Enterprise Strategy Group | **Getting to the bigger truth.™**

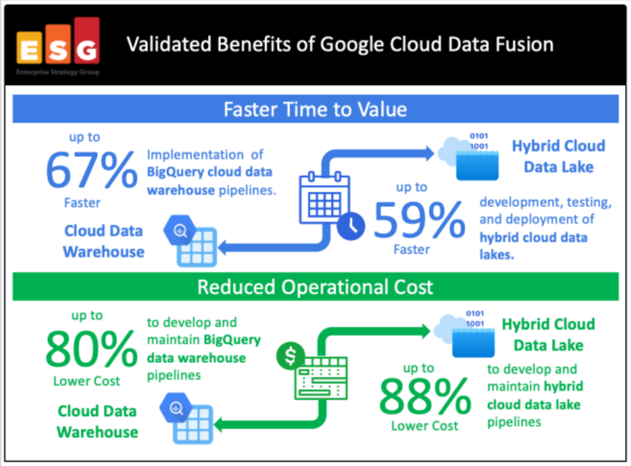
ESG Economic Validation

The Economic Benefits of Google Cloud Data Fusion

By Aviv Kaufmann, Senior Validation Analyst

May 2021

Executive Summary



Validated Benefits of Google Cloud Data Fusion

Faster Time to Value

up to 67% Faster Implementation of **BigQuery cloud data warehouse** pipelines.

Cloud Data Warehouse

Hybrid Cloud Data Lake

up to 59% Faster development, testing, and deployment of **hybrid cloud data lakes.**

Reduced Operational Cost

up to 80% Lower Cost to develop and maintain **BigQuery data warehouse** pipelines

Cloud Data Warehouse

Hybrid Cloud Data Lake

up to 88% Lower Cost to develop and maintain **hybrid cloud data lake** pipelines

In today's data-driven environment, organizations need to use all of the data sources available to them to extract timely and actionable insights. Instead of wrestling with the design, testing, and remediation of complex data pipelines, however, organizations should spend the time providing analytics functionality to those who need to make decisions.

ESG validated that the fully managed, cloud-native Google Cloud Data Fusion service, which is powered by open source Cask Data Application Platform (CDAP), allows organizations to build and manage reusable data pipelines with code-free efficiency and speed. Integration with Google Cloud enables subscribers to take advantage of data mobility spanning on-premises, hybrid cloud, and multi-cloud to improve business-wide collaboration. The unified platform allows all services to have a common layer for monitoring, management, and security, so organizations can operationalize data faster while improving visibility, compliance, and control.

ESG's modeled scenarios for structured and unstructured data predict that organizations can capture substantial savings and accelerate time to value compared to alternative solutions. Up to 80% savings were predicted for organizations to deploy, manage, and maintain data pipelines for cloud-based enterprise data warehouses in BigQuery, and up to 88% savings were expected to operate a hybrid cloud data lake into which siloed on-premises platforms generating data across multiple markets were sanitized, normalized, secured, and integrated.

Introduction

This ESG Economic Validation is focused on the quantitative and qualitative benefits, such as operational savings, fast time to value, and other savings, that organizations can expect from unifying data silos, integrating disparate data, and enabling code-free deployment of ETL/ELT data pipelines.

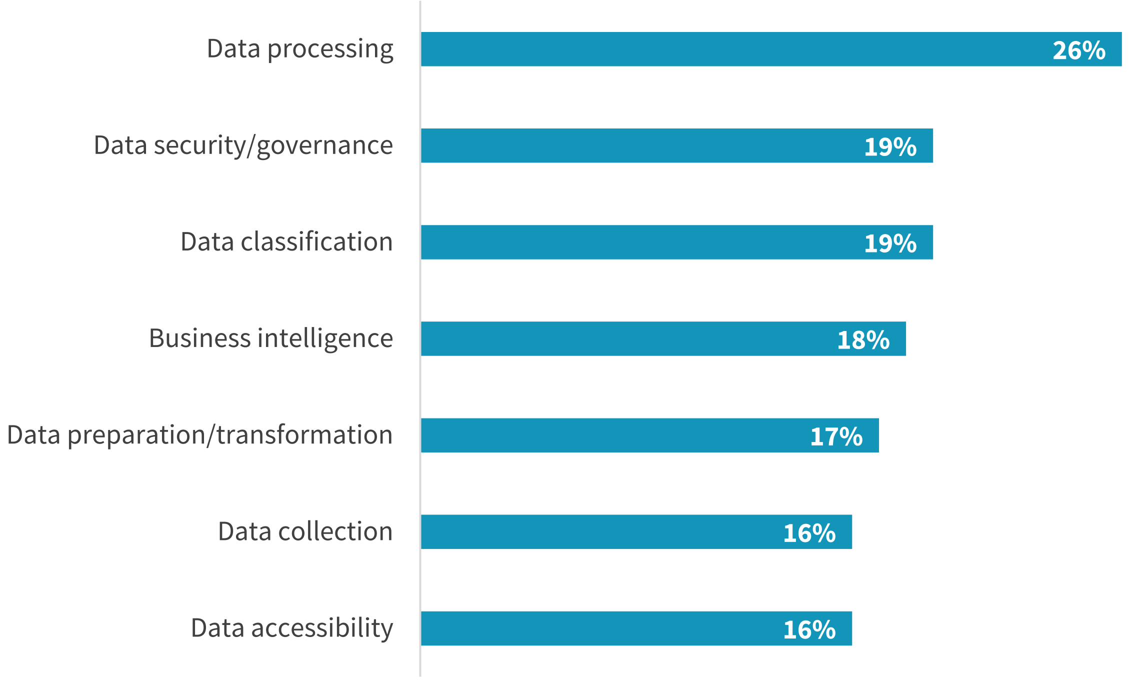
Challenges

Comprehensive near-real-time analytics depend on the integration of diverse data sources, but data integration is complex. Data pipelines typically are built by a few specialized resources who write and manage the necessary scripts and code. The process takes a lot of time, relying on siloed resources required to create, debug, test, update, and maintain pipelines. Development teams may "do their own thing" with little or no collaboration using different platforms, which doesn't bode well for scalability. The inevitable changes to source APIs must be handled by the same skilled resources who are backlogged and unable to respond quickly. Also, legacy databases and data warehouses may be straining to meet performance or usability requirements. Wherever and whenever the analytics process breaks down, data use is restricted, reports are delayed, users are unhappy, and decisions reflect aging or incomplete data.

ESG research into data analytics identified several reasons for data pipeline delays (see Figure 1). Twenty-six percent of respondents named data processing as a top factor, but other significant contributors were data security/governance, data preparation/transformation, and data collection and accessibility.[1] The average analytics exercise involves integrating six disparate data sources,[2] so pipeline challenges escalate when developers in different countries or regions write unique scripts. Additionally, survey respondents weighed in on the most time-consuming aspects of integration, citing understanding data due to lack of metadata (16%) and combining on-premises and cloud data sets (16%) as the top two.[3]

Figure 1. Top Seven Contributors to Data Pipeline Delays

What aspects of your data pipeline are most frequently responsible for causing delays? (Percent of respondents, N=310, three responses accepted)



Source: Enterprise Strategy Group

[1] Source: ESG Master Survey Results, *The State of Data Analytics,* August 2019.
[2] Ibid.
[3] Ibid.

Cloud-based data warehouses and data lakes provide significant benefits for modern data analytics strategies. But data integration into hybrid cloud- and multi-cloud-based analytics solutions is viewed mostly from a functional, not a cost, perspective. It is important to understand the resources, time, and costs involved to recognize the value in simplifying the design, testing, and remediation of complex data pipelines.

The Solution: Google Cloud Data Fusion

A fully managed, cloud-native, code-free data integration service, Google Cloud Data Fusion contrasts sharply with solutions designed for on-premises use and ported to work in the cloud—ported solutions do not function as well. By removing common data integration complexities and bottlenecks and reducing dependence on technical expertise, Cloud Data Fusion simplifies and speeds near-real-time analytics. The service unifies disparate data silos and allows organizations to transform and map data by building, debugging, and testing complex data pipelines in hybrid or multi-cloud environments with enterprise-level security and per-pipeline, rules-based governance. Everyone, regardless of location, works with data that is consolidated, standardized, and anonymized.

A unified platform on Google Cloud, Cloud Data Fusion allows all services to have a common layer for monitoring, management, and security. All delivery styles are enabled in the platform, helping streamline operations and minimize issues. Alternative solutions often support one delivery style at a time—with unique scripts for each service on each cloud and with separate monitoring, management, and security capabilities.

Additionally, Cloud Data Fusion is fully enabled with microservices and APIs to support modern containerized applications. All services have APIs for lifecycle management, ensuring seamless, fast implementation of CI/CD strategies. Solutions without microservices and APIs will require more effort and expense to work with modern platforms and may not perform as well.

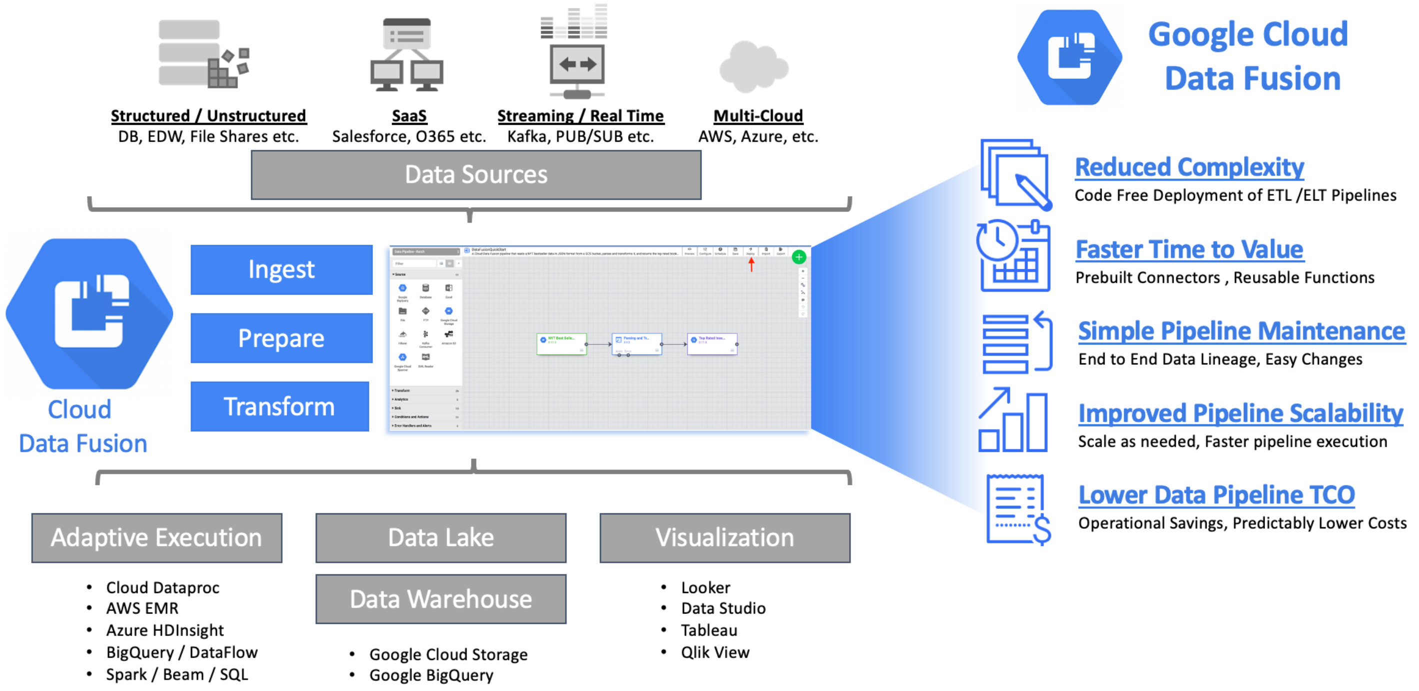
The Cloud Data Fusion dashboard provides centralized visibility and control of code-free pipeline activities. The intuitive GUI allows users to connect desired items in a visual list of taps and sinks with drag-and-drop ease to complete extract/ingest, transform, and load steps. Cloud Data Fusion can ingest data directly from databases, on-premises applications, SaaS and streaming applications, cloud services, mobile applications, sensors, Pub/Sub, and other sources.

Key features of Cloud Data Fusion include:

- Extensibility via templates and a library of 150+ prebuilt, preconfigured connectors and transformations, suitable for both batch and real-time processing—all at no additional cost.
- Efficient execution of pipelines on ephemeral instances of Cloud Dataproc and native integration with best-in-class Google Cloud services streamlines the use of tools such as BigQuery, Hadoop, and Dataproc.
- End-to-end data lineage and integrated metadata simplify root cause, impact analysis, and provenance.
- Integration with Cloud Composer, which organizations use to automate and orchestrate the entire end-to-end data lifecycle from a source to a staging area, through transformations, to archiving or retiring data.
- Pipeline reusability supports standardized analytics and metrics.
- Pipeline portability, based on the Cask Data Application Platform (CDAP) open source core, provides freedom of choice.

Cloud Data Fusion also comes with an established, growing CDAP community focused on data integration. Users help other users, submit ideas for features or improvements, review code, and engage in other ways. An active, collaborative community benefits all members by expanding freedom of choice, supporting multi-cloud, easing third-party application integration, and decreasing the amount of time and money organizations spend on training and/or hired expertise.

Figure 2. Google Cloud Data Fusion



Source: Enterprise Strategy Group

ESG Economic Validation

ESG's Economic Validation process is a proven method for understanding, validating, quantifying, and modeling the economic value propositions of a product or solution. The process leverages ESG's core competencies in market and industry analysis, forward-looking research, and technical/economic validation. For this analysis, ESG interviewed end-users, viewed product demos, and studied Google case studies to better understand and quantify how Cloud Data Fusion has impacted organizations, particularly in comparison with previously deployed solutions.

Google Cloud Data Fusion Economic Overview

ESG's economic analysis revealed that Cloud Data Fusion provided its customers with economic benefits in the following categories:

- Operational savings—Organizations decreased the number of hours required to design, build, test, maintain, and troubleshoot data pipelines and found they could achieve results using resources with fewer specialized skills.
- Faster time to value/insight—Removing technical bottlenecks enabled earlier integration of data pipelines, greater availability of data, and self-service capability. These outcomes contributed to greater business flexibility and agility.
- Other savings —Subscribers avoided large upfront investments, reduced risk, improved compliance, and increased collaboration with greater visibility, accuracy, and consistency.

Operational Savings

Cloud Data Fusion delivers operational savings based largely on simplicity and reusability. According to one customer, “With Cloud Data Fusion, we are able to run with fewer people and fewer consultants but with more capabilities and faster responsiveness.”

- **Fewer labor hours required—**ESG validated that organizations using Cloud Data Fusion obtained up to 75% labor savings. Customers were able to “just spin up Cloud Data Fusion instances.” One customer reported that a five-person team relying on other tools moved only 20% of the data that one Data Fusion engineer was able to move. Another customer pulled in data from a SQL relational database using default configurations in one-half the time it took with the previous solution. Customers also commented on reusability: “…able to reuse what we invested…to accelerate development time for other teams to create new pipelines—they don’t have to start from scratch." And “reusable components allow us to build once and deploy many—this was our vision." One organization built a plugin in one week and then customized and used it in seven countries within the next week. Pipeline reusability also enabled organizations to apply the same metrics consistently across the business.

> ***“We had work planned for 12 people, but with Cloud Data Fusion, we are able to run with 3, for 75% operational savings.”***

> ***"It is amazing…I have not seen any [other data integration] tool with this level of reusability."***

- **Less time spent maintaining and troubleshooting—**ESG validated up to 40% time savings for developers and up to 60% faster application operations as a result of faster identification of root causes, improved visibility, the ability to test preproduction pipelines, simpler changes and fixes, and fewer errors. The Cloud Data Fusion GUI, prebuilt connectors and transformations, among other Cloud Data Fusion features, were highlighted for simplifying pipeline deployment and operation. DIY connectors, in comparison, did not always follow best practices or have complete documentation, and this caused issues during pipeline use.

  A customer in a large service organization said that pipeline testing allowed engineers to document best practices and commented that changes could be deployed mid-sprint and in as little as a week. Another customer reported no outages since July of last year, and yet another reported that Cloud Data Fusion “has been so stable running that I don’t have to hire another specialist to monitor and address any issues with the source.” The Dataproc cluster proved easy to scale, according to a customer who verified that all issues (bumping up VMs, maintenance, and cleaning) related to the prior solution went away. Another customer commented on a positive benefit of Cloud Data Fusion notifications, which triggered an ITIL ticket and enabled business users to be notified of an issue before they reported problems seeing the data.

- Reduction in specialized skills—ESG verified that the use of Cloud Data Fusion enabled organizations to use less-experienced resources to build, deploy, and manage pipelines. Currently, data engineers, data scientists, and developers each do their part to design, construct, and maintain pipelines. The task of building complex connectors can take weeks. With Cloud Data Fusion, one customer went from "seven or eight data engineers to three plus one QA engineer," and others found they did not need to engage professional services or contractors. ESG also found that organizations lowered training costs going code-free with Cloud Data Fusion. Ultimately, customers were able to bring analytics capabilities to more departments in less time, less expensively.

"Cloud Data Fusion allows us to transfer skills from senior to junior data engineers. Best practices can be created into templates by senior engineers and then leveraged by junior data engineers for repeatable patterns."

Faster Time to Value/Insight

Fast time to value is all about getting things done more quickly, from creating pipelines to running analytics. With Cloud Data Fusion, customers experienced fewer bottlenecks and accelerated desired outcomes such as earlier revenue recognition or the elimination of negative impacts on revenue.

- Earlier integration of data pipelines—ESG found that earlier integration shortened the time to insights, accelerated decisions, and enabled greater scale to improve revenue. In one case, a customer stated that data availability went from weeks to minutes. Another commented, "We were able to get up and running in one month." One organization went to market in one-quarter the time in one country compared to the go-to-market experience in a different country using the previous solution.

- Operationalization of data—ESG validated increased customer productivity and collaboration. One organization completed sprints faster, saving time by 33% to 66%, and noted the ability to deploy mid-sprint. Additionally, data was pulled in less time, 1.5 hours compared to 3 to 4 hours. Another customer was able to "set new expectations around how fast we can accommodate changes to reports and accelerated the time to insight." KPIs could be added or changed rapidly.

"Not only does Cloud Data Fusion help bring our data to the cloud, but once there it makes integrating with other services very easy."

- Quicker remediation of data pipeline issues—ESG confirmed that customers identified pipeline issues faster through alerts and indicators like red text. Issues could be fixed prior to testing or production to ensure that everything would work smoothly. Fewer breaks or changes caused less impact on the business in terms of interrupted and delayed analytics and insights. One user noted that a one-hour fix with Cloud Data Fusion could have taken a day with the previous solution.

- Self-service capability—ESG found that the self-service model allowed more groups to understand more of the pipeline and essentially train themselves. Business analysts, for example, were able to go into a pipeline, pull a field, and change an item on their own. This enabled organizations to create data pipelines and bring them to new areas of a business without the need to bring in engineers and developers.

- **Improved flexibility and business agility**—ESG validated reduced time to market by up to 75% and confirmed other aspects of flexibility and agility. Alternative solutions restricted organizations through high prices and/or limited functionality. For example, tools built for on-premises environments and ported to the cloud produced mixed results, and tools built for business intelligence purposes were inflexible and difficult to use. The cloud-native and open source qualities of Cloud Data Fusion simplified customization and modification, specifically through the use of built-in REST APIs. ESG confirmed that customers created their own connectors and found it easy to customize Cloud Data Fusion compared to alternative solutions. One customer observed that because Cloud Data Fusion is open source, “We are free to move to other platforms or on-premises, if needed, which was quite important.”

  "Once we create the pipeline, it runs very nicely…getting from development to production is near instantaneous.”

  Customers commented on Google’s ongoing efforts to expand functionality and choice by creating new data sources and services and by integrating with other Google and non-Google cloud services. Every customer, whether subscribed to the basic or enterprise edition, has access to all Cloud Data Fusion features for unlimited users. As a result, organizations experienced fewer obstacles to accomplishing their business objectives.

Other Savings

Other savings contribute lasting benefits and value. Cloud Data Fusion subscribers received best-practice guidelines for integration that promote standardization and reusability. With no vendor lock-in, organizations retained freedom of choice.

- **Cost savings—**ESG confirmed that the subscription model eliminated large upfront investments by customers in software licenses, servers, and training. No extra charges were incurred for data preparation and transformations, which are performed by the Wrangler tool, or for the prebuilt data connectors. The all-inclusive monthly charge covered all of Cloud Data Fusion except for the cost of execution on Dataproc (ephemeral Dataproc instances greatly helped to keep costs low, as they were only charged when running Dataproc instead of 24x7). Storage also was a separate charge but one likely to be similar across solutions. One customer commented, “With other tools, anything you want to do, you need to pay. With Cloud Data Fusion, you pay once, and you have access to all of the tools and abilities. New features and capabilities are being added, and they are all included. Other tools make you pay for absolutely everything."

  “With Cloud Data Fusion, you pay once, and you have access to all of the tools and abilities…other tools make you pay for absolutely everything."

- **Lower risk to the organization—**ESG verified with customers that the fully managed, cloud-native architecture lowered risk. One praised the service for “unlocking the scalability, reliability, security, and privacy features of Google Cloud.” Another user stated that “security was a non-negotiable requirement, and Cloud Data Fusion covered everything we were looking for from a security perspective.” Other customers pointed out that they decreased risk by eliminating DIY scripting and errors, and “in one year running Cloud Data Fusion pipelines, we have had only two production issues and were able to recover in a day." Other stated benefits were a reduction in the amount of code written for ELT or ETL and greater consistency among engineering teams. ESG also found that businesses lowered risk by improving governance through use of the data lineage capabilities.

- Improved collaboration—ESG found that one organization that developed 130 pipelines completed knowledge transfer to another team in about one month, and that team wrote an additional 100 pipelines with minimal supervision. A customer observed that "Cloud Data Fusion offers the ability to create an internal library of custom connections and transformations that can be validated, shared, and reused across an organization." Pipeline reusability decreased development costs because each effort could be used by others to improve analytics capabilities—a finding important to organizations that expect to add or change personas.

"Data lineage helps to identify where information comes from and what transformations do and why, and it helps identify issues and remediate them faster. "

- Improved compliance, visibility, and control— ESG validated that the data lineage capabilities of Cloud Data Fusion met or exceeded customer expectations—especially for tracing end-to-end lineage with different engines using data that is sanitized, standardized, and checked for compliance (such as stripping out personally identifiable information). One customer reported that classic ETL tools could not provide the level of lineage that Cloud Data Fusion provided. The integrated business, operational, and technical metadata provided superior intelligence and insights, which enabled issues to be identified and remediated rapidly. Additionally, alerts and visual indicators in Cloud Data Fusion allowed customers to see problems and resolve them rapidly with fewer tickets. The ability to provide more data, via screenshots, to those in charge of a pipeline sped troubleshooting and resolution. The dashboard centralized control and provided consistent views to everyone.

"Data is sanitized, standardized, and checked for compliance."

ESG Analysis

ESG leveraged the information collected through vendor-provided material, public and industry knowledge of economics and technologies, and the results of customer interviews to create two modeled scenarios that compare the costs and benefits of solving cloud data integration challenges with Google Cloud Data Fusion against building hand-coded solutions and leveraging alternative data integration tools first designed for on-premises environments. ESG's interviews with customers who have recently made the transition to Cloud Data Fusion, combined with experience and expertise in economic modeling and technical validation of the solution, helped to form the basis for our modeled scenarios.

Scenario #1: Creating a BigQuery Data Warehouse

The first scenario compared the expected time and costs to develop and manage data pipelines for a cloud-based EDW on BigQuery over a one-year period. ESG assumed that a large enterprise organization was looking to create pipelines to bring in data originating from ten on-premises and/or cloud data sources with a mix of nightly run batch pipeline executions and real-time stream ingestions.

ESG assumed a team size of 9 developers would be required to build a DIY solution consisting of a series of hand-coded connectors, open source APIs, SQL and Python transformation scripts, etc. ESG assumed that by leveraging a data integration service, a team of only 4 developers would be required to develop and maintain the same number of pipelines. ESG also assumed a modest 8% lower average rate for those using a legacy data integration solution and a 15% lower rate for those using Google Cloud Data Fusion to reflect the increased level of expertise required to operate the development-intensive and proprietary tools.

Leveraging ratios validated through customer interviews, ESG assumed that a new data pipeline would take three full sprints of two weeks each (6 weeks total) to hand code, test, and troubleshoot with the DIY solution. ESG learned that this time could be reduced by half (3 weeks) by using an alternative data integration tool, with slightly more time required to integrate real-time and cloud-based data sources and to troubleshoot issues. By leveraging Cloud Data Fusion's prebuilt connectors, code-free interface, data wrangling capabilities, end-to-end data lineage, and metadata integration, ESG found that pipelines could be created, tested, and tuned in a single sprint of 2 weeks. This 67% improvement over a DIY implementation provides a faster time to value that has an immediate impact on both operations and revenue generation.

ESG similarly modeled the time required to maintain pipelines by making periodic changes to them (ESG assumed 4 annual major changes or enhancements per pipeline, each requiring 30% of the team's resources) as well as the time to identify and remediate issues (ESG assumed 6 annual major issues per pipeline, each consuming 30% of the team's resources). ESG found that major changes to the pipeline required an effort of about two full sprints (4 weeks) in the DIY case, a single sprint (2 weeks) using the alternative DI tool, and only a single week with Cloud Data Fusion, resulting in 50% to 75% faster implementation of changes. Similarly, major issues often took a full day or more with a DIY solution to remediate but could be identified and remediated in only 2 hours with an alternative DI tool and in less than 1 hour with Cloud Data Fusion. This results in up to 88% faster remediation of pipeline issues and greatly improves the stability of data pipelines.

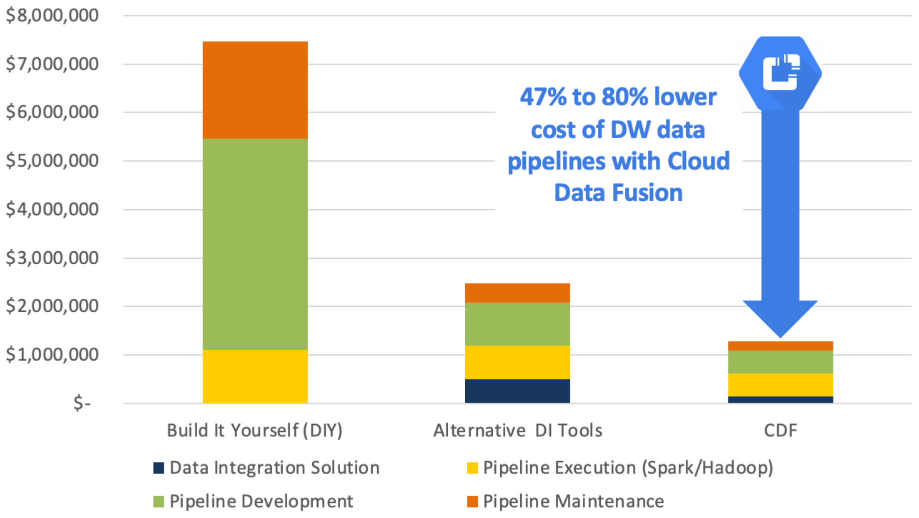
Why This Matters

Extracting, standardizing, and sanitizing data for use by applications and end-users is a complex and time-consuming operation that requires specialized expertise and coding skills.

ESG validated that Cloud Data Fusion can help organizations simplify the task of creating, maintaining, and troubleshooting complex data pipelines, allowing organizations to deliver pipelines faster and with less interruption. This enables organizations to deliver improved insight by making a greater volume and variety of data available to end-users and applications at a greater velocity.

ESG modeled out the expected price of the solutions based on known and published pricing, customer-reported spending ratios, and previously validated savings. While the cost of the alternative data integration solution can skyrocket very quickly based on licensing features, organizational requirements, and number of data connectors required (ESG validated annual spending as high as 12x versus Cloud Data Fusion), we conservatively priced the solution for this analysis. Finally, ESG calculated the expected cost to execute pipelines on a mix of 24x7 and batch job instances on on-premises or cloud-based Hadoop instances (for DIY) and cloud PaaS solutions like Cloud Dataproc. Our modeled analysis (shown in Figure 3) resulted in an 80% lower cost to create, manage, and maintain data warehouse pipelines compared to a build-it-yourself approach and a 47% lower cost when compared to alternative data integration solutions.

Figure 3. Modeled Cost to Develop, Manage, and Maintain Data Pipelines for a BigQuery Data Warehouse



47% to 80% lower cost of DW data pipelines with Cloud Data Fusion

up to 91% lower cost to maintain pipelines

up to 87% lower cost to develop pipelines

up to 57% lower pipeline execution cost

up to 71% lower cost of tools/services

Source: Enterprise Strategy Group

<u>What the Numbers Mean:</u>

- **Faster Time to Value:** Cloud Data Fusion allows organizations to begin creating work on the data warehouse sooner without the need for upfront investment in expertise, hardware, software licenses, or expensive connectors.
- **Improved Data Warehouse Velocity and Agility:** Cloud Data Fusion reduces the time and number of resources required to develop and maintain pipelines, allowing organizations to begin realizing value from their BigQuery DW earlier, add data from a greater variety of sources (including real-time), execute batch pipelines faster, make changes to existing pipelines and KPIs, and add application capabilities quicker and easier.
- **Improved Data Warehouse Integrity:** Cloud Data Fusion improves security and compliance, reduces the number of pipeline issues that occur, and speeds the time to identification and remediation of any issues that do arise, enabling greater availability of data pipelines.

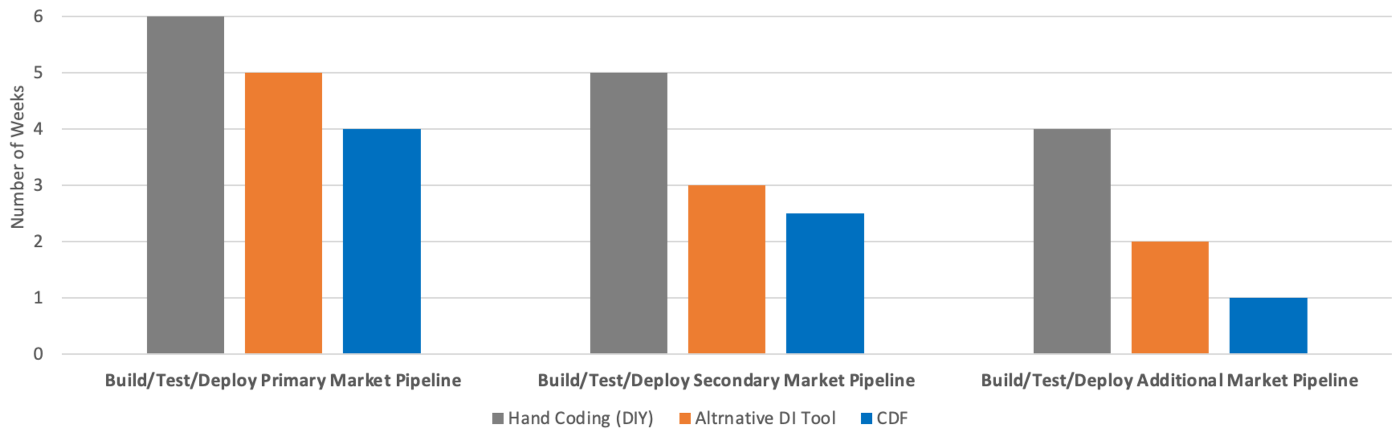
Scenario #2: Operating a Hybrid Cloud Data Lake

The second scenario compared the expected costs for a distributed organization to integrate data from siloed on-premises platforms across many geographies into a normalized, scalable, and centralized hybrid cloud data lake to leverage insight across the organization.

ESG assumed that a large enterprise organization operated across five global regions that previously operated independently, collecting, storing, and analyzing a variety of data across regionalized data centers. Each data center controlled its own set of disparate data, and gaining global insight was a difficult and time-consuming task. By normalizing and standardizing data into a centralized cloud data lake, it could be leveraged for insight by corporate resources, applications, and customers across the globe.

Based on validation with customers, we assumed that the organization would use a phased process to first bring in the most important sources in key markets and then bring in additional markets upon successful completion. With each successful implementation, lessons learned and reusable components would make subsequent markets faster to bring in. This is especially true with Cloud Data Fusion, where customers reported significant reusability of components for almost all tasks related to the pipeline. Leveraging what we learned from customer experiences, ESG assumed the time that would be required to build initial market, secondary market, and subsequent market pipelines, as shown in Figure 4.

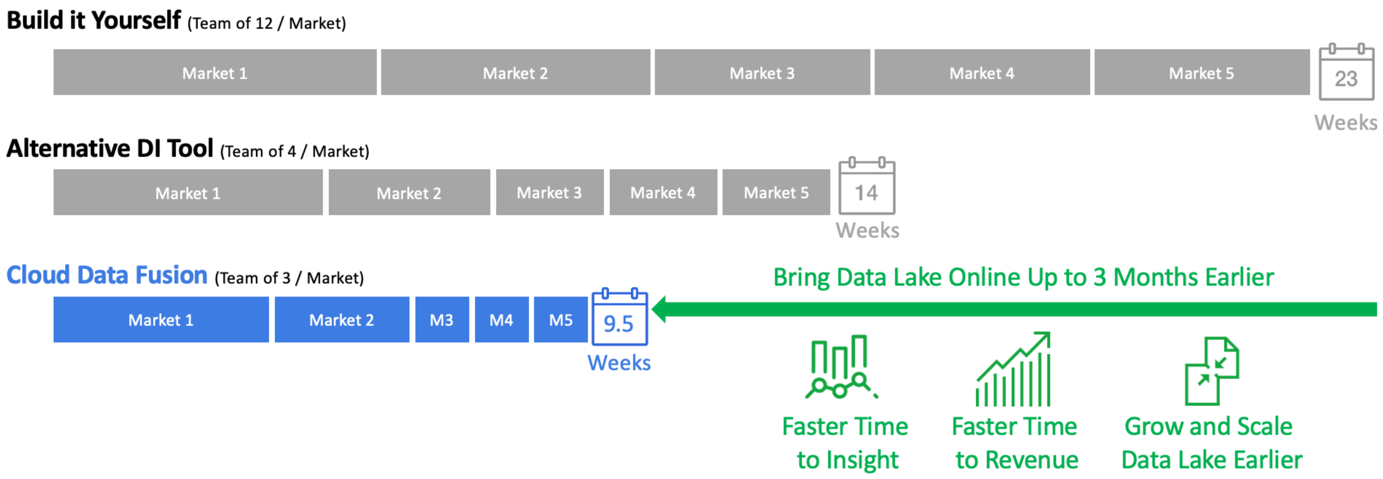
Figure 4. Modeled Time to Develop, Test, and Deploy Data Lake Pipelines for Initial and Subsequent Markets



Source: Enterprise Strategy Group

Leveraging this concept, ESG modeled the time required to develop, test, deploy, and migrate the critical integrations for each of the five markets into the data lake using a phased approach similar to that used by the customers we spoke with. Once all five markets had critical data and pipelines in place, the data lake would be ready for use for the most important cases. Additional sources would be added over time as the data lake capabilities scaled and grew. ESG found that the organization's data lake could be functioning and realizing value up to 3 months earlier using Cloud Data Fusion (see Figure 5).

Figure 5. Modeled Organization Time to Initial Hybrid Cloud Data Lake Functionality across Five Markets

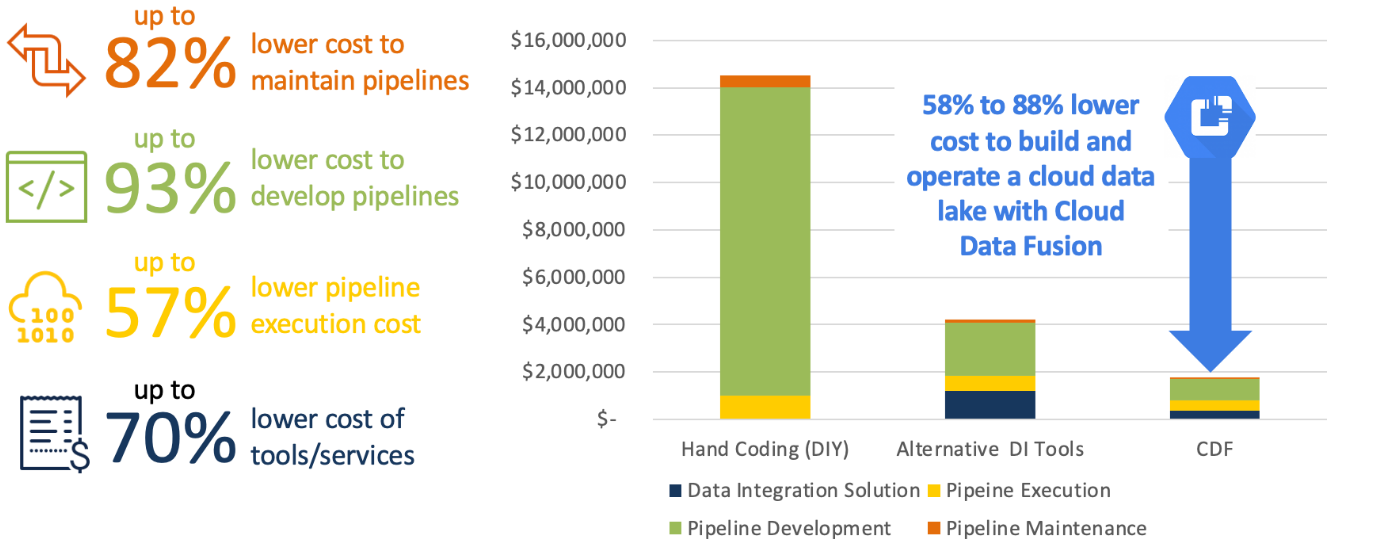

Build it Yourself (Team of 12 / Market)
Market 1
Market 2
Market 3
Market 4
Market 5
23
Weeks
Alternative DI Tool (Team of 4 / Market)
Market 1
Market 2
Market 3
Market 4
Market 5
14
Weeks
Cloud Data Fusion (Team of 3 / Market)
Market 1
Market 2
M3
M4
M5
9.5
Weeks
Bring Data Lake Online Up to 3 Months Earlier
Faster Time to Insight
Faster Time to Revenue
Grow and Scale Data Lake Earlier

Source: Enterprise Strategy Group

Based on the diversity of expertise required, ESG assumed a team size per region of 12 for build-it-yourself, 4 for alternative tools, and 3 for Cloud Data Fusion implementations. In addition, because the larger team sizes would be made up of serialized and siloed activities, ESG built in a productivity utilization adjustment factor of 30% for DIY, 80% for alternative DI solutions, and 90% for Cloud Data Fusion to accurately estimate the active person-hours required over the calendar time. Then using a similar modelling approach to the one described in scenario 1, we modeled the expected costs to build, test, and deploy the initial two major integrations for each market to create the initial data lake and then incrementally add five additional new data sources per region over a two-year period. In addition, ESG modeled the expected cost to maintain and make changes as needed to these pipelines, as well as to troubleshoot and remediate issues.

Our model predicted that building and operating the cloud data lake with Cloud Data Fusion resulted in a two-year total cost savings of 58% when compared to alternative data integration solutions and 88% when compared to build-it-yourself strategies consisting of hand-coding pipelines. In addition to a 70% lower cost of tools and services, Cloud Data Fusion resulted in a 59% lower cost of pipeline development, a 35% lower cost of pipeline maintenance, and a 32% lower cost of pipeline execution compared to the alternative DI solution. While both data integration solutions provided significant savings over hand coding, the majority of the development and maintenance savings resulted from Cloud Data Fusion's ability to reuse components across markets, provide data lineage and integrated metadata, and simplicity. When compared to DIY solutions, Cloud Data Fusion resulted in a 93% lower cost of pipeline development, an 82% lower cost of pipeline maintenance, and a 57% lower cost of pipeline execution. The results of our two-year modeled analysis are shown in Figure 6.

Figure 6. Modeled Cost to Build and Operate a Cloud Data Lake for a Distributed Organization



Source: Enterprise Strategy Group

<u>What the Numbers Mean:</u>

- **Faster Time to Insight and Value:** Cloud Data Fusion allows organizations to achieve the goals of their data lake projects months earlier than with a DIY approach and weeks earlier than with alternative DI tools. With critical data sources synced to the lake and pipelines running reliably, organizations can begin to realize the operational and financial benefits of the data lake earlier and begin adding sources and reports to grow the capabilities earlier.
- **Economies of Scale for Distributed Organizations:** Probably the largest benefit seen was the reusability of components and expertise. Instead of repeating efforts from scratch, as teams built and perfected pipelines in one market or region, it greatly sped the time for subsequent markets to repeat similar work. The more work that is put into the data lake, the faster it gets to bring the functionality to new sources, markets, and regions.
- **Streamlined Workflows and Operational Efficiency:** Cloud Data Fusion allows organizations to build data lakes with a streamlined workflow, using smaller teams, with less specific areas of expertise. Without the need for siloed developers performing serialized tasks, data lake teams operate more efficiently and achieve faster results.

Issues to Consider

It should be noted that, in our analysis, for simplicity and to focus on the pipeline-related costs, we did not include those costs of the data warehouse and data lake solution that we thought would be incurred equally in all three cases, such as the cost of cloud storage, database services, Big Query processing, network egress costs, etc. Also, while ESG's models are built in good faith upon conservative, credible, and validated assumptions, no single modeled scenario will ever represent every potential environment. ESG recommends that you perform your own analysis of available products and research Google Cloud Data Fusion and alternative offerings to understand the differences between the solutions hopefully proven through your own proof-of-concept testing.

The Bigger Truth

Google's acquisition of Cask and CDAP was an important step toward the goal of expanding enterprise capabilities on Google Cloud. The vision was to help organizations shift the focus from code and integration to insights and actions. Mission accomplished.

As data engineers, data scientists, ETL developers, and business analyst teams think through how best to build, manage, and maintain data pipelines, they will find ample reasons to consider Cloud Data Fusion—not the least of which is that it hides the complexity of traditional pipeline development and maintenance behind code-free simplicity. By managing data sets and pipelines centrally on a unified platform, organizations use fewer specialized skills yet realize efficiencies on many fronts. A customer that will be using Google Cloud in 24 countries commented that "*Cloud Data Fusion is the enterprise tool we are using to drive our global transformation towards a common platform…allowing all the people working with different technologies to work off a common platform and share the best practices and knowledge*."

ESG's modeled scenarios for structured and unstructured data predict that organizations can capture substantial savings and accelerate time to value compared to build-it-yourself efforts and alternative data integration solutions. Our models demonstrated up to 80% savings to deploy, manage, and maintain data pipelines for cloud-based enterprise data warehouses in BigQuery. Up to 88% lower cost is forecast to operate a hybrid cloud data lake into which siloed, on-premises platforms are integrated across geographical markets. While actual savings and benefits would be dependent on scale and complexity of environments, we believe that Cloud Data Fusion will prove most beneficial to large organizations and/or those that do a substantial amount of ETL in a repeatable fashion.

By eliminating the complexities, cost, and delays associated with traditional data integration and ETL/ELT pipeline deployment, organizations can focus on becoming true data-driven organizations that operate with less red tape and more self-service. Consider Cloud Data Fusion to make this journey faster, easier, and better.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.