KPMG

Data mesh

A modern approach to building your data ecosystem

kpmg.com

Introduction

Technology leaders understand that data has become the lifeblood of their company and that they have the primary responsibility to provide the business with easy, secure, and reliable access to that data.

The challenge is that the data is typically stored in dozens or even hundreds of siloes scattered across the enterprise—in legacy on-premise solutions, in the cloud, across departments, business units, lines of business, and so on. By some estimates, data scientists spend as much as 80 percent of their time looking for and obtaining access to the data they need, instead of spending that valuable time building models and delivering insights. IT, too, spends much of its time in one-off projects providing that ad-hoc access.

In recent years, many companies have implemented data lake technology to help solve the problem. The concept is simple and compelling: break down those siloes by pulling all that data, both structured and unstructured into a single, centralized platform.

A data lake also provides what many see as the holy grail of data: a single source of truth. Each division in a large enterprise, for example, might have a different definition of revenue or a customer's worth to the business. Centralizing data eliminates such inconsistencies.

Data scientists spend as much as

80%

of their time looking for and obtaining access to data

Source: martinfowler.com, Zhamak Dehghani (December 3, 2020)

Problem solved?

While it may sound like an ideal solution, that isn't always the case. The siloes may be gone and the data now centralized, but more often than not, data consumers still require IT's help to access it. As with anything that attempts to centralize large and complex things, scalability and maintainability are significant challenges.

More importantly, data quality can suffer. The engineers responsible for it will be intimately familiar with the lake's plumbing—the technologies and services used to centralize the data—but they have little or no experience with or functional understanding of the data itself. When data enters the lake, ownership of or responsibility for it is severed. It loses domain-specific knowledge about its origin—what it means or why it's needed. Is the data the most up-to-date version? What does a certain field mean? Data provenance becomes more challenging.

Further, data scientists have come to realize that creating a "single source of truth" can often do more harm than good. What had been seen as undesirable inconsistencies across different data sources may actually contain valuable insights—there's a reason different divisions define a customer's worth differently, for example. Eliminating those differences can eliminate the insights.

As many have discovered, the result can be more of a messy "data swamp" than a lake, into which they pour increasing amounts of money without realizing any of the promised returns.

A new way of thinking

The disappointment with traditional data lake architecture has spurred technology leaders to seek alternate architectures to build their data lakes. The most promising of these comes from a lesson learned in our journey to the cloud. When the cloud first appeared, most companies treated it as a replacement for their expensive and difficult-to-maintain data centers. The "lift and shift" of on-premise applications to the cloud was the only goal.

Then technology leaders had an epiphany. They realized that the cloud held far more value. Unlocking that value would require a new way of thinking about software development and deployment. It would require a new operating model and new technologies designed to support it, such as Agile development, containerization, and continuous integration/continuous deployment.

Instead of building big, complex, monolithic applications, we now break down those monoliths into much smaller, discreet microservices, with each providing a narrowly focused set of capabilities. These individual, stand-alone microservices can communicate with each other via standardized application programming interfaces (APIs), and so can be dynamically combined in multiple ways to form complete, robust solutions.

While those big monolithic applications required large teams of software developers to build, microservices use small, independent teams comprising both technology and business professionals with domain-specific expertise related to each microservice's function. This enables each team to treat its microservice as their product, and so team members collaborate to continually improve it. Because each team can deploy new features and enhancements independent of all other teams, they're able to innovate more rapidly with greater agility.

While this radical new approach has firmly taken hold in software development, not much had changed with how we approach data—that is, until 2019, when Zhamak Dehghani published her now-famous white paper arguing that the same approach should be applied to data—a concept she called data mesh.

In 2019, Zhamak Dehghani published her now-famous white paper arguing that a microservices approach should be applied to data— a concept she called data mesh

Enter the data mesh

The concept of data mesh is a modern way to build a data lake. There are four hallmarks:

1 Domain owner control

With a data mesh, all of the independent data creators own and maintain their own data. They can define it as they like and as they find most useful. Instead of creating a silo, however, data mesh enables each data owner to easily share or "publish" their data via standardized APIs to others who want to consume it on a read-only basis. As a result, domain-specific knowledge is preserved, and it can be applied to improving data usefulness, or for identifying data errors or anomalies that others less familiar with the data might miss.

Along with the data itself, each data owner must maintain and publish a catalog of metadata, which details what each field means, what it's used for, where is comes from (its lineage), how often it's updated, etc.—all things that are best maintained by those most familiar with the data.

2 Self-service consumption

Anyone who wants access to the data can request it directly from the data owner and use it as they see fit. The metadata catalog provides these data consumers with all of the information they need to understand it. Because the data is always accessible, there's no need to replicate it—thus avoiding the problems that often arise when there are multiple copies determining if a version is still accurate or up to date.

3 Federated security and access

Data owners are responsible for the security of their data. They're responsible for handling access requests from data consumers and for provisioning that access—another key element of the self-service model. IT is no longer burdened with such requests.

4 Federated governance

To pull this off, data mesh requires strong, federated governance. Each data product must be fully interoperable with the others, and so a governing body—usually a group of internal stakeholders—defines which tools or technologies can be used. It defines the standard for the metadata catalog. It provides guiding principles for the handshake mechanisms and RESTful APIs used to connect one system to another, and for the identity and access management systems, protocols, and procedures used to provide secure access to them. Service level agreements are used to help ensure each data owner maintains a high level of quality and availability.

Solving the challenge of data access

Instead of creating a big, complex, monolithic data lake to provide universal access to data, organizations can build a modern data lake by leveraging the concepts of data mesh architecture. The data mesh concept borrows a page from the microservices playbook. Just like microservices, the individual data sources can be dynamically combined to create incredibly rich data resources specific to each data consumer's needs — all on a self-service basis without IT intervention.

As with cloud-based software development, a significant change in thinking is required to truly unleash the value of all this data. Data owners must view their data as a product. They are responsible for maintaining quality and incented to continually enhance it by making it more useful and easier to access and consume.

New tools and operating models are required to enable this change in thinking. Some can be borrowed directly from software development—"data as code," for example—while others are specific to data. Thankfully, we now have a host of such solutions and operating models that have been tested and honed by real-world implementations.

The data mesh distributed model appears to solve the challenge of data access while avoiding all the issues that come from dumping data into a single, centralized lake. If you're struggling to provide data access or dealing with the pain of a data lake, it's definitely worth a look. If you'd like to hear more, shoot me a message. I'll be happy to connect.•

As with cloud-based software development, a significant change in thinking is required to truly unleash the value of all this data. Data owners must view their data as a product.

How KPMG can help

KPMG is here to help. At KPMG, we have the digital transformation talent, technology experience and advanced tools to help you execute on your technical debt management and modernization initiatives. Our business process and operating model acumen can help you map the right path to take your enterprise into the future. Count on us.

To learn more about how KPMG could help your business, please contact:

Kevin Martelli
Principal, Advisory
kevinmartelli@kpmg.com

Saravanan Subbarayan
Advisory Managing Director
ssubbarayan@kpmg.com

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

kpmg.com/socialmedia