The era of hybrid environments is here: Is your data refined and ready?

Consistently govern, manage and integrate your data wherever necessary—on-premises or in the cloud

IBM

Introduction

Today, all consumers can obtain any piece of data at any point in time. **This experience represents a significant cultural shift: the beginning of the democratization of data.**

However, the data landscape is increasing in complexity, with diverse data types from myriad sources residing in a mix of environments: on-premises, in the cloud or both. How can you avoid data chaos?

The good news is that there's an emerging approach that supports these dynamic new scenarios: the data refinery. This e-book describes how a data refinery can make trusted data available quickly and easily to people and systems across your organization. It includes simple steps you can take to start exploring—and implementing—this strategy for handling hybrid data environments.

The democratization of data

Businesspeople expect more from their applications. After all, businesspeople are consumers, too. They know the devices they carry in their pockets can bring them the latest news, track stock performance, help them find their way in a strange city, show them the best products to buy and answer questions on every imaginable subject—all in seconds, with no special training or expertise required.

So it's not surprising that in a work setting, they also expect access to enterprise data whenever and wherever they need it. The expectation set by the democratization of data becomes even more intense as business leaders encourage teams across the enterprise to achieve greater agility, efficiency and market responsiveness.

The rising popularity of cloud-based business applications and environments is also helping to accelerate the democratization of data. Organizations can deploy cloud applications and environments more quickly than on-premises ones—and without the same level of IT cost overhead.

Now, however, it's easier for start-up organizations and other businesses with flat organizational structures to attain data democratization than their larger, more conventional counterparts. **According to Forrester, "Agile start-ups and flat organizations lead the charge in changing expectations around data use. A hallmark of this new approach is that customer data is no longer the exclusive domain of data analysts and customer insights teams . . . [E]very employee [should have] access to customer data."**[1] These types of organizations don't need to evolve from more traditional attitudes and physical approaches to data where IT has historically been the keeper of data, and data is often trapped inside data silos and legacy systems.

Data democracy or data chaos?

History has shown that evolution within societies can lead to either democracy or chaos. Likewise, the evolution of attitudes toward democratic data access and the ability of infrastructure to support such access—in a constructive and governed way—face a similar challenge.

For example, line-of-business (LOB) teams understand the advantages that cloud solutions offer for spinning up projects that will quickly yield results, such as running cloud-based marketing campaign analytics. As a result, increasing numbers of IT teams are struggling to maintain order and control as they are bombarded with cloud requests—or as teams go off and try to establish these projects on their own.

In this barrage of requests, IT teams see the risk of data chaos. They have three key areas of concern:

1. **Creation of cloud silos** by standing up unique cloud projects, leading to an unhealthy and unstainable architecture; after all, many teams are still trying to untangle data silos on-premises—they don't want to create cloud silos as well
2. **Loss of visibility into and control over data** quality, lineage and governance
3. **Contamination of on-premises systems** by chaotic and untrustworthy cloud data

The underlying obstacle here is that "the route to improved agility, streamlined operations and customer focus is slowed by the lack of integrated data and an increasingly distributed data landscape of both cloud and on-premises systems."[2]

But adopting cloud environments is not the same as abandoning on-premises environments. At most large organizations, more than 90 percent of data exists in on-premises environments[3]—and that data powers mission-critical systems. As a result, cloud applications cannot exist in a vacuum or in a silo. For example, you can purchase a software-as-a-service (SaaS) cloud application, but it still needs on-premises data to work.

In your move to a more distributed environment, you run the risk of creating data chaos if you don't invest in modernizing your information integration infrastructure. You also need to plan ahead to ensure the integration infrastructure can support cloud environments and requirements. That's because "the complexity of an ever-increasing distributed data landscape and a rapid growth in demand for new data sources of varying data types is making it harder to know where data is and to integrate data in a timely manner."[4]

Drive business growth in a hybrid environment

To support a successful move to a distributed environment that includes cloud infrastructure, you need to renew your investment in information integration infrastructure. Analyst firm Ovum notes that "enterprises need to undertake integration infrastructure modernization to effectively exploit the quartet of digitalization, mobility, cloud and Internet of Things (IoT) for driving business growth."[5]

The key to supporting data democratization—and minimizing data chaos—is to implement an approach that has the flexibility to mix data from different environments.

At the same time, you should be able to maintain control over that data, ensuring it is still integrated, trusted, governed and available throughout the enterprise.

Making data useful to everyone

The data refinery approach provides data access and refinement services that help you to successfully realize the democratization of data (see Figure 1). A data refinery makes it easy for IT teams to support cloud, on-premises and, importantly, hybrid environments (which mix both cloud and on-premises data). It also can empower other teams, such as business analysts and application developers, by providing self-service data access and data refinement services. Plus, any team can build these readily available services into any cloud application.

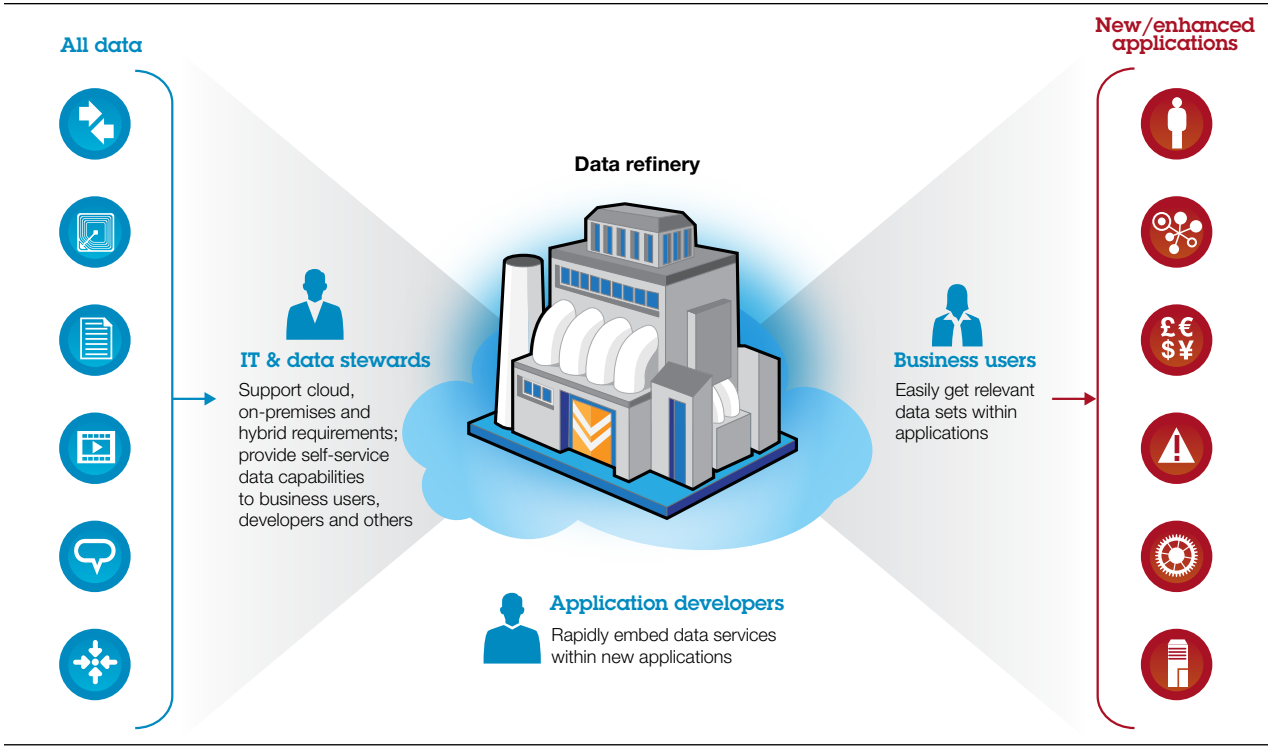


Figure 1. On-premises or in the cloud, a data refinery makes data useful and available to everyone.

IBM® DataWorks™ provides cloud-based data access and refinement services through REST application programming interfaces (APIs). You can use the services in several ways:

- **For developers:** Developers can use the APIs to quickly enrich applications, improve the quality and usability of data in the application, and speed development time.
- **For business analysts:** The APIs can be embedded in other applications, which means business analysts can leverage DataWorks capabilities quickly and easily.
- **For IT teams:** In addition to supporting cloud and hybrid environments, the APIs can be embedded into IT-oriented applications such as IBM dashDB.

IBM DataWorks supports the new on-premises/cloud landscape that you may already have or are considering adopting. It is designed to integrate data from on-premises and cloud sources, refine that data and then share that data with on-premises targets, applications that consume cloud-based services (such as IBM Watson™ Analytics and IBM dashDB) or other cloud targets (see Figure 2).

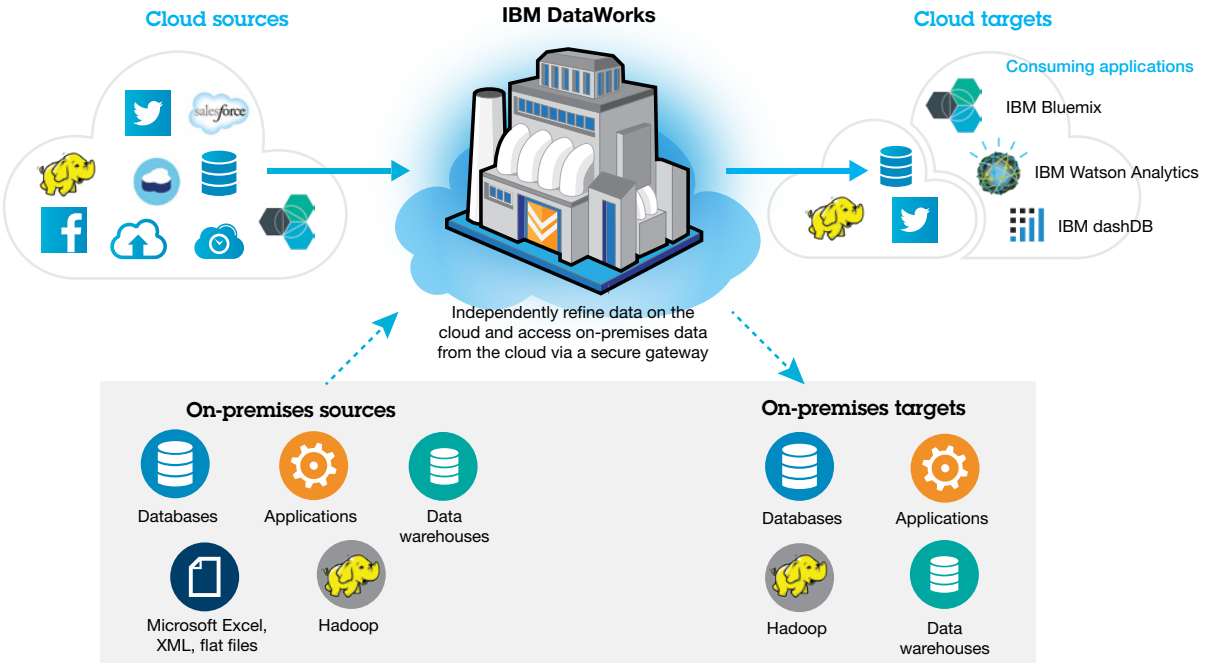


Figure 2. IBM DataWorks delivers refinement services on the cloud.

Rise of the hybrid environment

A hybrid cloud/on-premises environment will be the path forward for many, since it's the most cost-effective and least-disruptive approach over time. This paradigm supports your need to extend into the cloud, but also helps you leverage on-premises data and investments, such as existing data refinement capabilities delivered by the IBM InfoSphere® Information Integration and Governance (InfoSphere IIG) portfolio (see Figure 3).

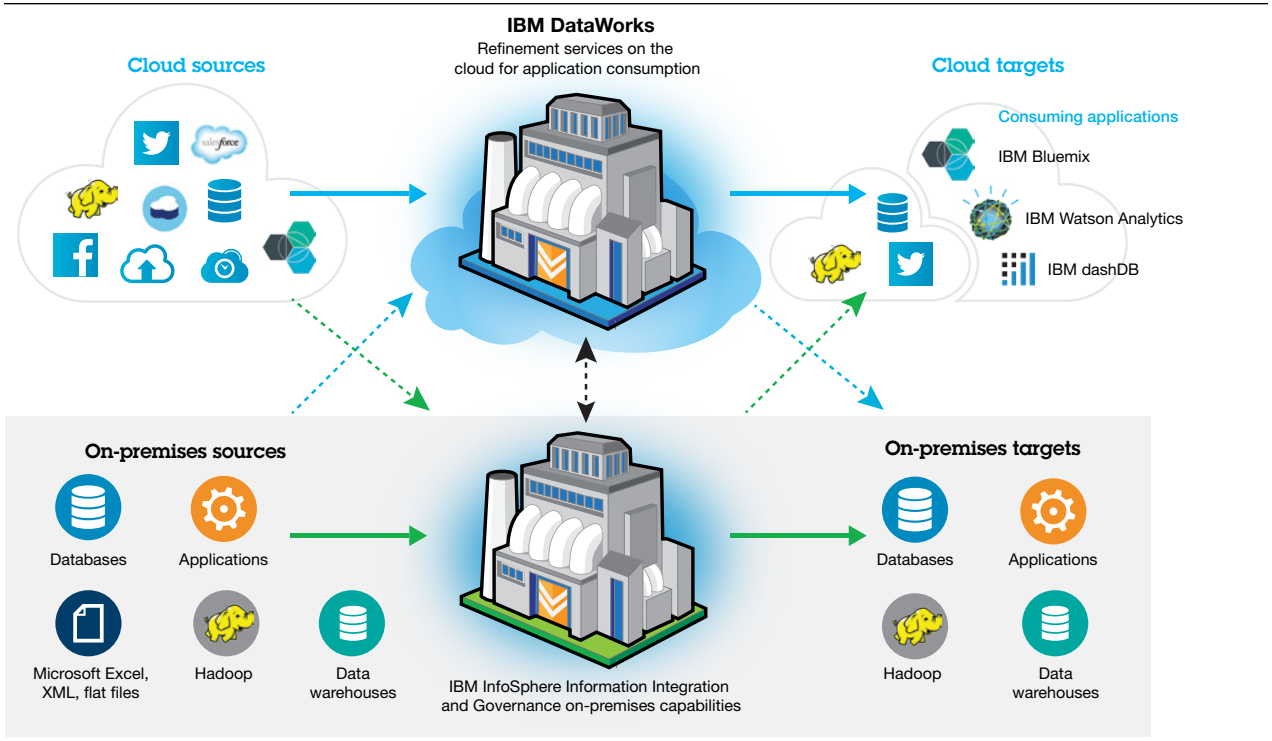


Figure 3. Hybrid environments can utilize IBM data refinement services and governance capabilities that operate on-premises and in the cloud.

The InfoSphere IIG portfolio includes:

- **Metadata, business glossary and policy management:** Define metadata, business terminology and governance policies with IBM InfoSphere Information Governance Catalog.
- **Data integration:** Handle all integration requirements, including batch data transformation and movement, real-time replication and federation.
- **Data quality:** Parse, standardize, validate and match enterprise data.
- **Master data management:** Act on a trusted view of your customers, products, suppliers, locations and accounts.
- **Data lifecycle management:** Manage the data lifecycle from test data creation through retirement and archiving.
- **Data security and privacy:** Continuously monitor data access, protect repositories from data breaches and support compliance.

All elements of the data access, refinement and integration scenario are supported in this hybrid world:

- **On-premises environment:** Data is integrated from on-premises and cloud sources and delivered to on-premises targets.
- **Cloud environment:** Cloud and on-premises data is integrated and refined on the cloud and can be delivered to either cloud or on-premises targets.
- **Hybrid environment:** Mixed on-premises and cloud data is processed wherever necessary, and data access and refinement capabilities flexibly deliver data wherever and whenever it's needed.

As part of this new approach, it's important to either establish an on-premises data refinery or evaluate your existing one. For example, you may need to expand or update current capabilities while extending your infrastructure to support cloud environments.

The data refinery approach lets IT teams take advantage of the respite provided by a self-service environment. Instead of fulfilling data access and data provisioning requests from business users and developers, IT teams can focus on improving data governance, curation, protection and more.

Example scenario: Deploying a hybrid environment

The IT team at a large retailer was being bombarded by cloud requests from multiple LOB teams. The teams had urgent cloud-based projects they wanted to launch so they could perform analytics more quickly.

The IT team was worried that all these separate, siloed projects would create an unhealthy, unsustainable architecture. One-off projects could lead to IT losing visibility into and control over data quality and lineage—in other words, data chaos. They were also divided about where data processing should occur: in the cloud or on-premises.

After discussing their options and architectural choices, the IT team decided to take a holistic and integrated approach. In fact, it was a strategy that started on-premises, because the team felt that if it didn't have control of known, on-premises data, it wouldn't be able to govern and keep up with cloud data as well.

The IT team also took a pragmatic, middle ground when it came to data processing, since most of the company's data resided on-premises. The team decided to integrate some of the key on-premises data to enrich the cloud-based analytics requested by the LOB teams. The results would be integrated back into the on-premises systems for consistency. As the amount of cloud-based data increases, the team will be able to dynamically shift where the processing occurs.

In addition, the IT team focused heavily on enhancing on-premises integration, quality and governance capabilities. At the same time, it began supporting cloud-based marketing campaign analytics that leveraged Salesforce.com data as well as on-premises data.

The previous example demonstrates one approach you can take to successfully deal with increasingly distributed environments. Ultimately, the solution is to "bring it all together on a single platform that embraces the distributed data landscape across the cloud and on-premises and that facilitates communal, collaborative data governance whereby business users, data stewards, IT developers and IT architects can all contribute to the provision and reuse of trusted, secure, well-managed data."[6]

IBM DataWorks is designed to support these needs while complementing existing on-premises data access and refinement capabilities (see Table 1).

Service	What it does
Load data	Gets users the data they need by moving data between cloud data stores.
Provision masked data	Uses the data load API to mask sensitive data at the source while it is moved.
Profile data	Profiles data structure and content, with details such as data types, lengths, formats and value frequencies.
Classify data	Classifies each field in a data domain to identify fields that contain sensitive data or those to use for statistical or predictive analysis.
Secure on-premises load to cloud targets	Uses the data load API to access on-premises data using a secure gateway that helps protect against security intrusions. Provides a quick and easy configuration process to access on-premises data sources transparently from the cloud.
Cleanse addresses	Standardizes and enriches addresses to improve data quality.
Probabilistic matching	Matches customer information within and across sources for operations and analytics. Patented algorithms return accurate search results and relevancy scores to identify duplicate customer records and support building a 360-degree view of the customer.

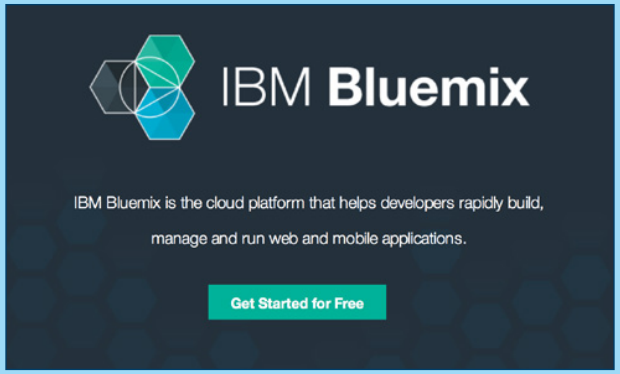
Table 1. Examples of IBM DataWorks capabilities.

Get started with IBM DataWorks

Ready to explore IBM DataWorks? It's as easy as registering for IBM Bluemix™ and discovering the available services. Follow these five steps to get started.

STEP 1:

Go to **ibm.com**/bluemix and sign up.



STEP 2:

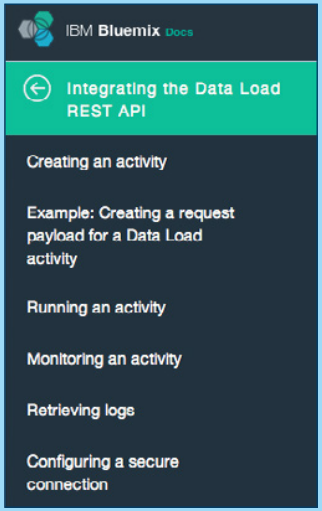
In the catalog, select the Data Management category, and choose IBM DataWorks.

STEP 3:

Explore IBM DataWorks capabilities and select an API to load data, cleanse addresses, profile data or classify data.

STEP 4:

Incorporate the selected API into a test application, following the online guide.



STEP 5:

Proceed with developing and testing your own application.

Resources

Now is the time to start integrating your on-premises and cloud data and making that data easily available to the business users, application developers and others who need it. Check out these resources or contact your IBM representative to learn more about IBM DataWorks and the data refinery approach.

- Infographic: It's time to evolve the way the world works with data
- Data Access and Refinement: Making data useful and available to everyone
- Revolutionizing the developer experience
- Revolutionizing the business analyst experience
- Revolutionizing the data experience for IT

[1] Costa, Tony, et al. "The Data-Driven Design Revolution." Forrester Research. August 5, 2014. www.forrester.com/The+DataDriven+Design+Revolution/fulltext/-/E-RES115903

[2] Ferguson, Mike. "The Enterprise Data Refinery." Intelligent Business Strategies. November 2014. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov29848&S_TACT=109HF54W&dynform=13048&lang=en_US

[3] Based on IBM customer experiences.

[4] Ferguson, Mike. "The Enterprise Data Refinery." Intelligent Business Strategies. November 2014. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov29848&S_TACT=109HF54W&dynform=13048&lang=en_US

[5] Sharma, Saurabh. "2015 Trends to Watch: Integration and Middleware." Ovum. November 10, 2014. www.ovum.com/research/2015-trends-to-watch-integration-and-middleware

[6] Ferguson, Mike. "The Enterprise Data Refinery." Intelligent Business Strategies. November 2014. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov29848&S_TACT=109HF54W&dynform=13048&lang=en_US

Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
March 2015

IMM14173-USEN-00