TRIFACTA

BRIEF

Managing Data Wrangling in Hybrid Cloud and On-premise Analytics Environments

The cloud is no longer just a future vision; it's finally becoming real. Now that companies are experiencing real benefits from using a cloud infrastructure—like reduced data center needs, more flexible computing capacity, increased business agility, and lower costs—it's no wonder the rush is on to transition data from traditional on-premise platforms to cloud-based environments.

Even still, there are hurdles along the way. For some companies, a total migration of their computing infrastructure to the cloud is not feasible, and in some cases may be undesirable. For example, firms may not wish to expose any data to a public cloud, while others may want to keep only their most sensitive data on-premise, fearing a loss of data security control by moving to the cloud. Even for those companies comfortable moving to a 100 percent cloud-based computing infrastructure, the transition is often a multi-year process.

As a result, most organizations now operate in a hybrid environment: a synchronous blend of cloud, on-prem, and private cloud computing environments. The same study reported that adoption of hybrid cloud solutions grew 3x in the last year, increasing from 19% of organizations surveyed to 57%.

Data Wrangling Difficulty Is Increasing

No matter where data is stored, organizations have always needed to be able to be prepare it for various downstream applications. This process—discovering, structuring, and cleaning the contents of data for various analytic outputs, or what we call data wrangling—has long been considered the most challenging part of analytics. Organizations typically report that 80% of any data project is spent wrangling data, while only 20% is left for analysis. As organizations trend toward self-service solutions, and an increasing number of business users demand access to the diverse data they need, patience for data wrangling delays are wearing thin.

Delivering and managing a data wrangling solution for business users becomes even more challenging across hybrid cloud and on-premise analytics environments. Despite clear architectural differences between cloud and on-premise computing infrastructure, business users won't be satisfied switching between different technologies to wrangle data in each—learning how to navigate two interfaces is challenging and, frankly, business users shouldn't have to create different processes for different environments. Along with the challenge of maneuvering between technologies, it is near impossible to compare data that doesn't share commonalities in metadata, wrangling logic, and data lineage across these environments.

Not only is siloed data preparation challenging for business end-users, but a nightmare in data governance. Different naming, formats or update rates may create various representations of the same dataset across these different environments, leaving organizations paralyzed while they wait for the "source of truth" to emerge. This prevents companies from leveraging the insights and opportunities within their data efficiently, as well as managing their risks and threats. Meanwhile, increased regulation is driving a demand for superior data lineage tracking and transparency, leaving many IT organizations scrambling to uncover a solution.

Interoperable Data Wrangling Solves Hybrid Data Challenges

Siloed data sets are the enemy of effective data-driven business. The solution? Interoperable data wrangling that functions seamlessly across computing environments. By utilizing shared logic across all sources of data, an interoperable application not only enables more efficient data preparation, but also better data governance, leading to higher and quicker returns on big data investments. Here's how it's done:

Common Metadata

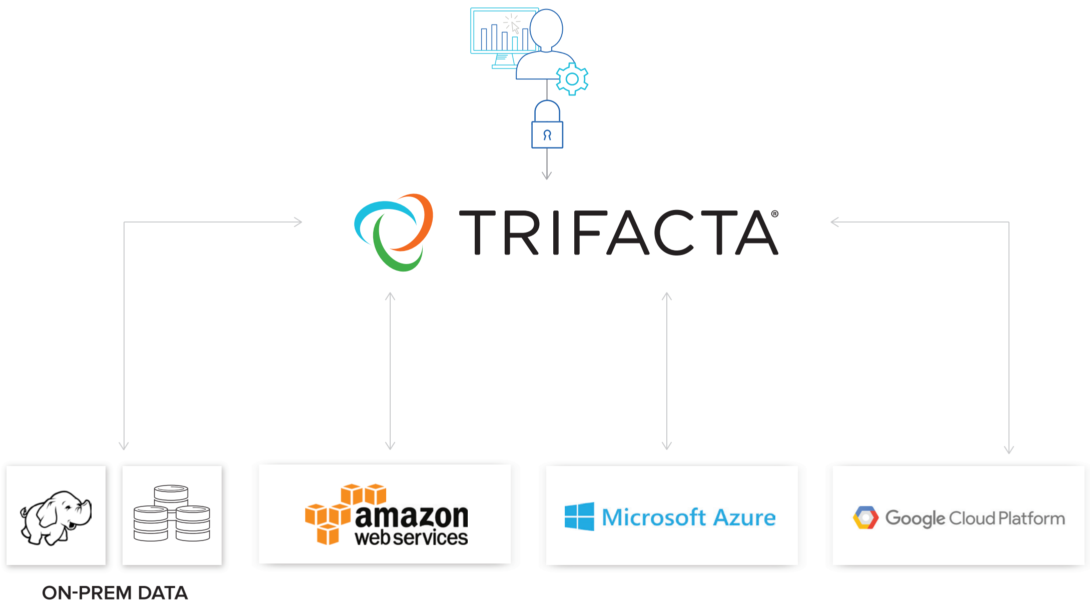
Metadata is "data about the data," and should be consistent across an entire hybrid system. When data comes from different applications, sources, and locations, differing metadata can cause confusion as to how the data should best be leveraged for analysis. An interoperable application allows for a common framework and language for managing metadata.

Uniform Wrangling Logic

Organizations can use a variety of languages to create data wrangling logic—Python, SQL, Visual Basic, or even Excel macros—but when different languages are used, the potential for errors and inaccuracies rises dramatically. When companies are able to prep diverse data across these systems using a single language, preparation logic can be saved, shared, and reused. This accelerates the end-to-end analysis process while ensuring more accurate results.

Common User Experience

Business users need to be able to quickly and easily prepare data for a variety of different business processes. Freeing them and their analysts from extremely slow technical data preparation efforts—as well as empowering them with a user interface that remains consistent across computing environments—leads to a markedly faster analytics process. A first-rate interoperable data preparation product will also present all data, regardless of origin, within a common user interface. Ideally, that same product will enable data democracy so that any business user or analyst can intuitively leverage the application. When everyone can explore and prepare data, teams are empowered to collaborate more effectively and efficiently, further accelerating speed to transformative insight.



Trifacta: Data Wrangling, Everywhere

Trifacta is a data preparation solution built to streamline the process of getting analysis-ready data to stakeholders faster, thereby enabling companies of all sizes to gain competitive advantage from their data. Designed from the ground up to be interoperable with both legacy and modern analytics platforms and tools, both in the cloud and on premises, Trifacta balances self-service for analysts and business users with powerful data exploration and extraction tools. Taking a better approach to data wrangling, Trifacta brings together the latest techniques in data visualization, machine learning, and human-computer interaction, making data preparation—whether in the cloud or on-prem—more intuitive and efficient.

Trifacta Ensures Interoperability Regardless of Environment

Agnostic Processing

Trifacta's intelligent execution architecture supports a variety of different data processing environments whether in-browser, on a desktop, or in different computing environments such as Google Cloud Dataflow, Amazon EMR, and Apache Spark Microsoft Azure HDI. The ability to support a variety of different processing frameworks in different on-prem and cloud environments enables organizations to work with the best-fit engine for their data and workload.

Fast Iteration via Sample Data

Trifacta first uses a smaller data sample subset for crafting initial wrangling logic. Iterations and transforms happen quickly on this representative sample, allowing for fast exploration and experimentation, and is followed by execution later at full scale.

Engaging, Intuitive User Interface

No matter how complex a data source is or how sophisticated the transformation, users can wrangle data consistently across the organization within the same interactive, intuitive visual interface.

Reusable Wrangling Logic

Every wrangling step created and executed in Trifacta is reusable across environments, enabling consistent extraction of insights.

Comprehensive Data Governance

Trifacta provides extensive support for open source and vendor-specific security, metadata management and governance frameworks, providing a grassroots approach to how organizations have visibility and administration over the data wrangling analysts are performing.

A Market Leader In Data Preparation

Analysts at over 7,300 companies in 143 countries utilize Trifacta to more quickly explore and prepare diverse data for a variety of analytic purposes. And Trifacta is routinely recognized as the top data wrangling vendor by leading industry analysts. In the 2017 Forrester Wave on Data Preparation, Trifacta was recognized as the leading vendor in the category. Dresner Advisory Services has ranked Trifacta as the top data preparation vendor over the past three years: 2015, 2016 and 2017.

Google Cloud Dataprep + Trifacta

Trifacta's ability to integrate with cloud infrastructure has been verified by one of the leading technology vendors, Google. Google Cloud Dataprep is a new managed data service built in collaboration with Trifacta, and enables analysts and data scientists to visually explore and prepare data for analysis in seconds within the Google Cloud Platform.

Google Cloud customers leverage this collaboration for their data preparation needs in the cloud, but can also use Trifacta Wrangler Enterprise for their on-premise data wrangling. Google Cloud Dataprep and Trifacta's on-premise solution, Wrangler Enterprise, provide a truly interoperable data preparation solution across cloud and on-premise environments. Customers benefit from consistent transformation logic, user experience, workflow, metadata management, and comprehensive data governance across these environments.

Google Cloud Dataprep In Action: Global Banking

The world's largest international bank is not unlike many other innovative, data-driven Google Cloud Dataprep customers. Even though its corporate strategy is "Cloud First," this financial services powerhouse is heavily regulated, and therefore still relies upon traditional on-premise computing infrastructure including Hadoop.

The bank happily uses Google Cloud Dataprep for workloads in the cloud in addition to Trifacta Wrangler Enterprise for on-prem workloads. Using Trifacta for the on-premise data was an easy and ideal choice for this bank because they could keep their user interface consistent while gaining all the benefits of interoperable data applications—all recipes (data wrangling scripts), metadata, lineage, and UX function across both their Google Cloud Dataprep AND on-prem Trifacta deployment. As a bonus, using Trifacta allows even their most non-technical of analysts to pull and align data consistently across all sources and reuse them across the organization.

After discovering how the Google Cloud and Trifacta combination could work for them, this financial services leader then successfully tackled anti-money-laundering compliance workloads, and has begun to migrate other workloads to Google Cloud Dataprep, including finance liquidity reporting, risk analysis and reporting for complex Monte Carlo simulations, and even valuations services.

Trifacta: Data Preparation Interoperability For Hybrid Environments

To seamlessly operate across multiple computing environments, companies must adopt technologies that address the inherent conflicts that will inevitably arise. Trifacta unifies data across all major data management providers—both in the cloud and on-prem—allowing organizations to tear down data silos, democratize access to business data, and reach critical business insights more quickly. As the data wrangling partner of choice for Google Cloud Dataprep, Trifacta is uniquely suited to meet the interoperability requirements of hybrid cloud and on-prem data systems. Now more than ever, firms looking to modernize and accelerate their analytics practices can increase consistency, transparency, speed, and scale across their data with Trifacta, no matter where the data is located.

Download our free desktop product, Wrangler, to get started with Trifacta today. Email team@trifacta.com for more details on how your organization can implement Trifacta across cloud and on-prem environments.