

TREND REPORT

What to Consider When Choosing a Cloud-Centric Analytical Data Platform

Key Questions to Ask Before Choosing a Cloud-Ready Data
Warehouse or Combined Lake/Warehouse Platform



Doug Henschen

Principal Analyst and Founder

TABLE OF CONTENTS

Executive Summary..... 3

Market Description..... 4

Importance to Buyers..... 7

Recommendations.....14

Analyst Bio17

About Constellation Research.....18



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 report explores recent trends in analytical data platforms, including the move to cloud DBMS services and the emergence of lake query engines and combined lake/warehouse platforms. Most importantly, it identifies key organizational and technology strategy considerations every company should review before even considering product-specific buying criteria. Technology buyers should use this report as a starting point for an analytical data platform review leading to a short listing of candidates and final proof-of-concept projects.

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.

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.

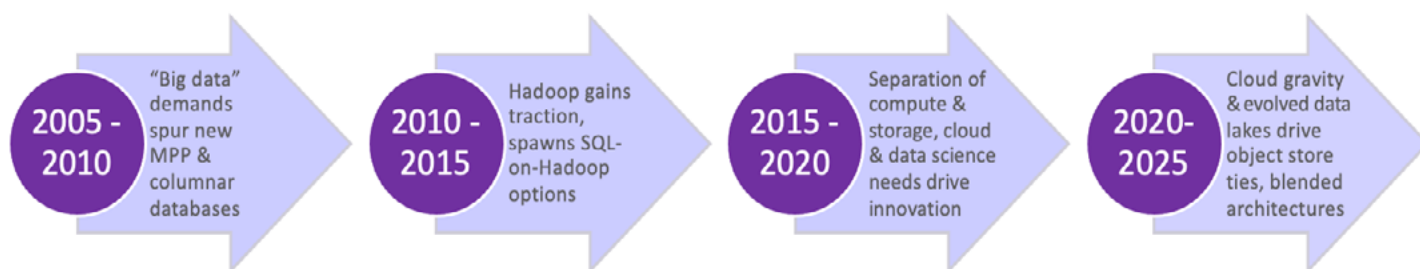
Starting in 2010 and gathering steam by 2012, organizations were captivated by the promise of low-cost storage in Hadoop. Deployments multiplied, but the complexity of this new platform limited access to data. Analytical DBMS technologies were soon adapted (and multiplied yet again) to support SQL-on-Hadoop 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. 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.

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



Source: Constellation Research

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.

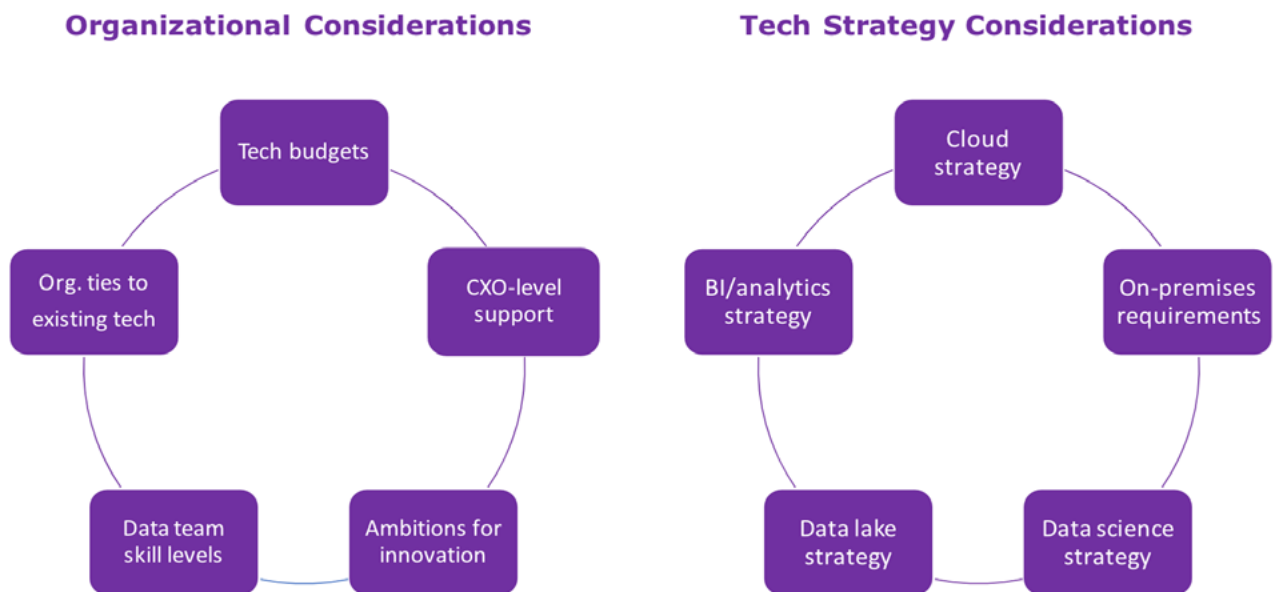
- Online and brick-and-mortar retailers analyze customer behavior data at scale to deliver targeted cross-sell and upsell offers and personalized services.
- Healthcare organizations analyze admissions trends, the efficacy of treatments, and internal policies and procedures to ensure better patient outcomes.
- 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

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



Source: Constellation Research

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.

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.

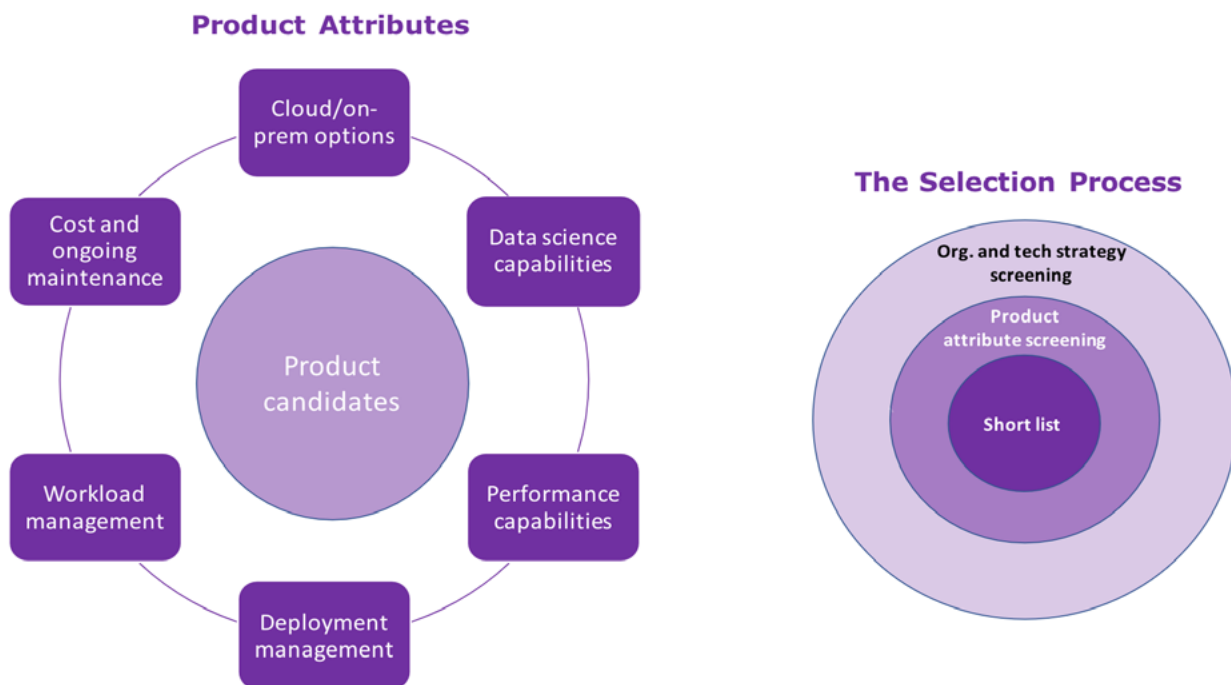
Once the context of the organization and its tech strategies are well understood, the team should be able to narrow things down to specific technology categories (meaning DBMSs versus data lake query platforms versus combined lake/warehouse offerings). And when it comes to the tech, an understanding 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

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



Source: Constellation Research

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?

- **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 and 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.

RECOMMENDATIONS

Any organization dealing with, or expecting to grow into, high-scale analytical workloads ranging from the tens of terabytes into the petabytes should look at the latest generation of analytical data platforms. (For product-specific insight, see Constellation's October 2021 Market Overview report "[What to Look for in Analytical Data Platforms for a Cloud-Centric World](#).") It's rare to see greenfield deployments at this scale, so new 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](https://twitter.com/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](https://twitter.com/ConstellationR)



info@ConstellationR.com



sales@ConstellationR.com

Unauthorized reproduction or distribution in whole or in part in any form, including photocopying, faxing, image scanning, emailing, digitization, or making available for electronic downloading is prohibited without written permission from Constellation Research Inc. Prior to photocopying, scanning, and digitizing items for internal or personal use, please contact Constellation Research Inc. All trade names, trademarks, or registered trademarks are trade names, trademarks, or registered trademarks of their respective owners.

Information contained in this publication has been compiled from sources believed to be reliable, but the accuracy of this information is not guaranteed. Constellation Research Inc. disclaims all warranties and conditions with regard to the content, express or implied, including warranties of merchantability and fitness for a particular purpose, nor assumes any legal liability for the accuracy, completeness, or usefulness of any information contained herein. Any reference to a commercial product, process, or service does not imply or constitute an endorsement of the same by Constellation Research Inc.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold or distributed with the understanding that Constellation Research Inc. is not engaged in rendering legal, accounting, or other professional services. If legal advice or other expert assistance is required, the services of a competent professional person should be sought. Constellation Research Inc. assumes no liability for how this information is used or applied nor makes any express warranties on outcomes. (Modified from the Declaration of Principles jointly adopted by the American Bar Association and a committee of publishers and associations.)

Your trust is important to us, and as such, we believe in being open and transparent about our financial relationships. With our clients' permission, we publish their names on our website.

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.