Hype Cycle for Data Management, 2020

Published: 15 July 2020 **ID:** G00450207

Analyst(s): Donald Feinberg, Robert Thanaraj

This Hype Cycle will help data and analytics leaders interested in data management solutions to understand the evolutionary pace of maturing and emerging data management technologies. Most technologies have passed the Peak of Inflated Expectations, while many are approaching or are on the plateau.

Table of Contents

List of Tables

List of Figures

Analysis

What You Need to Know

Increasingly, data and analytics leaders have to cope with the requirements of digital business, the impact of cloud as a platform and ecosystem complexity. Data is becoming increasingly distributed across multiple systems, especially in a multicloud and intercloud architecture. As a result, the technologies that support these ecosystems, and the "connect" versus "collect" architectures supporting information, must evolve. Data and analytics leaders (including chief data officers [CDOs] and other senior roles) can use this Hype Cycle to identify technologies that are less mature but offer significant differentiation (some first movers), and technologies that are now mature that can be evaluated for broader adoption. Although there are only a few new technologies entering the Hype Cycle, there is no lack of innovation in the more mature technologies approaching the Plateau of Productivity.

To succeed with data and analytics initiatives, enterprises must develop a holistic view of critical technology capabilities. There are five Hype Cycles for 2020 that cover the disciplines, practices and technologies for data and analytics. Together, they contain the necessary elements for data and analytics leaders to form this holistic view.

Hype Cycles covering data and analytics:

- "Hype Cycle for Analytics and Business Intelligence, 2020"
- "Hype Cycle for Data Science and Machine Learning, 2020"
- "Hype Cycle for Enterprise Information Management, 2020"
- "Hype Cycle for Data and Analytics Governance and Master Data Management, 2020"
- "Hype Cycle for Data Management, 2020"

As organizations make the move to digital business, they must evaluate several major trends:

- **Cloud and platform as a service (PaaS).** As a deployment model, the cloud has become the default option for all data management technologies. This includes not only pure cloud

deployments but also a hybrid approach, embracing both on-premises and cloud together in a single, cohesive strategy (see “The Future of the DBMS Market Is Cloud”).

- **Metadata.** Metadata tools are critical for both connect and collect data architectures, as well as to support evolving demands for data security. Data catalogs must not only describe data but also include location information that is necessary for the connection and protection of data from disparate sources (see “The State of Metadata Management”).
- **Machine learning (ML).** ML is increasingly being used in data management for increasing the automation of tools, and for the optimization, modeling and execution of routine tasks. Combined, ML and artificial intelligence (AI) are the functionality added to data management for augmented data management, across all tools.
- **Internet of Things (IoT).** As the IoT is becoming ubiquitous throughout organizations, technologies supporting the access to and management of IoT data — not only in IT but also all lines of business — are becoming a priority. The IoT will require new language, policies and approaches to achieve integration of this data — otherwise, a new series of silos will be created (see “10 Critical Data Management Implications for Your Internet of Things Initiative”).
- **Data integration.** Data integration is expanding beyond traditional tools, with technologies such as data hubs, data virtualization and event streams diversifying data integration, allowing both the connection and collection of data (see “Data Hubs: Understanding the Types, Characteristics and Use Cases”).

The use of diverse data management infrastructure technologies, including both hardware and software (e.g., software in silicon, nonvolatile memory and use of graphics processing units [GPUs]), is necessary to support the digital enterprise and the modernization of data management infrastructure (see “Data Management Solutions Primer for 2020”).

The Hype Cycle

The Hype Cycle for data management covers the broad aspects and technologies that describe, organize, integrate, share and govern data. During the past few years, we have seen many new technologies introduced, including augmented data management, data fabric, cloud-based technologies, augmented data management and time series. Technologies represented here for managing new processing models are from both new and mature suppliers.

This year, it is notable that seven technologies are either on the Innovation Trigger or moving toward the Peak of Inflated Expectations (up from six in 2019): DataOps, ledger DBMS, augmented data quality, cloud data ecosystems, private cloud dbPaaS, distributed transactional databases, and data fabric. This aligns with several of our Magic Quadrants:

- “Magic Quadrant for Operational Database Management Systems”
- “Magic Quadrant for Data Management Solutions for Analytics”
- “Magic Quadrant for Data Integration Tools”
- “Magic Quadrant for Metadata Management Solutions”

- “Magic Quadrant for Data Quality Solutions”

Most of these Magic Quadrants continue to show a shift from vision to execution, in line with digital business in general moving from innovation to execution at scale. However, we are beginning to see the focus on execution revert toward a more balanced approach, shifting from execution at scale to a balance of execution and vision. This shift is enabled by most data management vendors shifting to a cloud-first focus, allowing for more rapid deployment of innovations. Examples include: augmented data management (across all data management technologies), data fabric, cloud data ecosystems and ledger DBMS.

We have three new technologies this year: database audit and protection, cloud data ecosystems and distributed transactional databases. Of these, cloud data ecosystems is of particular interest. This architecture is driven by enterprises’ desire to move from “some assembly required” to an integrated set of cloud services (see “Cloud Data Ecosystems Emerge as the New Data and Analytics Battleground”). The following innovation profiles are new to the 2020 data management Hype Cycle:

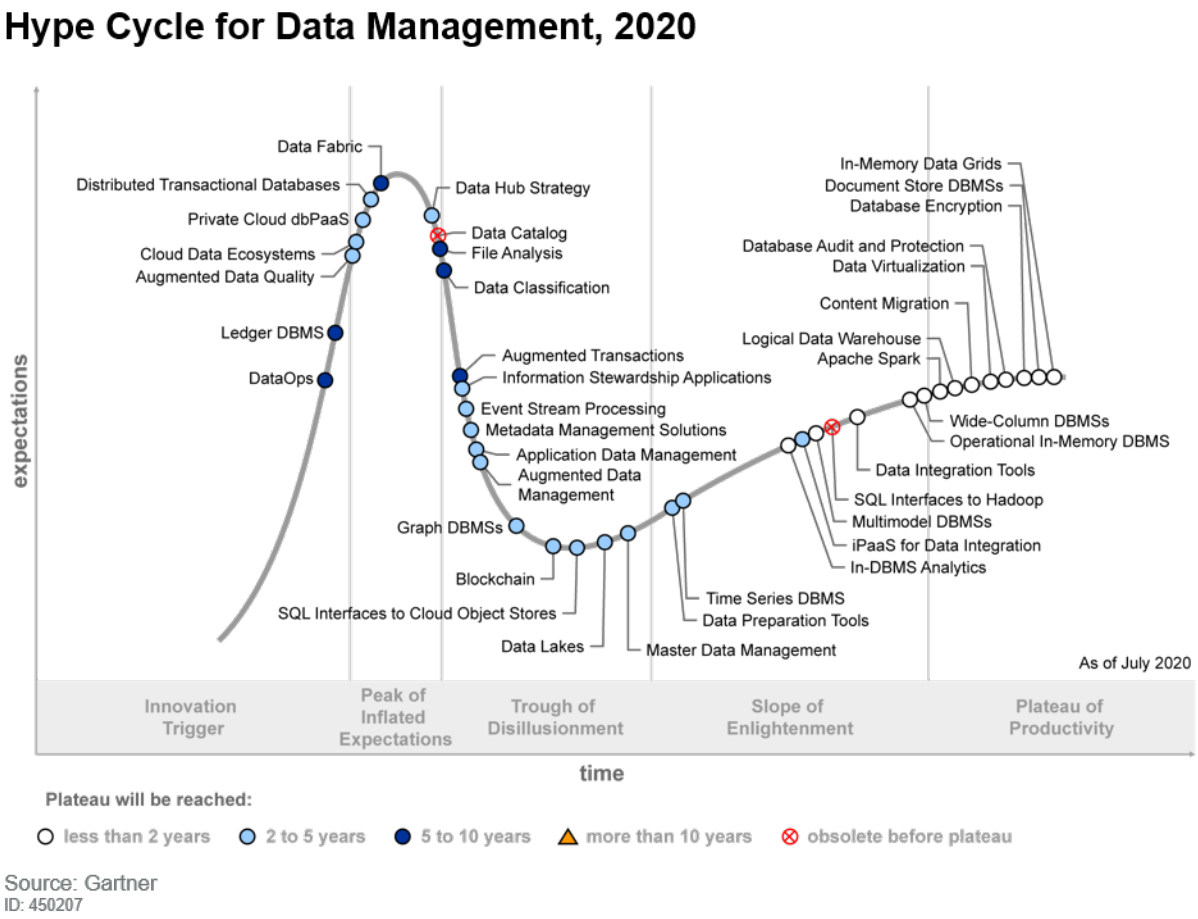
- Database Audit and Protection
- Cloud Data Ecosystems
- Distributed Transactional Databases

Organizations seeking new assets for specific purposes, such as supporting a variety of data types in increasing volumes and the IoT, can assess the relative maturity and availability of the products and technologies. Multimodel DBMSs are challenging the other nonrelational DBMSs and have the potential to support both relational and nonrelational use cases, while reducing the number of disparate DBMS products in an organization.

We also see many technologies moving to or along the Plateau of Productivity, such as multimodel DBMSs, in-memory data grids, document store DBMS, logical data warehouse and data virtualization. This indicates strong adoption and the likelihood that they will soon be off the Hype Cycle.

Finally, technologies at or near the Peak of Inflated Expectations, such as data fabric, cloud data ecosystems, private cloud dbPaaS and distributed transactional databases, are poised to begin their descent into the trough. We have already seen this happen with blockchain, augmented data management and metadata management systems, as appropriate use cases emerged and the technologies matured. Technologies at the peak are generally disruptive or are modern replacements for older technologies that have not yet gone through the rigors of reality, or the trough. Examples include metadata management solutions replacing the tool-specific catalogs, and SQL interfaces to cloud object stores replacing Hadoop distributions.

Figure 1. Hype Cycle for Data Management, 2020



Source: Gartner
ID: 450207

The Priority Matrix

For several years, most data management technologies have remained in the two- to five-year time frame for moving through the Hype Cycle, and 2020 is no different. This time frame is the most important on the Hype Cycle, balancing risk and reward over penetration in the target audience. During the next two to five years, we believe that many of these technologies will move onto, or get close to, the Plateau of Productivity, as can be seen by the number of technologies approaching or in the trough. Market adoption continues to increase, and organizations should be evaluating these resources for cost-efficiency, infrastructure simplification and new use cases, such as augmented transactions. Also, as the DBMS expands its capabilities beyond managing databases, so it becomes an information management platform — integrating with other functionality to support governance, metadata management, integration, and more.

Technologies in the transformational category offer a high risk-to-reward benefit. Although some, such as data fabric and blockchain, are two to five years out (or more), they may offer immediate value to organizations willing to accept the risk associated with their lesser maturity. Generally in

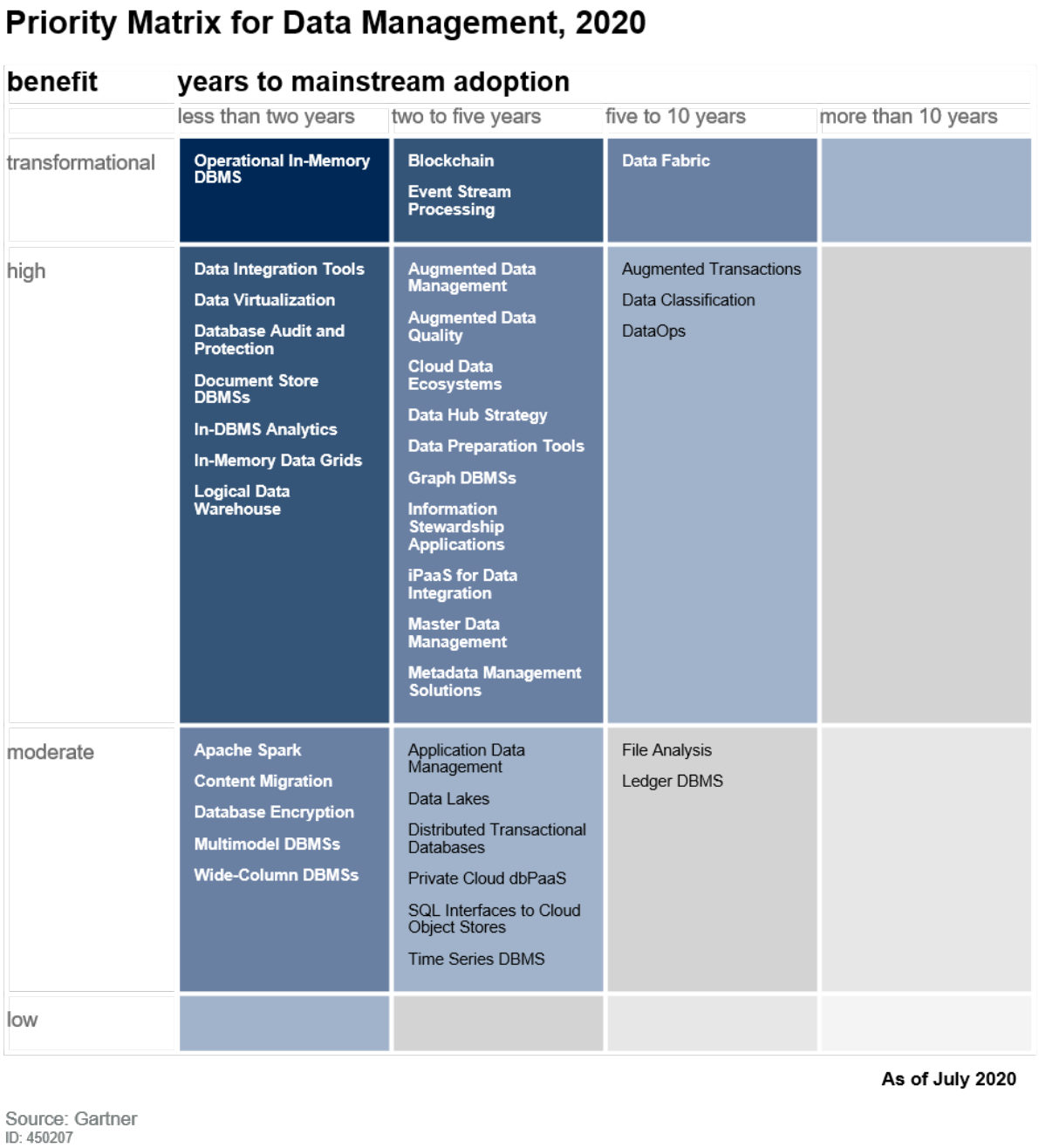
these markets, it takes at least two to five years for technologies to become more widely adopted, mature and show value.

Data management continues to be the central theme in the move toward digital business and the cloud. As requirements change within the architecture of an organization, a greater demand is placed on its underlying technology. Examples include cloud deployment and migration, nonrelational DBMS (especially time series and graph), and the use of ML (augmented data management) across all data management. Many of these technologies are rapidly advancing in capabilities and maturity, causing them to accelerate toward the plateau. Technologies can descend rapidly into the trough but then rise gradually to the plateau, normally due to market penetration. From a data management perspective, one of the most significant effects of digital business maturity is the demand for a combined data store and data integration style of data management. Moving data from Point A to Point B is not always the answer. The demand for balancing data access in place with data duplication approaches has increased to a level never before seen.

As vendors move to a cloud-first or cloud-only delivery model, many technologies are advancing rapidly and becoming pervasive, partially due to continuous innovation in the cloud. Yet even in cloud-only scenarios, the potential for multicloud providers raises an even more complex set of data management requirements, especially around integration and financial governance. Because these are really more delivery platforms than technologies, they can move rapidly to and off the plateau.

Some of the slower-moving technologies, such as augmented transactions, will require significant changes in the way applications are designed. Although they may take longer to be adopted, they *will* arrive at the plateau rather than becoming obsolete beforehand. Others, such as SQL interfaces to cloud object stores, are moving rapidly to the plateau, again due to cloud adoption.

Figure 2. Priority Matrix for Data Management, 2020



Source: Gartner
ID: 450207

Off the Hype Cycle

The following profile has been removed from the Hype Cycle as it reached the plateau:

- Analytical in-memory DBMS

The following innovation profiles have changed as stated:

- Data preparation has been renamed to data preparation tools.
- Machine learning-enabled data quality is now called augmented data quality.
- Distributed ledgers has been dropped and is now covered in blockchain.
- Apache has been renamed to Apache Spark

On the Rise

DataOps

Analysis By: Nick Heudecker; Ted Friedman; Alan Dayley

Definition: DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization. The goal of DataOps is to create predictable delivery and change management of data, data models and related artifacts. DataOps uses technology to orchestrate and automate data delivery with the appropriate levels of security, quality and metadata to improve the use and value of data in a dynamic environment.

Position and Adoption Speed Justification: The concept of DataOps hasn't changed much since it was introduced to the Hype Cycle in 2018. Best practices have not yet been developed and the market hype is almost entirely generated by a diverse array of technology and service providers applying the term to existing products. DataOps has not yet evolved as a practice or discipline that an organization can identify as providing value or improving delivery.

Early conversations with Gartner's end-user clients indicate an interest in DataOps, but it's mostly driven by an increasing awareness of broken processes and procedures. However, clients can use the term to start driving conversations around collaboration, education and tooling evolution. Because DataOps is new and still largely undefined, organizations can engage in parts of the DataOps concept to influence change around data and analytics processes without expending material budget.

DataOps as a concept has rapidly garnered awareness within the vendor space. While the hype is a long way from peaking, Gartner expects vendors to begin to use this term to denote an entire category of tools encapsulating self-service data preparation, streaming data integration and, likely, traditional data integration. Some early vendors are already doing this. Add in some level of orchestration, automation or machine learning, and end users can expect a massive amount of hype, competitive messaging and overpromises, all in the near term. The net result will be the dilution of DataOps as a concept. However, DataOps is a practice, not a technology or tool; you cannot buy DataOps as an application. It's a cultural change that is supported by tooling, and many of your existing tools may be adequate to bring in the required levels of automation demanded by DataOps.

Much like DevOps, DataOps is not some rigid dogma. Instead, DataOps is a principles-based practice influencing how data can be provided and updated to meet the dynamic needs of your organization's data consumers.

User Advice: As a new practice, DataOps will be most successful on projects targeting a small scope with some level of executive sponsorship, primarily from the CDO or other top data and analytics leaders. Executive sponsorship will be key as DataOps represents a new way of servicing data consumers. Practitioners will have to overcome the resistance to change existing practices as this concept is introduced. They will also have to exploit other emerging practices, like data literacy, to deliver on the promise of DataOps.

Data and analytics leaders focused on modernizing their data management strategies and solutions should:

- Enable greater reliability, adaptability and speed by leveraging techniques from agile application development and deployment (DevOps) in your data and analytics work.
- Enable collaboration across key roles (DBAs, data engineers, integration architects, data stewards, etc.) by including them in a common process, providing an infrastructure for shared metadata, establishing/formalizing an "operations focused" role, and providing channels for regular communication and feedback.
- Begin by introducing these capabilities with a focus on requirements definition, development and monitoring — these are the activities where collaboration, communication and consistency are most relevant and important.
- Maximize the chances of successful introduction of these approaches by choosing data and analytics projects that are struggling due to lack of collaboration or are overburdened by the pace of change — these create the best opportunity to show value.

Business Impact: DataOps focuses on changing the organizational speed in delivering data management and integration solutions to the enterprise. Shifting to continuous and reliable delivery of data will impact the speed and effectiveness of analytics and BI initiatives, continuous intelligence efforts, operational data uses, and data provided for machine learning applications, among other data-reliant activities. The focus on automating and providing reliable data pipelines should reduce risk and increase opportunities for data to be used within organizations. This will also improve the productivity of data engineers facing repetitive data tasks.

Given the early state of DataOps, it is difficult to determine if the benefit will be high or transformational. At this point, the authors believe the benefit will at least be high for most businesses, and very well may become transformational over time. DataOps will, however, be delivered deep within the overall infrastructure data practices and its success will be measured by it being less and less visible.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: Composable Analytics; DataKitchen; Delphix; Hitachi Vantara; IBM; Informatica; Nexla; Saagie; Unravel

Recommended Reading:

"Introducing DataOps Into Your Data Management Discipline"

"Accelerate Your Machine Learning and Artificial Intelligence Journey Using These DevOps Best Practices"

"Innovation Insight for DataOps"

Ledger DBMS

Analysis By: Donald Feinberg; Nick Heudecker

Definition: A ledger DBMS is an append-only, immutable DBMS with an embedded cryptographically verifiable audit trail. A ledger DBMS is useful for private and permissioned "blockchainlike" applications where distributed consensus is not required and one entity has control over the ledger and designates which parties, other than the owner, have access to and may add to it.

Position and Adoption Speed Justification: Ledger DBMSs provide many of the benefits of a blockchain platform, like data tampering detection and auditing without the complexity of configuring and managing decentralized environment. They are managed by a single entity, which makes their implementation and management far easier and more secure. Despite their utility, adoption of ledger DBMSs is evolving slowly as the benefits and potential use cases are being understood by the broader market. This technology has yet to reach the Peak of Inflated Expectations. It will take demonstrable market successes and additional product introductions to drive hype.

Most ledger DBMSs are new and relatively immature. The Amazon Quantum Ledger Database (QLDB) is an exception, because the service has been used internally by Amazon for many years (see "Amazon QLDB Challenges Permissioned Blockchains"). During the next few years, we believe there will be a number of new ledger DBMS products (especially for dbPaaS), which will increase the choices available. Thus far, Oracle has announced support for a feature it calls "blockchain tables" in version 20c of its eponymous DBMS.

User Advice: Review your organization's tactical and strategic requirements and be cognizant of the benefits and challenges of centralized, decentralized and distributed systems, so that you select the most business-appropriate platforms. Compare the suitability of ledger DBMSs against permissioned blockchain technologies by carefully considering your need for:

- A centrally administered data-auditing capability versus a decentralized network of peers.
- A single organization as the "source of truth" versus multiple potential validators.

Keeping business processes, and their associated data, private to a single organization versus the value of sharing business processes executed through smart contracts.

Business Impact: Today, many blockchain projects are forced to use public blockchain technologies when a DBMS would suffice. Ledger DBMSs represent a choice that is more manageable and easier to implement. This will enable businesses to use ledger technology where immutability is required in use cases such as audit trails, data lineage, digital assets and sharing data. The same goes for IoT and other use cases requiring smart contracts for digital signing of software updates. Ledger DBMSs will also have better performance, as there is no need for massive public participation in the consensus process.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Amazon Web Services; Fauna; Fluree; Oracle

Recommended Reading: "Amazon QLDB Challenges Permissioned Blockchains"

At the Peak

Augmented Data Quality

Analysis By: Melody Chien; Ankush Jain

Definition: Augmented data quality refers to the application of AI/ML across data quality (DQ) products (e.g., profiling, matching, linking, merging, cleansing, monitoring and issue resolution) offered by vendors in their data quality solutions. Augmented data quality extends conventional data quality features to reduce manual tasks with autorecommendations on "next best actions." Augmented DQ is also complemented by NLP to understand and translate business requirement into data quality rules based on business context and definition.

Position and Adoption Speed Justification: As organizations accelerate their digital transformation and innovation initiatives and take advantage of distributed environments with the huge number of data assets (internal and external) available to their enterprise, the challenge of managing trust at scale has increasingly become a limiting factor. While the connection between the quality of data and good decision making, process efficiencies, reduced risk and increased revenue is understood, the ability of existing DQ tools and practices is rapidly diminishing.

Vendors of DQ tools have recognized this and have responded by augmenting and extending their existing data quality capabilities with AI, machine learning and other advanced technologies. For the past several years, DQ vendors have been heavily investing in AI/ML to differentiate themselves in the market. There are many data quality tasks that have been automated with ML technology, such as data matching, data transformation and enrichment, and business rule suggestions. During the COVID-19 pandemic, some of mainstream DQ vendors demonstrated this augmented data quality

capability. They responded by quickly integrating, transforming and standardizing huge volumes of data from various sources in various data types and formats (structured or unstructured) for outbreak detection and contact tracing. The market shows fast adoption of this technology, especially in public health and healthcare industries.

As these underlying technologies (AI/ML, NLP, graphic analytics, predictive analytics) mature over time and become more widely adopted, we are expecting to see a broadening out of this support to the entire spectrum of data quality tasks, for greater productivity and higher efficiency. As augmented data quality continues to expend and develop, it is possible that end users will see some occurrences of functionality conflict, as providers in adjacent markets (e.g., metadata management, and master data management) also seek to extend their capabilities with machine learning. While data quality tools have become mainstream, innovations such as data quality machine learning are emerging and we expect the hype for this to rapidly increase in the market in the immediate future.

User Advice: Organizations faced with challenges of time-consuming and manually intensive DQ processes should first assess the existing augmented DQ capabilities within their data quality solution, if they have one. Starting first with existing use cases, an assessment of how their data quality tools and practices are applied and their limitations in addressing complex business logic and workflow, and large, distributed and fast datasets (e.g., streamed data) must be understood. Then, opportunities to close this gap by leveraging the augmented DQ functionality, where it exists, must be explored. Work with vendors to explore their augmented data quality capabilities, and determine skills, processes or training required to implement the features. Depending on vendors' technology maturity, it's very likely that some degree of custom development may be required to fully leverage the features.

Some organizations may discover that their data quality tool provider does not provide augmented DQ components within their offering. In such cases, if the opportunity cost of not having augmented data quality tools is low or the business requirement is not immediate, discussions with the vendor to understand what will be offered and when it will become commercially supported should take place. However, if the opportunity cost is high and the business requirements make machine learning supported data quality improvement a high priority, immediate discussions with the vendor should be initiated and architectural options evaluated.

Business Impact: As organizations accelerate the pace of change, seek to exploit new markets or improve customer experiences, the complexity of their operations and the creation and consumption of huge and diverse datasets will increase. This means that the need for good data quality is greater than ever. In addition, growing regulatory requirements, from government and from industry, add more restriction to organizations in how to manage personal data properly. Organizations now are accountable for any personal data they are holding. So, how to incorporate regulatory requirement into the architecture of their products and services in order to comply is a big challenge. Many existing DQ tools that lack the ML-enabled features present difficulty in managing data privacy and compliance requirements at scale. Therefore, organizations that are quick to exploit augmented data quality features are likely to have a greater competitive advantage with greater automation and further insights. However, because the risk associated with adoption of

machine learning for data quality is also commensurately large, it is imperative that organizations also step up their game in the governance of their data and analytics and metadata management.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Ataccama; IBM; Informatica; MIOsoft; Oracle; Precisely; SAP; SAS; Talend

Recommended Reading:

"Magic Quadrant for Data Quality Tools"

"Critical Capabilities for Data Quality Tools"

"Build a Data Quality Operating Model to Drive Data Quality Assurance"

"Predicts 2020: Data Management Solutions"

"Modern Data and Analytics Requirements Demand a Convergence of Data Management Capabilities"

Cloud Data Ecosystems

Analysis By: Adam Ronthal

Definition: Cloud data ecosystems provide a cohesive data management environment that aims to unify exploratory or experimental workloads with production delivery of those workloads. They have a common governance and metadata management framework, unified access management, and integrate augmented data management capabilities with a set of services accessible by the business user.

Position and Adoption Speed Justification: Many data and analytics leaders report that the cloud experience today requires a significant integration effort to ensure that components all work well together. Cloud service providers (CSPs) and independent software vendors (ISVs) are starting to respond with more refined and mature cloud data ecosystems as the market moves from "some assembly required" to a more packaged platform experience. Cloud data ecosystems emerged in earnest in 2019 with products like Azure Synapse Analytics, SAP HANA Cloud Services, and the Cloudera Data Platform Public Cloud Services. They promise faster time to delivery for data and analytics initiatives via a packaged solution experience.

While cloud data ecosystems have a vision of unifying data and analytics environments (exploratory analytics and production delivery, operational systems and analytics systems) with common governance, security and metadata, today there is still significant work to be done to make this a reality. Gaps exist in data integration, data quality, metadata, and governance. These will need to be addressed either through native CSP offerings or partnerships with ISVs to fully realize the cohesive vision of the cloud data ecosystem.

User Advice: Data and analytics leaders will need to assess the maturity of these offerings, and the degree to which they deliver on the promise of a unified environment. End users report that cloud data ecosystems are still early in their maturity, with new features and capabilities emerging on a regular basis. Early adopters should be sure to:

- Assess points of integration between various components to determine how cohesive the resulting ecosystem is. A less cohesive ecosystem will require significantly more integration time and effort.
- Ensure that your cloud data ecosystem has a well-articulated path to production for the full data life cycle (from discovery to production optimized delivery).
- Define what you expect CSPs to deliver as part of the solution, and what capabilities you expect to obtain from third-party ISVs.

Business Impact: The integration of augmented data management capabilities to streamline the delivery of data and analytics to the business in a unified offering is an attractive proposition. These offerings promise to unify the exploratory world of data science with the production delivery of traditional data warehouses. They promise to unify operational and analytics systems. They promise a holistic management framework for the full data life cycle, and aim to address key operational use cases such as data integration, interenterprise data sharing, master data management, and other core data and analytics requirements. If they are able to deliver on these promises, we expect to see improvements in the delivery of data and analytics projects to the business, with shorter time frames, better quality and a far-reduced operational footprint.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon Web Services; Cloudera; Databricks; Google Cloud Platform; IBM; Microsoft Azure; Oracle; SAP

Recommended Reading:

“Cloud Data Ecosystems Emerge as the New Data and Analytics Battleground”

Private Cloud dbPaaS

Analysis By: Adam Ronthal

Definition: Private cloud database platform as a service offerings bring the self-service and scalability of public cloud dbPaaS to a private cloud infrastructure, without external exposure. They can be deployed and managed as part of an existing private cloud management framework. Private cloud dbPaaS offerings should provide similar benefits to their public cloud counterparts — a database management system or a data store engineered as a scalable, elastic, multitenant service, ideally with subscription of chargeback pricing models.

Position and Adoption Speed Justification: Private cloud database platform as a service (dbPaaS) offerings continue to emerge in the marketplace, with a number of offerings from both database management system (DBMS) and infrastructure vendors small and large. These offerings may leverage the existing container-based infrastructure common to many private cloud offerings, either as a private IaaS or as PaaS frameworks. They may also be self-contained products in an appliance form factor, such as IBM Cloud Pak for Data System and Oracle Cloud at Customer offerings, or proprietary software frameworks such as Robin Cloud Platform. Finally, they may be extensions of existing cloud service provider offerings like AWS Outposts, Microsoft Azure Stack, Alibaba Cloud Apsara Stack or Google Cloud Platform Anthos.

Private cloud dbPaaS offerings promise a marketplacelike experience for a range of DBMS offerings: commercial and open source, relational and nonrelational. Most are still maturing to offer services that go beyond self-provisioned developer environments to true production-class environments with high availability, elastic scalability and solid performance. Hybrid private/public cloud databases are still nascent or rare.

User Advice: When evaluating private cloud dbPaaS offerings, focus on the following capabilities to ensure they meet your requirements:

- **Breadth of DBMS services offered —** Not all offerings support a full range of database types (open source, proprietary, relational and nonrelational).
- **Storage model —** Especially in container-based services, a scalable, persistent data storage tier will be required to effectively use these offerings in a production capacity. Storage that is internal to the container may not meet production-level requirements.
- **Pricing model —** These services often provide pay-as-you-go pricing models. If you are setting up the service as part of a shared service environment where you need to track chargeback, make sure the product provides or enables services to do this.
- **Elastic scalability —** Make sure that the offering can scale effectively, both up and down. Doing so in an automated way is a bonus. Ensure that the elastic scalability extends to financial governance considerations.
- **Production capabilities —** Evaluate the high availability and disaster recovery capabilities to make sure they can meet your requirements.
- **Deployment model —** Ensure that the appliance versus infrastructure offerings as well as potential hybrid deployment options meet your requirements.
- **Disconnected operations —** Many of these offerings have a cloud-based control plane that is part of the management stack. If connectivity to the cloud is unreliable, ensure that the selected offering meets any requirements for disconnected operations — including length of time it can be disconnected.

Business Impact: Private cloud dbPaaS offerings can provide a cloud experience in an on-premises data center, and can play the role of a transition technology as organizations develop their long-term cloud strategy. They will appeal to those organizations that are unable or not ready to move to public cloud offerings, due to security, regulatory or other concerns.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Alibaba Cloud Apsara Stack; AWS Outposts; Google Cloud Platform (Anthos); IBM; Microsoft Azure Stack; OpenStack; Oracle Cloud at Customer; Robin; VMware

Recommended Reading:

"Rethink Your Internal Private Cloud"

"2020 Planning Guide for Cloud Computing"

"Competitive Landscape: Container Management Software, 2019"

"Market Guide for Container Management"

"Decision Point for Selecting Stateful Container Storage"

"Market Guide for Database Platform as a Service"

Distributed Transactional Databases

Analysis By: Rick Greenwald

Definition: A distributed transactional database is a database that allows for write 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 distinguishes this technology.

Position and Adoption Speed Justification: Transactional databases are widely used for operational use cases. In the event of a network interruption, a database has to choose between offer users availability and risking a loss of data integrity, or choosing to enforce data integrity and facing a loss of database availability. Distributed transactional databases provide high levels of data integrity while minimizing the loss of availability through a combination of software and hardware infrastructure.

A robust implementation of a distributed transactional database allows for transactions to be performed on any of the distributed nodes of the database instance. Recent combinations of hardware and software can now allow these implementations, with few or no limitations on user interactions.

Some distributed transactional systems do not allow complete transparency for transactional activity, requiring some compromises in design and implementation.

Prior to the introduction of these newer systems and their production readiness, most systems would not even attempt to implement these requirements. With the capabilities needed now available, organizations are more open to considering the benefits that a truly global transactional system can provide when appropriate, without excessive system development work. Consequently, we expect to see more adoption of this technology, especially for key business systems. The position on the Hype Cycle is representative of the adoption of the technology, which is driven by use cases, rather than the maturity of the technology from the leading vendors.

Geographically distributed transactions are still not requirements in many transactional systems, so these systems do not require the extra functionality offered by distributed transactional databases. Implementers can design systems and data schemas to reduce or eliminate the impact of potential losses of data integrity with distributed systems. Based on these options, the need for a distributed transactional database will not be widespread, but scenarios where this technology is required will derive significant benefits.

User Advice: Organizations looking to implement globally consistent transactions should:

- Select a distributed transactional database if your use case requires data integrity implemented over a widely distributed database instance.
- Consider the use of design compromises, rather than a distributed transactional system, if your need for distributed integrity your use case can accommodate these compromises without excessive development and maintenance.
- Ensure developers understand the consequences of lost data integrity, both near- and long-term effects.

Business Impact: Organizations who need to implement transactional applications that span geographic distances will be able, for the first time, to implement systems without compromise or work arounds.

- Systems such as global trading applications will be able to use of distributed transactional systems to expand their scope and reach, and potentially create new markets.
- Globally implemented transactional systems will allow organizations to expand their use of transactional systems and provide competitive advantage.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: CockroachDB; FaunaDB; FoundationDB; Google (Cloud Spanner); NuoDB; RavenDB

Recommended Reading: “Data Consistency Flaws Can Destroy the Value of Your Data”

Data Fabric

Analysis By: Ehtisham Zaidi; Robert Thanaraj; Mark Beyer

Definition: A data fabric is an emerging data management design concept for attaining flexible, reusable and augmented data integration pipelines, services and semantics, in support of various operational and analytics use cases delivered across multiple deployment and orchestration platforms. Data fabrics support a combination of different data integration styles and utilize active metadata, knowledge graphs, semantics and ML to augment data integration design and delivery.

Position and Adoption Speed Justification: The data fabric — as a data management design concept — is a direct response to long-standing issues now being aggravated by digital transformation. These include the multiplicity of data sources and types, the soaring data volume, the increasingly complexity of data integration and the rising demand for real-time insights. Simply put, a data fabric is a design that leverages existing tools and platforms and adds metadata sharing, metadata analysis and metadata-enabled self-healing along with orchestration and administration tools to manage the environment. As a data fabric becomes increasingly dynamic, it evolves to support automated data integration delivery. Data fabrics are almost at the Peak of Inflated Expectations due to the hype in the market and the inherent confusion on how to deliver these. A data fabric is not in itself a tool/platform that can be purchased — it is a design concept that requires a combination of tools, processes and skill sets to deliver. Yet, we witness various tools being developed and sold under the data fabric tag which do not provision all the requirements needed to fulfill a data fabric. Not least the ability to integrate existing data integration technologies together to deliver a dynamic data integration design that uses active metadata to auto-adjust to new use-case requirements.

Data fabrics will, at the very least, need to collect all forms of metadata (not just technical metadata) and then perform machine learning over this metadata to provide recommendations for integration design and delivery. This capability is typically achieved through the augmented data catalog capabilities of a data fabric. Advanced data fabrics have the capability to assist with graph data modeling capabilities (which is useful to preserve the context of the data along with its complex relationships), and allow the business to enrich the models with agreed upon semantics. Some data fabrics come embedded with capabilities to create knowledge graphs of linked data and use ML algorithms to provide actionable recommendations and insights to developers and consumers of data. Finally, data fabrics provide capabilities to deliver integrated data through flexible data delivery styles such as data virtualization and/or a combination of APIs and microservices (and not just ETL). These are capabilities that together make up a data fabric and will mature over time as more vendors move away from point-to-point and static data integration designs and adopt more dynamic data fabrics.

User Advice: Data and analytics leaders looking to modernize their data management solutions must:

- Invest in augmented data catalogs. These will help you to inventory all types of metadata — along with their associated relationships — in a flexible data model. Enrich the model through semantics and ontologies that make it easier for the business to understand the model and contribute to it.

- Combine different data integration styles to incorporate a portfolio-based approach into the data integration strategy (for example, not just ETL, but a combination of ETL with data virtualization).
- Establish a technology base for the data fabric and identify the core capabilities required before making further purchases. Start by evaluating your current tools (such as data catalogs, data integration, data virtualization, semantic technology and DBMSs) to identify the existing or missing capabilities.
- Invest in data management vendors which exhibit a strong roadmap on augmented capabilities, i.e., embedded ML algorithms that can utilize metadata and provide actionable recommendations to inform and automate parts of data integration design and delivery.

Business Impact: By leveraging the data fabric design, data and analytics leaders can establish a more scalable data integration infrastructure that can provide immediate business impact and enable new use cases, such as:

- Data fabrics provide a much needed productivity boost to data engineering teams that are struggling with tactical, mundane and often redundant tasks of creating data pipelines. Data fabrics once enabled will assist data engineering teams by providing insights on data integration design and will even automate repeatable transforms and tasks so that data engineers can focus on more strategic initiatives.
- Data fabrics also support enhanced metadata analysis to support data contextualization by adding semantic standards for context and meaning (through knowledge graph implementations). This enables business users to be more involved in the data modeling process and allows them to enrich models with agreed upon semantics.
- Over time, the graph develops as more data assets are added and can be accessed by developers and delivered to various applications as needed. This allows organizations to integrate data once and share multiple times thereby improving the productivity of data engineering teams.
- Data fabrics provide improved decisions for when to move data or access it in place. They also provide the much sought-after capability to convert self-service data preparation views into operationalized views that need physical data movement and consolidation for repeatable and optimized access (in a data store such as a data warehouse, for example).

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Cambridge Semantics; Cinchy; CluedIn; data.world; Denodo; Informatica; Semantic Web Company (PoolParty); Stardog; Talend

Recommended Reading:

“Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration”

“Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders”

“Modern Data and Analytics Requirements Demand a Convergence of Data Management Capabilities”

“Top 10 Data and Analytics Technology Trends That Will Change Your Business”

“Magic Quadrant for Data Integration Tools”

“Critical Capabilities for Data Integration Tools”

Data Hub Strategy

Analysis By: Ted Friedman; Andrew White

Definition: A data hub strategy effectively determines where, when and how data needs to be mediated and shared in the enterprise. It layers data and analytics governance requirements atop data sharing demands to establish the technology approach for enabling data flow. The strategy drives implementation of one or more data hubs that link the work of data and analytics governance and sharing. Deployment of data hubs involves various types of integration technology, governance-related tools, metadata and, possibly, data persistence capabilities.

Position and Adoption Speed Justification: A data hub is a logical architecture which enables data sharing by connecting producers of data (applications, processes and teams) with consumers of data (other applications, process and teams). Examples of common data hubs today are master data management solutions (master data hubs) and data integration hubs. Endpoints, such as business applications, data warehouses or data lakes, interact with the data hub(s), provisioning data into it or receiving data from it. The hub then provides a point of mediation and governance, and visibility to how data is flowing across the enterprise.

The position on the Hype Cycle relates to those organizations that are new to the idea; for those who have been applying these principles (perhaps without the name “data hub”), the hype is really not relevant. The hype related to data hubs has reached a peak — technology providers and practitioners of various types have a focus on this topic, but often with very different definitions, principles and goals. This is due to several reasons:

- The opportunity created from the general failure of modern data and analytics efforts to cope with high complexity across large and diverse landscapes of data, applications and processes.
- The misinformation created by vendors that sell capabilities referred to as “data hubs” or “hubs” that have little to do with this modern design pattern.
- The confusion created internally when enterprises don’t clearly understand and communicate the definition and purpose of data hubs.

A data hub does not imply a central physical repository. A hub is like a transit station on a rail network; it is not a place where all passengers converge. A hub is a small component, part of the infrastructure; it is not an endpoint like a data warehouse or data lake. Once data and analytics

teams get this point, the idea makes a lot of sense. However, vendor messaging will continue to confuse many organizations.

User Advice: Data and analytics program leaders, including chief data officers (CDOs) and information architects, should:

- In your architectures and business plans, consider all applications, databases, data warehouses and data lakes as possible endpoints. The purpose of a data hub strategy is to focus on key points where you can gain benefits by applying data and analytics governance more effectively across sets of endpoints that need to share data.
- Design a data hub strategy to understand data and analytics governance and sharing requirements, and to drive integration efforts.
- 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.
- Start by using Gartner's Adaptive Data and Analytics approach to align the governance approach to the use case. Then follow up with Gartner's Value Pyramid to align data efforts to outcomes, and Gartner's Three Rings of Information Governance to identify the data that is most frequently used or is most important with most business value.
- Iterate changes to your data hub landscape as business requirements for data and analytics governance, data sharing and data integration change; perhaps even specializing certain hubs on specific kinds of data being governed, shared and/or integrated.

There are many types of data hubs in practice. One common example is an MDM implementation, whereby master data is shared and governed through a hub. Others include application data management, customer data platforms (CDPs) and general-purpose integration hubs. This helps explain the low penetration: an explicit data hub strategy is very new and not well-penetrated even though numerous hubs themselves have likely been adopted for discrete, even siloed purposes.

Business Impact: In coordinating programs and projects, and "connecting the dots" with a hub strategy, the business benefits will tend to grow over time as more endpoints are connected to data hubs, and possibly more hubs are adopted. This is because, without a hub strategy, complexity and cost of data sharing grows exponentially — with the hub, it grows linearly. Organizations should focus on the most high-value or complex areas first in order to gain a significant business benefit impact through the deployment of the initial hub. A formal set of hubs, managing the trusted flow of data across the entire landscape of applications and warehouses and lakes, will also expose more trusted lineage information.

If there are no effective data and analytics governance, data sharing or effective integration programs in place, the benefits of starting with a data hub strategy (compared with trying to retrospectively fit one into an established environment) will be greater.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Recommended Reading:

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

"Implementing the Data Hub: Architecture and Technology Choices"

"Data Hubs: Understanding the Types, Characteristics and Use Cases"

Data Catalog

Analysis By: Guido De Simoni; Ehtisham Zaidi; Robert Thanaraj

Definition: A data catalog is a technology capability that is used to manage an inventory of heterogeneous and distributed data assets through the discovery, organization and description of the enterprise datasets. It provides context to help data architect, developers, data analysts, data engineers, data scientists, data stewards and other data consumers to locate a relevant dataset and understand what it means, in order to determine and extract business value from it.

Position and Adoption Speed Justification: While a data catalog continues to be viewed as a critical capability in broader data management and analytics solutions, point catalogs will be successful in accessing and inventorying metadata only within the context of these narrow or use-case-specific solutions. Just like the market ended up introducing data silos, there is a growing concern of introducing metadata silos due to these embedded data catalogs in broader solutions. Customers are looking for independent/stand-alone catalogs that are application neutral and more capable of cataloging data across the organizations data assets. Gartner also believes that there is still room for specialized data catalogs that are more mature in their usage of machine learning to automate parts of the data catalog implementation process. Therefore, while we do believe that many data and analytics offerings will increasingly include data cataloging capabilities, data is catalogued to achieve a business outcome. Which usually means faster data quality, faster data integration, more informed analytics, support for data stewardship, enhanced productivity, improved productivity of data lake initiatives and more recently better cloud migrations. Best-of-breed, stand-alone catalogs will become less relevant over time since the real value is not in cataloging but in what you do with the results — as in the use case. In the long term, the data catalog will be an AI-automated feature. As technology capability is still valid to state that data catalog will be obsolete before plateau but moving in the Hype Cycle in alignment to the expansion of markets and use cases requiring the capability.

User Advice: The overall complexity and sophistication of the business data environment — along with the number of datasets, their volumes and distributed nature — are rapidly overwhelming analysts. This is particularly true with the increasing need to incorporate and correlate exogenous datasets in support of innovative use cases driven by digital business and IoT.

- Data and analytics leaders should exploit this emerging category of tools or solutions that embed data catalogs as a capability or as a stand-alone offering and that are present in the market under a variety of different names.

- Evaluate and leverage catalog capabilities in existing data management tools before investing in external tools.
- Functionality requirements should be balanced with other aspects such as vendor execution and vision, service and support, requirements for information security, data and analytics governance, and total cost of ownership.
- Proceed in the knowledge that tool-specific embedded data catalogs (like those delivered as part of a Hadoop distribution, a cloud-based data lake, etc.) will improve data usability, trust and shareability only in the context of that tool.

Data catalogs are just that — catalogs. The knowledge gleaned from cataloging information assets (of all kinds) can be used in many use cases, and each use case will have other technology-enabled requirements that need to be evaluated independent of the catalog itself. As such, vendors offering "catalog" capability are not all "catalog" vendor per se, some might be focused on analytics development, some on data and analytics governance. Be aware and evaluate like vendors accordingly. Finally, give due preference to business focused interfaces for catalogs that appeal to the business teams and use embedded machine learning capabilities to rapidly simplify and (in some cases) even automate the data catalog process.

Business Impact: Data catalogs will:

- Contribute to the ability to achieve insight from critical business data that is currently difficult to integrate and analyze due to the inability of organizations to inventory and curate their distributed, heterogeneous data assets.
- Support evolving nonrelational data initiatives (including for example data lake and/or graph database initiatives) by highlighting the data that is available.
- Enhance the organization's ability to share and curate the data at its disposal across teams, functions, environments and processes.
- Coordinate and enhance data and analytics governance processes as a business-enabling capability.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Alation; Collibra; IBM; Informatica

Recommended Reading:

"Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders"

"Modern Data and Analytics Requirements Demand a Convergence of Data Management Capabilities"

"Magic Quadrant for Metadata Management Solutions"

File Analysis

Analysis By: Michael Hoeck

Definition: File analysis software analyzes, indexes, searches, tracks and reports on file metadata and file content. FA solutions are offered as both on-premises and SaaS options. FA software reports on detailed metadata and contextual information to enable better information governance, risk management and data management actions against unstructured data.

Position and Adoption Speed Justification: File analysis (FA) solutions assist organizations in managing the ever-expanding repository of unstructured "dark" data. This includes file shares, email databases, content services platforms, content collaboration platforms and cloud-based productivity platforms, such as Microsoft Office 365 and Google G Suite. The primary use cases for FA software for unstructured data environments include:

- Organizational efficiency and cost optimization
- Regulatory compliance
- Risk mitigation
- Text analytics

The desire to mitigate business risks (including security and privacy risks), identify sensitive data, optimize storage cost and implement information governance are key factors driving the adoption of FA software. The hype associated with the growing trend of privacy regulations such as GDPR and CCPA has greatly raised the interest and awareness for FA software. When exposed through the use of FA software, the potential value of contextually rich unstructured data is capturing the interest of data and analytics teams.

User Advice: Organizations should use FA software to better grasp the risk of their unstructured data footprint, including where it resides and who has access to it, and to expose another rich dataset for driving business decisions. Utilize for cleanup of old file shares containing ROT data, which can be defensively disposed of or relocated to optimize data infrastructure. Data visualization maps created by FA software can be presented to better identify the value and risk of the data. This, in turn, can enable IT, line of business (LOB) and compliance organizations to make better-informed decisions regarding classification, data governance, storage management and content migration. For example, apply to regulatory compliance projects, such as CCPA, GDPR or other privacy regulations and readily identify sensitive personal data or key corporate intellectual property data.

Business Impact: FA software reduces business risk and inefficiencies hidden in unstructured data sources often considered "dark" data. They improve management of information governance and operational efficiency practices by:

- Eliminating or quarantining of sensitive data
- Identifying access permission issues and protecting intellectual property
- Optimizing storage utilization by finding and eliminating redundant and outdated data

- Feeding data into corporate retention initiatives through the utilization of standard and custom file attributes

FA software also assists in classifying valuable business data so it can be more easily found and leveraged, as well as supporting e-discovery, data migration and analytics.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Active Navigation; Adlib; Condrey; Ground Labs; Index Engines; SailPoint; Stealthbits Technologies; Titus; Varonis; Veritas Technologies

Recommended Reading:

"Market Guide for File Analysis Software"

"Top 5 Emerging Cost Optimization Opportunities for Storage and Data Protection"

"Beyond GDPR: Five Technologies to Borrow From Security to Operationalize Privacy"

Data Classification

Analysis By: Alan Dayley; Bernard Woo; Bart Willemsen

Definition: Data classification is the process of organizing information assets using an agreed-upon categorization, taxonomy or ontology. It enables effective and efficient prioritization for data and analytics governance policies that spans value, security, access, usage, privacy, storage, ethics, quality and retention. Data classification typically results either in the development of a large repository of metadata useful for making further decisions, or applying a "tag" to a data object to facilitate use and governance of data during its life cycle.

Position and Adoption Speed Justification: There are many reasons to classify data, and the decisions about who or what system establishes and enforces classification differ accordingly. Approaches include categorization by data type, owner, regulation or classification by data sensitivity or retention requirements. Data classification momentum continues for digital business transformation, artificial intelligence/machine learning (AI/ML), security concerns and increased regulations and concerns for privacy. Data protection regulations have added needs to classify data around individuals and entities, with privacy-centric classification efforts, and as a foundation for data life cycle management. Datasets residing and migrating across on-premises and multicloud has thrust data classification into the forefront for both organizations and cloud providers.

Data risk and value assessments rely heavily on data classification. For example, security teams use classification technology to assign risk profiles to data. Likewise, the emergent field of infonomics requires classification capabilities to assign value and liability to varying datasets.

Data classification has been around for some time, just not widely adopted. The position in the Hype Cycle reflects a majority of partial or siloed adoption, and still developing capabilities.

User Advice: Data classification attempts can be difficult at best: to identify, tag and store all of an organization's data, without first taking into account the utility, value and risk of that data. To start, organizations should assess information assets for value and risk, for example using infonomics measures. The purpose is to isolate the data that has minimal or no value to the organization, optimize data management costs and address any archival, retirement, destruction, risk and security requirements. Use an ongoing, adaptive and iterative approach instead of one-off or intermittent audits, continued discovery of data assets helps to perpetuate the process and help with cross-organizational engagement. The key is to start somewhere that will have a business impact and build out the data catalog over time.

Data classification must also be an ongoing aspect of changing organizational behavior within the data-driven culture, and be supported by data governance and metadata management. SRM leaders and chief data officers (CDOs) should collaboratively architect and use classification capabilities.

Data classification recommendations include:

- Implement data classification as part of a funded data and analytics governance program.
- Determine organizationwide classification use cases and efforts, and, at a minimum, keep all stakeholders informed.
- Combine privacy regulation adherence efforts with the security classification initiatives overseen by the chief information security officer (CISO). As such, information can be categorized by nature (e.g., what is PII, PHI or PCI), or by type (e.g., contract, health record, invoice). Regardless, records should also be classified by risk categories as to indicate the need for confidentiality, integrity and availability. Finally, records can be indicated to serve specific purposes. Both CIA and purpose classifiers may change during the record life cycle.

Business Impact: Targeted classification, combined with the capabilities of various data management tools, will enable organizations to produce faster, more reliable and efficient data use for discovery, risk reduction, value assessment and analytics. This enables organizations to focus security and analytics efforts primarily on their important datasets. Identity-centric classification efforts can assist when managing concerns for the European General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA) and other privacy regulations.

Classification enables CDOs to drive information asset management as a value-adding opportunity to support better business outcomes, rather than being an approach driven by compliance and records-keeping requirements. In addition, CDOs, CISOs, compliance and other involved functions should work with each other and other pertinent functions to ensure mutual awareness of classification activities.

Data classification can be used to support a wide range of use cases:

- Privacy compliance

- Risk mitigation
- Master data and application data management
- Data stewardship
- Content and records management
- Data catalogs for operations and analytics
- Data discovery for analytics and application integration
- Efficiency and optimization of systems, including tools for individual DataOps

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Boldon James; Collibra; Dathena; IBM; Informatica; Microsoft; Netwrix; OpenText; Titus; Varonis

Recommended Reading:

"How to Overcome Pitfalls in Data Classification Initiatives"

"Ignition Guide to Data Classification"

"Building Effective Control Documents for Sensitive Data Classification and Handling"

"Market Guide for Enterprise Data Loss Prevention"

"How to Successfully Design and Implement a Data-Centric Security Architecture"

Sliding Into the Trough

Augmented Transactions

Analysis By: Donald Feinberg; Massimo Pezzini

Definition: Augmented transactions use various forms of augmented analytics — advanced analytics, artificial intelligence (AI) and machine learning (ML) — enabling concurrent analytical and transaction processing within a transaction or process. In-memory computing (IMC) technology is a key enabler for augmented transactions.

Position and Adoption Speed Justification: Augmented transactions were formerly called in-process hybrid transactional/analytical processing (HTAP), but we have broadened the definition to now include AI (including ML) with augmented analytics. The transactional and analytical processing dimensions are designed together in the context of an individual transaction or process. They are delivered as a single application, where transaction processing and analytics functions are designed

to interplay, thus also leading to a streamlining of the overall technology infrastructure. Typically, these applications include decisions based on analysis requiring analytics, AI and ML. These could be assisted by an automated recommendation based on a real-time analysis of the impact of different options on key business indicators (for example, inventory levels, profitability and customer satisfaction).

Certain types of augmented transactional applications can be implemented using traditional data management technologies; practically, however, augmented transactions often require the use of IMC technologies. Performance and scalability limitations have historically prevented advanced analytics from running concurrently with or within transaction processing. However, IMC, and specifically in-memory DBMS, makes it possible to relax the restriction of using external or after-the-fact analytics and to implement even the most sophisticated augmented transactions.

Although an increasing number of packaged application providers (such as Oracle, SAP and Workday) have products in the market with augmented transactions, adoption is still in the early stages. This is especially true with the addition of AI and ML, and is the reason why the position of this technology continues to progress slowly toward the trough. Although we are seeing an increasing number of internal applications developed on DBMSs, such as MemSQL, SAP S/4HANA remains the most widely installed packaged application with augmented transactions.

For augmented transactions, there was little movement on the Hype Cycle this year, despite its potential for dramatic business innovation. It will take more than five years for augmented transactions to reach mainstream adoption, largely due to the following factors for DBMS vendors:

- Skills for augmented transactions are still limited to the most leading-edge organizations. Best practices have not yet crystallized, and skills are hard to find.
- When retrofitting existing (custom or packaged) applications at the core of business operations, augmented transactions require massive, expensive and risky reengineering efforts.

User Advice: Application leaders responsible for modernizing application architecture, information management and business analytics strategies should:

- Pilot augmented transactions in individual “system of innovation” projects.
- Discuss with strategic information management and business application providers their vision, roadmap and technology for augmented transactions in their products.
- Use augmented transactions for use cases involving observation data (such as IoT) or interaction data (such as log data) for which real-time operational analytics is required.
- Educate business leaders about augmented transactions and IMC concepts and importance, by brainstorming with them to identify concrete opportunities to rethink business processes and create applications that could not be implemented using traditional architectures.

Application leaders should, however, plan for a long-term coexistence of augmented transactions with traditional architectures. A traditional approach often meets business requirements well, and migrating to an architecture with augmented transactions may not be justified and is often not even

desirable or possible. Moreover, it will be practically impossible to migrate many established applications toward augmented transactions — due to the massive organizational and technical efforts needed, and because of the inertia of most packaged business application vendors.

Business Impact: Augmented transactions will redefine the way some business processes are executed, in real time, assisted by analytics, AI, and ML (for example, planning, dynamic repricing, forecasting and what-if analysis). The augmented transaction becomes an integral part of the business process itself, rather than a separate activity performed after the fact.

Augmented transactions will make it possible for business leaders to perform, in the context of operational processes, much more advanced and sophisticated real-time analysis of their business data than feasible with traditional architectures. In many cases, this will also reduce the complexity of the overall architecture by eliminating unnecessary duplication of data and infrastructure. Large volumes of complex business data can be analyzed as the processes unfold, thus:

- Enabling business users to make more-informed operational and tactical decisions in real time
- Opening up the possibility of driving prescriptive decisions without human intervention
- Improving business leaders' situation awareness in operations, and providing constantly updated forecasts and simulations of future business outcomes

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Aerospike; GigaSpaces; GridGain Systems; IBM; MemSQL; Microsoft; Oracle; SAP; ScaleOut Software; VoltDB

Recommended Reading:

"Predicts 2018: In-Memory Computing Technologies Remain Pervasive as Adoption Grows"

"How to Enable Digital Business Innovation via Hybrid Transaction/Analytical Processing"

"Delivering Digital Business Value Using Practical Hybrid Transactional/Analytical Processing"

"Critical Capabilities for Operational Database Management Systems"

Information Stewardship Applications

Analysis By: Guido De Simoni; Andrew White

Definition: Information stewardship applications are solutions that support the work of information stewards. Application capabilities may include monitoring of data governance policy performance, such as data quality, access to data models, a business glossary (integrated with a data dictionary), tasks, workflow, exception management, business rules and policies, audit trails and lineage, and

analytics. Additionally, these applications may include playbooks, preloaded templates and other capabilities to help make this business role more effective.

Position and Adoption Speed Justification: Currently, information stewardship applications leverage the capabilities of more-technology-oriented solutions that are already on the market. However, they are packaged to meet the requirements of information stewards who support the work of information policy enforcement as part of their normal, business-centric, day-to-day work in a range of use cases.

Information stewardship applications are in the Trough of Disillusionment and maturing very slowly. They are supporting the evolution of, and lessons learned from, the information stewardship discipline, but they have been disrupted as a stand-alone market by several new drivers relating to the demand for data and analytics governance operationalization. The complexity that is emerging — associated with, for example, increased interest in data lakes and Internet of Things analytics — extends the need for policy enforcement in analytical use cases. The overlap between the data governance board and the analytics center of excellence, which is now discovering that it needs to comply with and respect policy set by others, has not been captured in the market. Finally, it is fast becoming clear that, as stewardship gets established, such programs are in fact initiatives focused on business process integrity and outcome improvement and much less about data for data's sake or data at a point in time. This variety of requirements affects vendors' experimentation with and assessment of information stewardship applications. In particular, we observe clear market disruptions related to the adoption of data catalogs and organizations scrambling to work within the privacy management requirements of regulations such as the EU's General Data Protection Regulation (GDPR). Even more so with the potential convergence of capabilities in the context of data and analytics governance; hence, information stewardship applications are in a state of stall on the Hype Cycle.

User Advice: COVID-19 is pushing all organizations to address understanding of curated data in a more operationalized and automated way. Data and analytics leaders should work with their technology providers to help them understand what works for information stewardship that can be made operational with appropriate technology. If you need to steward data outside a data and analytics governance program, tread more carefully, because the lack of unifying drivers for master data management (MDM), records management or enterprise content management could result in few technology choices.

Data and analytics leaders should:

- Evaluate the capabilities needed from fit-for-purpose, business-user-oriented information stewardship and other solutions, as compared with IT-centric data management tools, including data quality, metadata management and federation/integration capabilities.
- Run a proof of concept of vendor solutions involving all contributing roles, such as business users, information governance board members, information architects, information stewards and business analysts.
- Focus on all dimensions (people, process, technology and data) when addressing the information stewardship use case. These dimensions are relevant for effective use of a solution

to maximize your ROI through reuse, while also minimizing administrative costs and errors due to inconsistencies across technologies.

Business Impact: Data and analytics governance is a core component of any enterprise information management (EIM) discipline. Such governance cannot be sustained and scaled without an operational information stewardship role and function. At worst, the lack of effective stewardship, and hence ineffective governance, will lead to the failure of EIM initiatives; at best, it will result in lower-than-desired benefits. A successful stewardship routine will lead to sustainable and persistent benefits in support of programs and projects such as EIM, MDM, application data management, analytics and business intelligence. These benefits include increased revenue, lower IT and business costs, reduced cycle times (for example, for new product introductions), improved trust in organizational data and increased business agility.

In particular, the impacts of information stewardship applications include:

- Encouragement to share data, increased data reuse, improved consistency and accelerated time to value because of the use of existing data dictionaries to identify areas of synergy between data used for different business initiatives (both data content and meaning).
- More effective understanding and communication of the semantic meaning of data. This will facilitate resolution of contention between business teams when inconsistency arises and reduce the amount of time and effort wasted on reconciliation, so that efforts can focus on new business actions.
- Intelligent decisions about the information life cycle, from data interoperability and standards to archiving, disposal and deletion.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: ASG Technologies; Collibra; Global IDs; Infogix; Informatica; Prodago; SAP

Recommended Reading:

“Market Guide for Information Stewardship Applications”

“The Role of Technology in Data and Analytics Governance”

Event Stream Processing

Analysis By: W. Roy Schulte; Nick Heudecker; Pieter den Hamer

Definition: An event stream is a sequence of event objects arranged in some order, typically by time. Event stream processing (ESP) is computing that is performed on event objects for the purpose of stream data integration or stream analytics (also called complex-event processing [CEP]). ESP is typically applied to data as it arrives (data “in motion”). It provides information about

emerging threats or opportunities for near-real-time alerts, dashboards and sense-and-respond processes, or it stores data in a database for use in subsequent analytics.

Position and Adoption Speed Justification: Three factors are driving the expansion of ESP:

- The growth of the Internet of Things (IoT) and digital interactions (including clickstream data) is making event streams ubiquitous.
- Business is demanding continuous intelligence for better situation awareness and faster, more personalized decisions.
- Vendors are bringing out new and improved products, many of them open source or partly open source.

Companies have access to more streaming data from sensors, meters, control systems, corporate websites, transactional applications social computing platforms, news and weather feeds, data brokers, government agencies and business partners. ESP technology is maturing rapidly. It will eventually be adopted by multiple departments within every large company. ESP will reach the Plateau of Productivity within several years, largely by being embedded in IoT platforms, SaaS solutions and off-the-shelf packaged applications.

User Advice: Data and analytics leaders should:

- Use ESP when traditional DBMS architectures cannot execute fast enough to provide real-time information from high-volume event streams.
- Acquire ESP functionality by using a SaaS offering or an off-the-shelf application that has embedded CEP logic (but only if a product that addresses your specific business requirements is available).
- Build your own application on an ESP platform if an appropriate off-the-shelf application or SaaS offering is not available.
- Use free, community-supported, open-source ESP platforms if your developers are familiar with open-source software and languages such as Java, Scala or Python, and license fees are the primary consideration. Use vendor-supported closed-source platforms or products that mix an open-source core with value-added closed-source extensions for mainstream applications that require enterprise-level support and more complete sets of features.
- Use on-premises ESP in preference to cloud event processing services for low-latency applications such as IoT edge computing and financial trading, and for applications where most of the data originates on-premises.
- Use ESP technology that is optimized for stream data integration to ingest, filter, enrich, transform and store event streams in a file or database for later use.

Business Impact: Stream analytics provided by ESP platforms:

- Can support situation awareness through dashboards and alerts by analyzing multiple kinds of events in real-time.

- Enable smarter anomaly detection and faster responses to threats and opportunities.
- Can help shield businesspeople from data overload by eliminating irrelevant information and presenting only alerts and distilled versions of the most important information.

ESP is one of the key enablers of continuous intelligence and other aspects of digital business. It has transformed financial markets and become essential to smart electrical grids, location-based marketing, supply chain, fleet management and other transportation operations. Much of the growth in ESP usage during the next 10 years will come from three areas where it is already somewhat established: IoT, customer experience management and fraud detection applications. More than 40 ESP products or cloud event stream processing services are available on the market.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Apache Software Foundation; Confluent; Evam; IBM; Microsoft; Oracle; SAS; Software AG; TIBCO Software; Ververica

Recommended Reading:

"Market Guide for Event Stream Processing"

"Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements"

"The Five Levels of Stream Analytics — How Mature Are You?"

"Technology Insight for Event Stream Processing"

"Innovation Insight for Continuous Intelligence"

Metadata Management Solutions

Analysis By: Guido De Simoni

Definition: Metadata management solutions are software that includes one or more of the following: metadata repositories, a business glossary, data lineage, impact analysis, rule management, semantic frameworks, and metadata ingestion and translation from different data sources. Metadata describes various facets of an information asset and its role in the information architecture in support of four use cases: data governance, security and risk, analytics and data value.

Position and Adoption Speed Justification: Metadata management solutions (MMSs) are going through a major market shift. Modern MMSs are supporting organizations that manage distributed and varied information assets. Moreover, demands for accessing and using data are no longer limited to IT, and new data-oriented citizen roles are emerging in the business. In addition, data and

analytics leaders are facing greater regulatory pressures, such as privacy requirements, that force new approaches to data management.

Demand and hype for data catalog capabilities are growing, but many of these are solely limited to a single application and often primarily to an inventory of data assets. MMS vendors promise to deliver much broader metadata capabilities than catalogs and address all use cases, thus relieving the current issues emerging from metadata silos. As a result, MMSs are accelerating due to innovation generated by active metadata that leverages machine learning. Active metadata enables real-time analysis of the applicability of data, checks on the veracity of data sources used, and monitoring of the ways users act as they form dynamic communities that support outcomes from positive "tribal" behaviors while simultaneously avoiding exclusivity. In this case, we refer to how informal and formal teams emerge and slowly convert to community participation with as much automation as possible to ease and simplify this process. These demands are only now starting to be addressed by vendors, with modern metadata management practices slowly being established within organizations.

We expect MMSs to take two to five years to reach the Plateau of Productivity as the technology continues to expand in terms of both capabilities and support for all four use cases. While still emerging in many organizations, the metadata management practice continues to grow, due to ever-expanding data and information volumes, along with regulatory compliance and business requirements to catalog and manage that information. Ultimately, metadata will become a critical input to machine learning approaches for dynamic data management solutions, and metadata solutions will evolve toward a graph-based analytical approach.

User Advice: Most organizations will find their current metadata management practices differ across applications, data and technologies, and that these practices are siloed by the needs of different disciplines and even software applications — each with their own governance authority, practices and capabilities. Data and analytics leaders who have already invested in data management technologies should first evaluate the metadata management capabilities of their existing data management tools, including data integration, data quality and even master data capabilities, before buying a modern MMS. However, if dealing with emerging use cases, including data and analytics governance, security and risk, support for analytics and augmented data value, they should learn about, and build pilot implementation using MMSs. In addition, the introduction of "active metadata" concepts (to metadata in 2019, but prior to that for integration in 2017) means that some of the more basic catalog capabilities no longer differentiate solutions in the market. Data and analytics leaders must now consider using metadata from platforms, tools, third-party providers and a widely divergent range of data sources and user experiences.

Business Impact: MMSs are relevant to several of the business requirements of enterprises, such as:

- **Management of complexity:** The complexity of data management depends on the complexity of data needs arising from applications, the variety of information and a growing number of information management use cases. MMSs help to break down and reduce the complexity often inherent in data.

- **Automation of processes:** Because data is subject to change, there are numerous recurring activities that MMSs may enable or streamline by (partial) automation — for example, creation, publication, approval and revision. MMSs enable these activities and processes by the application of technology.
- **Collaboration:** At an enterprise level, metadata requires the contribution of numerous people from different divisions, countries, etc. An MMS can provide a multiuser environment able to address complex collaboration requirements. Additionally, an MMS can facilitate collaboration among data consumers and providers by enabling business-driven development of data management and its metadata.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Alation; Alex Solutions; ASG; Collibra; erwin; IBM; Infogix; Informatica; Oracle; SAP

Recommended Reading:

"Magic Quadrant for Metadata Management Solutions"

"Critical Capabilities for Metadata Management Solutions"

Application Data Management

Analysis By: Andrew White; Malcolm Hawker

Definition: Application data management (ADM) is a technology-enabled discipline where business and IT work together to ensure uniformity, accuracy, stewardship, governance, semantic consistency and accountability for data in business application or suite, such as ERP, 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, or prices.

Position and Adoption Speed Justification: Organizations have long struggled to manage data within the context of business applications, even though almost all business applications assume authority of their own data. The tools offered by application vendors and developed by app developers have rarely focused on what is needed to govern and steward application data. The focus has shifted recently to data quality and entity resolution as newer uses of graph technology, even augmented with machine learning, to improve the speed and efficacy of entity discovery and identification. As a result this innovation profile has moved back up the down-swing of the Hype Cycle being drawn by growing hype in related technologies such as graph.

With the COVID-19 reset, we anticipate a balancing of price power between rightsized MDM and ADM implementations. Organizations will realize they are different value propositions, business value, and you don't always need both at the same time.

User Advice: Starting with a focus on business outcomes to identify what data matters most, organize, classify and govern data based on which drives the most important business outcomes:

- **Master data and MDM —** The data that matters most to the most impactful business outcomes. Master data is the least number of attributes that define business entities such as customer, citizens, products, services etc., are governed in a master data hub for the broadest use across all applications and uses and are thus context-free
- **Application data and ADM —** The data that matters most to a specific set of use cases supported by one application or suite like e-commerce or customer data platform. Application data is the rest of the descriptive or reference data that describes business entities and other entities such as price, unit of measure etc. used in an application or suite for uses specific to those applications or suites and are thus context-specific.

Demand from your packaged (on-premises or cloud) application provider the necessary capability to set (that is, govern) and enforce (that is, steward) information policy pertaining to data used in the application or suite. When this is lacking, look to MDM vendors to support this capability. Note that even if you obtain such capability from your application vendor, you may still need to integrate it into your MDM hub infrastructure.

Design your overall program independent of application data management capabilities to support your enterprise application architecture and landscape. It is also possible that some MDM solutions will negate the need for an additional application data management solution. Note that some vendors do not use the term “ADM” and instead use other names that are more appropriate to the context of the user, such as customer data hub, or customer interaction hub.

Implement ADM alongside any MDM program so that they can operate at their own speed and benefit. They do align and share metadata in support of a wider EIM program.

Business Impact: The primary benefactors of this discipline are business users, as in material planners, production planners and customer service reps., or marketers, but not IT users. Business users will finally be able to steward and govern the application data needed within their specific business application or suite. Thus, this is a good first step in support of a widening of your MDM program, since ADM can now be coordinated with the very same governance and stewardship work that is part of an MDM program. If you don’t have an MDM program yet, you can still adopt ADM for each application, but your integration challenges with the shared (application or master) data will likely persist.

Many users of large packaged or industry-vertical applications believed that these applications or suites already helped them do a good job of managing the data used in them. They might, therefore, be shocked to find their strategic vendor partners developing solutions that accomplish what they thought the application package had been doing all along. Most often lacking is the capability’s focus on governance and stewardship of the business rules, workflows and metrics reporting on data consistency across the application for the entirety of the data life cycle. However, the need to manage this data formally has emerged only recently due to the increasing complexity of application environments (even those labeled ERP) and the growing need to ensure a trusted view for data across organizations.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Chain-Sys; Epicor Software; Oracle; PiLog; SAP; Tealium; Utopia Group; Winshuttle (EnterWorks)

Recommended Reading:

"Design an Effective Information Governance Strategy"

"Toolkit: How to Classify Information Assets to Be Governed in Applications"

"Use the 7 Building Blocks of MDM to Achieve Success in the Digital Age"

"The Role of Technology in Data and Analytics Governance"

Augmented Data Management

Analysis By: Donald Feinberg; Merv Adrian; Ehtisham Zaidi

Definition: Augmented data management refers to the application of AI and ML for optimization and improved operations. AI and ML are applied — based on the existing usage data — to tune operations and to optimize configuration, security and performance. They are also applied to create, manage and apply policy rules within the different products, such as metadata management, master data management, data integration, data quality and database management systems.

Position and Adoption Speed Justification: Artificial intelligence (AI) and machine learning (ML) can automate capabilities, altering job roles, product design and overall data management processes. Cost-based query optimization based on the collection of operational and usage statistics across large numbers of deployed instances, especially in the cloud, has accelerated the pace of vendor delivery. 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 suggest and implement new designs, schemas and queries. They can even infer the semantics and associations of the data in order to recommend structural improvements.

The adoption of these capabilities is gated by the movement of the product categories to the cloud, where they are delivered first. Offerings in data integration, data quality, master data management (MDM), metadata management and DBMS software are proliferating and maturing rapidly. This profile therefore represents an aggregate view and position of what is happening across data management. Its slow movement on the Hype Cycle (despite ML's longevity in the DBMS) is due to the numerous use cases in addition to query optimization. Products such as Amazon Aurora and Oracle Autonomous Database, and data management software that has made the transition to the cloud, are steadily gaining more users. In these platforms, enormous volumes of user data on a consistent infrastructure improve the applicability of the results and offer opportunities for the continuous training and retraining of models. As a result, they are being aggressively used to drive

competitive improvements and some of the features are making their way into on-premises, private cloud deployments as well. We believe the improvements, initially available in cloud platforms, will neutralize the distinction between cloud and on-premises over the next few years.

User Advice: For data and analytics leaders focused on data management capabilities, we recommend you:

- Question the vendors of your data management tools about their roadmap for the introduction of AI and ML into their 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.
- Create a business case for using these new tools, and be sure to model and measure the benefits realized from the resources that will be released for other functions of greater business value.
- Plan for roles to change: Provide new skills training to add value as responsibilities evolve.
- Make augmented capabilities a “must have” selection criterion for new purchases of data management products.
- Begin seeking data management solutions that share design and performance metadata for use.

Business Impact: Augmented data management will offer benefits in the following areas:

- 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.
- Data integration — To automate the integration development process, by recommending or deploying repetitive integration flows.
- MDM — MDM solution vendors will increasingly focus on offering AI- and ML-driven configuration and optimization of record matching and merging algorithms as a part of their information quality and semantics capabilities.
- Data quality — AI and ML will be used to extend profiling, cleansing, linking, identifying and semantically reconciling master data in different data sources, to create and maintain “golden records.”
- DBMS — In addition to enhancing 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.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Cinchy; CluedIn; IBM; Informatica; Microsoft; Oracle; SAP; SnapLogic; Teradata

Recommended Reading:

"Cool Vendors in Data Management"

"Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration"

"Market Guide for Data Preparation Tools"

"Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders"

"Automating Data Warehouse Development"

Graph DBMSs

Analysis By: Mark Beyer; Henry Cook; Nick Heudecker

Definition: Graph DBMSs represent relationships by separating storing data elements and their relationships to one another. After storing the data, processes such as complex traversal operation of the data network are difficult to do at scale with traditional RDBMSs. Most graph DBMSs use basic graph theory and are suitable for general-purpose uses, such as processing the complex many-to-many connections found in social networks. Other systems use triplets or network databases for more-specialized applications.

Position and Adoption Speed Justification: Graph DBMSs are entering the Trough of Disillusionment and may languish in the trough for longer than previously expected. Users have realized that modeling, loading, processing and analyzing graph data requires skills. Graph analysis enables multiple, simultaneous models in parallel over the same data, which describe dynamic and complex domains more naturally and efficiently than is possible in JOIN-based processing, thus avoiding the performance and complexity challenges using JOIN-only logic. More user-friendly interfaces are required due to the difficulty of evaluating graphs versus tables. Use cases to exploit graph data models and graph analytics were starting to become visible. But, with the slow adoption of better tools, as well as an inability of practitioners to adequately explain the benefits of graph, many of the use cases are now moving to embedded functionality in other tools (AI, metadata, data integration lead the way). Gartner's inquiries through May 2020 were up 40% over 2019, but mostly reflected the difficulties in applying graph.

New vendors continue to enter the market as cloud providers offer solutions and highly mature DBMS products often embed graph DBMS to augment traditional relational database functionality. A number of vendors in the advanced analytics platforms, metadata management and metadata analytics space have introduced graph DBMSs as back ends, and there has been an expansion of knowledge graph use cases in the market due to the increasing importance of semantics. Graph DBMSs will increasingly be used to navigate both existing and newly discovered relationships more efficiently than relational processing over the next two to three years.

The hype around graph DBMSs has revolved around ad hoc discovery of relationships, but the majority of graph use cases remain predominantly for relationships already defined. Graph support for supply chain, transportation/logistics and epidemiology is now being reinforced, in addition to the early adoption in financial fraud detection, telecommunications network analysis and master data management. Graph support of metadata use cases is becoming increasingly important. Its availability in open-source versions and on cloud platforms continues to drive experimentation. Graph capabilities have been introduced as the first additional option in many new multimodal DBMS offerings. Finally, there are correlations across multiple sensors in asset-intensive Internet of Things use cases and scenarios, where write capability — even ACID-compliant write is necessary — continues to fuel increasing marketplace interest. It is imperative to remember that graph DBMS does not replace relational DBMS.

User Advice: Data and analytics leaders should:

- Assess graph DBMSs' capabilities when RDBMS performance requirements for highly nested or relational data fall outside current processing capabilities. The core advantage of graph DBMSs is the relationships they support — and increasingly discover — in data, which they then persist in models.
- Use open-source graph DBMS projects or community editions of commercial ones to experiment and gain experience.
- Ensure that any open-source products you use in production environments are commercially supported. Graph DBMSs can render the relationships and traversal of data (as discovered by a data scientist or data science team) into a reusable form for data miners, data engineers and business analysts. Scaling and reliability are another matter.

Business Impact: Graph analytics will benefit from embedded graph DBMS solutions. Open-source projects continue to gain traction and compete with proprietary offerings for first projects, as open-source options and standards may reduce the need for multiple proprietary skill sets. Also, cloud-first provides a good experimental approach for graph databases. The overall impact of graph DBMSs is high, but skills in graph modeling are not widely available. Graph DBMSs represent a radical alternative to how information is organized and used; this change will see slow adoption until additional industry-specific use cases emerge, the skills shortage abates and newer tools gain acceptance. Semantic mapping is an example; graph technology is already used for data hub use cases and inside some commercial software offerings for metadata management.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services (AWS); Cambridge Semantics; DataStax; MarkLogic; Microsoft; Neo4j; Oracle; TigerGraph

Recommended Reading:

“COVID-19 Demands Urgent Use of Graph Data Management and Analytics”

“An Introduction to Graph Data Stores and Applicable Use Cases”

“Innovation Insight for Extended Detection and Response”

“Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration”

Blockchain

Analysis By: David Furlonger; Rajesh Kandaswamy; Christophe Uzureau

Definition: A blockchain is an expanding list of cryptographically signed, irrevocable blocks of records shared by all participants in a peer-to-peer (P2P) network. Each block of records is time stamped and references links to previous data blocks. Anyone with access rights can historically trace a state change in data or an event belonging to any participant. Distributed ledgers are design limited and lack decentralized and tokenized elements.

Position and Adoption Speed Justification: A blockchain is an expanding list of cryptographically signed, irrevocable blocks of records shared by all participants in a peer-to-peer (P2P) network. Each block of records is time stamped and references links to previous data blocks. Anyone with access rights can historically trace a state change in data or an event belonging to any participant. Distributed ledgers are design limited and lack decentralized and tokenized elements.

User Advice:

- Educate senior leaders about the opportunities and threats that blockchain capabilities introduce. Use clear language and definitions in internal discussions about how distributed ledgers may or may not improve existing systems and processes.
- Continue to develop proof of concepts (POC) — especially in the context of market ecosystems. Identify integration points with existing infrastructures (for example, digital wallets, core systems of record, customer service applications and security systems). Analyze the role, maturity and interdependence of synergistic technologies such as artificial intelligence (AI) and the Internet of Things (IoT) as key levers in the evolution of blockchain complete and enhanced solutions.

Executives planning on deploying blockchain solutions must:

- Ensure sufficient innovation capacity is applied to the evolution of distributed ledgers and blockchains outside of your immediate industry.
- Identify integration points with existing legacy infrastructures. Evaluate the total cost of ownership against existing systems. Be very cautious about vendor lock-in and merely replatforming the enterprise without any additional value.

Business Impact: Blockchain provides an opportunity for enterprise leaders to imagine new kinds of business models and revenue flows. In particular, leaders decentralize commercial exchange, thereby reducing friction and cost, by monetizing multiple forms of assets. Enterprise leaders also

face a threat from startups and businesses that can use the five core elements of the blockchain concept to disrupt and disintermediate markets and industries. This is done by offering capabilities like identity portability, trustless interactions, smart contracts and new forms of value exchange. These opportunities and threats will evolve over the next 10 years in varying degrees, affording strategic planners an opportunity to proactively address opportunities and threats. Regulation will play a significant role in the speed of evolution. Recent developments around the framing of compliance for token use and initial coin offerings (ICOs) are to be watched, as well as general consumer behavior toward, and acceptance of, multiple forms of assets. Progression with identity management will change the power structure in many industries and should be viewed through both a business and technology lens.

For distributed ledger concepts to really transform industry operating models via automation, and for economies of scale to improve efficiency, there needs to be a wholesale reimagination of digital transformation. COVID-19 may provide that catalyst. However, rather than encourage collaboration, it may increase fragmentation, making it harder for cross-industry and cross-jurisdiction consortia to be successful. Multiple business use cases have yet to be proven, and accurate value outcomes have yet to be calculated. It is unclear whether current approaches for using distributed ledgers provide sufficient differentiation compared with existing, proven messaging and data technologies. Clearly, interoperability of both technologies and processes is a significant requirement.

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Algorand; Block.one; Cardano; Ethereum; Hyperledger; Neo; R3; Zilliqa

Recommended Reading:

“Understanding the Gartner Blockchain Spectrum and the Evolution of Technology Solutions”

“Guidance for Blockchain Solution Adoption”

SQL Interfaces to Cloud Object Stores

Analysis By: Adam Ronthal

Definition: SQL interfaces to cloud object storage provide the ability for enterprises to interact with data residing in cloud storage using familiar SQL syntax. An object store differs from block storage, in that it manages data as objects rather than the lower-level sectors and tracks. Typically delivered as a cloud storage service, object stores are well-suited to storing large volumes of multistructured data and are often used in support of data lakes.

Position and Adoption Speed Justification: Cloud object stores represent an alternative to technologies like the Apache Hadoop Distributed File System (HDFS), which are often used in support of analysis on multistructured data, as in the case of a data lake. As the Hadoop stack

continues to disaggregate, increased adoption of technologies and scenarios that substitute for some of the core Hadoop elements are rising in prominence. SQL interfaces to cloud object stores fall into this category and aim to fulfill a desire for easier access to data residing in cloud-based data lakes on native cloud storage like Amazon Simple Storage Service (Amazon S3), Microsoft Azure Blob Storage and Azure Data Lake Storage, and Google Cloud Storage. Offerings from various vendors are widely available, but the ability to provide sophisticated optimization for queries is highly variable. Because of this variability (both in capabilities and compliance to SQL standards) which can lead to a mismatch of expectations and real-world capabilities, SQL interfaces to cloud object stores remain firmly in the Trough of Disillusionment this year. We expect rapid continued adoption as these solutions continue to mature and find specific use-case applicability.

Offerings from established vendors include Amazon Athena, Amazon Redshift Spectrum, Apache Spark, Starburst (Presto), Cloudera Impala, Google Cloud Platform's BigQuery, and Microsoft Azure Data Lake Analytics, as well as core Hadoop functionality offerings like Apache Hive. New entrants to the market like Varada provide some performance optimizations on top of existing functionality in open-source Presto. Additionally, many established DBMS and Hadoop offerings are providing the ability to reach into cloud object stores, often using an external table definition.

In 2021, the profiles for SQL interfaces to cloud object stores and SQL interfaces to Hadoop will be combined, since specialization is eroding and the categories are converging.

User Advice: End users should evaluate these capabilities in incumbent products to see what benefits they can offer, at the same time, exploring synergies to related complementary technologies already in use, like relational data warehouse offerings. While the level of SQL support may initially be limited, tight integration with relational offerings may provide robust functionality in some cases; testing core functionality and integration with third-party BI and analytics tools, and applications to validate functional support and performance expectations will be essential.

These offerings will initially align to exploratory, data discovery ad hoc use cases (as expected from an interface to a data lake), as well as the ability to access "cooler" data via SQL and federation at lower cost and performance. Operational use cases with advanced performance and concurrency SLAs may emerge in the future, but do not expect this capability to be present in current offerings, or to be mature if it does exist. This approach does not offer the same level of performance optimization capabilities available in a traditional RDBMS.

Given their widely variable uses and consumers, cloud object stores need a thoughtful approach to governance that the stores themselves do not provide. Ensure access is passed through layers and tools that close this gap.

Business Impact: Cloud object storage is increasingly filling a role of the data management glue that ties other cloud services together. As such, it tends to be a clearinghouse of data that may be ingested into enterprise cloud environments. Some data will simply pass through to other cloud services, traditional data warehouse offerings, for example. Other data will remain and serve as the basis for a cloud-based data lake where data is ingested in its raw, native form for later exploration and use. SQL interfaces to object stores allow ready access to this data from a standard interface, sometimes even integrated with existing traditional relational offerings, as in the case of Amazon Redshift Spectrum and Microsoft SQL Server or Azure SQL Data Warehouse with PolyBase. They

will play a key role in democratizing access to the data lake and will become the standard method for implementing data lakes in the cloud, replacing HDFS-based approaches.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services; Apache Software Foundation; Cloudera; Databricks; Google Cloud Platform; IBM; Microsoft Azure; Snowflake; Starburst; Varada

Recommended Reading:

"Assessing the Optimal Data Stores for Modern Architectures"

"How to Avoid Data Lake Failures"

"Best Practices for Designing Your Data Lake"

"The Practical Logical Data Warehouse: A Strategic Plan for a Modern Data Management Solution for Analytics"

"Solution Path for Planning and Implementing the Logical Data Warehouse"

Data Lakes

Analysis By: Nick Heudecker; Henry Cook

Definition: A data lake is a concept constituting a collection of storage instances of various data assets combined with one or more processing capabilities. Data assets are stored in a near-exact, or even exact, copy of the source format and in addition to the originating data stores.

Position and Adoption Speed Justification: Though data lakes have started emerging from the Trough of Disillusionment, a majority of the market still exhibits significant confusion over the data lake concept, how it compares to concepts like data warehouses and data hubs, and how it supports different user groups and service-level agreements. Another portion of the market is embracing packaged data lake offerings from cloud providers and other vendors. These packaged offerings help enterprises conceptualize both what a data lake is and where the data lake fits into their data estate. Adoption of these products has pushed data lakes through the Trough of Disillusionment and toward the Slope of Enlightenment.

This progression has come at a cost. Data lakes have already run their course for many organizations. Some companies struggled to determine the return on investment for their data lake projects, failing to uncover a single meaningful outcome that originated from their lake. Others found some success in their experiments but struggled to evolve those experiments into production for a variety of reasons. Many of these organizations gave up on their data lakes, preferring to use

infrastructure that accommodated diverse analytics consumers, rather than solely accommodating data scientists.

Despite progression along the Hype Cycle, data lake success is far from guaranteed. Infrastructure is only one part of the data lake equation. Data and analytics leaders must design and implement a pipeline to move projects into production, ensure high quality, reproducible outcomes, and develop highly skilled individuals that can derive value from datasets with varying levels of context, quality and format.

User Advice:

- The fundamental assumption behind the data lake concept is that everyone accessing a data lake is moderately to highly skilled at data manipulation and analysis. Before implementing a data lake, ensure you have either the necessary skills, such as data science or engineering, or a plan to develop them.
- Recognize that results will likely be difficult to reproduce between analysts. By definition, data stored in data lakes lacks semantic consistency and data governance of any kind. This makes data analysis highly individualized (a consumerization of IT goal) at the expense of an easy comparison or contrast of analytic findings (also indicative of consumerization of IT).
- There are certain SLA expectations that can be served by data lakes. However, most end-user SLAs for analytics rely on repeatability, semantic consistency and optimized delivery. Once data lake efforts confront these SLAs, it is time to explore alternative information management architectures, such as the logical data warehouse, to rationalize how information is stored with how it is used.
- Evaluate a variety of implementation options. Cloud-based data lake offerings are increasingly popular choices and provide a simple pattern for data ingestion and consumption, but no two data lakes are the same. Your users' needs may require a radically different implementation than prepackaged services. Expect your data lake to be a portfolio of processing capabilities.
- Many organizations think of a data lake to share data within the organization, roughly equivalent to data as a service. This frequently results in multiple copies and lineages of data — exactly what many data lake advocates said wouldn't happen. Alternative architectures, like data hubs, are often better fits for such use cases (see "Use a Data Hub Strategy to Meet Your Data and Analytics Governance and Sharing Requirements").

Business Impact: The data lake concept has the potential to have a high impact on organizations, but its effect is only moderate at present. To get full value from a data lake, its users must possess all the skills of a system analyst, data analyst *and* programmer. They should also have significant mathematical and business process engineering skills — otherwise it will still have a significant impact, but a highly undesirable one.

Depending on the method of implementation, a data lake can be a low-cost option for massive data storage and processing. Processed results can be moved to an optimized data storage and access platform, based on business requirements and tool availability. However, the potentially high impact of this will be diluted by vendors seeking to use the term "data lake" merely as a means of gaining entry to the highly mature analytics and data management markets. This presents the potential for

some very real lost opportunities and large sunk costs, as a balanced warehouse/services/lake architectural approach would be the better solution.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Cambridge Semantics; Cazena; Google Cloud Platform; IBM; Informatica; Microsoft; Oracle; Zaloni

Recommended Reading:

"How to Avoid Data Lake Failures"

"Solve Your Data Challenges With the Data Management Infrastructure Model"

"Efficiently Evolving Data From the Data Lake to the Data Warehouse"

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

"Building Data Lakes Successfully"

"Metadata Is the Fish Finder in Data Lakes"

Master Data Management

Analysis By: Sally Parker; Simon Walker

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 consistent and uniform set of identifiers and extended attributes that describes the core entities of an enterprise, such as customers, citizens, suppliers, products, assets, sites, hierarchies and the chart of accounts.

Position and Adoption Speed Justification: MDM is focused on the consistency and quality of data that describes the core entities of an organization — data that sits at the heart of the most important business decisions. The need for consistency of this 'master data' across business silos and the recognition that MDM is the cornerstone of digital business success continues to drive interest in MDM. Organizations having invested in establishing an enterprisewide, trusted, view of their master data benefit from a greater agility to predict and respond to unexpected events — such as COVID-19 triggered changes in customer buying patterns. Interest in MDM has spiked courtesy of COVID-19.

The market penetration of MDM as a whole is still low due to perceived complexity and cost. Efforts are underway by MDM vendors to lower the barrier to entry to adoption with subscription pricing, cloud offerings, simpler products, and rapid deployment tools. However, MDM remains a complex and maturing undertaking because technology is often mistakenly seen as a panacea — technology alone is insufficient to solve a problem that traverses people, process and technology across the enterprise. Confusion also remains regarding what constitutes master data and discipline in keeping the MDM program lean when in the pursuit of a "360-degree view" of critical data.

MDM is leaving the Trough of Disillusionment as organizations better understand the challenges, but they still most often are unable to overcome them without external guidance.

User Advice: Organizations with complex or heterogeneous application and information landscapes typically suffer from inconsistent master data, which in turn weakens business-process integrity and outcomes. Business applications affected may include customer-facing, supplier-facing, enterprisewide and value chain applications. If your business strategy depends on the consistency of data within your organization, you will likely consider MDM as an enabler of this strategy.

Companies investigating MDM should treat MDM as a technology enabled business initiative and consider the following:

- Secure executive sponsorship prior to proceeding.
- Ensure a clear "line of sight" to tangible business benefits.
- Prioritize, and agree upon which business initiatives will benefit most from trusted master data as a starting point.
- Stay lean and focused — classify only the most widely shared application data as master data "the least amount of data governance on the least amount of data that has the greatest impact on business outcomes."
- Identify the architectural role that each implemented MDM solution will play in your approach to enterprise information management (EIM). Use MDM as an opportunity to implement sound information architecture fundamentals, such as canonical transaction formats for master data domains as part of a well-managed data integration practice.
- Factor services into costs, over 90% of organizations leverage services for support with their MDM strategy and/or implementation.

Business Impact: Leading organizations that create a strategy to implement MDM and supporting technology that is well-thought-out, holistic and business-driven will be able to deliver significant business value. They will do so in terms of enabling competitive differentiation and business growth, improved customer experience, reduced time to market and delivery on operational efficiency as well as by meeting governance, risk management and compliance requirements.

MDM strategies that are linked to strategic IT enterprise transformation efforts (such as ERP and CRM implementations) provide significant additional value to those efforts. Conversely, MDM-centric business cases are often used to highlight opportunities for significant business process optimization.

In some cases, we have seen the need for MDM to trigger improvements in areas such as data quality, information governance, enterprise metadata management, although conversely, we have also seen those programs initiate the need for better master data management.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Ataccama; IBM; Informatica; Profisee; Reltio; Riversand Technologies; SAP; Semarchy; Stibo Systems; TIBCO Software

Recommended Reading:

"Magic Quadrant for Master Data Management Solutions"

"Critical Capabilities for Master Data Management Solutions"

"Mastering Master Data Management"

"Create a Master Data Roadmap With Gartner's MDM Maturity Model"

Climbing the Slope

Data Preparation Tools

Analysis By: Ehtisham Zaidi; Sharat Menon

Definition: Data preparation is an iterative and agile process for finding, cleaning and transforming raw data into curated datasets for data integration, data science/ML and analytics/BI. Data preparation tools promise faster data delivery time by allowing business users to integrate internal and external datasets for their use cases. Further they allow users to identify anomalies, patterns and review their findings in a repeatable fashion. Some tools embed machine learning algorithms that reduce or even automate certain repeatable data preparation tasks

Position and Adoption Speed Justification: Businesses continue to demand faster time to insight to remain competitive. They seek this insight in increasingly diverse and complex combinations of data from both internal as well as less-known external sources. This is driving a shift from traditional, IT-centric analytics/BI toward augmented data profiling and integration that assist business users and domain experts in data integration design and delivery. This approach backed by data preparation tools enables them to quickly find important patterns in data with reduced IT support.

This is where data preparation tools provide subject matter experts and other citizen users the ability to find and discover data during the development of analytic and operational data use cases. In the past year we have seen organizations realize the value of data preparation and vendors have

moved quickly to incorporate data preparation as an embedded capability within broader data and analytics platforms. Most vendors have already responded by including data preparation as an embedded capability within their data integration, analytics/BI and data science/ML platforms whereas others have acquired stand-alone data preparation tools to plus this gap in their solutions.

Gartner views data preparation as a critical capability in modern data and analytics platforms. There is still the need for a stand-alone data preparation tool is still relevant for organizations using multiple analytics and data science tools (to enable the preparation of data once and utilization across multiple data consumption tools). Stand-alone data preparation tools are also needed by organizations that require a specialized data preparation IP for specific data types/use cases (such as data preparation for IoT or streaming data or data preparation for industry/domain specific use cases). However, beyond these specific needs for an independent, stand-alone data preparation tools, we now see most organizations utilizing the embedded data preparation capabilities of their existing data integration, analytics/BI and data science/ML tools for their general-purpose use cases. Therefore, data preparation has experienced swift progression along the Hype Cycle this year. As usage scenarios become clearer and vendor offerings have consolidated to include other key capabilities including data catalogs and hybrid integration and deployment options for distributed data sharing and governance.

User Advice: Data and analytics leaders focused on data management solutions and analytics strategies should:

- Consider data preparation to help business analysts, data scientists, citizen integrators and other analytics content developers to enhance and streamline their data preparation time and effort, while improving data sharing, reuse and governance.
- Evaluate stand-alone data preparation tools when your use case is a general-purpose one, needing integration of data for different analytics/BI and data science tools. Contrastively, evaluate the embedded data preparation capability of incumbent tools if you need data preparation only in the context of those tools, platforms or ecosystems.
- Evaluate data preparation tools for their ability to scale from self-service models to enterprise-level projects. Give preference to tools that can coexist with other data management tools (such as data quality or data governance) and can capture, analyze and share metadata (and lineage) with them, to ensure security, governance and compliance. Also look for tools that can enable users to share and rate prepared datasets with each other for promoting a trust-based governance approach.
- Deploy data preparation tools to complement traditional data integration approaches. These are well-suited in environments that are composed of widely varying data management skills within your analytics users. Enable your most skilled analysts but be mindful of providing yet another platform to undisciplined users who still require external design and hygiene controls.

Business Impact: Data preparation has significant business impact potential because it addresses one of the biggest challenges that your most skilled analysts, citizen data integrators, data engineers and other analytics content authors face. This reduces the time-to-data-delivery for their key data and analytics use cases. Moreover, when used by IT to create curated datasets for a broad

range of content authors in a repeatable manner, it has the potential to impose a level of trust on widely deployed and distributed analytics initiatives.

With the inclusion of machine learning (to automate tedious data integration tasks) and data catalogs (to inventory distributed datasets), data preparation has disrupted IT-centric data integration. It has also forced mature vendors to begin including similar features and functionality within their broader tools. Mature data preparation tools now include basic data quality, governance and metadata management capabilities along with the ability to access and prepare data across a hybrid and multi-cloud environment.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Altair; Alteryx; Boomi; DataRobot; Infogix; Informatica; SAP; SAS; Talend; Trifacta

Recommended Reading:

“Market Guide for Data Preparation Tools”

“Data Engineering Is Critical to Driving Data and Analytics Success”

“Toolkit: Evaluate Data Preparation Tools Across Key Capabilities”

“Magic Quadrant for Data Integration Tools”

“Toolkit: Job Description for the Role of a Data Engineer”

Time Series DBMS

Analysis By: Rick Greenwald

Definition: Time series DBMSs are designed to provide rapid ingestion of data, manipulation of data based on its position in a time series, and aggregation of data outside of an active window of data. Once data leaves the active window, downsampling provides an aggregate view of historical data.

Position and Adoption Speed Justification: Time series database technology has been around for many years with vendors such as InfluxDB, and precursors such as kdb+, originally designed for financial use cases. The expansion of Internet of Things (IoT) data sources has sparked new interest in this technology, as organizations look to both perform more sophisticated analytics on IoT data and combine that data with more-traditional sources. Leading cloud vendors, such Microsoft and Amazon Web Services, are adding time series capabilities or services to their platforms.

Time series DBMSs are well-suited for many types of IoT and financial services systems. The requirements of rapid ingestion of data combined with real-time analysis of this incoming data are

the target tasks for time series DBMSs. These products typically keep detailed data for a period of time, and then aggregations based on that data over a longer period of time. Time series DBMSs can also compare and analyze data across multiple time streams, which is difficult to do outside of this technology.

Time series databases may also store data in other DBMSs, with extended internal models and functions. Time series DBMSs also include time-specific functions designed for the type of analysis typical for this use case, and generally include compression designed for optimal effect on time series data.

User Advice: Choose time series DBMSs when you need high levels of performance for specific time series tasks. If your use case fits the target profile — real-time or historical analysis of append-only event data that is in time order — a time series DBMS can add increased flexibility, performance and agility for processing the data. If your use case does not have demanding requirements for ingestion or analysis, the time series capabilities of more standard DBMSs may be adequate for your needs.

Business Impact: Time series DBMSs will be useful for some real-time and financial systems where time series and rapid ingestion of data are critical. Time series DBMSs fulfill a niche use-case role and will be used with other DBMS offerings as part of an overall data solution. Cloud provides best fit solutions to many groups of use cases without significantly increasing management overhead, so time series cloud offerings can be a more viable approach.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; IBM; InfluxData; Machbase; Microsoft; OpenTSDB; Redis Labs

Recommended Reading: “Time Series Database Architectures and Use Cases”

In-DBMS Analytics

Analysis By: Henry Cook

Definition: In-DBMS analytics (also known as in-database analytics or in-database processing) constitutes the integration of analytics into the database management system (DBMS) platform. 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.

Position and Adoption Speed Justification: In-DBMS analytics offerings have been available from data warehouse software vendors for many years and are now becoming much more widely used. Most of today’s mainstream DBMS vendors — both cloud and traditional — are offering in-DBMS analytics capabilities with libraries of many algorithms. In addition, some analytics vendors such as

SAS and IBM (SPSS software) can push their analytics processing down into a suitable DBMS. Also, some vendors such as Alteryx and Fuzzy Logix provide analytics libraries that can be used with DBMS from more than one vendor.

As machine learning becomes more commoditized and its use spreads beyond specialist data scientists, in-DBMS machine learning is an excellent enabler for this wider group of developers and users — who may not realize that it exists. Adoption by traditional and specialist data scientists is more sporadic due to their preference for more sophisticated tools like R, Python, and notebooks (Project Jupyter, Apache Zeppelin). However, now that an increasing number of DBMS vendors are integrating in-DBMS analytics and these can be used with popular analytics and query tools, this attitude is starting to shift. In-DBMS analytics provide a very good solution for moving analytic models to production with model generation, administration and execution all in the same environment. Due to this, we believe in-DBMS analytics is almost on the plateau and will be in mainstream adoption in less than two years.

User Advice: Data and analytics leaders should:

- Consider in-DBMS analytics as a viable option for making large-scale business analytics available to a wider audience. In-DBMS analytics helps by embedding machine learning capabilities in familiar and pervasive platforms that can deliver rapid insights on both historical and incoming data. By avoiding the need to move data out of the DBMS to build analytic models, in-DBMS analytics allows for more flexible experimentation and efficient development of models and applied use cases that can serve as a foundation for broader, more robust and reliable delivery.
- Review your data science development process. Evaluate whether it can be better enabled through in-DBMS analytics, especially for deployment (as deploying machine learning to production can be a challenge).
- When evaluating DBMS systems check whether in-DBMS analytics are supported and, if so, the range of algorithms offered. Also evaluate how these are implemented and do a proof of concept running them against your data volumes. Some vendors embed the algorithms in the DBMS, some allow you to call a library of functions, some interface to a separate analytics server, for example an R server — and combinations of these approaches are also possible. Some vendors have adapted the analytic algorithms they provide to take advantage of parallel processing, others have not. Where parallelism is enabled this may not apply to all of the algorithms. Check these factors for all algorithms you are likely to use. For smaller volumes of data, these factors are unlikely to be a problem, but application across very large volumes of data may limit their usefulness.

Business Impact: In-DBMS analytics benefits enterprise business analytics in the following ways:

- Realizes cost savings by capitalizing on existing investments in DBMS and analytics platforms through reuse of existing assets. You may find that you already own data science capability within your existing DBMS system. It supplies an easy way to add machine learning functionality to your data management architecture and infrastructure or introduce it to your developers.

- Enables users to develop and consume analytics directly from the DBMS, removing the need to move data to a separate dedicated analytics environment.
- Enables a self-contained and robust development, testing and deployment of models. This can include version control, ownership, administration of models and security controls. In particular, it gives you easier and faster deployment since derived models can be invoked in production instances of the DBMS. Given the difficulty that some enterprises encounter moving machine learning analytics into production, this can be of high value.
- Enables delivery of rapid insights by using DBMS performance features such as parallel processing and in-memory processing.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Google; IBM; Micro Focus; Microsoft; Oracle; relationalAI; RapidMiner; SAP; Teradata; VMware (Pivotal)

Recommended Reading:

"5 Useful Ways to Use Artificial Intelligence and Machine Learning With Your Logical Data Warehouse"

"Magic Quadrant for Data Science and Machine Learning Platforms"

"Magic Quadrant for Data Management Solutions for Analytics"

"Why In-DBMS Analytics Deserves a Fresh Look"

iPaaS for Data Integration

Analysis By: Eric Thoo

Definition: Increasingly iPaaS technologies provide data integration capabilities such as ETL, data replication and data virtualization as cloud services. The use of iPaaS technology for data integration applies to diverse scenarios, including the integration of data in cloud, multicloud, on-premises, between enterprises, at the edge, and hybrid deployments.

Position and Adoption Speed Justification: Data integration initiatives continue to grow in importance and have long been addressed via on-premises software technologies. However, the pressures of digital business transformation are adding complexity to data integration strategies in terms of achieving rapid deployments and diversified ways of provisioning data. The adoption of hybrid and multicloud deployment models, cloud data stores, as well as the movement of enterprise data into cloud applications, fuels demands for using cloud services for data integration. Increasingly, on-premises data integration tool vendors are actively competing in the iPaaS market. These providers target specific data integration opportunities, such as integrating SaaS endpoints

and supporting IoT and multicloud solutions, via a cloud-native platform or by offering an iPaaS rendition of existing on-premises integration platforms. In many cases, users of iPaaS favor offerings that support both data and application integration within a single toolset. Midsize organizations — plus business roles and citizen integrators outside of IT in larger organizations, and teams focused on event-driven IT capabilities — are using these capabilities to move and synchronize data that includes cloud sources. Some SaaS providers embed an iPaaS to make it easier and faster to integrate their services with the rest of the application portfolio customers use. In mainstream adoptions, iPaaS for data integration use cases and capabilities expands alongside growing vendor choices. As the use of hybrid delivery models continues to grow, organizations are increasingly considering iPaaS as a strategic component, or as an extension to their data integration infrastructure to enable a hybrid integration platform strategy.

User Advice: Consider iPaaS for data integration as an extension of the organization's data integration infrastructure, and as an enabling technology for hybrid integration platform capabilities (see "How to Deliver a Truly Hybrid Integration Platform in Steps"). In supporting a data integration workload involving cloud integration, use iPaaS as a way of accelerating time to value, minimizing costs and resource requirements relative to on-premises models, and simplifying the deployment of data integration infrastructure with an adaptive or incremental approach. Those in less-technical roles who perform data integration, but are not inclined to make significant customizations, should consider using iPaaS. However, recognize that iPaaS may not address the full scope and complexity of broader data integration requirements, such as extremely large-scale bulk/batch workloads for on-premises enterprise data warehouse support. Be aware that some offerings may be purpose-built for integration problems involving cloud data, where sources and targets are mainly common structures, such as data residing in popular SaaS or operational databases. Because some iPaaS providers are limited in their partnering and metadata extensibility to operate with broader data management infrastructure capabilities such as data quality, plan to ensure that data integration processes are governed, traceable through data lineage support, and reconfigurable for a broad range of integration use cases. Migrating data from one part of the cloud to another, or into internal applications, may elicit data integrity or quality issues (including incomplete or inaccurate data) that aren't necessarily addressed by an iPaaS solution.

Business Impact: The use of iPaaS for data integration offers benefits for most organizations in relieving some of the common challenges concerning flexibility, developer productivity, skills availability, and entry cost. Resource-constrained businesses can apply data integration capabilities offered in iPaaS to targeted issues — such as the synchronization of data between on-premises and off-premises applications, and the composition of integrated views of data sources residing both inside and outside the firewall. Organizations experienced in dealing with SaaS and other cloud services are expanding their techniques to create a more dynamic computing environment for data integration workloads. An example is simplifying deployment and expanding access to data sources in the cloud. With increasing interest in cloud-based solutions for data warehousing and analytics, requirements for access to, and delivery of, datasets utilizing the cloud environment present growing opportunities to leverage iPaaS for data integration. In complementing established on-premises platforms and supporting an HIP strategy, iPaaS for data integration will become an increasingly common component of an organization's data management and application infrastructures.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: DBSync; Dell Boomi; Etlworks; Informatica; Jitterbit; Oracle; SAP; SnapLogic; Talend; TIBCO Software

Recommended Reading:

"Magic Quadrant for Enterprise Integration Platform as a Service"

"Magic Quadrant for Data Integration Tools"

Multimodel DBMSs

Analysis By: Merv Adrian

Definition: Multimodel DBMSs support several data engines, relational and/or nonrelational (e.g., document, key value, graph, time series, wide column), in a single database. They reduce complexity and enhance usability and performance by providing a common access mechanism for different persistence types, each optimized for the nature of the data being used.

Position and Adoption Speed Justification: Multimodel DBMSs typically are ongoing enhancements of existing products. Although multimodel is rarely used as a leading marketing message by vendors, every leading vendor now has a flagship offering that support two, three or more types. Adoption will be driven by customers' familiarity with their choices — simplifying either the development process, integration requirements, performance, or the management of a portfolio of vendors. Gartner inquiries regarding multimodel DBMSs remain limited, reflecting minimal awareness of multimodel as category — it is an analysts' abstraction, not a market label (except for Oracle's "Converged" message). With gaining momentum for cloud-based best-of-breed portfolios as a leading influence in the DBMS world, multimodel offers an alternative in both deployment modes.

Integration issues with services-based architectural approaches in the cloud are leading vendors to pursue both strategies, adding both separate products and multimodel capabilities to their flagship offerings to capture both approaches. An example is Oracle's collection of Oracle Database (itself multimodel, including relational, document and graph), Oracle NoSQL Database and others. The same vendors may also offer a different approach via virtualization capabilities such as Microsoft PolyBase, a SQL Server feature where storage is separate, but the interface is unified. Microsoft also offers document and graph capabilities within SQL Server and multimodel capabilities on a nonrelational base with CosmosDB.

The growth of actual multimodel usage will depend on adoption by developers. However, vendors continue to rapidly add capabilities to their products, so market adoption as measured by the Hype Cycle moves multimodel well ahead in this year's report toward the Plateau of Productivity.

User Advice: Sometimes, it will be sufficient to use the capabilities of a multimodel product, but in other cases a "best fit" approach will be much better. Data and analytics leaders should:

- Assess multimodel capabilities of your DBMSs when planning augmented transaction processing and logical data warehouse deployments due to their multipurpose capabilities, which may simplify development challenges.
- Use multimodel DBMSs to create a stable Mode-1-style platform that is suitable for responding to and stabilizing agile, Mode 2 explorations of alternative data paradigms.
- Plan for training because skills in multimodel DBMSs are lacking. Demand that vendors provide assistance with the transition to new models of design and deployment.
- Choose vendors whose stability and track record are demonstrable, and whose roadmaps are consistent with your planned use cases.

Business Impact: Multimodel DBMSs can reduce the complexity of existing portfolios of production systems. They can often provide some of the needed auditing, concurrency controls, versioning, distributed data complexity management, points of governance and security that specialty products lack, permitting delivery of these functions across use cases. They offer a potential solution to vendor proliferation and complexity, as well as performance benefits form reduced layers of external integration, but also create new skills requirements and potentially competing alternatives.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services; DataStax Enterprise; EnterpriseDB; Google Cloud Platform; IBM; MarkLogic; Microsoft; MongoDB; Oracle; SAP

Recommended Reading:

"IT Market Clock for Database Management Systems, 2019"

"The Impact of Modern Data Solutions on IT Modernization Decisions"

SQL Interfaces to Hadoop

Analysis By: Merv Adrian

Definition: SQL interfaces for Apache Hadoop provide the ability for enterprises to interact with data residing in HDFS, often in data lakes, using familiar tools for analysis — most of which use SQL as their query language.

Position and Adoption Speed Justification: SQL interfaces to Hadoop move ahead slightly from 2019 as Hadoop adoption has continued to grow; the Data and Analytics Survey 2019 showed 75%

of surveyed enterprises already using Hadoop, with another 14% planning to in 12-24 months. Following the merger of Cloudera and Hortonworks, Apache Hive (batch-oriented) and Apache Impala (interactive) are joined by Apache Arrow, Apache Drill, Apache Presto, and Apache Spark among others, all of which have varying degrees of support for multiple target file and DBMS data.

As users learn the limitations and ideal use cases for various versions of these components across batch, interactive and streaming workloads for operational and analytic use cases, growth and competition begin to expand. Some data stores are suitable for basic to moderate queries but lack the ability to handle truly complex queries at scale and/or with high concurrency. Features like cost-based optimizers, indexing and sophisticated joining techniques remain early in their development and will not provide substantial performance improvement across all cases. In some cases, they are paired with specialized data stores, such as Apache Kudu for use with Apache Impala or LLAP for use with Apache Hive. But proliferating multiple interfaces can create additional complexity as well as lock-in. There are also continuing limitations around user concurrency, as well as occasional issues with software quality, with the various marketplace offerings.

SQL on Hadoop functionality may also be embedded within analytic products that use SQL for multiple targets, such as Amazon Athena, Microsoft Azure Data Lake Analytics and Google Cloud Platform's BigQuery — all available for cloud-specific targets. These tend to be more sophisticated and inherit more DBMS-related technologies. However, interfaces to HDFS alone are less relevant as data is also placed into DBMSs and, as Hadoop usage itself moves to the cloud, placed into native cloud object stores instead of HDFS. SQL Interfaces to Hadoop thus are subsumed into broader access tools and are no longer a separate category. They are now considered Obsolete Before Plateau. Access to these multiple data stores with different capabilities is discussed in the technology profile SQL Interfaces to Cloud Object Stores.

User Advice:

- Hadoop SQL interfaces aren't general-purpose query engines. Optimizing multiple workload types may require several SQL interfaces, or a more broadly capable multitarget layer.
- Don't rely on benchmarks as an indicator of performance characteristics. Test SQL queries against your data and workloads requirements for throughput and ability to scale, including concurrency and performance SLAs.
- Test with your third-party BI tools and applications to validate functional support; implementations are unlikely to implement ANSI SQL standards in their entirety. Test not just for SQL compatibility, but whether a representative set of your queries can run to completion.
- Choose the broadest alternative weighed by the most value-bearing use case, since vendors are attempting to differentiate. Look for support for creating user-defined functions to facilitate the use of multiple offerings in a consistent way.
- Evaluate SQL interfaces provided by your incumbent DBMS vendor that extend their data access capabilities to external environments like Hadoop and nonrelational data stores.

Business Impact: Business impact is improving, and will be enhanced further as data stored on Hadoop clusters, file systems and cloud object stores becomes accessible to a broader base of business analysts than BI tools have hitherto delivered due to the previously stated performance

and scalability problems. These benefits will be made available to broader groups of users, such as data scientists and senior business analysts, rather than whole enterprises.

Benefit Rating: Moderate

Market Penetration: More than 50% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services; Cloudera; Databricks; Dremio; Google Cloud Platform; Hewlett Packard Enterprise; IBM; Microsoft; Oracle; SAP

Recommended Reading:

"Best Practices for Designing Your Data Lake"

"Selecting SQL Engines for Big Data Workloads"

Data Integration Tools

Analysis By: Ehtisham Zaidi; Mark Beyer; Robert Thanaraj

Definition: Data integration tools enable the construction and implementation of data access and delivery capabilities that allow independently designed data structures to be leveraged together. This market has matured from its traditional focus on batch integration but continues to evolve to support a combination of modern delivery styles (such as data virtualization and stream data integration), and for hybrid and multi-cloud integration scenarios. Some tools also support data fabric designs that augment parts of data integration design and delivery.

Position and Adoption Speed Justification: Organizations now need to enable a data integration architecture that can support distributed data management and deliver data at all latencies, granularities and across a range of use cases. These use cases include MDM, Analytics/BI, data science/ML, data warehousing, maintaining data consistency between applications, and multi- and hybrid cloud data access.

Tightly integrated suites in which all components share metadata (both active and passive), design environment, administration and data quality support remain an area for improvement in the data integration tools market. Activities for data preparation by skilled data engineers and citizen integrators and other non-IT roles spur requirements for new interfaces. The demand for a seamless integration platform that spans and combines multiple data delivery styles, multiple deployment options (hybrid and multi-cloud) and multiple personas currently exceeds the capabilities of most offerings.

Data Integration tools have now become more mature on technical metadata ingestion and analysis to support data integration activities. However, there is still room for improvement and maturity for data integration vendors to introduce capabilities to harness and leverage "active" metadata. The ability to execute data integration in a hyper connected infrastructure (irrespective of structure and

origins)and the ability to recommend and automate transformations through embedded ML capabilities are the most important differentiators between traditional and modern data integration tools .

Here it is important for data integration tools to enable data fabric architectures that support the much needed automation in data integration and data preparation design and delivery through analysis of active metadata, embedded knowledge graph capabilities and ML. See “Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration.” Dynamic data fabric designs which will bring together physical infrastructure design, semantic tiers, prebuilt services, APIs, microservices and integration processes to connect to reusable integrated data. Driven to expand capabilities, vendors continue to add data integration functionality or acquire technology in these areas.

User Advice: We recommend:

- Assess your data integration approach and capability needs to identify gaps in critical skill sets, tools, techniques and architecture needed to position data integration as a strategic discipline at the core of their data management strategy.
- Review current data integration tools to see what added capabilities they may offer. Such as the ability to deploy core elements (including connectivity, transformation and movement) in a range of different data delivery styles driven by common metadata, modeling, design and administration environments.
- Identify and implement a portfolio-based approach to your integration strategy that extends beyond consolidating data via extraction, transformation and loading (ETL) to include stream data integration, event recognition and data virtualization.
- Be aware of the ongoing synergy with the metadata management, data quality, application integration and data preparation tools, and ideally seek solutions that align capabilities in a complementary manner.
- Make automation of data integration, ingestion and orchestration activities as your No. 1 goal for the year. Favor those vendors offerings that can provide actionable evidence of enabling data fabric architectures (that utilize active metadata, embed knowledge graphs and ML toolkits) and support automation in integration design and delivery. This will immediately increase the productivity of your data engineering teams.

Business Impact: Data integration tool suites with broad applicability will bring value to all modern and existing data and analytics initiatives. Organizations adopting data integration tools increasingly exploit comprehensive data delivery capabilities and harness benefits in the form of time-to-data-delivery, cost savings, productivity improvements, quality enhancements and flexibility provided by these tools. However, the popularity of data preparation and other self-service ingestion tools, along with the sole focus on analytics use cases demonstrated by these tools, will create some confusion in the market, slowing the advance of data integration tool suites. Advanced business users will use data preparation functionality, which can then extend or forward their innovations to IT implementers for governed data management and integration. Data integration tool suites are

increasingly expected to deliver simpler interfaces and use-case-oriented functionality for less-experienced roles and self-service functionality.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Denodo; IBM; Informatica; Microsoft; Oracle; SAP; SAS; Syncsort; Talend; TIBCO Software

Recommended Reading:

"Magic Quadrant for Data Integration Tools"

"Critical Capabilities for Data Integration Tools"

"Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration"

"Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements"

"Modernize Your Data Integration Architecture for Digital Business by Combining Data Delivery Styles"

Operational In-Memory DBMS

Analysis By: Donald Feinberg

Definition: An operational in-memory database management system (IMDBMS) is a DBMS (including both relational and nonrelational) used for transactions where all necessary data is stored in memory (DRAM) on a server — not simply an in-memory disk-block cache. It has all the necessary structure in memory and all DBMS operations — such as select, update and delete — are performed in memory. An operational IMDBMS can scale vertically (in a single server) or horizontally (in a cluster of servers).

Position and Adoption Speed Justification: The adoption of IMDBMSs for transactions has been slower than we expected. Technology for operational IMDBMSs is maturing and growing in acceptance, with most DBMS vendors offering an IMDBMS or in-memory option. Several now also support Intel Optane DC Persistent Memory. Although the costs of the necessary infrastructure have been declining and price/performance is improving, acquisition costs remain higher than for their disk- and flash-based counterparts. This is being remedied by major cloud providers now offering support for IMDBMS, generally at a lower total cost of ownership (TCO) than on-premises.

An inhibitor to the growth of operational IMDBMS technology is the need for new persistence models that support the high levels of availability required to meet transactional SLAs. DRAM is volatile because data is lost if the power is lost. Solid-state drives (SSDs) and hard-disk drives

(HDDs) can be used for persistence, but this must be synchronous with transactions to achieve consistency and can degrade performance. Delivery of nonvolatile RAM (NVRAM) — such as Intel Optane DC persistent memory in April 2019 — and adoption of persistent memory by DBMS vendors will have a major impact on high availability/disaster recovery (HA/DR) and IMDBMS restart. Although IMDBMS will need to be mature, more available and also come down in price.

Increasing number of DBMS vendors offering IMDBMSs, coupled with the declining cost of in-memory infrastructure, is accelerating adoption of operational IMDBMS offerings. The major use cases are:

- Consolidation — driving down the TCO
- Augmented transactions — combining operational, analytic processing and AI/ML using the same process or transaction, allowing new real-time applications
- Operational intelligence — real-time dashboards and reporting performed on the operational database
- However, all of these forces (including the speed at which independent software vendors adopt the technology) support the position on the Hype Cycle and the time frame of less than two years to mainstream adoption.

User Advice: Data and analytics leaders should:

- Identify a specific application that would benefit from an IMDBMS — particularly if you are an early technology adopter willing to accept the risks.
- Evaluate applications to use operational IMDBMS for both business value and the possibilities of implementing systems of innovation.
- Monitor maturing technology and published case studies for opportunities suitable for your use cases. For some organizations, operational IMDBMSs remain several years away.
- Consider new DBMS entrants into the space (such as Aerospike and MemSQL), because they may provide operational IMDBMS functionality at attractive price points and more flexible implementations.
- Monitor the use of the new NVRAM products, such as Intel Optane DC persistent memory (by DBMS vendors), for simplifying HA/DR and the restart of IMDBMS.

Business Impact: Operational IMDBMSs for transactions have the potential to make a tremendous impact on business value in primarily four ways:

- IMDBMSs can speed up transactions by two or three orders of magnitude, especially in throughput (shown by current vendor benchmarks and validated by many Gartner client inquiries).
- Augmented transactions (using IMDBMS technology) enable an entirely new set of real-time applications (such as real-time repricing and power grid rerouting, planning and forecasting), since latency issues do not prevent the use of transaction data for analytics.

- IMDBMSs are also used to support operational intelligence, real-time reporting directly on the operational database, removing the need to move the data first to an ODS, data warehouse or data mart.
- The value of operational IMDBMSs will be magnified by consolidation, because most organizations have few applications, if any, that require this much speed. A cluster of small servers running an operational IMDBMS can support most or all of an organization's applications, drastically reducing operating costs for cooling, power, floor space and resources for support and maintenance. This will drive a lower TCO during a three- to five-year period and will offset any higher total cost of acquisition from the more expensive servers.

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Aerospike; IBM; MemSQL; Microsoft; Oracle; Redis Labs; SAP; VoltDB

Recommended Reading:

"Delivering Digital Business Value Using Practical Hybrid Transactional/Analytical Processing"

"Magic Quadrant for Operational Database Management Systems"

"Critical Capabilities for Operational Database Management Systems"

Wide-Column DBMSs

Analysis By: Merv Adrian

Definition: Wide-column DBMSs store rows of data in tables, similar to relational DBMS but without referential integrity and often without JOINs. They support flexible schema definitions, making wide-column DBMSs popular for storing semistructured data, like log and sensor data, and relaxed consistency make them popular for geographically distributed deployments with challenging SLAs. Although wide-column DBMSs do not typically support relationships or transaction semantics between rows or tables, they are effective where these are not required.

Position and Adoption Speed Justification: Wide-column databases are increasingly used by enterprises processing growing amounts of structured and unstructured data in less transaction-intensive distributed scenarios. Thirty-one percent of respondents to the 2019 OPDBMS Magic Quadrant survey are using them; another 15% plan to within a year. Cloud deployment is likely to accelerate now that Amazon Web Services (AWS) Keyspaces will compete with DataStax Enterprise's Apache Cassandra-based offering. They join Google's Cloud Bigtable, Alibaba Cloud Tablestore, Instaclustr, Microsoft Azure Cosmos DB and others in a crowded market. Apache HBase is offered by Cloudera and Hewlett Packard Enterprise's (HPE's) API-compatible MapR Database.

The complexity of deployment and operation has been a challenge for most wide-column offerings but the move to the cloud is adding much improved and often automated, even serverless operations. The pace of functional enhancement will also pick up as cloud vendors battle for differentiation and deliver new releases more rapidly.

User Advice:

- Model expected workloads and data volume to gain an accurate understanding before deployment. The right hardware or cloud instance types for wide-column databases are critical for cost-effective delivery to your SLAs.
- Ensure that your network has no bottlenecks preventing efficient distributed operations for read and write requests, including replication. Additionally, conflict resolution requires time synchronization across the cluster, as well as between applications interacting with individual nodes.
- Develop the right data model to enable decisions about how trade-offs should be made between consistency and latency. Some DBMS solutions allow such trade-offs to be made, even on a per-operation basis; developers must incorporate this into applications. Because of the absence of JOINs, operational use cases are more suitable targets.
- Leverage open-source offerings for early on-premises wide-column DBMS experiments and pilots. Users of these technologies should engage with vendors offering commercial support before moving to the production stage.
- Investigate the security capabilities of wide-column DBMSs before implementing applications with significant risk profiles, since these features still lag behind mature RDBMS.

Business Impact: The current business impact of wide-column DBMSs is moderate. The ability to distribute massive amounts of semistructured data addresses several existing and emerging use cases, particularly in support of more-complex Internet of Things initiatives. Although vendors offering products, training and support are driving adoption and accelerating development, multimodel DBMSs may provide similar function with more flexible opportunities — many include wide-column capability.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Sample Vendors: Alibaba Cloud; Amazon Web Services; Cloudera; DataStax; Google Cloud Platform; Hewlett Packard Enterprise; Microsoft Azure; ScyllaDB

Recommended Reading:

“Assessing the Optimal Data Stores for Modern Architectures”

“Assessment of DataStax Enterprise With Cassandra”

"Choosing the Right Path When Exploring Hadoop's Future"

Entering the Plateau

Apache Spark

Analysis By: Svetlana Sicular

Definition: Apache Spark is an open-source in-memory distributed computing framework for large-scale data processing and analytics. It uses a memory-centric processing model to support performance for highly iterative tasks, such as query processing, data streaming, machine learning and graph analysis. Spark can be deployed both on-premises and in the cloud.

Position and Adoption Speed Justification: The hype around Apache Spark is subsiding, as it is entering the Plateau of Productivity. Apache Spark is becoming a productive part of infrastructure, when it is needed, and is working reliably for its target audience. Spark is now embedded in many data management and analytics products on-premises and in the cloud, mostly simplifying the infrastructure challenges.

Currently, every component of Spark has a formidable alternative, but the proven reliability of Spark and availability of Spark skills make it less hyped, but nevertheless, popular and well-understood in data engineering and data science. Data scientists are notoriously "shiny object"-driven, so Spark may not be the coolest framework, but rather, another tool in rapidly proliferating toolboxes of data pipelines' development and machine learning. Spark is used extensively for data management, and again, end users don't see Spark, because it is the engine for serverless Databricks, AWS Glue, Amazon EMR and Microsoft Azure Data Factory.

Meanwhile, the creators of Spark (mostly employed by Databricks) are working on new commercial and open-source capabilities, such as MLflow, Delta Lake and ML Runtime, outside Apache Spark, leaving this open-source framework less directed. Spark may become less popular if vendor competition slows the Apache Spark open-source project, or if a disruptive alternative to Spark draws talent and attention from the current Spark ecosystem.

User Advice: We recommend the following:

- Use Spark as a general-purpose engine if you have large datasets or if your vendor relies on Spark in its own solution. D&A leaders who seek flexibility of programming languages for large-scale data processing and analysis in various ways, such as SQL, ETL, data streaming and machine learning, should employ Spark as a general-purpose distributed compute engine, even given the volatility in the machine learning space.
- It takes time and effort to get the Spark design right, but once implemented, Spark is a relatively low-risk, portable method of experimentation with machine learning, when underlying data is represented in large datasets. Spark could be also viewed as a gateway to other popular software frameworks, for example, TensorFlow, Keras or PyTorch. Spark is less effective for

small datasets and for general-purpose BI and analytics due to limits in result presentation and retrieval.

- Spark is available on all major cloud platforms, and therefore, it could be the means of portability among the clouds. However, D&A leaders should monitor the rapidly changing cloud landscape for alternative solutions that could be disruptive to Spark or could represent a best fit for their use cases, rather than a general-purpose solution.
- Apache Spark doesn't have its own storage, and disruptive storage or platform offerings could overturn Spark's popularity.

Business Impact: As Spark is available in multiple cloud and other third-party offerings, the fact that Spark is open source is overshadowed by commercial solutions that support it. Therefore, customer selections of data engineering and ML tools should follow a "fit for purpose" reasoning, rather than a choice of Spark as a starting point.

Benefit Rating: Moderate

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Sample Vendors: Amazon Web Services; Cloudera; Databricks; Google Cloud Platform; IBM; Microsoft Azure

Recommended Reading:

"Magic Quadrant for Data Science and Machine Learning Platforms"

"Critical Capabilities for Data Science and Machine Learning Platforms"

"Magic Quadrant for Data Management Solutions for Analytics"

"Market Guide for Database Platform as a Service"

Logical Data Warehouse

Analysis By: Henry Cook; Adam Ronthal

Definition: The logical data warehouse (LDW) is a best-practice analytics data management architecture and design that combines multiple physical analytics engines into a logically integrated whole. Data and analytics leaders can use the LDW to cover the full range of modern analytic requirements within a logically unified system, rather than a single physical server. It is effective for on-premises as well as cloud deployments.

Position and Adoption Speed Justification: In 2020, full or partial LDW deployments increased to more than 25% of all data warehouse deployments. The LDW is an architecture and not an off-the-shelf product; the preintegration of LDW components (particularly the data warehouse and data

lake) is now the norm. Most major data warehouse software suppliers support this architectural approach, in the cloud and on-premises.

Modern analytics workloads need to support many types of data, analytical processing techniques, types and numbers of users and service levels. Data warehouse designers meet these requirements by integrating multiple analytic servers and services using data virtualization, data transports and common metadata, while achieving a single logical view of all data. The LDW enables enhanced enterprise agility and maximizes return on investment for both development and runtime. Data and analytics leaders are encouraged to adopt an LDW architectural approach as it has emerged as a best practice for data management in analytics environments.

User Advice: Data and analytics leaders, and business intelligence (BI) and data warehouse architects, should:

- View analytic servers, such as the data warehouse, data marts and the data lake, not as competing solutions but as collaborating engines. Use the LDW architecture as a template to ensure that all data requirements, processing and users can be satisfied within that architecture. Use the minimum number of data servers to satisfy the maximum number of requirements while dramatically reducing redundancy. The LDW can cover both exploratory and production-optimized delivery use cases among its multiple components.
- Remember that the LDW does not alter the underlying performance and design characteristics of the data stores that it accesses. Use the LDW architecture to make it easy to move data and processing to where price performance can be optimized. Place data firstly according to the type of data, and then according to the service level needed for the analysis of that data. The LDW standard interface will be suitable in most cases. However, for optimal performance, LDW components may be accessed directly as local direct access can be more efficient in meeting particularly stringent performance requirements. Different key measures can be accommodated — the data lake for optimizing cost per terabyte, the warehouse and marts for optimizing cost per query. The blended costs and benefits will deliver maximum return on investment.
- When evaluating and procuring systems, seek those that can support the LDW architecture, either as a controlling hub, or as a participating component. Look for the ability to use multiple processing engines and techniques, queries that can span multiple server types, integrated metadata and easy ways to transfer data within the architecture. Evaluate the benefits of these in how they can support more requirements with faster and more agile development, thus maximizing benefits delivered.
- Use the Gartner Data and Analytics Infrastructure Model (DAIM) and other LDW-oriented research as tools for planning your analytical processing.

Business Impact: The LDW allows for faster exploration of new data assets by qualified and skilled users. At the same time, it provides a framework that allows the work of highly skilled analysts and data scientists to be “promoted” into production runs. Those production runs usually result in metrics, indicators or lists that are more easily rendered in production dashboards or reporting systems for less-skilled analyst users. The LDW is both an evolution and an augmentation of existing data warehouse architecture practices. It is also an approach for starting a data lake

initiative and building “backward” to combine it with traditional data warehouse solutions as needed. It reflects the fact that not all analytical, querying and reporting needs can be supported by a traditional, centralized, repository-style data warehouse, nor, conversely, by data lake implementations. It implies a much broader and more inclusive data management solution for analytics.

The LDW provides a more reliable way to respond to new analytical or reporting demands with short time to delivery, and with a large number of datasets made available via query tools and applications. In this way, it accelerates data warehouse modifications and provides a rapid deployment capability for new sources with gradually maturing use cases. The LDW can use a data lake as a source, or one of the other underlying data stores.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services; Cloudera; Databricks; Denodo; IBM; Microsoft; Oracle; Snowflake; Teradata; VMware (Pivotal)

Recommended Reading:

“Solve Your Data Challenges With the Data Management Infrastructure Model”

“The Practical Logical Data Warehouse: A Strategic Plan for a Modern Data Management Solution for Analytics”

“5 Useful Ways to Use Artificial Intelligence and Machine Learning With Your Logical Data Warehouse”

“Organizing Your Teams for Modern Data and Analytics Deployment”

“Critical Capabilities for Data Management Solutions for Analytics”

“Efficiently Evolving Data From the Data Lake to the Data Warehouse”

“Adopt the Logical Data Warehouse Architecture to Meet Your Modern Analytical Needs”

Content Migration

Analysis By: Gavin Tay

Definition: Content migration refers to the process of consolidating and transferring unstructured content (files, documents, objects) — along with related metadata, permissions, certificates (DRM, encryption), structure and linked components — stored permanently in one or more content repositories, to a new environment (cloud content services). During the migration process, enterprises typically choose to cleanse their content repositories by archiving old and outdated content.

Position and Adoption Speed Justification: Content migration refers to the process of consolidating and transferring unstructured content (files, documents, objects) — along with related metadata, permissions, certificates (DRM, encryption), structure and linked components — stored permanently in one or more content repositories, to a new environment (cloud content services). During the migration process, enterprises typically choose to cleanse their content repositories by archiving old and outdated content. The demand for, and complexity of, content migration has increased sharply as organizations reevaluate their content services platform (CSP) investments due to market consolidation, cloud office and digital workplace initiatives. The associated complexities are threefold:

- Failed CSP initiatives have caused organizational content to be distributed across myriad content repositories, on-premises and in the cloud.
- Content collaboration platforms (CCPs) used to satisfy basic content management needs are typically on several clouds and transitioning to workstream collaboration tools.
- Infrastructure is being modernized by moving content repositories to new environments, such as the cloud, to optimize operations. Common use cases are file-share elimination, M&A/ divesture activity (multicloud tenant consolidation or splits), information life cycle management, and content services application modernization.

Content migration to CSP alternatives will leverage connectors for a one-time, one-way transfer of large volumes of content stored in obsolete repositories. However, as content is increasingly becoming distributed and progressive, constant synchronization of content from CSPs to cloud alternatives and vice versa have become the norm going forward. Migration tools are occasional and typically one-way bulk loaders, although they are sometimes very robust and increasingly becoming more granular in approach with the ability to redefine the content.

User Advice: Choose CSPs, applications and services that offer standardized and easily accessible repositories, instead of proprietary content databases. At present, migration tools that support the movement of large volumes of content from end-of-life repositories to newer technologies, such as CCPs or purpose-built content migrators that transform content during the migration, have stabilized. While such tools add immediate value, some organizations are choosing not to migrate content given the cost involved and the lack of available IT budget. Accessible repositories allow for content integration and federation rather than migration in these cases.

Evaluate opportunities to employ in-house IT expertise with the acquisition of migration tools. Alternatively, hire a system integrator that would use its own migration frameworks and experts. Using in-house staff may absorb the cost involved, but external expertise should ease the rigorous efforts required to perform complex migrations and bring valuable experience to the process. Many CSP and CCP vendors now include built-in migration options for a common set of legacy repositories. Limiting factors of in-built tools often inhibit the fulfillment of content transformation expectations. Such expectations include improving and evolving content governance iteratively, or introducing an organizational taxonomy driven by business outcomes.

Although each migration is dependent on business needs, many organizations have been highly successful by simply archiving old environments and moving onto a new “greenfield” environment,

without bringing across any baggage. This approach not only eliminates the migration cost burden or the need for migration tools; it also separates most of the dependencies between the old and new systems, allowing for more autonomy when introducing the new solution into the business environment.

Business Impact: Content migration technology isn't only about pushing information and users toward cloud platforms. Cloud office migrations have expanded their scope toward content classification, governance and compliance, to ensure such shifts offer an exit strategy subsequently.

Content migration tools have evolved from on-premises to cloud scenarios during the past year to cater for cloud-to-cloud migration, but also multicloud tenant consolidation and splits. However, the absence of standards and the uncertainty of service availability make cloud-to-cloud migrations complex. Few organizations are embarking on such efforts today, but migrations are set to increase as cloud content becomes commonplace. This complexity is also exacerbated by the fact that hybrid content architectures are left to inherently manage the security risks associated with the integration of public cloud or hosted repositories (as they have done with hosted email).

Companies in highly regulated industries, such as financial services and health sciences, that need to account for all of their content at all times will benefit from stringent, auditable and automated processing using content migration tools. Organizations that have conducted an inventory of all their content repositories and found that more than half of their content is outdated or casual should not waste their IT budgets on content migration tool and service purchases. Rather, they should focus on archiving instead.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Sample Vendors: AvePoint; Binary Tree; Proventeq; Quest Software; ShareGate; Simflofy; SkySync; T-Systems (Vamosa Technologies); Tervela (Cloud FastPath); Xillio

Recommended Reading:

"Market Guide for Cloud Office Migration Tools"

"The Most Common Justifications for a Move to Cloud Office"

"How to Manage Multitenant Cloud Office Deployments"

"How to Successfully Migrate Documents and Collaboration Processes to Office 365"

"Cloud Office Deployments Require Governance of Applications, People and Content"

Data Virtualization

Analysis By: Ehtisham Zaidi

Definition: Data virtualization is one style of data delivery in the broader data integration tools market. It is based on the execution of distributed queries against heterogeneous data sources, federation of query results into cached virtual views, and consumption of these views by applications or other infrastructure components. It can be used to create integrated views of data in-memory (rather than physically storing integrated views in a target data store) and provides a layer of abstraction above the physical implementation of data.

Position and Adoption Speed Justification: The use cases that data virtualization enables are those specifically related to connecting to data in-place instead of collecting the data in a data store. Architectural components of this technology include adapters of various data sources and a distributed query engine that accepts queries and provides results in a variety of ways. For example, as SQL row sets, RESTful or web services interface, APIs or even microservices.

Data virtualization is now an established data integration style which has significantly matured across overall adoption (a customer reference survey conducted to support the 2019 “Magic Quadrant for Data Integration Tools” shows around 30% of organizations using data virtualization). Data virtualization tools have matured across connectivity options, performance optimization, hybrid deployment options and data security. But most importantly customers have now understood the use case applicability and the strengths and limitations of data virtualization technology. Some tools have now added an in-memory cache (for superior performance optimization), data catalogs (for providing users with an inventory of connected data assets) and integrated massively parallel processing support (to enhance query acceleration). This has further driven adoption and the successful movement of use cases to production environments.

Even though data virtualization tools have matured rapidly in the last five years there was still some confusion around its performance. This was mainly because organizations were trying to use data virtualization as a replacement to ETL (or other data movement techniques) for all use cases. Another reason for confusion was that a few applications such as analytics/BI tools continued to tout their localized semantic virtual tiers (which have specific limitations related to performance and query optimization) as a replacement to stand-alone data virtualization technology. This hype and confusion around semantic virtual tiers versus independent data virtualization technology led to performance challenges and incorrect expectations. However, most organizations that have consulted Gartner are now able to identify the right use cases for data virtualization and distinguish between independent data virtualization technology and other tool specific data abstraction/ virtualization techniques. Such tool specific embedded data virtualization techniques include semantic virtual tier connectivity options that come embedded within broader analytics/BI and data science/ML platforms. This has led to a rapid reduction of hype and confusion leading to maturity and confidence in data virtualization.

User Advice: We recommend:

- Make data virtualization a must have data integration component within the broader data management and integration portfolio to assist with agility, reusability and cost optimization in data integration design and delivery.

- Evaluate the current state of your data integration. Set proper expectations, select the right use cases and document the agreed-upon SLAs to separate out when to collect data (using ETL, for example) versus simply connecting to it using data virtualization, before starting your data virtualization journey.
- Use data virtualization for enabling logical data warehouse architectures, deployment of a data access layer via a data fabric architecture, and composition of integrated "single views" of master data objects. Newer use cases include cloud data migrations and federation of data across hybrid and multi-cloud environments.
- Be prepared to respond to monitoring and auditing of data virtualization jobs to determine when they have evolved toward commonly shared models. Create a plan for moving that common virtual model to a more traditional physical data integration process such as batch data integration for consolidation of data.
- First investigate, and use, your existing data integration tool vendors or DBMS vendors for their embedded data virtualization capabilities. Approach a stand-alone data virtualization tool in support of advanced data virtualization capabilities not currently supported by your existing vendors.

Business Impact: Businesses that identify a demand for rapid accessibility to newly added data or for experimental data integration for analytics and to a lesser extent operational data use cases will find data virtualization to be one flexible option for data integration. Data virtualization will allow them to quickly connect to heterogeneous data sources for faster time to data delivery. Data virtualization can also be used to build a "third-party" API layer for accessing data stores. This style of technology will be used to deliver virtualized data services — those that formulate an integrated view of data from multiple datasets and enable this view to be accessed via a service interface. The value of this technology to the business is also inherent to data engineers, citizen integrators and business analysts. This is where data virtualization can provide a flexible way of integrating siloed data sources to provide data services, which can be used in experimental/discovery style use cases or in enterprise class production infrastructure.

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Sample Vendors: Data Virtuality; Denodo; Dremio; Gluent; IBM; Informatica; Oracle; TIBCO Software

Recommended Reading:

"Market Guide for Data Virtualization"

"Toolkit: Assessing the Applicability of Data Virtualization to Your Data and Analytics Use Cases"

"Adopt Data Virtualization to Improve Agility and Bimodal Traits in Your Aging Data Integration"

"Magic Quadrant for Data Integration Tools"

Database Audit and Protection

Analysis By: Brian Lowans; Joerg Fritsch

Definition: Database audit and protection (DAP) provides centralized and consistent security across a variety of database management systems (DBMSs) for both relational and NoSQL databases. Most vendors are now expanding support for cloud-based DBMS. DAP sets and manages DBMS user privileges, monitors user behavior with data and provides database audit logs. It can detect unusual activity and send alerts to prevent a breach. DAP is a critical security control to help meet data residency and data protection and privacy compliance requirements.

Position and Adoption Speed Justification: The DAP is maturing for on-premises deployments. Its ability to control user and administrator privileges, as well as monitor user activity with data, uniquely provides prebreach, real-time breach, and forensic or incident response analytics. This user analytics with data context is not addressed by tools such as identity and access management (IAM), security information event management (SIEM), or user and entity behavior analysis (UEBA). The core DAP monitoring capabilities are maturing. Several products offer machine learning (ML) to enhance behavior analytics, identify the user account and gain deeper insight into the environment. Support for cloud-based databases, big data and NoSQL platforms (e.g., Apache Hadoop, MongoDB and Cassandra) are growing. Vendors continue to develop capabilities for algorithmic detection of malicious activity, due to hacking or insider misuse and emerging capabilities to help with monitoring and audit for privacy requirements.

User Advice: DAP provides a comprehensive, uniform and cross-platform database security suite across heterogeneous database environments. Clients should implement DAP functionality to mitigate privacy risks and data breaches resulting from user and administrator activities, database vulnerabilities, and poor segregation of duties (SOD), especially when DBMS vendors don't provide adequate capabilities to do so. DAP provides unique security functionality because it intercepts all communication paths to the database to then analyze and/or modify SQL commands and/or responses. It should be used as part of a broader, risk-based strategy by using the data security governance framework to help select and deploy complementary security products. Use DAP for five common use cases:

- **Unification of data security policies —** Do not rely on siloed and independent native database security functions. Apply and monitor unified database security policies across large-scale heterogeneous database environments to maintain data protection and privacy across geographic jurisdictions.
- **User activity monitoring —** Identify and assess the who, what, why, where, when and how of all users, including administrators and highly privileged application users. Monitor the activity of all users with data context to detect any privilege changes, as well as unusual data access and security policy violations, either accidental or malicious, that might lead to data breaches.
- **Enforcement of access controls —** Enforce the SoD of privileged and application users. If access to sensitive data is not permitted, then particular fields can, for example, be blocked or

redacted. Some vendors may even be able to anonymize the field (e.g., using masking, tokenization and encryption). If privileges change, access can also be blocked until verified.

- **Attack prevention —** Identify and mitigate open vulnerabilities in each DBMS and verify configuration or schema changes to prevent malicious activity. Applies virtual patches to block SQL attacks.
- **Audit and forensic analysis and reporting —** Use the audit report to provide a full record of activity for compliance reporting. Traditional database auditing has a significant impact on performance of the database servers; however, DAP technologies collect the same amount of data (or even more), with low to negligible performance impact. This is a best practice for all data stores in scope of regulatory frameworks that require audit records and breach notification. Many regulations require notification of datasets affected. Privacy laws, such as the General Data Protection Regulation (GDPR). require detailed reporting of data breach details within 72 hours.

If not provided by the vendor, data protection tools, such as format preserving encryption (FPE), tokenization and dynamic data masking, should be used in parallel with DAP to restrict access and protect data at rest and/or in use. This allows a greater focus on and monitoring of privileged users.

Business Impact: DAP is an important addition to enterprise data security governance programs, because it provides data context against user privileges and activity; other tools, such as IAM, SIEM or UEBA, do not. It is a critical investment to protect large and/or heterogeneous database infrastructures or Hadoop deployments containing regulated or business-critical data. It is important to address typical audit recommendations, such as enforcement of segregation of duties, vulnerability management, user activity monitoring and providing a forensic audit record of all activities. With increasing risks caused by data residency and privacy, hacking and insider abuse, DAP is a critical preventive and detective technology.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Beijing DBSec Technology; DataSunrise; IBM; Imperva; McAfee; MENTIS; Oracle; SecuPi; Trustwave; WareValley

Recommended Reading:

“Use the Data Security Governance Framework to Balance Business Needs and Risks”

“Market Guide for Data-Centric Audit and Protection”

“Securing the Data and Advanced Analytics Pipeline”

“The Future of the DBMS Market Is Cloud”

“Consuming DBaaS Securely: Comparing Options for Securing On-Premises and Cloud Databases”

Database Encryption

Analysis By: Brian Lowans

Definition: Database encryption solutions protect the column, table or database on-premises instances of relational database management systems (RDBMSs).

Position and Adoption Speed Justification: There is an increasing focus on database encryption as a risk-based access control (RBAC) by data protection and privacy laws and to address data residency issues. Encryption is growing in importance to minimize the risk of a data breach and to maintain privacy through data protection, the enforcement of segregation of duties (SOD) and access control.

User Advice: Authorized users and database administrators (DBA) with database access privileges have access to all data, unless encryption is used as a blunt-force data access control. Organizations should use the data security governance (DSG) framework to decide whether to implement encryption at the database, column or field level. They should also consider complementary tools to monitor and audit all user and DBA access to sensitive data with database audit and protection (DAP) tools. Although several RDBMS vendors are offering native transparent database encryption (TDE) capabilities (at the database, table or column), these are unique to that platform, with localized key management. They are typically managed by the DBA. Database encryption vendor products, however, apply security policies across multiple RDBMS platforms with a centralized enterprise key management (EKM). DBAs should not have management responsibility for encryption, but EKM will provide consistent security policies across the different RDBMS platforms. When considering database encryption, conduct a careful assessment to identify:

- What data needs to be protected, based on perceived risks, threats and compliance requirements?
- What is the overall data security policy? Should encryption be combined with DAP?
- How will SOD and access control be handled?
- Should encryption be applied to protect the whole database instance, the tablespace, or individual columns or fields?
- If protecting columns, is format-preserving encryption (FPE), tokenization and dynamic data masking (DDM) needed?
- How will the chosen database encryption integrate with an EKM strategy?

Most deployments focus on anonymization of specific types of regulated data — such as credit card numbers, personally identifiable information (PII), protected health information (PHI) and financial data. Mature users then branch out using risk-based approaches to include critical, but nonregulated data. Evaluate any impact on performance and functionality of applications accessing the RDBMS, and be aware that other security and database functionality, such as data discovery, can be affected.

Business Impact: When implemented consistently and aligned with the correct risks, database encryption, can offer a strong level of control against unauthorized access to data. It is a blunt-force tool that limits access to specific user accounts. However, more comprehensive security can be provided if combined with other tools, such as DAP. Consequently, concerns about the privacy of PII and PHI, data breach disclosure regulations, and the PCI Data Security Standard (DSS) are putting pressure on organizations to make greater use of encryption. Data residency across borders is also driving the need for additional anonymization options to protect data in use. Consider combining or replacing TDE with field protection methods such as FPE, tokenization or DDM to enforce stronger SOD. RDBMS encryption is increasingly recommended by auditors, but may not specify which method is required. Organizations need to use the DSG framework to review how other compensating security controls can be implemented as part of a broader, data-centric security strategy.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Sample Vendors: eperi; IBM; Micro Focus; NetLib; Oracle; Penta Security Systems; PKWARE; Protegrity; Thales eSecurity; Townsend Security

Recommended Reading:

"Use the Data Security Governance Framework to Balance Business Needs and Risks"

"Develop an Enterprisewide Encryption Key Management Strategy or Lose the Data"

"Prioritize Enterprisewide Encryption for Critical Datasets"

Document Store DBMSs

Analysis By: Merv Adrian

Definition: Document store DBMSs contain objects stored in a hierarchical, inverted, treelike format. Documents in these DBMSs often lack a predefined formal schema and do not have references to other documents within the collection. Documents are commonly self-described with JavaScript Object Notation (JSON).

Position and Adoption Speed Justification: Enterprise awareness of the benefits of document stores is widespread. Fifty-three percent of surveyed organizations in the 2019 OPDBMS Magic Quadrant survey are using document DBMS; another 12% plan to within 12 months. Enterprise features such as security, SQL-like query languages, backup and restore capabilities, and management tools are increasingly available. Competition is driving growth and innovation, especially in the cloud where MongoDB Atlas has shown strong growth and CSPs are trying to displace it. Leading vendors now all stress multimodel capabilities that broaden the addressable use cases.

The growth of use cases that require semistructured data and configurable levels of transactional consistency has advanced document store DBMSs along the Hype Cycle. Successes outweigh failure stories as products' maturity continues to grow, and widespread skills and knowledge have made deployment and operations easier. The continued growth in Gartner inquiries on this topic demonstrates the continued movement of document store DBMSs into the Plateau. However, many mainstream DBMSs also have fairly robust JSON processing capability, and other RDBMSs offer automatic JSON conversion. So document databases are not necessary for all JSON use cases.

User Advice:

- Planners and architects should continue to shortlist document store DBMSs for their schema flexibility based on use of the JSON data type, their price advantages and the wide appeal to developers. Traditional DBMS vendors' presence in the market raises their attractiveness further, and will make it easier to use competition to get the right features and price/performance.
- Design applications for document store DBMSs with their architecture in mind. A modeling approach different than the one normally used for RDBMSs is required. Simply porting applications and databases from an RDBMS to a document store DBMS is rarely advisable. Porting is not possible without significant investment in schema and application migration and redesign.
- Use mature document store DBMSs for agile web-scale applications that require large data stores with high performance, especially when activities are read mostly or not complex. For transactions that do not require ACID properties, and that have complex, mixed data types, these databases can be very effective. Where transactional uses are part of the picture, test performance carefully at appropriate scale to ensure SLA delivery.
- Note that while skills and applications are not typically transferable between document DBMS products, the conversion and compatibility tools from Amazon, Microsoft and others provide increasing flexibility. Bear in mind that the simplified model offered by document-style DBMSs frequently lacks the richness available in their relational counterparts.

Business Impact: The overall impact of document store DBMSs is widespread. In the move to the cloud, additional use cases are increasingly being seen as operational capabilities (such as backups and monitoring) improve and as data integration with data warehousing tools expands. Use cases such as network management data collection and event collection for logistics reflect this more mature, scalable profile.

Document store DBMSs allow for easy mapping to applications that must frequently scale dynamically and change rapidly. They are most applicable to web applications as the data exchange format is JSON-based. Certain Internet of Things architectures also have characteristics that make them ideal for document store DBMSs.

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Sample Vendors: Amazon Web Services; Couchbase; Google Cloud Platform; Hibernating Rhinos; IBM; MarkLogic; Microsoft; MongoDB

Recommended Reading:

"Assessing the Optimal Data Stores for Modern Architectures"

"Who's Who in NoSQL DBMSs"

"Decision Point for Selecting the Right NoSQL Database"

In-Memory Data Grids

Analysis By: Massimo Pezzini

Definition: In-memory data grids (IMDGs) provide a distributed, reliable, scalable and consistent in-memory object store — the data grid — that is shareable across multiple distributed applications. IMDG-based applications concurrently perform transactional and/or analytical operations in the low-latency data grid, thus drastically reducing the use of conventional, high-latency storage. IMDGs maintain data grid consistency, availability and durability via replication, partitioning and durable storage persistency.

Position and Adoption Speed Justification: IMDG technology is quite mature in terms of functionality, manageability, high availability and support of complex use cases. Atop the classic caching-oriented capabilities, IMDG products often provide in-memory DBMS, stream processing and real-time analytics capabilities, thus extending their use beyond the traditional advanced caching and high-performance transaction processing scenarios.

During the 15+ years since when this technology has emerged, providers have accumulated multiple thousands of clients worldwide. In terms of new customer adoptions, the market is mostly dominated by pure play vendors, often providing open source software-based products, as several general-purpose vendors either withdrew from the market or scaled down their commitment to IMDG, although some still enjoy a large and profitable installed base.

The requirements for scale-out architectures to support low-latency/high-throughput digital scenarios as well as new use cases such as stream analytics, along with their bundling into software products and cloud services, will continue to drive IMDG growth. Technologies such as containers make IMDG scale-out deployments easier, thus further favoring adoption by mainstream organizations.

However, the factors that have chronically inhibited widespread IMDG — including small skills pool, limited ISV support, deployment and management complexity, data stewards' conservatism — will continue to limit IMDG penetration in the less technically skilled organizations.

User Advice: Organizations in multiple verticals successfully leverage IMDG to enable large-scale business-critical systems, typically hyper-scale, high-performance transaction processing applications, such as digital banking, financial trading, e-commerce and travel reservation systems.

However, most recently, IMDG products have been extended with analytical capabilities. Therefore, IMDGs are increasingly used in large-scale event processing, real-time decisions and augmented transaction applications. Several vendors have also added to their products' in-memory DBMS capabilities, such as SQL support, thus expanding the use cases their IMDGs can address.

Therefore, application leaders responsible for modernizing their application infrastructure should consider IMDGs when they are implementing scale-out architectures to:

- Boost established application performance and scalability (thanks to IMDG support for popular APIs such as REST, Hibernate, JPA, JDBC and SQL)
- Enable high-scale/high-performance use cases such as the digital integration hub
- Develop native IMC applications, such as hyper-scale systems, low-latency event processing and real-time analytics
- Implement augmented transaction applications that support stream analytics use cases

Business Impact: By facilitating scale-out architectures, whether stand-alone or embedded in other software (e.g., application platforms), IMDGs help organizations retrofit established applications to:

- Enhance business process efficiency
- Speed up operations
- Support low-latency requirements
- Improve user/customer/supplier/employee satisfaction
- Extend application reach to larger user constituencies
- Enable event-based, real-time decisions

IMDGs also enable organizations to develop innovative, hyper-scale, high-performance and low-latency applications that are based on, for example, the augmented transaction, event brokering or digital integration hub paradigms, which often cannot be supported by traditional platforms alone.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

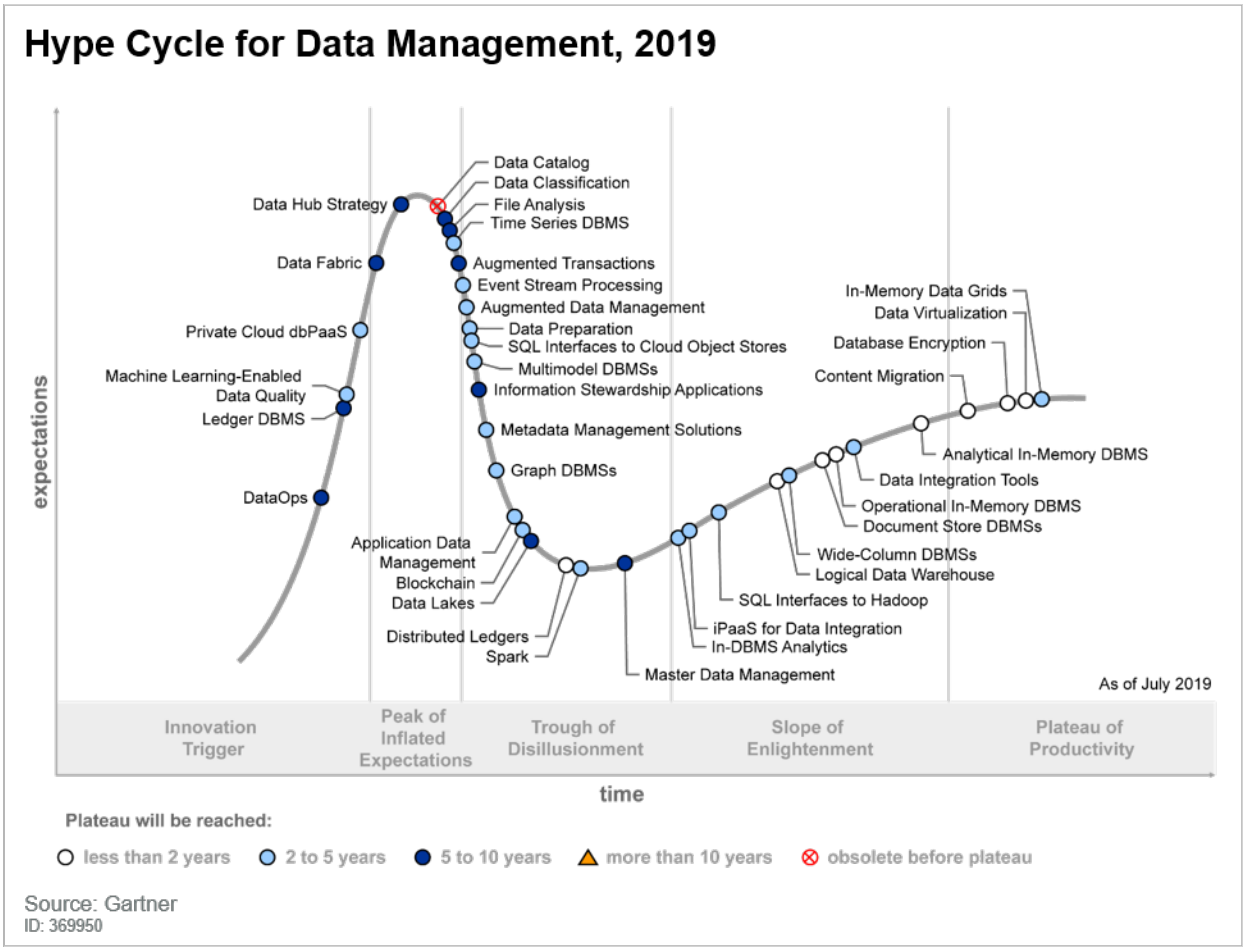
Sample Vendors: GigaSpaces; GridGain Systems; Hazelcast; Oracle; Red Hat; ScaleOut Software; Software AG; TIBCO Software; VMware

Recommended Reading:

"Predicts 2019: In-Memory Computing at a Turning Point, Driven by Emerging Persistent-Memory Innovation"

Appendixes

Figure 3. Hype Cycle for Data Management, 2019



Source: Gartner
ID: 369950

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 1. Hype Cycle Phases

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant press 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 technology is pushed to its limits. The only enterprises making money are conference organizers and magazine publishers.
Trough of Disillusionment	Because the technology 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 technology'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 technology 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.
Years to Mainstream Adoption	The time required for the technology to reach the Plateau of Productivity.

Source: Gartner (July 2020)

Table 2. Benefit Ratings

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 2020)

Table 3. Maturity Levels

Maturity Level	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 ■ Cost of migration constrains replacement	■ Maintenance revenue focus
Obsolete	■ Rarely used	■ Used/resale market only

Source: Gartner (July 2020)

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

Understanding Gartner's Hype Cycles

Magic Quadrant for Data Management Solutions for Analytics

Magic Quadrant for Metadata Management Solutions

Magic Quadrant for Data Integration Tools

Magic Quadrant for Data Quality Tools

Magic Quadrant for Master Data Management Solutions

Market Guide for Data Virtualization

"Market Guide for Data Preparation Tools"

More on This Topic

This is part of an in-depth collection of research. See the collection:

- 2020 Hype Cycle Special Report: Innovation as Strategy

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp