Taxonomy

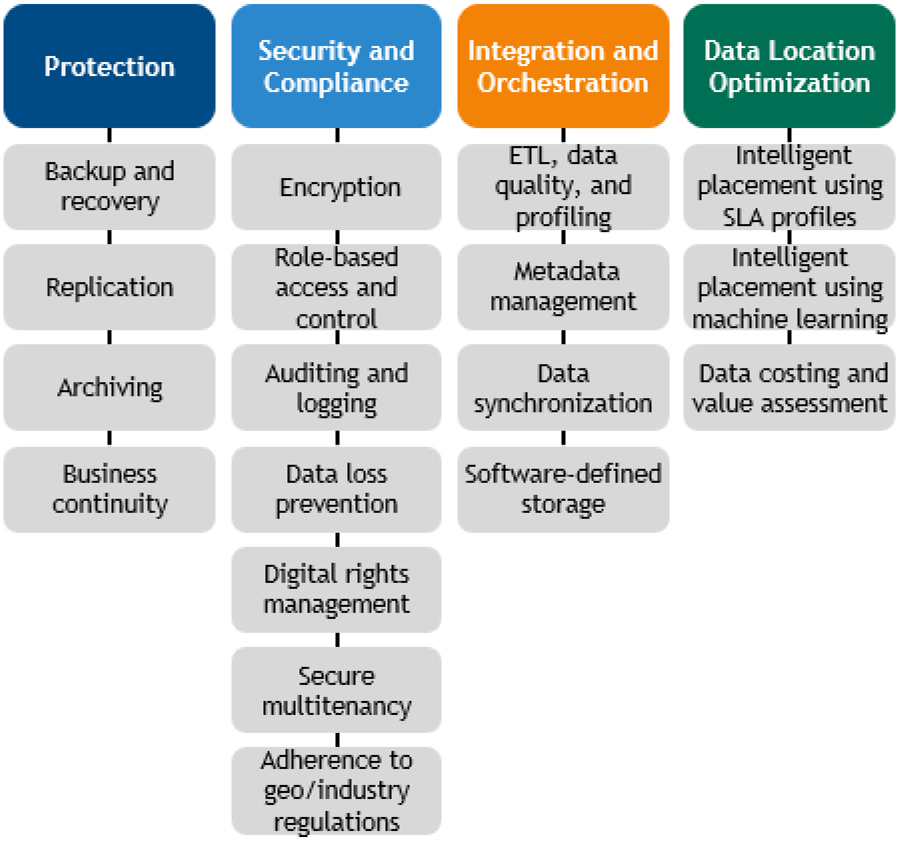
IDC's Worldwide Data Services for Hybrid Cloud Taxonomy, 2017

Phil Goodwin
Laura DuBois
Sean Pike
Ritu Jyoti
Dan Vesset

IDC'S WORLDWIDE DATA SERVICES FOR HYBRID CLOUD TAXONOMY

FIGURE 1

Data Services for Hybrid Cloud Primary Segments


Protection
Backup and recovery
Replication
Archiving
Business continuity
Security and Compliance
Encryption
Role-based access and control
Auditing and logging
Data loss prevention
Digital rights management
Secure multitenancy
Adherence to geo/industry regulations
Integration and Orchestration
ETL, data quality, and profiling
Metadata management
Data synchronization
Software-defined storage
Data Location Optimization
Intelligent placement using SLA profiles
Intelligent placement using machine learning
Data costing and value assessment

Source: IDC, 2017

June 2017, IDC #US42598817

DATA SERVICES FOR HYBRID CLOUD TAXONOMY CHANGES FOR 2017

Data services for hybrid cloud taxonomy is a new taxonomy created in 2017, so there are no changes to reference from.

TAXONOMY OVERVIEW

We are fast approaching a new era of the data age. IDC forecasts that by 2025 the global datasphere will grow to 163ZB (that is a trillion gigabytes). That's 10 times the 16.1ZB of data generated in 2016. All this data will unlock unique user experiences and a new world of business opportunities.

Data has become critical to all aspects of human life over the course of the past 30 years; it's changed how we're educated and entertained, and it informs the way we experience people, business, and the wider world around us. Powered by this wealth of data and the insight it provides, enterprises around the globe will be embracing new and unique business opportunities.

Data Age 2025 is an IDC study sponsored by Seagate, and it describes five key trends that will intensify the role of data in changing our world, namely:

- **The evolution of data from business background to life critical.** In fact, IDC estimates that, by 2025, nearly 20% of the data in the global datasphere will be critical to our daily lives and nearly 10% of that will be hypercritical. Critical meaning disruptive to our processes and stream of life such as the electricity going off. Hypercritical is potentially life ending such as data powering self-driving cars, embedded medical device monitoring biometrics, and data to inform adjusting insulin or shocking your heart.
- **Embedded systems and the Internet of Things (IoT).** As standalone analog devices give way to connected digital devices, the latter will generate vast amounts of data that will, in turn, allow us the chance to refine and improve our systems and processes in previously unimagined ways. New data-specific roles will also come to fruition. Big data and metadata (data about data) will eventually touch nearly every aspect of our lives — with profound consequences.
- **Mobile and real-time data.** Increasingly, data will need to be instantly available whenever and wherever anyone needs it. Industries around the world are undergoing digital transformation (DX) motivated by these requirements. By 2025, more than a quarter of data created in the global datasphere will be real time in nature and real-time IoT data will make up more than 95% of this.
- **Cognitive/artificial intelligence (AI) systems that change the landscape.** The flood of data enables a new set of technologies such as machine learning, natural language processing, and artificial intelligence — collectively known as cognitive systems — to turn data analysis from an uncommon and retrospective practice into a proactive driver of strategic decision and action. IDC estimates that the amount of the global datasphere subject to data analysis will grow by a factor of 50 to 5.2ZB in 2025; the amount of analyzed data that is "touched" by cognitive systems will grow by a factor of 100 to 1.4ZB in 2025!

- **Security as a critical foundation**. All this data from new sources opens new vulnerabilities to private and sensitive information. There is a significant gap between the amount of data being produced today that requires security and the amount of data that is actually secured, and this gap will widen — a reality of our data-driven world. By 2025, almost 90% of all data created in the global datasphere will require some level of security but less than half will be secured.

As data grows in amount, variety (structured, semistructured, or unstructured), and importance, business leaders must focus their attention on the data that matters the most. Not all data is equally important to businesses or consumers. The enterprises that thrive during this data transformation will be those that can identify and take advantage of the critical subset of data that will drive meaningful positive impact for user experience, solving complex problems, and creating new economies of scale. Business leaders should focus on identifying and servicing that unique, critical slice of data to realize the vast potential it holds. Data identification will require human data specialists in combination with cognitive systems.

Enterprises worldwide are contending with an accelerated pace of digitization where reliable and flexible information technology (IT) infrastructure could mean the difference between winning and losing customers. The groundbreaking agility, flexibility, simplicity, and power of cloud computing has businesses exploring ways to adopt cloud functionality and economics. Hybrid cloud deployments are the new norm.

The proliferation of application deployment models, including newer, cloud-native software as a service (SaaS), IoT, mobile, and hybrid cloud, plus traditional on-premises applications has resulted in organizational data being widely and unpredictably spread across multiple repositories. This proliferation of data types and repositories creates numerous and increasing challenges for IT staff, ranging from knowing what data is where to changes in fundamental data protection, security, governance, and infrastructure management. In many cases, organizations must purchase and manage numerous overlapping tools simply to address unique environments. Integrating the information from these similar, yet disparate, tools may be either time consuming or impossible. As a result, these organizations are not able to harness the value and the totality of the information within their organization, an issue that is even more important than the inherent inefficiency of this scenario.

Moreover, the lines between previously discrete IT activities are being erased. Data protection is a clear case in point, where backup, disaster recovery, and high availability are evolving from disciplines to points on a continuum. Similarly, security is a requirement that pervades nearly every element of the IT stack. This situation challenges product development requirements for vendors to provide needed functionality without succumbing to product creep. It also challenges end-user organizations to develop a coherent hybrid cloud data strategy of complementary, integrated, and cloud-enabled products/solutions that optimizes the value of organizational data.

Data Services for Hybrid Cloud Solutions

IDC has developed the data services for hybrid cloud taxonomy to assist both vendors and IT organizations develop a structure to the concepts of hybrid cloud data services and to articulate how the various components interrelate to form a common management platform. This taxonomy defines a *competitive market* rather than a functional market (see the Definition of Data Services for Hybrid Cloud section). As such, we do not expect any product or company to fulfill all elements. Rather, it is a combination of products, specific features/functionalities of products, and markets that add up to the totality of the taxonomy.

In the future, this taxonomy will be used to create a total competitive market value, which IDC will publish on an annual basis along with a forecast for growth. In addition, specific components that can be defined and quantified separately will be broken out accordingly. The empirical data used to build these market models will be derived from IDC trackers and QViews.

Definition of Data Services for Hybrid Cloud

Data services for hybrid cloud is location- and infrastructure-independent software that understands and performs various protection, security, integration, and optimization functions on data for the purposes of business control or SLAs. These functions can be performed in place or following data movement.

Data services for hybrid cloud covers data services (including protection, security, compliance, integration, orchestration, and data location optimization using SLA profiles or machine-learning-based cognitive/AI capabilities) that operate on structured, semistructured, and/or unstructured data and work across location (on-premises and public cloud stacks) and infrastructure.

Further, a central tenant of these solutions is their operation on data, specifically files, objects, or application instances rather than physical or logical storage constructs such as LUNs, volumes, or devices. Data services for hybrid cloud solutions do not provide the persistence capabilities themselves nor do they provide the schema, structure, or repository in which the data is stored. Data service functions operate on data that can reside in structured, semistructured, or unstructured repositories.

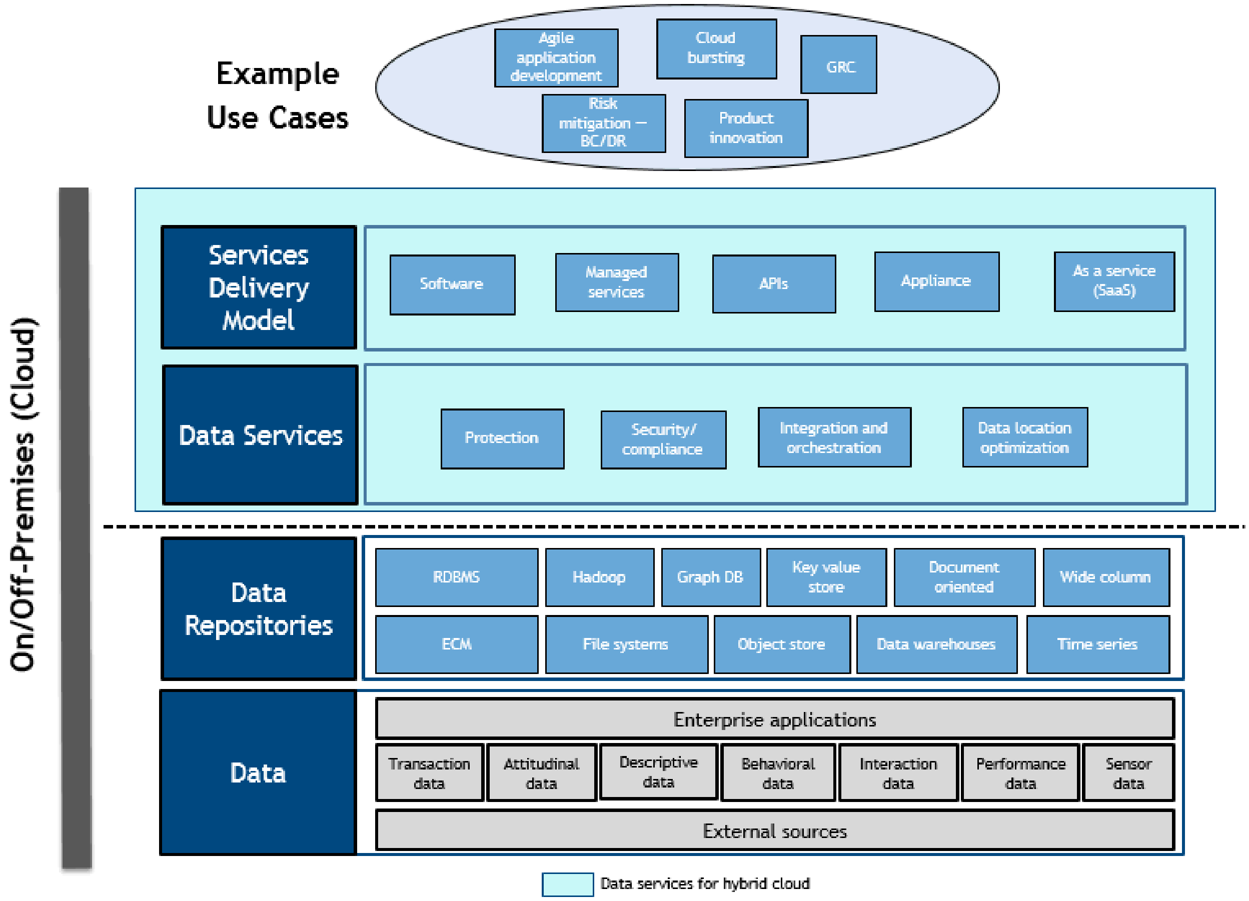
Figure 2 illustrates the relationship between major capabilities in data services platform for the hybrid cloud as they relate to use cases. It also identifies how these solutions, as they evolve and come to market, will be made available for consumption. While the innovation and development of these data services will be instantiated through software capabilities, this software may be offered for delivery in a variety of formats including:

- **Software.** Software can be deployed in traditional IT or private or public cloud infrastructure-as-a-service configurations. Data services for hybrid cloud software will be available on perpetual and software subscription terms, for installation on industry-standard computing platforms.
- **Managed services.** The software will be used to deliver managed services to customers that seek to gain greater control over their data security, protection, governance, and insight.
- **Application programming interfaces (APIs).** IDC expects that, in the longer term, as the market develops, data services for hybrid cloud capabilities will be made available through extensive APIs. This would allow for integration of the capabilities on one product with the capabilities of another product. For example the security capabilities of one product could be integrated with the tokenization/masking capabilities of another product. Rather than data services for hybrid cloud reinventing capabilities that already exist, the solutions would call upon services from other products as integration becomes available and would likely be OEMed.

- **Appliance.** Convergence in the datacenter across software and hardware and between compute, storage, and networking is being driven by reducing cost and accelerating time to market. Convergence is also occurring within software markets while blurring the line between data protection and security. It's expected that data services for hybrid cloud capabilities will be offered in appliance form factors.
- **Software as a service.** Security, data protection, and disaster recovery offerings are already available through SaaS delivery. As data services capabilities are built out, it would be a natural evolution for existing offerings to add these data services features and for new entrants with data services for hybrid cloud to make the capabilities available as a service.

FIGURE 2

Data Services for Hybrid Cloud — Conceptual Architecture



Source: IDC, 2017

Inclusion Criteria for Data Services for Hybrid Cloud Taxonomy

Data services for hybrid cloud covers data services (including protection, security, compliance, integration, orchestration, and data location optimization using SLA profiles or machine learning-based cognitive/AI capabilities) that operate on structured, semistructured, or unstructured data and work across location (on-premises and cloud stacks) and infrastructure. These data services will draw in part or in whole from the worldwide software functional markets as shown in Figures 3-6.

Listed here are the software/functional market inclusion criteria to this new competitive market:

- **Step 1**: Does the software operate on structured, semistructured, or unstructured data represented as files, objects, application files, or application instances and is stored in one of the data repositories noted in the conceptual architecture (refer back to Figure 2)?
- **Step 2**: If the answer to step 1 is "yes," then does the software operate on both in-place data and a replicated copy of the data?
- **Step 3**: If the answer to step 2 is "yes," then does the software work across location (on- or off-premises and across different cloud stacks) and across infrastructure?
- **Step 4**: If the answer to step 3 is "yes," then evaluate whether the software provides one or more of the following functions:
  - **Protection** including backup and recovery, replication (including copy data management and disaster recovery), archiving, availability, and clustering
  - **Security and compliance** including encryption (including tokenization and data masking), role-based access and control (RBAC), auditing and logging, secure multitenancy, and policy controls to meet geo/industry regulations
  - **Integrate and orchestrate** for integration of data between data repositories (This includes functions such as extract, transform, and load (ETL); data quality and profiling; metadata management; data synchronization across location [on-premises and cloud stacks]; and programmatically or on demand spin up resources.)
  - **Optimize data location** with intelligent placement of data using SLA profiles and predictive or cognitive/AI approaches as well as data costing and value assessment
- **Step 5:** Include in this new competitive market "data services for hybrid cloud" if the answer from step 4 is "yes" for at least one of the functions.

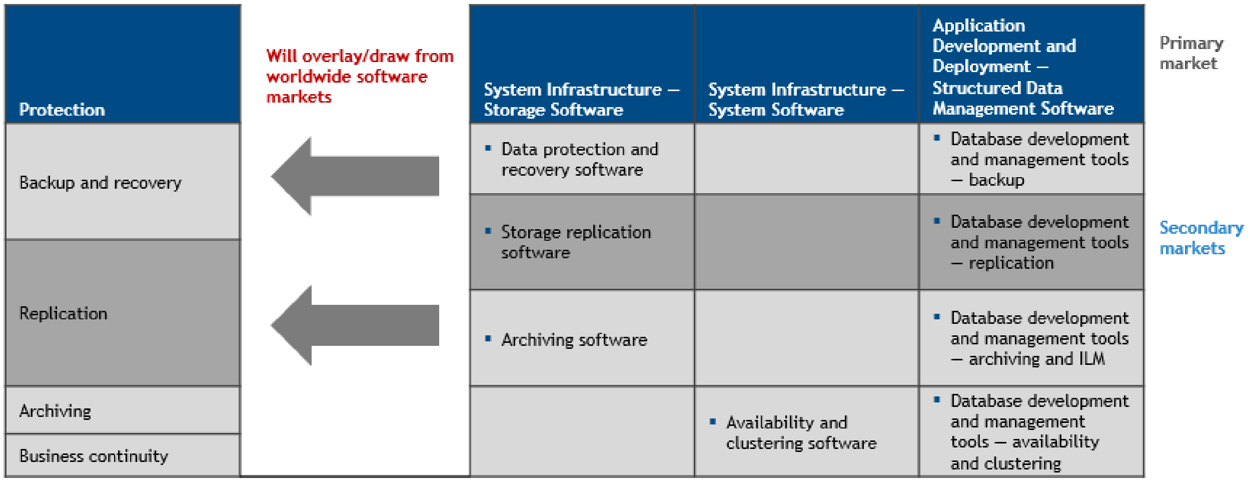
Protection

Data protection includes software and services related to the avoidance of data loss. These products, individually or collectively, address the range of data loss threats, including human error (i.e., accidental deletion), malicious attack, hardware failure, system software failure, datacenter plant and equipment failure (e.g., power, cooling, and environmental support), and entire site failure (e.g., fire, flood, and natural disaster). Note that this does not include the software used to detect such events, such as malware detection software, but the mechanism for assuring that data is not lost and is recoverable. This category includes both short-term and long-term retention (i.e., archiving) and the means of moving and storing data between locations to ensure data safety. Also included is software necessary to ensure data availability.

The protection data services will draw in part or in whole from the worldwide software functional markets, as shown in Figure 3.

FIGURE 3

Protection Data Services



Protection	System Infrastructure — Storage Software	System Infrastructure — System Software	Application Development and Deployment — Structured Data Management Software
Backup and recovery	▪ Data protection and recovery software		▪ Database development and management tools — backup
Replication	▪ Storage replication software		▪ Database development and management tools — replication
	▪ Archiving software		▪ Database development and management tools — archiving and ILM
Archiving			
Business continuity		▪ Availability and clustering software	▪ Database development and management tools — availability and clustering

Source: IDC, 2016

Backup and Recovery

Backup and recovery in this taxonomy includes software, which includes backup/recovery software, continuous data protection software, and modules that integrate snapshot or cloning capability with traditional backups. These solutions provide continuous and point-in-time copy functionality for defined sets of data to tape, disk, optical, or cloud-based data repositories. Representative vendor products are:

- Veritas NetBackup
- IBM Spectrum Protect

Backup as a service has emerged as a major cloud-related data protection scheme, whereby backup service providers deliver the mechanisms to transfer data from an on-premises repository to the cloud or protect data within the cloud (intracloud or intercloud). The cloud provider also manages the infrastructure and backup operations. Representative vendor products are:

- NetApp Cloud Control
- Barracuda Cloud-to-Cloud Backup

Replication

Storage replication includes hardware appliances (bundled hardware and software) as well as standalone software designed to create image copies of virtual machines (VMs), volumes, or files via techniques such as hypervisor-based replicas, clones, mirrors, and snapshots. Replication may reside on storage systems, application servers, and hypervisors; be fabric or appliance based; and occur locally or between remote sites (including cloud sources and targets), potentially separated by long distances. Replication and snapshot software are often used in conjunction with backup software to improve data protection or performance. This market does not include data replication software that operates at the database, table, or record level.

Replication — Cloud Gateways

Cloud gateways are bundled hardware and software appliances used to facilitate the transfer of data from cloud source (public, private, or managed) to a cloud target (public, private, or managed). They are analogous to PBBAs in that they are standalone devices designed for a single purpose. These devices differ from other hardware-based replication mechanisms (i.e., PBBA to PBBA or array to array) in that they translate data from the source software replicator to the target format using target-specific protocols or APIs (e.g., S3, OpenStack, and RESTful). Representative vendor products are:

- NetApp AltaVault
- CETRA cloud storage gateway

Replication — Host or Hypervisor-Based Replication Software

This software typically resides at the hypervisor, file system, or logical volume level within the operating system and makes a point-in-time copy or snapshot of the data set to persistent storage so it can be used for disaster recovery, testing application development, or reporting. In recovery, replication eliminates the intermediary step of a restore process. Representative vendor products are:

- Commvault Simpana Replication
- VMware vSphere Replication

Replication — Database Based

This software is used for maintaining an exact copy of a live database or a subset of a live database typically for recoverability, high availability, or nonstop maintenance purposes, or to distribute the database workload or isolate workload components by segregating them and assigning them to separate, replicated instances. Representative vendor products are:

- Dell SharePlex
- Oracle GoldenGate

Replication Management Software

This software is used to control, monitor, and/or schedule the point-in-time copies made by the replication product. It may automate various replication tasks, such as sync, split, and mount. Copy data management (CDM) is also included within replication management software. Copy data management software optimizes the number of data copies required, such that all use cases and service levels are served while eliminating superfluous copies of the original data. CDM typically includes data masking and role-based access to protect sensitive data. It may also include data discovery and mapping. Copied data refers to any copies of original data created by any mechanism such as snapshots, mirrors, and replication that are redundant to the primary copy. Copied data does not include RAID 10 copies. Representative vendor products are:

- Actifio CDS
- Dell-EMC Enterprise Copy Data Management

Archiving

Data archiving is often closely related to data governance, as data is retained primarily because it is required by law or regulation. Archiving is generally thought to be the retention of data for more than one year, although the actual practice is determined by the organization's requirements. Archived data is also generally considered to be unaltered (i.e., no longer used for application or transaction processing), though archive products and services may or may not enforce this through write once, read many (WORM) technology. The WORM capability is normally a separately available option. Representative vendor products are:

- Veritas Enterprise Vault
- EMC SourceOne

Archiving as a service has emerged as a major methodology for IT organizations to archive data in the cloud. Cloud providers manage the infrastructure needed to archive the data. Examples include:

- HPE Digital Safe
- Veritas enterprisevault.cloud

Database Archiving and ILM Software

Products in this segment are used to manage the evolution of data from its creation to removal from the database and include database sub-setting, data masking, and test data-generation tools as well as tools that build and maintain archives of databases, often allowing transparent access to archived data, preserving original schema information about archived data and intelligence for selecting referentially complete subsets of data for archiving. (Such products can also be used to create referentially complete subsets of databases for populating subset or test databases.) Representative vendor products are:

- Hewlett Packard Enterprise (HPE) HP Integrated Archive Platform
- IBM Optim

Business Continuity — Availability

In today's times, more and more organizations are looking for 24 x 7 access to their data and adherence to their SLAs and derive timely insights. Organizations need business continuity — availability and clustering software that virtualizes the system services of multiple systems (physical servers or virtual servers) and across location (on-premises or on public cloud) so that they appear in some sense as a single computing resource. They need failover clustering software, which maintains a "heartbeat" between the linked servers and restarts workloads on alternate servers if the heartbeat (or the lack of it) signals that one of the servers is offline. It also includes cluster managers and compute farm managers, as well as load balancing software and application virtualization software that stand between the user request and the processors or systems that are supporting applications or services. This software determines which processor or system has the most available capacity and routes the workload to that computing resource. Representative vendors and products include the following:

- Microsoft Windows Server 2008 Failover Clustering (formerly Microsoft Cluster Service [MSCS])
- Symantec Veritas Cluster Server (VCS)

Security and Compliance

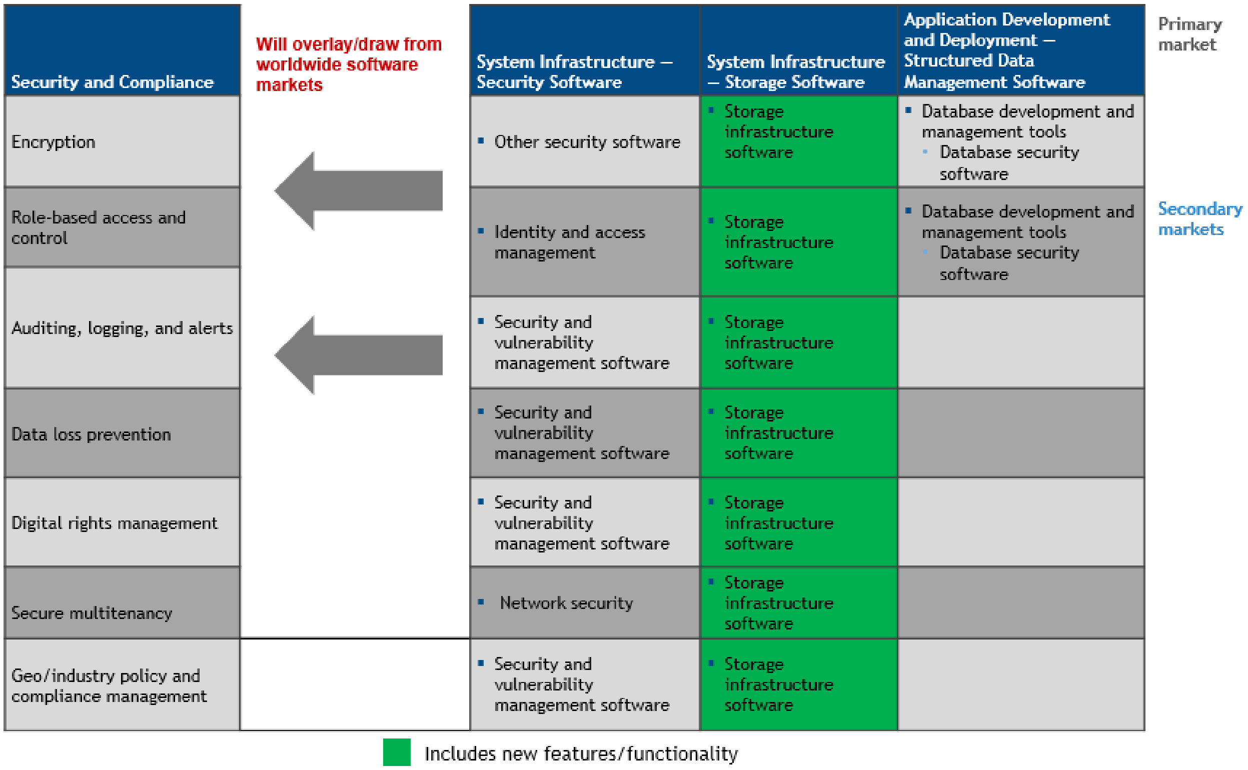
Data security and compliance are disciplines gaining ever more importance and attention as governmental regulations expand and penalties for noncompliance increase. In many cases, the regulations are jurisdictional, meaning that multinational companies must be prepared to treat data differently, depending upon the country in which data physically resides, and in some cases, ensure that the data does not cross an international border. Organizations must also take affirmative steps to ensure that data cannot be accessed by any not authorized to view it.

The burden of proof is on the company in most cases, making the ability to audit and prove compliance imperative. In the era of digital transformation, data is everywhere and data risks are compounded. IT has limited to no visibility on data access and is challenged to find data on time or expire it on time to comply with data regulations. This taxonomy takes into consideration the security and compliance for the data itself and does not include physical datacenter access or system access.

The security and compliance data services will draw in part or in whole from the worldwide software functional markets as shown in Figure 4.

FIGURE 4

Security and Compliance Data Services



Security and Compliance	Will overlay/draw from worldwide software markets	System Infrastructure — Security Software	System Infrastructure — Storage Software	Application Development and Deployment — Structured Data Management Software
Encryption		▪ Other security software	▪ Storage infrastructure software	▪ Database development and management tools ▪ Database security software
Role-based access and control		▪ Identity and access management	▪ Storage infrastructure software	▪ Database development and management tools ▪ Database security software
Auditing, logging, and alerts		▪ Security and vulnerability management software	▪ Storage infrastructure software	
Data loss prevention		▪ Security and vulnerability management software	▪ Storage infrastructure software	
Digital rights management		▪ Security and vulnerability management software	▪ Storage infrastructure software	
Secure multitenancy		▪ Network security	▪ Storage infrastructure software	
Geo/industry policy and compliance management		▪ Security and vulnerability management software	▪ Storage infrastructure software	

Primary market

Secondary markets

🟩 Includes new features/functionality

Source: IDC, 2016

Encryption

Current data management best practices require that data be encrypted both at rest and in flight. Failing to do so increases the risk that personally identifiable information (PII) or other restricted data can be viewed in plain text by unauthorized persons. Because data traverses numerous devices across location (on-premises and public cloud) as it moves through servers, disk arrays, switches, and so on, processing overhead is a price to be paid. It also raises the need in some cases for enterprise key management solutions.

Encryption can be software or hardware based and file, object, or volume/full-disk based, and techniques vary by product, including IES, AES, and NIST. Most of the modern infrastructure stack support software-based encryption for data at rest and in flight and integrate with standard commercial key management solutions. Some of them may provide basic key management functionality to simplify the process but are not intended as a standalone key management solution to be used by third-party solutions. Likewise, infrastructure-level management software integrate/bundle with data tokenization and masking software to ensure that the appropriate security checks are in place for the data copy made for test/data prep on-premises or in the public cloud. Tokenization is used to generate a random number or alphanumeric value out of plain text. It is a common technique used in credit card transaction systems. Representative vendor products are:

- Gemalto SafeNet
- HPE Voltage

Data masking solutions replace sensitive data with fictitious data elements in databases and file systems for use in development, testing and, mainly but not exclusively, nonproduction environments. Vendors that provide data masking solution are:

- Informatica Dynamic Data Masking
- IBM Optim masking

This segment also includes encryption software that provides obfuscation of database objects (tables/columns) to prevent exposure of information due to physical loss of the media.

When data is in motion, it is more susceptible to being breached or accessed for interception or unauthorized access. Examples of data in transit are users browsing the internet or accessing a database, migrating data throughout a virtual environment, and accessing third-party or hybrid cloud data from a private environment. Typically, the most common solution for data in motion is the use of Secure Sockets Layer (SSL) and Transport Layer Security (TLS), most often seen and referred to as using the HTTPS protocol. In addition, virtual private network (VPN) solutions (typically IPSec based) are utilized.

This market will also include the big data security segment that includes data protection products designed to specifically address digital security within dynamic data management systems, including scalable data collection managers (the most common being Hadoop) and dynamic DBMSs. Some solutions provide authenticated encryption, a management platform to manage the issuance and protection of keys, policy management, and access control.

This taxonomy does not differentiate between techniques, leaving that as a buyer criteria issue. Representative vendor products are:

- Gemalto Encryption for Enterprises
- Sophos SafeGuard Encryption

Role-Based Access Control

Access control is a foundational element of security. Functions such as RBAC enable organizations to determine who has data access and to what extent they have such access. This capability limits vulnerabilities and exploits, including data exfiltration and escalation of privileges.

RBAC at the infrastructure level allows administrators to limit or restrict users' administrative access to the level granted for their defined role. This feature allows administrators to manage users by their assigned role. For example, access to employees' salary, grades, or performance information may be limited to only HR and the employees' managers. Database security software enhance security through database log analysis for the detection of improper data access and database access control.

In today's multicloud and hybrid cloud environments, organizations need to ensure that access controls throughout cloud stacks are consistent with the organization's security policies. Infrastructure-level management software integrates with data access and policies to ensure that in-place/copied data is only accessible by the users/entities that require access to the data and to the extent to which they need that data. Representative vendor products are:

- Imperva SecureSphere Management
- Protegrity Data Security Platform

Auditing, Logging, and Alerts

Infrastructure systems have increased their footprint in today's threat landscape. Therefore, the visibility provided by audit functions remain critically important. Infrastructure software stack have increased the number of auditing events for files and objects. Creation, modification, deletion, successful access, failed attempts, folder permission changes, and so forth are logged and alerts are triggered to track anomalous activities. These are essential to protect data against malicious activities. Representative vendor products are:

- Varonis DatAlert
- Netwrix Auditor

Data Loss Prevention

Data loss prevention (DLP) technologies include a broad range of solutions designed to discover, monitor, and protect confidential data wherever it is stored or used. DLP includes solutions that discover, protect, and control sensitive information found in data at rest, data in motion, and data in use. The systems are designed to detect and prevent the unauthorized use and transmission of confidential information. Representative vendor products are:

- Digital Guardian Platform
- Forcepoint DLP

Enterprise Collaboration — Digital Rights Management

These solutions provide controls to enforce data governance policies, such as enterprise or digital rights management products and file synchronization and sharing platforms. These solutions typically contain reporting and analytics on user activity and content use. Representative vendor products are:

- Citrix ShareFile
- Varonis DatAnywhere

Secure Multitenancy

In the hybrid cloud and multicloud adoption era, organizations need to be overcautious and ensure that while they share same applications and hardware in the public cloud, the information from their respective logical environments is isolated and never shared. Multiple VMs share compute, storage, memory, and other resources. The sharing of such resources inherently increases the attack surface and therefore increases risk. Organizations need to understand how their service providers segment and isolate each customer in their cloud infrastructure. It is imperative because segmentation applies at the logical/application level as well as the physical level. Any solutions in the hybrid cloud data services platform, which are typically a subset of the infrastructure management software stack, should make sure that data from various users is segmented. Typically, administrators set the resource limits and QoS levels for each tenant. Tenants have administrative control of their provisioned environment while remaining isolated from others. Network segmentation makes sure of logical separation and therefore isolation and reduction of the attack surface while providing key points of visibility. Representative vendor products/product features are:

- VMware vShield
- NetApp ONTAP Storage Virtual Machines

Geo/Industry Policy Management

Governance, data privacy, and data sovereignty must be aligned. Because data stored in binary or digital form is subject to the conditions of the country in which the data resides, it is imperative that organizations understand the nature of their cloud architecture and interactions. Moreover, it is important to understand the rules, legislation, and regulations that apply to various countries/industries as they apply to personal information, data privacy, and data protection. For example, the Payment Card Industry Data Security Standard (PCI DSS) is a set of security standards designed to ensure that *all* companies that accept, process, store, or transmit credit card information maintain a secure environment. Similarly, the General Data Protection Regulation (GDPR) by the European Union (EU) addresses the export of personal data outside the EU. The primary objectives of the GDPR are to give citizens and residents back control of their personal data and to simplify the regulatory environment for international businesses by unifying the regulation within the EU.

With data spread across multiple on-premise and off-premise repositories, IT organizations need to understand where data is, what the data is used for, and what the data's life cycle is. Functionality will include data discovery, data mapping to applications, trends and forecasts, and capacity management as well as data destruction and the end of its useful life. While many products may not physically delete or destroy data, they should at least identify data that is eligible for destruction based on preset policies. Note that data visibility and control in this context does not include the data itself but rather information about the data.

Most of the infrastructure stack management software integrates with commercial GRC suites employed by customers to ensure compliance. They act upon synched as well as cloned copies of the data, whether it is on public cloud or on-premises. Some incorporate basic policy definition to support bare bones compliance validation. Representative vendor products are:

- ngCompliance Sherlock
- SAI Global (Compliance 360)

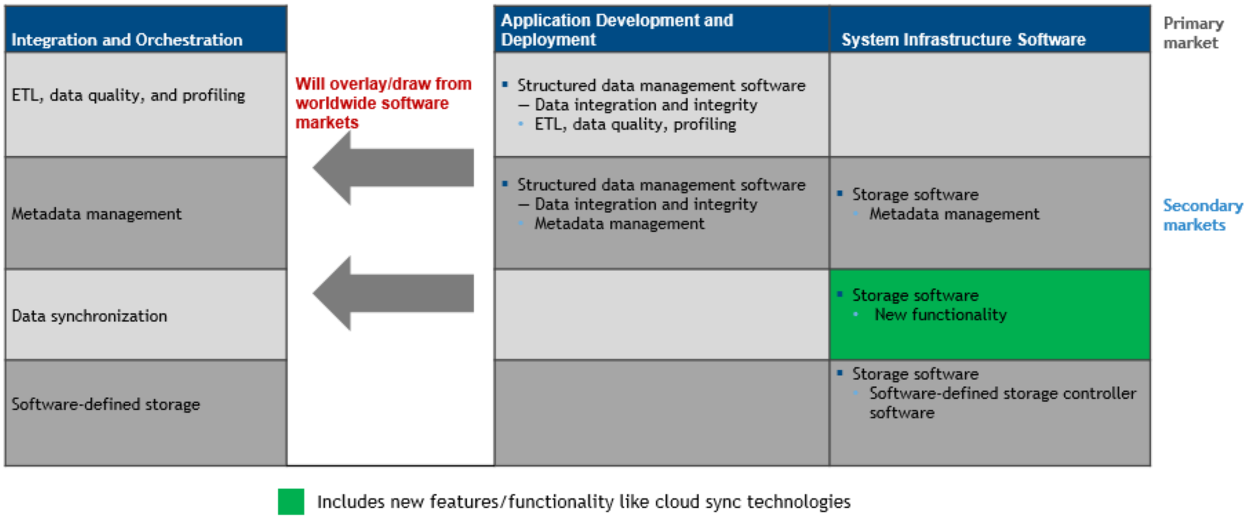
Integration and Orchestration

In a hybrid cloud setup, data portability is vital. Integration and orchestration services integrate data across repositories, perform bidirectional data movement across locations (on-premises and public cloud stacks), transform data to the appropriate format, optimize data placement, and make data accessible for additional data services (e.g., analytics).

The integration and orchestration data services will draw in part or in whole from the worldwide software functional markets as shown in Figure 5.

FIGURE 5

Integration and Orchestration Data Services



Integration and Orchestration		Application Development and Deployment	System Infrastructure Software	Primary market
ETL, data quality, and profiling	Will overlay/draw from worldwide software markets	▪ Structured data management software – Data integration and integrity ▪ ETL, data quality, profiling		
Metadata management		▪ Structured data management software – Data integration and integrity Metadata management	▪ Storage software Metadata management	Secondary markets
Data synchronization			🟩 ▪ Storage software ▪ New functionality	
Software-defined storage			▪ Storage software Software-defined storage controller software	

🟩 Includes new features/functionality like cloud sync technologies

Source: IDC, 2016

ETL, Data Quality, and Profiling

ETL software, commonly referred to as extract, transform, and load software, selectively draws data from source databases, transforms data into a common format, merges data according to rules governing possible collisions, and loads data into a target. A derivative of this process, extract, load, and transform (ELT), can be considered synonymous with ETL for the purpose of market definition and analysis. ELT is an alternative commonly used by database software vendors to leverage the inherent data processing performance of database platforms. ETL and/or ELT software normally runs in batch but may also be invoked dynamically by command. It could be characterized as providing the ability to move many things (data records) in one process. Representative vendor products are:

- IBM InfoSphere DataStage
- Informatica PowerCenter

The data quality software submarket includes products used to identify errors or inconsistencies in data, normalize data formats, infer update rules from changes in data, validate against data catalog and schema definitions, and match data entries with known values. Data quality activities are normally associated with data integration tasks such as match/merge and federated joins but may also be used to monitor the quality of data in the database, either in near real time or on a scheduled basis. The data quality software submarket also includes purpose-built software that can interpret, deduplicate, and correct data in specific domains such as customers, locations, and products. Vendors with domain-specific capabilities typically offer products that manage mailing lists and feed data into customer relationship management and marketing systems. Representative vendor products are:

- Experian Data Quality
- Talend Open Studio for Data Quality

Metadata Management

Metadata is data about data. It is commonly associated with unstructured data, as a structure in which to describe the content. Object storage offerings include comprehensive metadata management. However, it is increasingly being used in the structured data world, as the data becomes more complex in its definition, usage, and distribution across an enterprise. At a basic level, metadata includes definitions of the data — when, how, and by whom the data was created and last modified. More advanced metadata adds context to the data, traces lineage of the data, cross-references where and how the data is used, and improves interoperability of the data. Metadata has become critical in highly regulated industries as a useful source of compliance information. The metadata management submarket has grown out of the database management, data integration, and data modeling markets. Metadata management solutions provide the functionality to define metadata schema, automated and/or manual population of the metadata values associated with structured data, and basic analytic capabilities for quick reporting. These solutions also offer application programming interfaces for programmatic access to metadata for data integration and preparation for analytics. Representative vendor products are:

- IBM InfoSphere Information Governance Catalog
- Object Storage Metadata Management:
  - NetApp StorageGRID Webscale
  - EMC ECS
  - Scality RING

Data Synchronization

In a hybrid cloud/multicloud environment, organizations are looking to constantly move data back and forth between on-premises and public cloud to run desired cloud services (e.g., AWS EMR and RDS), and get the data back to wherever it is needed. The organizations need a simple, efficient, secure, and automated way to rapidly move data across locations while keeping the data in sync and without the complexity of data formats (e.g., files to AWS S3 and vice versa). Some solutions automatically trigger the return of the results back to the original location of the data, whether on-premises or in the cloud. Representative vendor products are:

- NetApp Cloud Sync
- Datadobi DobiSync

Software-Defined Storage

Much of the attraction associated with cloud is the ability to provision resources on demand. Doing so requires a class of products designed to provision these resources, regardless of location (on-premises or public cloud) and infrastructure. Software-defined storage (SDS) fundamentally alters how storage systems are provisioned. IDC refers to software-defined storage as complete systems that deliver the full suite of storage services via a software stack that uses (but is not dependent on) commodity hardware built with off-the-shelf components. SDS solutions should offer a full suite of data access interfaces, storage, and data management services (included federation services). Software-defined storage controller is extensible and autonomous and allows data access via known and/or published interfaces (APIs or standard file, block, or object interfaces). A rich set of standard SDS APIs enable agile provisioning and application/data portability across on-premises and public cloud. For example, developers can use self-service SDS APIs to easily spin off resources on public cloud for their innovation/development and then move back to on-premises if and when desired. Representative vendor products are:

- NetApp ONTAP Cloud and ONTAP Select
- IBM Spectrum Accelerate and Spectrum Scale

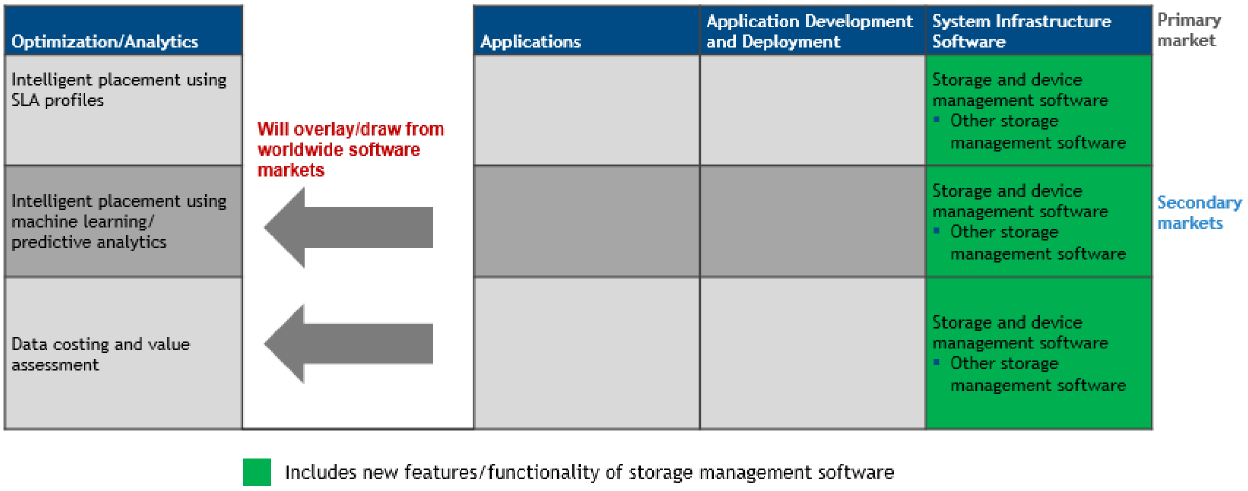
Data Location Optimization

In the era of digital transformation, organizations worldwide are looking to do more with less and are exploring to optimize data location for cost, availability, and reliability. Various tools and technologies are being leveraged to do data costing and value assessment and automate the placement.

The data location optimization services will draw in part or in whole from the worldwide software functional markets, as shown in Figure 6.

FIGURE 6

Data Location Optimization Services



Optimization/Analytics		Applications	Application Development and Deployment	System Infrastructure Software	Primary market
Intelligent placement using SLA profiles	Will overlay/draw from worldwide software markets			Storage and device management software ▪ Other storage management software	
Intelligent placement using machine learning/ predictive analytics	←			Storage and device management software ▪ Other storage management software	Secondary markets
Data costing and value assessment	←			Storage and device management software ▪ Other storage management software	

🟩 Includes new features/functionality of storage management software

Source: IDC, 2016

Intelligent Data Placement Using SLA Profiles or Machine Learning

With the deluge of data and that data being varied and spread across locations (on-premises and public cloud stacks), it is impossible for humans to optimize data location for cost, availability, and reliability. Data optimization using predictive analytics and machine learning is being put to use. "Learning systems" that interrogate metadata, analyze access patterns, and learn from the changing context of data are being used to help assign appropriate data value. Administrators can help train the "learning system" by providing sample files and labelling types of data as having different value. This could drive data-value-based dynamic decision on storage media (flash, disk, or tape) and location (on-premises or the cloud). Automated data tiering can be applied to public cloud storage as well and could use the same technology to automatically tier data between say AWS EBS block storage and AWS S3 object stores. A hybrid cloud data tiering could support all-flash arrays (AFAs) on-premises and AWS S3 object stores in the public cloud. It could also help dynamic policies-based data life-cycle management. There machine learning techniques are applied and data is analyzed for improvements in usage, service, and placement outcomes. The benefits include not only improving service and support but also allowing customers to organize and optimize their data management practices.

Currently, most of the infrastructure software stack offerings do policy-driven data placement along with a basic understanding of the access patterns of the data, enabling the stacks to intelligently and automatically move data to the location or storage tier that best suits the observed access patterns. IDC believes that, in the longer term, the vendors will enhance these capabilities by either integrating with cognitive applications like IBM Watson or building their capabilities organically.

This market would also include policy-driven automated storage tiering software that enables automated movement of data sets between differing tiers of storage resources. This may occur at the volume level or at a sub-volume level. Tiers may be defined by performance, capacity, and/or resiliency requirements of the data/applications. These capabilities will help the customer benefit by getting the best of all the worlds: SSD-level performance for hot data, HDD prices for capacity, geographic and industry compliance by leveraging the right cloud, leverage of cloud-based analytics and, above all, automatic data placement to exploit all the benefits and cost savings. Representative vendor products are:

- Komprise Intelligent Adaptive Data Management
- NetApp OnCommand Insight (Machine Learning Based) and NetApp ONTAP FabricPool (SLA Based)

Data Costing and Value Assessment

Organizations worldwide are scrambling to assess the value of data and get cost metrics based on data location (on-premises or public cloud) and storage media (HDD, SSD, memory) as well as the associated services. Organizations are using these to make the appropriate selection of cloud services and seamlessly adhere to industry and geographic regulations.

Most of the infrastructure providers are building this capability in their hybrid IT stack in partnership with the service providers, and some are making the capability available as a SaaS offering. Representative vendor products are:

- AWS Trusted Advisor and AWS Cost Explorer
- Cloudyn for Enterprise

Industry Developments

Suppliers in infrastructure-related markets (incumbents and start-ups, as well as on-premises or cloud stack offering providers) are stepping up to help businesses harness the wealth of data, create new value, improve customer experience, and gain competitive edge with limited time, skills, and budget. They are taking a data-centric view for the core infrastructure management functions and consolidating them along with important data-related functions built organically or in partnership. They are working to provide a simple and consistent set of tools/technologies across data formats, infrastructure, and location (on-premises and across cloud stacks). In the short term, the offerings are/will be focused on unstructured data, and longer term, some suppliers have the vision to expand their offerings to structured data. Likewise, some of the offerings currently work on-premises and on a public cloud stack. Longer term, the offerings would be expanded to multiple cloud stacks. No single vendor product offers all the functions described in this competitive market today. Examples of early entrants launching data services for the hybrid cloud include:

- **Commvault.** The cornerstone of Commvault's Cloud Data Management is its data protection software, realizing the backup data sets can be leveraged for a multitude of different purposes. Commvault provides cloud data protection, cloud DR, cloud workload migration, cloud dev/test, and process automation. The company has also developed a set of APIs into its backup repositories to enable third-party vendor solutions, such as data classification and discovery.
- **Datadobi.** Datadobi is an enterprise-class software for heterogeneous file and object migration and replication. DobiSync synchronizes file and object data between heterogenous storage

systems either on-premises, remote, or in the cloud easily, quickly, cost effectively, and accurately. DobiMigrate enables movement of data between storage platforms. DobiReplicate replicates file and object data between heterogenous storage systems either on-premises, remote, or in the public cloud.

- **Dell-EMC.** Dell-EMC's data protection solutions for Federation Enterprise Hybrid Cloud synthesize the strengths of private and public cloud, enabling enterprise IT to perform the central role of cloud services broker. IT gains the control and visibility it needs while business users are empowered to provision standardized protection services to meet their business needs. Further, EMC's broad portfolio of data protection, archiving, and copy data management offerings together with RSA offerings present an opportunity.
- **Komprise.** Komprise is an intelligent analytics-driven data management solution that analyzes data growth and usage across a customer's current storage, projects the ROI of moving inactive/cold data to secondary storage such as cloud/object, and then moves data based on customer-defined policies transparently.
- **NetApp.** NetApp's vision for data management is a data fabric that seamlessly connects different clouds, whether they are private, public, or hybrid environments. Data Fabric simplifies and integrates data management across cloud and on-premises to accelerate digital transformation. It delivers consistent and integrated data management services and applications for data visibility and insights, data access and control, and data protection and security. NetApp is leveraging the right combination of technology, partnership, business model, and vision to be at the forefront of delivering integrated and consistent data services for the hybrid cloud.
- **Varonis.** Varonis storage solutions detect insider threats and cyberattacks by analyzing access events (from Common Event Enabler [CEE] and many other systems); prevent disaster by locking down sensitive and stale data, reducing access, and simplifying permissions; and sustain a secure state by automating authorizations, migrations, and disposition. Varonis has long-standing partnerships with leaders in the file-based storage market such as NetApp and Dell-EMC.
- **Veeam.** Veeam's cloud data management strategy revolves around its data protection for virtual machines and workload migration. The company also offers Veeam Availability Orchestrator to assist partners in using Veeam for automating DRaaS failover, non-disruptive DR testing, and compliance documentation. Also available is Veeam DR for Azure, which is made up of Veeam PN (powered network) to automate the establishment of network connectivity from a datacenter to Azure.
- **Veritas.** The Veritas 360 Data Management offering provides unified data protection with visibility, enables cost-effective long-term data retention, and provides unified resiliency – predictive recovery for applications and integrated copy data management for rapid and secure access to data.

DEFINITIONS

- **Functional market.** Discrete, measurable total sales value for specific product types for which both vendor market share and growth rates can be established

- **Competitive market.** A combination of functional markets reflecting the consumption patterns of product categories by end customers for which total value and growth rates can be established, but not individual vendor market share
- **Cloud-native application.** Applications developed specifically for deployment in public or managed cloud environments, relying on virtual computing, cloud storage, cloud protocols (e.g., RESTful, S3, and OpenStack) designed to deliver on-demand availability of resources, and services (scale up or down) via a subscription engagement model
- **Hybrid cloud.** Application deployment environment that utilizes both on-premise private cloud resources (i.e., local datacenter) and off-premise public or managed cloud resources to deliver the totality of the application functionality
- **Multicloud.** Infrastructure deployment environment that utilizes two or more off-premise public or managed cloud resources for complete or partial application delivery
- **Data services.** Location and infrastructure-independent software that understands and performs various protection, security, integration, and optimization functions on data for the purposes of business control or SLAs (These functions can be performed in place or following data movement.)

LEARN MORE

Related Research

- *Worldwide Disk-Based Data Protection and Recovery Forecast, 2017-2020: Continued Moderate Growth* (IDC #US42237416, March 2017)
- *IDC's Worldwide Storage Software Taxonomy, 2016* (IDC #US41593216, July 2016)

Synopsis

This IDC study examines and defines the hybrid cloud data services (including protection, security, compliance, integration, orchestration, and data location optimization using machine-learning-based cognitive/AI capabilities) that operate on structured or unstructured data and work across location (on-premises and cloud stacks) and infrastructure to help data architects, LOB data owners, DevOps, and IT accelerate digital transformation (DX) within their organization.

"Digital disruption is real. According to Innosight, 75% of S&P 500 will be replaced by 2027. To survive, companies need to embrace and accelerate DX. Leading digital organizations are exploiting data insights to deliver personalized value services, optimize customer experience, explore new opportunities, and reduce the overall cost of doing business. The world's most admired and best-run businesses use IT for their competitive advantage. The groundbreaking agility, flexibility, and power of cloud computing has businesses exploring ways to adopt cloud functionality and economics. Hybrid cloud/multicloud deployments are becoming the new norm." – Ritu Jyoti, research director, IDC's Storage team

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com