Data Lakes vs. Data Mesh:

Navigating the Future of Organizational Data Strategies

AWS CLOUD ENTERPRISE STRATEGY BLOG

Data Lakes vs. Data Mesh: Navigating the Future of Organizational Data Strategies

For more than a decade, organizations have embraced data lakes to overcome the technical limitations of data warehouses and evolve into more data-centric entities. While many organizations have used data lakes to explore new data use cases and improve their data-driven approaches, others have found the promised benefits hard to achieve. As a result, the effectiveness and ROI of many data lake initiatives are now under scrutiny.

Navigating the Pitfalls: When Data Lakes Turn into Data Swamps

The tech community's view of data lakes has evolved as some organizations face challenges around managing vast data stores and avoiding "data swamps," where data is stored but not used. These data swamps are massive repositories where data is dumped indiscriminately, leading to problems with discoverability and usability. Centralization can create bottlenecks that slow access and analysis, and without rigorous governance, data quality can quickly deteriorate. In addition, the one-size-fits-all approach of data lakes fails to address the specific needs of different business domains. The potential of data lakes often remains untapped because users struggle to extract value due to a lack of appropriate tools or the complexity of the data itself.

Well-Implemented Data Lakes	Poorly Implemented Data Lakes
Single Source of Truth	Data silos that are difficult to access
Cost-effective storage	Lots of unnecessary data stored ("Data is the new oil")
Data democratization	Specialized skills necessary to access data
Flexibility in data formats	Poor data quality and inconsistency
Advanced analytics and machine learning	Difficulties in deriving value from vast amounts of unstructured data
Faster insights	Lengthy communication and a lack of tools, interfaces, and skills
Streamlined data management	Management becoming cumbersome as the lake grows

AWS CLOUD ENTERPRISE STRATEGY BLOG

Data Lakes vs. Data Mesh: Navigating the Future of Organizational Data Strategies

For more than a decade, organizations have embraced data lakes to overcome the technical limitations of data warehouses and evolve into more data-centric entities. While many organizations have used data lakes to explore new data use cases and improve their data-driven approaches, others have found the promised benefits hard to achieve. As a result, the effectiveness and ROI of many data lake initiatives are now under scrutiny.

Navigating the Pitfalls: When Data Lakes Turn into Data Swamps

The tech community’s view of data lakes has evolved as some organizations face challenges around managing vast data stores and avoiding “data swamps,” where data is stored but not used. These data swamps are massive repositories where data is dumped indiscriminately, leading to problems with discoverability and usability. Centralization can create bottlenecks that slow access and analysis, and without rigorous governance, data quality can quickly deteriorate. In addition, the one-size-fits-all approach of data lakes fails to address the specific needs of different business domains. The potential of data lakes often remains untapped because users struggle to extract value due to a lack of appropriate tools or the complexity of the data itself.

Well-Implemented Data Lakes	Poorly Implemented Data Lakes
Single Source of Truth	Data silos that are difficult to access
Cost-effective storage	Lots of unnecessary data stored (”Data is the new oil”)
Data democratization	Specialized skills necessary to access data
Flexibility in data formats	Poor data quality and inconsistency
Advanced analytics and machine learning	Difficulties in deriving value from vast amounts of unstructured data
Faster insights	Lengthy communication and a lack of tools, interfaces, and skills
Streamlined data management	Management becoming cumbersome as the lake grows

Data Producers vs. Consumers: The Organizational Chasm

The root cause of these issues is the organizational interaction between data producers and consumers on the one side and the central data lake team on the other. Data producers are often more motivated to develop new application features than to make data available for analytical use cases. Their focus on transactional rather than analytical workloads means that their shared data can be of poor quality. They also lack a connection to the consumers of their data, leading to a mismatch between what is produced and what is needed.

Data lake teams have their own problems: They are overwhelmed with data sources, have to perform complex technical maintenance, and constantly juggle changing priorities. Exacerbated by their limited analytical capabilities and disconnect from data producers, consumers are frustrated by the lack of transparency, unclear value, and low prioritization of their needs. Data consumers and producers do not usually interact directly; this communication is blocked by the data lake's proxy organization, where all data is stored centrally.

Empowering Teams with Data Mesh: The Path to a Distributed Data Ecosystem

The discussion in the tech community has shifted to a more nuanced and adaptable data strategy called data mesh. It aims to overcome some limitations of centralized data lakes by promoting a more distributed, human-centric, and context-specific approach to data management.

Data mesh is an alternative approach to centralization issues. It assigns responsibility for analytical data to the domain-specific teams that build and run applications and produce transactional data, such as e-commerce teams, and those that consume data and use it to gain insights. For example, the team that owns the checkout page in a web shop and stores the sales data in a transactional database is also responsible for making that sales data available for analytics, reporting, and AI/ML use cases, such as marketing or finance. Data mesh makes it easier and simpler for consumers to use this analytical data.

It is not just another interface implemented; the data is made available as an independent data product that provides a specific benefit to an actual consumer. This data product consists of the specific data and its metadata, the source code required to prepare and deliver the data, the necessary test and production infrastructure (as IaC), and its configuration.

Fostering Data Literacy: Introducing New Roles in Data Mesh Teams

Teams that create and use data, such as the e-commerce checkout team and marketing department in my example, often lack the expertise to develop and manage data for analytics. However, their deep knowledge of the business context of their data is invaluable. In a data mesh framework, it's essential to capitalize on this knowledge by upskilling these teams to implement analytical use cases. This includes providing extensive training to existing members and creating additional specialized roles. Two key roles are critical: a data product owner to guide the strategic direction of the data and a data engineer to handle the technical aspects of building and managing these data products.

A data product owner is a business-oriented data person who knows the business domain very well from a transactional and analytical perspective. They communicate directly with the data product's consumers and define the product, its strategy, and its roadmap.

A data engineer is a broad generalist in data engineering and data science with deeper expertise in a data-related area needed by the business. This person builds the actual data products and is the point of contact for technical questions from other teams.

Creating a Foundation for Success: The Data Mesh Platform

To realize the full potential of data mesh, I recommend embedding both roles directly in the producing and consuming teams. A valid but suboptimal variant, because it reintroduces a proxy team, is to establish a separate data mesh team for each business domain (e.g., e-commerce). A data mesh platform supports producers and consumers, making their work easier and more efficient. The data mesh platform teams do not create data products or store or process data.

The data mesh platform has three roles: (1) to provide tools and infrastructure such as a data catalog, access control, CI/CD pipeline, monitoring, and preparatory development and test environments; (2) to train and advise producers and consumers and, if necessary, support them with additional development capacity; and (3) to moderate common standards and procedures in a federated approach that must be adhered to throughout the organization. The mission of the data mesh platform is to make life simple, efficient, and stress-free for producers and consumers.

Unfortunately, running a platform successfully and sustainably is not as easy as some in the tech community suggest. I have summarized my experiences in my blog post on how to set up a platform that effectively supports your development teams.

When done correctly, the data mesh model promotes a proactive approach to maintaining data quality, relevance, and accessibility, as well as tailoring data products to meet the unique needs of different business units. By closely aligning analytical data with its operational context, a data mesh facilitates more effective use and sharing of data across the organization. It leverages modern distributed architecture principles, such as those derived from microservices architectures, to not only store data more efficiently but also make it readily available for consumption, driving actionable insights closely aligned with business objectives.

For a good example of a data mesh showcase, check out GoDaddy's Chief Data and Analytics Officer Travis Muhlestein's presentation on building data mesh architectures on AWS from re:Invent 2022.
—Matthias

Thank you for downloading this AWS and Vendor Resource! Carahsoft is the distributor for AWS public sector solutions available via GSA, NASPO, The Quilt and other contract vehicles.

To learn how to take the next step toward acquiring AWS's solutions, please check out the following resources and information:

For additional resources:
carah.io/AWS-Resources

For additional AWS solutions:
carah.io/AWS-Solutions

To set up a meeting:
AWS@carahsoft.com
888-662-2724

For upcoming events:
carah.io/AWS-Events

For additional public sector solutions:
carah.io/AWS.Solutions

To purchase, check out the contract vehicles available for procurement:
carah.io/AWS-Contracts