The Economic Advantage of Google BigQuery On-Demand Serverless Analytics

By Aviv Kaufmann, Senior ESG Validation Analyst; and Nik Rouda, Senior Analyst
April 2017

This ESG White Paper was commissioned by Google and is distributed under license from ESG.
Contents

The Challenge: Economical Insight ........................................................................................................... 3
The Solution: Google BigQuery .................................................................................................................. 4
Google BigQuery versus Alternative Solutions .......................................................................................... 5
ESG’s Economic Value Audit Process ........................................................................................................ 6
Economic Benefits of Google BigQuery .................................................................................................... 6
ESG Economic Validation ........................................................................................................................... 8
  Modeled Scenario #1: Small Organization ............................................................................................... 10
  Modeled Scenario #2: Medium Organization ........................................................................................... 11
  Modeled Scenario #3: Large Organization ............................................................................................... 12
  Model Considerations: Pricing Options .................................................................................................... 13
The Bigger Truth .......................................................................................................................................... 15
The Challenge: Economical Insight

When it comes to gaining insight from your data, traditional business intelligence and data warehouse solutions have leveraged years of innovation to arrive at intelligent solutions and methodologies to gain valuable insight from structured sets of data. The success of these solutions has conditioned organizations to the value of collecting and analyzing data—resulting in initiatives to collect and process even more. Today, the sheer volume and rate of data generated dwarfs that of days past and the potential value of this data leaves little incentive to throttle back. Organizations are now faced with the reality of staying ahead of the ever-growing scale and velocity of the data they are generating. To better do so, many have turned to big data solutions powered by Hadoop and data lakes for most or even all of their data discovery, organization, analytics, and reporting initiatives.

Big data solutions do not come cheaply, simply, or quickly. Not only is a large upfront monetary investment in hardware and supported software required, but also an upfront investment in time to plan, purchase, install, configure, and test the solution is needed before delivering any value to the business. Expert administrators and operators are required to administer the system. Storage capacity requirements grow rapidly and massive amounts of compute power are essential to processing data quickly. It is nearly impossible to decouple storage capacity and compute power to scale independently. Systems must be greatly overprovisioned for redundancy and future growth, and run the risk of being obsoleted quickly. Quite simply: Purchasing or building an on-premises big data solution comes at a big cost with a big risk of impacting time to insight.

Time to insight is an important metric in today’s rapidly evolving, knowledge-powered industries. ESG’s annual IT spending survey reveals that nearly four out of ten organizations prioritizing big data initiatives in 2017 expect to allocate funding to enhancing their business intelligence capabilities and customer insights, which remains a priority as businesses seek to differentiate from their competition by enabling a smarter workforce (see Figure 1).

![Figure 1. 2017 Data Analytics Spending Priorities](image)
It is important to remember that most organizations are not in the business of building and operating big data solutions, but rather they are in the business of generating data and extracting valuable insight from this data. Similar to what IaaS did for physical infrastructure, cloud-based big data solutions can offer a cost-effective and rapidly scalable alternative to DIY or integrated on-premises solutions. To best meet the ever-growing requirements of their big data initiatives, organizations must keep a close eye on and truly understand the costs and benefits involved with the on-premises and cloud service solutions available to store their data, run queries, and extract insight.

The Solution: Google BigQuery

Google BigQuery is a cloud-based, fully managed, serverless enterprise data warehouse that supports analytics over petabyte-scale data. It delivers high-speed analysis of large datasets without requiring investments in onsite infrastructure or database administrators. BigQuery scales its use of hardware up or down to maximize performance of each query, adding and removing compute and storage resources as required.

Google BigQuery, part of the Google Cloud Platform, is designed to streamline big data analysis and storage, while removing the overhead and complexity of maintaining onsite hardware and administration resources. Some of the specific advantages of Google BigQuery for businesses that work with big data include:

- **Time to Value** - Users can get their data warehouse environment online quickly and easily, without requiring expert-level system and database administration skills by eliminating the infrastructure and reducing the management (known as “No-Ops” or “Zero-Ops”).
- **Simplicity** – Complete all major tasks related to data warehouse analytics through an intuitive interface without the hassle of managing the infrastructure.
- **Scalability** – Scale up to petabytes or down to kilobytes depending on your size, performance, and cost requirements.
- **Speed** – Ingest, query, and export PB-sized datasets with impressive speeds using the Google Cloud Platform as the underlying cloud infrastructure.
- **Reliability** – Ensure always-on availability and constant uptime running on the Google Cloud Platform with geo-replication across Google data centers.
- **Security** – Protect and control access to encrypted projects and datasets through Google’s cloud-wide identity and access management (IAM).
- **Cost Optimization** – Predict costs with transparent flat rate and/or pay-as-you-go pricing, and contain costs through the use of project and user resource quotas.

Google BigQuery is self-scaling; it identifies resource requirements for each query to finish quickly and efficiently, and provides those resources to meet the demand. Once the workload has completed, BigQuery reallocates those resources to other projects and other users. Both in transferring data in, and in processing that data for results, BigQuery delivers tremendous speeds even at petabyte scales. For enhanced data durability, BigQuery provides high availability and reliability through geographic replication that is completely transparent to its users, and without the requirement to obtain the physical resources and space to house it all.

Ultimately, Google BigQuery enables organizations to address the cost and complexity challenges associated with building and maintaining a fast, scalable, and resilient big data infrastructure. By leveraging Google BigQuery’s cloud-based approach, the time and cost traditionally dedicated to protecting data and guaranteeing uptime is nearly eliminated. With Google handling scalability, replication, protection, and recovery, organizations can focus more on gaining valuable insights, as opposed to infrastructure management.
Google BigQuery versus Alternative Solutions

The purpose of this paper is to help organizations understand and compare the direct and indirect costs that should be considered when choosing a solution to store their big data and perform queries against it. This paper compares a do-it-yourself (DIY) on-premises Hadoop cluster deployment with instance-based cloud services from AWS (Redshift with Kinesis) as well as the on-demand cloud service from Google (BigQuery).

With an on-premises Hadoop cluster, the organization must plan, deploy, maintain, and configure the physical hardware and software required to store the data and power the queries. Hadoop nodes are comprised of commodity servers populated with large NL-SAS drives that are used to store and protect the data. Substantial work must be done to administer, configure, and optimize both the hardware solution and the Hadoop/Hive software. AWS Redshift helps to greatly simplify the management and eliminate the need to physically administer the hardware. Like a Hadoop cluster, the AWS solution is based on the concept of nodes (albeit virtual nodes). To scale the deployment, similar nodes of a fixed compute and storage capacity are added simultaneously, sometimes resulting in provisioning more compute or storage capabilities in order to meet the requirements of the other.

Google’s BigQuery solution is completely serverless from the customer perspective. There are no nodes to plan, configure, or scale. The complexity of sizing, managing, and maintaining the physical infrastructure is handled behind the scenes by Google, so the burden is removed from the end-user. Figure 2 depicts the three solutions compared in this analysis and how they each implement compute, storage, administration, and query management.

Figure 2. Hadoop, AWS, and Google Implementations
ESG’s Economic Value Audit Process

ESG’s Economic Value Audit (EVA) 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. The EVA audit process leverages interviews with real-world customers who have had experience with both Google BigQuery and alternative big data solutions to help qualitatively and quantitatively validate the benefits that Google BigQuery has brought to their operations. This information is then applied to help estimate the costs and benefits in ESG’s modeled scenarios depicted in this paper.

Economic Benefits of Google BigQuery

Google BigQuery’s serverless, on-demand query service was designed and priced to provide customers with insight quickly and economically. ESG’s Economic Value Audit process revealed that BigQuery can provide significant cost savings and economic benefit opportunities. ESG found that BigQuery customers have enjoyed significant economic and operational savings when compared with both on-premises Hadoop-based deployment and AWS’s Redshift cloud-based big data solution. These benefits fall mainly into three categories: elimination of upfront capital investment; operational and administrative savings; and lower cost of cloud services.

Upfront Capital Investment Savings

- An on-premises Hadoop deployment requires a very large capital investment to purchase nodes, networking infrastructure, software, and licenses.
- An on-premises Hadoop deployment requires a significant amount of planning, purchasing, configuration, and testing prior to providing any big data benefit to the organization. This results in additional upfront operational expenses and potentially less revenue due to a much longer time to value when compared with BigQuery.
- For both an on-premises Hadoop deployment (hardware nodes) and an AWS Redshift deployment (AWS virtual instances), the storage capacity is directly tied to the compute power and memory, and none can be scaled independently of the other. This can potentially result in overprovisioning of compute or storage resources in an effort to scale the other, due to the inflexibility of pre-defined machines.
- For both an on-premises Hadoop deployment and an AWS Redshift deployment, the organization must spend time to plan and size the deployment, often overprovisioning to accommodate the worst-case scenario.
- AWS Redshift customers often choose to pay upfront, reserved instance pricing and benefit from significant discounts versus running Redshift instances on demand. Customers can save up to 75% over the on-demand pricing by choosing 3-Year All-Upfront reserved instance pricing.

BigQuery is serverless and payments are often made completely on-demand, based only on the amount of data processed and stored per month. This means upfront payments, planning, purchasing, installation, configuration, node management, or testing are not required. BigQuery can provide value in the form of insight as soon as the data is made available on GCP storage. Multiple customers who ESG spoke with reported that they were up and running queries on BigQuery in a matter of hours rather than days, weeks, or even months. Other customers who had experience with multiple solutions stated:

“...We spent over $1M on a Hive-based warehouse that was always running out of memory and required constant tuning and maintenance. Upgrades created instability and completely took the system down.”

“...of course we chose to pay the upfront reserved instance price. It doesn’t make much sense not to.”

Removing the financial barrier of a large upfront investment in terms of time and money makes BigQuery an attractive platform for organizations looking to simply try out the service, or quickly scale their analytics capabilities.
Operational and Administrative Savings

- An on-premises Hadoop deployment requires more people to administer the solution, often including a hardware and software administrator, and one or more Hadoop administrators and/or operators.
- With an on-premises Hadoop solution, analysts usually must work with the Hadoop operators to translate and execute queries rather than executing the queries themselves.
- An on-premises deployment must be powered and cooled, and requires floor space—increasing the operational cost of the deployment.
- On-premises deployments often require the services of consultants to help tune the solution for optimal performance because those skills go beyond those of the typical Hadoop administrator.
- On-premises deployments must be maintained, including updating firmware, the OS, security patches, and Hadoop releases, as well as troubleshooting and resolving issues with the hardware and software.
- AWS Redshift is similarly based on the concept of nodes containing a fixed set of resources. This may make planning, deployment, and growth more complex than with BigQuery’s serverless service.
- AWS Redshift requires the user to login to a “leader node” in order to run requests against the pool of Redshift instances. This added complexity means that less technical analysts may continue to leverage an operator to handle queries rather than performing the queries themselves.

BigQuery does not require dedicated administrators, and customers reported that it is simple enough for analysts to run queries on their own by cutting and pasting queries into a web-based self-service portal. In comparison, to use Redshift, analysts that are comfortable using VPN can log into a proxy server running in AWS to connect to the Redshift deployment. This proxy server can make it more prohibitive for multiple groups on different ACLs to share the same Redshift deployment, resulting in the need to deploy and manage separate instances for each department and further increasing the management complexity of the solution. When asked about operational costs, the customers with whom ESG spoke agreed that Redshift was far easier to manage, operate, and maintain compared with an on-premises Hadoop deployment, and also agreed that BigQuery was simpler to operate and manage than their Redshift deployment had been.

“The kind of people that can debug Hadoop exceptions are typically not close to the business. The error messages that come out of Hadoop are not well written and are difficult to troubleshoot. This led to retention issues—we lost good people because they were sitting around dealing with issues.”

“The web-client is a huge bonus. Analysts do not have to use an ODBC/JDBC, they just need a URL.”

The simplicity of BigQuery enables analysts and operators within the company to become citizen data scientists—empowering them to take control of their own queries, removing their dependence on others, and ultimately producing higher quality insight for the organization in a shorter amount of time.

Cloud Cost Savings

- Both Google BigQuery and AWS Redshift provide very well documented price lists and web-based pricing calculators to determine estimated cloud costs.
- Google BigQuery pricing is simple: Pay for your storage capacity and pay for the TBs processed each month. There is no complexity of figuring out sizing or risk of overprovisioning.
- Redshift On-Demand pricing is based on deploying virtualized instances (nodes) with fixed compute and storage requirements, meaning they cannot be scaled independently.
- Powering down Redshift nodes requires use of snapshots or migration of data. Only the smallest deployments generally benefit from paying for less than 24 x 7 operation.
- Redshift On-Demand pricing is generally quite high compared with BigQuery and also when compared with AWS Reserved Instance pricing options.
- Reserved Instance pricing on AWS provides a particular instance type, making it difficult to take advantage of the latest CPU or storage offerings over the length of your contract. BigQuery always makes use of the latest available...
technologies automatically and transparently.

BigQuery pricing is designed with simplicity and economics in mind. There are absolutely no upfront costs, and you simply pay for what you store and what you query. This is true “on-demand” pricing, without the need to plan, configure, tune, or update nodes. Customers may be wary of paying an on-demand per-TB processing fee without knowing exactly how much data they will process in a given month. ESG validated with customers and with internal data collected by GCP that the amount of data processed on average is very often equivalent to the amount of data stored. In many cases, this results in a very favorable cost advantage over Redshift (in which you must pay for enough instances to cover the storage requirements).

With BigQuery, pricing is simplified and planning is not required because it is not based on the concept of deploying preconfigured instances as big DataNodes. BigQuery users do not have to estimate the hardware requirements of their configurations, and never pay for what is not required. They do not have to chose between paying very high monthly on-demand costs or locking themselves into a one- or three-year contract with a large upfront investment. The customers whom ESG spoke with shared their concerns with reserved instance pricing:

“...we made the switch to BigQuery and today we still have 9 months left in reserved instances on our Redshift nodes.”

ESG’s analysis in the following sections of this paper further illustrate the operational and pricing advantages that BigQuery holds over an AWS Redshift deployment.

**ESG Economic Validation**

To validate the economic benefits of Google BigQuery when compared with both an on-prem Hadoop deployment and AWS Redshift, ESG created requirements for three modeled organizations that are representative of typical small, medium, and large organizations performing queries against their data. The requirements of these organizations are shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Modeled Scenario Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Data Citizens</strong></td>
</tr>
<tr>
<td>5 Analysts</td>
</tr>
<tr>
<td><strong>Amount of Data Stored</strong></td>
</tr>
<tr>
<td><strong>TBs Queried per Month</strong></td>
</tr>
<tr>
<td><strong>Streaming Applications</strong></td>
</tr>
</tbody>
</table>

ESG modeled and compared the costs that the small, medium, and large organization might expect to pay over a three-year period to plan, purchase, deploy, operate, administer, and maintain a solution to perform their queries. The assumptions and costs used in the scenarios were based on the results of detailed interviews with BigQuery customers who have experience with two or more of the solutions compared and could help to quantify the costs and relative differences between their deployments.

As will be described in more detail later in this paper, there are many factors to consider when it comes to cloud services pricing. Because AWS’s three-year reserved instance (RI) pricing provided the lowest total costs over three years when compared with other AWS pricing options, ESG based the TCO analysis on this upfront pricing model. ESG also sized and priced on-premises Hadoop nodes and AWS Redshift instances that would be expected to provide similar performance results when compared with the BigQuery solution.
ESG leveraged knowledge of markets, the industry, and vendor solutions, as well as detailed interviews with Google and its customers to model and predict the costs to deploy, administer, manage, maintain, and operate each of the solutions. These costs were generally based on the number of employees and contracted services that were required for different sized deployments, as well as for the individual solutions. Wherever possible, direct comparisons between the solutions were used to gauge relative differences in man-hour requirements. ESG’s modeled three-year TCO analysis considered the high-level cost categories depicted in Table 2 and created detailed models to estimate the costs of the small, medium, and large scenarios for each of the three big data solutions.

**Table 2. Modeled Scenario Requirements**

<table>
<thead>
<tr>
<th>Cost Category</th>
<th>Description</th>
<th>Hadoop On-premises</th>
<th>AWS Redshift</th>
<th>Google BigQuery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upfront Costs</td>
<td>Payment made prior to realizing value from the solution</td>
<td>• Cost of hardware, software, and licenses for Hadoop nodes</td>
<td>• 3-year reserved instance pricing paid in full</td>
<td>• No upfront costs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cost of networking</td>
<td>• Adjustment for time-value of money at 8% WACC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-year On-demand Cloud Costs</td>
<td>Expected monthly cloud services costs (sum of 36 monthly payments)</td>
<td>• No monthly cloud service costs</td>
<td>• No monthly cloud service costs for instances after paying upfront RI</td>
<td>• Cost of GCP storage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Cost of streaming inserts paid to AWS Kinesis service (medium and large</td>
<td>• Cost of total monthly TBs processed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>scenarios only)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Cost of streaming inserts (medium</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>and large scenarios only)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operational Costs</td>
<td>Resources (money and manpower) used to get the system functioning and keep</td>
<td>• Cost of power, cooling, and floor space</td>
<td>• Man-hour costs to plan deployment size and purchase Redshift instances</td>
<td>• Man-hour costs to migrate data,</td>
</tr>
<tr>
<td></td>
<td>the solution operating (not including administration and queries)</td>
<td>• Man-hour costs to plan and purchase Hadoop cluster</td>
<td>• Man-hour costs to configure, test, and perform POC</td>
<td>configure, test, and perform POC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Man-hour costs to install, test, troubleshoot, and perform POC on cluster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrative Costs</td>
<td>Expected costs to administer the solution on a daily basis</td>
<td>• Cost of query orchestration</td>
<td>• Cost of query orchestration</td>
<td>• Cost of query orchestration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cost of Hadoop administration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cost of database administration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cost of consultants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance Costs</td>
<td>Cost of maintenance and support contracts</td>
<td>• Cost of maintenance and support contracts for Hadoop nodes and networking</td>
<td>• No cost to maintain the solution</td>
<td>• No cost to maintain the solution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The results of the comparison showed that by leveraging the BigQuery service for their needs, the modeled organizations could save a total of between $532K and $2.7M over a three-year period, with the added benefits of utilizing a solution that integrates better into their existing environment, no upfront costs, and empowering citizen data scientists within the organization by removing the reliance upon an administrator to prepare, schedule, and execute queries.

**Modeled Scenario #1: Small Organization**

For the first scenario, ESG started by calculating the requirements to store a 100TB dataset and perform a mix of batch and ad-hoc queries that on average processed 50 TB of data per month (50% of the stored data capacity) for each of the three solutions (on-prem Hadoop, AWS Redshift, and Google BigQuery). ESG figured that the on-premises Hadoop infrastructure would require the deployment of a single Hadoop NameNode, and eight DataNodes. ESG assumed that Hadoop would leverage the standard triple redundancy and made sure that there was 20% extra storage capacity available for data growth (for both the on-premises and AWS instances). While this may sound unfair, it is in fact a common and necessary practice for both solutions as users would deploy more nodes to avoid filling 100% of the available capacity. In contrast, it is unnecessary to pay for storage growth in advance with BigQuery’s on-demand storage. The Hadoop upfront costs also include software licenses for the operating system, Hadoop, and Hive distributions. ESG calculated the expected maintenance and support contracts on the hardware, as well as the expected power, cooling, and floor space requirements, accounted for as operational costs.

ESG estimated that due to the capacity requirements, a Redshift customer would need to deploy eight ds2.8xlarge compute nodes as well as a single leader node for VPN access to the cluster and to generate queries. Because the three-year reserved instance pricing is paid upfront, ESG chose to account for this cost (as well as an additional 8% APR cost of capital adjustment) as an upfront cost, rather than an on-demand cloud cost. ESG calculated that the small organization could alternatively pay monthly on-demand cloud costs instead of paying all costs upfront. However, over a three-year period, the organization’s cumulative cost would be roughly 75% higher for the AWS solution.

Because there is no hardware to install, configure, maintain, power, and cool, once deployed, the operational costs for both the Redshift and BigQuery solutions would be minimal. ESG estimated the one-time man-hours to migrate the data and test both solutions before putting them into production. Beyond this the BigQuery solution is completely free of additional maintenance and operational costs. The AWS Redshift solution, because it is based on the concept of deploying virtual nodes, would also be expected to require an upfront investment in man-hours to research, plan, and size the deployment to ensure it meets the requirements.

Administration costs for both Redshift and BigQuery were estimated to be far lower than those for the on-premises Hadoop deployment. With BigQuery, administration would be minimal—a single administrator would be expected to manage the account, manage user access, investigate advanced issues, and help guide other analysts to run their own queries via a web-interface. ESG estimates that this would require far less than a full-time hire for smaller deployments, and could be easily accomplished by allocating responsibility to an existing analyst. The Redshift solution would be simpler to manage than the on-premises solution; however, access to the query engine is more complex than for BigQuery, placing more of a burden on the administrator to assist others, and making self-service more difficult for non-technical analysts.

ESG’s model estimated that Google’s BigQuery solution could save a small organization $881K over a three-year period when compared with investing in an on-premises Hadoop cluster, and over $532K when compared with using AWS Redshift. The results of ESG’s three-year TCO analysis for the 100TB “small” modeled organization for each of the three solutions is shown in Figure 3.
**Modeled Scenario #2: Medium Organization**

The medium-sized organization required enough storage to store and protect a 500TB dataset, and performed a mix of batch and ad-hoc queries that on average processed 500 TB of data (100% of the stored data capacity). It should be noted that this 1:1 ratio of stored to processed data is a solid standard for most organizations to plan on, and was validated as the typical case across all BigQuery deployments. The medium-sized deployment also made use of streaming data services, which could be representative of data generated through web or mobile applications, or IoT sensors for example.

ESG estimated that both the on-premises Hadoop cluster and AWS Redshift deployment would require a total of 39 nodes (including the NameNode and leader node). Upfront maintenance and operational costs were modeled for the three systems in the same manner described in the small scenario, scaled of course to accommodate the larger deployment size. Administration of the medium-sized deployment required two full time administrators for the on-premises Hadoop deployment to oversee operations, assist analysts with queries, and administer the Hadoop deployment, database, and hardware. ESG assumed that a single query manager could manage the Redshift and BigQuery deployment in roughly the same number of man-hours as the small deployment. However, ESG assumed a 44% higher hourly rate for the manager based on the scale and complexity of the organization.

In order to accommodate the streaming service, the AWS Redshift solution made use of the AWS Kinesis service to continuously load data onto the Redshift nodes for querying. The Kinesis service offers on-demand pricing only, which can be complex to calculate. To predict pricing, one must consider throughput requirements, payload size, and data retention (number of shard-hours required, number of PUT payload units per million, and optional requirement to keep data longer than 24 hours). In contrast, BigQuery streaming inserts are simply billed per GB inserted.

ESG’s model estimated that Google’s BigQuery solution could save a medium-sized organization more than $2M over a three-year period when compared with investing in an on-premises Hadoop cluster, and over $1.7M when compared with using AWS Redshift. The results of ESG’s three-year TCO analysis for the 500TB “medium” modeled organization for each of the three solutions is shown in Figure 4.
Modeled Scenario #3: Large Organization

The large-sized organization required enough storage to store and protect a 1PB dataset, and performed a mix of batch and ad-hoc queries that on average processed an enormous 4 PB of data (400% of the stored data capacity). The large-sized deployment also made use of a larger percentage of streaming data services (5% of queries processed).

ESG estimated that both the on-premises Hadoop cluster and AWS Redshift deployment would require a total of 76 nodes (including the NameNode and leader node). Upfront maintenance and operational costs were modeled for the three systems in the same manner described in the small scenario, scaled to accommodate the larger deployment size. Administration of the large-sized deployment required three full time administrators for the on-premises Hadoop deployment to oversee operations, assist analyst with queries, and administer the Hadoop deployment, database, and hardware. ESG assumed an 11% higher hourly rate (compared with the medium-sized scenario) for the administration of the BigQuery and Redshift solutions based on the large scale and complexity of the organization.

ESG’s model predicted that Google’s BigQuery solution could save a large organization more than $2.7M over a three-year period when compared with investing in an on-premises Hadoop cluster, and over $2.3M when compared with using AWS Redshift. The results of ESG’s three-year TCO analysis for the 1PB “large” modeled organization for each of the three solutions is shown in Figure 5.
What the Numbers Mean

ESG’s models show that Google’s BigQuery service provides a simple and economical solution for organizations of all sizes. ESG’s models predict that organizations can expect to generate insight at a cost that is 60% to 88% lower than deploying and managing an on-premises Hadoop deployment, and 56% to 82% lower than utilizing the AWS Redshift service over a three-year period.

The big win, however, may be some of the “softer” benefits that BigQuery provides. BigQuery does not require an upfront investment, provides a faster time to value, is easier to manage for departmental use, and allows for more analysts to perform their own queries. The result is that organizations can spend less time managing software, hardware, and queries, and more time generating valuable insight.

Model Considerations: Pricing Options

On-demand cloud services often make sense for young and smaller organizations that are looking to get started in analytics without making a large upfront investment. Larger and more established organizations, however, often struggle to make sense of the numerous and complex pricing schemes that some cloud providers offer. When analyzing the investment options between an on-premises and AWS Redshift deployment, over 20 options may have to be modeled and considered (on-demand versus upfront pricing; one-year or three-year term; and nothing, partial, or all upfront, across four instance types, plus the Kinesis service). In contrast, modeling BigQuery pricing is simple.

ESG’s models found that when looking to minimize costs over a three-year period, it almost always makes the best financial sense to pay the three-year term, all upfront pricing option for AWS Redshift. The savings of up to 75% nearly force organizations to opt to pay long in advance of value realization. Figure 6 shows a sample analysis based on the “medium” scenario, comparing the highest and lowest cost AWS Redshift pricing options to BigQuery’s on-demand pricing.
The differences in these costs are more apparent when they’re broken out into one time “upfront” costs that are paid before any value is realized from the solution, and recurring “monthly” costs, which are paid as the value is realized from the solution. The differences between the solutions are shown in Figure 7.

Figure 7. Upfront versus Monthly Cost Comparison for “Medium” Scenario
Why This Matters

If a cloud service is not easy to plan, purchase, and deploy, organizations will have a harder time justifying the investment to decision makers. While some organizations may enjoy the option of paying upfront, or find value in consistent predictable pricing, the decision should be straightforward and should not result in being locked into a technology or vendor. This defeats the purpose of a service.

Google understood this when it created a simple, serverless on-demand pricing model for its BigQuery service. Organizations that demand dedicated resources or price predictability can decide to pay the monthly flat rate fee, or increase their compute requirements if needed. However, the majority of customers will benefit from the simple on-demand pricing and advances in server technology, with the freedom to change their mind at any time.

The Bigger Truth

Even though the smoke-filled days of the industrial age have long past, the lessons learned remain. The challenges of industry once included how to quickly and cost-effectively mine, transport, and convert raw materials into usable products—businesses today are similarly tasked with collecting, manipulating, and converting raw data to yield actionable intelligence and valuable insight. Time has not transformed the technique for success: keep costs low and predictable, while simplifying operations, avoiding bottlenecks, and delivering maximum value in the minimal amount of time.

Like our industrial-age predecessors, the primary methodology employed by today’s organizations is to make a large upfront investment in physical infrastructure (a large factory), sized to meet the expected demand for the foreseeable future at great financial risk of over- or under-provisioning equipment (servers, network, storage, and software) and under- or over-staffing human resources (hardware and software administrators and experts). As the industrialists eventually found out, products can be produced with greater agility and less risk by leveraging massive global operations that specialize in production—while the company could spend more cycles maximizing the monetary value of the end-product.

Today’s analytics challenges are no different. Many organizations are simply not large enough to justify spending valuable time, resources, and money building, managing, and maintaining an on-premises Hadoop infrastructure when Google’s BigQuery on-demand, serverless technology can provide insight more cost-effectively, in a greatly simplified manner, and with less financial risk to the organization.

ESG’ models, built on the results of validation with BigQuery customers, show that organizations can expect to save between $881K and $2.7M over a three-year period by leveraging BigQuery instead of planning, deploying, testing, managing, and maintaining an on-premises Hadoop cluster. The models also show that BigQuery’s serverless design and simple pricing can provide a solution that is simpler to manage at a total cost that is between 56% and 82% less expensive than using AWS Redshift to store data and perform queries. Perhaps more importantly, the simplicity of the BigQuery solution accommodates economies of scale across departments and enables analysts to execute their own queries, allowing for greater organizational efficiency and quicker insight.

If your organization is looking to spend more time generating more value from your data, and less time managing the solution, while keeping costs to a minimum, then ESG suggests you consider letting Google handle the infrastructure while empowering your organization with the ability to focus on the insight.