Hype Cycle for Data Management, 2024

8 July 2024 - ID G00808751 - 137 min read

By Aaron Rosenbaum

Learn how Gartner can help you succeed.

Become a Client

Accelerated cloud adoption continues to enable enormous innovation across data management technologies, with some new disruptors like generative AI now entering the mix. Data and analytics leaders can use this research to strategize their system and architecture roadmaps and investments.

Analysis

What You Need to Know

Data and analytics (D&A) leaders must cope with the evolving requirements of digital business, the deepening impact of cloud and increasingly capable data ecosystems. Today, data is rapidly moving to the cloud. This migration is creating opportunities for innovation and accelerating the pace of change across many aspects of data management.

To address these challenges, new entrants on the Hype Cycle reflect two key themes:

- The push to distribute responsibility and accountability for data management across lines of business and domains
- The impact of generative AI (GenAI) on data management

D&A leaders should use this Hype Cycle to identify promising technologies and decide when it is appropriate to evaluate them for adoption. Some technologies are less mature but offer significant innovation or differentiation. Others have recently matured and are near-ready for mainstream use.

Gartner provides several Hype Cycles that cover data management, analytics and related fields. Together, these Hype Cycles draw a holistic view of the emergent technologies and practices across the entirety of data and analytics:

- **Hype Cycle for Data and Analytics Programs and Practices, 2024**
- **Hype Cycle for Data and Analytics Governance, 2024**
- **Hype Cycle for Analytics and Business Intelligence, 2024**
- **Hype Cycle for Data Science and Machine Learning, 2024**
- **Hype Cycle for Artificial Intelligence, 2024**
- **Hype Cycle for Data Security, 2024**
- **Hype Cycle for Privacy, 2024**

The Hype Cycle

A Hype Cycle uses both current level of adoption and number of years to mainstream adoption to measure the maturity of technologies. With this in mind, we can see various levels of maturity (and various kinds of change) in this year's Hype Cycle. Here are some highlights.

Changes in the 2024 Hype Cycle

New entrant: The 2024 Hype Cycle for Data Management includes just one new entrant, open table format, which enters approaching the Peak of Inflated Expectations.

Maximum hype: At or just over the Peak of Inflated Expectations are some of the most hyped, albeit less mature, technologies in data management today. New entrants to maximum hype are vector databases and open table formats. Open table formats are entering the hype cycle at this placement while vector databases, while moving rapidly for now, may get subsumed by other technologies before the plateau.

Entering the Trough of Disillusionment: Future data architecture technologies and disciplines are currently heading into or passing through the trough: DataOps, data fabric, active metadata management, and augmented data cataloging/metadata management solutions. Lakehouse, which was at maximum hype last year, has rapidly achieved significant market penetration and is moving quickly through the hype cycle, moving to midway through the trough (despite any actual decline in market hype).

Plateau of Productivity: Both in-database analytics and event stream processing are nearing the plateau, and were also there last year. This is typical; it's unusual to move through the plateau in a single year. Both seem very likely to achieve the plateau in the near future.

Obsolete before plateau: This evolution to "obsolete before plateau" is frequent for database features and doesn't reflect on the utility of these models, but the evolution of the products in that category. This Hype Cycle has just one innovation that is obsolete before plateau: data mesh. While it continues to grow in expectations, we expect the core capabilities of data mesh to be subsumed by data products and data marketplaces and exchanges.

The Current State of Data Management Maturity

This year's Hype Cycle for Data Management has a wide distribution of technologies, reflecting an extremely high pace of innovation and change that has accompanied the migration to the cloud. Generative AI in the cloud makes it easier to innovate with very new technologies, and also accelerates adoption by enterprises and maturity by vendors, moving technologies through the Hype Cycle more quickly. Lakehouse and DataOps, for example, moved significantly since last year.

Fast Movers

Some of the areas within data management that are driving innovation and pushing the need to mature rapidly include:

- **Data ecosystems:** The whole environment for data to be used in analytics, business intelligence, AI and machine learning is changing. The desire for less complex management of disparate data sources has led to the concepts of data ecosystems and data fabric. These concepts are being realized in technologies such as lakehouse, augmented catalogs and DataOps, all of which are advancing to useful products very quickly. Of course, all are tied together with metadata management.
- **Generative AI:** While very early, the potential for transformational impact from GenAI is high. The benefits of GenAI, if realized, are significantly disruptive to many current practices within data management. The GenAI market is, however, very immature and extremely fast-moving, so any near-term deployment should be tempered by appropriate risk management.
- **Data management operations (DataOps and FinOps):** Due in part to the growth in cloud along with intercloud and edge computing, operations management for data management is changing. The need to manage and govern across disparate platforms is increasing the need for DataOps, data engineering, augmented FinOps and more.

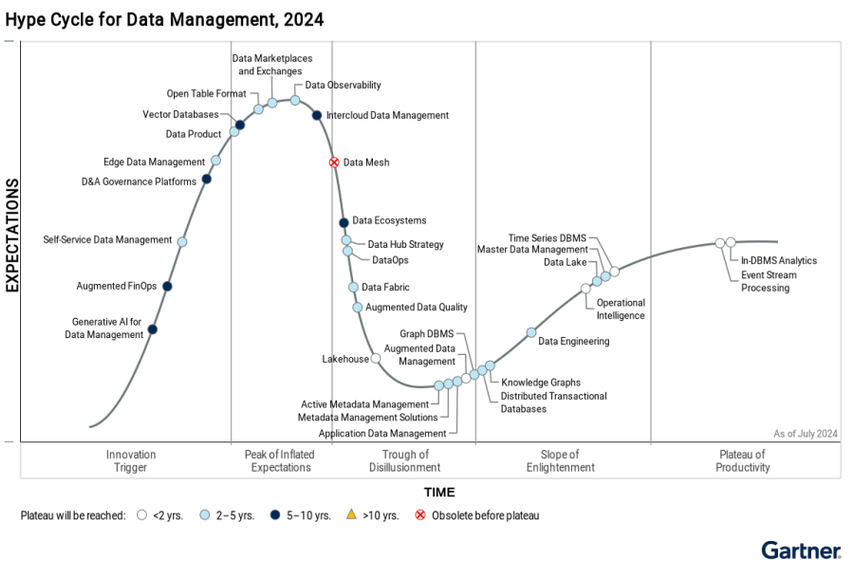
Stuck in the Trough

There is a cohort of technologies that is not moving smoothly through the Hype Cycle. At a point in time, it may seem that a technology is maturing based on both market adoption and technology maturity, unless it is poised to become obsolete. Some technologies move slowly through the Hype Cycle, some quickly. When a technology is seen as two to five years from the plateau, it's unusual to not move smoothly through the rest of the Hype Cycle. However, we have a group of technologies that we predicted in 2017, or even earlier, would reach the plateau in two to five years, which are now "stuck" either in or just past the Trough of Disillusionment. These technologies include:

- Graph DBMS
- Active metadata management systems
- Metadata management solutions
- Knowledge graphs

Had these technologies not had significant value, they'd already be obsolete — they all represent significant value for a wide array of enterprises and use cases. What all of these technologies have in common is a strong need for both solid technical tooling and significant implementation skills that span both technical and domain knowledge. But those successes have been driven by a combination of robust, deep platforms and project leaders who have both significant technical and domain knowledge. This is in stark contrast to the technologies currently moving most quickly to maturity — they are all making operations simpler and are often successfully deployed by enterprises of all sizes and levels of maturity.

Figure 1: Hype Cycle for Data Management, 2024



The Priority Matrix

The Priority Matrix for Data Management, 2024 positions each technology tracked in this Hype Cycle according to two dimensions — business benefit, and years to mainstream adoption and maturity. In the matrix, benefit is quantified as transformational, high, moderate or low.

Transformational benefit: The adoption of some hyped technologies could dramatically change how data management is performed, as well as transform organizations that depend on data management:

- Open table formats — the only new technology introduced to the Hype Cycle this year — and data products are both already having a very significant impact on how data management is performed in many organizations. Both seem well on a path to maturity in two to five years.
- Data ecosystems, data fabric and generative AI, while still five to 10 years from maturity, are having a very rapid incremental impact on data management. The move to cloud drives data ecosystem adoption and enables extremely rapid absorption of new capabilities into the existing (cloud) architectures.
- Data fabric, data ecosystems and generative AI for data management are three spokes of a flywheel that is gaining speed. Data fabric explicitly requires AI as well as the rich metadata most easily gathered in a data ecosystem; generative AI is accelerating data ecosystem adoption.
- Augmented FinOps, while lagging behind, is required to actually realize the benefits of the other technologies.

High benefit: Many data management technologies (the majority of which will be mainstream in five years or less) promise a high level of business benefit. A set moving rapidly through the Hype Cycle, expected to mature in less than two years — event stream processing, augmented data management, lakehouse, operational intelligence — is moving much more rapidly in the cloud than they did when requiring on-premises technology commitments. All were quite specialized/niche for very high-value use cases with skilled personnel, but in the cloud, their operations have become much simpler. Meanwhile, a set of high-benefit technologies that is potentially "stuck" in the trough and not moving as quickly may be subsumed into other technologies — data fabric in particular.

Moderate benefit: Most of the technologies in the moderate category have clearly established the value of their benefits already. But, for all the technologies in this category, it is very likely their capabilities get subsumed by technologies elsewhere on the Hype Cycle. Vector databases are one example. While vector embeddings are extremely important for generative AI, they are so important that every data management technology is implementing them, reducing the impact and benefit of stand-alone vector databases.

Low benefit: Because of the overall criticality of data management, we consider all data management technologies to deliver at least moderate benefit.

Table 1: Priority Matrix for Data Management, 2024

Enlarge Table

Benefit	Years to Mainstream Adoption			
	Less Than 2 Years	2 - 5 Years	5 - 10 Years	More Than 10 Years
Transformational		Data Fabric Data Product Open Table Format	Augmented FinOps Data Ecosystems Generative AI for Data Management	
High	Augmented Data Management Event Stream Processing In-DBMS Analytics Lakehouse Operational Intelligence	Active Metadata Management Augmented Data Quality Data Engineering Data Hub Strategy Data Marketplaces and Exchanges Data Observability DataOps Graph DBMS Knowledge Graphs	D&A Governance Platforms Intercloud Data Management Vector Databases	

Source: Gartner (July 2024)

Off the Hype Cycle

In 2024, a few innovations have changed or left the Hype Cycle for Data Management:

- **Fully mature innovations:** SQL interfaces to object stores and multimodel databases attained maturity on the Plateau of Productivity, and they have now moved off this Hype Cycle into mainstream adoption.
- **Dropped from the Hype Cycle:** Augmented data catalog/metadata management has been dropped, as its technologies have been spread across many other categories. Intercloud data management, as a stand-alone technology, has also been dropped.
- **Name changes:** "Augmented data cataloging and MMS" is now called "augmented data catalog/metadata management." "Cloud data ecosystems" has been renamed "data ecosystems." "Augmented transactions" has been renamed "operational intelligence."

On the Rise

Generative AI for Data Management

Analysis By: Roxane Edjlali, Ehtisham Zaidi, Rita Sallam, Mark Beyer

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Generative AI (GenAI) technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. GenAI applied to data management will have a profound impact: it will enable data specialists and consumers to manage data using natural language interfaces and automate manual tasks. It will reduce the skills barrier to data management and increase the productivity of data management specialists.

Why This Is Important

Although leveraging AI and machine learning (ML) in augmented data management is not new, using GenAI may transform the discipline, making data management activities accessible to much larger audiences. Through integration with active metadata management tools and enrichment from semantic tools and knowledge graphs using retrieval-augmented generation (RAG) or by training large language models (LLMs) on enterprise- or industry-specific corpora, it will become possible to have conversational interfaces to perform data management tasks.

Business Impact

The 2024 Gartner CDAO Survey identifies GenAI as one of the technologies that gathers the greatest level of interest with 23% of organizations having deployed it so far. Interest in applying this technology to enterprise data is high. Using conversational interfaces will reduce the pressure on highly skilled roles, such as data engineers, and ease the job of data stewards and even database administrators (DBAs). The expectation is that natural language becomes the data language. This will result in a transformation of the organizational models, roles and data management practice.

Drivers

GenAI technology embedded in data management solutions will:

- Allow organizations to innovate and implement data and analytics use cases faster.
- Open data management activities to new roles by reducing the skills barrier.
- Offer human-centric interfaces and self-service interaction with data.
- Increase productivity of data management specialists.
- Generate and correct code supporting data engineering, software development and DBAs.
- Support metadata discovery and management, semantic modeling and ontology mappings.
- Support augmented data operations and optimization using natural language interfaces.
- Offer data stewards the option to express data quality rules in natural language and have the tool generate corresponding scripts or pipelines automatically.
- Enable augmented capabilities for testing, validation, error detection, schema drift detection and correction, personally identifiable information (PII) detection and masking, and impact analysis reporting. These are facilitated through interfaces that make it easier for subject matter experts to become a part of the data management and engineering journey.

Obstacles

- GenAI techniques applied to data management are still in their infancy and will take a long time to gain both maturity and significant value.
- Like any other use of GenAI, when applied to data management, these solutions are subject to hallucinations. The queries or code generated can be erroneous, and hence the data discovery process could be inaccurate.
- When using GenAI with enterprise data, trust, governance and security are a concern.
- When connecting to enterprise data, additional risks related to security, privacy or even exposing metadata can become additional challenges.

- The ease of use and productivity these technologies promise do not fix the challenges in the data. The accuracy, quality and trust derived from these will require mature governance, data and metadata management.
- Adoption will require dedicated training, education and development of new skills such as prompt engineering.

User Recommendations

- Monitor adoption of GenAI as part of your data management tool portfolio.
- Evaluate solutions provided by your key data and analytics product suppliers.
- Consider GenAI capabilities across overall data management when selecting a cloud data ecosystem from a single cloud provider as part of that evaluation.
- Encourage your most curious team members (data engineers, data quality engineers, DBAs or data analysts) to experiment with GenAI and explore possibilities.
- Build skills for GenAI technique as part of your data management team, such as prompt engineering or RAG frameworks.
- Evaluate use cases that would benefit the most from GenAI and assess build vs. buy based on the uniqueness of the value.

Sample Vendors

Amazon Web Services; Databricks; Google; illumex; Informatica; Microsoft; SnapLogic

Gartner Recommended Reading

Innovation Insight: How Generative AI Is Transforming Data Management Solutions

Successful Generative AI Projects Require Better Metadata Management

The Impacts of Data Ecosystems: A Cloud Architectural Perspective

Innovation Insight: State of Data Management Support for Self-Service

Augmented FinOps

Analysis By: Adam Ronthal, Dennis Smith

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

FinOps applies the traditional DevOps concepts of agility, continuous integration and deployment, and feedback to financial governance, budgeting and cost optimization efforts; it is both technical and cultural. Augmented FinOps is the use of artificial intelligence (AI) and machine learning (ML) practices — predominantly in the cloud — to enable environments that automatically optimize cost based on defined objectives expressed in declarative terms.

Why This Is Important

In the cloud, assessing the cost of a specific workload or collection of workloads assigned to a project is possible. However, correlating cost to performance — a primary measure of cloud value — is difficult: many differently metered and priced resources in the cloud may be involved to deliver a unit of output. Augmented FinOps addresses this issue by using AI/ML to automatically optimize cost against performance, based on defined business objectives expressed in declarative terms.

Business Impact

Algorithmically driven cloud budget planning and financial operations will enable businesses to express their objectives — ideally in natural language or other declarative forms — and allow their cloud ecosystems to automatically optimize the underlying cloud resources to accomplish those objectives. This will result in efficient resource utilization and therefore optimal spending by reducing misaligned or poor use of cloud infrastructure and service offerings.

Drivers

- Cloud practitioners are increasingly realizing that rapidly changing cloud costs necessitate automated controls.
- Cloud adopters have a strong desire for transparency in cloud spending.
- Buyer inexperience is leading to either overprovisioning and spending more than needed, or underprovisioning and associated resource contention.
- Vendors are positioning cost-effectiveness as a competitive differentiator in their go-to-market strategies. However, achieving cost-effectiveness requires complex and continuous operations.
- Unpredictability of cloud consumption — and by extension spending — requires practitioners to seek the means of reducing this via automation.
- Business stakeholders are demanding that cloud leaders provide a way to correlate technical costs with business results.
- Solid financial governance frameworks are positioning organizations to take advantage of cloud financial management models, including FinOps.
- Emergence of specific roles — including FinOps practitioner or cloud economist — focusing on FinOps practices and cost optimization means organizations have the expertise to address augmented FinOps.
- The programmability and the granularity of cost-tracking data, coupled with the large amount of data that cloud environments produce, make augmented FinOps possible.
- Core FinOps capabilities are being delivered in three ways: homegrown solutions, cloud service provider (CSP) native tooling and third-party vendors. Practitioners often find that homegrown solutions require greater efforts to build and maintain, and thus pursue third-party or CSP tools to address their needs. As basic cost optimization techniques become standard features in various provider and third-party FinOps tools, vendors are adopting augmented capabilities as a means of competitive differentiation.
- The use of generative AI applied to FinOps is making it easier to engage with complex optimization problems, albeit at very early stages of maturity.

Obstacles

- CSP pricing models remain needlessly complex and diverse.
- Cloud ecosystems are (and will remain) open to third-party participants, which implies multiple commercial arrangements with multiple providers.
- Standards for cloud cost, usage and billing data, such as the FinOps Foundation's FOCUS proposal, have yet to be broadly adopted. APIs for communicating performance data within the context of a broader ecosystem have yet to emerge. Both of these are required to assess the primary measure of success: price/performance ratio.
- Full-scale augmented FinOps capabilities require transparency and accountability to build trust. This may take years to achieve and will require foundational tracking of recommendations and their results.
- Gaps remain in the intersection between CSPs and independent software vendors in delivery coordination, making it difficult to provide holistic FinOps solutions and SLAs that span the full scope of the cloud ecosystem.

User Recommendations

- Seek out service offerings to automate (via AI/ML) performance, consumption and pricing options. Increasingly, incorporate these capabilities into cloud data ecosystems that will learn from consumption patterns as they seek to optimize the underlying resources, and by extension, cloud spending through orchestration and optimization.
- Apply Gartner's FinOps Maturity Model (see The State of FinOps for Data and Analytics) to assess FinOps offerings in terms of their ability to observe, report, recommend, predict and optimize.
- Plan for using multiple tools to address the full scope of requirements. Many tools are broad in reach, but do not go deep into prescriptive recommendations. Others are tightly scoped and provide very targeted optimizations. Expect to spend time combining multiple tools to achieve broad and deep capabilities.

Sample Vendors

Acceldata; Anodot; Capital One; CloudBolt Software; Enteros; Finout; IBM; Sync Computing; Unravel Data

Gartner Recommended Reading

Control Cloud Costs With a FinOps Approach

Beyond FinOps: The Gartner Framework for Public Cloud Financial Management

The State of FinOps for Data and Analytics

Use Active Metadata to Measure Business Value for FinOps

Financial Governance for Successful Cloud Data and Analytics

Self-Service Data Management

Analysis By: Nina Showell, Ehtisham Zaidi, Michele Launi, Roxane Edjlali

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Self-service data management is a technology-led discipline that provides the ability for nontechnical users, typically working in lines of business, to complete data management tasks with minimal IT support. These activities include data preparation and integration, data cataloging, data quality and DBMS administration and management.

Why This Is Important

Because of increased demand for data throughout lines of business domains, self-service is no longer optional; it is a necessity. Most organizations experience bottlenecks in their data management processes, especially due to overreliance on their IT department's data engineering or a centralized data and analytics team, which struggles to keep pace. Increases in data literacy, generative AI (GenAI) and augmented capabilities are reducing the skills barrier, expanding both the demand and support for self-service.

Business Impact

The 2023 Gartner CDAO Survey ranks talent shortages as a major challenge. In parallel, two out of three teams are boosting their investments in data management. Self-service data management reduces the skills barrier, powered by augmented data management, including GenAI. Lowering the skills barrier allows less-skilled business teams to efficiently use data on their own, with reduced IT support. This frees up IT personnel to work on other, more strategic data management tasks.

Drivers

- The primary driver of self-service data management is a growing need for data within lines of business. Combined with an organizational desire to decentralize or federate data management functions, new approaches such as data mesh encourage domain-led data management activities.
- Previously, organizations had difficulty supporting self-service data management, usually due to the complex nature of data management tasks and low levels of data literacy. Within the past year, substantial improvements in GenAI have lessened this skills gap, demonstrating that self-service data management is realistically possible.
- Improvements to augmented data management (including GenAI) will make even the most technical data management tasks more accessible, including activities like data integration and DBMS administration.
- Data management is beginning to focus on the data life cycle as opposed to highly segmented technical processes. This allows organizations to support business-led and self-service activities at appropriate points in time, including tasks that happen upstream in the data life cycle.
- Cloud data management services and cloud data warehouses have made it possible for lines of businesses to become more autonomous. One example from data integration is using an extract, load and transform (ELT) integration pattern as opposed to extract, transform and load (ETL). The shift to ELT creates flexibility and allows more input from business-users into how the data is transformed, which promotes efficiency.
- Adoption of data literacy programs across enterprises over the last few years is creating both greater demand for data and analytics and a greater base level of competence across the business. Combined with augmented and AI capabilities, the skills gap will continue to narrow and vendors will promote "under-the-hood" enhancements that simplify their products to support broad use by nontechnical users.

Obstacles

- Concerns about control are the largest barrier. IT departments often want to maintain control, and lines of business may be too protective of "their" data, which impedes collaboration.
- Data governance is also an obstacle. Organizations need to balance centralized versus decentralized governance, which may require a shift in operating models. Decentralization can increase data silos, making sharing and reuse a problem.
- Self-service introduces new risks related to data governance, quality, security and privacy. This results in a false sense of confidence, allowing mistakes to compound.
- Work may begin as self-service (such as data prep). As it moves through the data life cycle, IT departments must support the ability to promote it to production (that is, operationalization through metadata access, governance and automation).
- Even with improved ease of use, data literacy and GenAI, the distribution of skills and technical maturity across disciplines or business domains continues to be a challenge.

User Recommendations

- Set up clear policies about who has access to which data, how often and the context in which data can be reused as a way to proactively mitigate risks related to security, privacy and data loss.
- Rethink organizational models by balancing a centralized and decentralized approach. Use adaptive data governance and focus on outcomes and agility as opposed to control.
- Make ease of use a priority since it will benefit both technical and less-technical users. Don't select a tool because it looks pretty. Select a tool that works well in practice. You should also evaluate existing tools; if needed, provide more training.
- To address business-driven pain points, involve users at earlier points in the data life cycle. Maintain open lines of communication between central IT teams and lines of business.

- Develop and communicate clear change management practices for promoting self-service processes to production.
- Don't overlook data literacy. A data-driven culture is required to avoid pitfalls.

Sample Vendors

Adeptia; Alteryx; Atlan; Collibra; Microsoft; Precisely; SnapLogic

Gartner Recommended Reading

Innovation Insight: State of Data Management Support for Self-Service

Innovation Insight: How Generative AI Is Transforming Data Management Solutions

6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh

Successful Generative AI Projects Require Better Metadata Management

D&A Governance Platforms

Analysis By: Guido De Simoni, Anurag Raj

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

A data and analytics (D&A) governance platform represents a set of integrated technology capabilities that help govern and steward a range of policies spanning security, quality, access, retention, privacy and ethics. It exposes a user experience for policy setting and enforcement to all relevant participants (such as data stewards in business roles, business analysts, line of business users, data scientists and governance board members).

Why This Is Important

The most complex governance challenges can no longer be met with siloed approaches. Today, the execution of D&A governance is inconsistent, with different organizations using different types of technology. These disparities increase costs and also impede the success of digital business initiatives. Governance needs have grown more diverse and complex at the same time, due to growth in things like sovereign data strategies. All aspects of governance for all types of policies can benefit from cohesive technology support.

Business Impact

- Enhanced productivity and efficiency in governance processes, more rigor in enforcement of policies, and therefore more control and trust in data and analytics.
- Ability to mitigate risk from most complex, cross-organizational governance challenges.
- Emerging augmented data management capabilities that discover data and its relationships to seed and power various governance work efforts.
- Converging long-term, discrete markets that will collide into one.
- Mitigation and eventual retirement of significant IT debt bound up in solid implementations.

Drivers

- Organizations want to have automated, synchronized, integrated, cost-effective and efficient D&A governance solutions with a central design, yet a distributed deployment. This requirement is driven by the growing recognition that the work of data and analytics governance is different from the work of data management, but that augmented data management supports the growth of these platforms of convergence (also potentially enhanced by generative AI).
- Increasing complexity from sovereign data and digital strategies is forcing organizations to simplify and coordinate governance efforts globally across privacy, security, storage, access, use and sharing. The recent EU AI Act is a good case in point.
- All of these aspects are operationalized, and more efficiency is gained when identification of data sources, curation of data, application of workflow, harmonization, reporting and visualization are provided in a coherent platform with automation. For example, you can address autogeneration of data quality rules using a number of methods. These include rule definitions and automated execution of data quality checks, AI-assisted data curation and association of business terms to technical artifacts, automated classification of sensitive data, and build subject registry.

Obstacles

- D&A governance today is served by discrete markets, each with its own solution, knowledge and IT debt. Inertia and sunk costs will slow down the emergence of this newer market.
- The current convergence within data management may not satisfy the needs of organizations across D&A governance.
- Incompatibility between what vendors can support and what different customer environments require will likely necessitate multiple metadata management solutions.
- Data management executes the policy that D&A governance sets. The work — policy setting, enforcement and execution — is different, so the technology capacities, roles and value propositions of the platforms are different.
- Other obstacles reside in the cultural shift that many organizations must address in utilizing the inherent value of D&A governance. When organizations are committing to data and analytics initiatives aligned to mission-critical priorities, such obstacles can jeopardize the adoption of these platforms as enablers for continuous improvement. We estimate that this innovation will reach the Plateau of Productivity in more than 10 years, even if the market is now adolescent and not just emerging.

User Recommendations

- Align use case and "work to do" for your specific D&A governance priorities, and engage the emerging D&A governance platform market from that launch pad.
- Design proofs of concept that will capitalize on the required critical technology capabilities. Identify the relevance of these technologies and their connection to business outcomes as a first step. Then, look into their ability to support specific use cases (such as risk management and compliance).
- Minimize the number of tools and solutions deployed by analyzing your strategic approach to D&A governance and by using available market technology capabilities in end-to-end scenarios supported by D&A governance platforms.

Sample Vendors

Alex Solutions; Collibra; data.world; Global Data Excellence; IBM; Informatica; OvalEdge

Gartner Recommended Reading

The Role of Technology in Data and Analytics Governance Policy Management

Market Guide for Data and Analytics Governance Platforms

Tool: Vendor Identification for Data and Analytics Governance Platforms

Edge Data Management

Analysis By: Aaron Rosenbaum

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge data management comprises the capabilities and practices required to capture, organize, store, integrate and govern data outside of traditional data center and public cloud environments. An increasing number of digital business use cases, including those based on IoT solutions, will leverage data in edge environments. This expansion creates tremendous opportunities to optimize resources and drive real-time decisions and actions, but also brings complexity and governance challenges.

Why This Is Important

Valuable data is increasingly generated and used outside of traditional data centers and cloud environments. This data often has a shorter useful life span, requiring value to be captured near the place and time of its origin. This is the role of edge-computing environments deployed closer to assets in the physical world. Edge data management will both impact and enable data and analytics leaders and their teams, requiring new capabilities and skills while also opening up new opportunities to deliver value.

Business Impact

Edge data management creates value in various ways:

- Dynamic factory operations, power usage efficiency, autonomous vehicles, and increased health and safety in mining and energy all rely on edge data management.
- More solutions, such as for Internet of Things (IoT) use cases, must operate in disconnected (or intermittently connected and low-bandwidth) scenarios.
- It enables smarter physical assets and collections of assets, including remote management or autonomous behavior, via edge data.
- It addresses inconsistencies, protection, sovereignty and other governance issues arising from siloed edge environments.

Drivers

- **Low-latency:** By placing data at the edge, data management capabilities and analytics workloads at optimal points, ranging all the way out to endpoint devices, enterprises can enable more real-time use cases. In addition, the flexibility to move data management workloads up and down the continuum from centralized data centers or the cloud to edge devices will enable greater optimization of resources.
- **Data gravity:** Bandwidth costs and scenarios with limited or intermittent connectivity demand the ability to organize and process data closer to the edge.
- **Expanded scale and reach:** By using distributed computing resources and spreading the load across the ecosystem, enterprises can broadly scale their capabilities and extend their impact into more areas of the business. These areas include use cases and outcomes traditionally managed only via operational technology (OT) teams, such as those managing equipment in industrial settings. Dedicated hardware for edge processing of data will continue to amplify these benefits.
- **Resiliency:** Pushing data management capabilities toward edge environments can bring benefits in the form of greater fault tolerance and autonomous behavior. If edge environments do not require centralized resources, then connectivity issues or unplanned downtime of those centralized resources don't disrupt processes that rely on local edge capabilities.

Obstacles

- **Management of distributed data architectures:** Data management has been largely based on principles of centralization — bringing data to central data stores (for example, data warehouses), and then processing that data to create value. Edge environments break that model via distributed data architectures, raising complex choices about where to locate and aggregate data on the continuum of cloud/data center to edge. Determining the right balance of latency and consistency is one such choice.
- **Governance and security:** With the distribution and complexity of edge environments, data governance and security become challenging. Organizations should extend their governance practices and policies to address edge-resident data storage and processing capabilities, including disposal of ephemeral or nonvalue event data.
- **Data access and integration:** The opportunities — and constraints — encountered when architecting very focused edge applications are often significant barriers to enabling outside access and utilization of that same data.
- **Broadband satellite internet service:** Broadband satellite internet using low Earth orbit now enables sufficient bandwidth to reduce the need for edge data storage and management for many applications.

User Recommendations

- Identify use cases where data management capabilities in edge environments can enable differentiated products and services by collaborating with OT and IT personnel working in edge locations.
- Identify data volumes and use cases that require edge processing as opposed to consolidation.
- Expand the skill sets of IT and OT teams to include edge platforms and the technologies required to manage data and data-intensive workloads on them.
- Place a greater emphasis on end-to-end system design. Understanding the dependencies between all components of distributed data pipelines, analytics workloads and AI models will be crucial to success.
- Ensure safety and control by extending existing governance capabilities to edge data environments.

Sample Vendors

Couchbase; DuckDB; EDB; InfluxData; Microsoft; MongoDB; ObjectBox; Oracle

Gartner Recommended Reading

Plan for Edge Computing in Data Management

Building an Edge Computing Strategy

Innovation Insight: How Enterprises Are Applying Edge Computing

Forecast: Internet of Things, Endpoints and Communications, Worldwide, 2021-2032, 1Q24 Update

At the Peak

Data Product

Analysis By: Ehtisham Zaidi, Robert Thanaraj, Lydia Ferguson

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

A data product is an integrated and self-contained combination of data, metadata, semantics and templates. It includes access and logic-certified implementation for tackling specific data and analytics (D&A) scenarios and reuse. A data product must be consumption-ready (trusted by consumers), up to date (by engineering teams) and approved for use (governed). Data products enable various D&A use cases, such as data sharing, data monetization, analytics and application integration.

Why This Is Important

In today's business environment, D&A leaders seek faster cycle times for integrated and governed data that is ready for consumption by other teams. However, data engineering teams still struggle with provisioning integrated data for mission-critical use cases within the agreed-upon SLAs. Data products are becoming important, because they provide business teams and subject matter experts (SMEs) with integrated and prepared data that is easily findable, trusted, self-contained and certified for reuse.

Business Impact

- Data products enable self-service, quicker response times and business accountability over data.
- Data products are cataloged, discoverable, preintegrated and targeted at specific use cases, enabling domain teams to focus on outcomes, rather than wait for IT to integrate, prepare and share data.
- Successful data products enable reuse. They not only serve as common inputs to internal analytics and data science projects, but also open up new revenue streams through data sharing and data monetization.

Drivers

- **Time to analytics (speed):** Organizations that operate with a product mindset for data (with associated operating models, roles and skills) are more likely to achieve faster turnaround times for their analytics use cases. Data products include preintegrated data that is approved for use (and reuse) by governance teams. Hence, this data is trusted by consumers, improving the cycle time to analytics delivery.
- **Reuse (scale):** Product management principles support iterative and agile best practices for data product creation, maintenance, operationalization and life cycle management. Thus, data products can be reused and scaled across use cases quickly when needed.
- **Trust:** Data products enable domain teams to get closer to data by providing prepackaged and preintegrated datasets with well-documented schema, semantics, ontology and usage guidance. This allows domain experts, analytics and SMEs to use the data products to support their business outcomes.
- **New revenue streams (monetization):** Monetizing data, by sharing it across data exchanges and marketplaces, requires the creation of data products.
- **DataOps support (agility):** One of the biggest challenges with data pipelines is that they break, and it takes a long time to find issues and fix them. Another challenge is that pipelines are developed through waterfall approaches, leading to painfully long wait times for the business. A critical premise behind data products is that they must support DataOps practices for agile data integration, continuous integration/continuous delivery, change management and issue resolution (such as schema drift).

Obstacles

- **Data product washing:** Every integrated dataset is not necessarily a product, and not every use case needs a data product. Data products must be created to support established, focused and repeatable business challenges, not one-offs. If created without scalability and reusability in mind, data products may become expensive to maintain and eventually abandoned.
- **Operating model shift:** Data products require a shift in D&A organizational structure and the introduction of product management principles. They need a formal data product manager role to oversee their life cycle.
- **Governance and value justification:** Organizations struggle to train SMEs on proper adaptive governance data products across domains. Moreover, teams struggle to reach consensus on KPIs, which are necessary to measure the success of data products.
- **Poor planning:** Organizations that create one-off data products and don't plan for management of the entire data product life cycle usually struggle with data products not delivering value.

User Recommendations

- Collaborate with domain teams to identify use cases that are firmly established and widely understood. These are ideal use cases where data products can be beneficial.
- Begin with business-critical use cases that need to be scaled. Then, correlate them with use cases that are impeded by central IT's delivery bottlenecks. These would be ideal starting points.
- Prioritize delivering a minimum viable product that teams can improve over time, rather than focusing on perfecting a data product for specialized delivery.
- Ensure that producing and consuming teams reach a consensus on KPIs for measuring data product success.
- Invest in the data product manager role. Besides being skilled in product management, data product managers must be adept at translating business requirements for data engineering teams.
- Train SMEs and analysts on data products usage. Ensure that data engineers are trained in DataOps tools and practices to enable agile maintenance of data products.
- Measure and communicate the success of data products to garner continued investment and support from business domains.

Sample Vendors

Agile Lab; IBM; Informatica; K2view; The Modern Data Company; Nexla; One Data; Revelate

Gartner Recommended Reading

Data and Analytics Essentials: How D&A Leaders Can Accelerate Data Mesh Adoption

Quick Answer: Do D&A Teams Need a New Delivery Model for Data Products?

6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh

Quick Answer: How Should CDAOs Approach D&A Product Management?

A Technical Professional's Guide to Governing Data Products

Vector Databases

Analysis By: Arun Chandrasekaran, Radu Miclaus

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Vector databases store numerical representations of data. In such databases, each point is represented by a vector with a fixed number of dimensions, which can be compared via mathematical operations, such as distance measures. Vector databases are commonly used in machine learning (ML) solutions, where vectors represent data features/attributes, such as text embeddings. Storing these vectors in a database enables users to search for similar data points at scale with low latency.

Why This Is Important

Vector databases serve use cases such as similarity search and product recommendation. Emerging approaches in generative AI (GenAI) have made vector databases often part of retrieval-augmented generation (RAG), a popular way of grounding AI with enterprise data. When customers adopt GenAI models, vector databases store the embeddings for fast retrieval. With the indexed and stored vector embeddings, the database can do a similarity search, which matches a prompt (the question) with specific or similar vector embedding.

Business Impact

Vector databases enable digital business by enhancing semantic search, personalizing recommendations and supporting AI applications with efficient management of unstructured data. They enable deeper insights, scalable data handling and better accuracy of AI, crucial for data-intensive operations. They can be valuable in leveraging data for competitive advantage, particularly in industries where data volume, complexity, and the need for precision in search and analysis are growing exponentially.

Drivers

- **Popularity of vector embeddings in GenAI for grounding large language models:** The surge in AI and ML applications across industries necessitates databases that can efficiently handle complex, high-dimensional data. With the rise of AI foundational models, embeddings have become the cornerstone for semantic search and RAG architectures. Hence, they are the working inputs for grounding large foundation models with enterprise data.
- **Advanced search and personalization needs:** As users demand more personalized experiences and relevant search results, enterprises need databases capable of semantic search and personalization at scale. Vector databases enable searching based on the meaning behind data, rather than just matching keywords, allowing businesses to deliver personalized content and product recommendations, thus enhancing user engagement and satisfaction.
- **Performance and scalability needs:** The applications looking to embed generative methods that use embeddings-based models need back-end services that can respond with low latency to high-concurrency requests (prompts) and responses (completions) for GenAI use cases.
- **Regulatory compliance and data security:** As regulatory requirements around data privacy and security tighten globally, enterprises must ensure their data management practices comply with laws. Vector databases can offer better control over data access, processing and storage, helping enterprises navigate the complex landscape of data privacy regulations more effectively.
- **Developer focus:** Developers of new applications are driving the demand for vector databases by presenting use cases that cannot scale without the ability for embeddings to be stored in an optimized structure for high-throughput production applications.

Obstacles

- Enterprises lack an understanding of what vector databases do and the unique use cases they enable. Moreover, know-how of data chunking and retrieval optimization is needed to ensure vector databases deliver value within GenAI applications.
- Vector databases are superspecialized databases that may cause challenges around data migration, integration and limited extensibility across use cases, although many large database vendors are now starting to offer vector support in existing databases.
- Most vector databases are delivered as cloud-managed services. The complexity of deploying, configuring and operating them outside cloud environments requires deep technical skills and know-how.
- Vector databases can be expensive to implement, given the newness of the technology and lack of industry skills on ongoing optimization.
- The vector database market is nascent and populated mostly by startups, which may not have extensive experience working with enterprise clients or have proven product-market fit. Also, it is uncertain whether this will be a separate market category in the long run.

User Recommendations

- Determine whether your functional requirements can be satisfied by incumbent vendors that can support the storage and retrieval of vector embeddings — you may not always need a purpose-built vector database.
- Prioritize developer experience, ecosystem integration, use-case fit, reliability and performance as important selection criteria, and validate them thoroughly via a proof-of-concept process.
- Select managed, cloud-based vector databases as deployment modes, unless you have stringent requirements and deep technical skills for an on-premises, self-managed deployment mode.
- Conduct internal training and education on the appropriate use cases for vector databases, how to leverage their true potential, and effective ways to optimize their deployment and maximize their value.

Sample Vendors

Amazon Web Services (AWS); DataStax; Google; KX; Microsoft; MongoDB; Pinecone Systems; Qdrant; Redis; Weaviate

Gartner Recommended Reading

Innovation Insight: Vector Databases

A CTO's Guide to the Generative AI Technology Landscape

How to Choose an Approach for Deploying Generative AI

Innovation Guide for Generative AI Models

Open Table Format

Analysis By: Aaron Rosenbaum

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Open table format (e.g., Apache Hudi/Apache Iceberg/Delta Lake) is a data abstraction layer and potential "game changer" in data and analytics (D&A) architectures and platforms. It is common to all lakehouse architectures. It creates logical table structures (in metadata form) around all data types (i.e., structured, semistructured and unstructured) and atomicity, consistency, isolation and durability (ACID) transaction; change data capture (CDC); and data versioning on low-cost object storage.

Why This Is Important

Multistructured data and metadata in scalable, low-cost object storage can have many use cases. Open table formats support ACID transactions, CDC and data versioning, extending into use cases involving streaming, analytic and AI/ML workloads. It supports the logical data warehouse pattern at lower cost and higher performance than was achievable using approaches such as external tables or virtualization. It lowers the cost of available data for new use cases versus storage-based cloning, external tables or export/import.

Business Impact

Open table formats decrease costs and increase value. With proper governance, many organizations experience higher efficiency in data engineering and operational costs through open table formats and lakehouse technologies. Higher efficiency drives decreased labor costs, or, more typically, decreased time to value, which accelerates gains from data insights. This increases value, making possible opportunities that couldn't be pursued at the previous pace of change and insight.

Drivers

- Many popular cloud analytics platforms, and all of the leading data lake/lakehouse platforms, use open table formats as their primary storage format. There is significant market adoption of these platforms, which is, in turn, driving the adoption of the open table format.
- There is significant dependence on open table formats through many of the tool chains used in the highest growth platform vendors in data management and AI. In particular, the three largest cloud service providers (CSPs) have made significant public and architectural commitments to open table formats throughout their analytics and data prep/governance tooling.
- Leading technology companies outside the enterprise vendor space support and develop open table formats. Organizations such as Apple, Netflix and LinkedIn are contributing to and furthering the development of the technology, driving standards independent of the vendors.
- The use of Parquet for data interchange has increased greatly, due to increased portability and efficiency versus other approaches. Open table formats are built on Parquet and are immediately compatible.
- Many enterprises are concerned with proprietary product lock-in. The use of open table formats is a popular mitigation strategy, because it supports multiple processing engines without migration of the data.
- Open table formats enable data to be used for AI/ML, data science and analytics without requiring separate silos of the data, potentially accelerating reuse.

Obstacles

- Advanced indexing and tuning for some analytic use cases can't be implemented in open table formats.
- The openness of these formats is relative to the implementation. Not all vendors interpret the standards the same way or keep up with the same versions. This results in a lack of interoperability across engines or a lack of functionality, which requires a "lowest common denominator."
- Tooling to manage and leverage an enterprise's data in world-addressable object storage is not in place. Many enterprises conduct governance as part of the process of moving data among silos. Removing this friction without proper governance creates risk and errors.
- Many organizations may be unable to govern data outside their DBMS-centric tooling. Adoption of new tooling (e.g., new or existing toolset versions) might be required prior to widespread adoption.
- Interoperability among different formats is often implemented at the query engine level. There are not good solutions for managing an estate with multiple data formats.

User Recommendations

- Use open table formats to leverage data across multiple tools and engines in a bounded ecosystem that is led (but not necessarily supplied entirely) by a single vendor. Typically, this will be the supplier of the technical data catalog in that environment (e.g., Databricks Unity Catalog or AWS Glue).
- Use data engineering, pipeline modernization and data science to achieve the most significant and immediate benefits and ROI from the adoption of open table formats.
- Test portability and vendor independence scenarios before they are needed. For example, if you want to use your cloud supplier's SQL engine, but use an open-source engine as a fallback, it's insufficient to rely on documentation or a single test. Ideally, you'll want to use the alternative platform continuously to detect and mitigate issues as soon as possible.

Sample Vendors

Amazon Web Services (AWS); Cloudera; Databricks; Dremio; Google; Microsoft; Starburst; Tabular

Gartner Recommended Reading

How to Leverage Lakehouse Design in Your Data Strategy

Comparative Assessment of Data Virtualization Solutions

6 Tips for Successful Logical Data Warehouse Projects

Data Marketplaces and Exchanges

Analysis By: Annette Zimmermann, Jim Hare, Robert Thanaraj

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Data marketplaces and exchanges (DMEs) provide infrastructure, transactional capabilities and services for consumers and providers of data assets. Marketplaces prioritize data monetization via one-time or recurring subscription transactions, while exchanges prioritize sharing. Internal data exchanges facilitate enterprise data sharing and remove silos to cross-organization data product provision and access.

Why This Is Important

AI's need for large, varied and specialized datasets with which to train models has increased the demand for greater convenience for data sharing, purchase and consumption. Although adoption remains in the early phases, DMEs provide liquidity to the data products space — enabling the sale, purchase or exchange of data products with relative ease. Participants are attracted to diverse third-party asset selection, simplified procurement, reduced operating/transaction costs and the opportunity to enrich existing data assets.

Business Impact

Easier access to third-party data assets via external DMEs benefits sellers through market liquidity and buyers through consolidated access and management of third-party data products. Organizations value internal data exchanges to simplify the sharing of data while supporting the rationalization of third-party data spend. Buyers of data products benefit through reduced provider exclusivity of specific data products from increased competition across marketplaces — driving more favorable pricing.

Drivers

- Data and analytics (D&A) products are a key enabler of digital business outcomes — increasing the demand for third-party data and reflecting the growing role of business ecosystems/AI marketplaces that may simplify the sharing or purchase of data.
- Vendors are introducing platforms supporting DME capabilities that reduce the burden on enterprises seeking to participate in or directly operate marketplaces and exchanges.

- Enterprises are seeking to rationalize their spend across third-party data projects that use internal data exchanges as a central point of managing and accessing external third-party data products across the organization.
- Vendors are adding marketplace capabilities to their ecosystem in adjacent markets (including data management, data integration and data governance solutions).
- The emergence of AI marketplaces offering a wide spectrum of AI products such as models and applications drives further demand for data marketplaces for model training.
- There is increased awareness across public and private sectors in terms of the value associated with internal and external data assets and products.
- Third-party data acquired from DMEs has become increasingly attractive as a part of feature engineering efforts for AI model features and accuracy lift.
- There is increased adoption of cloud service providers (CSPs), which has reduced on-premises data gravity limitations that slow the physical movement and integration of data across parties.
- The number of public and private data providers for DMEs continues to increase — providing both an increased level of specialization and breadth in terms of available third-party D&A product offerings.

Obstacles

- Data privacy legislation and unknown downstream data sharing risks impede the pursuit of monetizing and productizing specific data, reducing participation within DMEs.
- "Last mile" payment, integration and governance efforts for third-party D&A products acquired through DMEs are often underestimated by enterprises — leading to acquisition delays.
- Data ethics and sharing standards along with complexity in the management of data contracts and sharing agreements can prevent the adoption of third-party data.
- Enterprise procurement processes and CSP account structures provide friction for lines-of-business user spend within the CSP marketplaces and ecosystems.
- Terms and conditions associated with D&A products vary. This can complicate the use and derivative asset creation that often results from deep analytics.
- Funding and pricing of data products remain an obstacle — particularly for organizations that embrace data products and data monetization solely for secondary revenue streams.

User Recommendations

- Promote organizational participation in your DME of choice to accelerate the time to business value over the use of independent data asset providers or consumers.
- Adopt prebuilt DME platforms or utilize marketplace operators for the transactional infrastructure when seeking to monetize D&A products to build differentiated products and services.
- Leverage providers that operate within or are optimized for your cloud providers of choice to reduce data movement complexity and improve integration consistency.
- Include resources and support for integrating data from DMEs with established internal data as a part of your budgeting and planning process.
- Establish enterprisewide guidelines for the acquisition and management of third-party data products and analytics to rationalize and maintain budgetary oversight of third-party data spend.
- Explore the value of third-party data to increase analytic insights by either adding context as new attributes or through additional data science model features.

Sample Vendors

Amazon Web Services; Dawex Systems; Defined.ai; DemystData; Explorium; Harbr; Narrative I/O; Nomad Data; Revelate; Snowflake

Gartner Recommended Reading

Tech CEO Market Foundations for Data Marketplaces and Exchanges

Top Trends in Data and Analytics, 2023

Quick Answer: How Do I Get Started With Data Monetization?

6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh

Data Observability

Analysis By: Melody Chien

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Data observability (DO) enables organizations to understand the health of their data, data pipelines, data landscape and data infrastructure by continuously monitoring, alerting and troubleshooting issues to reduce and prevent data errors or system downtime. It provides detection and monitoring, validation, issue investigation, impact analysis and solution recommendation. It improves data reliability by enabling users to observe changes, discover unknowns and take appropriate actions.

Why This Is Important

Organizations are looking to ensure data quality across different stages of the data life cycle. However, traditional monitoring tools are insufficient to address unknown issues. DO technology learns what to monitor and provides insights into unforeseen exceptions. It fills the gap for organizations that need better visibility into data health and data pipelines across distributed landscapes well beyond traditional network and infrastructure monitoring.

Business Impact

Data observability:

- Calculates data quality metrics and identifies quality issues, such as missing data, anomalies and outliers, so businesses can resolve issues before they propagate.
- Provides insights into existing or potential data pipeline issues so the data team can quickly respond, thus improving productivity, reducing system downtime and avoiding critical business impacts.
- Provides a detailed view of data lineage, including where data comes from, where it is used and who is using it. This helps ensure proper use and governance of data to comply with regulations.

Drivers

- Data and analytics leaders face a growing number of mixed data stacks, diverse datasets, unexpected data drifts such as change in schema or business context, high demand for data quality, and near-zero tolerance of downtime. All these add to data management challenges. They need a holistic view of the state of data quality and data pipelines within interconnected systems.

- Data pipelines move data from point to point and deliver data to consumers. This journey can be disrupted by unexpected events such as data quality issues or a lack of infrastructure resources. The data flowing through these pipelines needs to be monitored for loss of quality, performance or efficiency. Organizations need to be able to identify points of failure before they propagate. DO automatically detects important events and analyzes various signals to troubleshoot the issues, and provides actionable insights for what to do next.
- Traditional monitoring tools require users to know what they're looking for. Thus, they are insufficient to address new issues that were not previously understood or detected. DO technology learns what to monitor and provides insights into unforeseen exceptions.
- DO provides a multidimensional view of data including performance, quality, usage and financial impacts to the downstream applications. Leveraging active metadata, lineage of data and AI/ML, DO generates real-time insight by monitoring the business context and analyzing data pattern, comparing history, and developing a semantic understanding of the data. It provides an end-to-end observability to better equip organizations to handle critical events and prevent business disruptions.
- DO capability is essential to facilitate GenAI adoption by providing observability over the data and making sure data is AI-ready and fit for purpose.

Obstacles

- There is no standard definition of what constitutes a DO tool. Vendors offer a range of capabilities often branded as DO, which is causing confusion in the market.
- The current vendor landscape is very fragmented in terms of coverage areas and data environments supported. Most vendors focus on monitoring data quality and data pipelines, and are less concerned about infrastructure and costs. Individual vendors seldom offer full end-to-end observation.
- Most DO tools only support the modern data stack. This limits their application in large enterprises with more complex data environments that in many cases use legacy data management tools.
- Most DO tools target the data engineer persona and are positioned as IT tools. Though business users receive critical insights from the tools, they may find them less user-friendly.
- Organizations are embracing the concept of "observability." However, the actual adoption of the tools is not straightforward. The consideration of connecting to the entire ecosystem and aligning to data governance strategy is still a concern.
- It is not clear who should own the tools and how to make them work with existing data infrastructure; the specifics of collaboration among stakeholders.

User Recommendations

- Identify which data elements or data pipelines require high standards in quality, uptime, latency and performance. Pinpoint the gap between current versus desired monitoring capabilities to support the requirements.
- Evaluate DO tools available in the market to enhance your observability based on priority of business requirements, primary users and interoperability with the enterprise data ecosystems.
- Pilot data observability program by building a monitoring mechanism as a starting point to increase visibility into data health. Invest in observability capabilities in a cloud environment first, as it's commonly supported by vendors and is faster and easier to demonstrate value.
- Include both business and IT perspectives when evaluating DO tools by engaging with both personas early on in the evaluation process and securing specific requirements from both sides.
- Partner with business stakeholders to evaluate and demonstrate business value of DO practices by tracking improvement in data quality, reduction in downtime and ability to meet SLAs to show tangible benefits.

Sample Vendors

Acceldata; Ataccama; Bigeye; Collibra; IBM; Informatica; Monte Carlo; Soda; Telmai; Unravel Data

Gartner Recommended Reading

Data and Analytics Essentials: Data Observability

Innovation Insight: Data Observability Enables Proactive Data Quality

Deliver Trust by Adopting Data Observability Practices

Magic Quadrant for Augmented Data Quality Solutions

Quick Answer: What Is Data Observability?

Intercloud Data Management

Analysis By: Adam Ronthal

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Intercloud data management is the process of actively managing data across multiple cloud providers as part of a cohesive application and data management strategy. It builds on the foundation of multicloud capabilities, but adds the ability to access and use data across clouds in an operational context. It can be enabled at the object storage layer, database management system (DBMS) or application tier.

Why This Is Important

The vast majority of organizations using the public cloud are storing data on more than one cloud. Today, most of that data remains siloed — accessed and managed in the context of a single cloud environment. As data's center of gravity shifts to the cloud (or multiple clouds), data and analytics leaders will seek out ways to unite that data in a logical and cohesive consumption tier to improve integration, optimization, efficiency, flexibility and insight.

Business Impact

The ability to access data — regardless of where it is located — is a potent capability that will serve to break down barriers to access for end users and applications. Intercloud data management is cloud-agnostic and will allow enterprises to access their data in any cloud at any time. It will enable globally distributed applications that span cloud providers and geographies, providing resilience and avoiding lock-in to any single cloud service provider.

Drivers

Though actual market penetration remains minimal, several drivers affect adoption, including:

- **Need for integrated data:** Distributed data must be integrated (in storage and/or logically) to achieve maximum business value, which can be challenging. Intercloud data management capabilities start to address this need on a global scale.
- **Regulatory requirements:** Some industries are starting to mandate the use of multiple clouds for resilience and availability reasons, and some countries have strict data sovereignty laws.

- **Global applications:** Applications that operate on a global basis may require the use of multiple clouds to meet latency and performance requirements.
- **Distributed teams:** Organizations with distributed teams may find value in working with data in more than one cloud and providing continuity in their multicloud environments.

Obstacles

- Enterprises seeking to adopt intercloud data management approaches tend to be large, global companies with specific application requirements; most small to midsize organizations do not have these requirements. Thus, market penetration is minimal.
- Intercloud data management requires technologies built for distributed data management. Some established cloud service providers — most notably Google Cloud Platform and Microsoft Azure — are entering the space with recent offerings. However, most of the leading data management technologies are not designed to manage data across multiple clouds.
- Intercloud data management requires management of traffic between the clouds. Multiple methods exist, each with different cost, performance and security considerations.
- Any intercloud data management application will be subject to the laws of physics. Practitioners will need to be aware of both performance implications and trade-offs in consistency and availability, and ensure that their applications are both aware of and designed with these trade-offs in mind.
- As with any data, intercloud data must be governed and integrated in storage and/or logically to achieve maximum business value.

User Recommendations

Weigh trade-offs in optimization and flexibility when making design decisions, as intercloud data management can occur at three different levels:

- **Object store:** Distribute data at the object storage layer and replicate between clouds when flexibility and diversity of choice are required. Data science teams, for example, can choose whatever optimization layer they would like for last-mile delivery, as all can read and write to the local cloud object store.
- **DBMS:** Select a DBMS that distributes and manages data in a geodistributed cluster when global operational efficiency and resilience are required. Nodes can reside in multiple clouds and/or on-premises. This approach can not only support local, low-latency read/write applications with global read capabilities, but also enforce data sovereignty requirements.
- **Application:** Use the application tier when data is already in multiple clouds and you require a means of bringing it together. This approach essentially defers data integration to the point of consumption.

Sample Vendors

Cirata; Cockroach Labs; Couchbase; Google; Microsoft; MongoDB; Oracle; PingCAP; Snowflake; Yugabyte

Gartner Recommended Reading

Distributed Cloud: Does the Hype Live Up to Reality?

What Are the Different Provider Approaches to Sovereign Cloud Demands?

How to Decide Between a Single-Cloud or Multicloud Strategy

What Are the Key Factors to Consider When Choosing a Cloud Data Management Architecture?

Magic Quadrant for Strategic Cloud Platform Services

Sliding into the Trough

Data Mesh

Analysis By: Michele Launi, Ehtisham Zaidi, Mark Beyer, Roxane Edjlali

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Data mesh is an emerging data management operating model. It supports a business-domain-led practice for defining, delivering, maintaining and governing data products. It aims to decentralize data delivery and support business-driven data and analytics initiatives. Data mesh is composed of four underlying principles: data products, domain ownership, self-service platform and federated computational governance.

Why This Is Important

Data mesh represents a potential complement to centralization-only data management strategies for analytics and other use cases. Organizations continuously seek a means to balance data requirements for repeatability, reusability, governance, authority, provenance and optimized delivery of data. Data mesh is an organizational approach that shifts responsibility, governance and management back to business stakeholders in their respective domains.

Business Impact

The goal of data mesh is to provide ready access to data through data products. Business representatives are assumed to have the greatest experience in capturing, preparing and using data within their domain of expertise. They are responsible for determining guidance and processes for creating, managing and governing data products, preventing unnecessary proliferation. From a governance and authority perspective, data mesh relies on a federated governance approach that aims to balance central and local authority.

Drivers

- Organizations question the success of data centralization, which can't meet all the planned use cases. Due to an increased need to manage data within lines of business, data mesh has emerged as a decentralized model for data management.
- Data mesh provides domains the flexibility they need to build data products that meet their required use cases. It gives domains more control over the use of those data products across the enterprise.
- Data mesh emerges as a compromise to address delivery issues, budget constraints, and misunderstandings between central teams and lines of business. Centralized approaches are often detached from the broader business domain requirements.
- Data mesh can reduce the time and effort required to enable data reuse throughout the enterprise by leveraging existing assets instead of centralizing the data architecture. It asserts remediation for flexibility, scalability and accountability issues in approaches like centralized data warehouses, data lakes and data hubs.

Obstacles

- Data mesh will be obsolete before reaching the Plateau of Productivity. We believe the underlying principles of data mesh on their own hold merit and will mature independently. Data mesh as the encompassing concept is experiencing slow depreciation as the market explores the approach.
- Maturity in data management and data governance within lines of business (LOBs) is required. The associated skills may not be present or insufficient in business-led teams.

- Inappropriate identification of either data details or correct integrity for combining them may cause data product proliferation, thus increasing management and maintenance and necessitating reengineering to reconcile different interpretations of the data.
- Data mesh implementations and practices do not follow any specific guidelines. They cannot be vetted against standardized, or even competing, approaches. Implementations vary and may incorporate multiple approaches (for example, marketplace experiences, virtualized views or subject-specific data marts).
- Improper split from data fabric implies choosing one approach over another instead of adopting both as complementary solutions.

User Recommendations

- Assess business domains readiness and partner with the respective stakeholders. Independent domains with broad utility across the enterprise may provide lower risk for initial efforts.
- Commit to building a distributed data management team, as the data mesh concept is highly dependent on the organizational model and the distribution of skills across central IT and LOBs.
- Assign a product owner from the business domain to each data product and adopt product management best practices across the whole product life cycle.
- Control data product proliferation by assessing their fit for purpose and measuring their consumption, ensuring that data products continuously evolve to meet changes in usage and scope.
- Start a metadata management program in parallel with data mesh, and collect passive metadata. This approach will allow data mesh to evolve toward metadata activation over time, guiding data product operationalization, value justification and greater transparency in the uses of data.

Gartner Recommended Reading

Quick Answer: What Is Data Mesh?

Data and Analytics Essentials: How D&A Leaders Can Accelerate Data Mesh Adoption

6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh

Data and Analytics Essentials: Data Fabric and Data Mesh

Data Ecosystems

Analysis By: Adam Ronthal, Robert Thanaraj, Aaron Rosenbaum

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Data ecosystems provide a cohesive data management environment that supports all data and analytics workloads. They have a common governance and metadata management framework, and unified access management. They integrate augmented data management capabilities with a set of services accessible by the business user. Third-party independent software vendors (ISVs) also participate in data ecosystems.

Why This Is Important

Data and analytics leaders report that the cloud experience today requires a significant integration effort to ensure that components work well together. Cloud service providers (CSPs) and ISVs are responding with more refined data ecosystems as the market moves from “some assembly required” to a “packaged platform experience.”

Business Impact

Data ecosystems are delivered both by a combination of native CSP offerings as well as ISV components that provide enhanced capabilities. They unify data management with streamlined delivery and easy integration within a holistic management framework. They address key data management disciplines such as data integration, data quality, data sharing, governance, metadata and observability via augmentation (including generative AI [GenAI]), and provide the basis for operational and analytics capabilities.

Drivers

- Data ecosystems address a strong demand for operational simplicity and ease of use.
- Data and analytics architectures are under significant stress due to pressure from CSPs to manage all data within a single cloud environment, and organizations commonly wanting (or needing) to use multiple and/or hybrid cloud environments.
- Cloud practitioners need to rationalize data silos, which span multiple deployment environments and frequently require different and potentially conflicting operating models.
- Enterprises are looking to unify the way they engage with different data persistence models, platforms and use cases to improve efficiency and time to value for data-driven initiatives.
- Built on a common foundation of governance, metadata and emerging data fabric design, data ecosystems enable new practices like DataOps, FinOps and PlatformOps. Data ecosystems will become self-optimizing and self-tuning, and support financial governance efforts through cost optimization.
- Data ecosystems promise improved productivity and ROI by nullifying the need for explicit data and application integration as they are based on a common set of services.
- Cohesive data ecosystems will allow implementers to leverage GenAI with greater efficiency and accuracy.

Obstacles

- End users may require additional integration effort when combining native CSP offerings with third-party ISV offerings. This undermines the core concept of a unified, holistic data ecosystem.
- While data ecosystems have a vision of unifying data management with common governance, security and metadata, significant work is still needed to make it a reality. The gaps existing in data integration, data quality, metadata and governance need to be addressed either through native CSP offerings or partnerships with ISVs.
- While CSPs are working with third-party ISVs to provide open ecosystems, their initial focus remains on ensuring their own components are working well within their own cloud and addressing the basic needs of their customers.
- For a cloud ecosystem to function well, CSP and ISV components must have standard interfaces that exchange metadata (including financial metadata) bidirectionally. Participating ISVs and CSPs agreeing on common metadata-sharing standards, while essential, has not been realized yet.

User Recommendations

- Assess the maturity of data ecosystems and the degree to which they deliver a unified environment.
- Define what CSPs need to deliver and what capabilities to obtain from third-party ISVs; expect spending more time on integration efforts when combining CSP and ISV offerings.

- Evaluate integration points between various components (data persistence, use cases, data integration, observability, governance and metadata capabilities) to determine how cohesive the resulting ecosystem is. A less cohesive ecosystem will require significantly more integration time and effort.
- Ensure that the data ecosystem has a well-articulated path to production for a full data life cycle (from discovery to production-optimized delivery).
- Evaluate data ecosystems based on their ability to resolve distributed data challenges, and access and integrate with data sources outside their immediate environment.
- Ensure organizational data maturity, skills and literacy are at a foundational level to best leverage advanced features like GenAI.

Sample Vendors

Amazon Web Services (AWS); Cloudera; Databricks; Google; Hewlett Packard Enterprise; IBM; Microsoft; Oracle; SAP

Gartner Recommended Reading

The Impacts of Data Ecosystems: A Cloud Architectural Perspective

Innovation Insight: Data Ecosystems Will Reshape the Data Management Market

Strategic Roadmap for Migrating Data Management Solutions to the Cloud

Data Hub Strategy

Analysis By: Andrew White, Thornton Craig

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

A data hub strategy effectively determines where, when and how data needs to be mediated, governed and shared in an enterprise. It layers data and analytics governance requirements atop sharing demands to establish the patterns for data flow. The strategy drives the implementation of one or more data hubs — architectures that enable data sharing and access by connecting data producers (applications, processes and teams) with data consumers (other applications, processes and teams).

Why This Is Important

While data management leaders are fixated on trying to mitigate siloed infrastructures with data mesh or data fabric, the need to integrate data and analytics (D&A) governance and data sharing policy with data management has only grown. A coordinated, connected approach is needed. A data hub strategy provides that connected approach to streamline and simplifies how all D&A governance and sharing policy is implemented across the entire information infrastructure, wherever it resides.

Business Impact

- Increased operational efficiency by aligning and integrating previously siloed D&A programs with governance elements such as data quality, master data management (MDM), application data management (ADM), metadata management, analytics governance, product management and data catalogs.
- Increased return on all D&A investments through more effective and targeted efforts on implementing governance of D&A information assets — such as data, analytics and models — across the entire mesh or fabric.
- Reduced complexity and cost and increased need for data sharing across overall information infrastructure and data fabric or mesh.

Drivers

- Increased demands for consistent and reliable sharing of critical data between the organizations and things that comprise the extended enterprise — for example, in support of Internet of Things solutions and new digital products.
- Desire for cost optimization and need to increase productivity of data management infrastructure.
- New and burgeoning demands to implement AI governance that should extend D&A governance efforts but with new tools, technology and organizations.
- Better collaboration across business-oriented (governance) and IT-centric (integration) roles concerned with delivering data to points of need across the enterprise.
- Longtime and continued frustration of business stakeholders over the lack of consistency and trust of data driving strategic business outcomes. A data hub strategy enables more-focused application of governance controls, as compared with trying to align governance approaches inside many endpoint systems.
- Growing need for a flexible and governable architecture that complements centralized data stores such as data lakes and data warehouses.
- Desire of many organizations to leverage the concepts and successes of MDM (sharing) and ADM (using in applications) programs toward governance and sharing of other types of critical data — includes coupling MDM and ADM across the enterprise.

Obstacles

- Inability to modernize D&A governance programs and shift away from legacy domain and data-centric or IT-focused efforts to an outcome-based program.
- Resistance from teams or business units that are ill-equipped or prefer to retain control over their choices regarding how data is shared and governed.
- Inability to enable collaboration and agreement of critical stakeholders on data sharing and governance requirements across boundaries in the enterprise.
- Overreliance on technology and viewing governance and sharing of data as purely an implementation issue instead of an operating model and human/talent issue.

User Recommendations

- Identify the data that is most frequently used or is most important with most business value, and that requires effective governance and sharing. This might be a lean MDM or ADM/ERP program.
- Design a data hub strategy to understand data and analytics governance and sharing requirements, and to drive integration efforts across multiple use cases.
- Include any master data, application data, reference data, analytics data hubs or other intermediaries (such as customer data platforms) in your overall data hub strategy.
- Iterate changes to your data hub strategy as requirements for governance, sharing and integration change.

Sample Vendors

IBM; Informatica; Profisee; Progress (MarkLogic)

Gartner Recommended Reading

Data Sharing Is a Business Necessity to Accelerate Digital Business

Use a Data Hub Strategy to Meet Your Data and Analytics Governance and Sharing Requirements

Data Hubs: Understanding the Types, Characteristics and Use Cases

Data Hubs, Data Lakes and Data Warehouses: How They Are Different and Why They Are Better Together

DataOps

Analysis By: Robert Thanaraj, Michael Simone, Sharat Menon

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

DataOps is an agile and collaborative data management practice focused on improving the communication, integration, automation, observability and operations of data flows between data managers, data consumers and their teams across the organization. The goal is to assist data and analytics (D&A) leaders in driving operational excellence in data delivery of business outcomes via optimized data management.

Why This Is Important

DataOps eliminates various inefficiencies and misalignments between data management and consumption use cases by optimizing data delivery processes and operationalizing data workloads. DataOps practices:

- Improve organizational speed and trust in delivering data
- Manage interdependencies across business processes
- Increase reusability of data engineering work
- Provide reliable data delivery service levels
- Enable data governance best practices leading to trust and confidence among consumers

Business Impact

- Data engineers benefit from increased productivity and robust change management, ensuring adherence to data delivery service levels such as quality, lineage and security.
- Data consumers benefit from reduced cycle time of accessing ready-to-use data and improved data trust.
- Organizations thrive on collaboration, data literacy, productivity gains, self-service enablement and a collaborative culture.
- DataOps eliminates wasted data delivery efforts by focusing on value flows tied directly to business impact.

Drivers

- Organizations strive to improve speed and efficiency of producing trusted and usable data. DataOps practices reveal bottlenecks in the current D&A delivery process and guide toward improving the lead time (process efficiencies) and cycle time (technical efficiencies).
- DataOps improves the shareability and reusability of the data across the organization. It involves formal processes around data architecture, quality and modeling, and ensures that the data governance requirements are being applied as part of the operational processes. Otherwise, the initial data pipeline would only be designed for narrow use cases.
- DataOps practices enable reorganization of teams, which helps overcome the challenges caused by fragmented teams and processes, and delays in delivering data in consumable forms.
- DataOps tools eliminate the various inefficiencies and misalignments across data management technologies by streamlining data delivery processes and operationalizing data workloads. DataOps is both practices and tools. And the tools market is emerging.

Obstacles

- Setting up DataOps is a challenge as it needs efforts to justify operating model shifts. Focus on metadata management practices and continuing to provide effort or cost over value justifications (like ROI), which require aligning business outcomes to DataOps activities.
- Most organizations have substantial domain expertise and siloed functional capabilities. It is challenging to retain the advantages of the legacy approaches to data management and the people holding those skills while also aggressively pursuing DataOps.
- Organizations lack a holistic view of various stand-alone technologies that are often managed by multiple teams with varying levels of operational maturity.
- Many strive for end-to-end automation of pipelines driven by code and integration of diverse technologies and approaches, which is complex as it involves diverse skill sets.

User Recommendations

- When introducing DataOps, target projects that are struggling due to lack of collaboration or overburdened by the pace of change, or where service tickets from data consumers are piling up.
- Track metrics such as time to deploy changes, degree of automation, developer productivity, code quality, failure rates in production, cost-efficiencies and business impact in dollar amounts.
- Plan for data pipeline operations ownership and service-level management. DataOps teams generally own the full development life cycle from inception to production. They must be cross-functional teams that combine software engineering, data engineering and analytics — some organizations even hire for new roles such as data product manager.
- Automating data pipelines, tracking the health of pipelines and preventing pipeline failure — all need metadata. A lot of DataOps do not happen unless D&A leaders set up and connect with a good metadata management practice.

Sample Vendors

Astronomer; BMC; DataKitchen; DataOps.live; IBM; iCEDQ; RightData; Unravel Data

Gartner Recommended Reading

Data and Analytics Essentials: DataOps

Market Guide for DataOps Tools

5 Ways to Enhance Your Data Engineering Practices

How to Apply DevOps and Value Stream Mapping to Data, Analytics and AI

Toolkit: Delivery Metrics for DataOps, Self-Service Analytics, ModelOps and MLOps

Data Fabric

Analysis By: Ehtisham Zaidi, Robert Thanaraj, Mark Beyer

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

A data fabric is a data management design for attaining flexible and reusable data pipelines, services and semantics. The fabric leverages metadata analysis, data integration and preparation, knowledge graphs and semantics, and ML over metadata. Data fabric aims to overturn the traditional approach to data management which is "build to suit" and replaces it with "observe and leverage."

Why This Is Important

Data fabric:

- Leverages traditional approaches while enabling the enterprise to adopt technology advances and avoids "rip and replace."
- Capitalizes on sunk costs and simultaneously provides prioritization and cost control guidance for new spending in data management.
- Leverages existing platforms/architectures or implementation approaches. It offers flexibility and extensibility in infrastructure for humans or machines to ensure that their integrated data is consumable across multiple use cases and across a multicloud or hybrid architecture and edge ecosystem.

Business Impact

Data fabric:

- Acts as an intelligent orchestration engine on top of distributed data stores to provide guidance on optimal data access, integration, and delivery without always recommending further data replication or rip-and-replace of existing systems.
- Provides insights to data engineers by automating repeatable integration tasks and supporting design and delivery guidance.
- Adds semantic knowledge for context and meaning.
- Reduces maintenance and support costs associated with managing data.

Drivers

- The dearth of new staffing or qualified personnel for data management roles and the attrition of experienced professionals leaving the practice area has increased the demand for more efficient data reuse.
- Demand for rapid comprehension of new data assets has risen sharply and continues to accelerate, regardless of the deployed structure and format.
- Increasingly, more data is being generated and accessed, integrated and analyzed at the edge to support Internet of Things and streaming use cases. Here, the data fabric is needed to manage the access to and integration of data across the edge, data center and a hybrid cloud environment to support IT/OT data management.
- Organizations have found that one or two approaches to data acquisition and integration are insufficient. Data fabrics provide capabilities to deliver integrated data through a broad range of combined data delivery styles including bulk/batch ETL, data virtualization, message queues, use of APIs, microservices and more. Data fabric design is a recognized use case for data integration tools. (See **Critical Capabilities for Data Integration Tools**.)
- Increased demand for data tracking, auditing, monitoring and reporting (especially given the rise of GenAI) in a business domain or for an organization as a whole.
- Data catalogs alone are insufficient in assisting with data self-service. Data fabrics capitalize on machine learning (ML) to provide recommendations for integration design and delivery, reducing the amount of manual labor that is required.
- Significant growth in demand and utilization of knowledge graphs of linked data, as well as ML algorithms, can be supported in a data fabric. This assists with graph data modeling capabilities and use case oriented semantics that support search and discovery.

Obstacles

- Vendors use differing design and semantic metadata access and sharing standards (i.e., closed metadata), which affects effective metadata analysis and sharing, pivotal to delivering and maintaining a data fabric.
- Active metadata management practices lag behind data fabric adoption but are critical to its implementation.
- Fabric needs analytics and ML capabilities to infer missing metadata. This will be error-prone at first, with staffing and resources assigned to competing demands in advanced analytics, data science and AI near the data consumption layer.
- Diverse skills and platforms demand a cultural and organizational change, ranging from data management based upon analysis, requirements and "design then build" to discovery, response and recommendation based upon "observability and leveraging."
- Organizations often apply budget or staff to one-off and point-to-point integration solutions without thinking of data management solutions as a broader design.

User Recommendations

- Build a metadata management practice where tools share metadata across the entire data pipeline. Use tools to remove duplicate metadata and share metadata bidirectionally with other tools.
- Invest in an augmented data catalog that permits multiple ontologies over top of business data taxonomies.
- Prioritize using knowledge graphs in targeted areas where adequate metadata management practices already exist.
- Ensure that business process experts can enrich knowledge graphs with business semantics.
- Use data integration platforms that take active metadata based insights from the data fabric to stitch data pipelines through a combination of all major data integration modes i.e., batch, streaming, virtual and replication.
- Utilize insights and automations from the fabric to support your orchestration demands and improve operational excellence through DataOps.
- Make the fabric the cornerstone of your data product creation, delivery, maintenance and sharing to enable the data mesh operating model.

Sample Vendors

Cambridge Semantics; Denodo; Google; IBM; Informatica; Microsoft; Nexla; Semantic Web Company; Stardog; Stratio Big Data

Gartner Recommended Reading

Data and Analytics Essentials: How to Define, Build and Operationalize a Data Fabric

Quick Answer: What Is Data Fabric Design?

Data and Analytics Essentials: Data Fabric and Data Mesh

Augmented Data Quality

Analysis By: Melody Chien

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Augmented data quality (ADQ) solutions provide capabilities aimed at improved insight discovery, next-best-action suggestions and process automation. They leverage artificial intelligence/machine learning (AI/ML) features, graph analysis and metadata analytics. Each technology can work both independently and cooperatively to create network effects, which can then be used to increase data quality automation and effectiveness across a wide range of data quality use cases.

Why This Is Important

Data quality is a top priority in data and analytics initiatives. Due to the rapid expansion of the contemporary data environment, the multitude of data types and rising generative AI adoption, organizations seek transformative techniques that are fast, affordable, scalable and easy to implement to tackle data quality issues. ADQ technologies revolutionize conventional and time-consuming manual procedures by increasing automation in the data quality process and enhancing insights.

Business Impact

Augmented data quality can:

- Automate repetitive processes to solve common data quality issues across the enterprise to reduce manual effort and improve efficiency.
- Perform in-depth and rapid data quality actions using AI/ML algorithms.
- Connect data semantically, track lineage and ownership, and create domain data mapping to quickly identify impacts and possible solutions with knowledge graphs.
- Empower users with natural language processing (NLP) enabled by large language models (LLMs) in data quality processes within their own business context, along with built-in workflow, knowledge bases and recommendations.

Drivers

- Data quality is essential for improving business efficiency, enhancing analytics for better insights, and enabling trusted, data-driven decisions.
- The rise of AI models creates a new demand for high-quality, AI-ready data. Data quality is a competitive advantage that data and analytics leaders must continuously pursue to achieve these goals.
- ADQ enables organizations to scale and unify data quality efforts for enterprisewide success, which is often a challenge, due to limited internal capabilities and strategies.
- Current data quality technologies and practices are often manual and insufficient for the speed and scale needed by generative AI (GenAI) and other use cases.
- Organizations need seamless integration, agile deployment and bidirectional exchange of intelligence with related data management functions. This is core to ADQ.
- ADQ tools are designed to share profiling and technical metadata analysis with other data management tools. This transparency helps users justify price-to-performance ratio and allows ML to adjust compute, thereby optimizing financial governance for data management tasks.

Obstacles

- **The coexistence of traditional data quality products and ADQ in the market** creates confusion for end users about the actual benefits of ADQ. Some vendors offer both, while others are still innovating their existing data quality products. This can lead to mixed messaging about ADQ's capabilities.
- **The lack of explainability and transparency of AI/ML models** can lead to user reluctance, especially regarding GenAI-embedded features. Vendors use different technologies (supervised or unsupervised algorithms) or different OpenAI sources. Drifts (such as data or schema drifts) may occur over time and affect AI models. Continuous improvement through active learning and human supervision is crucial.
- **The inclusion of data and analytics governance is crucial** when implementing ADQ tools. While AI-driven automation offers users independence, embedding governance requirements into the AI models is essential to mitigate data-related risks.

User Recommendations

- **Evaluate current data quality capabilities:** Assess manual effort and complexity needed to support use cases. Identify improvement areas ADQ can address. This will help determine ADQ adoption requirements.
- **Explore ADQ capabilities:** Investigate the features, setup process, required skills and constraints associated with ADQ solutions. Assess offerings from data quality vendors, identify any gaps and explore product roadmaps for potential enhancements.
- **Establish data cataloging:** Implement practices to collect and analyze metadata for automation and efficient data quality processes. Enhance data asset management and facilitate ADQ integration.
- **Align with data governance:** Partner with stakeholders to monitor ADQ solutions. Ensure adherence to governance requirements and frameworks. Establish metrics to show the benefits and business value of ADQ.
- **Collaborate with solution providers:** Provide feedback, share experiences and suggest enhancements. Engage in user forums to contribute to improvements and shape the development roadmap to meet your needs.

Sample Vendors

Ataccama; CluedIn; Collibra; Datactics; DQLabs; Experian; IBM; Informatica; Precisely; Qlik

Gartner Recommended Reading

Magic Quadrant for Augmented Data Quality Solutions

Critical Capabilities for Augmented Data Quality Solutions

Quick Answer: How to Boost GenAI Impact on Data Quality Initiatives

Data and Analytics Essentials: Data Quality

Innovation Insight: Data Observability Enables Proactive Data Quality

Lakehouse

Analysis By: Aaron Rosenbaum, Adam Ronthal, Roxane Edjlali

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

A lakehouse is a converged infrastructure environment that combines the semantic flexibility of a data lake with the production optimization and delivery of a data warehouse. It supports the full progression of data from its raw, unrefined state, through the steps of refining it, to ultimately deliver optimized data for consumption.

Why This Is Important

Lakehouses provide a single infrastructure that supports different analytical workloads such as data engineering, data science, artificial intelligence/machine learning engineering and business intelligence. Lakehouses have quickly become the design pattern of choice for analytics modernization in the cloud. Organizing around a single platform offers significant operational efficiency for many organizations.

Business Impact

Businesses will benefit from streamlined delivery, rapid access to data, and a consolidated data management platform that supports many different use cases and user personas at both lower cost and lower time to value than many alternatives. A lakehouse provides a well-defined path from discovery-oriented analytics and analytic model development (via the lake portion of the lakehouse) to the delivery of analytic insights and quantification to end users (via the warehouse portion).

Drivers

- Data lakes and data warehouses are optimized for different things. Lakes enable data science, other analytics and the management of any data structure, latency or container. Warehouses excel with refined data that are high quality and accurate, and require an audit trail or special data structures (dimensions, time series and hierarchies). Between the two, the lakehouse supports a long list of data requirements and business use cases.
- The highest growth cloud analytics platform providers are exclusively lakehouse architectures or have lakehouses at the center of their product strategy going forward.
- Many cloud data warehouse solutions and almost all cloud data lakes already leverage semantically flexible cloud object storage as their storage of record. It is a natural progression to unify these storage environments, thus reducing disparate and duplicate infrastructures.
- Data lakehouse concept is maturing fast as it benefits from the market understanding of data warehousing and data lakes, and their respective pros and cons.
- Enterprises consistently seek rapid and unencumbered access to data, and struggle with the processes and perception of delayed delivery associated with the data warehouse. The lakehouse is, inaccurately, often positioned as a silver bullet to solve this problem.
- Operationalizing data science projects, so their insights are shared broadly, continues to be a challenge. A lakehouse unifies the exploratory and production environments for maximum business use and value.

Obstacles

- While interest and usage in open-table formats supporting lakehouses is increasing, they still do not not support the full range of transaction consistency or indexes.
- The most complex data warehousing workloads are still likely to be beyond the scope for most lakehouse solutions that do not incorporate already-developed functionality.
- Lakehouses promise increased data utilization across multiple use cases that will remain out of reach to organizations with weak governance capabilities.
- The full scope of optimization includes data quality, security, governance and performance, and most importantly, good metadata management and data integration. Only a few lakehouse platforms address all of these.
- There is significant vendor messaging around lakehouse products that often is ahead of product capabilities.

User Recommendations

- Employ a targeted use-case approach that solves specific problems and expands from there for long-term success.
- Avoid overpromising and underdelivering by testing candidate solutions thoroughly to ensure that you can actually deliver reliable and high-performance workloads on data lake infrastructure.
- Do not assume that established workloads on traditional enterprise data warehouses will run in the same SLAs within the lakehouse.
- Run your most complex workloads on the evaluated target platform in a proof of concept to make better-informed decisions about when a lakehouse approach is sufficient and when a dedicated data warehouse may still be required.
- Choose a logical data warehouse (LDW) approach when addressing a broader data and analytics scope. A lakehouse is a subset of the LDW built opportunistically. The LDW remains the mature and best practice.
- Evaluate metadata management, security and governance capabilities to ensure they meet your enterprise standards and data requirements.

Sample Vendors

Amazon Web Services (AWS); Cloudera; Databricks; Google; IBM; Microsoft; Oracle; Snowflake; Teradata

Gartner Recommended Reading

Quick Answer: What Is a Data Lakehouse?

How to Leverage Lakehouse Design in Your Data Strategy

Critical Capabilities for Cloud Database Management Systems for Analytical Use Cases

Active Metadata Management

Analysis By: Mark Beyer, Ehtisham Zaidi, Guido De Simoni

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Early mainstream

Definition:

Active metadata management is the continuous analysis of all types of metadata to determine the alignment and exceptions between “data as designed” and “operational experiences.” Its utilization includes operationalizing analytic outputs, operational alerts and recommendations. Active metadata specifically includes metadata from one process being used in a second process to change operations, then confirming activity in both processes.

Why This Is Important

Active metadata management changes how data tools, platforms and systems respond to instructions from each other in a continuous learning loop. It enables augmented and automated solutions, AI-ready, machine learning (ML), DataOps, low-code/no-code and dynamic data governance (security, quality and sharing). Previously, metadata management was simply too large for most human analysts to leverage and was therefore a highly selective and preferential process. ML and graph analytics thrive on expansive information to improve their accuracy and analysis to determine data relevance and validity relative to use cases. Active metadata management enables cross-platform orchestration of data tools, cross-industry validation and verification processes, and the identification of flawed data capture, inappropriate usage, logical fallacies and newly encountered data.

Business Impact

Active metadata management can help businesses:

- Identify commonalities among users, use cases, reporting and analysis models across an organization. The business result has increased human awareness to perform machine-assisted, collaborative data engineering based on observable, common data needs and operational requirements.
- Support self-service analytics and application development by automating data content, structures, availability and discovery of data assets.
- Automate orchestration for data access, locations, performance, processing requirements and resource allocation, adding financial governance benefits (FinOps).
- Accelerate data management deployment while reducing ongoing administration and maintenance efforts.

Drivers

- Changing requirements from both business and IT are driving demand for privacy awareness amid demands for data sharing, data quality tools, data catalogs, metadata management solutions and data integration tools in one comprehensive solution.
- Human-driven data utilization must be augmented to adapt quickly to the demand for the rapid discovery, access and incorporation of new data assets throughout an enterprise or organization.
- Data integration is required even from sometimes distantly removed vertical industries but is complicated by third-party data and data utilized from enterprise partners. Active metadata management introduces awareness of these complex issues — especially with the growing nationalized data management focus that are complicating international data usage.
- Intercloud data demands are increasing rapidly. The large-scale capabilities in cloud-based deployments have enabled the broadest diversity of data structures, processes and use cases to date. Active metadata management introduces awareness of changes in patterns, forms and structures that previously require human data engineering skills to resolve.
- Demands are emerging for organizations to be able to isolate data anomalies and classify them as errors, outliers or actual data design changes that are undocumented.

Obstacles

- Metadata management standards are significantly lacking in today's market, so it is difficult to share metadata bidirectionally across tools and platforms.
- Access to all available metadata (such as social, operational, technical and business) is required for active metadata practices to reach full potential. Automated cross-platform and tools orchestration will be inhibited by data management tools that refuse to share metadata.
- Vendors/suppliers have begun using the terminology to describe any periodic update to passive metadata as "active." This may discredit active metadata, forming resistance due to skepticism that pushes active metadata into the trough.
- Data management solution providers are reluctant to make their metadata assets available to — much less accept instructions conveyed from third-party tools that change their internal operations and processes.
- Human designers, implementers and users may resist this approach, assuming humans are the best interpreters of data value, which also deters metadata management maturity.

User Recommendations

- Introduce an enterprise data catalog, and expand it to ingest metadata beyond your data warehouses/lakes, such as from master data management, data quality, data integration, data preparation tools and analytical tools. Attach it to the catalog entries to accumulate metadata for analysis.
- Develop an operational knowledge graph of which data is used, how often, by whom, for what purpose and on which platform. Accumulate operational metadata (such as runtime logs and system workloads) from the wider data and analytics ecosystem. Analyze the logs for patterns of data combinations in use cases. Examine user or connection strings, queries and views executed, as well as resource allocation.
- Acquire a prototype that combines at least three different disciplines from data management to enable metadata notification between tools. Deploy a UI to reconfigure metadata repositories for analysis by data engineers. This will change the data architecture and design culture into one that involves observing and analyzing instead of designing from scratch — over and over again.
- Pilot the use of LLMs to learn the significant potential of utilizing GenAI and ML together to accelerate data management deployment while reducing ongoing administration and maintenance efforts.

Gartner Recommended Reading

Case Study: An Active Metadata Augmented Data Classification System to Boost Analytics Efficiency

How D&A Leaders Can Use Metadata for Better Business Outcomes

Successful Generative AI Projects Require Better Metadata Management

Metadata Management Solutions

Analysis By: Guido De Simoni, Robert Thanaraj, Brian Lowans

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Metadata management solutions are software with one or more of the following capabilities: metadata repositories, business glossary creation, data lineage, impact analysis, rule management, semantic frameworks, and metadata ingestion and translation from different sources. Modern AI-driven, augmented data cataloging is part of the solutions, automating technical metadata discovery, ingestion, translation and enrichment, and creating semantic relationships between business and security metadata.

Why This Is Important

With data-oriented citizen roles emerging, demand for data access, usage and sharing is not limited to IT. Moreover, data and analytics (D&A) and security leaders face growing security and privacy risks, requiring new data management approaches. Metadata management solutions support organizations managing varied data assets. With use of data across a distributed landscape by citizens, augmented data catalogs — as part of the broader management solutions market — can support metadata discovery and management automation.

Business Impact

D&A leaders investing in metadata management solutions will see benefits from:

- **Collaboration:** Metadata requires contributions from multiple users. Metadata management solutions provide a multiuser environment to address this requirement.
- **Process automation:** As data changes, metadata management solutions can streamline many recurring activities via automation.

- **Cost optimization:** Metadata management solutions ensure that organizations understand the datasets, workloads, queries and tools being utilized, and highlight redundancies.

Drivers

- Metadata management solutions can remediate suboptimal results in an organization's use of data, arising from improperly managed metadata. This saves time, effort and money, while ensuring organizations are not exposed to unnecessary risks.
- Innovation generated by active metadata is not only reducing human effort in inventorying and managing data assets but also accelerating augmented data cataloging and metadata management solutions. Humans are primarily validators as opposed to doers for the operational tasks associated with the metadata management solutions.
- Informal and formal teams emerge and convert into community participation with automation supported by metadata management solutions and leveraged by AI. These demands are only starting to be addressed by vendors, with modern metadata management practices being established within organizations.
- Several vendors are starting to offer enterprise data cataloging techniques that combine business, security and privacy metadata. This enables cross-functional operations for both D&A and security products by leveraging the same metadata management solutions.

Obstacles

- Historically, enterprises have struggled to understand, define and use metadata showing business value, leading to the lack of maturity of strategic business conversations about metadata management solutions.
- The integration of metadata management solutions in multivendor environments is necessary but expensive. This barrier is being addressed by vendors and community initiatives relating to openness and interoperability (for example, the open-source project Egeria).
- The lack of metadata management solutions with capabilities that meet the current and future requirements of specific use cases.

User Recommendations

- Drive metadata insights to mitigate risks by directing business leaders and their teams to ask relevant questions of the data.
- Recognize that the metadata management solutions market will take two to five years to reach the Plateau of Productivity as the technology continues to expand both capabilities and support for existing and emerging use cases.
- Evaluate the metadata management capabilities of your company's existing tools, including data integration, data quality and master data capabilities, before buying a modern solution.
- Pilot the use of metadata management solutions for emergent use cases, including D&A governance, security and risk, data fabric, data mesh, and augmented data value for analytics.
- Invest in metadata management solutions to augment manual efforts needed to access all types of metadata and analyze the data to support your company's data fabric designs for automation.
- Cooperate with security and privacy teams to establish the benefits of enterprise data cataloging techniques.

Sample Vendors

Alation; Alex Solutions; Atlan; Collibra; data.world; IBM; Informatica; Zeenea

Gartner Recommended Reading

How D&A Leaders Can Use Metadata for Better Business Outcomes

Application Data Management

Analysis By: Andrew White, Tad Travis

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Application data management (ADM) is a technology-enabled discipline where business and IT work together to ensure uniformity, accuracy, and semantic consistency for data in an application or suite, such as ERP, supply chain management, customer data platform or custom-made app. Application data is the consistent and uniform set of identifiers and extended attributes used within an application or suite for things like customers, products, bills of materials or prices.

Why This Is Important

Many SaaS and business application providers are currently marketing capabilities to support governing data in those applications. However, clients remain, overall, disillusioned with such capabilities. Success stories are uncommon. Business application vendors typically misunderstand the needs of governing application data. Mistaking application data for master data and master data management (MDM) serves as a great example of that confusion; trying to sell data catalogs or data ownership programs compounds the confusion. Every business application needs some data governance — otherwise, the business process it supports may fail due to data issues.

Business Impact

ADM can offer the following benefits:

- Application data, once identified, ensures that governance effort is appropriately aligned with the intended business impact of the application.
- Stewardship roles in the business, and in operational and analytical use cases, can be determined more effectively.
- Business process and decision integrity, assumed by the use of the business application, is more reliable and assured.
- MDM programs connected to or built on ADM will help govern such application data that is shared with other applications.

Drivers

- The majority of "successful" go-lives of business applications, such as ERP, CRM or custom-built applications, do not include any qualification of data and analytics governance. The result, very often observed in client inquiry, is that, on average, seven months after the "go live," organizations spot the vast array of small but noticeable business issues held hostage by the lack of governed data. Business performance and process integrity fail, and business outcomes start to be negatively impacted.
- Many MDM programs are poorly defined and instead seek to govern all application data. This leads to misaligned expectations, investments that don't deliver the targeted ROI, and less positive business impact from the use of the business application. Rightsizing the work of application data and its governance is therefore key to any new or modernized application program.
- Digital business success hinges not on the quality and governance of all data equally, but on a graduated, efficient means to classify data and apply adaptive data and analytics governance. Such growing demand on scaling digital business will, of necessity, drive increased need to recognize and adopt ADM.

Obstacles

- Many misnamed MDM programs associated with large, global ERP, CRM and SCM implementations still mistakenly centralize all the work related to governing application-specific data. MDM is meant only for that narrowest definition of master data that is shared between applications. Others create a hybrid organization across business and IT, and mistakenly call it all data management. Alternatively, these programs conflate MDM and ADM, making both too slow, expensive and unwieldy. As a result, neither program is a success.
- On average, seven months after the "go live" of a business application implementation (such as ERP), clients tell Gartner, "We have lost control of our data." This loss of control resulted from a lack of ADM implementation as part of the ERP roll out, and so over time, poor data quality and governance erodes trust and integrity of business processes. This situation has become almost acceptable because, overall, most organizations don't fail mortally from poor data quality.
- The organization's ability to support agility is held back, and consequently, budgets are set that even support mediocrity via poor governance practices. This is an unacceptable way to run an organization, but too few data and analytics leaders speak up against this approach.
- Traditional top-down governance programs lead to the same misunderstandings and poorly scoped initiatives.

User Recommendations

- Start with a focus on business outcomes to identify what data matters most, then organize, classify and govern data based on which data drives the most important business outcomes.
- Identify your application data to scope ADM. That is, identify the data that matters most to a specific set of use cases supported by each application or suite like ERP, CRM, or e-commerce, product information management or customer data platform.
- Examine reusing MDM technology to support your ADM implementation — even if it is in a distinct instance. The business requirements are very similar — but the value propositions are different.
- Demand from your business application provider (and those in the cloud) the necessary capability to set (that is, govern) and enforce (that is, steward) information policy pertaining to data used in the application or suite.
- Implement ADM for application-specific governance alongside any MDM for core data shared between applications. This allows you to operate at your own speed and benefit while aligning and sharing metadata in support of a wider enterprise information management program.

Sample Vendors

ChainSys; Epicor; Oracle; PiLog Group; Tealium; Utopia Global

Gartner Recommended Reading

How D&A Leaders Can Assess and Advance Master Data Management Maturity

Augmented Data Management

Analysis By: Roxane Edjlali

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Augmented data management is the application of AI and machine learning (ML) in data management technology for optimization, automation and improved operations. Based on existing metadata and usage data, AI and ML are applied to tune operations and optimize configuration, security and performance. Augmented data management applies to all tools and technologies used as part of data management, such as metadata management, master data management (MDM), data integration, data quality and DBMS.

Why This Is Important

AI and ML can automate time-consuming and error-prone data management activities, altering job roles, product design and overall data management processes. These solutions are being used not only to tune and optimize the use of the products themselves based on actual usage, including failures and poor performance, but also to suggest and implement new data pipelines, designs, schemas and queries. These solutions can even infer the semantics of the data to recommend structural improvements. Generative AI (GenAI) is increasingly combined with augmented data management capabilities, exposing them through natural language interfaces. However, GenAI techniques are not as mature as some of the other augmented capabilities, slowing down the progress of augmented data management. Moreover, vendors are increasingly marketing all augmented capabilities under GenAI, adding to the total hype and reducing the overall market understanding.

Business Impact

Augmented data management offers benefits in metadata management, data integration, MDM, data quality and DBMS. This will assist those engaged in data management by automating many manual, repetitive tasks performed today, and increasing accuracy and reducing time spent on these tasks. This allows valuable resources to perform other tasks with far more business value. As these capabilities span full data ecosystems, augmented data management becomes a competitive differentiator among full cloud service provider ecosystems, enabling the utilization of metadata from all the uses across the various services.

Drivers

- **Metadata management:** Increasingly, AI and ML are used to explore and define metadata from the data, helping the analysts to evaluate metadata more rapidly, accurately and with reduced redundancy. Similarly, augmented data management functions can automatically catalog, tag and annotate data elements during data extraction, access and processing.
- **Data integration:** Augmented data management automates the integration development process, including the development of data pipelines, by recommending or deploying repetitive integration flows, such as source-to-target mappings.
- **MDM:** MDM solution vendors will increasingly focus on offering AI- and ML-driven configuration and optimization of record matching and merging.
- **Data quality:** AI and ML will be used to extend profiling, cleansing, linking, deduplication, identifying and semantically reconciling data in different data sources.
- **DBMS:** In addition to enhancing performance and cost-based query optimization, AI and ML are being used to automate many current manual management operations, including the management of configurations, elastic scaling, storage, indexes and partitions, and database tuning.

Obstacles

- The adoption of augmented data management capabilities is gated by the movement of the product categories to the cloud and their participation in data ecosystems. Offerings in data integration, data quality, MDM, metadata management and DBMS software are delivering augmented data management capabilities but at different speeds.
- As augmented data management represents an aggregate view and position of what is happening across data management, some areas are more mature in augmented data management than others and some data management areas will benefit more from augmentation than others.

User Recommendations

- Create a business case for using new data management capabilities, and be sure to model and measure the benefits realized from their implementation.
- Question the vendors of your data management tools about their roadmap for the introduction of AI and ML into their products. Make augmented capabilities a "must have" selection criterion for new purchases of data management products.

- Begin testing the components of augmented data management products (where visible) to understand their capabilities and the validity of the automated functionality. Audit the results; with any new functionality, there is the risk of introducing errors and reduced performance.
- Plan for role transitions by providing new skills training to add value as responsibilities evolve.
- Seek data management solutions that share design and performance metadata for external usage.

Sample Vendors

Amazon Web Services (AWS); Google Cloud; IBM; Informatica; Microsoft; Oracle; SAP; SnapLogic

Gartner Recommended Reading

Market Guide for Active Metadata Management

Innovation Insight: How Generative AI Is Transforming Data Management Solutions

Innovation Insight: State of Data Management Support for Self-Service

Critical Capabilities for Data Integration Tools

Graph DBMS

Analysis By: Masud Miraz, Aaron Rosenbaum

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Graph DBMSs are designed to be multirelationship databases that store data in collections of nodes and edges, instead of columns and rows. Most graph DBMSs use basic graph theory and are suitable for multiple types of interactions, including complex multihop queries, reasoning and inference, and algorithms for analytical workloads.

Why This Is Important

Graph DBMSs have built-in commands, controls and structures to reduce thousands of lines of relational code into dozens of graph-based instructions. Relational DBMSs (RDBMSs) are best suited to specific relationships and optimized to follow preapproved pathways. Graph DBMSs allow business users to describe and query dynamic and complex domains more naturally and efficiently. This reduces the need to create complex looping controls that do not scale well.

Business Impact

Graph support has widely demonstrated value. Customer 360, supply chain, transportation/logistics, financial fraud detection and network analysis are typical use cases. Correlations across multiple sensors in asset-intensive Internet of Things (IoT) use cases are fueling increasing marketplace interest. Embedded graph support of metadata use cases is becoming increasingly important.

Drivers

- The challenges of connecting a broader variety of data to uncover new insights and hitting the wall with RDBMS because of performance challenges have accelerated adoption.
- Cloud-first developments are increasing, facilitated by ingest and conversion of existing data in a low-cost, low-risk environment, and graph DBMSs are widely available in cloud platforms.
- The increasing importance of semantics in organizing and operating larger and more complex data fabrics is accelerating experimentation, adoption and training.
- The increasing visibility of use cases that exploit graph and achieve successes in solving previously intractable operational problems is raising interest even further. Mature, widely distributed multimodel DBMS products are embedding graph DBMS engines and/or functions, and both open-source and commercial graph DBMS offerings are maturing and gaining adherents.
- Many vendors are testing the use of graph stores to enable knowledge graphs, and then using GraphQL queries to input data into large language models (LLMs), which are giving better contextual results.

Obstacles

- Modeling, loading, processing and analyzing graph data require new and uncommon skills, including the sometimes steep learning curve of a graph query language (GQL) like Gremlin.
- Dueling architectures (knowledge graphs versus labeled property graphs), languages and APIs, and varying support among vendors for libraries of machine learning algorithms make choosing a product complex.
- Existing analytical tools can often be successful for graph use cases without requiring a dedicated DBMS, making graph DBMS sometimes unnecessary.
- The inability of relatively more technical practitioners to adequately explain the benefits of graph to business sponsors continues. Moreover, many use cases are now moving to embedded functionality in other tools (AI, metadata and data integration lead the way), making the decision to plunge into an unfamiliar type of product less attractive.

User Recommendations

- Assess graph DBMS capabilities when performance requirements for highly nested or connected data fall outside current RDBMS processing capabilities. Consider your current DBMS's graph functions or evaluate other multimodel databases that have broader applicability and can cover graph use cases.
- Use open-source graph DBMS projects, or community editions of commercial ones, to experiment and gain experience. Ensure any open-source products you use in production environments are commercially supported. Scaling and reliability cannot be relied on with open-source offerings.
- Use graph DBMSs to render the relationships and traversal of data into a reusable form for data miners, data engineers and business analysts.
- Select graph DBMSs based on your analytics requirements. Factor in the capabilities of Resource Description Framework (RDF)-based and property graph-based architectures when making your selection.
- Familiarize data teams with the just released and much anticipated International Organization for Standardization (ISO) GQL and associated SQL to property graph (Property Graph Query Language [PGQL]) standard. It will enhance interoperability between graph and RDBMS platforms.

Sample Vendors

Amazon Web Services (AWS); ArangoDB; Cambridge Semantics; Hypermode (Dgraph Labs); Microsoft; Neo4j; Ontotext; OpenLink Software; TigerGraph

Gartner Recommended Reading

Graph Steps Onto the Main Stage of Data and Analytics: A Gartner Trend Insight Report

Working With Graph Data Stores

Understanding When Graph Analytics Are Best for Your Business Use Case

How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications

Market Guide for Graph Database Management Systems

Climbing the Slope

Distributed Transactional Databases

Analysis By: Henry Cook

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

A distributed transactional database allows transactions to be performed on any of a distributed set of database instance nodes. The ability to accept writes from a geographically distributed set of nodes while maintaining data integrity and consistency and delivering acceptable performance distinguishes this technology.

Why This Is Important

A distributed DBMS allows transactions to be performed across a set of distributed nodes/servers, regarded as one database instance with high throughput and comparable performance and integrity to nondistributed systems. Nodes and locations can be dynamically added or removed and the system adjusts. Thus, high availability and disaster recovery are a natural part of their operation — not a separate and complex set of recovery mechanisms.

Business Impact

Organizations that need to implement applications whose transactions span geographic distances will need systems that don't compromise data integrity. Systems such as global trading applications use distributed transactional systems to expand their scope and reach, and may provide competitive advantage by allowing organizations to deliver unique capabilities. However, their configuration and operation must be well-understood.

Drivers

- Global business operations increasingly require widely distributed data models and active transactional capabilities across the distributed nodes. For example, applications that provide a global financial position to their users.
- Cloud platforms that provide underlying infrastructure for data distribution make it easier to deploy widely distributed platforms for globally active transactional systems. This allows workloads to be dynamically moved between locations between clouds, on premises and clouds. This allows of processing platform and dynamic movement of workload between them.
- Distributed data allows more local data access and lower latency worldwide for systems that need reduced latency in operations spread over a distributed set of data and that are still subject to write transactions. Electronic gaming applications can provide high throughput by distributing their work across many distributed nodes — while maintaining an overall consistent view of the game state.
- Systems that can support distributed transactions can also offer the highest levels of resilience and fault tolerance. Coping with node failure is a standard part of distributed working, so high-availability applications such as mission-critical manufacturing or finance applications are a good fit.

Obstacles

- Geographically distributed transactions are still not requirements in the majority of transactional systems, so they may not be considered.
- Data that is not required for distributed transactional scenarios can be processed with parallel processing and be implemented with a much wider variety of systems — still delivering scalability and performance advantages but without the use of specialized distributed engines.
- Implementers need to understand how distributed database systems are configured to ensure that they operate and recover while maintaining data integrity.

User Recommendations

- Select a distributed transactional DBMS if your use case requires data integrity implemented over a geographically distributed database instance. Note that some regular relational database systems are now adding distributed capability, so adoption is not as daunting as it has been in the past.
- Examine the use of distributed transactional systems for high-throughput applications and where the need for rolling upgrades and high availability are valued — distributed database systems cope with these requirements as part of their normal operation.
- Ensure that a distributed transactional DBMS includes policies for update contention resolution that are appropriate for your business use cases. Several distributed transactional DBMS have built-in contention avoidance or resolution which provides atomicity, consistency, isolation and durability (ACID) levels of DBMS integrity.

Sample Vendors

Aerospike; Cockroach Labs; Couchbase; Dassault Systèmes (NuoDB); DataStax; Google; MongoDB; Oracle; Yugabyte

Gartner Recommended Reading

Data Consistency Flaws Can Destroy the Value of Your Data

Magic Quadrant for Cloud Database Management Systems

Critical Capabilities for Cloud Database Management Systems for Operational Use Cases

Knowledge Graphs

Analysis By: Afraz Jaffri

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Knowledge graphs are machine-readable representations of the physical and digital worlds. They include entities (people, companies and digital assets) and their relationships, which adhere to a graph data model — a network of nodes (vertices) and links (edges/arcs).

Why This Is Important

Knowledge graphs capture information about the world in a visually intuitive format yet are still able to represent complex relationships. Knowledge graphs act as the backbone of a number of products, including search, smart assistants and recommendation engines. Knowledge graphs support collaboration and sharing, exploration and discovery, and the extraction of insights through analysis. Generative AI models can be combined with knowledge graphs to provide context for more accurate outputs in a technique becoming known as GraphRAG or G-RAG.

Business Impact

Knowledge graphs can drive business impact in a variety of different settings, including:

- Digital workplace (such as collaboration, sharing and search)
- Automation (such as ingestion of data from content to robotic process automation)
- Machine learning (such as augmenting training data)
- Investigative analysis (such as law enforcement, cybersecurity and risk management)
- Digital commerce (such as product information management and recommendations)
- Data management (such as metadata management, data cataloging and data fabric)

Drivers

- The need to complement AI and machine learning methods that detect only patterns in data (such as the current generation of foundation models) with the explicit knowledge, rules and semantics provided by knowledge graphs.
- The desire to make better use of unstructured data held in documents, correspondence, images and videos, using standardized metadata that can be related and managed and provide the foundation for AI-ready data.
- The increased usage of knowledge graphs with large language models (LLMs) to provide enhanced contextual understanding when answering questions on large quantities of enterprise data.
- The increasing awareness of the use of knowledge graphs in consumer products and services, such as smart devices and voice assistants, chatbots, search engines, recommendation engines and route planning.
- The emerging landscape of Web3 applications and the need for data access across trust networks, leading to the creation of decentralized knowledge graphs to build immutable and queryable data structures.
- The need to manage the increasing number of data silos where data is often duplicated, and where meaning, usage and consumption patterns are not well-defined.
- The use of graph algorithms and machine learning to identify influencers, customer segments, fraudulent activity and critical bottlenecks in complex networks.

Obstacles

- Awareness of knowledge graph use cases is increasing, but business value and relevance are difficult to capture in the early implementation stages.
- Moving knowledge graph models from prototype to production requires engineering and system integration expertise. Methods to maintain knowledge graphs as they scale — to ensure reliable performance, handle duplication and preserve data quality — remain immature.
- The graph DBMS market is fragmented along three properties: type of data model (Resource Description Framework or property), implementation architecture (native or multimodal) and optimal workload (operational or analytical). This fragmentation continues to cause confusion and hesitation among adopters.
- Organizations want to enable the ingestion, validation and sharing of ontologies and data relating to entities (such as geography, people and events). However, making internal data interoperable with external knowledge graphs is a challenge.
- In-house expertise, especially among subject matter experts, is lacking, and identifying third-party providers is difficult. Often, expertise resides with vendors of graph technologies. Skills in scalability and optimization are also hard to acquire.

User Recommendations

- **Create a working group of knowledge graph practitioners and sponsors** by assessing the skills of data and analytics (D&A) leaders, practitioners and business domain experts. Factors like use case requirements, data characteristics, scalability expectations, query flexibilities and domain knowledge of knowledge graphs should be addressed.
- **Run a pilot to identify use cases that need custom-made knowledge graphs.** The pilot should deliver not only tangible value for the business, but also learning and development for D&A staff.
- **Create a minimum viable subset that can capture the information of a business domain to decrease time to value.** Assess the data, both structured and unstructured, needed to feed a knowledge graph, and follow Agile development principles.
- **Utilize vendor and service provider expertise** to validate use cases, educate stakeholders and provide an initial knowledge graph implementation.
- **Include knowledge graphs within the scope of D&A governance and management.** To avoid perpetuating data silos, investigate and establish ways for multiple knowledge graphs to interoperate and extend toward a data fabric.

Sample Vendors

Cambridge Semantics; Diffbot; eccenca; Fluree; Neo4j; Ontotext; Stardog; TigerGraph; TopQuadrant

Gartner Recommended Reading

How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications

3 Ways to Enhance AI With Graph Analytics and Machine Learning

How Large Language Models and Knowledge Graphs Can Transform Enterprise Search

Data Engineering

Analysis By: Robert Thanaraj, Ehtisham Zaidi, Roxane Edjlali

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Early mainstream

Definition:

Data engineering is the discipline of translating raw data into usable forms by building and operationalizing data pipelines across various data and analytics systems meeting use-case requirements, data governance principles and SLAs. Data engineering is a team competency that brings together three different practices — data management, software engineering and I&O management — to achieve frictionless, trustworthy data delivery with agreed-upon business and technical SLAs.

Why This Is Important

Data engineering caters to the full data supply chain, provisioning trusted, quality data to be used at the right business moments with agreed service levels. It enables the creation, operationalization and maintenance of data pipelines across heterogeneous environments, aimed at delivering integrated data as per consumer needs. It also manages data delivery debt around reusability, governance, compliance and operational readiness through data cataloging, data quality and data observability efforts.

Business Impact

- Contextualized data leads to efficient data utilization.
- High-quality data improves data trust among data consumers.
- It becomes faster to onboard new data to existing analytics and data science models with robust data pipeline change management.

- It becomes easier to fulfill regulatory requirements to meet data transparency expectations through cataloging efforts in tracking sensitive data and enabling data governance enforcement.

Drivers

- **Productivity through automation:** Data management teams spend most of their time on data preparation, data integration and operationalization. As a result, these are the primary candidates for automation.
- **Agility:** Organizations are forced to change the way they traditionally work with the waterfall methodology of data pipeline delivery because they are unable to keep up with the demand increase and skills shortage.
- **Self-service and customer experience:** Organizations seek successful consumer experiences, the last mile in the "data insights decisions" continuum. Data engineering enables self-service data management among citizen users and domain experts, and also provides guardrails and established best practices to follow. See Lesson No. 2 in 6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh.
- **Cloud data management:** As organizations struggle to manage multiple data gravities across their on-premises and multicloud setup, the data engineering practice plays a major role in balancing the collect and connect approaches in distributed architectures.
- **Analytics/business intelligence (BI) and data science success:** Organizations are bound to fail if they launch data science initiatives without onboarding the necessary data engineering skills because of high technical debt associated with data management (such as governance, compliance and operations readiness), which must be managed.

Obstacles

- **Inefficient legacy practices:** Baggage around poor integration and operations practices hurts and/or delays data engineering practice adoption.
- **Unrealistic expectations:** Many think a data engineer can "do it all," catering to the full spectrum from data management to software engineering to infrastructure and operations. Data engineering is a team competency involving data architecture, data modeling, data stewardship, software engineering skills (such as DevOps engineering, Python development or test engineering), domain knowledge, incident management and operationalization expertise. Sometimes, organizations might even need adjacent roles for test engineering and infrastructure automation.
- **Inability to scale:** Data engineering roles are a critical part of data engineering teams. However, adding more data engineers in response to increasing data demands is not sustainable. Organizations need DataOps practices to streamline and scale data engineering delivery.

User Recommendations

- Establish a data engineering discipline with roles that support end-to-end delivery of data pipelines aligned to business use cases, SLAs and governance requirements.
- Catalog an inventory of data assets and make them searchable. Use metadata to drive automation of data pipelines and related artifacts. Study data usage patterns among consumers and systems, and employ this metadata to improve efficiency and optimize delivery.
- Evaluate progress measures — such as time to market, productivity, CI/CD automation of data pipelines, code quality, and cost efficiency of build and operations — quantitatively and regularly, and share them with your stakeholders. As a data engineering leader, add this to your communication plan.
- Foster a collaborative environment by adopting a hub-and-spoke operating model. This model establishes a centralized data engineering team that supports multiple decentralized data engineering teams embedded within various business functions and the data science lab.

Sample Vendors

Ab Initio; Ascend.io; Coalesce Automation; dbt Labs; Informatica; Nexla; Prophecy; Qlik; Upsolver

Gartner Recommended Reading

5 Ways to Enhance Your Data Engineering Practices

6 Lessons Data Leaders Can Learn From the Early Adopters of Data Mesh

Critical Capabilities for Data Integration Tools

Data and Analytics Essentials: DataOps

Operational Intelligence

Analysis By: Henry Cook

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Operational intelligence provides analytical processing within transactionlike workloads. Two forms of the technology exist. In the first, transactional systems extend to include advanced analytics, AI and machine learning (ML). In the second, analytical databases extend to enable highly concurrent transactionlike queries. In-memory computing (IMC) technology is a key enabler of operational intelligence.

Why This Is Important

Operational intelligence (formerly hybrid transactional/analytical processing [HTAP]) provides analytical processing within transactionlike workloads. Business processes can be steered in real time, assisted by analytics, AI and ML. Use cases include planning, dynamic repricing, forecasting and what-if analysis. The analytics become an integral part of the business process itself, rather than a separate activity performed after the fact having incurred a delay for extraction, transformation and loading.

Business Impact

Operational intelligence enables business users to make more informed operational and tactical decisions in real time. It opens up the possibility of driving prescriptive decisions through automated decision making aided by analytics — that is, without relying solely on a staff member's intuitive situational awareness of operations. Operational intelligence also provides constantly updated forecasts and simulations of future business outcomes.

Drivers

- Operational intelligence subsumes "augmented transactions" (the previous name of this innovation). In operational intelligence, the transactional and analytical processing dimensions are designed together in the context of an individual transaction or process, and are delivered as a single application. This packaging helps streamline the overall technology infrastructure. Operational intelligence and augmented transactions were formerly the two flavors of HTAP — named "in-process HTAP" and "point-of-decision HTAP."
- Because of in-memory database technology and other techniques like in-stream processing, organizations, for some types of analysis, reduce the reliance on analytical systems. This applies in particular to external or after-the-fact analytics. These techniques implement significant analyses within a transaction. For example, they can execute predictive models or run significant online analytical processing queries. Likewise, analytical systems have improved, allowing large numbers of transactionlike queries to be executed concurrently with fast response times.

- Automated recommendations, based on real-time analysis of the impact of different options on KPIs, can improve decisions. Performance and scalability limitations have historically prevented advanced analytics from running concurrently with or within transaction processing.
- Many DBMSs are enabling ML, vector search and generative AI within the database, increasing the functionality available for operational intelligence. Further, the growth of in-memory DBMSs, including the addition of in-memory functionality to general DBMSs, is enabling the use of operational intelligence.

Obstacles

- Operational intelligence may require significant migration of function when organizations are retrofitting existing (custom or packaged) applications at the core of business operations.
- Skills in operational intelligence are limited. This discipline requires a good understanding of designing for performance and the trade-offs involved. However, as the technology improves, the skills requirements are easing.
- Organizations may not be fully aware of the availability of mixed transactional and analytical processing features and the addition of some analytical functions to transactional systems. Not all use cases can or should be addressed through operational intelligence. Suitability must be determined on a use-case-by-use-case basis. Operational processes that need a modicum of analytics are the "sweet spot."

User Recommendations

- Educate business leaders about operational intelligence concepts and their importance. Brainstorm with them to identify concrete opportunities to rethink business processes and create applications that could not be implemented using traditional architectures.
- Ask strategic information management and business application providers about their vision, roadmap and technology for using operational intelligence within their products — that is, their support for real-time and near-real-time working. Selecting the right use cases is critical.
- Pilot operational intelligence in individual "system of innovation" projects to get a sense of how much mixed transactional and analytical processing is possible.
- Employ operational intelligence for use cases involving observation data (such as Internet of Things) or interaction data (such as log data) for which real-time operational analytics is required.

Sample Vendors

Aerospike; IBM; Microsoft; Oracle; Redis; SAP; SingleStore; Teradata

Gartner Recommended Reading

Magic Quadrant for Cloud Database Management Systems

Critical Capabilities for Cloud Database Management Systems for Operational Use Cases

Critical Capabilities for Cloud Database Management Systems for Analytical Use Cases

Data Lake

Analysis By: Masud Miraz, Roxane Edjlali

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Definition:

A data lake is a semantically flexible data storage repository combined with one or more processing capabilities. Most data assets are copied from diverse enterprise sources and are stored in their raw and diverse formats, so they can be refined and repurposed repeatedly for multiple use cases. Ideally, a data lake will store and process data of any structure, latency or container, such as files, documents, result sets, tables, formats, binary large objects (BLOBs) and messages.

Why This Is Important

Data lakes enable data collection and advanced analytics and complement traditional data warehouses. For example, the massive repository of source data in a data lake supports broad, flexible and unbiased data exploration, which is a prerequisite for data mining, statistics, ML and other analytics techniques. A data lake can also provide scalable and high-performance data acquisition, preparation and processing, either to be refined and loaded into a data warehouse or for processing within the data lake.

Business Impact

A data lake can be a foundation for multiple forms of business analytics. For example, data science is a common use case for a data lake, enabling predictive analytics to help a business retain customers, execute impact analyses, and anticipate issues in maintenance, logistics, risk and fraud. Similarly, data lake use for self-service data access is a growing business use case that contributes to programs for business transformation and digitization. Data lake is not a democratizing data storage form. It must be used by highly skilled users as unskilled users tend to drown in the data lake. This is what leads users to the data lakehouse concept.

Drivers

- Organizations are increasingly driven by data and analytics, so they can achieve their goals in business transformation, digitization, data democracy, operational excellence and competitiveness. A data lake provides the required data and supports analytics for these high-value goals.
- Established forms of analytics — namely reports, dashboards, online analytical processing (OLAP) and statistical analysis — will continue to be relevant and maintained by organizations. However, as maturity grows, organizations gradually expand their analytics programs, such as data mining, natural language processing (NLP), machine learning, AI and predictive analytics. A data lake provides the scale, as well as the structure-agnostic storage and processing options, that advanced analytics require.
- Data exploration and data engineering are now common practices. This is true for many user types, from data scientists and data analysts to business end users who are capable of self-service data preparation. A data lake, when designed properly, can provision data for the diverse exploration requirements of multiple user types and use cases.
- Data lakes can expand the data warehouse and address additional use cases, such as data exploration. In these cases, the warehouse and the lake are integrated by shared refined datasets, platform infrastructure (DBMS brands and storage, whether on-premises or cloud) and architecture components (data landing/staging).

Obstacles

- Data lake best practices are still evolving. There is still much confusion about how to design and govern a data lake, plus how to optimize a lake's data without losing its purpose as a repository for data science and advanced analytics. An ungoverned, disorganized data lake can result in an unwieldy "data swamp."
- The first data lakes, built on Hadoop, were meant to be used for data science only. They lacked metadata, relational functionality and governance. Now, data lakes are typically on the cloud, have different data storage types, and support multiple analytics techniques (not just data science). Data governance is crucial and cannot be neglected; it includes data quality, data catalog, data security and data life cycle management.
- Success in the data lake relies on having skilled users who can work with raw or less refined data. Many early data lakes failed because this was not recognized or acknowledged.

- One common architecture is to use a data lake along with a data warehouse. While currently popular, this architecture can lead to issues with stale data and high data integration costs (as data must be loaded into the data warehouse) and difficulty performing advanced analytics. As a result of these challenges, some organizations are adopting a lakehouse architecture instead of using a data lake along with a data warehouse.

User Recommendations

- Build a data lake as a foundation for establishing competency in data science and advanced analytics.
- Staff the data lake for maximum value by hiring data scientists, data engineers and analysts who have the skills required to conduct data exploration and analytics with the lake's data.
- Create business value by designing a data lake that addresses multiple high-value business use cases, such as data science, analytics, self-service data access or customer 360.
- Enable broad data exploration, multiple analytics techniques and machine learning by populating a data lake with broadly collected data in various structures, formats and containers.
- Modernize the whole data architecture to extend the data lake. Consider logical data warehouse and lakehouse concepts.
- Keep each data lake from becoming a data swamp by governing the use of data in the lake, curating the data allowed into the lake, and documenting data via metadata and other data semantics.

Sample Vendors

Amazon Web Services (AWS); ChaosSearch; Cloudera; Databricks; Dremio; Google; Infoworks; Microsoft; Snowflake

Gartner Recommended Reading

Building Data Lakes Successfully — Part 1 — Architecture, Ingestion, Storage and Processing

Building Data Lakes Successfully — Part 2 — Consumption, Governance and Operationalization

Data and Analytics Essentials: Data Warehouses, Data Lakes and Data Hubs

Market Guide for Analytics Query Accelerators

Master Data Management

Analysis By: Sally Parker

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

Master data management (MDM) is a technology-enabled business discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, governance, semantic consistency and accountability of the enterprise's official shared master data assets. Master data is the least number of consistent and uniform sets of identifiers and extended attributes that describe the core entities of an enterprise.

Why This Is Important

MDM is a cross-organizational collaborative effort that focuses on the consistency, quality and ongoing stewardship of master data. Master data is the subset of data that describes the core entities an organization requires to function — customers, citizens, products, suppliers, assets and sites. Master data sits at the heart of the most important business decisions, driving a need for a consistent view across business silos. Thus, it is where governance efforts yield their highest returns.

Business Impact

MDM is mainstream; MDM initiatives are a foundational component of digital transformation and business agility more broadly. Organizations seeking trust in their master data recognize it is a necessity. Leading organizations draw a causal link between their master data (parties, things and places) and business outcomes, including customer retention and satisfaction, supply chain optimization, and improved risk and regulatory compliance. Thus, vested-interest stakeholders exist across the enterprise.

Drivers

- **Business agility and resilience:** Rapidly evolving business needs, particularly in uncertain times, translate into greater demand for the benefits of trusted and connected data afforded by MDM.
- **Business process integrity:** Complex application and data landscapes often suffer from master data that is siloed and inconsistent, creating a lack of trust in the data and the decisions made using it. Organizations are increasingly recognizing the causal link between master data and business goals.
- **AI and analytics use cases:** Recent hype around AI has triggered a heightened need for a trusted, curated foundational layer of master data to ensure accuracy in AI and analytics programs.
- **Cloud:** Cloud- and subscription-based pricing have lowered the barrier to entry for commercial MDM solutions, rendering MDM a viable option beyond large enterprises. Today, almost all new MDM deployments are cloud-based.
- **Multidomain:** While customer-MDM and product-MDM remain the most common focus areas, interest levels are increasing across a broader range of business stakeholders that require additional MDM domains to be targeted to address business problems more holistically. For example, product-MDM, location-MDM, asset-MDM, employee-MDM and supplier-MDM collectively address environmental, social and governance compliance in supply chain.
- **Land and expand:** A prior hesitance to embark upon MDM initiatives, due to complexity and cost, is easing. Cloud, more readily consumable user interfaces, and a shift toward more granular, business-outcome-led MDM initiatives are the norm. Vendor strategies reflect this approach with their land-and-expand strategies.
- **Application modernization:** Digital transformation is forcing organizations to start or modernize their MDM programs to leverage more recent cloud-based offerings and new augmented MDM capabilities that are an expectation of MDM solutions today.
- **Variable geographic maturity:** North America is the most mature region, followed by Western Europe. The rest of the world is earlier in the maturity cycle, representative of markets primed for growth.

Obstacles

- **Inconsistent MDM vendor coverage:** Coverage is weakest outside North America and Northern Europe. Additionally, not all vendors suit all industries and use cases (see Market Guide for Master Data Management Solutions).
- **MDM is a discipline, not a tool:** The prevailing pitfall remains the instinct to treat MDM as a technology initiative in isolation. Technology alone won't solve a challenge that traverses people, processes and technology.
- **Stakeholder engagement:** Organizations that fail to proactively engage business stakeholders in scoping struggle to meet expectations of value and to establish an operational governance structure in service of MDM.
- **Perceived complexity:** Traditional approaches to MDM have failed to deliver value. A combination of modern, more approachable MDM solutions and more granular approaches to MDM programs have reduced both complexity and deployment times.
- **Skills:** Successful MDM implementations require business acumen, technology and governance capabilities. Finding the right balance and availability of these skill sets remains problematic and is driving a need for third-party services as the norm.

User Recommendations

- **Start with "why":** Approach MDM as a technology-enabled business-led initiative. Consider *why* you are doing MDM and what business outcomes it will impact.
- **Choose iterative rollout of business outcomes over serial rollout of technology:** Plan the rollout based on a prioritized list of business outcomes to deliver more granular, but incremental, business value.
- **Secure executive sponsorship:** Sponsorship facilitates required cross-organizational collaboration.
- **Communicate value:** Ensure the causal link between the MDM initiative and the business outcomes it supports is clearly understood and articulated.
- **Keep master data lean:** Your master data attributes should be lean and focused. Use business outcomes to identify the least amount of data with the greatest business impact.
- **Leverage services to fast-track time to value:** Most organizations leverage external support with their MDM strategy and/or implementation. Third parties offering industry expertise and accelerators can greatly impact time to value.

Gartner Recommended Reading

How D&A Leaders Can Assess and Advance Master Data Management Maturity

Data and Analytics Essentials: Master Data Management

3 Essentials for Starting and Supporting Master Data Management

How to Measure and Communicate the Business Value of MDM

Innovation Insight: The Rapid Evolution of Master Data Management

Time Series DBMS

Analysis By: Masud Miraz, Aaron Rosenbaum

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Time series DBMSs are designed to provide rapid ingestion and manipulation of data based on its position in a time series. With the growth of internet-connected mobile devices and the decreasing cost of storage, the demand for time series databases has grown over the last decade.

Why This Is Important

Time series DBMSs are well suited for high-volume transaction systems, such as Internet of Things (IoT), streaming data and financial services systems. The requirements of rapid data ingestion and high throughput, combined with real-time analysis of incoming data, are typical tasks for time series DBMSs. These products can keep detailed data for a short period of time and aggregations based on that data for a longer period. Time series DBMSs can compare and analyze data across multiple time streams, which is difficult to do outside of this database technology.

Business Impact

Time series DBMSs are useful for real-time IT infrastructure and operations, industrial operations, IoT, real-time trading, weather and financial systems, where time series measurements and rapid data ingestion are critical. Time series databases can efficiently store large volumes of time series data, often using data compression techniques to reduce storage footprint and thus save on storage costs. Built-in support for data retention policies and downsampling also provide cost savings.

Drivers

- The expansion of IoT data sources has established interest in this technology as organizations aim to conduct more sophisticated analytics on IoT data and combine it with more traditional sources.
- Customer demand for time series databases has prompted leading cloud vendors, such as Amazon Web Services and Microsoft Azure, to add time series capabilities or services to their platforms.
- The emergence of cloud data ecosystems will allow offerings with a specific focus, like time series capabilities, to be more easily integrated with the overall data landscape.

Obstacles

- Time series capabilities are being added to mainstream database offerings, and time series database services are being offered by major cloud providers. These dual developments, coupled with the ability to perform standard analysis without specific time series capabilities for many use cases, place pressure on DBMSs whose main benefit is providing time series capability.
- Technical professionals often struggle to explain real-world use cases and applicability of a time-series-only database.

User Recommendations

- Choose a time series DBMS when you need high levels of performance for specific time series use cases.
- Choose a time series DBMS if your use case fits the target profile — real-time, immutable or historical analysis of append-only event data that is in time order — to add increased flexibility, performance and agility for processing the data.
- Choose a time series DBMS if it can be easily integrated into your overall data ecosystem (see The Impacts of Data Ecosystems: A Cloud Architectural Perspective).
- Choose a more standard DBMS or a newer cloud service if your use case does not have demanding requirements for ingestion or analysis. In such cases, the time series capabilities offered by a standard DBMS may be adequate for your needs.

Sample Vendors

Amazon Web Services (AWS); CrateDB; GridGain Systems; IBM; InfluxData; KX; Microsoft; Redis; ServiceNow (Era Software); Timescale

Gartner Recommended Reading

Time Series Database Architectures and Use Cases

How to Use Real-Time and Right-Time Database Analytics

Entering the Plateau

Event Stream Processing

Analysis By: Peter Krensky, Pieter den Hamer

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Event stream processing (ESP) is computing that is performed on streaming data (sequences of event objects) for the purpose of stream analytics or stream data integration. ESP is typically applied to data as it arrives (data in motion). It enables situation awareness and near-real-time responses to threats and opportunities as they emerge, or it stores data streams for use in subsequent applications.

Why This Is Important

ESP enables continuous intelligence and real-time aspects of digital business. ESP's data-in-motion architecture is a radical alternative to the conventional data-at-rest approaches that have historically dominated computing. ESP platforms have progressed from niche innovation to proven technology, and now reach into the early majority of users. ESP will reach the Plateau of Productivity in less than two years and eventually be adopted by multiple departments within every large company.

Business Impact

ESP transformed financial markets and became essential to telecommunications networks, smart electrical grids, some Internet of Things (IoT), supply chain, fleet management and other transportation operations. However, most of the growth in ESP during the next 10 years will come from areas where it is already established, especially IoT and customer engagement. Stream analytics from ESP platforms provide situation awareness through dashboards and alerts, and detect anomalies and other significant patterns.

Drivers

- Organizations have access to ever-increasing amounts of low-cost streaming data from sensors, machines, smartphones, corporate websites, transactional applications, social computing platforms, news and weather feeds, and other data brokers. Many new AI and other analytical applications need this streaming data to satisfy business requirements for situation awareness and faster, more accurate decisions.
- The wide use of Apache Kafka and similar streaming messaging systems is reducing the cost and complexity of ingesting, storing and using streaming data.
- Conventional data engineering pipelines take hours or days to prepare data for use in analytics and business intelligence (ABI), causing delays that are unacceptable for some purposes. Therefore, an increasing number of data engineering pipelines are being reimplemented as real-time data flows (continuous extraction, transformation and loading) in ESP platform products or stream data integration tools with embedded ESP. These real-time data flows filter, aggregate, enrich and perform pattern detection and other transformations on streaming data as it arrives.
- ESP products have become widely available, in part because open-source ESP technology has made it less expensive for more vendors to offer ESP. More than 30 ESP platforms or cloud ESP services are available. All software megavendors offer at least one ESP product, and numerous small-to-midsize specialists also compete in this market. Cloud ESP platforms have lowered the cost of entry.
- Vendors are embedding ESP platforms into a wide variety of other software products, including industrial IoT platforms, stream data integration tools, unified real-time platforms (aka continuous intelligence platforms), insider threat detection tools and AI operations platforms.
- Vendors are adding highly productive development tools that enable faster ESP application development. Power users can build some kinds of ESP applications via low-code techniques and off-the-shelf templates.

Obstacles

- ESP platforms are overkill for many applications that process low volumes of streaming data (that is, under 1,000 events per second), or that do not require fast response times (that is, less than a minute). Conventional ABI tools with data-at-rest architectures are appropriate for most stream analytics with these less-demanding requirements.
- Many architects and software engineers are still unfamiliar with the design techniques that enable ESP on data in motion. They are more familiar with processing data at rest in databases and other data stores, so they use those techniques by default unless business requirements force them to use ESP.
- Some streaming applications are better implemented on unified real-time platforms that process both data in motion and data at rest. Some unified platforms use embedded open-source ESP platform products, while others get their ESP capabilities from custom internal code.

User Recommendations

- Use ESP platforms when conventional data-at-rest architectures cannot process high-volume streams fast enough to meet business requirements.
- Acquire ESP functionality through a SaaS offering, an IoT platform or an off-the-shelf application that has embedded ESP logic if a product that targets specific business requirements is available.
- Use vendor-supported closed-source platforms or open-core ESP products that mix open-source with closed-source extensions for applications that need enterprise-level support. Use free, community-supported open-source ESP products if developers are familiar with open-source software and license fees are more important than staff costs.
- Use ESP platforms or stream data integration tools to ingest, filter, enrich, transform and store event streams in a file or database for later use.
- Choose a unified real-time platform with embedded ESP capabilities over a plain ESP platform if the application uses both data at rest and data in motion.

Sample Vendors

Cloud Software Group (TIBCO); Confluent; EsperTech; Google; Hazelcast; IBM; Microsoft; Oracle; SAS Institute; Software AG

Gartner Recommended Reading

Market Guide for Event Stream Processing

Infographic: Four Techniques for Real-time Analytics on Streaming Data

5 Essential Practices for Real-Time Analytics

How to Use Real-Time and Right-Time Database Analytics

In-DBMS Analytics

Analysis By: Henry Cook

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

In-DBMS analytics is also known as in-database analytics or in-database processing. This approach pushes data-intensive processing — such as data preparation, online analytical processing, predictive modeling, operations and model scoring — down into the DBMS platform, close to the data, in order to reduce data movement and support rapid analysis.

Why This Is Important

In-DBMS analytics enables agility and productivity. It is a robust way of getting machine learning (ML) and other analytics results into production. It can also have performance benefits. This technology is relatively new to cloud DBMS systems. Use of cloud in-DBMS analytics will likely increase.

Business Impact

In-DBMS analytics provides a robust way of developing advanced analytics, such as ML. It is an ideal vehicle for moving ML models into production and monitoring their effectiveness. This technology not only provides productivity and agility benefits, but also enables ML to be productionized more readily.

Drivers

- Three forces are encouraging adoption: the drive for greater productivity in the use of ML, the demand for easier ML administration and maintenance, and the need to reliably move ML into production. Adopting in-DBMS analytics provides a very good solution for moving analytics models into production — with model generation, administration and execution all in the same environment.
- In-DBMS analytics eliminates the need for data copies and additional processing elements because those aspects can be managed within the database. Thus, in-DBMS analytics is cost-efficient from a data-consumption standpoint. In-DBMS analytics offerings have been available from on-premises data warehouse software vendors for many years. They are now increasingly featured in cloud DBMS offerings, which are gaining acceptance and adoption.
- By avoiding the need to move data out of the DBMS in order to build analytical models, in-DBMS analytics enables more flexible experimentation and efficient development.
- Most DBMS vendors are offering in-DBMS analytics capabilities. The means of doing this are well-understood. Algorithms can be embedded within the DBMS, invoked using external libraries or executed by integration with a separate data science service. In some DBMSs, the same facility also enables processing, such as load and transformation logic for ELT to assist with data preparation.
- ML is becoming more commoditized as its use spreads beyond specialist data scientists. In-DBMS ML is an excellent enabler for this wider group of developers and users.

Obstacles

Data scientists lack familiarity with DBMSs and tend to prefer R, Python and notebooks. DBMS professionals lack familiarity with ML and tend to prefer SQL. Thus, in many cases, lack of adoption may be due to development culture rather than technology issues or lack of potential benefit.

- In-DBMS analytics requires a sufficient range of analytical algorithms, and this is becoming the norm. However, organizations still need to validate how in-DBMS analytics will fit into their overall estate and, most importantly, how the analytics will be monitored and controlled.
- Some implementations have performance and scalability restrictions.
- The market is heading toward open standards/specs and composable architectures like lakehouse. New algorithms are being invented, ML users might feel constrained if they feel they have to choose from a predetermined list.
- Most ML capabilities are still handled through services outside the DBMS. Cloud providers can potentially charge more for separate services.

User Recommendations

- Use in-DBMS analytics for making large-scale business analytics available to a wider audience, and for embedding ML capabilities in applications to deliver rapid insights on historical and incoming data.
- Review your data science development process. It may be better enabled through in-DBMS analytics, especially for deployment to production.
- Check whether in-DBMS analytics is supported when you’re evaluating DBMS systems. If so, verify the range of algorithms offered. Experiment with use cases where it is more efficient to bring ML algorithms to the data at scale.
- Be aware that support for programming languages, such as Python, Scala, Java and JavaScript, varies among products. Before committing to a DBMS brand, be sure it supports the languages you need.

Sample Vendors

Amazon Web Services (AWS); Fuzzy Logix; Google; IBM; OpenText (Micro Focus); Oracle; SAP; Snowflake, Teradata; VMware

Gartner Recommended Reading

Tool: Vendor Identification for Data Science and Machine Learning Platforms

Magic Quadrant for Cloud Database Management Systems

Appendixes

See the previous Hype Cycle: Hype Cycle for Data Management, 2023

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

Enlarge Table

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation’s applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology’s target audience has adopted or is adopting the technology as it enters this phase.

Source: Gartner (July 2024)

Table 3: Benefit Ratings

Enlarge Table

Benefit Rating	Definition
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2024)

Table 4: Maturity Levels

Enlarge Table

Maturity Levels	Status	Products/Vendors
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments	Maintenance revenue focus

Source: Gartner (July 2024)

© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's research organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner research may address legal and financial issues, Gartner does not provide legal or investment advice and its research should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its research is produced independently by its research organization without input or influence from any third party. For further information, see "Guiding Principles on Independence and Objectivity." Gartner research may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

© 2024 Gartner, Inc. and/or its Affiliates. All Rights Reserved.