ESG

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

ESG Economic Value Audit

Analyzing the Economic Benefits of Google Cloud Dataproc Cloud-native Hadoop and Spark Platform

By Aviv Kaufman, Senior Validation Analyst

June 2018

Executive Summary

Companies faced with the reality of exponential growth in data have found Apache Hadoop and Apache Spark to be effective solutions for housing and managing their data processing and data lake needs. Hadoop is powerful; however, the complexity of effectively managing an environment growing at such a dramatic pace is a logistical and administrative nightmare.

Google Cloud Dataproc is a highly available, cloud-native Hadoop and Spark platform that provides organizations with a cost-effective, high-performance solution that is easy to deploy, scale, and manage. ESG validated that Dataproc provides cost-effective and agile managed Hadoop clusters that can easily be spun up and down as required and can be optimally configured for individual jobs. Through this validation process, ESG found that while there were significant cost benefits from shifting an on-premises Hadoop environment to Dataproc, customers also reported substantial benefits in the strategic value they were able to pull out of the data hosted in the Google Cloud.

This evaluation was created as a TCO analysis comparing an on-prem Hadoop and Spark strategy against hosting the same data with Cloud Dataproc. In the process, we were able to gain insights comparing Cloud Dataproc and Amazon EMR from customers who have experience using both cloud environments.

57% Lower TCO than On-Prem Hadoop

32% Lower TCO than AWS EMR

Our analysis shows that hosting the same data using Google Cloud Dataproc is 57% less expensive than using on-premises servers and 32% lower than using Amazon EMR. Additionally, we found dramatic benefits in the ability for customers to pull strategic information from data stored with Dataproc (when compared with the other two environments). While cost is always a driving factor, every customer ESG interviewed as part of our research identified business and revenue benefits that far outweighed the cost benefits.

This ESG Economic Analysis was commissioned by Google and is distributed under license from ESG.

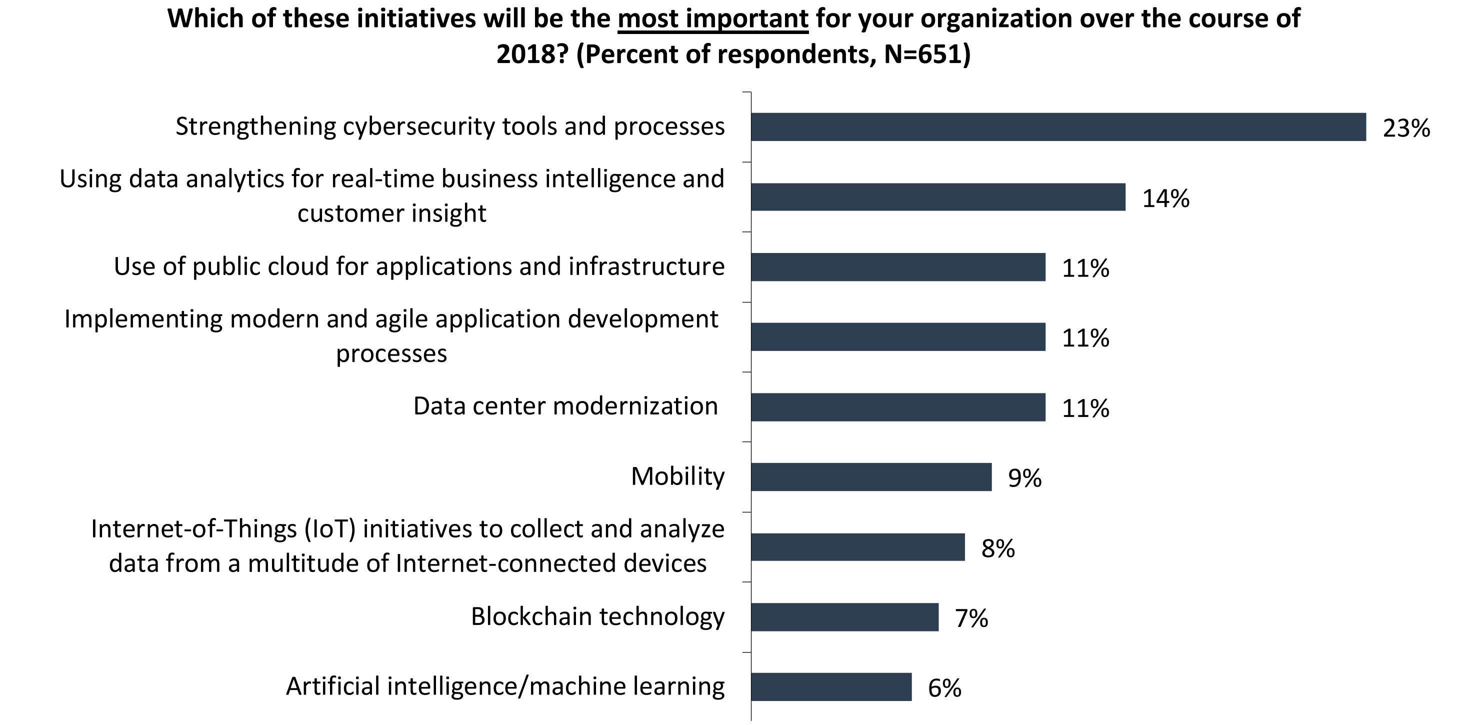
Introduction

This ESG Economic Value Audit focused on the quantitative and qualitative benefits organizations can expect from leveraging Cloud Dataproc instead of deploying on-premises (on-prem) Apache Hadoop and Apache Spark clusters or leveraging other cloud solutions. ESG created a modeled scenario that factored in cost of servers, storage, software, support, maintenance, and administration over a three-year period.

Challenges

In today's digital age, significant value is placed on gaining a better understanding of every aspect of the business. Scientific applications, online activities, mobile applications, IoT sensors, and other sources are generating data at an enormous rate. Sources report that up to 90% of all the data that has ever been generated was generated in the last two years.[1] Somewhere in this vast pool of data are the customer trends, business opportunities, and actionable insight that will define success for organizations in the coming years. ESG research shows that after strengthening cybersecurity, using data analytics for real-time business intelligence and customer insight is the second most important IT initiative for 2018 (see Figure 1).[2]

Figure 1. Most Important Initiative for 2018



Source: Enterprise Strategy Group

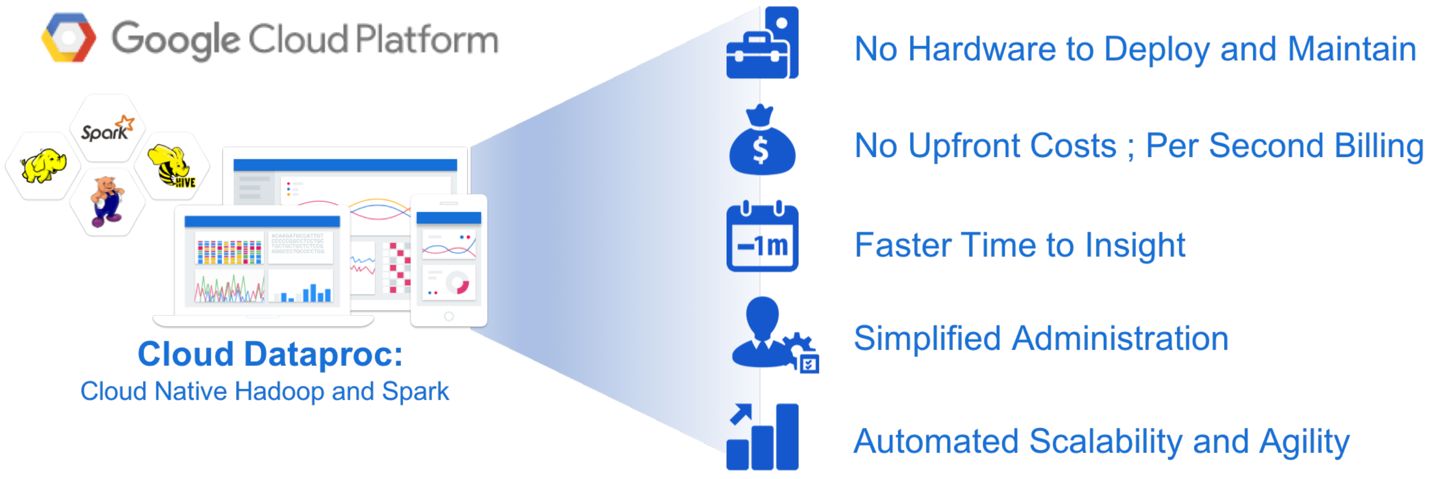
Gaining insight from data becomes progressively harder as the quantity and velocity of data increases. Organizations have turned to open source big data solutions powered by Spark and Hadoop to keep this vast information store organized and ready for processing. However, as the density of data grows, and the diversity of data evolves, the power and creativity needed to harness actionable insight becomes overwhelming. An upfront investment in an on-prem Hadoop or Spark cluster often results in a solution that lacks the required agility and is expensive to purchase, operate, and maintain.

[1] Source: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
[2] Source: ESG Research Report, *2018 IT Spending Intentions Survey*, February 2018.

The Solution: Google Cloud Dataproc

Cloud Dataproc is a highly available, cloud-native Hadoop and Spark service that provides organizations with a cost-effective, high-performance solution that is easy to deploy, scale, and manage. Dataproc nodes can be deployed and spun up in less than 90 seconds and can be easily customized and resized with the optimal resources required for individual jobs. The clusters access data stored in Google Cloud Storage (GCS), and can be leveraged in conjunction with Google's other big data solutions such as BigQuery, Dataflow, and TensorFlow to deliver a single platform for data processing, analytics, and machine learning. Google's Cloud Dataproc's key features are shown in Figure 2.

Figure 2. Google Cloud Dataproc


Google Cloud Platform
Spark
HIVE
Cloud Dataproc:
Cloud Native Hadoop and Spark
-1m
No Hardware to Deploy and Maintain
No Upfront Costs ; Per Second Billing
Faster Time to Insight
Simplified Administration
Automated Scalability and Agility

Source: Enterprise Strategy Group

ESG Economic Validation

ESG completed a quantitative economic analysis of Cloud Dataproc. Focus was placed on the economic benefits organizations can expect when leveraging Dataproc compared with on-prem Hadoop clusters and AWS Elastic MapReduce (EMR) managed Hadoop service.

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. ESG conducted in-depth interviews with end-users to better understand and quantify how Dataproc has impacted their organizations, particularly in comparison with previously deployed and/or experienced solutions. In addition to having experience with on-prem Hadoop and Spark solutions, some of the customers interviewed had migrated their Amazon EMR environments to Dataproc and were able to give detailed feedback on ongoing administration differences between the cloud solutions. The qualitative and quantitative findings were used as the basis for a simple economic model comparing the expected costs of on-premises and various cloud-based managed Hadoop services.

Google Cloud Dataproc Economic Value Overview

ESG's economic analysis revealed that an effective deployment of Dataproc can provide significant cost, administration, and agility benefits when compared with on-prem Hadoop and Spark deployments. Additionally, ESG found the flexibility and cost structure of Dataproc to provide savings and benefits even when compared with using Amazon Web Services

Elastic MapReduce (EMR) managed Hadoop service offering. ESG found that Dataproc provided its customers with significant savings and benefits in the following categories:

- **Hardware investment** – The drastically reduced need for onsite hardware results in lower upfront hardware cost while the simplified scalability of a Dataproc environment eliminates the need to over-purchase hardware.
- **Simplified administration** – ESG validated that administration, maintenance, and operation of a Dataproc environment is easier, faster, and more effective than managing Hadoop on-prem or with EMR.
- **Business agility** – Customers reported that Dataproc enabled enhanced business agility that allowed them to harness the value of their data to better service existing customers and open new revenue streams.

Lower Hardware Investment

As the ability to collect functional data points increases, data processing requirements and data lake size rapidly expand. Hadoop environments are required to grow at an exponential rate to store this growing data. Customers have found that by migrating their data to a Dataproc cloud they could:

> ***"We used to budget 25% over planned hardware capacity before we moved to Dataproc. Even worse, we sometimes had to tell our business groups to wait weeks on a project while we sourced and deployed more capacity to handle their needs. This was completely eliminated when we migrated to Dataproc."***

- **Reduce hardware spend by up to 71%** – Shifting to Dataproc eliminates the need for most of the server and storage hardware required for on-prem environments. When compared with EMR, Dataproc provided more predictable latency and customized instance types that can satisfy processing requirements with a smaller number of committed virtual instances.
- **Eliminate the need to overbuy capacity** – The ability to effectively spin up capacity in a Dataproc environment in minutes eliminates the need to project, procure, deploy, and maintain excess capacity to handle spikes, seasonal requirements, and company growth. The per second billing of Dataproc allows customers to pay for what they need at that moment instead of funding future growth. The virtually unlimited capacity of Google Cloud allows Dataproc customers to plan and sell without concern for the ability to store and access big data. Customers reported that the ability to quickly turn off a cluster reduces cost while giving them just-in-time access to the exact amount of resources needed to complete the job. The fundamentals of Moore's Law detail the doubling of computing power every two years; this means that the price of computing power reduces 25% each year. Dataproc allows you to pay today's hardware prices for today's needs, compared with on-prem requiring you to overbuy capacity, paying yesterday's prices for hardware that may already be outdated.
- **Achieve simplicity** – For companies managing on-prem Hadoop or Spark, the ever-expanding data lake and data processing requirements necessitate complex planning in both hardware needs and the physical space to house that hardware. Hardware costs go far beyond the cost of the actual servers. Companies are forced to plan and fund real estate to house servers and storage arrays while keeping them powered, cooled, and administered. The complexity of accurately projecting these needs is as much art as it is science. Dataproc eliminates this need as it allows you to focus on gaining insight from your data instead of planning for your data.

- **Realize flexibility** – Dataproc offers the ability to create custom machine types. This allows you to pick exactly how much processor power and memory you want for the job and pay for only what you need. Additionally, Dataproc's 90 second spin-up and scheduled deletion gives you the flexibility to do what you need, when you need to do it, without paying for wasted time or capacity.

Simplified Administration

Customers reported an 85% reduction in administration costs when moving Hadoop and Spark on-prem operations to Google Hosted Dataproc and a 34% savings in administrative costs when shifting from Amazon EMR to Dataproc. In addition to the elimination of many of the tasks associated with an on-prem Hadoop cluster, the simplicity of administering Dataproc allows less experienced and less expensive resources to complete most of the simplified tasks. For example, customers we interviewed reported that on-prem resources took an average of 30 minutes to prepare or configure clusters for jobs while creating the same capability with Dataproc takes only one to two minutes.

"We had 1 employee working 20 hours a week purely focused on just the EMR pipeline. This was a high salaried employee due to the complexity of EMR. This was completely eliminated when we migrated to Dataproc. We saved about $200K a year just on those basic administrative needs."

In traditional on-prem environments, there are requirements for procurement, deployment, configuration, scaling, tuning, testing, updates, and monitoring of equipment. The same amount of capacity can be administered with Dataproc and requires only minimal time for configuration—the rest of the time it is available to analyze and pull insights from your data.

"The instability in the EMR platform caused us to overprovision. EMR was just as painful as running our own clusters. We look to managed solutions to make things easier. EMR was just the opposite. This pain went away when we adopted Google Dataproc."

Customers interviewed by ESG who have experience with both Dataproc and Amazon's EMR reported reductions in the time and complexity of administrative tasks when comparing the two cloud-based services. One reported that they needed a high-level administrator who spent over 50% of their time just making sure the EMR pipeline was configured and available. This need was eliminated when the organization migrated to Dataproc.

Table 1. Simplified Management Helps Lower Administration Cost

IT Task	On-Prem Hadoop	AWS EMR	GCP Cloud Dataproc
Planning and Research	$ $ $	$	$
Justification & Procurement	$ $ $	◖	◖
Deployment	$ $ $	$	◖
Configuration	$ $ $	$	◖
Tuning & Optimizing	$ $ $	$ ◖	◖
Daily Job Administration	$ $ $	$	$
Troubleshooting Hardware	$ $ $		
Updates and Maintenance	$ $ $		
Capacity Planning	$ $ $		
Hardware Refresh	$ $ $		

Source: Enterprise Strategy Group

Business Agility

Data is only valuable when it can be effectively utilized to help meet, predict, or improve the business. ESG's customer interviews uncovered several ways in which replacing on-prem Hadoop clusters with Dataproc had helped to make the business more agile and better enable business processes through:

> ***"We are now able to give all employees access to data. This allows them to identify customer trends that we never would have spotted before and complete projects weeks faster than we could have in our old environment."***

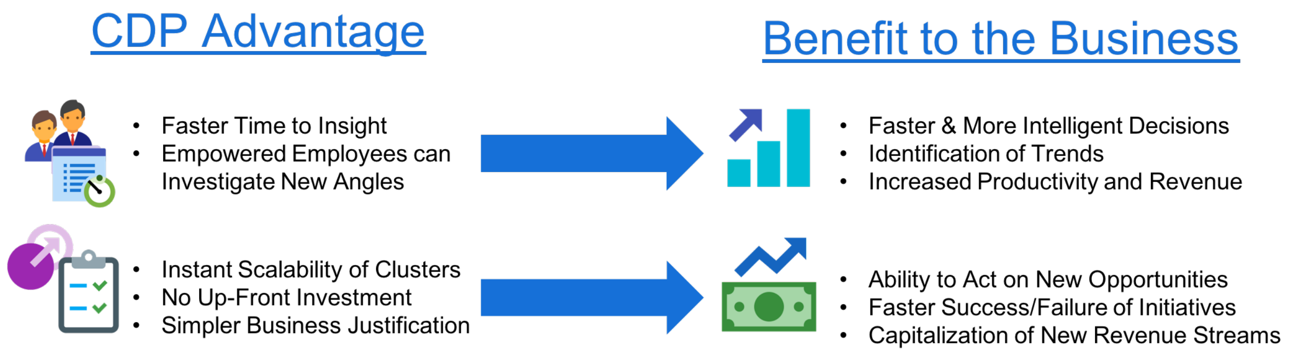
- Increase in current revenue streams – Customer interviews showed a marked improvement in the ability to capture revenue in a Dataproc environment. One customer reported, "We could make more intelligent decisions quicker and with more insight than we could in the past. We were able to empower all employees with the ability to run their own hypotheses which allowed us to identify customer trends. We saw gains of about 45% in revenue in these areas."

- Capitalization of new revenue opportunities – Identifying opportunities is only a part of the challenge. ESG has found that customers were able to act faster on new revenue opportunities in a Dataproc environment than they could in the past using on-prem Hadoop or Spark. The ability to instantly scale allows for fast processing of data and quick access to the value in their data lake while the flexibility of Dataproc gives companies the freedom to try new ideas without the need to commit to the long-term capacity cost that they saw running Hadoop on-prem.

> ***"With CDP we looked at data in ways we couldn't in the past. All employees were able to touch data through custom dashboards. This drove new initiatives and created new revenue streams."***

- Ability to quickly act on opportunities – Customers reported substantial value in the ability to act quickly on opportunities. Whether a seasonal observation or pivoting to a change in the marketplace, the agility brought by Dataproc not only opened new sales channels weeks faster than on-prem, but it also allowed for rapid changes in existing operations, which either increased sales or decreased problem areas.

Figure 3. Improved Business Agility with Cloud Dataproc



CDP Advantage

- Faster Time to Insight
- Empowered Employees can Investigate New Angles

- Instant Scalability of Clusters
- No Up-Front Investment
- Simpler Business Justification

Benefit to the Business

- Faster & More Intelligent Decisions
- Identification of Trends
- Increased Productivity and Revenue

- Ability to Act on New Opportunities
- Faster Success/Failure of Initiatives
- Capitalization of New Revenue Streams

Source: Enterprise Strategy Group

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 a three-year TCO/ROI model that compares the costs and benefits of satisfying a modeled organization's Hadoop requirements with Dataproc versus an on-prem Hadoop solution or with AWS EMR. ESG's interviews with customers who have recently made the transition, combined with experience and expertise in economic modeling and technical validation of Hadoop solutions helped to form the basis for our modeled scenario.

The modeled organization required 1 PB of stored data across 150 Hadoop data nodes. Each node consisted of eight processors and 30 GB of RAM. The cloud-based solutions required only 1 PB of storage that was protected and could be accessed by all 150 nodes, as well as a nominal amount of storage per node. In comparison, an on-prem Hadoop deployment would require three times as much storage to provide redundancy.

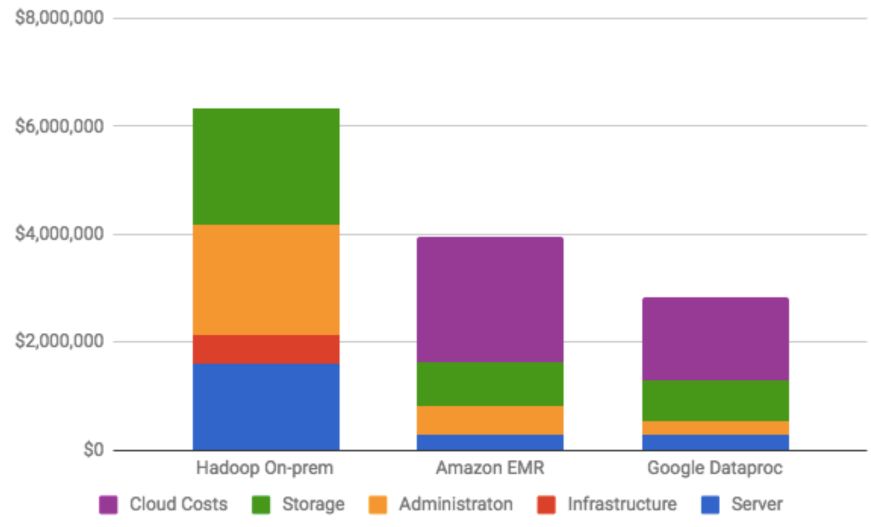
Why This Matters

On-premises Hadoop requires the purchase and maintenance of physical servers, usually on a three- or four-year lifecycle plan. This requires a planning and procurement cycle that can be limiting to most businesses.

Both Dataproc and EMR allow for dynamic provisioning of capacity. Dataproc adds a level of flexibility compared with EMR when looking at costing models because of Dataproc's speed in spinning up and deletion of capacity.

The nodes were expected to run for 16 hours per day on average over the year. An additional 50 nodes were periodically required to run to handle seasonal bursts and additional queries as demanded by the business. While the cloud-based solutions could spin up new nodes only when required, the on-prem Hadoop solution required deployment of the "worst case" scenario from a performance and capacity standpoint as well as the purchase of an additional 50 nodes at the time of initial deployment.

In addition to gaining feedback from the actual environments of the customers we interviewed, we obtained validation of expected administration time spent managing each solution. Customers we asked about the strengths and weaknesses of each solution and were given the opportunity to provide guidance that allowed ESG to model the administrative hours required for an on-premises deployment as well as the expected reduction in hours expected with Google Cloud Dataproc and Amazon EMR.



ESG found that the Dataproc solution could satisfy the needs of the modeled organization at a total cost that was 57% lower than an on-prem Hadoop deployment with no upfront costs, and 32% lower than an AWS EMR deployment.

The Dataproc solution greatly reduced or eliminated upfront costs and maintenance, support, and infrastructure spending (power/cooling/floorspace). And because administration was greatly simplified and did not require trained server, storage, or database administrators, cost of administration was reduced by 85% versus an on-prem deployment, and 48% versus an AWS EMR deployment. While every organization's requirements are different, and your particular savings may vary, ESG believes benefits will be realized to some degree by most organizations looking to lower cost and complexity, while getting the most out of their Hadoop deployment.

The Bigger Truth

After leveraging traditional, proprietary, siloed BI platforms for decades, Hadoop and Spark proved to be the critical catalyst in enabling organizations to unlock the next level of insight that was attainable only through the consolidation of multiple data sources. Scientific and business organizations alike quickly grew the infrastructure and expertise required to effectively process the data they had, while also growing their ability to capture increasingly complex streams of data in hopes of gaining even more insight. On-prem Hadoop clusters require a large upfront investment in terms of time and money, must be over configured to handle peak workloads, and lack the agility to easily scale or be reconfigured once deployed.

ESG validated that Dataproc provides cost-effective and agile managed Hadoop clusters that can be easily spun up and down as required and can be optimally configured for individual jobs. ESG validated the many benefits of Dataproc with Google customers and found that Dataproc not only provided their operations with significantly faster time to insight, but also provided substantial cost savings without the need for a long-term contract or an upfront investment. More importantly, Dataproc freed administrators to work on higher value initiatives and enabled businesses to perform analysis that they otherwise would not have been able to perform—positively impacting revenue. While some organizations may continue to find value in owning and operating their own cluster, Dataproc can still be used to augment existing operations by helping to gain insight in other areas.

ESG has performed economic evaluations of many of GCP's offerings such as Google Compute Engine, BiqQuery, and Advanced Networking services and has found that with each offering, Google provides its customers with products that offer simple and flexible solutions with fair and highly visible per second on-demand pricing. Dataproc offers an excellent opportunity for organizations to take advantage of the true agility and cost savings of a cloud-based architecture while leveraging their existing expertise in Hadoop and Spark. If your organization is looking to begin or expand the capabilities of your Hadoop and Spark infrastructure with a cloud-based solution that is high performing and cost effective, ESG recommends that you give Google Cloud Dataproc serious consideration.

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