Data Management for Research

Adam Robyak, Field CTO and Principal Engineer
Dr. Jeffrey Lancaster, Senior Higher Education Strategist

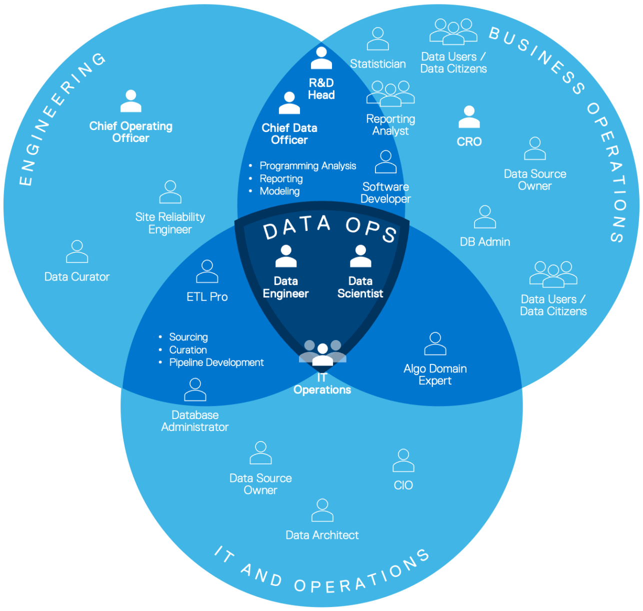
DELL Technologies

Data Management for Research

For researchers, the process of collecting information to formulate a hypothesis, conduct experiments, or analyze and iterate upon a research program can be a daunting task. The challenge compounds when the use of advanced technologies and big data are included, and it only gets more difficult with increased pressure from regulations and security constraints. To address these challenges, it is imperative to employ **data management** strategies. But what are best practices for data management in an ever-evolving landscape of approaches, tools, and threats?

What is Data Management?

Data Management comprises all of the various disciplines related to managing data as a valuable resource (Data Management, n.d.). It should not be a surprise that data is increasingly integral to research now and into the future, and so it should also not be a surprise that many products, solutions, and applications have been developed to work with research data. Indeed, the rapid growth of the amount of data available to researchers is a driving force behind the ever-increasing pace of innovation in research and technology. Cutting-edge methodologies such as advanced analytics and artificial intelligence are being applied to data and leading to new insights. And although the research practice itself is benefiting from these advancements, new challenges are also surfacing; data is becoming increasingly distributed and it often requires stringent integrated security models that address other regulatory, privacy, and emergent challenges. Data management, therefore, is the process by which data is effectively and securely orchestrated in a cost-efficient manner.



DATA ENGINEERS
Data specialists sourcing / curating, rationalizing, integrating, moving and pipelining data for data science and business analytics.

DATA USERS
Business unit executives and teams seeking analysis and interpretation of data / reporting / data models to make revenue, growth and optimization decisions.

IT OPERATIONS
Information and software specialists supporting centralized and decentralized data science, data management, data protection and IT modernization initiatives.

DATA SCIENTIST
Trained scientists (various levels) leveraging vast data and datasets to activate data models to create unique revenue and operational insights for Business Operations.

Figure 1. Emerging data management archetypes

Within organizations and educational institutions there are four key personas who contribute to data management (**Figure 1**):

1. **Data Engineers** – Data specialists sourcing, curating, rationalizing, integrating, moving, and pipelining data for data science and business analytics.
2. **Data Scientists** – Trained scientists at various levels leveraging vast data and datasets to activate models to create unique revenue and operational insights for business operations.
3. **IT Operations** – Information and software specialists supporting centralized and decentralized data science, data management, data protection, and information technology modernization initiatives.
4. **Data Users** – Business unit executives and teams seeking analysis and interpretation of data, reporting, data models to make revenue growth and operations optimization decisions.

Each archetype has their own processes, purpose, and challenges; and each interacts with the others in a variety of ways:

Data engineers work to make data available and accessible to their internal customers: the data scientists and data users:

> "...for our users, the big challenge they have is being able to understand **what data is available and how to get access to data.**"
>
> —Director of IT Architecture, United Healthcare

Meanwhile, data scientists often struggle to get data from different sources to be able to craft insight:

> "**As data scientists, the biggest challenge is having to drag data from different silos**. We preferred not to pull the data from one place to the another, and then another to the third place. We don't want to do separate task at each level, we really want to be able to pull the data from multiple sources into one centralized level so that we can we can do everything in one place and not, you know, depend on the hardware or the environment to be able to limit us in doing what kind of tasks we do with our data."
>
> — Senior Optimization and Data Scientist, Delta Airlines

And for IT operations, getting a comprehensive view of all available data within the organization can prove daunting:

> "...for the past few years our biggest challenge has been **first to understand all the different data we have within the organization**. With 350 locations around the world, this effort alone took about 2 years to understand what data we had, to standardize the view and provide the organization visibility. It was then we could really show people what's possible with data."
>
> — Director of IT Magna Corporation

Table 1. Business and technical challenges for the key archetypes.

Business Challenges	Technical Challenges
· **Managing Risk** of all Data · **Grow and Realize** Value of Data · **Defining Policies and Standards** for Governance, Quality, and Normalization · **Visualize Compliance** of Platform and Data Standards · **Protecting IP** while leveraging the operational efficiencies of data management and data science · **Maximizing resource investment** in tools, technologies, and people to produce tangible and differentiating outcomes for **top use cases** · Leveraging **data management and data science best practices** to impact top and bottom-line operations	· **Find and Catalog** the Data and Data Sources siloed within the enterprise · **Preparing the Data** to the business and Data Scientist standards · **Data Cleansing & Data Quality** · **Enriching the Data** with Learnings and Metadata · **Simplifies Integration of Data** to Data Science Platforms/Workbenches · **Flexibly Create Data Pipelines** for Productization · **Simplifies Deployment, Management, Upgrade, Support and Scaling** of Infrastructure and Services needed by the CDO, Data Engineer, and Data Scientist · **Flexibility and Ease of Deployment** and Supports an Edge, Core, Multi-Cloud Scenarios · **Automatic Integration** of Data Sources both internal and external

The common challenges among all four personas align to the business and technical management of data. From a business perspective, all must manage risk, ensure compliance, and define standards and policies. Each finds an imperative to manage growth while maximizing investment in resources and protecting intellectual property. In parallel, common technical challenges include collecting, preparing, and cleansing data. All stakeholders must think through how to simplify integration to support merging with data pipelines, and optimization through automation and orchestration can be a tremendous opportunity. Finally, underlying the business and technical management of the data, the infrastructure on which the data ecosystem is built must be scalable, resilient, and easy to manage.

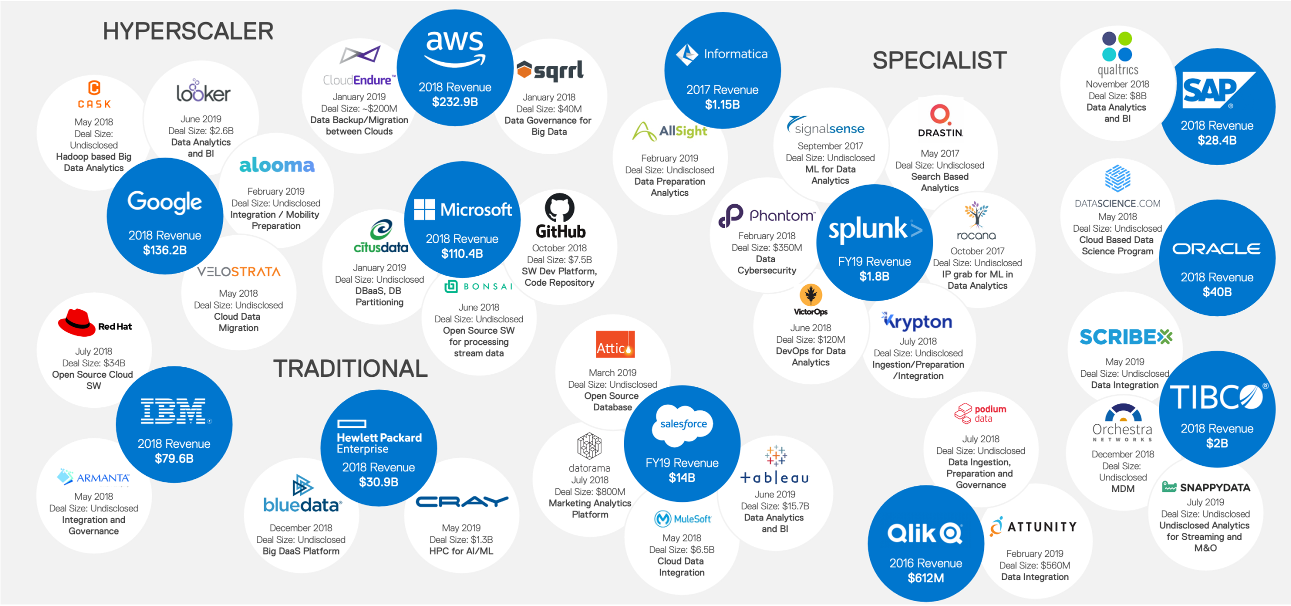


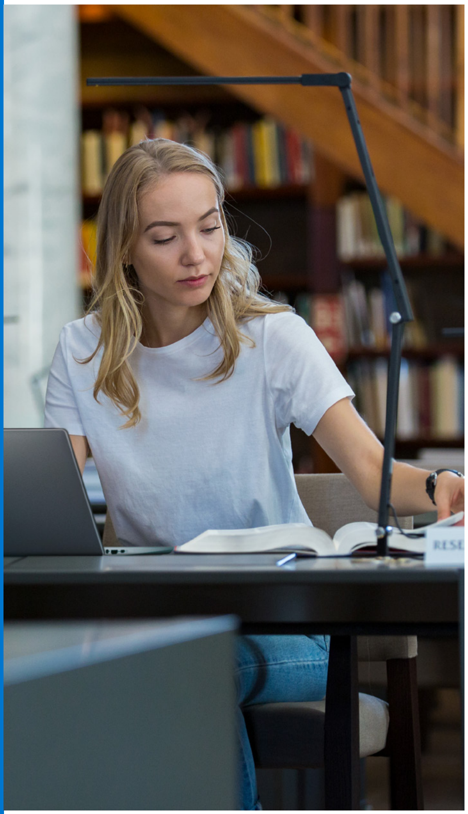
Figure 2. Consolidation of the data management industry

Over the past decade, the breadth of products and solutions in the data management ecosystem grew vast in response to these various challenges; it is a multi-billion-dollar industry with heavy competition and an often-confusing landscape.[1] Although that expansion has now given way to a period of contraction and consolidation, the ecosystem is, if nothing, ever-evolving and it still continues to shift rapidly: mergers, acquisitions, and displacements all impact the tools and platforms used to manage information **(Figure 2)**. New hardware and software tools can quickly upend best practices for how data ought to be managed.

To better understand the relevant tools, trends, and current best practices, it can be useful to break apart the monolithic concept of Data Management into a set of interconnected component parts. Taken as a whole, these components provide a structure through which to understand how the evolution of data management is impacting how research can be constructed and conducted, the skillsets necessary for the data science personas mentioned above, and what may be on the horizon for the data management ecosystem:

1. Data movement
2. Data locality
3. Metadata management
4. Data integration
5. Search capabilities
6. Data catalog(s)
7. Data pipeline(s)
8. Policy and governance
9. Intrinsic security and trust

[1] Research indicates that revenue from data science platforms and tools should reach $39.4 Billion in 2023. Artificial intelligence (AI) and machine learning (ML) are primary drivers of market growth for compute, networking, and storage, and data scientists and the ecosystems they leverage greatly influence buying decisions for research departments, organizations, and educational institutions.



Data Movement

Data doesn't come pre-packaged to include everything a researcher might want as they construct a model to understand the world, and it doesn't exist in a single location; data is increasingly located in disparate and disaggregated silos. A key first step for researchers is to identify, collect, and aggregate the data that they believe is necessary to build their deliverable. But how does one "get" data?

Data movement covers the ways data is moved from point A to point B, from source to destination. More broadly, it is both the process and workflow by which data is delivered and the set of tools and technologies supporting the complete data lifecycle. Functionally, data movement serves as an enabler of a data management strategy rather than a complete solution; it comprises a rich set of services that maximize the efficacy of a data management strategy. Data movement is responsible for populating data warehouses and data lakes, exchanging information between applications or with business partners, and ensuring that the right data is available to researchers so they can prepare it for subsequent use. Streaming, real-time analytics, and Internet of Things are other components of data movement.

> "**Data is increasingly being kept in hybrid cloud environments** that extend outside the traditional company firewall yet must continue to follow a common set of retention, security, and access control policies regardless of physical location." (451 Research, 2019)

> "**Through 2022, data management manual tasks will be reduced by 45%** through the addition of machine learning and automated service-level management." (Gartner, 2019)

Two trends are likely to impact how data movement will evolve over the coming years. First, organizations are adopting hybrid cloud environments in which data is stored in both on-premise infrastructure as well as with cloud providers, on remote devices, in sensors, and at edge gateways on top of on-prem and cloud services. As researchers seek to use that data, it will need to be both accessible and secure, no matter where it is stored. Second, machine learning is increasingly being used to automate manual tasks that had previously been the responsibility of IT professionals. As a result, those IT professionals can expect to spend less time on rote processes and more time monitoring resource allocation and troubleshooting at a distance.

Data Locality

The researcher's ability to leverage data will be severely disrupted if data is too siloed to obtain or too decentralized for the researcher to reach in a timely fashion. **Data locality** – where data resides and how it is accessed – determines how data movement is structured and will impact other components of the overarching data management strategy. Whether that data is generated and stored in the cloud, in a data center, on the edge, or somewhere in between, understanding data locality is critical to any data management strategy.

Edge computing is a newer consideration that has emerged in response to decentralized IT, Web 3.0, and disaggregated data where the computational advantage comes in pre-processing data so only key data, aggregate data, or pre-analyzed data is transmitted from the edge back to a data center. And in some cases, data doesn't need to make a round-trip to a data center; it can be wholly processed at the edge. Edge computing can be employed for a range of applications, from AI and analytics to inference and localized learning. Edge systems can also provide data aggregation from multiple endpoints and they can act as relays or nodes in a distributed network.

> "**Within two years**, half of all surveyed enterprises will be performing real-time data analytics at the edge." (451 Research)

> "**By 2022, over 40% of organizations' cloud deployments will include edge computing**, and 25% of endpoint devices and systems will execute AI algorithms." (IDC, 2019)

> "**By 2022, more than 50% of enterprise-generated data** will be created and processed outside the data center or cloud." (Gartner, 2019)

The focus on data locality is being impacted by technologies such as 5G and Next-Gen communications because low-latency, high capacity transactions are enabling more data to move through a network through "always on" mechanisms. As more data flows through a network, it may ultimately be less critical to consider a priori where data is located because it can be transmitted anytime, to anywhere.

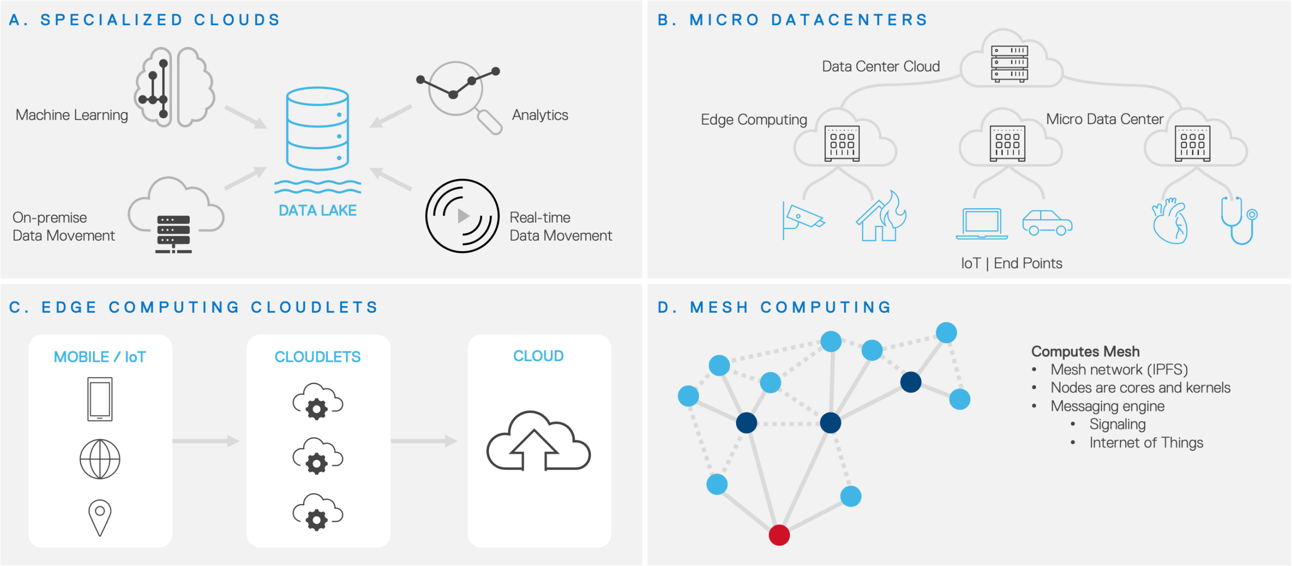


Figure 3. Emerging models of cloud.

Additionally, data locality needs are contributing to an evolution of cloud stacks. Beyond the traditional models, 4 new models of cloud are emerging (**Figure 3**):

- Specialized cloud models for solutions like High Performance Computing, data lakes and analytics, and even advanced DevOps
- Micro datacenters built for specific purposes that can be deployed in non-traditional environments
- Edge computing "cloudlets" intended to solve particular architectural needs for boutique, private, or specialized clouds outside of traditional cloud
- Mesh computing for disaggregated systems that are truly untethered compared to traditional architectures

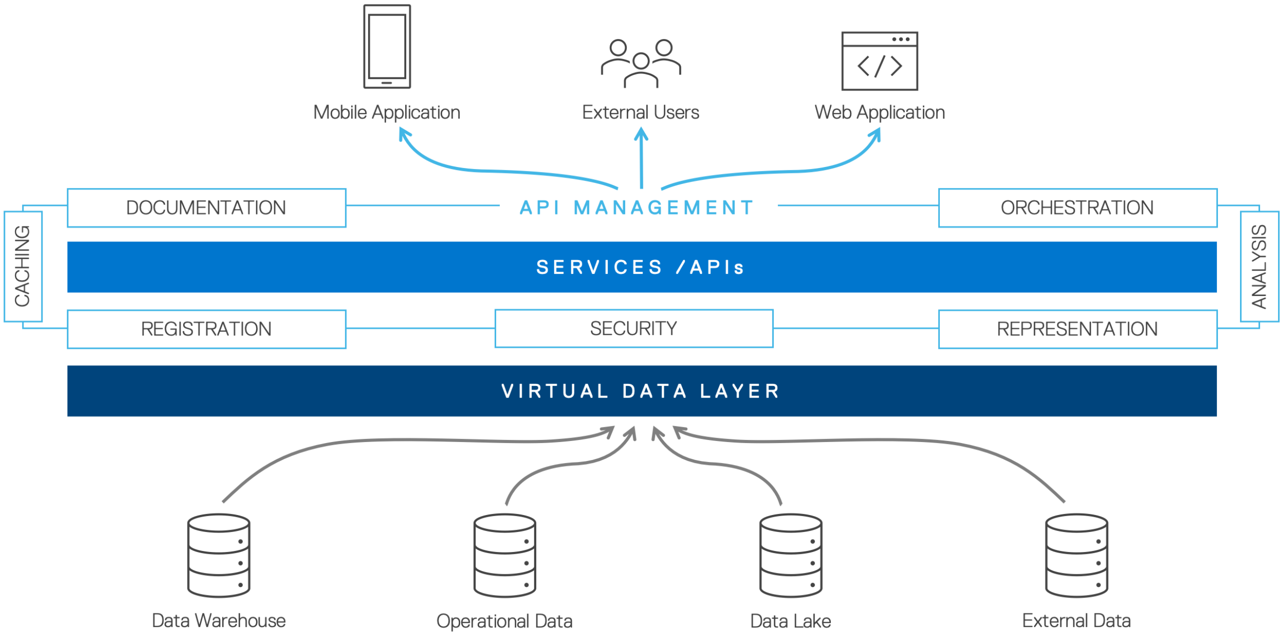


Figure 4. Data-as-a-Service

One sign of how critical data locality is to the data management enterprise are new Data-as-a-Service (DaaS) models that address some of the challenges arising from the automation and orchestration of data movement (**Figure 4**). DaaS promises on demand data to the user, regardless of the geographic or organizational separation between provider and consumer (Data as a Service, n.d.). Research information exchanges (Information Exchange, n.d.), data pipelines, and research-as-a-service models are all incorporating DaaS into their design.

DaaS models are leading to other disruptions to the conventional use of data in the form of low code and no code software tools. These tools provide avenues of application development to citizen developers, thereby enabling researchers to build applications and functions that leverage data despite having little-to-no software coding experience. Low code and no code tools are being applied to simple applications – to replace paper forms for simple data collection – and highly complex workflows – orchestrating entire business processes. DaaS models are also making serverless and Function-as-a-Service (FaaS) tools more common. In a FaaS tool, a function creates an output and that output can lead to data generation; the resultant data must live somewhere. Researchers have recognized the value in using serverless and FaaS to expedite research their workflows.

Regardless of where data might reside, its location and its manner of access are crucial pieces in the data management puzzle.

Metadata Management

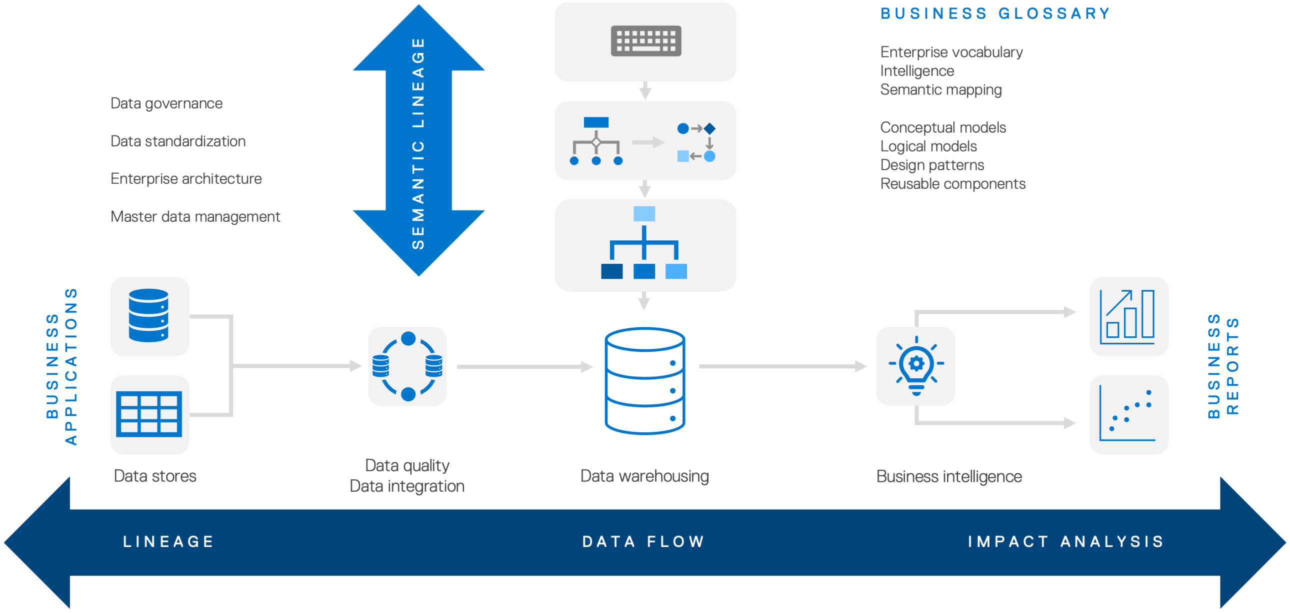


Figure 5. Metadata management flow

Metadata management – understanding data about data – provides researchers with information, knowledge, and intelligence (or wisdom) that extends the value of the underlying data (**Figure 5**). Managing metadata enables others to know where the data comes from, how it was produced, who or what produced it, whether it has been culled or pre-analyzed, whether or not it has been tampered with, and more. Ultimately, metadata enables researchers to better understand the context in which the data was generated, and when performing research, understanding context is crucial. Metadata has therefore rapidly enabled an extension of analytics that can be integrated into data pipelines and can be part of DevOps/MLOps automation.

> "By 2022, organizations utilizing active metadata to dynamically connect, optimize, and automate data integration processes will reduce time to data delivery by 30%" (Gartner, 2019)

Having the capability to manage metadata helps researchers better execute pipelines for shared information exchanges, and to build functions and applications that lead to insight.

Data Integration

Data is married to any research process, however, properly integrating that data is another crucial component of a data management strategy. Proper data integration allows for the timely access and availability of information when it is needed and is based on two characteristics: data mobility and data ingest.

Data Mobility

Data mobility refers to the immediate and self-service access to data. By providing data at the time of need, researchers can accelerate application development, test more quickly, get acceptance by decision-makers, develop more output, improve productivity, and decrease the time-to-impact of business intelligence (Tropeano, 2015). Without well-considered data mobility, the pace of a research enterprise will be throttled.

Data Ingest

To provide data at the time of need, data must be properly ingested into the systems, pipelines, and applications where it will subsequently be used. Methodologies of data ingest include ETL (extract, transform, load), ELT (loading data before transforming it), and other variations. Having the right mechanism in place to automate data ingest whenever possible makes overall data management easier and more efficient for researchers and organizations.

Search Capabilities

For researchers to take full advantage of data movement, ingestion, and integration, it may be necessary to have the capability to search across the necessary data and metadata. Finding the right information quickly enables an increased efficiency and improved cadence of research outcomes. At minimum, search functions such as content awareness, indexing, and query processing attributes should be included in a data management plan.

Data Catalog(s)

A **data catalog** can provide a repository of information about available data assets to researchers or organizations; in doing so, it enables the search capabilities mentioned above. The information in a data catalog should be collected automatically and may span multiple data sources. A data catalog should support self-service and collaboration, and it should be indexed and searchable.

Commonly, a data catalog will make use of data discovery behind the scenes. Often provided as functionality with a data profiling tool, data discovery helps identify the relationships existing between different data sets and within or across data sources (Howard, 2021). Data discovery commonly addresses regulatory compliance requirements (e.g. GDPR) and it plays an important role in (MDM) Master Data Management (Master Data Management, n.d.).

Data Pipelines

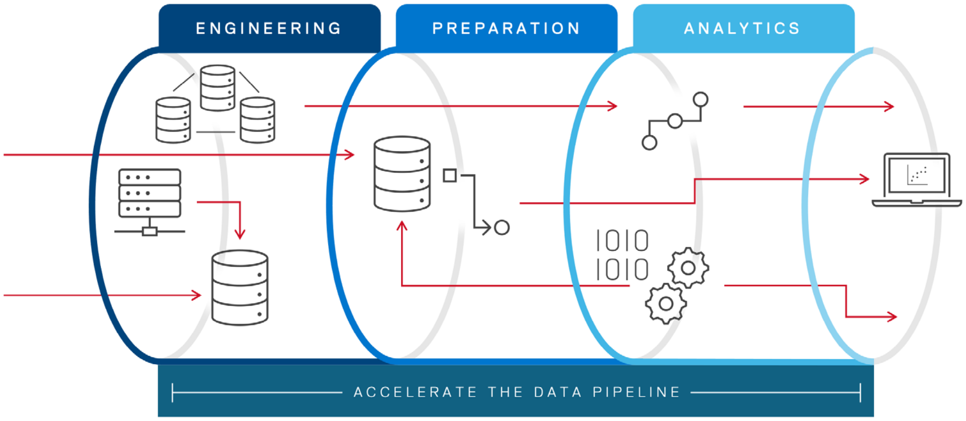


Figure 6. Data pipelines

Researchers and organizations can take full advantage of the concepts above by using data pipelines (**Figure 6**). Data pipelines are any set of processing elements that move data from one system to another, where the output of one is the input of the next, possibly transforming the data along the way (Raviv, 2017). Data pipelines provide an organized and often-efficient construct for delivery of information from data source to destination. Pipelines should be automated whenever possible and can leverage machine learning and artificial intelligence to aid in sourcing as well as ingest. To make the best use of data pipelines, researchers should be able to clearly articulate where, when, and how data is collected. Furthermore, multiple data pipelines are likely to be employed by researchers and organizations who have a mature data management strategy.

Policy and Governance

Although it may not be the most dynamic part of a data management strategy, **policy and governance** has become increasingly critical because of the legal responsibilities to which data stewards are being held by international governments. Researchers and research organizations are just as culpable as any other individual or entity when abiding by the policy and governance rules surrounding any piece of data. Existing regulations such as HIPAA (Health Insurance Portability & Accountability Act), FERPA (Family Educational Rights and Privacy Act), PCI DSS for payments, and others combined with newer or emerging regulations like GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act) impact how and what data can be stored and the ways it can be used.

> "By 2020, over 50% of multinationals worldwide will have **automated their cloud, IT infrastructure, and data governance setups** to comply with regional data regulations such as GDPR and PSD2." (IDC, 2019)

These regulations go beyond encouraging best practices; for some, best practices are mandated and carry real consequences if they are not adhered to: large fines and legal fees have been doled out for violations. Whereas it was once enough to simply state a policy, now strict governance mechanisms must be in place to ensure compliance with regulations. Common practices like de-identifying human data for patients or students have been codified, but so too have requirements to notify individuals when data is being collected and indicating what information is being captured. Whereas users once were required to opt out of data collection, new regulations require users to opt in. These changes have also led to the emergence of new revenue models whereby a person can monetize their own private data.

Policy and governance have also led to the expectation that researchers must have a plan for data management. The National Science Foundation and the National Institutes of Health, along with other Federal agencies in the United States, mandate the inclusion of a data management plan as part of grant applications. Universities and colleges thereby assume the responsibility for the proper stewardship of the data that is generated by the research enterprise. The burden on institutions continues to grow as the amount of research data for which they are responsible exponentiates.

Intrinsic Security and Trust

The requirement for data governance is enabled by the final component of a data management strategy: intrinsic security and trust. For all verticals, the demand for higher levels of trust across systems, services, and data is a result of both regulations and user expectations. And although compliance and data privacy laws continue to add to the technical complexity and operational cost of data management, cohesive and consistent security using existing solutions has still proven to be ineffective in highly (geo)distributed systems.

> "By 2020, 50% of servers will encrypt data at rest and in-motion, and over 50% of security alerts will be handled by AI-powered automation." (IDC, 2019)

The trust gaps associated with current solutions present an opportunity for new and emerging technologies to fill. The Internet of Things is being secured through a mix of edge and telemetry data collection and processing. Data provenance solutions are ensuring the accuracy and legitimacy of data, even for physical items procured through complex supply chains. Data security across hybrid cloud models is protecting data in transit. And even SecDevOps[2] – the process of integrating Security, development, and IT Operations into a contiguous and cohesive lifecycle management architecture – is a sign of the attention and importance afforded to the need for trust within data management.

[2] "By 2024, DevSecOps will be automated to the extent that 60% of new applications will have comprehensive security and compliance assessment included in the continuous delivery pipeline." (IDC, 2019)

> "By 2020, we expect that companies that are digitally trustworthy will generate 20% more online profit than those that aren't." (Gartner, 2019)

In summary, researchers and research organizations must carefully consider their data management strategy to enable the research enterprise effectively, to generate efficiencies, and to protect all data as valuable assets. By deconstructing the components of a data management strategy, researchers can ensure that they are both responsible stewards of the data and that they are employing best-in-class emerging technologies. Although the responsibility does not wholly fall on researchers – it must be shared by research administrators, students, and others – it is only through the collaborative cooperation of researchers, organizations, and IT operations that the optimal implementation of a data management strategy can be achieved for research.

References

Data Management. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Data_management

Data as a Service. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Data_as_a_service

Howard, P. (2021, March 3rd). ***Data Discovery and Catalogues***. Retrieved from Bloor Research: https://www.bloorresearch.com/technology/data-discovery-and-catalogues/#whatisit

Information Exchange. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Information_exchange

Master Data Management. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Master_data_management

Raviv, O. (2017, October 29th). ***What is the difference between a data pipeline and an ETL pipeline?*** Retrieved from Alooma: https://www.alooma.com/answers/what-is-the-difference-between-a-data-pipeline-and-an-etl-pipeline

Tropeano, G. (2015, May 4th). ***Essential Data Mobility – people, devices, data, business***. Retrieved from actifio.com: https://www.actifio.com/company/blog/post/data-mobility-people-devices-data-business/