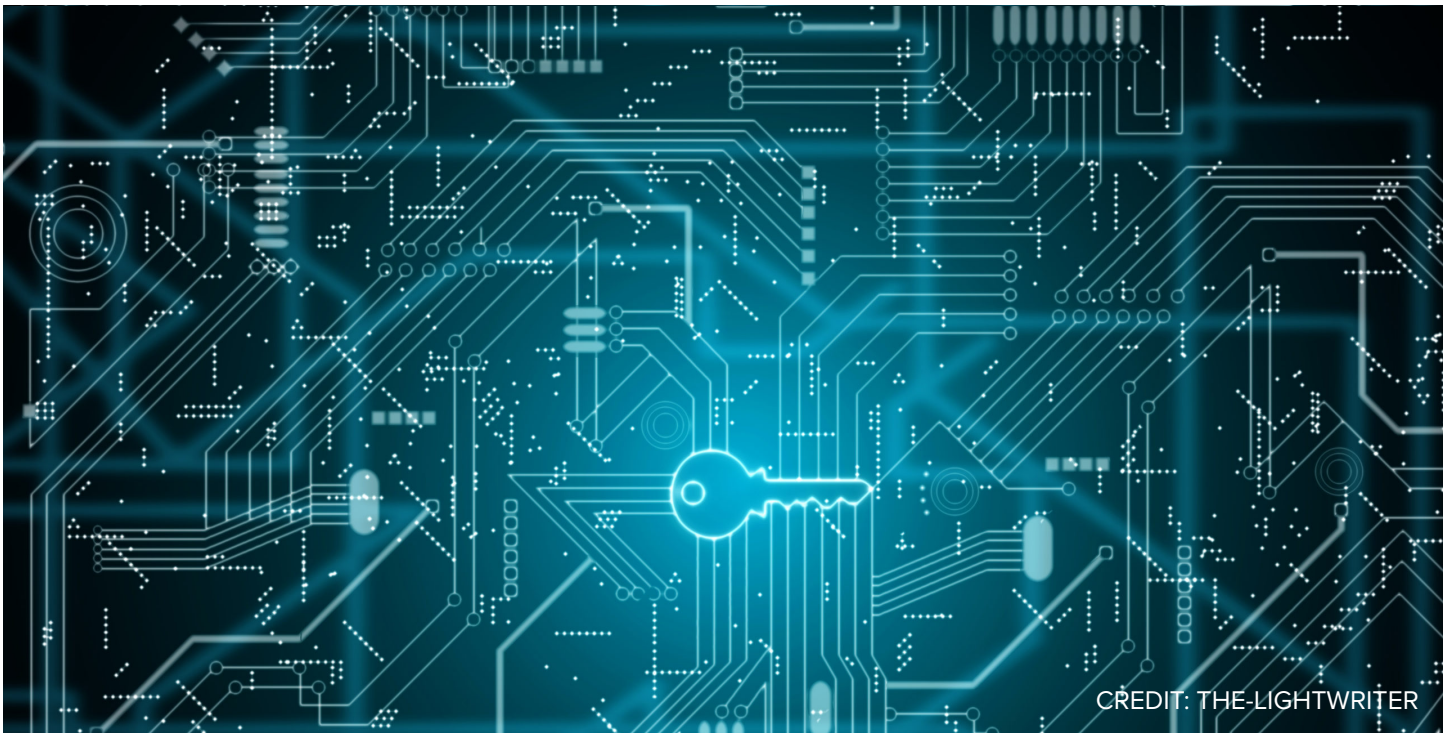
GIGAOM

MARKET LANDSCAPE REPORT

GigaOm Radar for Evaluating Data Warehouse Platforms

ANDREW J. BRUST AND YIANNIS ANTONIOU

TOPIC: DATA WAREHOUSE



GIGAOM

GigaOm Radar for Evaluating Data Warehouse Platforms

TABLE OF CONTENTS

1. Summary

For decades data warehouses have been the trusted technology for large-scale data storage and analytics in the enterprise. And that role has become only more vital, as traditional data warehouse vendors have modernized their products to provide advanced scaling capabilities, massive parallelism, enhanced ease of use, and reduced total cost of ownership. At the same time, vendors are evolving new features and advancing their architectures to leverage the native capabilities of the cloud.

Today, vendors are extending their products, moving from core data warehouse offerings to more integrated platforms with warehouse capabilities at their core. These include integrations with formerly discrete technologies such as data lakes, Hadoop and Spark, as well as AI operations, deep integration with data analytics and other BI tools, and easier integration with data engineering, data science and machine learning workflows. Built-in data governance, data quality and data preparation are also included.

Managing the full data cycle is now virtually impossible for an organization without a data warehouse platform. Evaluating the market's most important vendors is therefore vital for any organization. This process needs to examine both technical and non-technical considerations, and should be approached holistically.

This GigaOm Radar report examines and evaluates the most important data warehouse platforms in the market today. It looks at each vendor's approach, capabilities and, crucially, its ongoing development, and explores how each is poised to evolve over the next twelve months.

This report is designed to help you evaluate both the current and future position of solutions within the market. The aim is to help your organization make the best possible decision about the vendor it selects for its data warehouse.

Key findings:

- New platform development is focused on the cloud. Several cloud-native platforms have emerged, and more traditional vendors have modernized their platforms to transition from on-premises to the cloud.
- Vendors are investing heavily in hybrid cloud capabilities. Data sets spanning on-premises and the cloud can now be processed and analyzed, and disaster recovery capabilities have been enhanced.
- Most vendors are integrating data science, ML, and artificial intelligence (AI) capabilities into their base data warehousing architecture.
- Data warehouses are able to store and analyze data more quickly, allowing huge amounts of data to be explored more readily, and this trend is accelerating.

- SQL is the dominant querying language, though vendors do make their own additions or incorporate special variations.
- Integration with data lakes, query federation, and processing of remote data sets in-place are slowly becoming more prevalent, with several vendors implementing similar performance-enhancing features.

2. About the GigaOm Radar Report

HOW TO READ THIS REPORT

This GigaOm report is part of a series of documents that help IT professionals understand, explore, and evaluate a specific technology and its attendant market. It enables organizations to assess competing solutions in the context of well-defined criteria and metrics. For a fuller understanding, consider reviewing the following reports:

Key Criteria report: A detailed market sector analysis focused on a specific technology domain. The report enables IT decision-makers to make better decisions by defining key features and criteria for a product sector and assessing their impact on core evaluation metrics. This framework provides a strong overview of a technology sector and the solutions and vendors enabling it. The Key Criteria report is critical to informing the GigaOm Radar report.

Radar report: A market landscape analysis that provides a forward-looking evaluation of vendors and their solutions in a specific technology sector. The GigaOm Radar leverages scoring and qualitative analysis to plot a chart that depicts the relative value, character, and progression of vendors' solutions. The Radar report includes a breakdown of each vendor's offering in the sector.

Vendor Profile: An in-depth vendor analysis that provides an accessible, deep dive into a company's engagement with a technology sector. The analysis builds on coverage presented in the Key Criteria and Radar reports, drilling into details of the vendor's solution and assessing the company's strategy as it relates to the market sector. This analysis includes forward-looking guidance around both strategy and product.

3. Key Criteria Comparison

Following the general indications introduced with the Key Criteria for Evaluating Data Warehouse Platforms, **Table 1** summarizes how each vendor included in this research performs in the areas that we consider differentiating and critical for modern data warehouses. The objective is to give the reader a snapshot of the technical capabilities of different solutions and define the perimeter of the market landscape.

Table 1: Key Criteria & Evaluation Metrics Comparison

	KEY CRITERIA					EVALUATION METRICS				
	Multi-Cluster Capabilities	Managed Service	Data Quality & Governance	Data Science ML Integration	Data Lake Int. Query Fed.	Cloud-native Hybrid	Enterprise Focus	BI Tool Integration	Platform Ecosystem	Scaling & Elasticity
Actian	-	+	-	-	+++	+	+	+	+	++
AWS Redshift	-	++	-	++	++	++	++	++	+++	+++
Cloudera Data Warehouse	++	-	+	+	++	++	+	+	+	++
Google BigQuery	-	++	-	+	+	++	++	++	++	+++
IBM Db2	+	+	-	++	-	+	+++	+	+	++
MemSQL	-	+	-	-	+	++	+	-	+	++
Microsoft Azure	-	++	-	++	+	++	+++	+++	+++	+++
Oracle Autonomous	-	+	-	+	+	+	+++	+	+++	++
SAP Data Warehouse Cloud	-	+	+	+	+	+	+++	+	++	++
Snowflake	++	+	-	-	++	++	++	++	++	+++
Teradata Vantage	++	++	-	++	+	++	++	+	+	++
Vertica	++	++	+	++	+++	++	++	++	++	++
Yellowbrick	-	+	-	-	-	+	+	+	+	++

+++: strong focus and perfect fit of the solution
++: The solution is good in this area, but there is still room for improvement
+: The solution has limitations and a narrow set of use cases
-: Not applicable or absent.

Source: GigaOm 2020

4. GigaOm Radar

This report synthesizes the analysis of key criteria and their impact on critical metrics to inform the GigaOm Radar graphic in **Figure 1**. The resulting chart is a forward-looking perspective on all the vendors in this report, based on their products' technical capabilities and feature sets.

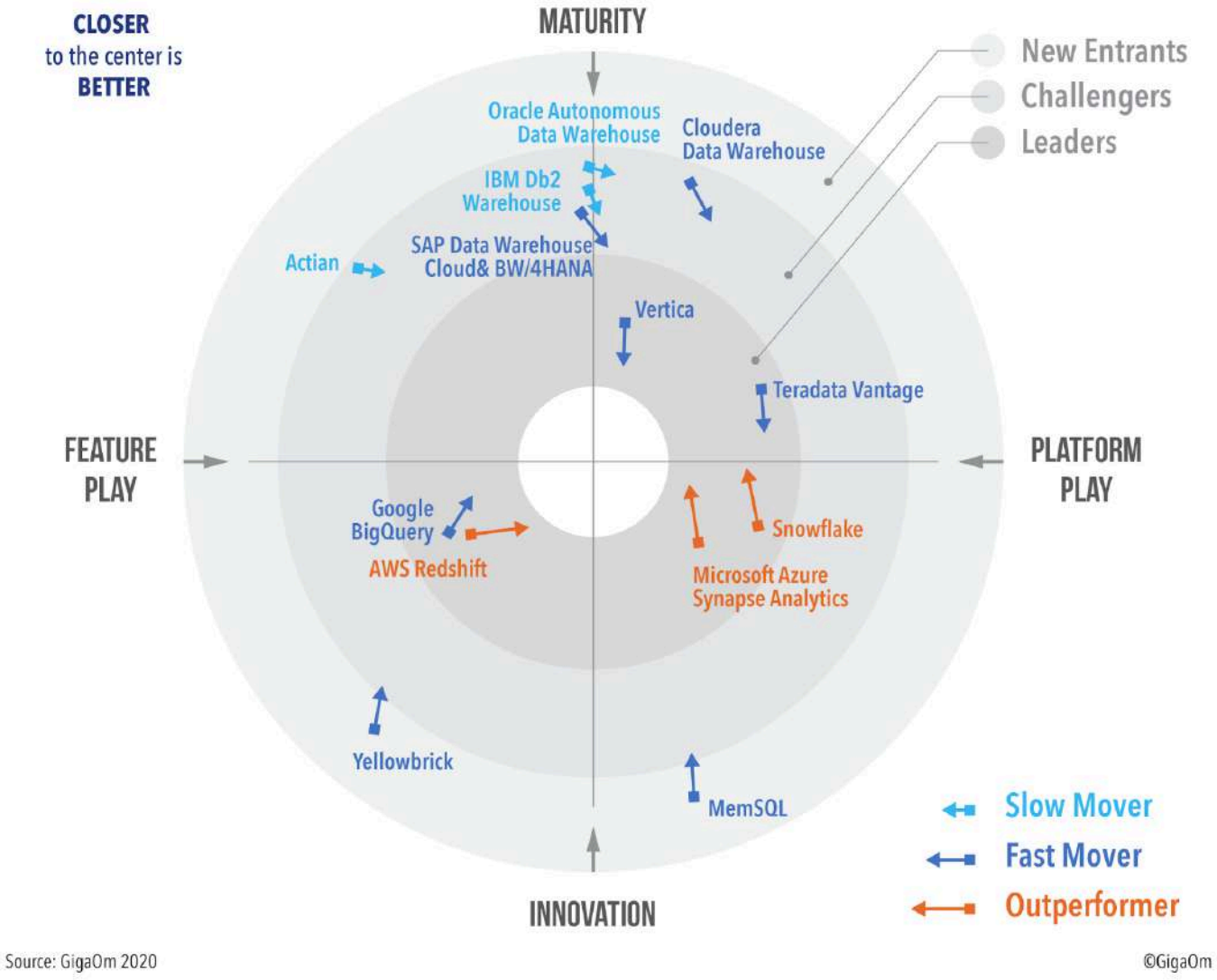


Figure 1: GigaOm Radar for Data Warehouse Platforms

INSIDE THE GIGAOM RADAR

The GigaOm Radar focuses on each vendor's technology roadmap, execution, and ability to innovate. It excludes vendor market share as a metric to yield a forward-looking analysis that emphasizes the value of innovation and differentiation over incumbent market position. The resulting graph plots the relative market position and movement of each vendor across three fundamental data points:

- The current position on the chart provides insight into the present state of each solution
- The direction models the impact of ongoing product strategy and development on the solution
- The vector module shows how quickly the vendor is executing on its vision and strategy

The GigaOm Radar aligns solutions along four characteristics, set in the chart as opposing pairs: Maturity and Innovation, and Feature Play and Platform Play. The closer a solution is to the axis line of a characteristic on the Radar chart, the stronger its execution in that regard:

Maturity: Expresses the stability and user acceptance of the solution, and overall ecosystem sustainability. Vendors on this axis may be more conservative in their approach.
Innovation: Indicates the level of differentiation of the solution from others in the market. Technical innovation and an aggressive approach to the market are often implied here.
Feature Play: Represents a focus on differentiating features and technical aspects, often advanced by niche players, point solutions, and new vendors leveraging cutting-edge tech.
Platform Play: Recognizes solutions that provide a broad, horizontal platform, with a comprehensive feature set and extensive ecosystem.

Finally, the GigaOm Radar is organized into three concentric circles around a bullseye. The closer to center, the better the solution. The three levels are:

Leaders: Vendors that are competing for market leadership in the metrics described above, even if they are competing in different market segments.
Challengers: Vendors with the potential to become a leader soon, niche or traditional players with an established market, and other companies that have interesting solutions but are still maturing.
New Entrants: Usually companies with a limited feature set and too little history to be included in the Leaders or Challengers categories, but with potential to move there soon.

The center-most circle of the GigaOm Radar is almost always empty, reserved for extremely mature and consolidated markets with very few competitors and mature solutions lacking space for further innovation.

5. Vendor Roundup/Overview

In this section, we present our analysis of key players in the data warehouse platform market. We have classified the platforms into two distinct categories:

- Platform offerings from traditional, incumbent vendors
- Platform offerings from newer vendors

Incumbent Vendors

Actian

Actian, originally founded in 1980 as the storied relational database platform Ingres, has gone through a variety of changes, acquisitions, and name changes over the years as it has evolved its offerings to compete in the market. Now owned by HCL Technologies and Sumeru Equity Partners, the Palo Alto, CA-based company offers the Actian Avalanche platform as a fully managed, hybrid cloud-based data warehouse solution, in addition to other OLTP and data integration products.

The Actian Avalanche platform positions itself as a performance leader in the cloud data warehouse space. The platform takes advantage of its built-in Vector engine with support for single instruction multiple data (SIMD) parallelism, columnar storage, CPU cache data compression, and other architectural elements to deliver query performance and real-time updates to data, as it accesses analytical queries.

Other characteristics of Avalanche include federated query capabilities that may be used across on-premises and cloud data sources used simultaneously and without the need to move data; native cloud object storage and data lake capabilities; connectivity to more than 200 on-premises and cloud applications such as Workday and Salesforce and platforms such as Kafka and Hadoop through the Avalanche Connect component; and strong interoperability with a variety of BI and data science tools.

Deployment is supported on-premises, on AWS and Azure in the public cloud (with support for GCP planned), and as a managed private cloud as well.

The Avalanche platform is appealing due to its performance, breadth of features, robust connectivity options, and well-thought-out deployment options. This combination of features should help it maintain and gain market share in an increasingly crowded marketplace.

IBM Db2 Warehouse

IBM's Db2 Warehouse is the Armonk, NY-based software and hardware giant's main on-premises and cloud-based data warehouse offering. The Db2 platform has a long and celebrated history in the

marketplace, starting in the early ‘80s as a transactional database system and evolving over subsequent decades into a set of modern warehouse and transactional systems.

The Db2 Warehouse platform offers highly scalable, columnar, in-memory processing, and SQL-based in-database analytics at its core. On top of this base architecture, data federation allows for the processing of local and remote data in place, significantly increasing performance. A set of capabilities the company calls Adaptive Workload Management also offers target-based resource scheduling and utilization optimizations to ensure further consistent performance in the system.

The platform offers strong support for data science workflows by incorporating the Apache Spark engine and supporting in-database analytics for the R, Python, and Spark SQL programming languages. Several predictive modeling algorithms for clustering, regression, classification, and more are built-in. Geospatial data types are also first-class citizens. Programming support for the C++ and Lua languages and support for the SPSS statistical platform are also present.

The platform is offered for on-premises, public and private cloud deployment through any Docker container-supporting environment, including IBM’s own OpenShift-based Cloud Pak for Data platform. In addition, the IBM Db2 Warehouse on Cloud product expands the platform’s capabilities into the cloud with a fully managed service that is currently available for IBM Cloud and AWS.

The company also offers the IBM Performance Server for PostgreSQL cloud-native product for migration from existing Netezza and PureData analytics appliances. Another hardware appliance product, the IBM Integrated Analytics System, is also available for hybrid SQL and Hadoop-based analytic workloads.

IBM’s main warehouse platform is a strong contender for enterprises with existing IBM investments in software or appliances. The platform should also strongly appeal to data scientists due to its strong integrated support for their workflows. We expect that the company will continue to invest in enhancing the cloud product by extending its deployment options to other public cloud providers and integrating it with their native capabilities.

Microsoft Azure Synapse Analytics

Microsoft’s Azure Synapse Analytics is the Redmond, WA-based software giant’s cloud-native data warehouse platform for the future. Announced at the end of 2019, it builds on and enhances the previous Azure SQL Data Warehouse solution. We examine the main features of the platform below.

The Azure Synapse platform is designed to blend concepts from data warehousing, big data, data science, and data integration to enable data analytics at scale. The platform features a columnar, compressed, MPP data storage architecture at its core, with layers of additional integrations on top to create a holistic platform. The aim is to move beyond core data warehousing concepts and integrate as much of Microsoft’s cloud offerings as possible into a unified, attractive platform.

Unsurprisingly, integration into the Azure ecosystem is particularly strong. Built-in ML through Azure ML, integration with the Power BI self-service data visualization platform, Apache Spark integration, loading and querying data from Azure Data Lake Storage (ADLS) and ingesting streaming data are all present. Other capabilities include: sharing snapshots of governed data with external customers through the Azure Data Share service; ONNX support enabling native ML scoring that is compatible with models trained in Azure ML and Apache Spark; and data loading through Azure Data Factory.

Programming language support includes T-SQL (the native SQL Server querying language) and Spark SQL, Python, Scala, C#, R and more. Data connectors to more than 80 data sources are available, and a visual environment called Synapse Analytics Studio is being developed to unify the data pipeline and development experience.

Finally, the platform supports both serverless operation through on-demand query execution as well as always-available, pre-provisioned instances. Workload isolation functionality provides consistent performance of concurrent, heterogeneous workloads.

Building from an already strong base, Microsoft is evolving its warehouse platform to offer a unified experience and consolidate some of its internal offerings. Some of the platform's newer capabilities are not fully developed at this point. That said, we anticipate interest from organizations with significant Microsoft investments as well as others wanting to modernize, unify and cloud-enable their analytic and data science workloads.

Oracle Autonomous Data Warehouse

Oracle's Autonomous Data Warehouse is the Redwood City, CA-based company's fully-managed, cloud-native data warehouse platform. The product boasts standard data warehousing features such as columnar storage, parallelism and data compression, and differentiates itself mainly on the basis of its potential for being operated mostly hands-off or with very light administration effort.

The platform fundamentally is based on the widely-used Oracle Database, tuned and optimized for cloud deployment. The database is automatically monitored, patched, upgraded, backed up, secured, repaired, tuned and scaled. The company markets these characteristics under the moniker of 'autonomy,' emphasizing that in most situations the platform will always be online and in good operational health without human intervention.

Integration in the Oracle application and cloud ecosystem is, as expected, a strong selling point. Other notable features include Oracle ML, a built-in web notebook for data querying and reporting; automatic elasticity; wide ecosystem support with general drivers for all major business intelligence tools; built-in data security; support for the property graph database model through the Oracle Graph Server; integration with external data lakes; a variety of standard Oracle capabilities for application development; and deployment in both shared and dedicated Exadata-based infrastructure in the Oracle Cloud. The company also offers an equivalent transactional workload using the same concepts and infrastructure that can be provisioned alongside the warehouse offering.

The platform will appeal strongly to organizations with existing Oracle staff, database and programming investments. The staff will appreciate the fairly straightforward way of modernizing workloads, moving them to the cloud and having them operate in a mostly hands-off manner.

SAP Data Warehouse Cloud & BW/4HANA

SAP's data warehouse offerings revolve around its HANA database and how it has evolved over the years to move to the cloud and provide first-class support for analytics. The Walldorf, Germany-based company offers two different products in this space that cover different segments of the market.

First off, the SAP Data Warehouse Cloud is the company's main data warehouse cloud platform going forward. Based on the in-memory HANA database, the platform is a managed service deployed on the SAP Cloud Platform and aimed at both business and IT users. The platform uses, but does not require, a semantic layer on top of the base data to abstract and simplify concepts for business consumption, and also boasts a high level of UX sophistication aimed at non-technical audiences. There are several well-integrated features in the platform, including isolated workspaces for different use cases; a graphical editor for wrangling and modeling data and relationships; SQL scripting for technical users; and visual storytelling capabilities around data. Integration with additional SAP cloud platforms such as the SAP Analytics Cloud enable built-in data visualizations support, and connections to SAP data sources are natively supported.

Secondly, the BW/4HANA platform is the company's main data warehousing platform for on-premises solutions, although deployment to the cloud is also supported. The platform is also based on the HANA database and has the same base semantic layer capabilities discussed above, in addition to a variety of other features. These include integration with external and SAP-based data sources and applications, data governance and lifecycle tools, pre-built partner applications, and sophisticated data modeling tools.

The SAP data warehouse story should continue to appeal to existing users of legacy SAP platforms as they look to modernize their applications. Building out the cloud capabilities and creating a well-executed path for migrating from, or co-existing with, BW/4HANA should keep the company busy in the next year and beyond.

Teradata Vantage

Teradata, the San Diego, CA-based pioneer of MPP technology, has evolved its venerable Teradata Database into the modern Vantage platform, offering advanced data and analytics on-premises and in the cloud.

The platform consists of a base relational data store, on top of which sits a variety of analytic engines that cater to different use cases. These engines today include a SQL analytic engine (i.e., the traditional Teradata Database engine), an ML engine, and a graph engine (based on its well-regarded Aster Data technology). In the future, the company expects to include additional engines, such as Spark, TensorFlow, and other custom engines.

Support for a large variety of programming languages sits on top of these analytic engines. Languages supported include: SQL, Python, and R, with Scala, Go, and JavaScript support expected in the future. In addition, a plethora of analytic tools sits at the top of the platform architecture. R Studio, Jupyter, and SAS support, along with Teradata's own SQL Studio and App Center, are built-in, and there is forthcoming support for the Knime and Dataiku data science platforms.

The platform offers significant flexibility in how it can be deployed. Options include: managed service and IaaS options in the Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) public clouds; the company's own Teradata Cloud as a service; on-premises through the VMware virtualization platform; and on-premises as a service operated by Teradata.

Other notable capabilities of the platform include: support for hundreds of pre-built analytic and graph functions; connectivity to a variety of data storage technologies, including object stores such as AWS S3 and Azure Blob Storage; IoT edge analytics and operational capabilities through Teradata's 4D Analytics product; a self-service data prep and data science platform named Vantage Analyst; and a customer data analytics tool named Vantage Customer Experience.

Vantage is a strong offering in the marketplace, backed by one of the stalwarts of the data warehouse industry. We expect that Teradata's strong enterprise presence, the platform's modular architecture, varied offerings, and deployment capabilities will continue to position it as an enterprise leader for years to come.

Vertica

Founded in 2005 and eventually acquired by Micro Focus in 2017 after a series of structural and ownership changes over the years, the Cambridge, MA-based organization offers the Vertica Analytics Platform as its main product in the data warehouse and analytics market.

The platform offers excellent concurrent loading and querying performance over massive data sets and a columnar data store with a high degree of effective compression. In addition, the platform offers a robust set of SQL-based in-memory analytics functions such as time series, pattern matching, geospatial analytics and many more.

Building on the native analytics capabilities, the platform offers tight integration with the data science workflow by offering in-database ML capabilities. Functions for data prepping, model training and testing, prediction, performance evaluation and management are all well-thought-out and presented. Algorithms such as various forms of clustering, regression, classification, decision trees and more are built-in and predictive analytics in general are a prevalent feature. The algorithms can be extended by user-defined functions written in the Python, R, Java or C++ programming languages.

Strong interoperability is available with external tools in the ETL, data visualization, streaming and data transformation ecosystems such as Kafka, Spark, Informatica, Tableau, Qlik and many more. Hadoop is also well-supported with native running of analytics on the big data platform through the Vertica SQL on Apache Hadoop product. Data in external stores, such as AWS S3, can be analyzed in place and

joined with other Vertica data through the Vertica External Data offering, while support for data catalog platforms such as Alation and Apache HCatalog is also present.

Deployment options are some of the most flexible in the industry. The platform is available for infrastructure as a service (IaaS) deployment on the AWS, Azure and GCP platforms. The company also offers different scalability models with cloud deployments, including Enterprise Mode for coupled compute and storage resources, and Eon mode for independently scaled and sized storage and compute resources. Deployment is also available on-premises, where Eon mode is also supported through a partnership with Pure Storage and its FlashBlade file and object storage platform.

Vertica's platform offers strong performance; a wide variety of built-in analytics functions; excellent integration with data science workloads and external data lakes; and notable flexibility in its deployment options. These characteristics have helped it gain a foothold in a variety of large enterprises and should continue to serve its clients well as the market continues to evolve. We also anticipate additional growth as migration from on premises hardware platforms such as Netezza continues.

Newer Vendors

Amazon Web Services Amazon Redshift

Amazon Redshift, the Seattle, WA-based cloud giant's data warehouse service, is probably the most widely used cloud platform in this market segment today. Released in 2013 and using the PostgreSQL relational database system as its base, the platform has evolved over the years to define and extend how analytics at scale are performed in the cloud.

Amazon Redshift is a fully managed, cloud-native data warehouse service that offers a variety of capabilities built around its base columnar, compressed storage model. Fast performance on querying petabytes of data is built-in, with automatic caching for repeated query patterns also present. In addition, a variety of predictive and observational techniques are employed behind the scenes to aid with query acceleration. The platform also offers a variety of storage options of varying speeds, some of which allow for compute and storage to be scaled independently, while others support a combined storage/compute model. Managed storage, flash and spinning disk storage types are all supported in various combinations.

Scaling is, as expected, a strong feature of the platform. Amazon Redshift offers the usual (and one would argue Amazon-invented) concepts of cloud elasticity and on-demand scalability. In addition to that, serverless-like features such as automatic concurrency scaling are available for maintaining steady query performance for a virtually unlimited number of concurrent users.

The platform is well-integrated with AWS data lake offerings through Amazon Redshift Spectrum, which allows for exabyte-scale querying of structured and unstructured data stored in Amazon S3 object storage. Without the need for data movement from the data lake to the data warehouse, performance is excellent, especially when a compressed columnar format like ORC or Parquet is used to store files

in the data lake. The forthcoming implementation of federated querying across AWS's operational databases such as Amazon Aurora is another feature that will extend the reach of the platform's in-place querying capabilities.

As expected, there is also excellent integration with the rest of the AWS analytic and data science ecosystem, as well as support from pretty much all business intelligence tools and programming languages in the market.

Amazon Redshift is a powerhouse in the cloud data warehouse market and its current and planned offerings should help it to maintain its leadership position, even as other vendors continue to evolve their offerings.

Cloudera Data Warehouse

The Cloudera Data Warehouse is the Palo Alto, CA-headquartered company's hybrid, multi-cloud data warehouse platform. It was unveiled in the fall of 2019, builds on open source technologies, and is part of the wide-ranging Cloudera Data Platform (CDP) set of solutions.

The warehousing platform at its core enables a SQL-based, auto-provisioned and auto-scaling data warehouse infrastructure that supports elasticity and workload concurrency. The base of the platform is a modernized combination of the Hortonworks and Cloudera legacy Hadoop platforms. Several architectural changes have been made to combine and optimize the platform for better performance. The platform is based on core open source engines and technologies such as Apache Hive's LLAP (Long Lived Analytical Processing), Impala and Kudu. Cloudera Data Warehouse integrates with other CDP technologies, like HBase and Spark as well.

Other features of the platform include: in-place query execution over the variety of its built-in data stores; data governance, control and security tools through the Cloudera SDX product; integrated ML capabilities through the Cloudera Machine Learning service; virtual warehouse support for concurrently executing different workloads on the same data; real-time processing and predictive analytics on streaming data through Cloudera Data Flow; and self-service data discovery capabilities.

The platform supports on-premises, hybrid and fully cloud-native deployments in AWS and Azure, with GCP support being planned. Kubernetes and Docker containers are at the core of both on-premises and cloud deployments.

The Cloudera platform has taken a great first step in modernizing its legacy Hadoop-based infrastructure, combining it with some of the best features from its merger with Hortonworks, providing flexible warehousing capabilities and supporting a variety of on-premises and cloud deployment options. The platform should appeal immediately to enterprises already invested in Cloudera and Hortonworks legacy offerings while also competing for new customers interested in a hybrid and multi-cloud data warehousing strategy.

Google BigQuery

Google's BigQuery is the Mountain View, CA-based company's fully-managed, serverless cloud-native data warehouse platform. This serverless platform naturally excels at ease-of-use, set up, management and administration. The platform also offers significant capabilities (analyzed below), and delivers a low total cost of ownership and excellent integration with Google's own cloud platform.

BigQuery offers excellent support for the base data warehouse, storage and scaling capabilities we have come to expect in a modern cloud-native product. An additional offering of the platform is its built-in ML capabilities, a set of features that the company calls BigQuery ML. These allow the platform's users to create, execute and manage ML models for forecasting, regression, classification, segmentation and other use cases. The platform innovates in this space by allowing the models to be accessed and trained through SQL rather than the more typical Python, R, Java and other languages that data scientists usually employ. The platform also allows for the model data to be accessed in place at the data warehouse, rather than needing to be exported. Finally, the ML capabilities are easily accessible through the standard BigQuery Web interface, allowing business users and other non-traditional data scientists easier access to these features.

Another set of significant capabilities, dubbed BigQuery BI Engine by the company, allows for in-memory interactive analytics of massive data sets with near-immediate response time. The capability integrates with Google's Data Studio and Looker to allow for visual analytics and dashboarding on data local to BigQuery, without the need for data pipeline solutions. The cloud platform's built-in data streaming capabilities are also neatly integrated throughout.

BigQuery also offers a Data Transfer Service that allows for migration of data from on-premises sources such as Teradata, cloud warehouses such as Amazon Redshift and other Google-related ecosystem products such as Ads and YouTube. More than 140 connectors to a variety of data warehouses, databases and SaaS applications round out this offering.

Other capabilities in the platform include: BigQuery GIS, a built-in geospatial analysis engine; a high-speed streaming insertion API; use of standard SQL queries; a storage API to aid with data lake convergence; connectivity to several other data sources, platforms and execution engines within the overall GCP ecosystem; built-in data governance capabilities; a programmatic REST API that allows connectivity from a variety of programming languages such as Python, Java, C# and many more; and a set of curated public and commercial datasets already loaded into the platform.

Overall, we find the Google BigQuery platform to be a very strong product, due to its serverless nature, specialized ML and other capabilities, its overall polish and strong integration with the rest of GCP. We expect Google to continue to invest in it and evolve its offerings to compete strongly in the marketplace.

MemSQL

MemSQL, a newer entrant in the data warehouse marketplace, released the first version of its platform

in 2013 and has experienced rapid growth since. The San Francisco, CA-based company's platform combines support for transactional and analytic workloads in the same engine, offers an interesting new view on storage architecture and is currently targeting the operational analytics and AI/ML market for future growth.

The platform supports both transactional and analytic workloads through the blended use of in-memory, row-based and on-disk columnar data structures. The company calls this design a SingleStore and considers it the key for the platform's current capabilities and future growth. Currently, this is a work-in-progress and the company has plans to blend the features from both row and column-based approaches into an automatic, transparent process in future versions. For now, the users must designate what kind of storage strategy they want to use for individual tables; however, the company has already implemented significant performance optimizations such as sparse row compression, columnar seek, and others to clearly set out its vision. The end result should be that the platform will allow both transactional and data warehouse workloads to cross-pollinate and be supported seamlessly.

In the current version, the platform focuses on performance, especially in cases where SingleStore optimizations have already been implemented. Other strong features of the platform include: query compilation for additional performance enhancement; strong support for AI and ML through built-in SQL-based functions; real-time streaming ingestion capabilities through support for programmatically creating data pipelines and connections to Apache Kafka and Spark; interoperability with BI and data science tools; distributed query execution; and support for geospatial, time-series and other data types. The platform is also compatible with the MySQL wire protocol for connecting from external applications and programming languages.

The platform is available for deployment on-premises directly on virtual machines, bare metal or in containers through Kubernetes. Cloud deployment is also available in every public cloud, again through Kubernetes. A managed service deployment named MemSQL Helios is also available for AWS and GCP, with Azure support being planned.

MemSQL should appeal to many organizations looking to modernize and perhaps combine their operational and analytic workloads. We expect significant future growth as its vision for blending these two architectures evolves. We also believe its multi-cloud approach through Kubernetes will appeal to several organizations who might balk at being locked in to any particular public cloud vendor.

Snowflake

Snowflake's eponymous platform, one of the newer cloud-native entrants in the market, has seen significant adoption since the San Mateo, CA-based company originally released it in 2014. The platform features a multi-cluster, shared-data architecture that allows simultaneous, consistent operations on data in the warehouse by different processes and workloads. It also features a columnar storage strategy that splits the data into multiple partitions on disk and compresses it for better performance, and to allow for better concurrency.

The platform includes support for virtual warehouses that accommodate different user needs and populations, allowing independent concurrent operations on the same data. Other major characteristics include: built-in, as-of data views for up to 90 days in the past; virtual zero-copy data cloning; excellent scalability and ease of use; autoscaling operations to manage concurrency and performance automatically; support for structured and semi-structured data types with schema-on-read and general data lake capabilities; and built-in encryption.

The company also offers data sharing of governed data with internal and external audiences through its Data Sharing, Public Data Exchange and Private Data Exchange capabilities. This allows for easy, SQL-based combination of data at customers' Snowflake accounts with live, governed local slices of approved data that the organization has designated as shareable. This is a differentiating feature for the company, which goes as far as providing a rebate to sharing organizations to promote its use and incentivize the sharing of more data.

Interoperability with the major business intelligence and ML platforms is excellent. Data can also be loaded through the company's Snowpipe offering, which allows for continuous ingestion of data from cloud object storage to the platform without the need for traditional ETL tools.

The platform is offered for public cloud deployment in the AWS, Azure and GCP clouds. A dedicated and isolated virtual private offering is also available on AWS for customers with enhanced security and privacy needs.

Snowflake's offering is compelling for its fresh architectural look at a modern cloud-based data warehouse platform. Its combination of scalability, performance, ease of use, and operational simplicity should continue to win it customers for the foreseeable future.

Yellowbrick

Yellowbrick is the newest entrant in the marketplace that we are examining in this report. Formed in 2014 and releasing its platform at the end of 2017, the Palo Alto, CA-based company offers a data warehouse platform geared toward hybrid cloud environments. The platform positions itself as an alternative to more established offerings by promising increased performance, ease of maintenance and cost advantages over more established competitors, in addition to its core hybrid cloud support value proposition.

The company's main offering is the Data Warehouse Appliance, a hardware, flash disk-based system that is installed in customers' data centers and runs the company's core data warehouse software platform. The software is based on PostgreSQL as its core, with additional performance and scalability enhancements developed by the company layered on top. Performance and data lake augmentation are two of the pillars that the company is banking on to drive adoption, with four different hardware sizing options available to handle scale. The system supports SQL for querying; major BI tools such as Tableau, SAP BusinessObjects and Informatica; data import from major database systems such as Teradata, Redshift, SQL Server, Oracle and others; and ODBC, JDBC, ADO.NET, Kafka, Hadoop and Spark connectors. The company also positions the appliance as a replacement for Netezza, supports

data import directly from IBM's appliance, and has seen significant success in the market given the sunsetting of that platform.

The company also offers the Yellowbrick Cloud Data Warehouse platform, which brings the core software offering to the company's private cloud. The company stresses the hybrid and secure capabilities of the platform, with built-in support for multi-cloud, including private connections to customers' data stores in the AWS, Azure and GCP clouds.

Finally, the platform can be augmented through the Yellowbrick Cloud Disaster Recovery managed service, providing database replication from on-premises installations to the company's private cloud for disaster recovery. Integration with the Protegrity data security platform is also on the roadmap.

The platform should primarily appeal to customers looking to migrate away from on-site Netezza appliances, as well as those interested in Yellowbrick's well-thought-out hybrid, multi-cloud strategy. We expect the company to mature its offering and expand its client list in the next year as more organizations become aware of the platform and its benefits.

6. Conclusion

In this report, we've examined the leading platforms in the data warehouse marketplace, described the fundamentals of the technology, identified key criteria and evaluation metrics by which organizations can evaluate competing platforms, described some potential technology developments to look out for in the future, and classified platforms across those criteria and metrics.

The data warehouse platform space can be characterized as both a mature market and one that is still fostering exciting new developments. The former is due to the relative maturity of the underlying theoretical constructs and data architectures being employed by today's platforms, such as columnar, compressed storage and massively parallel processing. The latter is due to a variety of enhancements to existing products or entirely new products that have appeared in the last few years, forcing intense competition as vendors scramble to fill holes in their technology portfolios and create new capabilities.

Some of the new enhancements that are particularly exciting include the proliferation of cloud-native, serverless or near-serverless multi-cloud platforms from newer vendors. As the first wave of cloud-native platforms have shown, there are better ways to build and manage a data warehouse for the modern era. We expect the traditional vendors to enhance their platforms and learn rapidly from their younger competitors. Some of the areas of focus we expect to see additional work on include new theoretical constructs such as unified storage models; new operational capabilities such as autoscaling and serverless operations; increased integration with ML and data science workflows; additional emphasis on ease-of-use and platform integration; and a focus on elasticity in both storage and computation, as well as operational cost management.

We believe there is room in the market for several competitors to thrive. Organizations must perform deep analyses of their needs and match them to the offerings in the marketplace. One way this can be done is by using this report as a guide and aid to helping them choose and implement an appropriate data warehouse platform. We recommend studying the current state of the platforms in the marketplace as we have detailed in this report, and studying the criteria and metrics that are most relevant to them, methodically evaluating the relevant platforms that seem more fitting to their requirements. We expect additional enhancements and new capabilities to emerge as the market continues its evolution, and organizations should use both this report and their own research to stay on top of what is proving to be an incredibly exciting and fast-moving market.

7. About Andrew Brust

Andrew has held developer, CTO, analyst, research director and market strategist positions at organizations ranging from the City of New York and Cap Gemini to Gigaom and Datameer. He has worked with small, medium and Fortune 1000 clients in numerous industries and with software companies ranging from small ISVs to large clients like Microsoft. Andrew's resulting understanding of technology, and the way customers use it, makes his market and product analyses relevant, credible and empathetic.

Andrew has tracked the Big Data and Analytics industry since its inception, as Gigaom's Research Director and ZDNet's lead blogger for Big Data and Analytics. Andrew co-chairs Visual Studio Live!, one of the nation's longest running developer conferences. As a longtime technical author and speaker in the database field, Andrew understands today's market in the context of its longtime Enterprise underpinnings.

8. About Yiannis Antoniou

Yiannis Antoniou is a technologist with over 20 years of global experience in the financial industry. He has served as a CTO in the asset management industry, as a management consultant in the banking & insurance industries and as a technical architect, project manager, development and infrastructure manager in major financial firms in the US and Europe. Major organizations in his tenure include Goldman Sachs, JPMorgan, AIG, Pacific Global Advisors, EY and BNY Mellon, holding various technical management positions in the Asset Management, Investment Management, Enterprise Risk Management, Strategic Planning, Investment Banking, Insurance, Innovation and Digital Transformation areas.

Yiannis delivers technology expertise in data & analytics, cloud, application development, technology architecture, technology operations, technology infrastructure, AI and ML, Blockchain, and DevOps in enterprise and startup settings. He combines ‘go to market’ expertise with practical application of agile product, project, program and portfolio management processes and has managed and implemented more than 200 programs, projects and engagements with cumulative budgets of over $500 million. Yiannis is a graduate of the National Technical University of Athens, Greece where he worked in a variety of European Union research projects in the fields of energy and financial modeling, built applications, taught database systems and design and published research papers in peer-reviewed journals.

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