Self-Service Analytics on the Google Cloud Platform: 5 Data Preparation Lessons Learned to Ensure Success
In today’s dynamic and complex data landscape, traditional data warehouse and business intelligence solutions no longer cut it. Even when you unlock traditional data warehouses with solutions like Tableau or Qlik, you still end up working in spreadsheets and needing IT’s support.

This paper describes five lessons Trifacta has learned from helping organizations like yours adopt self-service analytics. It focuses on the Google Cloud Platform, a comprehensive smart analytic suite that gives analysts the freedom to work with any form of data, at any scale or speed.

The answer is a self-service analytics solution.
When moving analytics to the cloud, many organizations take a lift-and-shift approach—analytics applications are installed on a virtual machine in the cloud. While this approach makes infrastructure maintenance easier, it doesn’t improve the way analytics are delivered, boost their business benefits, or reduce their operational costs. The process and tooling stay the same, and users see no real added value when it comes to self-service.

To be truly self-service, analytics solutions must be cloud-native. Processes for ingesting, storing, wrangling, and reporting on data must be designed to natively integrate with systems built exclusively for the cloud.

They must support cloud environments that are dynamic, elastic, scalable, and increasingly containerized and oriented toward providing microservices. Rather than simply recreate the monolithic data platforms they had on premises, successful organizations embrace cloud agility and use open-source systems, such as Kubernetes or Docker, to orchestrate containers and their data flows.

The Google Cloud Platform is attractive for self-service analytics because it provides a native serverless smart analytics suite. Each component is easily activated, dynamically scales or reduces resource allocation based on usage. Maintenance operations don’t have to be planned. With the freedom to leverage any analytics components, use resources flexibly, and control costs, organizations can focus on the data and the value it provides to their business.
Legacy data integration technologies, such as extract-transform-load (ETL), can be used to build the pipelines to move data stored on premise and in other clouds to the Google Cloud Platform. Alternatively, new cloud-dedicated data integration and streaming technologies, such as StreamSets, Fivetran, or Google Cloud Data Fusion, are effective in flowing data between systems and making it available for analytics in the Google Cloud Platform.

It may be tempting to continue using ETL after a data lake or data warehouse has been filled, but ETL is just not suitable for self-service analytics. It’s quite a technical and complex technology aimed at data engineers who work with well defined, recurring real-time and batch transformations for large-scale data movements. For business data-savvy professionals to access and use data in their analytics, extra data preparation is needed.
As part of its smart analytics suite, the Google Cloud Platform offers a data preparation solution: **Cloud Dataprep by Trifacta**. It assesses data quality; refines, standardizes, and cleanses data; and combines data and handles various data calculations.

**Cloud Dataprep by Trifacta** is designed to help a whole host of data professionals—data engineers, data analysts, business analysts, and other data-driven professionals—unlock a data lake or data warehouse by interacting with the content of the data to iteratively refine it and bring it together to feed downstream business-driven analytics.
Lesson #3: Self-Service is Not a Free-For-All

When more people can use self-service analytics, more people will self-service analytics. But that doesn’t mean it’s a data free-for-all. Empowering more users with self-service calls for careful governance to:

- Keep data from proliferating out of control
- Comply with regulatory requirements
- Maintain trust in the data used to drive business decisions

It’s important to strike the right balance between protecting data assets and enabling users to collaborate and derive value from data as they see fit.
Organizations need to strike the right balance between protecting data assets (from governance and security perspectives) and enabling users to collaborate and derive value from data as they see fit. There are three ways to achieve this balance.

**Reduce Silos**
When data storage and processing are centralized in the cloud with virtually unlimited scalability, and when end users are authorized to bring their own data in the cloud, you stop data silos from proliferating. Users collect data extracts, run their own preparation routines, and create their reports in and from the cloud instead of extracting data and duplicating it in spreadsheets.

**Use a Data Catalog**
Shared resources, like a central catalog or glossary, that manage data definitions, metadata, and knowledge about the data’s lineage, help users find data faster and enable organizations to govern data sources and monitor its lifecycle. Machine learning (ML) and artificial intelligence (AI) solutions automate the collection and management of metadata and related knowledge about the data.

**Track and Document Data Lineage**
Data lineage—that is, how data has been used and transformed and by whom—is important for regulatory reporting and audits. It’s also important to decision-makers, who need to understand the history of the data behind analytics, visualizations, and prescriptive recommendations, as well as the impact of new requirements on the data pipeline to produce the analysis.
Self-service data preparation invites organizations to review the roles and responsibilities of key stakeholders in the data analytics value chain and reimagine the rigid data pipeline process as a more flexible and agile flow. Each stakeholder leverages self-service data preparation differently for their needs while adopting a common design, communication, and collaboration framework to deliver analytics.

**DATA ANALYSTS**, in addition to business analysts, project managers, and related positions, are typically in close contact with business users. They often explore raw data to discover what may be useful to answering business questions. Their goal is to get answers as quickly and easily as possible.

**Cloud Dataprep by Trifacta** is ideal for data analysts because it’s easy to implement and doesn’t require significant technical expertise in coding. Data analysts can be involved early in the data preparation process, exploring and prototyping data to fit their business needs. When automation is required for sustainability and repeatability, they can collaborate with data engineers to orchestrate the end-to-end data pipeline.

**Lesson #4: Review the Roles and Responsibilities**

Self-service data preparation presents new ways for data engineers, data analysts, data scientists, data architects, and analytics-focused business executives and managers to collaborate with each other.
DATA ENGINEERS design, build, and manage data processing and data architecture to support analytics and data science. They’re closely involved in transforming data, including the exploration and profiling of raw data. Chief among data engineers’ goals is streamlining and automating data-related processes so they can manage more of them.

Cloud Dataprep by Trifacta is ideal for data engineers because they can operationalize and monitor the various data flows they or data analysts design. It makes it easy for data engineers to collaborate with all stakeholders to understand data infrastructure requirements and provide guidance to users to improve how they explore, analyze, model, and consume data.

DATA SCIENTISTS apply specialized knowledge and skills to design and model algorithms by leveraging machine learning and artificial intelligence. But up to 80 percent of their time is consumed with routine data preparation tasks, leaving little time for innovation.

Cloud Dataprep by Trifacta is ideal for data scientists because it’s a cloud-native data preparation solution. It simplifies routine data preparation tasks, allowing them to be delegated to more readily available and less expensive resources.
All stakeholders need to know they can trust the data that drives their analytics, visualizations, and algorithms. The outcome of any predictive model or analytic insight is only as good as the data that feeds it. Every organization should commit to making clean data a priority.

Every organization should commit to making clean data a priority.

While it would be ideal if data quality could be addressed at a corporate level, it’s unrealistic. The volume of data and the enormous variety of data sources (both internal and third-party), types, and context make it difficult to evaluate data quality, correct flaws, and monitor new data for quality problems. New data repositories, such as cloud data lakes and modern data warehouses, present more complex data quality challenges.

Addressing data quality at the data preparation stage is more effective. Cloud Dataprep by Trifacta improves the accuracy, consistency, and completeness of data by applying ML and AI to automate data cleansing procedures. Automation handles the scale of very large data repositories and quickly identifies data values that appear to be incorrect, invalid, missing, or mismatched. Outliers that merit closer inspection are automatically flagged.

Automated data profiling and cleansing routines spot inconsistencies across sources being integrated into a cloud data storage, highlight probable data duplication, and recommend how to correct data quality problems visually through code-free, automated transformations.

As new structured, unstructured, or semistructured data is ingested and integrated into the cloud, data quality is continuously validated. “Continuous validation” means that users don’t have to wait until the end of the validation process to view and test the results, a delay that’s incompatible with today’s agile development methodologies.
Conclusion

Trifacta has helped many organizations succeed with self-service analytics in the cloud. From this experience, we’ve learned five key lessons:

1. Using a simple lift-and-shift approach to move on-prem analytics to the cloud doesn’t work. Select a cloud platform that offers analytics components designed specifically for the cloud, like Google Cloud Platform with its comprehensive native serverless smart analytics suite.

2. ETL technologies can smooth the friction resulting from moving on-prem data to cloud repositories, but ETL is too technical and inflexible for self-service analytics. A better alternative is to adopt self-service data preparation solutions, such as Cloud Dataprep by Trifacta, to empower all business and technical users to leverage the data they need for analytics.

3. Self-service isn’t a lawless free-for-all zone. For self-service analytics to be successful, organizations must strike the right balance of careful governance and openness. Understanding data lineage and using a business data glossary are critical to success.

4. Adopt new agile processes that highlight collaboration and a common language among all stakeholders to transform data into a refined asset for their analytics.

5. Data quality is everyone’s responsibility. Cloud Dataprep by Trifacta offers automatic data quality assessment and continuous validation and resolution to convey trust in the data used for self-service analytics.
Ready to Learn More?

With seamless data preparation across any cloud, hybrid or multi-cloud environment, Trifacta Wrangler is the ideal self-service data preparation solution for the Google Cloud Platform.

Schedule a demo and learn more about our Google partnership.

575 Market St, 11th Floor
San Francisco, CA 94105
1 844 332 2821
www.trifacta.com

Follow Trifacta

@Trifacta  Trifacta  Trifacta

Google

450 Concar Drive
San Mateo, CA 94402
www.cloud.google.com