October 22, 2021

MARKET OVERVIEW

What to Look for in Analytical Data Platforms for a Cloud-Centric World

A Close Look at 14 Cloud-Ready Data Warehouse, Lake Query Engine, and Combined Lake/Warehouse Platforms

Doug Henschen
Principal Analyst and Founder

TABLE OF CONTENTS

EXECUTIVE SUMMARY

High-scale analytical data platforms are used by organizations to drive better decisions and actions and to provide differentiated products, services, and customer experiences. They provide historical and low-latency data for business intelligence (BI) and analytical analysis, supporting production-scale reporting and dashboarding; ad hoc query and analysis; and, in some cases, a foundation for data science, including machine learning (ML) and artificial intelligence (AI).

The high-scale analytical data platforms market historically has been dominated by database management systems (DBMSs) optimized for analytics and deployed on-premises as the backbone of data warehouses. Today customers are turning to cloud-based database services and emerging options such as data lake query engines and combined lake/warehouse platforms.

This market overview evaluates 14 high-scale analytical data platforms, including DBMSs, lake query engines, and combined lake/warehouse platforms. Most of the featured offerings are delivered as cloud services, but the report also details the availability of cloud marketplace offerings as well as software and specialized appliances that can be deployed on-premises. The report identifies key market trends and buying criteria and evaluates the key differentiators, functional capabilities, and strengths and weaknesses of each offering. Technology buyers should use this report to evaluate analytical data platforms for implementation.

Business Themes

 Data-to-Decisions

 Technology Optimization

MARKET DESCRIPTION

Market Definition

The high-scale analytical data platforms market has flourished and drastically evolved over the last two decades. Whereas the market once centered on a dozen DBMSs that were deployed on-premises, today most of the attention has turned to DBMSs, lake query engines, and blended lake/warehouse platforms offered as services on public clouds. Still available and still very relevant are choices including DBMS software that can be deployed on-premises, combined hardware/software (aka appliance) analytical platforms deployed on-premises, and DBMS and lake query engine marketplace offerings that can be deployed by customers on public clouds.

Data lake platforms represent another type of high-scale analytical platform, and today they're increasingly built on cloud-based object stores. Apache Spark and Hadoop are still prevalent, and these platforms are also deployed by customers, either on-premises or on public clouds, or are consumed as cloud-based services run on public clouds.

The new breed of combined lake/warehouse offerings supports data engineering, data science, and data warehouse/BI workloads against a shared storage environment. Their vendors invariably tout the simplicity of having a single security and access control scheme and unified governance of data on one platform.

DBMSs remain the backbone of the vast majority of data warehouses that support BI/analytical workloads. The newcomer to the market is SQL query engines designed to work with data stored in data lakes and/or distributed data fabrics. Their vendors tout the advantages of querying data where it already lives: either in a lake (most frequently) or in distributed stores accessed via a virtualized access/federation approach.

Whether it's an analytical DBMS or a query engine designed to work with lakes, customers will expect it to support the querying required for BI, including scheduled reporting and ongoing refresh of executive and operational dashboards that might have tight service-level agreements (SLAs). Thus, these DBMSs and query engines must offer query tuning, data tiering, and data caching capabilities to support

performant querying against high-scale data as well as lots of concurrent users and queries. DBMS and query engines might also be taxed with unpredictable ad hoc query-and-analysis workloads, adding yet more workload management challenges on top of the reporting and dashboarding SLAs.

Data lake platforms (and the lake side of combined lake/warehouse offerings) enable organizations to go beyond the structured and semistructured data typically associated with data marts, data warehouses, and SQL-centric querying. Lakes can ingest any data and provide a platform for data transformation and data science analysis at scale. Lakes routinely handle internet clickstreams, sensor data, log files, mobile data rich with geospatial information, and text extracted from customer relationship management (CRM) call records and social-network interactions.

The data lake's combination of data type flexibility and lower cost of storage (compared with DBMSs) has become a foundation for innovative and value-driving analyses. What's more, data lakes serve as a platform for data engineering at scale. Lake-based data processing is often used to feed structured data into warehouse platforms. Lakes also support predictive data science and ML workloads that would be difficult; costly; and, in some cases, technically impossible to support on SQL-centric platforms.

This market overview report surveys 14 prominent vendors of analytical database management systems, SQL query engines, combined lake/warehouse platforms, and associated data lake platforms. It also details the deployment options provided by these 14 vendors, including on-premises (including private cloud), customer-managed deployments of cloud marketplace offerings, and vendor-managed cloud services. Data warehouse and SQL-centric BI needs are at the forefront of the research and criteria for inclusion. The research does not include data lake vendors that do not also offer analytical DBMSs, SQL query engines, or combined lake/warehouse offerings.

Market Trends

The analytic platforms market was sleepy and consolidated before it exploded in the 2000s, as organizations increasingly grappled with rising quantities of data, a desire to develop novel insights from unused data, and expectations for ever-faster analysis. The leading general-purpose relational database management systems (RDBMSs) at the time—Oracle Database, Microsoft SQL Server, IBM Db2, and

MySQL—were being used for data warehousing, but they had yet to be highly adapted for analytical use. Pioneers of the high-scale analytical data platforms market harnessed massively parallel processing (MPP)—employed by Teradata beginning in the 1980s—and column-store architectures—harnessed by Sybase IQ (now SAP IQ) in the early 1990s.

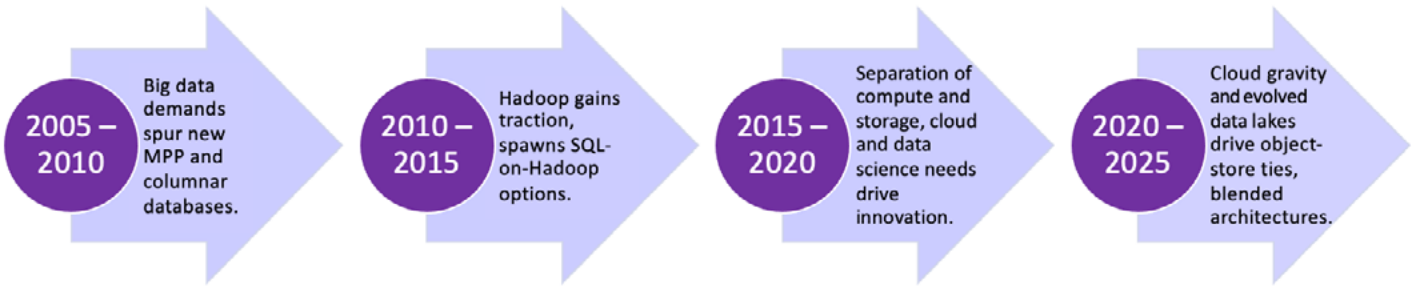
As shown in Figure 1, as the big data era emerged in the mid-2000s, a raft of new vendors and platforms burst onto the scene, with MPP, columnar architectures, and purpose-built analytic appliances coming to the fore. MPP (also known as scale-out architecture) soon became the cornerstone of many platforms, including yet more DBMS options, Apache Hadoop, Apache Spark, and many NoSQL stores.

Starting in 2010 and gathering steam by 2012, organizations were captivated by the promise of low-cost storage in Hadoop. Deployments multiplied, but the complexity of this new platform limited access to data. Analytical DBMS technologies were soon adapted (and multiplied yet again) to support SQL-on-Hadoop analysis, filling the accessibility void by bringing the familiarity of relational querying to BI-curated datasets within data lakes.

Innovations continued through the second half of the last decade, with yet more features introduced for in-database data science as well as more sophisticated (hot/warm/cold) data tiering and caching schemes aimed at optimal query performance.

The stampede to the cloud has been the most powerful market driver over the last decade, with organizations increasingly moving workloads—and, therefore, data to be analyzed—into public clouds.

Figure 1. The Evolution of High-Scale Analytical Data Platforms: 2005–2025



Source: Constellation Research

Tech flexibility and business and innovation agility have been the key draws to the cloud, although scalability and elasticity are also important when handling big data and spiky analytical workloads.

Cloud advantages and requirements have fueled yet more customer interest in new features, including automated "serverless" scaling, automated systems-management capabilities, and separation of compute and storage decisions. The move to the cloud (coupled with the complexity of managing Hadoop) also led to a new generation of data lakes built around low-cost cloud object storage.

With the new generation of object-store-based data lakes, we've seen two additional trends. First, query engine platforms have emerged that are geared to supporting SQL-centric BI workloads directly against data in data lakes. Second, combined lake/warehouse offerings have become available, supporting data engineering, data science, and data warehousing on a single, shared data platform.

As we move toward 2025, Constellation expects to see more extensive use of object storage, including as the foundational storage layer for databases as well as query engines and data lakes. Data fabrics are also gaining ground, with several vendors working on extended capabilities for accessing data where it lives and selectively moving compute to the data or data to the compute, as required, to meet performance demands.

IMPORTANCE TO BUYERS

Buyer Challenges

The most important challenge for any organization is harnessing data to drive better decisions and actions and to provide differentiated products, services, and customer experiences. The size, industry, and ambition of each organization influences what types of data it harnesses and the level of sophistication of its analyses.

- Insurers routinely gather and analyze driving behavior data from customer vehicles on a massive scale to drive dynamic policy-pricing decisions.

- Oil, gas, and mining companies routinely harness data from connected edge sensors to drive near-real-time decisions on drilling and mining operations.

- Manufacturers analyze shop floor sensor data to maintain product quality. They also analyze supply chain and logistics data to ensure plant productivity and responsive distribution of products in accordance with market demand.

- Telecommunications companies analyze data at scale to maintain and improve the reliability of their networks, monitor customer satisfaction, and trigger proactive actions to avoid customer churn.

- Online and brick-and-mortar retailers analyze customer behavior data at scale to deliver targeted cross-sell and upsell offers and personalized services.

- Healthcare organizations analyze admissions trends, the efficacy of treatments, and internal policies and procedures to ensure better patient outcomes.

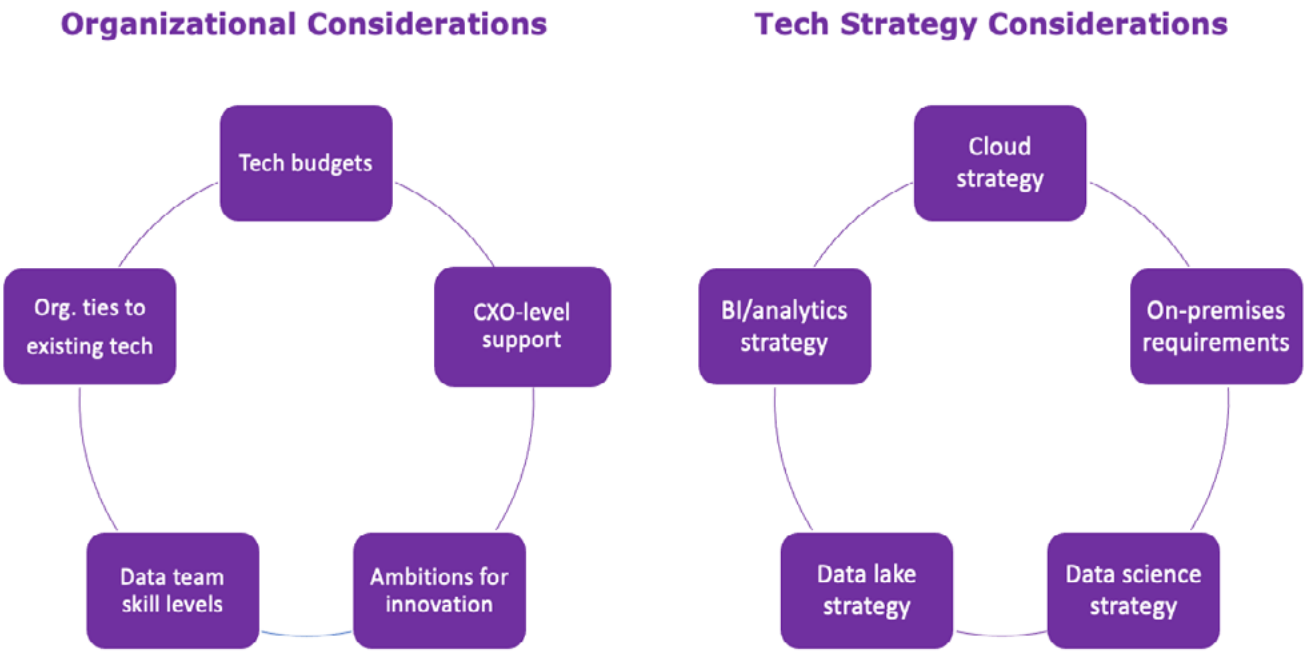
- Media and advertising companies analyze audience demographics and behaviors to guide programming and advise customers on where to spend their ad dollars.

Data-driven decision-making is a given in these and many other industries, so organizations have no choice but to keep up with (or try to surpass) their competitors when it comes to harnessing data.

Selection Criteria

The search for new technology selections should not start with the tech. Seek first to understand the organizational and technology strategy considerations. As shown in Figure 2, organizational considerations include existing tech budgets, CXO support for data-driven decision-making, the ambition (of CXOs and data teams) to step up innovation (and tech budgets), the existing skills and experience of data/analytics/data science teams, and organizational dependencies on existing tech investments.

Figure 2. The Selection Process Should Start With Understanding High-Level Organizational and Tech Strategy Considerations



Source: Constellation Research

Organizational Considerations

Where is the organization coming from? It starts with the understanding of budgets, available skills, and incumbent tech (what is and isn't changing). The existence and state of existing tech (in terms of age, effectiveness, and perceived value) have a lot to do with existing budgets, skills, organizational dependencies, and the ability to innovate. Significant expenditures for new products and skills will depend on executive ambitions and willingness to fund innovation.

Tech Strategy Considerations

Where is the organization going? Forward-looking technology strategy considerations include the following:

- **Cloud strategy.** What is the progress toward and commitment to moving into the cloud, and which cloud or clouds are part of that strategy? Are multiple clouds used to reduce business continuity risks or to meet data sovereignty requirements? What are the standards or preferences in terms of

self-managed versus vendor-managed services, use of virtualization or container technologies, and preferences for storage of data within vendor or customer cloud accounts? All of the above will shape the depth, breadth, and style of cloud-deployment options prioritized in a technology selection.

- **On-premises requirements.** Are certain applications and/or data types destined to remain on-premises for internal policy or external regulatory reasons? Are such requirements likely to be permanent, or might they increase (as with emerging national data residency requirements, for example)?

- **Data lake strategy.** Does the organization have a data lake or lakes? Are they on-premises and/or in clouds? Have new technology standards and migration paths been set, and what are the scale and diversity of data currently stored and expected to be added to the lake? Lake strategy will influence the selection of query engines designed to work with lakes or combined lake/warehouse offerings. It will also influence the extent and type of data science workloads that organizations pursue with data warehouse/mart environments.

  Constellation Research views modern lake architectures built on object storage as the most scalable, cost-effective, and flexible foundations for data lakes, supporting both a diversity of data (including unstructured data) and a vast array of possible data engineering and data science approaches.

  As for those emerging combined lake/warehouse offerings, some vendors have expanded into warehousing from their roots as data lake vendors. In other cases, database vendors have added lake capabilities for ingesting and cost-effectively storing variable data types in object storage without necessarily moving them into their database service. In general, the vendors with lake backgrounds tend to have broader support for data engineering and data science, whereas the database vendors with combined lake/warehouse architectures tend to focus mostly on drawing on data at scale to be structured and refined for SQL analysis use cases.

- **Data science strategy.** What's the existing and hoped-for level of data science sophistication, scale of analysis, and progress toward operationalization? Existing and hoped-for data engineering and data science activity will obviously shape data lake investments and ambitions.

Depending on which workloads are supported by data scientists on data lakes, the next question is what type of data science might be supported in-database within marts and warehouses. Many DBMSs support data science extensions of SQL, the use of data science languages such as Python and R, and in-database execution for workloads such as scoring/inferencing.

- **BI/analytics strategy.** BI and analytical capabilities are crucial, but they're often well established and sometimes overlooked and stuck in past practices. New investments in cloud adoption, data science, and data lakes should be coupled with a reexamination of the value and use of BI and analytics and the need for consolidation and new analyses. Indeed, data lake investments are often coupled with data warehouse optimization initiatives and BI/analytics upgrades. What's more, with augmented analytics features such as automated ML (AutoML) emerging, we're seeing a blurring of lines between BI/analytics and data science.

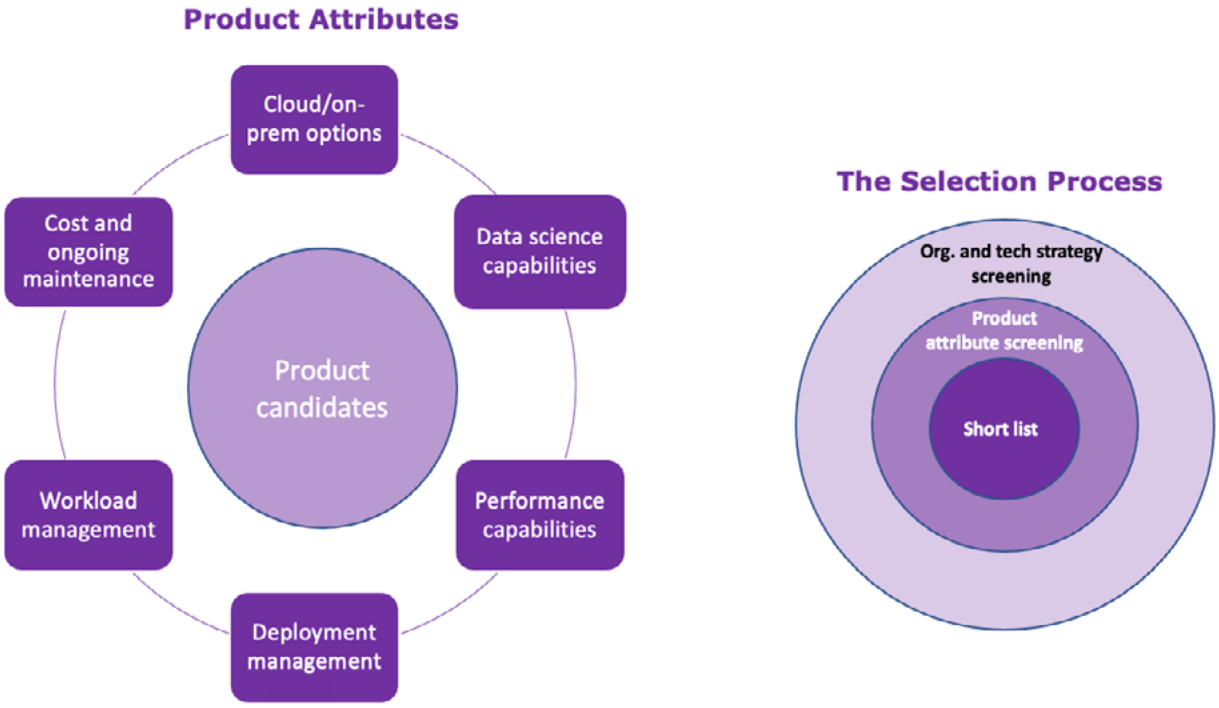
Once the context of the organization and its tech strategies are well understood, the team should be able to narrow things down to specific technology categories (meaning DBMSs versus data lake query platforms versus combined lake/warehouse offerings). And when it comes to the tech, an understanding of cloud strategy and on-premises needs will narrow down the selection among hybrid-, single-cloud-, and multicloud-capable products. Similarly, an understanding of data lake, data science, and BI/analytics strategy will inform the choice of DBMS, lake query engine, or combined lake/warehouse offerings.

Product Attributes

Eventually—and it may take weeks or months to narrow down from categories to candidates—you'll get down to considering the attributes of specific products. As shown in Figure 3, key product attributes to consider include:

- **Cloud and on-premises deployment options.** Can the product be deployed and/or is it available as a service in the cloud or clouds of your choice? Does it support cross-region or cross-cloud provider deployment to support business continuity and data sovereignty needs? Can the data reside within the customer's virtual private cloud account? Is there an on-premises option, and is it compatible/consistent with the cloud deployment option(s)? If there is no on-premises deployment option, what are the provisions for (and costs of) migrating data from/connecting to on-premises sources?

Figure 3. Organizational and Tech Strategy Considerations Determine the Candidates; Product Attributes Guide the Final Short List



Source: Constellation Research

- **Data science capabilities.** As noted earlier, this market overview is heavily focused on supporting BI/analytics workloads, whether supported by DBMSs, query-on-data-lake platforms, or combined lake/warehouse platforms. Although support for standard SQL is commoditized, support for data science on DBMSs, query engines, and the warehouse side of combined lake/warehouse platforms varies considerably. Does the vendor leave data science to the data lake and data science team, or does it support data science on its DBMS or query engine? Do these capabilities target exclusively data scientists, or can they be exploited by SQL-savvy analysts and power users? What types of data science are supported, and what stage of work (modeling versus scoring/inferencing) do you expect or want to do on which platform (meaning lake or warehouse)? What's the support for third-party data science platforms and ecosystems?

- **Performance capabilities.** There are many dimensions of performance, but where DBMSs, query engines, and the warehouse side of combined lake/warehouse offerings are concerned, the focus is on query performance. Performance will depend on the number, frequency, and sophistication of your queries and the number of concurrent users and their performance requirements and expectations.

Is the workload primarily predictable queries driving reports and dashboards at scale? Do you have tight service-level requirements? Will unpredictable ad hoc queries come into the picture? Can the platform easily isolate and sustain competing workloads? Where can or must the data reside in order to sustain a given level of performance? What's the ballpark node count and/or caching capacity (and associated cost) that might be required now and into the foreseeable future?

- **Deployment management.** Despite the marketing hyperbole, enterprise software is rarely, if ever, easy to configure and deploy. But just how onerous is the deployment experience? Is it software that must be deployed by the customer, either on-premises or in the cloud? Are marketplace offerings available to ease cloud deployment? Or is it a cloud service or multiple cloud services? Do you have to estimate capacity requirements up front, or is it a serverless offering that will automatically match your current scale and then scale up as data stores grow?

  Platform as a service (PaaS) versus software as a service (SaaS) is another dimension to consider. Is it a single platform running on multiple clouds, or multiple SaaS services with differences (and diminished portability) from cloud to cloud? Does your data live in your cloud account or in the vendor's SaaS service? If the offering depends on third-party storage or a third-party platform (such as an object-store-based data lake), what's entailed in integrating with that environment? Are container-based deployment options available that support consistent deployment and monitoring approaches across hybrid and multicloud footprints?

- **Workload management.** AI and automation technologies are in their infancy, so there's no such thing as a clairvoyant product that understands your workloads, workload priorities, and SLAs with zero guidance from humans. Some highly automated products offer simplified schemes for setting priority levels and assigning resources and letting the product make all sorts of query tuning, data tiering, and caching decisions behind the scenes. If and when performance falls short, the question is, do you have visibility into how choices are made, and do you have any performance-tweak options other than throwing more (cost-driving) cache and/or compute capacity at the problem?

  Many less-automated products give you any number of query performance tuning and tweaking options, but these might present challenges in terms of selections to be made and adjusted and rules

to be written and revised. In the (typical) case where new and competing workloads are commonplace, the skills required for workload management and the cost of these resources should not be overlooked or thought of as sunk costs.

- **Cost and ongoing systems management.** There's the capacity you think you will need, and then there's what you will actually use. Constellation Research has spoken to organizations that have tapped highly automated, cloud-based platforms that ended up needing more capacity than they anticipated. Keep in mind that some automation features consume compute cycles to monitor and optimize performance, so you may need more compute capacity than expected.

  Constellation has also encountered customers of nonautomated platforms that have to be manually sized and procured in advance to get one-year or three-year reserved capacity discounts that found that they were overprovisioned (spending more than they needed to). What's more, buyers of nonautomated systems also complain about people costs and the difficulty of finding and hiring skilled staff.

Experience is the best teacher, so if you're new to a product, a cloud service, or a cloud version of a product you've previously used on-premises, talk to existing customers about their performance and capacity-planning experiences. If possible, get references that have similar scale, analytical diversity, sophistication, concurrency demands, and service-level demands. Ask about the balance of automated capabilities versus any desire for greater control. Ask what the product enabled them to do that they could not do before.

If it's not an automated product, ask about the difficulty of management and tuning and the skill level required. Ask about their cost experience and any surprises that put a dent in their budget. Also look for burst-capacity options that enable you to meet peak workloads without resizing your system, stopping and restarting your system, and/or paying on-demand rates.

Once you've winnowed the selection to a short list of finalists, it's time to come up with proof-of-concept projects, giving all constituents of the would-be system hands-on experience with (or at least some degree of exposure to) the finalists.

Vendors and Offerings

Constellation has included 14 vendors in this market overview. Figure 4 lists the featured analytical data platforms and related products and services available from each vendor. The table includes on-premises deployment options, cloud marketplace options (and, in parentheses, on which clouds they are offered), database and data lake query engine services (and on which clouds they are offered), combined lake/warehouse offerings (and on which clouds they are offered), and cloud data lake offerings built on object storage or Hadoop (and on which clouds they are offered).

"The table includes on-premises deployment options, cloud marketplace options, database and data lake query engine services, combined lake/warehouse offerings, and cloud data lake offerings built on object storage or Hadoop."

As noted earlier, several vendors offer combined lake/warehouse offerings. Some of these vendors, such as Cloudera and Databricks, have added data warehouse capabilities to their existing data lake platforms. In contrast, combined lake/warehouse offerings with database roots, such as Amazon Redshift Spectrum, SAP HANA Cloud Data Lake, and Snowflake Data Cloud, added highly scalable, low-cost object storage as their first level of storage. In October 2021, Oracle added an object-store-based "Lakehouse" option alongside its Autonomous Data Warehouse service. Although many database vendors can query object stores, these combined lake/warehouse offerings govern the lake space, providing metadata management, governed access control over the lake to support reuse, variable transformation, and sharing of data for diverse use cases. See the analysis of each combined lake/warehouse offering for details on its warehouse capabilities, semistructured and unstructured data support, and data engineering and data science functionality.

Following Figure 4, are Constellation's vendor-by-vendor analyses, presented alphabetically, of offerings and capabilities, including our take on strengths and weaknesses. Note that the in-depth analyses are focused on the featured product/service, except where noted.

Figure 4. Constellation Research Analytic Platform Offerings and Deployment Options (Q3 2021)					
VENDOR	**On-Premises Analytic Platform Options**	**Cloud-Deployable Software and Marketplace Offerings (Cloud Options)**	**Analytic DB/Lake Query Engine Services (Cloud Options)**	**Combined Lake/ Warehouse Services (Cloud Options)**	**Cloud Data Lakes on Object Storage or Hadoop/Spark (Cloud Options)**
Amazon Web Services (P. 17)			*Amazon Redshift (AWS)	Amazon Redshift Spectrum/Lake Formation (AWS)	Amazon EMR (AWS)
Cloudera (P. 20)	CDP Private Cloud Edition	Cloudera Enterprise BYOL (AWS, Azure, GC, IBM)		*CDP Public Cloud (AWS, Azure, GC, IBM)	CDP Public Cloud (AWS, Azure, GC, IBM)
Databricks (P. 22)		Databricks Lakehouse (AWS – provisions SaaS in customer's cloud account)		*Databricks Lakehouse Platform (Alibaba, AWS, Azure, GC)	
Dremio (P. 24)	Dremio Software	Dremio Software (AWS, Azure)	*Dremio Cloud (AWS)		
Google Cloud (P. 26)			*BigQuery (GC), BigQuery Omni (AWS, Azure)		Dataproc and Dataproc Metastore (GC)
IBM (P. 29)	IBM Db2 Warehouse, Netezza Performance Server	Db2 Warehouse, Netezza, and BigSQL all on Cloud Pac for Data (deploy anywhere with OpenShift)	*IBM Db2 Warehouse on Cloud (IBM, AWS), Netezza as a Service (IBM, Azure), IBM SQL Query (IBM)		IBM Analytics Engine (IBM)
Incorta (P. 31)	Incorta Direct Data Platform	Incorta Direct Data Platform (AWS, Azure)		*Incorta cloud service (GC)	
Microsoft (P. 33)	SQL Server 2019	SQL Server on VMs (Azure)		*Azure Synapse Analytics (Azure), Azure Databricks (Azure)	Azure Data Lake Storage, HD Insight (Azure)
Oracle (P. 35)	Oracle Autonomous DW on Cloud@ Customer, Oracle Database	Oracle Database (AWS, Azure, OCI)	*Oracle Autonomous Data Warehouse (OCI)	Oracle Data Lake-house on OCI (OCI)	Oracle Big Data Service (OCI), OCI Data Flow (OCI), OCI Object Storage (OCI)
SAP (P. 38)	SAP HANA, SAP BW/4HANA, SAP IQ		*SAP Data Warehouse Cloud (AWS, Azure)	SAP HANA Cloud, Data Lake (Alibaba, AWS, Azure)	
Snowflake (P. 40)			*Snowflake Data Cloud (AWS, Azure, GC)	Snowflake Data Cloud (AWS, Azure, GC)	
Teradata (P. 43)	Teradata Vantage on Teradata IntelliFlex or VMware hardware	Teradata Vantage DIY (AWS, Azure)	*Teradata Vantage (AWS, Azure, GC)		
Vertica (P. 45)	Vertica Unified Analytics Platform, Vertica SQL for Data Lakes	*Vertica Unified Analytics Platform (Alibaba, AWS, Azure, GC)	Vertica Accelerator (AWS)		
Yellowbrick (P. 48)	*Yellowbrick Data Warehouse		Yellowbrick Cloud (Yellowbrick)		

Key: * = Featured product/service, AWS = Amazon Web Services, Azure = Microsoft Azure, GC = Google Cloud, IBM = IBM Cloud, OCI = Oracle Cloud Infrastructure

Source: Constellation Research

Amazon Web Services

Company: Public, founded 2006
Featured service:
Amazon Redshift (analytic database service)
Related services:
Redshift Spectrum/AWS Lake Formation (combined lake/warehouse service)
Amazon EMR (data lake service)
Public cloud option: AWS
On-premises option: Amazon EMR supported on AWS Outposts (Redshift not yet supported on Outposts)
Number of customers: More than 10,000 Redshift customers
Large deployment example (vendor-supplied): NASDAQ has a 1.1-petabyte-plus Redshift/Redshift Spectrum deployment supporting mixed reporting and ad hoc workloads and ingesting more than 4 terabytes of data per day.

Overview: Amazon Web Services (AWS) addresses both data lake and data warehouse needs with a range of services on its hyperscale public cloud. Data warehousing is addressed by Amazon Redshift, a distributed MPP columnar-database service that debuted in 2012. Redshift is based on the ParAccel Analytic Database, introduced in 2007, but since its launch, AWS has added many capabilities unique to Redshift. Scalability enhancing features, for example, include Redshift Spectrum and RA3 nodes. Performance has been enhanced via features such as concurrency scaling and the Advanced Query Accelerator (AQUA) caching layer. Data science capabilities are extended via user-defined functions (UDFs), using Python and Lambda, and via the 2021 general availability of Redshift ML. Redshift ML integrates with Amazon SageMaker to run custom algorithms within the database.

Redshift is joined at the hip with Amazon S3 object storage by way of Redshift Spectrum, a combined lake/warehouse architecture introduced in 2017, and next-generation RA3 nodes, introduced in 2019. Redshift Spectrum supports querying of high-scale structured and semistructured data in Amazon S3 without extract, transform, and load (ETL) operations. RA3 nodes effectively separate compute and storage charges and decisions, automatically moving cold data to S3 and prefetching hot data to

database-node-local solid-state drives (SSDs) according to query access patterns. Managed storage is charged at a fixed gigabytes-per-month rate, regardless of the split between S3 and SSDs.

AWS's most popular option for building data lakes is Redshift Spectrum in combination with AWS Lake Formation. The combination supports moving, storing, cataloging, cleansing, and securing data as part of a lake on S3, using the AWS Glue Data Catalog service, jobs, and crawlers. It addresses metadata management, storing structured data in Parquet or ORC formats, and defining centralized security and governance controls. Amazon EMR, a Hadoop and Spark service dating back to 2009, separates compute and storage in combination with S3. EMR addresses all data types, including unstructured data, and supports high-scale data engineering work with tools such as Spark, Hive, HBase, Flink, and Hudi.

Licensing/subscription terms: Amazon Redshift offers on-demand pricing with partial hours billed in one-second increments. AWS also offers capacity-based reserved-instance pricing with one-year and three-year discount levels as well as bidding-based spot pricing. Redshift Spectrum, the Concurrency Scaling feature, Redshift ML, and the AQUA caching layer are extra-cost options.

Amazon S3 object storage costs are measured in pennies per gigabyte per month, with discounts after 50 terabytes and 500 terabytes are exceeded. Extra charges apply for tiered fast access, data requests and retrievals, data transfers, replication, and certain analytic workloads. Amazon EMR is available on Amazon EC2, Amazon Elastic Kubernetes Service (EKS), and Amazon Outposts (on-premises) with on-demand, reserved (one-year/three-year discount), and spot pricing available.

Constellation's Analysis

Strengths:

- AWS provides a comprehensive portfolio of easily deployable, well-integrated, and highly scalable data warehouse and data lake cloud services headlined by Amazon Redshift, Redshift Spectrum/Lake Formation, S3, and Amazon EMR.

- Amazon Redshift scalability has been enhanced via Redshift Spectrum, and storage and compute decisions are effectively separated via RA3 nodes. Performance and predictability have been bolstered via the Concurrency Scaling feature and the AQUA caching layer.

- Redshift Spectrum and AWS Lake Formation have helped redefine and lower the cost of data lakes, yet these data stores can be readily accessed by myriad Amazon services, including Athena, Elasticsearch, EMR, Kinesis, and SageMaker.

Weaknesses:

- AWS offers no multicloud deployment options. On-premises deployment options for data-lake-related capabilities currently are limited to Amazon EMR on AWS Outposts. Redshift/Redshift Spectrum is not yet available on Outposts. Amazon S3 on Outposts currently is limited to archival storage.

- Amazon Redshift Spectrum and Concurrency Scaling are serverless (autoscaling) features that have been added to Redshift, but the underlying database service is not serverless. Customers must plan the size and compute power of their initial deployments and continue to monitor, manage, and scale resources as required.

- Redshift Spectrum is best suited to structured data, has limited support for semistructured data (in JSON format), and is not geared to unstructured data (although the latter is addressed by Amazon EMR).

- Redshift ML gives SQL jockeys a CREATE MODEL command supported by SageMaker Autopilot, but the related use of Amazon SageMaker and S3 is likely to require additional expertise and expense. Prebuilt ready-to-deploy data science options for non-data-scientists are lacking. Customers can import Python libraries and code their own user-defined functions.

Overall assessment: AWS covers all the bases with a comprehensive and well-integrated portfolio of data-warehouse- and data-lake-related services, but they're available only on AWS. "Multicloud" is not an option, and on-premises options on AWS Outposts are currently limited to Amazon EMR. Nonetheless, tens of thousands of customers are doing it all on AWS. The vendor is stepping up governance capabilities, but Constellation tends to see customers with collections of business-unit and/or use-case-focused warehouses and lakes rather than broad sharing and reuse of data supported by consolidated, centralized stores with fine-grained data-access controls and policy-enforcement regimes.

Cloudera

Company: Private, founded 2008
Featured service: Cloudera Data Platform (CDP) Public Cloud (combined lake/warehouse |platform/service)
Related products:
CDP Private Cloud
CDP Enterprise BYOL (marketplace offering)
CDP DataFlow (software)/CDP DataFlow for Public Cloud (streaming service)
Public cloud options: AWS, Azure, GC
On-premises option: CDP Private Cloud Plus
Number of customers: More than 2,000
Large deployment example (vendor-supplied)**:** SMG (Service Management Group) uses CDP data warehousing capabilities to support as many as 90 queries per second from its Experience Management reporting platform.

Overview: In its earliest days, Cloudera introduced the world to Hadoop and the concept of data lakes, and it led that market from 2008 into the middle of the last decade. Cloudera has since evolved to address huge technical shifts such as the separation of compute and storage and huge market shifts such as customer interest in cloud deployment and data lakes built on object storage.

Cloudera deserves credit as one of the first vendors to support data warehousing and data science on a single platform. This combination of capabilities on a single platform remains a point of differentiation backed by a single policy-based security, access control, and data governance scheme. Today's focus is on supporting hybrid and multicloud deployment requirements at scale. Recent acquisitions are aimed at accelerating the delivery of SaaS options in addition to the CDP Public Cloud PaaS option, already available in all three major public clouds.

Subscription terms: CDP Private Cloud is an on-premises offering with bare-metal clusters priced per node annually, with variable compute and storage costs above a cap and container-based analytic services whose price is based on compute and storage under management. CDP Public Cloud is a PaaS

deployed on AWS, Azure, and Google Cloud. Consumption can be billed hourly in a pay-as-you-go model or via (discounted) prepaid credits. CDP Public Cloud offers both virtual-machine-based and cloud-native container-based deployment options. Cloudera Machine Learning, Cloudera Data Warehouse, Cloudera Data Engineering, and Cloudera Operational Database are all built into the platform and can be used optionally, depending on customer interest in using these capabilities. All pricing includes support.

Constellation's Analysis

Strengths:

- This single consistent platform supports on-premises and multicloud deployment as well as a range of workloads, including data warehousing, ML/data science, data engineering, flow management, streaming analytics, and high-scale operational database requirements.

- The single policy-based security and governance regime (Cloudera SDX) covers data from ingest and transformation to query and visualization. The same policies and data tagging support all workloads and governance requirements for all users.

- The CDP PaaS architecture enables customer IT teams (in CDP Private Cloud deployments) and Cloudera (in CDP Public Cloud deployments) to manage the system control plane while business units (private)/Cloudera customers (public) control the data within their on-premises environments or cloud accounts.

- Cloudera's Apache Impala SQL query engine has demonstrated highly cost-effective performance for mixed query loads associated with BI/data warehouse deployments.

Weaknesses:

- Complexity and the continuing evolution of the CDP platform demand strong technical skills on the part of customers to understand workload requirements and plan corresponding resources and the evolution of deployments.

- The Impala SQL query engine does not address data science or ML. These workloads are separately addressed by the data-science-oriented lake/ML aspects of CDP.

- The 2021 acquisition of Cazena will accelerate the delivery of multicloud SaaS offerings, but these services (and their simplicity and ease-of-use advantages) will not be available until 2022.

Overall assessment: Cloudera's strengths are its hybrid and multicloud support with a single consistently secured and governed data lake platform that spans data warehousing, ML, data engineering, and operational database needs. The company is moving quickly to bring more automation, simplicity, and agility to a platform that has demonstrated cost-effective performance for the highest-scale enterprise-class workloads on the planet.

Databricks

Company: Private, founded 2013
Featured service: Databricks Unified Analytics Platform (combined lake/warehouse platform as a service)
Public cloud options: PaaS on Alibaba, AWS, Azure, GC
On-premises option: No standard offering (private-cloud deployment said to be available to select customers on a case-by-case basis)
Number of customers: More than 6,000
Large deployment example (vendor-supplied): One large unnamed Databricks Delta Lake deployment is said to support 32 concurrent users issuing 15,000 queries per hour.

Overview: Databricks is a leading proponent of the combined lake/warehouse platform. The Databricks Lakehouse Platform addresses data engineering, data science, data streaming, and data warehouse workloads on a single platform. The company introduced its Apache Spark–based high-scale data lake PaaS on AWS in 2013. A partnership with Microsoft led to the introduction of Azure Databricks in 2017. The platform originally catered to data engineers and data scientists, supporting high-scale data processing as well as streaming data and data science pipelines via a collaborative notebook-based workspace supporting coding in Python, R, Scala, and SQL. Model building is supported by Databricks-provided and bring-your-own ML frameworks as well as an AutoML feature that simplifies model

creation while exposing all experiments and notebooks to the user for refinement. Users track features, experiments, and models in one place, with MLflow easing handoffs from development to production.

In 2017 Databricks added Delta Lake, a "lakehouse" architecture supporting performant warehouse-style analysis on the same data and platform supporting data engineering, data science, and streaming data needs. Delta Lake tables support ACID transactions and unified streaming and batch processing. Queries are executed via the Photon Engine, a vectorized, MPP query engine with a cost-based optimizer. The architecture integrates into the Databricks Enterprise Security model, including cell-level access control, auditing, and HIPAA-compliant processing. Structured data is stored in the open, columnar Parquet or AVRO formats on low-cost object storage.

More recent announcements from Databricks include the 2020 launch of Databricks on Google Cloud (GC). Delta Live Tables, introduced in May 2021, simplify and ensure the performance of data pipelines, including ETL and ELT workloads.

Licensing/subscription terms: On-demand per-second pricing is based on compute capacity or reserved-instance pricing with one-year and three-year commitment discount levels.

Constellation's Analysis

Strengths:

- PaaS eases deployment and scaling on Alibaba, AWS, Azure, or Google Cloud, yet the data tier resides in the customer's account. The Databricks account console provides a unified view of data; workloads; and users across separate Databricks deployments on Alibaba, AWS, Azure, and Google Cloud.

- Platform unifies lake, warehouse, and streaming workloads on a single platform against a single copy of data/user-management and security regime. The platform delivers analytical, data science, data engineering, and data streaming tools and capabilities under a single subscription.

- Platform separates compute and storage, providing autoscaling of clusters and on-demand storage managed within customer-defined cost and performance guardrails.

Weaknesses

- Lakehouse deployments to date have yet to support the extremes of data warehouse concurrency—in terms of users as well as queries—often supported by conventional analytical databases. (One of the larger references available at this writing supports 32 concurrent users and 15,000 queries per hour.)

- Platform is available almost exclusively in the cloud (don't count on a one-off, Databricks on-premises support deal unless you are a large and/or influential customer).

- Announced Unity Catalog promises attribute-based data access control and governance as well as global metadata management across separate cloud instances, but don't expect general availability of this feature until early 2022.

Overall assessment: Databricks appropriately describes its platform as a lake-first offering. Indeed, the vast majority of customers to date were using Databricks first and foremost for its well-reputed lake and data science capabilities. It's early days for Delta Lake warehouse capabilities, with the 1.0 release in 2021. The warehouse capabilities are geared to production BI workloads, including complex joins and queries against large tables, but they're less well suited, at this time, to high-concurrency requirements and unpredictable ad hoc query workloads. Databricks in 2021 added an AutoML feature on the warehouse front, putting classification, linear regression, and time series forecasting capabilities in the hands of SQL-savvy programmers.

Dremio

Company: Private, founded 2015
Featured product/service: Dremio (lake query engine)
On-premises option: Dremio Software
Public cloud option(s): Dremio Cloud
Number of customers: More than 200
Large deployment example (vendor-supplied): A large financial services company has a 3-petabyte data lake and is using Dremio to support querying and interactive dashboards against terabyte-scale datasets.

Overview: Dremio is a proponent of combined lake/warehouse architectures, but it focuses entirely on providing the interactive SQL engine that directly queries data stored in high-scale object-store and Hadoop Distributed File System (HDFS)–based data lakes. It provides a semantic layer user interface that provides a single secure access point to data and a way to curate, analyze, and share datasets. Dremio employs multiple technologies to accelerate querying directly against data in data lakes, starting with Apache Arrow, the open source columnar in-memory data format that Dremio helped create. Other accelerators include Intel-chip-based vectorized parallel processing; caching; and Dremio Data Reflections, which precompute aggregations, sorts, and other data operations to minimize computation required at query runtime. Dremio is an alternative to (or a way to minimize investments in) traditional data warehouses and the associated costs and complexity of moving and copying data into data warehouses.

Subscription terms: Customers of Dremio's fully managed cloud offering pay on demand (billed at the end of each month) or prepay (at discounted rates) by purchasing Dremio Compute Units. Either way, customers are charged only when the engines are running/processing queries, based on time in use. Dremio Software (deployed by the customer in the cloud of its choice or on-premises) is subscription-based, priced per node annually.

Constellation's Analysis

Strengths:

- Works with open, structured file formats such as Parquet, ORC, CSV, Avro, JSON, and Excel; with open table formats, including Apache Iceberg and Delta Lake; or via a metastore such as Hive Metastore, AWS Glue Catalog, or Project Nessie.

- Creates a semantic layer and open, portable metadata that can be used by data analysts and data scientists to discover, curate, analyze, and share datasets.

- Accelerates querying directly against data in data lakes, exploiting Arrow/Gandiva engine acceleration; Dremio Columnar Cloud Cache for I/O acceleration; Dremio Data Reflections, which minimize data scanning; and query planning and filtering features that optimize query performance.

- Arrow Flight network protocol supports fast, high-scale data exchange among engines used on the lake, which is particularly useful for supporting data science workloads.

Weaknesses:

- Dremio Software lacks automated system and workload management options. Systems must be appropriately sized and managed (manually) and query queues and policies developed to govern performance and routing. Dremio Cloud eases deployment and supports automated replication and scaling within the bounds of user-defined rules.

- The list of prebuilt database connectors is short, and queries on sources external to the lake are often limited to JDBC-based data transfer. A configuration-driven framework supports the creation of custom relational connectors.

- Dremio is totally focused on SQL, leaving data science to scientists and their choice of third-party tools and processing options on the data lake. AutoML features are not available, and in-database processing options via user-defined functions are limited.

Overall assessment: Dremio is a small but well-funded vendor that raised $135 million in its last (D) funding round, in January 2021. The assumption with Dremio is that customers have or are building a high-scale data lake (using third-party cloud services and/or commercial/open source software). Dremio is added to that lake environment to support SQL-centric workloads, promising data-warehouse-like performance. Achieving required performance demands appropriate sizing, monitoring, and management of system resources and workloads. There's a bit of a learning curve, and it's not for newbies, but for the many Dremio customers that have deep engineering benches running their high-scale data lakes, management of Dremio is the least of their challenges.

Google Cloud

Company: Google Cloud, launched as unit of public Google/Alphabet in 2008
Featured service: BigQuery (analytic database service)

Related services:
BigQuery Omni (multicloud object store query option)
Dataproc and Dataproc Metastore (data lake services)
On-premises options: None
Public cloud options: BigQuery (GC), BigQuery Omni (AWS, Microsoft Azure), Dataproc (GC)
Number of customers: BigQuery has more than 10,000 customers.
Large deployment example (vendor-supplied): Verizon Media has more than 100 petabytes of data stored in BigQuery and executes more than one million queries per month, scanning more than 1 exabyte of data monthly.

Overview: BigQuery, introduced in 2011, was the early pioneer in the cloud data warehousing space, offering automated scaling and performance optimization. Customers don't have to concern themselves with the details of sizing this columnar analytical platform, even when deployments scale (routinely) into the petabytes.

Subscription terms: On-demand or fixed-rate (monthly/yearly) purchasing of compute capacity based on bytes scanned or "slot" compute capacity. Slot/byte scan capacity can be added in granular increments.

Constellation's Analysis

Strengths:

- Automated database deployment and scaling maintenance and performance tuning minimize administrative overhead, freeing data engineers and analysts to focus on harnessing new data sources and supporting new workloads.

- BigQuery Omni provides multicloud external table querying of data and retrieval of result sets from customer data in AWS or Azure object stores.

- BigQuery ML, an early AutoML feature introduced in 2018, has been adopted by more than 80% of the 100 largest BigQuery customers (by deployment size), putting data science capabilities in the hands of SQL programmers without requiring coding. The BigQuery BI Engine option adds an in-

memory analysis layer to ensure fast query performance in demanding scenarios involving complex data and high concurrency.

- BigQuery is well integrated with myriad Google Cloud services, including the closely related Google Cloud (object) Storage, Dataproc (Hadoop and Spark lake services), Dataflow (streaming analytics service), and Vertex AI platform.

Weaknesses:

- There's a learning curve to balancing BigQuery-automated scaling actions and performance optimization with available administrative tools for managing schemas, setting up cost guardrails, and ensuring performance for mission-critical workloads. Performance tuning becomes more challenging as the number of tables multiplies. Slot Estimator, now in preview, is designed to help with capacity sizing and cost optimization.

- Dataplex data fabric and data governance service, recently made generally available, is expected to fill current gaps in metadata management/cataloging, data lifecycle management, and data policy enforcement.

- Google Cloud is expanding its portfolio of data integration options—adding the Datastream change-data-capture service to Cloud Data Fusion, Data Transformation, and Dataflow—but Google BigQuery would benefit from more and deeper integrations with popular third-party data integration vendors.

Overall assessment: Google BigQuery is differentiated in terms of automation and ease of scalability. It has matured to cover data warehouse optimization and governance features demanded by enterprises. BigQuery Omni is a unique offering among the major public cloud providers, supporting external table querying with on-cloud querying of AWS and Azure object stores and egress of query result data subsets back to BigQuery on Google Cloud. Google Cloud has upped its enterprise focus over the last two years, making progress on delivering expected customer-support capacity and expanding integrations and solutions available through its third-party-vendor/partner ecosystem.

IBM

Company: Public, founded 1911
Featured service: IBM Db2 Warehouse on Cloud (analytic database service)
Related products and services(s):
Netezza as a Service and Netezza Performance Server IBM Db2 on Cloud Pak for Data (containerized deployment on Red Hat OpenShift)
IBM Analytics Engine (Spark Service on IBM Cloud)
IBM SQL Query (SQL query of structured data in object-store-based data lakes)
IBM BigSQL (lake query engine included with Cloud Pak for Data)
On-premises option(s): IBM Db2 on bare metal, Netezza Performance Server deployed on-premises, or IBM Db2 deployed as part of Cloud Pak for Data in containerized fashion on Red Hat OpenShift on premises
Public cloud option(s): IBM Db2 Warehouse on Cloud (AWS, IBM), Netezza as a Service (Azure, IBM), or IBM Db2 deployed as part of Cloud Pak for Data in containerized fashion on Red Hat OpenShift on any public cloud
Number of customers: More than 11,000 Db2 customers
Large deployment example (vendor-supplied)**:** GEICO has a 24-node instance of IBM Db2 Warehouse on Cloud. The deployment handles 480 terabytes of data (120 TB compressed), supporting 1,500 users with mixed reporting and ad hoc query workloads.

Overview: IBM is the birthplace of the relational database, and the roots of its flagship Db2 product extend back to a database for IBM mainframes introduced in 1983. A complete rewrite in 1993 for Linux, Unix, and Windows servers has since steadily evolved to embrace advances such as distributed and shared-nothing architecture, MPP, and columnar and in-memory performance.

With the massive shift to the cloud, IBM bet on enterprise interest in consistent hybrid and multicloud deployment via containerization with its $34 billion acquisition of Red Hat in 2019. IBM's unified cloud data management stack is Cloud Pak for Data, which includes Db2 as Db2 Cartridge. This platform can now be spun up and scaled on-premises or on any public cloud via the Red Hat OpenShift container platform. For customers not interested in (or ready for) container management, IBM also offers as-a-service options such as IBM Db2 Warehouse on Cloud. Behind the scenes, IBM uses the same container technologies to manage and scale its own cloud services.

Subscription terms: Db2 Warehouse on Cloud is priced on-demand, based on instance hours, virtual processor core hours, and volume-discounted storage hours.

Constellation's Analysis

Strengths:

- IBM Db2 deployment options span on-premises deployment, containerized deployment (on-premises or on any cloud) as part of Cloud Pak for Data deployed through Red Hat OpenShift, and as-a-service consumption as IBM Db2 Warehouse on Cloud (available at this writing on AWS and IBM Cloud).

- IBM Db2 supports multimode analysis spanning JSON documents, graph query, spatial analysis, and in-database ML functions (implemented as user-defined functions and stored procedures) for data exploration, data preprocessing, model training, model evaluation, model deployment, and scoring.

- IBM Db2 evolved for cloud deployment with REST API and the ability to query structured and columnar formats, such as Parquet, CSV, TXT, and Avro, in cloud or on-premises object storage and join results with DBMS block storage.

- Adaptive Workload Manager (job scheduling), introduced in 2021, improved Db2 workload throughput by as much as 30% and query performance by as much as 20%.

Weaknesses:

- Containerization and associated automation of Db2 management are works in progress, currently covering software installs and upgrades. Automated lifecycle management (including backup and failover) is planned for mid-2022. Automated management of logs, alerts, and metrics, plus automated horizonal and vertical scaling are on the roadmap for 2023 and beyond. For now, Db2 is not a highly automated product or service in terms of scaling and workload management.

- IBM Db2 Warehouse on Cloud service currently stores customer data within IBM accounts on AWS or IBM Cloud. An option to store data within customer cloud accounts won't be available until the second half of 2022.

- Choice of clouds for as-a-service offerings is limited at this writing, with IBM Db2 Warehouse on Cloud supported on AWS and IBM Cloud, Netezza as a Service on Azure and IBM Cloud, and IBM Analytics Engine and IBM SQL Query available only on IBM Cloud.

Overall assessment: No vendor is farther along than IBM when it comes to handing customers the keys to consistent, containerized deployment on-premises or in the cloud of their choice. For now, it's an approach that customers have to drive, but the release of IBM Cloud Satellite (OpenShift offered as a service, set for general availability in 2022) will simplify container deployment. IBM also can provide supporting services, but not all customers are interested in deploying, managing, and using OpenShift to manage deployments in multiple environments. With that in mind, IBM will continue to make as-a-service offerings such as IBM Db2 Warehouse on Cloud available. IBM's containerization investments will eventually bring more automated management and scalability to Db2, but the complete journey will take a few more years.

Incorta

Company: Private, founded 2013
Featured product/service: Incorta Direct Data Platform (combined lake/warehouse platform)
On-premises option: Subscription software
Public cloud options: Incorta cloud service (GC), marketplace deployments (AWS and Azure)
Number of customers: More than 90
Large deployment examples (vendor-supplied): Largest customer (an unnamed consumer tech manufacturer) has more than 9,000 concurrent users. Broadcom has more than 2,000 concurrent users running more than one million queries per month.

Overview: Incorta's Direct Data Mapping technology quickly connects to myriad enterprise applications and ingests, transforms, and synchronizes data to drive analytics, data engineering, and data science. Data is stored on a single platform that provides business semantics, security and access control, scheduling, monitoring, and alerting. The platform supports BI and analytics via a performant in-memory-optimized analytics engine that is also integrated with third-party analytics tools such as Microsoft Power BI and Tableau. Data engineering and data science are supported via preconfigured Spark clusters and job-scheduling capabilities. The emphasis is on low-latency insight, including self-service reporting, dashboarding, data visualization, and ad hoc exploration.

Subscription terms: Consumption-based pricing includes data management charges (list-priced at $120 per terabyte per month plus $2,000 per Incorta Compute Unit per month) and analytics charges (list-priced at $60 per "analyzer'" user per month and $15 per "viewer" user per month).

Constellation's Analysis

Strengths:

- Incorta's agile data-ingestion and direct data-mapping approach provides a fast on-ramp and platform for data with less of the technical friction associated with conventional ETL and data warehousing.

- Customers routinely expand on initial (often departmental) deployments launched to quickly solve problems and work around the slower-moving, conventional data-integration/warehouse/BI platform approach.

- Time to insight is decreased by prebuilt analytical applications for Oracle E-Business Suite, Oracle Enterprise Resource Planning (ERP) Cloud, Oracle's JD Edwards, Salesforce, and SAP as well as Incorta Blueprints for lead-to-cash cycle analytics, policy and claims analytics, general ledger analysis, order management analysis, accounts receivable analysis, and more.

Weaknesses:

- Incorta typically complements, rather than replaces, data warehouse and BI deployments within enterprises (although many midsize customers may use it as a warehouse alternative). Incorta embraces this role, offering integrations with the likes of Azure Synapse, Power BI, and Tableau.

- Incorta's analytics engine is focused on SQL-centric analysis and lacks AutoML features or extended in-SQL support for Python or R. Data science is left to data scientists using Incorta's Spark layer, packaged libraries, and their choice of tools.

- Incorta deployments, whether on-premises or in the cloud, must be sized and optimized by the customer via monitoring and management tools. Stepped-up automation is on the roadmap for the Incorta cloud service.

Overall assessment: Incorta is a small but well-funded vendor that presents a unique approach to ingesting, managing, and analyzing data. It typically gets its foot in the door when teams or projects hit a wall, and that wall may well be the time-constrained and overwhelmed incumbent data warehousing and BI platform team. Land-and-expand adoption validates Incorta's time to value and speedy query performance. There are converts that have replaced it all with Incorta, but the product more often plays a complementary role, particularly alongside the ERP, CRM, and other enterprise applications and systems for which there are prebuilt connectors and analytic content.

Microsoft

Company: Public, founded 1975
Featured service: Azure Synapse Analytics (combined lake/warehouse platform)
On-premises option: None
Public cloud option(s): Microsoft Azure
Number of customers: Not officially disclosed (estimated in the thousands for Synapse)
Large deployment examples (vendor-supplied): Dozens of unnamed customers are said to be running multipetabyte workloads.

Overview: Introduced in 2019, Azure Synapse Analytics (Synapse) is a converged data lake and data warehouse PaaS that also incorporates data integration (with more than 100 connectors and low-code ETL/ELT functionality) and integrated data catalog search capabilities. Synapse provides a single management interface for data integration, workload management, monitoring, and security and access control over a shared data platform that supports data integration as well as lake and warehouse workloads. The platform offers scalable Spark pools for data processing and data science work and SQL pools for BI workloads and ad hoc analytics.

Licensing/subscription terms: Serverless, on-demand managed service offers billing by the minute or hour (depending on compute) or by the gigabyte/terabyte (depending on storage choices). Also available is a pay-per-query option priced at $5 per terabyte. Dedicated service offers 37% and 65% discounts for one-year and three-year reserved instances, respectively. Dedicated service adds a performance-enhancing hierarchical cache option. Customers typically reserve dedicated capacity for expected workloads and can handle peak loads with serverless, on-demand capacity.

Constellation's Analysis

Strengths:

- Unified cloud-based platform for data supports high-scale warehouse, lake, and integration workloads with a single interface for workload management, monitoring, and security and access control.
- Synapse SQL PREDICT function supports in-engine runtime ML scoring by invoking any algorithm supported by the Open Neural Network Exchange (ONNX) ecosystem.
- Native catalog search integration with Azure Purview in the Synapse Studio experience simplifies data discovery for developers.
- SQL Polaris engine (currently limited to pay-per-query use) promises extreme query performance at multipetabyte scale.

Weaknesses:

- Current architecture constrains concurrency to 128 active queries per database instance.
- Currently lacks on-premises or multicloud deployment options. (Azure Arc container-service integration is not officially on the roadmap, but Constellation expects that it will eventually support on-premises and multicloud deployment of Synapse.)
- Spark is currently an ephemeral job service that does not support Spark Streaming. (Streaming scenarios currently are supported in Transact-SQL [T-SQL].)

Overall assessment: It's early days for Synapse. Constellation expects that Microsoft will soon address current query concurrency limitations and eventually address the single-cloud deployment limitation. The highly scalable Polaris SQL engine promises to differentiate SQL warehouse capabilities as it becomes available through dedicated services. Integration with Azure Purview simplifies data access within—and eventually beyond—Azure.

Oracle

Company: Public, founded 1977
Featured service: Oracle Autonomous Data Warehouse (analytic database service)
Related services:
Oracle Data Lakehouse on OCI (Combined Lake/Warehouse)
Oracle Big Data (Hadoop service)
Oracle Cloud Infrastructure (OCI) Data Flow (Spark service)
OCI Object Storage (object store service)
On-premises options: Oracle Autonomous Data Warehouse on Cloud@Customer, Oracle Database
Public cloud options: Oracle Autonomous Data Warehouse (on shared or dedicated infrastructure), Oracle Big Data, OCI Data Flow, and OCI Object Storage (on Oracle Cloud-OCI)
Number of customers: Hundreds of thousands on Oracle Database; Autonomous Database figures not disclosed
Large deployment example (vendor-supplied): Baxter, a global pharmaceuticals manufacturer, has a 65-terabyte Oracle Autonomous Data Warehouse deployment supporting more than 1,000 concurrent users and more than 1,000 queries per hour.

Overview: Built on the foundation of Oracle Database, Oracle Autonomous Database was announced in 2017 and became generally available in 2018 as the sibling cloud services Oracle Autonomous Data Warehouse and Oracle Autonomous Transaction Processing. Oracle later introduced Oracle Autonomous JSON Document Database. As the name suggests, automation built into Autonomous Data Warehouse (and Autonomous Transaction Processing) takes care of resource provisioning, scaling, and tuning, adding to the cloud-service benefits of not having to install, configure, upgrade, or patch software. Oracle says Autonomous Data Warehouse takes care of 90% of the tedious database administrative work that would otherwise be required to run Oracle Database, leaving DBAs to create and maintain connections to applications and administer users, groups, and access privileges. Oracle doesn't break out what percentage of its hundreds of thousands of customers use Oracle Autonomous Database, but it has been reporting triple-digit annual increases in adoption since the debut of the service.

Oracle also supports data lakes, through services on its Gen 2 OCI such as the Oracle Data Lakehouse on OCI (released days before this report was published and built on OCI Object Storage), Oracle Big Data, and OCI Data Flow. But these services get little promotion from Oracle compared with the emphasis the company puts on Oracle Autonomous Database as the future of its years-long run leading the relational database market.

Subscription terms: Autonomous Data Warehouse on Shared Infrastructure is a multitenant, metered subscription service with CPU and storage charges. CPU consumption is based on resource allocation and consumption per second. Storage metering is based on capacity (in terabytes) per hour. Existing Oracle Database customers have a bring-your-own-license option at discounted rates. Dedicated Infrastructure and Cloud@Customer deployments require a ¼ Exadata rack minimum. The Dedicated reservation is billed nonmetered, including the storage for Autonomous Database. The CPU consumption for Autonomous Database is based on resource allocation and consumption of Dedicated Infrastructure.

Constellation's Analysis

Strengths:

- Oracle Autonomous Database gives the hundreds of thousands of existing Oracle Database and Oracle applications customers a move-forward option with the cost-saving benefits of a vendor-managed cloud service and automated system and workload management. Complete compatibility with Oracle Database, migration tools, and a bring-your-own-license option ease the transition.

- The Dedicated Infrastructure option gives large customers their own isolated yet scalable and vendor-managed private cloud deployment. Cloud@Customer enables customers to deploy Exadata in the data center of their choice while retaining the benefits of vendor systems management and Autonomous Database services.

- Granular sizing of Autonomous Data Warehouse and Autonomous Transaction Processing enables customers to add one CPU at a time (on shared or dedicated infrastructure), avoiding overprovisioning. In addition, an autoscale feature adds as much as three times the customer's base

CPU capacity on demand to handle peak workloads. Customers pay only for capacity used at their prevailing metered rates.

- Oracle Database (and, therefore, Autonomous Data Warehouse) supports spatial, graph, and JSON analysis and more than 30 in-database data science options coupled with a new Oracle Machine Learning Notebooks service on Autonomous Data Warehouse, which includes a Python interface and REST API.

Weaknesses:

- Oracle Autonomous Database runs on Oracle's proprietary Exadata engineered system, so Autonomous Data Warehouse and Autonomous Transaction Processing services are available only on Oracle Cloud (or on Cloud@Customer). Autonomous Database services do not run on Azure, but a high-speed interconnect to Azure regions provides access to data without egress charges through a partnership with Microsoft.

- Dedicated (cloud) Infrastructure and (on-premises) Cloud@Customer deployments require a ¼ Exadata rack minimum. Cloud@Customer deployments offer less-granular scaling and require a four-year subscription term.

- Oracle Autonomous Database offers its best advantages to Oracle customers running (well-integrated) Oracle applications, Oracle Analytics Cloud, and Oracle data management options. Advantages diminish for customers with heterogeneous tech, application, and cloud landscapes.

Overall assessment: Oracle Database has topped the database popularity charts for years, and with good reason. Extensive enterprise-grade capabilities, constant investment in new features, and continuity and backward compatibility from release to release have kept Oracle Database at the top of the enterprise database market. Oracle faces its toughest competition on the data warehousing front rather than with OLTP. What's more, with operational data volumes and data lakes rapidly growing on rival public clouds, Autonomous Data Warehouse is handicapped by being available only on Oracle Cloud. It helps that there's a high-speed interconnect to Azure regions as part of a pact with Microsoft, but when the center of data gravity is on a rival public cloud, the complexities and costs of data

movement are likely to be unfavorable. As for Oracle-centric customers with lots of Oracle skills that are using Oracle applications and Oracle data management services, Oracle Autonomous Database is a compelling and obvious choice.

SAP

Company: Public, founded 1972
Featured services: SAP Data Warehouse Cloud (analytic database service)
Related products and services:
SAP HANA Cloud, Data Lake (combined lake/warehouse service)
SAP HANA (columnar in-memory database)
SAP BW/4HANA (data warehouse tied to SAP applications)
SAP IQ (high-scale columnar analytical database)
On-premises options: SAP HANA, SAP BW/4HANA, SAP IQ
Public cloud options: SAP Data Warehouse Cloud (AWS, Azure), SAP HANA Cloud, Data Lake (Alibaba, AWS, Azure)
Number of customers: SAP HANA: 56,240 overall; SAP Data Warehouse Cloud: not disclosed
Large deployment examples (vendor-supplied): An unnamed insurance company is enabling multiple use cases and mixed workloads, starting with a 10-terabyte deployment of SAP HANA Cloud, Data Lake for the first phase of the project and increasing to 20 terabytes in 2022.

Overview: Introduced in November 2019, the multicloud SAP Data Warehouse Cloud service (available on multiple regions of both AWS and Azure at this writing and planned for GC in 2022) is part of SAP HANA Cloud. SAP Data Warehouse Cloud is the high-performance platform, with columnar and in-memory performance and tiered storage options. SAP Data Warehouse Cloud is complemented by SAP HANA Cloud, Data Lake, a lake/warehouse that combines high-scale/low-cost file storage, governance, and access on object storage with SQL querying via the SAP IQ engine. SAP Data Warehouse Cloud can be combined with on-premises SAP BW/4HANA, SAP HANA data warehousing, or SAP IQ deployments.

Subscription terms: SAP Data Warehouse Cloud subscription charges are billed monthly, based on selected blocks of RAM, CPU, storage, and SAP HANA Cloud, Data Lake capacity. The approach ensures

predictable budgeting. Discounts are negotiated per customer, based on contract duration and overall spend with SAP. Pay-as-you-go pricing options are planned but not available at this writing.

Constellation's Analysis

Strengths:

- SAP Data Warehouse Cloud takes advantage of SAP HANA–native columnar and in-memory performance for "hot," frequently queried, data while also offering "warm" tiered storage options on SSD or SAP HANA Native Storage Extension. SAP HANA Cloud, Data Lake provides a high-scale object-store-based tier for cost-effectively storing structured, semistructured, and unstructured data.

- SAP Data Warehouse Cloud's "federated first" approach to building a warehouse enables business analysts to remotely access desired data sources and datasets—on-premises or in multiple clouds—without moving or copying data. A no-code user interface supports data modeling and development of business logic across these assets. Thereafter, select data can be replicated into tiered storage as needed, without the model changes, to support service-level requirements.

- A "Space" concept enables administrators to consolidate and assign workload priorities and blocks of RAM, CPU, storage, and HANA Cloud, Data Lake capacity to specific users, groups, projects, and/or lines of business, facilitating workload management and IT centralized service charges based on capacities assigned.

- SAP Data Intelligence offers data integration, data engineering, data catalog, and data science capabilities (including Python and R, Spark, PySpark, and Spark SQL) that can be used in conjunction with structured, semistructured, and unstructured data in SAP HANA Cloud, Data Lake.

Weaknesses:

- SAP Data Warehouse Cloud is limited to scale-up scaling (using more powerful cloud infrastructure) rather than the scale-out approach (adding more nodes of parallel processing). A scale-out "Elastic Read Node" option is on the roadmap. Single SAP Data Warehouse Cloud instances currently are

limited to 22 terabytes of combined memory and storage. Adjunct SAP HANA Cloud, Data Lakes support unlimited capacity on object storage.

- SAP Data Warehouse Cloud does not query CSV, Parquet, or ORC files directly (whether they reside in SAP HANA Cloud, Data Lake or in third-party object stores). Such files can be included in the federated data model, but files must be transformed and replicated as tables to support querying within SAP Data Warehouse Cloud.

- Automation is limited to cost-based query optimization. System sizing, scaling, and data tiering choices must be executed manually.

Overall assessment: At this writing, SAP Data Warehouse Cloud is nearing its second anniversary. Nonetheless, the service has come a long way quickly, building on the 10-year-old SAP HANA in-memory DBMS. SAP HANA has been proven in data warehousing scenarios on its own and as the backbone of multiple releases of SAP BW. At this writing, SAP Data Warehouse Cloud scaling limitations keep it out of the 100-terabyte-plus league (although adjunct SAP HANA Cloud, Data Lake deployments support unlimited scaling on object storage). With its native integration with SAP applications and prebuilt business content packages for lines of business and industries, think of SAP Data Warehouse Cloud as a next-generation complement to or replacement for SAP BW built for a multicloud world.

Snowflake

Company: Founded 2012; initial public offering, 2020
Featured service: Snowflake Data Cloud (analytic database service and combined lake/warehouse service)
Public cloud option(s): AWS, Azure, and GC
On-premises option: None
Number of customers: 4,900+
Large deployment examples (vendor-supplied): Adobe and Nielsen have 2.6-petabyte and 3.7-petabyte deployments, respectively. Keboola has a 378-terabyte deployment supporting 8,494 users and 435,793 data engineering jobs per day.

Overview: Snowflake started out as a cloud-based data warehouse company that was a pioneer in separating compute and storage decisions. The company has since evolved and expanded its vision and the scope of its capabilities to do more with data. Today Snowflake bills the Snowflake Data Cloud as a data lake platform, a data science platform, and a data marketplace as well as a data warehouse platform.

Snowflake is a single platform that runs on multiple clouds (rather than presenting a choice of independent as-a-service offerings built on the same database). It's a highly automated platform that hides all the underlying infrastructure and service details of the specific clouds that customers choose to run on, whether that's AWS, Azure, Google, or combinations among the three. Customers can add automated replication and failover options that work across cloud regions or even across clouds. With separation of compute and storage, customers can set up an unlimited number of virtual warehouses on top of the same shared storage, with each warehouse providing workload isolation, security and governance, and workload-appropriate performance guarantees and cost controls.

Subscription terms: Snowflake Data Cloud uses consumption-based pricing, either on demand (as compute is consumed) or via prepurchased capacity credits with discounted pricing tiers. Smaller customers and customers just starting out often choose the on-demand model, but the majority of customers end up converting to the capacity approach.

Constellation's Analysis

Strengths:

- Customers use one consistent Snowflake service, regardless of which cloud(s) they choose to run on. Behind the scenes, Snowflake storage and single-cluster or autoscaling multicluster virtual warehouses are built on infrastructure on the customer's choice of cloud(s), but the details of deployment are taken care of by the system and system management is highly automated.

- Infinite numbers of virtual warehouses/workloads run against shared storage built on low-cost cloud object storage. Each warehouse supports workload-appropriate governance and performance ensured via automated storage tiering/caching and optimization capabilities.

- Autosuspend policies used in combination with task-specific virtual warehouses enable customers to sleep selected workloads, thereby saving money. ETL and data-loading workloads, for example, can be turned off immediately, whereas warehouses employing query caching might be suspended more selectively to ensure performance on demand.

Weaknesses:

- Data stored in Snowflake resides in Snowflake-managed cloud accounts, not the customer's public cloud account. You can still get cloud discounts by buying Snowflake through the AWS or Azure marketplaces, but not so through GC. (A Virtual Private Snowflake option enables customers to use their own public cloud VPN and dedicated resources.)

- Snowflake offers no option for running on-premises. Data not loaded into Snowflake (whether on-premises or in customer cloud accounts) can be queried via external table access, with the option of adding materialized views. Performance will vary, depending on connections and the complexity of the data.

- SnowPark data engineering and data science options are in preview at this writing and are limited to Java and Scala. Constellation expects to see (previously announced) Python support in preview in 2021, with general availability in 2022. Support for unstructured data is also in preview.

- Some automation features consume compute capacity, so expect incremental costs beyond ETL/ELT and query workloads.

Overall assessment: Snowflake helped open the industry's eyes to the advantages of separating compute and storage and automating scaling and management. It remains a leader on both fronts, and its use of object storage and automated multicluster scaling are being copied. Ease of use has led to fast adoption and, as a result, a bit of sticker shock for some customers. In response, Snowflake has come a long way with resource monitoring features and cost management guardrails. Snowflake's data lake vision is well ahead of the current reality when it comes to semistructured- and unstructured-data support and (non-SQL-centric) data engineering and data science capabilities (the company is counting

heavily on partners to deliver on these fronts). As a multicloud-native unified and highly automated data warehouse platform, Snowflake is unique in the market and is attracting customers with good reason.

Teradata

Company: Public, founded 1979
Featured service: Teradata Vantage in the Cloud (analytic database services on AWS, Azure, and GC)
Related products:
Teradata Vantage cloud-deployable "DIY" marketplace offerings on AWS, Azure
Teradata Vantage on-premises on Teradata or VMware hardware
On-premises option: Teradata Vantage (on Teradata IntelliFlex or VMware commodity hardware)|
Public cloud options: Teradata Vantage in the Cloud (on AWS, Azure, GC)
Number of customers: 1,400
Large deployment examples (vendor-supplied): Groupon uses Teradata Vantage on AWS for a 400-terabyte-plus deployment supporting 80 to 100 concurrent users executing 2.5 million queries per day.

Overview: Teradata has a more-than-40-year history in data warehousing and pioneered the use of MPP. Many of the largest and most demanding data warehouse deployments—in terms of data scale and user and query concurrency—continue to run on Teradata. The company has shifted to a cloud-first strategy over the last five years; its analytic and data management platform was first made available as a service on AWS in 2016, followed by availability on Azure in 2017 and Google Cloud in 2020. Diverse analytical capabilities are a calling card, with 200-plus data science functions available for in-database processing.

Subscription terms: Teradata offers two cloud pricing models: blended (capacity-based) and consumption (usage-based). Blended subscription options include monthly, one-year, and three-year durations. Consumption subscription options include one-year and on-demand. Capacity and usage are measured by compute and storage and can be scaled independently in granular increments as small as 5% of current system capacity.

Constellation's Analysis

Strengths:

- Teradata Vantage software is identical, whether delivered as a vendor-managed cloud service, deployed and managed by customers through "DIY" cloud marketplace offerings, or deployed on-premises on Teradata IntelliFlex systems or commodity hardware. This eases migration and promotes hybrid and multicloud portability.

- Teradata QueryGrid data fabric enables any Teradata Vantage system to access and query external databases, systems, and platforms without copying or moving of data. Options include transactional databases, other analytical systems, object stores, and Hadoop, including querying of columnar Parquet and CSV files and semistructured JSON files. QueryGrid can also push down processing to external systems to ensure adequate performance without data's being copied into Teradata Vantage.

- Robust automated tiered storage and workload management features optimize performance within admin-defined cost and performance guardrails. The system ably supports challenging mixed query loads with high user and query concurrency.

- Extensive library of 200-plus data science functions can be invoked through SQL, Python, or R (in Python notebooks or language-native tools, such as Teradata Studio and RStudio) for in-database processing.

Weaknesses:

- System deployment and management are not automated, so Teradata Vantage deployments (including cloud services) must be sized and scaled (and schemas and indexes developed and maintained) by administrators. Data tiering and workload optimization tasks are automated according to administrator-defined cost and performance rules.

- Total cost of ownership is lowest when the demands on the system are at their highest. Customers that don't have high scale (or that don't expect data to grow into the hundreds of terabytes), high

concurrency (numbers of users/queries), and challenging mixed workloads aren't likely to realize the full value of the platform.

- Cloud-based deployments are limited to 128 nodes at this writing. Teradata offers Vantage on VMware, but hybrid/multicloud container-based deployment options have yet to be developed for state-of-the-art deployment consistency, flexibility, scaling, and workload portability and isolation.

Overall assessment: Teradata remains a leader when it comes to scalability and performance for the most demanding use cases, with the combination of tight SLAs, high concurrency, and mixed workloads. Over the past five years, it has added multicloud deployment and separation of compute and storage decisions to the robust data tiering and workload management capabilities already in place. Automation is not a strong suit. Automated deployment sizing, scaling, and system and workload management are on the roadmap, so for now you'll need a skilled team to exploit the strengths of the platform. Constellation has spoken to customers that have tried more automated options but that have returned to Teradata Vantage, seeing it as the most cost-effective way to meet exacting requirements.

Vertica (a Product Group of Micro Focus)

Company: Micro Focus: public, founded 1976; Vertica: founded 2005, acquired by HP in 2011 and by Micro Focus in 2017
Featured product: Vertica Unified Analytics Platform, supporting Vertica in Enterprise Mode (analytic database software) and Vertica in Eon Mode (analytic database and lake query engine)
Related products/services:
Vertica Accelerator (analytic database/lake query engine as a service)
Vertica SQL for Data Lakes (on-premises lake query engine)
On-premises options: Software on X86 hardware, including Dell and HPE; Eon Mode supported on a variety of S3-compatible on-premises object stores
Vendor-developed Docker images available on Docker Hub
Public cloud options:
Marketplace offerings on Alibaba, AWS, Azure, Google Cloud
Vertica Accelerator (analytic database/lake query engine as a service)

Number of customers: More than 2,000
Large deployment example (vendor-supplied): TradeDesk runs two separate 15-petabyte 640-node Vertica instances on AWS (with plans to move one instance to Azure). The deployments ingest 1 terabyte of data per day and process 40,000 automated reports per day while also supporting ad hoc analysis by more than 300 business analysts and engineers.

Overview: Vertica was cofounded in 2005 by database innovator Michael Stonebraker. The database, introduced in 2007, was designed to manage vast data volumes, scaling into the petabytes and harnessing MPP and columnar storage for fast query performance. Vertica has steadily evolved its product, first adapting it to work in conjunction with high-scale data lakes with the addition of bulk data loaders and support for direct querying of file formats such as ORC and Parquet in HDFS and later adapting it to work with object stores. More recently the focus has been on supporting data science workloads and cloud deployment.

Data science support has been extended and includes geospatial, time series, and ML analysis; flex tables for schema-on-read flexibility; and more extensive in-database support for ML algorithms and languages such as Python. Cloud-oriented enhancements include ready-to-deploy images on leading cloud marketplaces and prebuilt integrations with native cloud services such as object storage services.

Vertica in Eon Mode, introduced in 2018, separates compute and storage decisions for cost savings and improved elasticity in public or private cloud deployments. The Vertica Accelerator as-a-service offering was previewed on AWS in June 2021 and became generally available in late September. An Azure-based Accelerator service is expected in 2022.

Licensing/subscription terms: Vertica is available with perpetual or subscription-based licensing on per-node or per-terabyte terms. Subscriptions are time-based, starting with a by-the-hour option (on AWS and GC) and extending (on all clouds) to monthly, one-year, two-year, or three-year options, with discounts for longer terms. Vertica Accelerator subscriptions are time-based per node.

Constellation's Analysis

Strengths:

- Vertica has no shortage of petabyte-league customers relying on the database's MPP and columnar architecture for massive scalability and data compression and workload management for fast query performance.

- Vertica in Eon Mode on-premises deployments and Vertica Accelerator service separate compute and storage, eliminating copies of data; reducing ETL and data-movement requirements; and adding options for cost-saving elastic scaling, workload isolation, and dynamic workload management.

- Analytical capabilities span advanced SQL functions; multidimensional analysis; and model training, evaluation, and scoring via in-database predictive and ML functions. Analyses include outlier detection; linear and logistic regression; k-means; Naive Bayes; and random forest analysis supported by prebuilt SQL extensions implemented in C++, Java, R, and Python. Predictive Model Markup Language (PMML) import and export extend data science capabilities to frameworks such as TensorFlow.

Weaknesses:

- Optimization and recommendation features assist DBAs, but automation is focused on optimizing known workloads according to user-defined rules and cost and performance thresholds. Automated deployment sizing, scaling, and system and workload management are on the roadmap.

- Customer-managed Vertica in Eon Mode deployments rely on third-party storage. Storage and caching plans must be carefully set to ensure performance and compliance with demanding service-level requirements. The Vertica-managed Accelerator service handles these challenges behind the scenes with autoscaling, autoscheduling, and autoshutdown options, but scaling is currently capped at 48 nodes (using up to I3 16XL AWS EC2 compute instances).

- Vertica offers Docker images on Docker Hub, but it's up to customers to manage and monitor deployments independently. Deployment, monitoring, and management tools spanning hybrid and multicloud deployments are on the roadmap.

Overall assessment: Vertica should be considered by organizations dealing with (or expecting to grow into) high-scale deployments ranging into the hundreds of terabytes and beyond and preferring to tightly manage, optimize, and govern their workloads without the aid of automation. The Vertica Accelerator service eases the burden of administration with autoscaling and autoshutdown options. With Eon Mode, Vertica supports hybrid and multicloud deployments and works in combination with object-store- and Hadoop-based data lakes. Vertica addresses diverse data science requirements via built-in database functions and PMML import and export capabilities supporting in-database execution of models developed in R, Java, Python, and myriad frameworks such as TensorFlow.

Yellowbrick

Company: Private, founded 2014
Featured product: Yellowbrick Data Warehouse (analytic database platform)
On-premises option: Yellowbrick Data Warehouse (customer-managed MPP database appliance)
Public cloud option: Yellowbrick Cloud Data Warehouse (hosted service with high-bandwidth connections to customer public cloud instances)
Number of customers: Not available
Large deployment examples (vendor-supplied): Large, unnamed financial services customer is said to have a multipetabyte-scale deployment supporting as many as 500 concurrent users (4,000 users overall) handling mixed query loads; unnamed insurance customer is said to have a multipetabyte deployment supporting 500,000 queries per day.

Overview: The Yellowbrick Data Warehouse became generally available in 2017, promising extreme analytical performance via MPP scaling and novel exploitation of state-of-the-art hardware capabilities such as NVMe and flash storage combined with hybrid row/column query execution and operating-system-level bypass of RAM directly to L3 cache. Yellowbrick says its combination of technologies delivers in-memory performance at multipetabyte scale as well as the ability to support real-time scenarios with as many as three million inserts per second.

Today's performance is achieved via a 6U appliance that is deployed and managed by customers on-premises. Alternatively, Yellowbrick Cloud Data Warehouse is a vendor-managed service (hosted from Equinix and Digital Realty data centers) that is connected to customer cloud instances in AWS, Azure, or Google Cloud via low-latency private links. The company is well along in developing a software-only offering that's available today as a single-node test drive on AWS. A multinode Kubernetes-based MPP offering is on the roadmap (on AWS first, followed by other clouds), but it won't be generally available until 2022.

Subscription terms: Whether deployed on-premises as a hardware/software offering or consumed as a service on Yellowbrick Cloud Data Warehouse, it's a subscription service that starts at $10,000 per month for up to 10 terabytes of data under management.

Constellation's Analysis

Strengths:

- Yellowbrick's software strategy is based on Kubernetes-based deployment from single nodes on edge devices up to multipetabyte-scale deployments on-premises or in private or public clouds. Yellowbrick provides the hybrid and multicloud control plane (with Yellowbrick Manager console for monitoring and management), and data remains in customer data centers or public cloud accounts.

- The latest-generation Andromeda architecture adds field-programmable gate arrays (FPGAs) to the mix of performance-enhancing technologies that will continue to be available for on-premises deployments, as required to meet data residency and other governance demands.

- Yellowbrick offers strong support for low-latency and streaming use cases such as fraud and risk analytics and aggregation and filtering of data from edge environments to centralized warehouse instances.

Weaknesses:

- Software-only Kubernetes-centric offering is still in development and won't emerge until 2022 and beyond. New customers should investigate the preview (expected on AWS in Q4 2021) before committing to the Yellowbrick Cloud Data Warehouse hosted/private link model.

- Yellowbrick natively connects to (S3 and ADLS) object storage, but at this writing, it handles only structured data such as CSV files. Support for reading and loading Parquet, ORC, and JSON is on the roadmap.

- Customers can develop user-defined functions to go beyond SQL functionality, but Yellowbrick is focused mainly on traditional data warehousing and does not offer ML or data science capabilities built into or packaged with the database.

Overall assessment: Yellowbrick is in the middle of a 12-month-long journey to providing the sort of software-only public-cloud-deployable capabilities that most customers expect. It's not that the company's on-premises (appliance-based) offering is going away. Indeed, that specialized hardware/software combination is a point of deployment and performance differentiation for customers that still prefer on-premises systems. However, the circa-2021 Yellowbrick Cloud Data Warehouse offering is stuck with fixed compute-to-storage ratios that will go away once the software-only offering is available directly on public clouds in 2022 (beyond the single-node test drive now available on AWS). We like Yellowbrick's visionary commitment to Kubernetes-based deployment and hybrid/multicloud systems monitoring and management, with data always living in customer accounts. Yellowbrick is a vendor to watch in 2022 and beyond.

Other Vendors Not Included in This Report

Not included in this market overview but deserving mention are Exasol, a high-scale analytical DBMS vendor with its largest market presence in Europe, and lake/fabric query engine startups Starburst Data and Ahana, founded in 2017 and 2020, respectively.

Differentiation

The vendor landscape above offers plenty of insight into how vendors are differentiated, starting with Figure 4, on page 16, which details what types of offerings each vendor provides and where and how they can be deployed, covering on-premises and public cloud deployment options.

As is evident in Figure 4, large vendors, including IBM and Oracle and the hyperscale public cloud vendors AWS, Google Cloud, and Microsoft, address both DBMS/warehouse requirements and data lake needs (and they also have supporting data science and data management offerings not covered in this market overview).

Many database platforms can access and query structured data residing on object storage, but Constellation doesn't consider it a combined lake/warehouse platform unless that product manages the lake space, handling the governance and access control. The combined lake/warehouse vendors vary in capabilities, with Cloudera and Databricks offering deeper support for data engineering, data science, and analysis of semistructured and unstructured data. Incorta, SAP, and Snowflake are more focused on using their lakes for storing data at scale inexpensively on object storage and querying, transforming, and reusing that data primarily for SQL-centric analysis (although both Incorta and SAP do offer Apache Spark services and data science libraries). The Oracle Data Lakehouse on OCI was announced as a generally available service in October 2021, days before this report was published and, therefore, too late to be assessed in this market overview.

Look to the independent vendors for hybrid and multicloud deployment options. In contrast, the analytical data platform services available from the hyperscale public cloud providers are effectively tethered to their mother clouds. Yes, there are early on-premises deployment options from the hyperscalers (AWS Outposts, Azure Arc, Google Anthos), but Amazon RedShift, Azure Synapse, and Google BigQuery do not yet run on these options, nor are they on their respective public roadmaps to be delivered through these options.

AWS doesn't support multicloud and isn't talking about the option, but its two biggest competitors are gearing up to support deployments on rival clouds (and on-premises). Microsoft, for example, is moving

toward multicloud deployment on Arc, but, again, Synapse is not yet an option on Arc or on the Arc roadmap. Google BigQuery Omni is a multicloud option that Google runs on AWS and Azure, but it's not the same as the BigQuery database service running on Google Cloud. Rather, Omni supports remote querying of object stores on either AWS or Azure, with query results returned to the mother BigQuery instance running on Google Cloud.

When considering data warehousing on Databricks, Dremio, Microsoft Azure Synapse, or Yellowbrick, keep in mind that these solutions have been on the market for less than five years. Look for customer references that match the scale of your data, the sophistication of your queries, and the concurrency and service-level demands you are likely to face. We would offer the same advice to any customer considering services based on venerable analytical DBMSs that have been brought to the cloud in the last five years (such as Oracle Database, IBM Db2, SAP HANA, Teradata Vantage, and Vertica). Performance and infrastructure requirements in the cloud are invariably far different from what they are on-premises.

As for specific product recommendations, Constellation Research offers two ShortLists™ specific to the high-scale analytic data platforms market. As the name suggests, ShortLists are Constellation's recommendations on what to put on your short list for consideration.

Constellation ShortList: Hybrid-Cloud and Multicloud Analytical Database Management Systems (published February 2021)

- Oracle Database
- SAP HANA/SAP Data Warehouse Cloud
- Teradata Vantage
- Vertica

Constellation ShortList: Automated Cloud Data Warehouse Services
(published October 2021 in conjunction with this report)

- Google BigQuery
- Oracle Autonomous Database
- Snowflake

Constellation has not yet published a ShortList on combined lake/warehouse offerings, but we expect to do so in Q1 2022.

RECOMMENDATIONS

Any organization dealing with, or expecting to grow into, high-scale analytical workloads ranging from the tens of terabytes into the petabytes should consider these analytical data platforms. It's rare to see greenfield deployments at this scale, so these platforms are most often considered as upgrades or replacements for existing platforms that are failing to meet requirements due to:

- Performance constraints tied to growing data volumes and/or aging on-premises infrastructure
- Growing data-analysis requirements in public clouds not adequately addressed by on-premises platforms
- Increasingly sophisticated data science requirements not addressed by incumbent platforms
- Growing interest in and reliance on data lakes that are not well integrated with or supported by incumbent platforms

As noted in the "Selection Criteria" section on page 8, would-be buyers should begin their assessment with organizational and tech strategy considerations before considering specific vendor offerings. Organizational considerations include existing budgets; existing technology skills; incumbent technology

dependencies; and the desire (and executive and budgetary commitment) to innovate with data, new sources of data, and more advanced analytics and data science. Tech strategy considerations include cloud strategy, on-premises requirements, data lake and data science strategy, and BI and analytics ambitions.

It all starts with a clear understanding of where the organization is coming from and where it wants to go—and on which clouds and with what level of commitment to technology spending, skills building, and analytical innovation. With these understandings, you can look for products that meet known requirements, such as:

- Spanning hybrid-cloud and multicloud deployment requirements
- Addressing diverse analytical and data science requirements handling advanced SQL as well as ML and predictive analytics via built-in algorithms
- Operationalizing custom algorithms via AutoML features or user-defined functions (UDFs) supporting in-database execution of models developed in Python, R, other languages, or data science frameworks
- Unifying querying against lakes, including legacy Hadoop clusters and modern object-store-based data lakes

Based on conversations with dozens of organizations that have deployed high-scale analytical data platforms, Constellation offers the following cautions and suggested best practices:

- **Think big and long-term.** It's all too common for organizations to outgrow deployments within just a few years, through either unanticipated organic growth or business-changing acquisitions. Don't ignore history, but look beyond it to consider future possibilities and plan deployments that will stand the test of time and emerging requirements.
- **Look for deployment consistency and flexibility.** Does the analytical platform you are considering support on-premises deployment as well as cloud and/or multicloud deployment? What's the level

of consistency from deployment mode to deployment mode, and are unifying administrative, data access, and workload-management interfaces available? Are licenses or subscriptions portable, so you can leverage training and financial investments? Is there flexibility to mix and change deployment modes?

- **Be prepared for differences in on-premises and cloud performance.** Don't base cloud configurations and performance expectations on your on-premises experience. Plan for higher capacities to overcome the bandwidth, virtualization, and latency differences that are inevitable in deployment on any public cloud. Consider the guidance available from the vendor, including documentation, best practices, and the level of activity and topics discussed on customer forums and community pages.

- **Consider available skills and training resources.** Evaluate your existing talent, the availability of training, and the cost and availability of professionals experienced with the platforms you are considering. There are plenty of SQL-savvy data management professionals out there, but how many have experience deploying, managing, and/or working with the specific platforms you are considering? Take into account the size of each vendor's customer community and its level of activity.

- **Seek out reference customers.** Look for reference customers with similar data, data scales, analytical needs, and workload requirements. Talk to them at length about the strengths and weaknesses of the platform and supporting vendor you are considering. Do all of the above before mounting pilot projects with each short-listed vendor to test your own data and key workloads.

Doug Henschen

Vice President and Principal Analyst

Doug Henschen is a vice president and principal analyst at Constellation Research focusing on data driven decision-making. His Data to Decisions research examines how organizations employ data analysis to reimagine their business models and gain a deeper understanding of their customers. Data insights also figure into tech optimization and innovation in human-to-machine and machine-to-machine business processes in the manufacturing, retailing, and services industries.

Henschen's research acknowledges the fact that innovative applications of data analysis require a multidisciplinary approach, starting with information and orchestration technologies; continuing through business intelligence, data visualization, and analytics; and moving into NoSQL and big data analysis, third-party data enrichment, and decision-management technologies. Insight-driven business models and innovations are of interest to the entire C-suite.

Previously Henschen led analytics, big data, business intelligence, optimization, and smart applications research and news coverage at *InformationWeek*. His experiences include leadership in analytics, business intelligence, database, data warehousing, and decision-support research and analysis for *Intelligent Enterprise*. Further, Henschen led business process management and enterprise content management research and analysis at *Transform* magazine. At *DM News* he led the coverage of database marketing and digital marketing trends and news.

@DHenschen **constellationr.com/users/doug-henschen** **linkedin.com/in/doughenschen**

ABOUT CONSTELLATION RESEARCH

Constellation Research is an award-winning, Silicon Valley–based research and advisory firm that helps organizations navigate the challenges of digital disruption through business model transformation and the judicious application of disruptive technologies. Unlike the legacy analyst firms, Constellation Research is disrupting how research is accessed, what topics are covered, and how clients can partner with a research firm to achieve success. Over 350 clients have joined from an ecosystem of buyers, partners, solution providers, C-suite, boards of directors, and vendor clients. Our mission is to identify, validate, and share insights with our clients.

Organizational Highlights

- Named Institute of Industry Analyst Relations (IIAR) New Analyst Firm of the Year in 2011 and #1 Independent Analyst Firm for 2014 and 2015.
- Experienced research team with an average of 25 years of practitioner, management, and industry experience.
- Organizers of the Constellation Connected Enterprise—an innovation summit and best practices knowledge-sharing retreat for business leaders.
- Founders of Constellation Executive Network, a membership organization for digital leaders seeking to learn from market leaders and fast followers.

www.ConstellationR.com

@ConstellationR

info@ConstellationR.com

sales@ConstellationR.com

San Francisco Bay Area | Boston | Colorado Springs | Ft. Lauderdale | Los Angeles | New York Metro
Northern Virginia | Portland | Pune | Sacramento | San Diego | Sydney | Toronto | Washington, D.C.