GIGAOM

William McKnight, Jake Dolezal
June, 2021

Cloud Analytics Platform Total Cost of Ownership

Cloud Analytics Platform Total Cost of Ownership

Table of Contents

1. Summary

Organizations today need a broad set of enterprise data cloud services with key data functionality to modernize applications and utilize machine learning. They need a platform designed to address multi-faceted needs by offering multi-function data management and analytics to solve the enterprise's most pressing data and analytic challenges in a streamlined fashion. They need a selection that allows a worry-less experience with the architecture and its components.

The platform chosen should bring a multitude of data services onto a single cohesive platform. A key differentiator is the overarching management, deployment, governance, billing, and security. This aspect reduces complexity in administration and scaling data pipelines. As more components are added and more integration points among those components arise, complexity will increase substantially. Greater complexity will lead to more technical debt and administrative burden to cobble together and maintain the flow of data between point solutions.

We decided to take four leading platforms for machine learning under analysis. We have learned that the cloud analytic framework selected for an enterprise, and for an enterprise project, matters to cost.

By looking at the problem from a cost perspective, we've learned to be wary of architectures that decentralize and decouple every component by business domain, which enables flexibility in design, but blows up the matrix of enterprise management needs.

Some architectures look integrated, but in reality, may be more complex and more expensive. When almost every additional demand of performance, scale, or analytics can only be met by adding new resources, it gets expensive.

Based on our approach described in the next section, and using the assumptions listed in each section mimicking a medium enterprise application, Azure was the lowest cost platform. It had a cost of $1.6M for a one-year (annual) cost to purchase the analytics stack. AWS was 9.8% higher, Google 46% higher and Snowflake was 2.5 times higher.

Highlights of the Azure stack include Synapse, Synapse SQL Pool, Azure Data Factory, Azure Stream Analytics, Azure Databricks Premium Tier, HDInsight, Power BI Professional, Azure Machine Learning, Azure Active Directory P1, and Azure Purview. The AWS stack includes Amazon Redshift, Glue, Kinesis, EMD, Spectrum, Quicksight, SageMaker, IAM, and AWS Glue Data Catalog. The Google stack is BigQuery, Dataflow, Dataproc, Cloud IAM and Google Data Catalog. We labeled the 4th stack Snowflake since that is the featured vendor for dedicated compute, storage, and data exploration, but it is really a multi-vendor heterogeneous stack. This includes Talend, Kafka Confluent Cloud, Azure Databricks Premium Tier, Cloudera Data Hub + S3, Tableau, SageMaker, Amazon IAM, and Alation Data Catalog.

Azure was also the lowest cost platform for large enterprises at $4.7M one-year (annual) cost to purchase. AWS was 19% higher, Google 31% higher and the Snowflake stack was nearly 2.5 times

higher.

Dedicated compute is the largest configuration cost, ranging from 43% for the AWS stack to 79% for the Google stack. The Data Lake is second in three stacks but interestingly associated with minimal cost in the Google stack. The data catalog is expensive in the Snowflake stack due to the use of Alation.

In a 3-year total cost of ownership, which includes people cost, for medium enterprises, Azure offers the lowest cost of ownership at $6M. AWS is $7M, Google $11M, and Snowflake $15M. For large enterprises, Azure is $17M, AWS $24M, Google $27M, and Snowflake $42M.

2. Modernizing Your Use Case

Calculating the total cost of ownership in projects is something that happens formally or informally for many enterprise programs. It is also occurring with much more frequency than ever. Sometimes, well-meaning programs will use TCO calculations to justify a program but the measurement of the actual TCO on the flip side can be a daunting experience, especially if the justification TCO was entered into lightly.

Perils of TCO measurement aside, enterprise applications should be attaining high returns. However, if the application is not being implemented to a modern standard, using a machine learning platform as described herein, there are huge inefficiencies and competitive gaps in the functionality. Therefore, many enterprises are considering leveling up or migrating these use cases now, and reaping the benefits.

This paper will focus on the platform costs for medium-and large-sized configurations, broken down by category, across four major platforms: Synapse, Snowflake, Redshift and BigQuery.

The categories, or components in a modern enterprise analytics stack, that we included in our TCO calculations are as follows:

- Dedicated Compute
- Storage
- Data Integration
- Streaming
- Spark Analytics
- Data Exploration
- Data Lake
- Business Intelligence
- Machine Learning
- Identity Management
- Data Catalog

A performance test using the Gigaom Analytic Field Test queries (derived from the TPC-DS) was used to establish equivalency for our pricing and help determine the medium- and large-sized configurations of the four platforms.

Since each platform prices their services differently (with no way to align on hardware specifications), we did our best to align on overall price, so as to not give any of the four platforms a bottom-line advantage in this category. For time-based pricing (e.g., per hour), we assumed 24/7/365 operations. We leave it to the reader to judge the fairness of all our decisions.

In addition to these configuration components, the labor cost factors for the following functions are estimated using our cost multipliers for migrating to data warehouse ecosystems on AWS, Azure, AWS, Google, and Snowflake:

- Data Migration
- ETL Integration
- Analytics Migration
- On-going Support and Continuous Improvement

We then rated each of the platforms across complexity of maintenance, complexity of setup and complexity of operation and administration and came up with a support and improvement cost, to arrive at our final three-year total cost of ownership figures for the study.

3. Performance Comparison

A performance test was used to establish equivalency for our pricing of the four platforms. The GigaOm Analytic Field Test, used across all vendors, was designed to emulate the TPC Benchmark 122; DS (TPC-DS)[1] and adhered to its specifications. **This was not an official TPC benchmark.** The queries were executed using the setup, environment, standards, and configurations described below. For more details on how the testing was conducted, see the Appendix.

Field Test Results

This section analyzes the query results from the fastest of the three runs of the GigaOm Analytic Field Test queries (derived from the TPC-DS) described in the Appendix. The primary metric used was the aggregate total of the best execution times for each query. Three power runs were completed. Each of the 103 queries (99 plus part 2 for 4 queries) was executed three times in order (1, 2, 3, … 98, 99) against each vendor cloud platform, and the overall fastest of the three times was used as the performance metric. These best times were then added together to obtain the total aggregate execution time for the entire workload.

As previously mentioned, the best total aggregate execution time was taken for the entire workload. The following chart shows the overall performance of each platform in terms of total time it took to execute the entire set of 103 queries in the GigaOm Analytic Field Test.

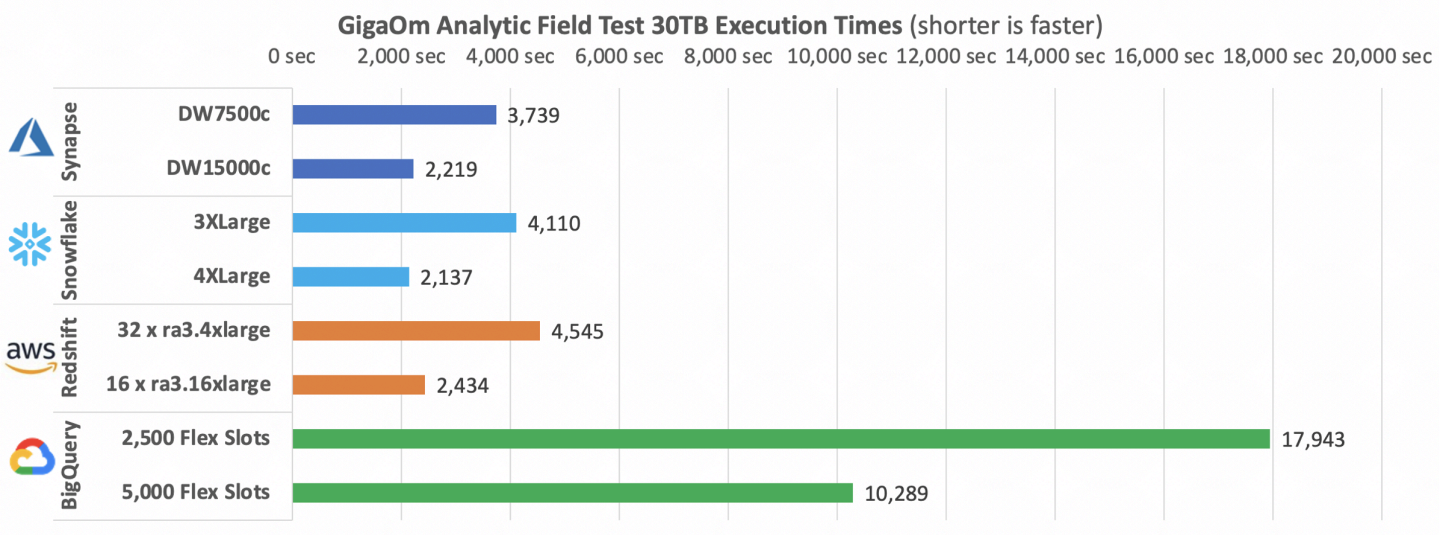

Figure 1: Analytic Field Test 30TB

Price Per Performance

The next step was to determine the price-for-performance. Our goal was to have price-performance

be as close as possible across all four platforms. This would allow us to choose the right size of platform to include in our TCO calculations.

The price-performance metric is dollars per query-hour ($/query-hour). This is defined as the normalized cost of running the GigaOm Analytic Field Test workload on each of the cloud platforms. It was calculated by multiplying the best on-demand rate (expressed in dollars) offered by the cloud platform vendor (at the time of testing) times the number of computation nodes used in the cluster and by dividing this amount by the aggregate total of the best execution times for each query (expressed in hours). Table 1 gives the price of each platform configuration we tested. Below that, if you run all 103 of these queries contiguously to completion of the set, the price-performance of the workload is indicated in figure 2 below.

Table 1: Price per Platform

Synapse		Snowflake		Redshift		BigQuery	
DW7500c	DW15000c	3XLarge	4XLarge	32 x ra3.4xlarge	16 x ra3.16xlarge	2,500 Flex Slots	5,000 Flex Slots
$90.00	$180.00	$256.00	$512.00	$104.32	$208.64	$100.00	$200.00

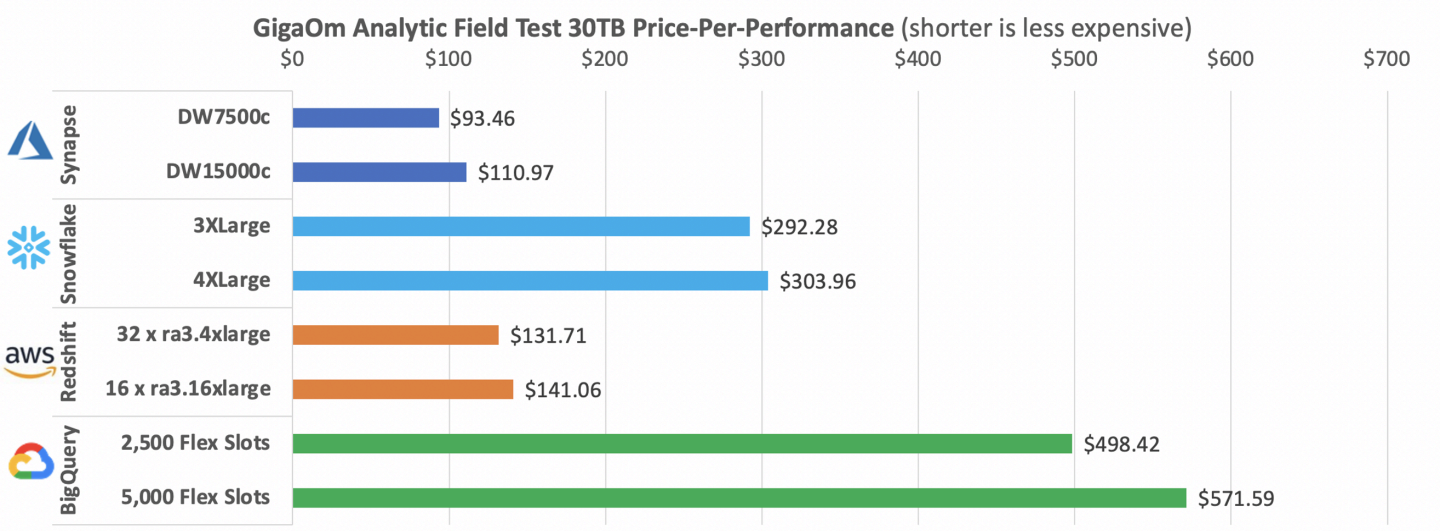

Figure 2: Price Performance of the Workloads

As you can see, we could not achieve similar price-per-performance. For Azure, we used the DW7500c for the Medium-sized configuration and DW15000c for Large that we originally tested. For

Snowflake, 3XLarge for Medium and 4XLarge for Large are as large a configuration as Snowflake offers, so we used those in our TCO calculations. For Redshift, price-performance was closer, but still higher than Azure. We've found Redshift to perform more closely to our Azure and Snowflake Medium-sized configurations with 38 nodes of ra3.4xlarge. Thus, we rounded this result to an even 40 nodes of ra3.4xlarge for Medium and 20 nodes of ra3.16xlarge for Large. Finally, BigQuery was the hardest to estimate, since its price-per-performance was so expensive. Ultimately, we opted for a 10,000 annual slot commitment for Medium organizations using BigQuery and a 20,000 annual slot commitment for Large enterprises.

---

1. More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

4. Total Cost of Ownership

A full analytics stack in the cloud is more than just a data warehouse, cloud storage, and a business intelligence solution. This TCO study required consideration of eleven (11) categories to establish both equivalence among the four analytics platforms' offerings and a fair estimate of pricing. In our experience, all of these components are essential to having a full enterprise-ready analytics stack.

Though the capabilities of the components across the stacks are not equal, all four stacks have been utilized successfully to build machine learning applications. Every component has six enterprise needs which must be met. These are: Security and Privacy, Governance and Compliance, Availability and Recovery, Performance and Scalability, Skills and Training, and Licensing and Usage. The place where the capability difference is made up is in labor (addressed below) using our "cost multipliers."

These stacks can be used for a variety of machine learning applications including customer analytics, fraud detection, supply chain optimization and IoT analytics. Of course, each application could use a slightly different set of components, or quantity of each component. Our vision for developing our stacks is based on customer analytics.

Primary Components

The categories or components in a modern enterprise analytics stack that we included in our TCO calculations are as follows:

- Dedicated Compute
- Storage
- Data Integration
- Streaming
- Spark Analytics
- Data Exploration
- Data Lake
- Business Intelligence
- Machine Learning
- Identity Management
- Data Catalog

Dedicated Compute

The dedicated compute category represents the heart of the analytics stack—the data warehouse itself. A modern cloud data warehouse must have separate compute and storage architecture. The power to scale compute and storage independently of one another has transitioned from an industry trend to an industry standard. The four analytics platforms we are studying have separate pricing models for compute and storage. Thus, this TCO component deals with the costs of running the compute portion of the data warehouse. Storage of data comes next.

Table 2: The Four Offerings of the Vendor Stacks for Dedicated Compute

	Vendor Offering	*Pricing Used*
	Azure Synapse Analytics Workspace	Pay as you go ($1.20/hour per 100 DWU)[2]
	Amazon Redshift RA3	1-year commitment all-upfront ($8.61 effective hourly)[3]
	Google BigQuery	Annual slot commitment ($1,700 per 100 slots)[4]
	Snowflake Computing	Enterprise+ ($4.00 per hour per credit)[5]

For Azure, we opted for the recently released unified workspace experience in Azure Synapse Analytics. While you can purchase reserved capacity for a dedicated SQL pool resource under their legacy pricing model, at the time of this writing, reserved capacity pricing was not available for Azure Synapse Analytics Workspace. For AWS, we chose its latest RA3 family of clusters. Redshift RA3 includes addressable Redshift Managed Storage in its price. However, there is a separate storage charge that falls into the separate Storage category discussed in the next section. For Google, we chose BigQuery with dedicated slots, which is much more economical than their on-demand pricing model that charges $5 for each TB scanned by each query. For a stack built on Snowflake, the only choice is, of course, Snowflake. We used its Enterprise+ Azure pricing model, which offers multi-cluster capabilities, HIPAA support, PCI compliance, and disaster recovery.

Azure Synapse and Amazon Redshift both have the ability to pause compute (and therefore billing) manually when the resource is not needed. With Snowflake, it automatically pauses itself after a period of inactivity determined by the customer (e.g., 5 minutes) or upon first query with auto-wake. These three platforms also allow you to scale the compute size up and down. Utilizing these features could result in some savings. With a BigQuery annual slot commitment, you get the best pricing, but there is no way to "pause" billing—you would need more expensive Flex Slots for that.

To determine the number of DWU, nodes, slots, and credits needed for Synapse, Redshift, BigQuery, and Snowflake, respectfully, we used our field test to determine a like-for-like performance equivalence test. See the section on Performance Comparison for a rundown of this methodology and results.

We assumed the data warehouse in a modern enterprise will be running 24/7/365. Thus, we priced it as running 8,760 hours per year.

Storage

The dedicated storage category represents storage of the enterprise data. In former days, this data was tightly-coupled to the data warehouse itself, but modern cloud architecture allows for the data to be stored separately (and priced separately).

Table 3 : The Four Offerings of the Vendor Stacks for Storage

	Vendor Offering	*Pricing Used*
	Azure Synapse Analytics SQL Pool	$0.023 per GB-month[6]
	Amazon Redshift Managed Storage	$0.024 per GB-month[7]
	Google BigQuery Storage	$0.023 per GB-month (uncompressed)[8]
	Snowflake Computing Storage	$0.04 per GB-month[9]

For all four vendor stacks, we chose the de facto storage that comes with each Dedicated Compute component discussed above. While these come with the compute resources, they are priced separately according to the size of the customer's data.

We priced storage at 30TB of uncompressed data (compressed to 7.5TB) for the medium-tier enterprise and 120TB of uncompressed data (compressed to 30TB) for the large-tier enterprise for Synapse, Redshift, and Snowflake. BigQuery prices data storage for uncompressed size.

Data Integration

The data integration category represents the movement of enterprise data from source to the target data warehouse through conventional ETL (extract-transform-load) and ELT (extract-load-transform) methods.

Table 4: The Options for Each of the Vendor Stacks We Chose for Data Integration

	Vendor Offering	*Pricing Used*
	Azure Data Factory (ADF)	$0.25 per DIU-hour + $1.00 per 1,000 activity runs[10]
	AWS Glue	$0.44 per DPU-hour[11]
	Google Dataflow (Batch)	$0.0828 per worker-hour[12]
	Talend Cloud Data Integration	$12,000 per user per year[13] + compute (Azure VM E16a v4 at $1.008 per hour)[14]

For Azure, we considered the Data Factory pipeline orchestration and execution using integration runtime pricing and Data Integration Unit (DIU) utilization. The compute power of a DIU is not published on Azure's website. AWS Glue was priced for ETL jobs using Data Processing Units (DPU). A single Glue Data Processing Unit (DPU) provides 4 vCPU and 16 GB of memory. Google Dataflow pricing units are worker-hours. A default Dataflow worker provides 1 vCPU and 3.75 GB memory. Snowflake does not offer a Data Integration solution. Thus, a third-party solution was chosen—Talend Cloud Data Integration. The Talend pricing we used was by user per year, but you would also incur a

separate charge from a cloud vendor for the use of virtual machines on which to run Talend.

For Glue and ADF, we considered 64 units for the Medium-tier enterprise and 256 units for Large of DPU and DIU, respectively, running for 8,760 hours per year (24/7/365). For Dataflow, we used 256 for Medium-tier and 1,024 for Large (since its default workers have one-fourth the compute power of a DPU). For ADF, we also priced 1,000 activity runs per month for Medium and 4,000 runs per month for Large. For Talend on Snowflake, we chose 16 users for Medium and 64 for Large. We also figured in 16 Azure E16a VMs to run Talend on for the Medium organization and 64 for Large (since they have 4x the compute power of a Glue DPU).

Streaming
The streaming category represents the movement of enterprise data via a streaming workload from event-driven and IoT (Internet of things) sources.

Table 5: The Options for Each of the Vendor Stacks We Chose for Streaming

	Vendor Offering	*Pricing Used*
	Azure Stream Analytics (for Analytics) and Azure Event Hubs	$0.11 per streaming unit (SU) per hour[15] + $0.03 Standard Throughput unit (1 MB/s ingress, 2 MB/s egress)[16]
	Amazon Kinesis Data Analytics	$0.11 per KPU-hour + $0.10 per GB-month for running application storage[17]
	Google Dataflow (Streaming)	$0.352 per worker-hour[18]
	Confluent Cloud (Kafka)	$1.50 base per hour + $0.12 per GB write + $0.05 per GB read[19]

For each of the four platforms, we made reasonable assumptions about the workload requirements of the Medium- and Large-tier configurations. Since each platform prices its Streaming services in vastly different ways (with no way to align on hardware specifications), we did our best to align on overall price, so as not to give any of the four platforms a bottom-line advantage in this category. For time-based pricing (e.g., per hour), we assumed 24/7/365 operations. We leave it to the reader to judge fairness in our decisions.

Spark Analytics
The Spark analytics category represents the use of Apache Spark for data analytics workloads.

Table 6: The Options for Each of the Vendor Stacks We Chose for Spark Analytics

	Vendor Offering	*Pricing Used*
	Big Data Analytics with Apache Spark	$0.143 per vCore-hour[20]
	Amazon EMR + Kinesis Spark	$1.26 for EMR on r5.4xlarge per hour[21] + $0.015 per shard-hour for Kinesis
	Google Dataproc	$0.01 per CPU-hour + $1.0481 per hour (n2-highmem-16)[22]
	Amazon EMR + Kinesis Spark	$1.26 for EMR on r5.4xlarge per hour[23] + $0.015 per shard-hour for Kinesis

Again, we made reasonable assumptions about the workload requirements of the Medium and Large-tier enterprises. Since each platform prices its Spark services in vastly different ways (with no way to align on hardware specifications), we did our best to align on overall price, so as to not give any of the four platforms a bottom-line advantage in this category. In this case, the one exception is Google Dataproc, which is priced significantly lower than the other three competitors. For time-based pricing (e.g., per hour), we assumed 24/7/365 operations. We leave it to the reader to judge fairness in our decisions.

Data Exploration

Table 7: The Four Offerings of the Vendor Stacks for Data Exploration

	Vendor Offering	*Pricing Used*
	Azure Synapse Serverless	$5 per TB-scanned
	Amazon Redshift Spectrum	$5 per TB-scanned + compute ($8.61 effective hourly)[24]
	Google BigQuery	$5 per TB-scanned (On demand rate) [25]
	Snowflake	Enterprise+ ($4.00 per hour per credit)[26]

Only Azure Synapse and Google BigQuery have a “serverless” pricing model that allows users to run queries and only pay for the data they scan and not an hourly rate for compute. Redshift has the Spectrum service to scan data in S3 without loading it into the data warehouse; however, you pay for the data scanned, plus you need a running Redshift cluster at an additional charge. For Snowflake, you pay for the compute, but not for data scanned. For all these scenarios (except Snowflake), we assumed 500TB scanned per month for the Medium-tier enterprise and 2,000TB scanned for Large organizations.

Data Lake

The data lake category represents the use of a data lake that is separate from the data. This is common in many modern data-driven organizations as a way to store and analyze massive data sets of “colder” data that don’t necessarily belong in the data warehouse.

Table 8: The Options for Each of the Vendor Stacks We Chose for Data Lake

	Vendor Offering	*Pricing Used*
	Azure HDInsight	$1.313 per hour (HDI on E16a v4)[27]
	Amazon EMR	$1.26 per hour (EMR on r5.4xlarge)[28]
	Google Dataproc	$0.01 per CPU-hour + $1.0481 per hour (n2-highmem-16)
	Cloudera Data Hub + AWS S3	$0.5333 per hour (CDP)[29] + $1.008 per hour (r5.4xlarge)[30] + $0.023 per GB-month (AWS S3)[31]

For this comparison, we set aside performance considerations and aligned on price. The exception,

again, is Google Dataproc, which is priced considerably lower than its competitors per hour.

Business Intelligence
The business intelligence category represents the use of a BI tool to complement the data warehouse for business users.

Table 9: The Options for Each of the Vendor Stacks We Chose for Business Intelligence

	Vendor Offering	*Pricing Used*
	PowerBI Professional	$9.99 per Developer per month[32]
	Amazon Quicksight	$18 per Author per month + $5 per Reader per month[33]
	Looker with Google BigQuery BI Engine	$0.0416 per GB-month[34]
	Tableau	$70 per Creator per month + $42 per Explorer per month + $15 per Viewer per month[35]

For this comparison, we did not consider features and capabilities. Amazon Quicksight and BigQuery BI Engine are not as mature or as fully-capable as PowerBI or Tableau, and therefore may be less effective for the workload. That characteristic is difficult to quantify.

We did however use the same number of users for each. Note that PowerBI is the most economical because it only charges for Developers (the same as Creator/Author in Quicksight and Tableau). For Developer/Creator/Author, we chose 100 for Medium-tier and 500 for Large-tier enterprises. For Quicksight, we chose 1,000 and 5,000 Readers for Medium/Large, respectively. For Tableau, we divided this population into 800 Viewers and 200 Explorers for Medium and 4,000 Viewers and 1,000 Explorers for the Large tier. For BigQuery BI Engine, we priced it to scan the entire data warehouse once a week (30TB for Medium and 120TB for Large), although this usage is difficult to predict and budget for.

Machine Learning
The machine learning category represents the use of a machine learning and data science platform on top of the data warehouse and/or data lake.

Table 10: The Options for Each of the Vendor Stacks We Chose for Machine Learning

	Vendor Offering	*Pricing Used*
	Azure Machine Learning	$0.504 per hour (ML on E8 v3)[36]
	Amazon Sagemaker	$0.504 per hour (ml.r5.2xlarge) + $0.202 per hour (Sagemaker)[37]
	Google BigQuery ML	$5 per TB-scanned + $25,000 M/$100,000 L (estimated) per year for model creation[38]
	Amazon Sagemaker	$0.504 per hour (ml.r5.2xlarge) + $0.202 per hour (Sagemaker)[39]

For this comparison, we set aside feature considerations and aligned on price once again. Azure ML is

free as a service, so you only pay for the compute. Snowflake and Amazon Sagemaker have some integrations that allow you to couple them together. BigQuery ML uses its on-demand pricing model outside of the slot commitment you may have purchased. BigQuery ML also charges different fees for model creation. We chose an estimated $25,000 budget per year for the Medium-sized configuration and $100,000 for Large to cover these fees. For Azure ML and Sagemaker, we priced 16 nodes for Medium and 64 nodes for Large enterprises. Also note that with BigQuery ML, you pay additional fees per model created. We did not factor this into our pricing, because it would impact the bottom line much less than 1% of the overall total.

Identity Management
The identity management category represents the integration of users through IAM (identity and account management).

Table 11: The Options for Each of the Vendor Stacks We Chose for Identity Management

	Vendor Offering	*Pricing Used*
	Azure Active Directory	$6 per user per month[40]
	Amazon IAM	free
	Google Cloud IAM	free
	Amazon IAM	free

For this comparison, only Azure charges for the use of Active Directory. While on the surface, the free IAM services of the other platforms are attractive, many organizations have sophisticated security and IAM that need to integrate with on-premise security and single sign-on (SSO). For this reason, Azure Active Directory is a popular choice, especially among organizations that already used Windows Active Directory for on-premises security.

Data Catalog
The data catalog category represents the use of data governance and a centralized data catalog for all data assets.

Table 12: The Options for Each of the Vendor Stacks We Chose for Data Catalog

	Vendor Offering	*Pricing Used*
	Azure Purview	$0.342 per Capacity unit-hour[41]
	Amazon Glue Data Catalog	$1 per 100K objects[42]
	Google Data Catalog	$10 per 100K API calls[43]
	Alation Data Catalog	$198K for 25 contributors[44] + $1.008 per hour (r5.4xlarge)

For this comparison, we used our best educated estimation. Snowflake is at the greatest cost disadvantage because we chose Alation. Alation is an excellent (but relatively costly) data catalog

platform with full data governance features. Alation is a Snowflake partner.

Annual Subtotals

Taking all the above pricing scenarios into consideration, the following figure and tables are the one-year (annual) cost to purchase the analytics stack from each cloud vendor.

Table 13: Medium Size Enterprise One-Year (Annual) Cost to Purchase an Analytics Stack

	aws			
Medium Enterprise				
Dedicated Compute	$753,360	$788,400	$2,242,560	$2,040,000
Storage	$2,160	$2,070	$3,600	$8,280
Data Integration	$246,682	$152,160	$333,281	$185,771
Streaming	$65,270	$78,490	$74,810	$74,012
Spark Analytics	$52,560	$40,086	$52,560	$42,332
Data Exploration	$105,336	$30,000	$140,160	$30,000
Data Lake	$353,203	$367,780	$440,337	$338,658
Business Intelligence	$81,600	$11,988	$328,800	$64,896
Machine Learning	$98,953	$70,641	$98,953	$147,880
Identity Management	$0	$72,000	$0	$0
Data Catalog	$1,200	$2,999	$413,660	$12,000
Medium-Tier Annual Subtotal	**$1,760,324**	**$1,616,613**	**$4,128,722**	**$2,943,829**

Table 14: Large Size Enterprize One-Year (Annual) Cost to Purchase an Analytics Stack

	aws			
Large Enterprise				
Dedicated Compute	$1,508,472	$1,576,800	$4,485,120	$4,080,000
Storage	$8,640	$8,280	$14,400	$33,120
Data Integration	$986,726	$608,640	$1,333,125	$743,083
Streaming	$261,082	$313,958	$259,822	$296,047
Spark Analytics	$210,240	$160,343	$210,240	$169,329
Data Exploration	$421,344	$120,000	$560,640	$120,000
Data Lake	$1,412,813	$1,471,119	$1,736,509	$1,354,632
Business Intelligence	$408,000	$59,940	$1,644,000	$259,584
Machine Learning	$395,812	$282,563	$395,812	$591,520
Identity Management	$0	$144,000	$0	$0
Data Catalog	$12,000	$29,959	$827,320	$36,000
Large-Tier Annual Subtotal	**$5,625,129**	**$4,775,603**	**$11,466,988**	**$7,683,316**

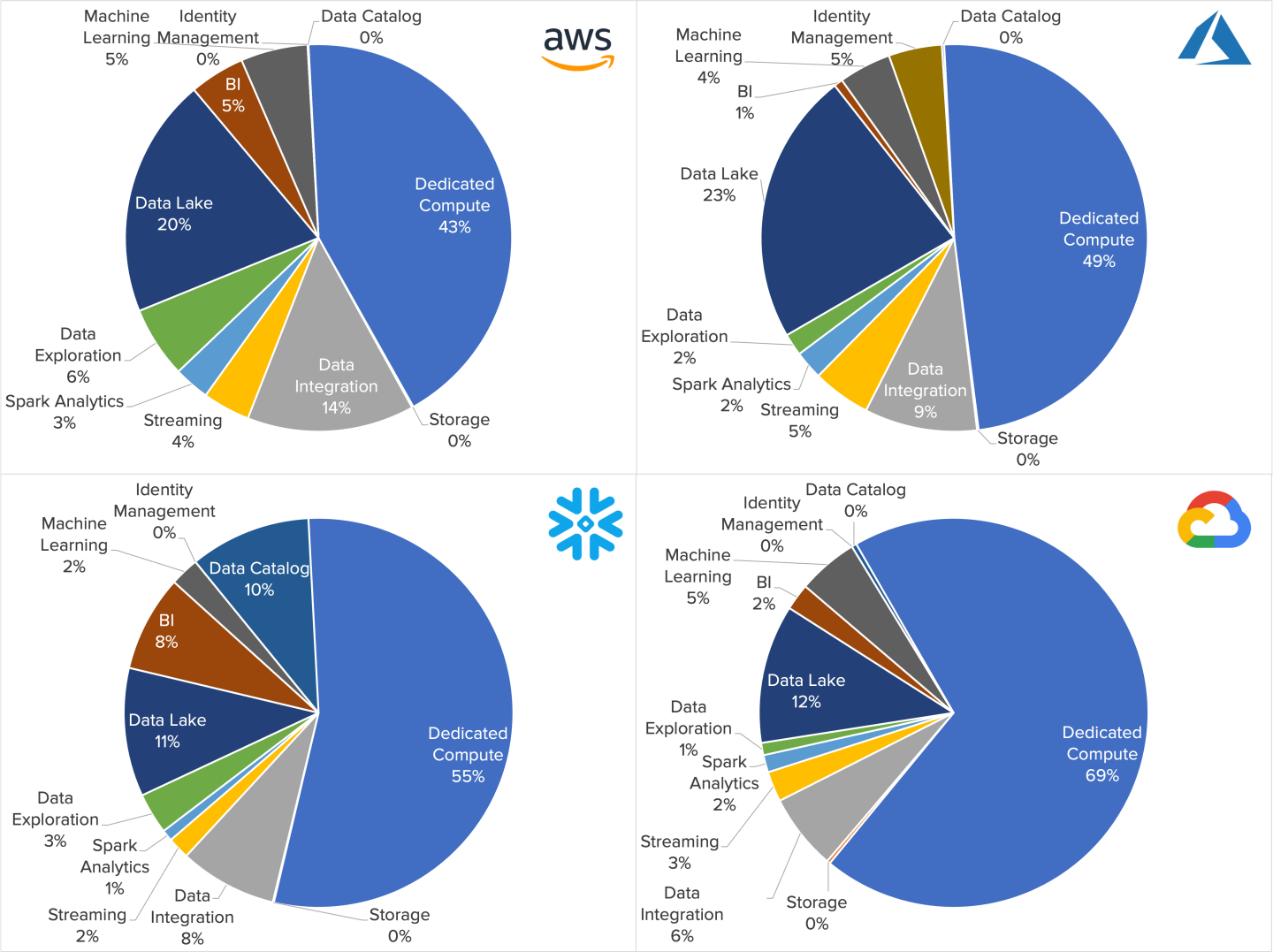

Figure 3: Breakdown by Category (using Medium Enterprise)

Other Costs

In addition to these components, other cost factors were considered. We realize that it takes more than simply buying cloud platforms and tools—namely, people—to achieve a production-ready enterprise analytics stack. These resources are required to build up and maintain the platform to deliver business insight and meet the needs of a data-driven organization.

The additional cost factors include:

- Data Migration
- ETL Integration
- Analytics Migration
- On-going Support and Continuous Improvement

Labor Costs
To calculate TCO, one cannot overlook the people factor in the cost. It is people who build, migrate, and use the analytics platform. To figure labor costs for our TCO calculations, we use a blended rate of both internal staff and external professional services.

Table 15: Estimate of Labor Costs for the Cost Categories Based on Our Industry Experience

Internal Staff	$66.89 per hour
External Services	$150 per hour

For internal staff, we used an average annual cash compensation of $100,000 and a 22% burden rate, bringing the actual cost to a blended $122,000. We also estimated the year to have 1,824 working hours, which gives us an effective hourly rate of $66.89. For external professional services, we chose a nominal blended rate of $150 per hour.

To support both the migration to the cloud data analytics platform and the ongoing maintenance and continuous improvement, we estimated a mixture of internal and external resources.

Table 16: Mixture of Internal and External Resources

	Migration Phase	*Improvement Phase*
Internal Staff	50%	75%
External Services	50%	25%
Blended Rate	$108 per hour	$88 per hour

Migration

To calculate the **base** cost for migrating from an existing platform to the cloud analytics solution, we used the following calculations.

Table 17: Base Cost to Migration from Existing to Cloud Analytics

Migration Component (and primary cost driver)	Items (Medium)	Items (Large)	Migration Effort*	Base (Medium)	Base (Large)
Data Migration (# of DW objects)	819	3,276	0.65	$53,828	$215,310
ETL Migration (# of ETL feeds/workflows)	82	328	7.64	$67,840	$271,359
Analytics Migration (# of reports)	410	1,640	1.21	$53,828	$215,310
Total Migration Base Cost				**$175,495**	**$701,980**

* Average hours to migrate each item

Data migration includes significant data objects (tables, views, stored procedures, triggers, functions, scripts, etc.) that need to be exported from the old data warehouse, transformed, and loaded into the new data warehouse. ETL migration includes the feeds (connections to source app databases/ systems) that need to be migrated to support ongoing data imports into the new data warehouse. Analytics migration includes reports and dashboards that need to be modified to work on (connect to) the new data warehouse.

The migration effort, or the average hours to migrate each item, varies due to the complexity of the environment's legacy artifacts. Migrating from more modern on-premise platforms might be easier than, say, a legacy mainframe. Each organization should do a deep source system analysis to understand the challenges and complexity factor for their given situation. The situation presented here is within what we, in our experience, consider relatively "typical" although your mileage will vary.

In addition, we consider the complexity and difficulty of migrating to the cloud and each of the four platforms we are considering in this paper. Based on experience and the rigorous independent assessment of each platform, we have developed multipliers which measure the degree of difficulty or complexity for each of these three migration categories with the vendor platforms we priced here.

Table 18: Our cost multipliers for migrating to data warehouse ecosystems on AWS, Azure, AWS, Google, and Snowflake

	aws			
Data Migration	2.0	2.0	4.0	3.0
ETL Migration	2.0	1.0	4.0	3.0
Analytics Migration	1.0	0.2	1.0	1.0
Medium Total	**$297,162**	**$186,260**	**$418,830**	**$540,497**
Large Total	**$1,188,649**	**$745,042**	**$1,675,318**	**$2,161,987**

In terms of data migration, most enterprise data warehouses (e.g., Oracle, Netezza, Teradata) have these features: indexing, constraints, replicated tables, heap tables, and user-controlled partitioning, in addition to security features such as row + column security, masking, and column encryption. Redshift, BigQuery, and Snowflake have limited support for them. Without these features, migration can be more difficult.

For ETL migration, code conversion is a big pain point. This can be felt in a smaller way with Redshift because its older PostgreSQL 8 syntax creates limitations or with BigQuery and its own flavor of SQL creating significant time for conversion.

For analytics migration, we mostly considered BI conversion. We assert that PowerBI already has excellent integration with Synapse and that will save conversion time.

For the overall three-year TCO calculation, migration costs will only be applied once.

Ongoing Support and Continuous Improvement
The cost of maintaining and improving the total analytics platform has to be considered as well. This includes support work: database administration, disaster recovery, and security. However, no analytics platform is (or should be) static. Maintenance and operations includes ongoing work in security/privacy, governance and compliance, availability and recovery, performance management and scalability, skills and training, and licensing and usage tracking.

The environment needs constant improvement and enhancement to grow with business needs. The work of project management and CI/CD integration are considered here as well.

The factors involved that can impact these costs include:

- Complexity of maintenance
- Complexity of setup
- Complexity of operation and administration

We rated each of the platforms across these metrics and came up with a support and improvement cost multiplier applied to the base cost (minus migration costs calculated in the previous section).

Table 19: Ongoing Support and Continuous Improvement Cost Multipliers

	aws			
Maintenance	Hard	Easy	Easy	Easy
Setup	Medium	Medium	Hard	Easy
Operation and Administration	Medium	Medium	Hard	Medium
Support/Improvement Multiplier	**35%**	**25%**	**35%**	**20%**
Medium Annual Total	**$616,113**	**$404,153**	**$825,744**	**$1,030,340**
Large Annual Total	**$1,968,795**	**$1,193,901**	**$2,293,398**	**$2,689,160**

Three-Year TCO

Finally, we arrive at our final three-year total cost of ownership figures for the study. The following is a three-year breakdown and grand total for each of the four cloud vendors' full analytics stacks.

Table 20: Three-Year Breakdown with Grand Totals

	aws			
Year 1				
Medium	$2,673,600	$2,207,026	$4,514,666	$5,373,296
Large	$8,782,573	$6,714,545	$12,534,463	$15,435,703
Year 2				
Medium	$2,376,438	$2,020,766	$3,974,169	$4,954,466
Large	$7,593,924	$5,969,503	$10,372,476	$13,760,385
Year 3				
Medium	$2,376,438	$2,020,766	$3,974,169	$4,954,466
Large	$7,593,924	$5,969,503	$10,372,476	$13,760,385
3-YEAR TOTAL COST OF OWNERSHIP				
Medium Enterprise	**$7,426,475**	**$6,248,559**	**$12,463,004**	**$15,282,229**
Large Enterprise	**$23,970,420**	**$18,653,551**	**$33,279,416**	**$42,956,474**

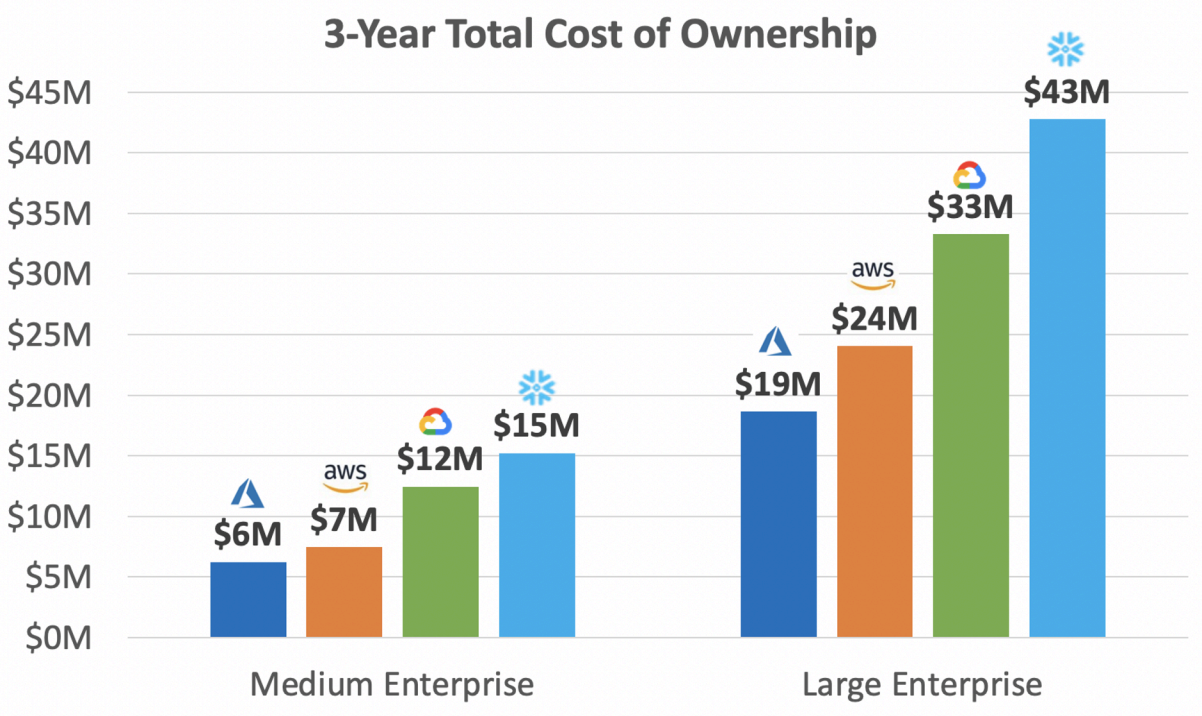

Figure 4: Visual of Our Three-Year Breakdown with Grand Totals

---

2. https://azure.microsoft.com/en-us/pricing/details/synapse-analytics/
3. https://aws.amazon.com/redshift/pricing/
4. https://cloud.google.com/bigquery/pricing
5. https://www.snowflake.com/pricing/
6. https://azure.microsoft.com/en-us/pricing/details/synapse-analytics/
7. https://aws.amazon.com/redshift/pricing/
8. https://cloud.google.com/bigquery/pricing
9. https://www.snowflake.com/pricing/ NOTE: We used on-demand storage and not up-front capacity storage pricing so as to be consistent with the on-demand pricing of the other platforms.
10. https://azure.microsoft.com/en-us/pricing/details/data-factory/data-pipeline/
11. https://aws.amazon.com/glue/pricing/
12. https://cloud.google.com/dataflow/pricing
13. https://aws.amazon.com/marketplace/pp/B0829B4MT2
14. https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/
15. https://azure.microsoft.com/en-us/pricing/details/stream-analytics/
16. https://azure.microsoft.com/en-us/pricing/details/event-hubs/
17. https://aws.amazon.com/kinesis/data-analytics/pricing/
18. https://cloud.google.com/dataflow/pricing
19. https://www.confluent.io/confluent-cloud/pricing/
20. https://azure.microsoft.com/en-us/pricing/details/synapse-analytics/
21. https://aws.amazon.com/emr/pricing/
22. https://cloud.google.com/dataproc/pricing
23. https://aws.amazon.com/emr/pricing/
24. https://aws.amazon.com/redshift/pricing/

25. https://cloud.google.com/bigquery/pricing
26. https://www.snowflake.com/pricing/
27. https://azure.microsoft.com/en-us/pricing/details/hdinsight/
28. https://aws.amazon.com/emr/pricing/
29. https://www.cloudera.com/products/pricing.html
30. https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/
31. https://azure.microsoft.com/en-us/pricing/details/storage/blobs/
32. https://powerbi.microsoft.com/en-us/pricing/
33. https://aws.amazon.com/quicksight/pricing/
34. https://cloud.google.com/bi-engine/pricing
35. https://www.tableau.com/pricing/teams-orgs#online
36. https://azure.microsoft.com/en-us/pricing/details/machine-learning/
37. https://aws.amazon.com/sagemaker/pricing/
38. https://cloud.google.com/bigquery-ml/pricing
39. https://aws.amazon.com/s3/pricing/
40. https://azure.microsoft.com/en-us/pricing/details/active-directory/
41. https://azure.microsoft.com/en-us/pricing/details/azure-purview/
42. https://aws.amazon.com/glue/pricing/
43. https://cloud.google.com/data-catalog/pricing
44. https://aws.amazon.com/marketplace/pp/Alation-Inc-Alation-Data-Catalog/B07GVL2T46

5. Conclusion

Based on our approach, and using the assumptions listed in each section mimicking a medium enterprise application, Azure was the lowest cost platform. It had a cost of $1.6M for a one-year (annual) cost to purchase the analytics stack. AWS was 9.8% higher, Google 46% higher and Snowflake was 2.5 times higher.

Azure was also the lowest cost platform for large enterprises at $4.7M one-year (annual) cost to purchase. AWS was 19% higher, Google 31% higher and the Snowflake stack was nearly 2.5 times higher.

In a 3-year total cost of ownership, which includes people cost, for medium enterprises, Azure is the platform with the lowest cost of ownership at $6M. AWS is at $7M, Google $21M and Snowflake $15M. For large enterprises, Azure is $19M, AWS $24M, Google $33M and Snowflake $43M.

With Azure Synapse there is a single point of management for significant services shared with the Azure platform. Performance and scale for all analytic capabilities are managed together. Skills and training are greatly simplified.

6. Appendix: GigaOm Analytic Field Test

A performance test was used to establish equivalency for our pricing of the four platforms. The GigaOm Analytic Field Test, used across all vendors, was designed to emulate the TPC Benchmark™ DS (TPC-DS)[1] and adhered to its specifications. **This was not an official TPC benchmark.** The queries were executed using the following setup, environment, standards, and configurations.

The GigaOm Analytic Field Test is a workload derived from the well-recognized industry standard TPC Benchmark™ DS (TPC-DS). From tpc.org:

The TPCDS is a decision support benchmark that models several generally applicable aspects of a decision support system, including queries and data maintenance. The benchmark provides a representative evaluation of performance as a general-purpose decision support system... The purpose of TPC benchmarks is to provide relevant, objective performance data to industry users. TPC-DS Version 2 enables emerging technologies, such as Big Data systems, to execute the benchmark.

The data model consists of 24 tables—7 fact tables and 17 dimensions. To give an idea of the data volumes used in our field test, the following table shows row counts of fact tables in the database when loaded with 30TB of GigaOm Analytic Field Test data:

GigaOm Analytic Field Test Table	Scale Factor 30,000 30TB Row Count
Catalog Returns	4,319,925,093
Catalog Sales	43,200,404,822
Inventory	1,627,857,000
Store Returns	8,639,952,111
Store Sales	86,399,341,874
Web Returns	2,160,007,345
Web Sales	21,600,036,511

The GigaOm Analytic Field Test is a fair representation of enterprise query needs. The GigaOm Analytic Field Test testing suite has 99 queries—4 of which have two parts (14, 23, 24, and 39). This brings it to a total of 103 queries. The queries used for the tests were compliant with the standards set out by the TPC Benchmark™ DS (TPC-DS) specification[2] and included only minor query modifications as set out by section 4.2.3 of the TPC-DS specification document. For example, minor query modifications included vendor-specific syntax for date expressions. Also in the specification, some queries require row limits and, thus, vendor specific syntax was used (e.g., TOP, FIRST, LIMIT, and so forth) as allowed by section 4.2.4 of the TPC-DS specification.

Cluster Environments

Our benchmark included four different cluster environments for medium and for large configurations.

	Azure Synapse Analytics Workspace	Amazon Redshift	Google BigQuery	Snowflake
Medium Configuration	DW7500c	ra3.4xlarge (32 nodes)	2,500 Flex Slots	3XLarge
Large Configuration	DW15000c	ra3.16xlarge (16 nodes)	5,000 Flex Slots	4XLarge

1. More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.
2. The TPC Benchmark™ DS (TPC-DS) specification we used was found at http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-ds_v2.1.0.pdf.

7 About William McKnight

William McKnight is a former Fortune 50 technology executive and database engineer. An Ernst & Young Entrepreneur of the Year finalist and frequent best practices judge, he helps enterprise clients with action plans, architectures, strategies, and technology tools to manage information.

Currently, William is an analyst for GigaOm Research who takes corporate information and turns it into a bottom-line-enhancing asset. He has worked with Dong Energy, France Telecom, Pfizer, Samba Bank, ScotiaBank, Teva Pharmaceuticals, and Verizon, among many others. William focuses on delivering business value and solving business problems utilizing proven approaches in information management.

8 About Jake Dolezal

Jake Dolezal is a contributing analyst at GigaOm. He has two decades of experience in the information management field, with expertise in analytics, data warehousing, master data management, data governance, business intelligence, statistics, data modeling and integration, and visualization. Jake has solved technical problems across a broad range of industries, including healthcare, education, government, manufacturing, engineering, hospitality, and restaurants. He has a doctorate in information management from Syracuse University.

9. About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.

10. Copyright