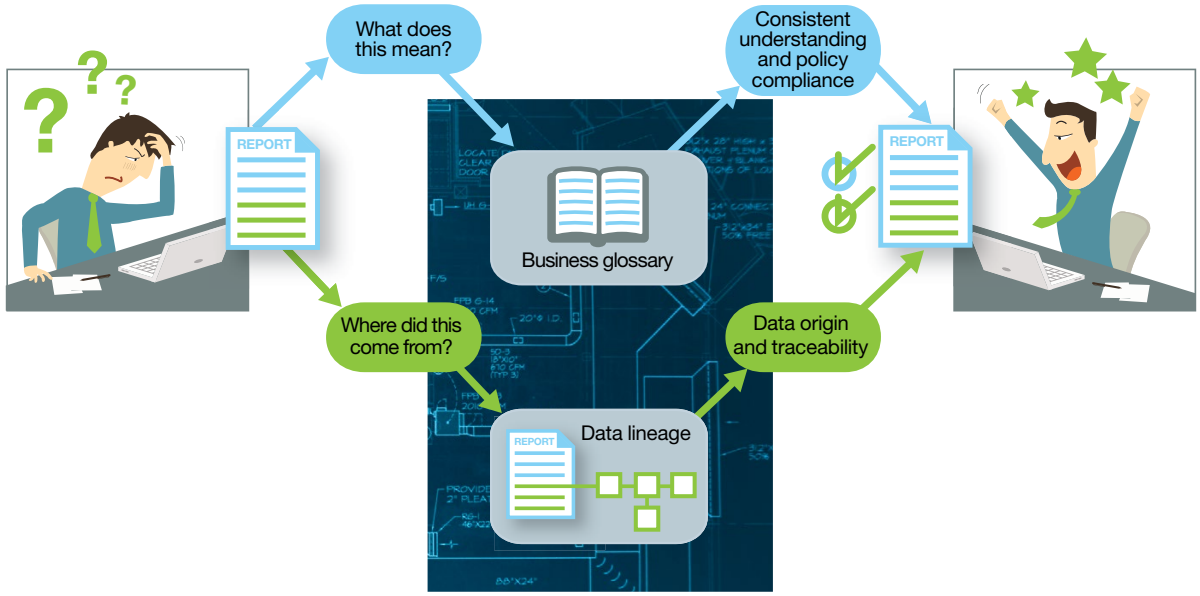


Figure 3. Business glossaries and data lineage tracking enables users to quickly answer questions about the meaning and source of information.

Pillar #2: Maintaining and monitoring assets: Quality stewardship

With so many possible data sources now so readily available, the classic problem of "garbage-in, garbage-out" is amplified. This data explosion is not limited to structured data; in fact, most of the added volume flows from unstructured sources, such as email, images and documents. Missing, inaccurate or incomplete information can generate high costs and reduce productivity when workers have to hunt for information or reconcile data.

An organization must be able to manage its supply chain of information, and then integrate and analyze it to make business decisions, as shown in Figure 4. Unlike a traditional supply chain, an information supply chain has a many-to-many relationship. For example, data about the same person can come from many places—that person may be a customer, an employee and a partner—and the information can end up in many reports and applications. In addition, various systems may define the same information differently.

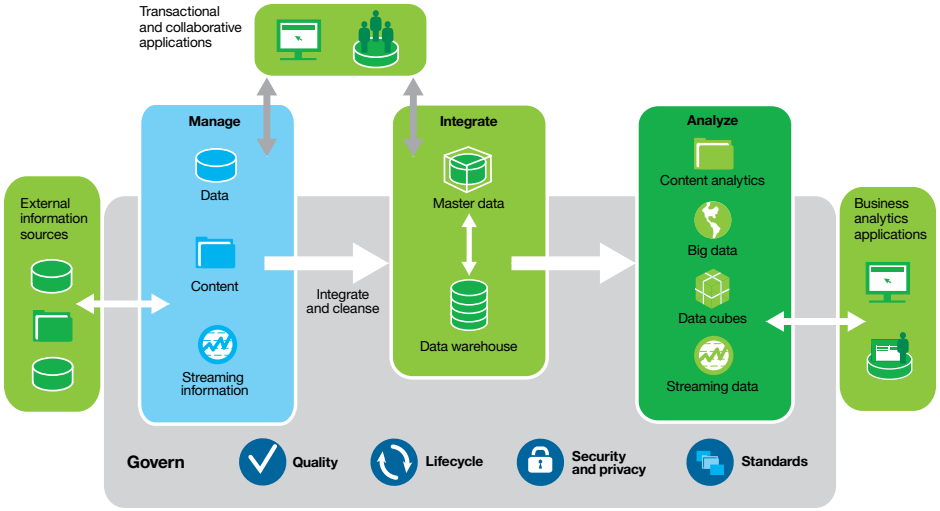


Figure 4. Governance enhances the quality, availability and integrity of the information supply chain.

Given this complexity, integrating information, ensuring its quality and maintaining a master of the information are crucial tasks. Information needs to be transformed into a trusted asset and governed to maintain quality across its lifecycle. The underlying systems must be cost-effective, easy to maintain and perform well for the workloads they handle, even as information continues to grow at exponential rates.

Effective information governance can enhance the quality, availability and integrity of an organization's data by fostering cross-organizational collaboration and structured policy making. Governance balances functional silos with enterprise-level oversight, directly affecting four factors critical to an organization: increasing revenue, lowering costs, reducing risk and increasing confidence.

Excellent data quality is achieved through several essential attributes:

- **Completeness:** All related data must be linked from all possible sources.
- **Accuracy:** Data must be correct and consistent, with common data problems remediated, such as misspellings or abbreviations.
- **Availability/timeliness:** Data must be available upon demand.
- **Mastery:** Critical data must be managed and maintained.

The ability to define, monitor and maintain these data quality attributes is essential in a hybrid environment.

Pillar #3: Securing information assets: Privacy and compliance

Organizations should centralize security controls in hybrid environments and ensure a separation of duties so the data administrator doesn't also become the security administrator or auditor. Key elements of a hybrid environment security strategy include:

- **Understanding where the data exists:** Organizations can't protect sensitive data unless they know where it resides and how it's related across the enterprise.
- **Safeguarding sensitive data,** whether it is structured or unstructured, online or offline, with the appropriate technologies and with the right access requirements established. This is especially important when working with third-party data, or if internal data is being stored, processed or managed in the cloud by third parties.
- **Monitor data activity in real time:** The security solution should identify unauthorized or suspicious activities by continuously monitoring access to databases, data warehouses, Hadoop and file share platforms in real time.
- **Audit and validate compliance:** Simplifying Sarbanes-Oxley Act, PCI-DSS and data privacy processes with preconfigured reports and automated oversight workflows helps satisfy mandates.
- **Demonstrating the compliance to pass audits with pre-built reports for auditors:** Holistic protection strategies for private cloud environments should provide alerts to security administrators of suspicious behaviors such as unusual network activity.

Organizations should also consider data security solutions that provide audit reporting capabilities and sign-offs to streamline the compliance process. Data security processes for hybrid environments must continuously track data across the cloud and on-premises systems to provide insight into who is accessing the data across applications, databases and other repositories. Such an approach ensures a 360-degree lockdown of all organizational data, no matter where it resides, in every stage of its utilization.

To ensure data is protected in hybrid environments, organizations must understand what data is going into these environments, how access to this data can be monitored and what types of vulnerabilities exist. Protections should be built into hybrid environments from the start with the goal of helping organizations demonstrate compliance.

When choosing data security solutions, organizations should select those that are scalable and unified across IT infrastructures—protecting physical, virtual and cloud environments from malicious external attacks, fraud, unauthorized access and insider breaches (see Figure 5). These solutions must work in a hybrid environment without any special setup, configuration or added expense. Such an approach will provide an efficient platform for data security and privacy delivery, help manage costs by reducing data security resources, and provide greater agility and flexibility with self-service features for security and privacy.

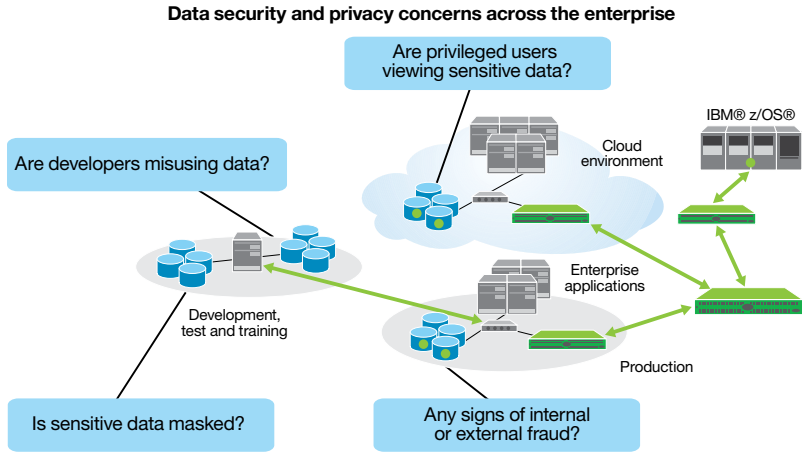


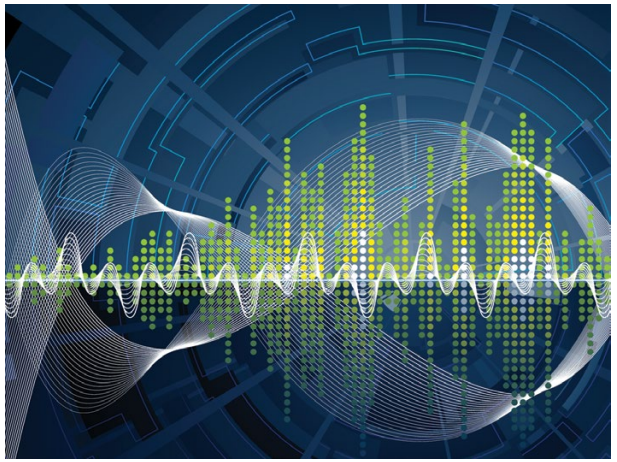
Figure 5. Organizations must secure data wherever it resides, including test, production, on-premises and cloud environments.

Pillar #4: Turning data into information: Integration and lifecycle strategy

In a hybrid data environment, torrents of data inundate IT organizations and can easily overwhelm the business managers who must sift through it all to glean insights that help them grow revenues and optimize profits. Yet, after investing hundreds of millions of dollars into new enterprise resource planning (ERP), customer relationship management (CRM), business intelligence (BI) and data warehousing systems or big data environments, many companies are still plagued with "dysfunctional" data: a massive, expensive sprawl of disparate silos and unconnected, redundant systems that fail to deliver the desired single view of the business.

The cloud portion of the hybrid environment offers a possible solution to the considerable expense of investing in these various information management systems, because many of them can be provided at less expense. However, the fundamental problem of data integration remains. What data do you need for analysis? Where is that data being fed? What will you do with the information that is created? How long do you need to maintain the data? Where will you maintain the data and the information created from it?

These are all critical aspects of data integration and lifecycle management. Even if your organization is only beginning to explore the business potential of a hybrid infrastructure, now is the time to clearly define your data integration and lifecycle strategy.



To meet the business imperative for enterprise integration and lifecycle planning, companies must manage the increasing variety, volume and velocity of new data pouring into their systems. They need to bring all their corporate data together, deliver it to end users as quickly as possible to maximize its value, and integrate it at a more granular level than ever before—focusing on the individual transaction, rather than on general summary data. As data volumes continue to explode, clients must take advantage of a fully scalable information integration architecture that supports any type of data integration technique, including extract, transfer and load (ETL), data replication and data virtualization.

A solution that supports a true hybrid data integration and lifecycle strategy should have the following characteristics:

- **A data flow architecture** supporting data pipelining that allows data to be processed from input to output without landing to disk, as data is moved between different operations such as profiling, cleansing and transformations
- **Dynamic data partitioning** and in-flight repartitioning of data
- **Scalable hardware environments,** portable across symmetric multiprocessing (SMP) clustered environments and massively parallel processing (MPP) platforms that do not require modifications of the data flow design
- **High performance and scalability for bulk and batch movement,** and for real-time data replication and processing
- **Extensive tooling** to support resource estimation, performance analysis and workload management
- **Policies and an enforcement mechanism** governing where and for how long data and information will be maintained
- **Ability to view complete business objects**
- **Capabilities to analyze and identify dormant data**

Next steps: Continuing the cloud governance discussion

Cloud-based data and processing services present too much opportunity for business users to ignore, and IT is charged with maintaining the integrity of internal, on-premises transactional and reporting systems. Charting a governance strategy for a hybrid environment is not something to consider at a future date. It needs to happen now.

This e-book provides an overview of four pillars for successful hybrid environment governance. For a deeper look at each of the pillars, download one or all of the e-books in this series:

- Make sense of your data
- Prepare and maintain your data
- Securing data in the cloud and on the ground
- Developing a data integration and lifecycle management strategy for a hybrid environment

For additional information on IBM governance thought leadership and supporting technologies, **ibm.com**/analytics/us/en/technology/agile/

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
July 2016

Please Recycle

IMM14172-USEN-01