ESG Economic Validation

Analyzing the Economic Benefits of Google Cloud Dataproc

By Nathan McAfee, Economic Analyst

April 2022

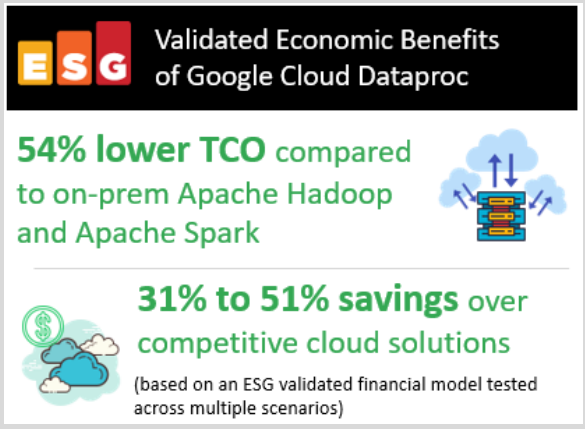
Executive Summary

The ideas that drive revenue and innovation are hidden in the stored (secondary) data of most organizations. However, the explosive year-on-year growth in the size of data stores makes the collection and storage of this data a challenge. Many companies utilize on-prem Apache Hadoop and Apache Spark to store and query their data but find the cost and complexity, both in planning and managing on-prem Hadoop, to be overwhelming. The result is missed opportunities to pull value out of datastores, increased risk to the security of their data, and a CapEx costing model that requires heavy periodic investments.

Google Cloud Dataproc is a cloud-native solution that runs over 30+ open source tools and frameworks including Apache Hadoop and Apache Spark. Dataproc is a highly performant solution that lowers the costs of storing and querying big data, improves the speed and overall quality of data insights, and reduces complexity when compared to on-prem solutions.

ESG created a financial model comparing the TCO and ROI of Google Cloud Dataproc when compared to on-premises solutions and competitive IaaS scenarios.

While benefits are realized through a combination of technology changes and best practices, ESG research finds that organizations can increase the value of insights queried in data, while lowering costs and complexity, by deploying Google Cloud Dataproc.



Introduction

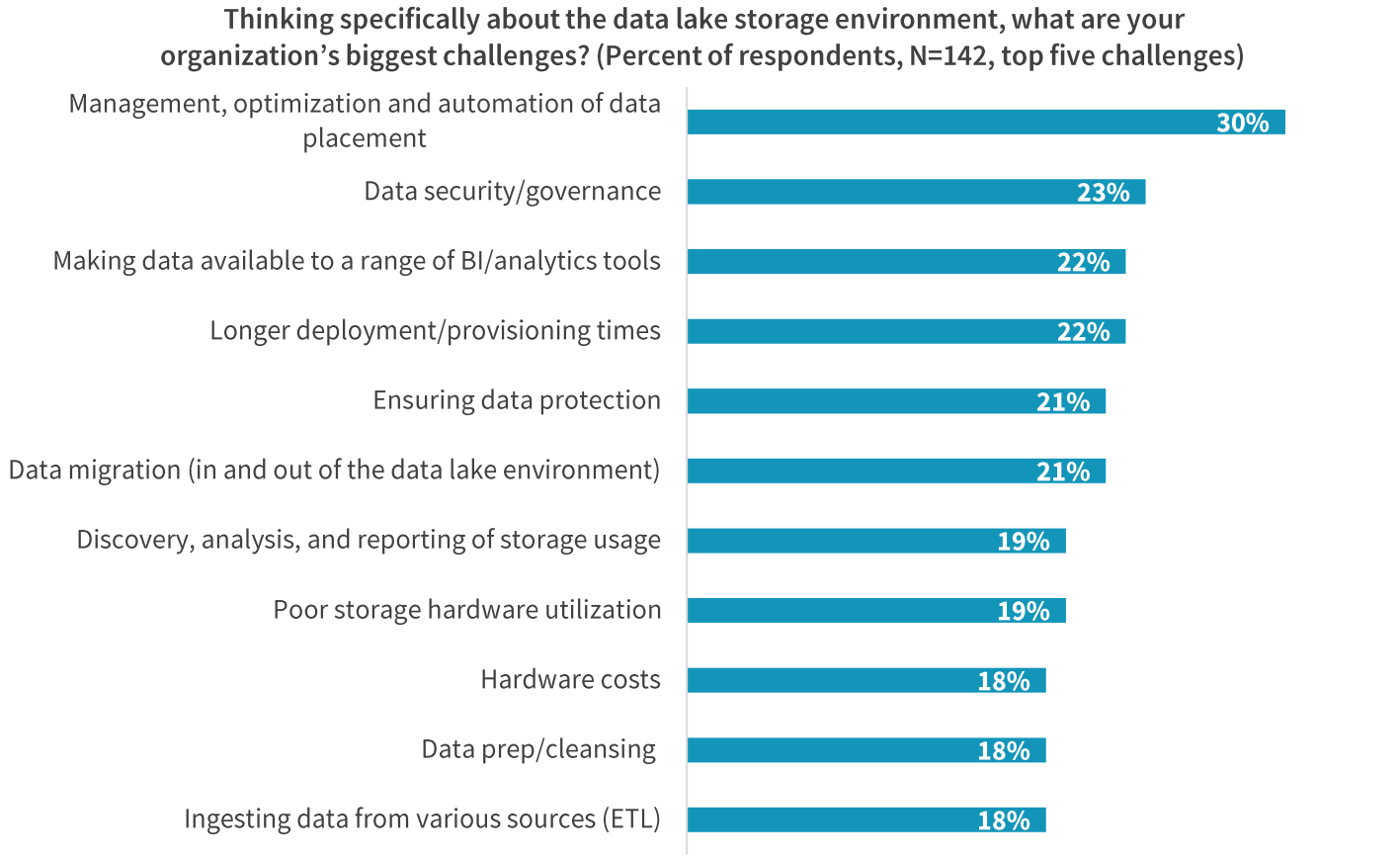
This ESG Economic Validation focused on the quantitative and qualitative benefits organizations can expect from using Google Cloud Dataproc when compared to on-premises deployments of Apache Hadoop and Apache Spark clusters or other cloud solutions to create, manage, and access clusters. Assumptions in this paper are based on a modeled scenario that considers the cost of servers, storage, software, maintenance, administration, support, and both hard and soft benefit areas.

Challenges

Most companies understand that the pathway to future growth is contained in their customer and operational data. However, the sheer volume of data that is generated by businesses makes the ability to analyze trends and opportunities overwhelming. Data is growing at an explosive rate, both in the velocity of new data creation and the constant addition of new data types and sources. Rapid change in stored data forces organizations to focus on the core fundamentals of data security and cost, too often ignoring the inefficiencies experienced when trying to utilize the data to enable future revenue. The result is an unrealized value stored in secondary data.

Many have adopted on-prem Apache Hadoop to store and process data along with Apache Spark. While these solutions provide the base functionality needed to store and access data, they quickly become limiting in a rapidly changing environment. ESG research shows that the top challenges organizations face when setting their data lake strategies include data placement optimization and automation (30%), security and governance (23%), and making data available to BI/analytics (22%). See Figure 1 for more information.[1]

Figure 1. Top Eleven Data Lake Storage Challenges



Source: ESG, a division of TechTarget, Inc.

[1] Source: ESG Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.

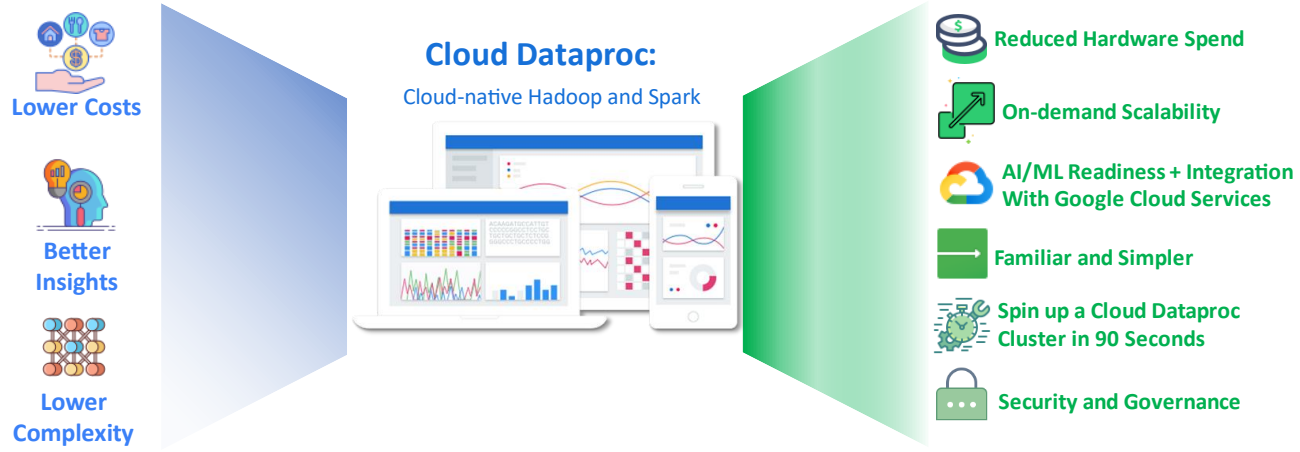
Massive scale in data storage is no longer enough. Companies that want to fully capitalize on the value contained in their data must find a better solution than on-prem Hadoop and Spark.

The Solution: Google Cloud Dataproc

Google Cloud Dataproc is a fully managed service created for data lake modernization. Dataproc shifts the focus away from hardware configuration and maintenance to allow organizations to concentrate on pulling value from their data to increase revenue and solve business problems.

Dataproc is a cloud-native service that runs over 30+ open source tools and frameworks, including Hadoop and Spark. Dataproc lowers costs and enables better and faster insights, all while lowering the complexity when compared to on-premises data environments. Dataproc customers will find a familiar Google interface that combines high performance, low maintenance, rapid scalability, and the security of the Google Cloud (see Figure 2), including the ability to use Dataproc on GKE to execute big data applications using the Dataproc jobs API on GKE clusters. Customers just getting started with Google Cloud Products will find variations of Dataproc to meet their specific needs, including serverless options for organizations looking to reduce infrastructure management and costs. Dataproc on Google Compute Engine (GCE) offers organizations the control to create the configuration to meet their specific needs.

Figure 2. Google Cloud Dataproc



Source: ESG, a division of TechTarget, Inc.

ESG Economic Validation

ESG completed a qualitative economic analysis of Google Cloud Dataproc, evaluating the economic benefits that a company can realize with Dataproc. A comparative environment of on-premises Hadoop clusters and competitive cloud-hosted, managed Hadoop were used to project benefits.

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 used custom and public research, existing case studies, internal industry and product analysts, and in-depth subject matter expert interviews to form the guidance contained in this paper. Economic results contained are a combination of changes in products, delivery environments, and best practices.

Google Dataproc Economic Overview

ESG's economic analysis revealed that Cloud Dataproc provides substantial benefits in lowered costs, faster time to value, and higher quality insights, while reducing complexity when compared to on-premises Hadoop and Spark deployments.

- Lower Costs – Elimination of hardware requirements, combined with billing flexibility, results in lower costs when compared to on-prem solutions.
- Faster Time to Insight and Higher Quality Insights – A solution that is quick to start, scale, and shut down brings agility that results in faster insights and the ability to use an entire set of data results in higher quality insights.
- Reduced Complexity – The elimination of hardware shifts the focus of FTEs to pulling valuable insights out of their data instead of building complex on-prem ecosystems.

Lower Costs

On-prem data warehouses require substantial capital expenditures that necessitate accurate prediction of future need and expensive administrative resources to configure and maintain in categories including:

- Reduced Hardware Spending – Dataproc is a cloud-based solution, eliminating the majority of the hardware spending associated with on-prem big data. In addition to the cost of the hardware, hidden costs such as hardware procurement, capacity planning, power and cooling, and physical space costs are virtually eliminated with Dataproc. ESG-verified financial models show up to 73% reduced hardware spending when comparing Dataproc to on-prem solutions and the ability of Dataproc to provide services more efficiently than competitive cloud solutions.

ESG financial models show up to a 73% reduction in hardware spending with Dataproc while shifting from an investment heavy CapEx model to a predictable open billing method.

- Lower FTE/Administrative Costs– Managing on-prem big data is complex and expensive. Highly paid IT resources must consistently configure, scale, tune, update, and monitor hardware. Each of these processes also injects risk into overall data operations. In addition to day-to-day maintenance, on-prem instances can take 15 to 30 minutes each to create, compared to 90 seconds to start and scale in Dataproc. This means less time administering and more time pulling value from your data with Dataproc. ESG-verified financial models show an 85% reduction in administration costs when moving Hadoop and Spark from on-prem to Dataproc and a 36% savings in administrative costs when shifting from competitive cloud solutions to Dataproc.

- **Elimination of the Need to Overprovision**– On-prem Hadoop and Spark environments are expanded through a traditional "stair step" model where periodic, large capital expenditures are necessary to acquire hardware. The exponential growth of data makes predicting this demand more challenging. In this stair step model (as seen in Figure 3), capacity is normally kept at 30% above projected peak demand, resulting in capital dollars spent that do not drive decisions or revenue. However, with additional capacity coming in large chunks, on-prem organizations see overprovisioning of up to 40% at some points in their demand curve. 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. Adding to this pain point, many companies are seeing up to a 6-month supply chain to add additional capacity once the decision to expand is made. Dataproc allows you to pay today's hardware prices for today's needs, compared with an on-prem model that requires you to overbuy capacity, paying yesterday's prices for hardware that may already be outdated.

  "Google allowed us to decouple compute and storage. Before, we were forced to predict and provide compute and storage at the same time. This caused us to overprovision at times and created challenges when demand for one resource was higher than the other."

  Scalability is a key tenet of Dataproc. With the ability to quickly spin up whatever instances are needed, as well as extend (or reclaim) GPUs and SSDs to match demand, additional capacity is always available with Dataproc without the need to overprovision.

Figure 3. On-prem Hadoop and Spark Require a "Stair Step" Model of Capacity Forecasting and Capital Expenditures

A traditional hardware model (stair step model) involves predicting demand and accompanying capital expenditures to increase capacity. However, most demand models do not follow the same pattern. The result is a pattern that sees overprovisioning at most points on the use curve. At critical times (as seen in the red circle), when demand outpaces capacity, service-level agreements (SLAs) are missed and organizations need to throttle business.

At times in this stair step model, companies can see resources overprovisioned an average of 40%

Source: ESG, a division of TechTarget, Inc.

- **Flexible Billing Model** – Dataproc is billed at 1 cent per virtual CPU per hour with the added ability to include preemptible instances, which can reduce costs up to an additional 91%. These preemptible instances are for

applications that are fault-tolerant and can withstand instance preemptions.[2] Additionally, the ability to quickly spin up and terminate instances helps eliminate excess charges for capacity that is not being actively used. With Dataproc, usage is billed by the second and a properly shut down instance does not incur charges.

Faster Time to Insight and Higher Quality Insights

- Ability to Use Entire Set of Data – Decisions made on complete sets of data are far more likely to be accurate than those made on partial data. In on-prem environments, the cost of aggregating and querying data can be high—in dollars, resources, and time. This results in many decisions being made through sampling or queries on partial data sets. With Dataproc, the cost in time and dollars is substantially lower, allowing entire data sets to be used to provide better answers.

There is substantial business value in increasing the number of queries on a data set. By lowering the cost requirements (both in dollars and time) of accessing big data, Dataproc customers find they can open data exploration to more employees. More eyes on data equates to more value pulled from data.

- Automatic Scalability – There is substantial value in querying data when a question is fresh. The ability to quickly provision needed capacity and alter resources during a job enhances the likelihood that queries will result in definitive answers. With Dataproc, capacity is available almost immediately instead of in hours with on-prem or, in the case where capacity demand means new hardware, months in traditional Hadoop and Spark. Capacity is always available without the planning, provisioning, and deployment delays associated with on-prem growth. One customer shared with ESG, ***"It used to take between 3 and 6 months to provision a server. We cannot get resources immediately when our needs change. We sometimes found that by the time capacity was provisioned, the need had changed."***

- Faster Validation/Fast Fail of Hypotheses – Faster validation is a key metric for almost all data warehouses. Performance is measured in query response speed. What is generally lost in this measurement is the value of fast fails. The ability to quickly test a hypothesis can be challenging in an on-prem environment. With Dataproc, ideas can be validated faster, eliminating cost and resource constraints that prevent the testing of a "random idea" in a traditional environment. History is filled with quick questions that led to new products or entire lines of business.

"Small ideas can become large opportunities when you have more data scientists working with your data. With Dataproc, we are able to provide resources faster and at a much lower cost than in the past." -Head of Analytics, International Communication Company

Reduced Complexity

Dataproc is built on the Google Cloud. This familiar interface presents a dramatic reduction in complexity when compared to the hardware configurations necessary for on-prem Hadoop and Spark.

[2] Source: Google Cloud, *Spot VMs*.

- Familiar and Simpler – Dataproc takes the complexity out of familiar open source tools. Integration with other solutions in the Google Cloud ecosystem reduces the learning curve for those exploring tools such as Hadoop and Spark, all while adding the assurances that come with the backing of Google.

- No Hardware to Manage– Big data hardware management is costly and complex. Capable administrators are in very high demand, often hard to find, and command high annual salaries when they can be found. By eliminating the need for hardware administrators, Dataproc organizations benefit from the removal of the complexity that big data hardware brings.

- Secure by Default– Google Cloud products are built with the mantra of being "secure by default." By shifting the burden of security from hardware administrators to Google, organizations can focus their energy on what they do best to drive revenue initiatives.

ESG Analysis

ESG leveraged the information collected through vendor-provided material; public and industry knowledge of economics and technologies; and previous economic and technical studies of Google Dataproc, Hadoop, and competitive big data solutions to create the financial models that were used as a foundational component for this analysis. The financial model was created using a sample company with the following profile: 300 nodes, 16 cores and 256 GB per node, 36 TB storage per node, and a total of 9 PB of data. A replication factor of 3 and an average daily CPU usage of 50% were used. In addition to the Hadoop/Spark nodes, the scenario included Hive, Cassandra/HBase, Elasticsearch, and Kafka.

The resulting numbers showed between 31% and 51% savings when compared to similar competitive cloud situations with Google Cloud Dataproc, showing a higher ROI in each of the 10 tested scenarios.

The Bigger Truth

The concepts that create new revenue streams and expand existing ones are contained in the stored data of most companies. The challenge is to find those concepts amidst the noise of a huge pool of data that is increasing in size at rates unseen in history. Many organizations have adopted Apache Hadoop and Apache Spark to store and query their data. These open source solutions provide quite a bit of power but come with a level of complexity that often forces limits on data access and a hardware-reliant model that requires large capital investments and costly administration and forces overprovisioning that results in cash outlays for capacity that sits unutilized.

ESG validated Google Cloud Dataproc, a cloud-based solution that addresses many of the challenges that companies face with big data. Through study of financial models, case studies, ESG analyst views, and industry research, ESG found that customers can realize reduced costs, improved quality of insights, and reduced complexity with Dataproc when compared with on-prem solutions. ESG found that Dataproc customers can shift the focus of IT resources that would be tasked with below-the-line hardware maintenance activities with an on-prem solution to above-the-line activities that better align the value hidden in data with business initiatives.

When comparing the overall cost metrics of Google Cloud Dataproc with that of on-prem hardware, the ESG-validated financial model shows a 54% lower TCO for Dataproc. When comparing the TCO between Dataproc and competitive cloud solutions, the simplicity of Dataproc pricing was clear, and the resulting TCO numbers were between 31% and 51% lower for Dataproc, depending on the specific scenario.

The realized value of Dataproc is a combination of cost savings and best practices that enhances the value of secondary data. While each customer's situation is unique, ESG believes that organizations with on-premises data stored in Hadoop could benefit from a transition to Google Dataproc.