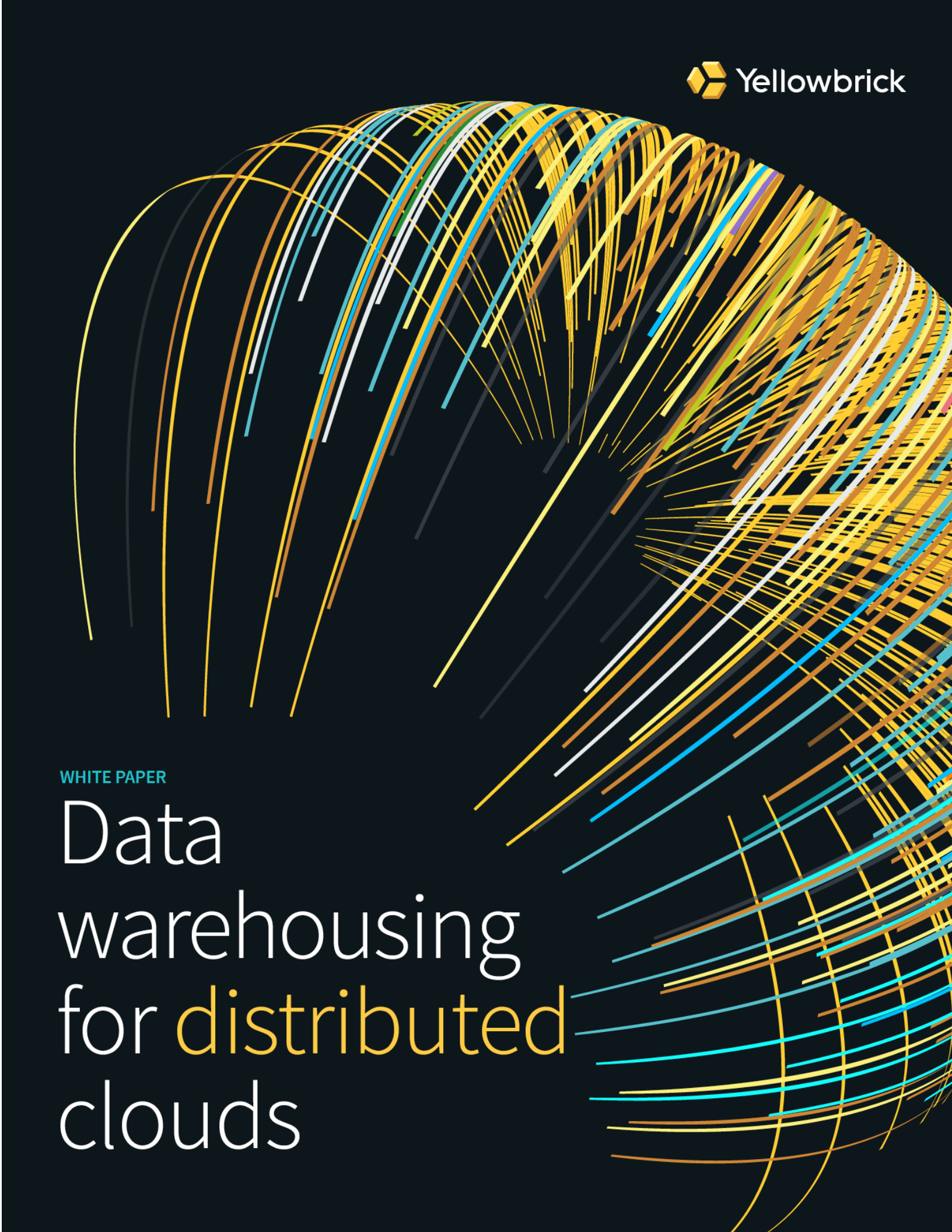
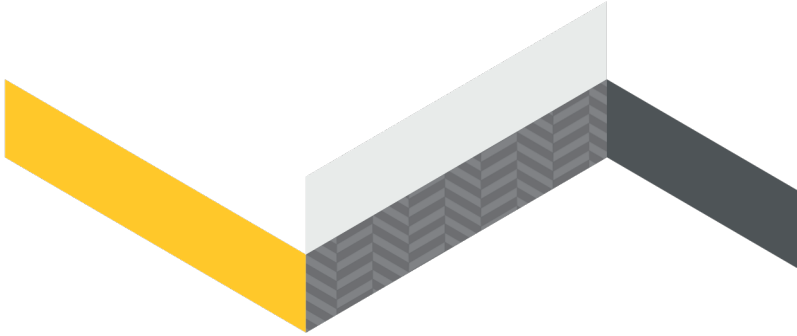
Yellowbrick



WHITE PAPER

Data warehousing for distributed clouds



The *Gartner Top Strategic Technical Trends for 2021* report suggests that the distributed cloud model will emerge to address the explosion of data growth, particularly at the network edge. Distributed clouds are characterized by the deployment of cloud software and hardware stacks outside of the public cloud provider's data center to provide a mesh of interconnected cloud resources to form a best-of-breed logical cloud. These stacks enable the ability to run applications developed for the public cloud in a company's own data center and in other locations, such as in multi-access edge computing centers connected to 5G cell tower groups, or on the factory floor in support of IoT applications in manufacturing.

The distributed cloud model offers reduced latency, increased data sovereignty, higher security, addresses data gravity requirements, and provides uniformity in terms of infrastructure and services. Distributed clouds also provide support for cloud applications in remote locations, which may be only intermittently connected to the internet, if at all. Adopters of distributed clouds include:

- SaaS companies that combine private and public infrastructure to ensure always-on access to their services
- Companies in regulated industries like banking, insurance, and healthcare (as well as public sector agencies) with use cases that require a hybrid approach for satisfying regulatory and data sovereignty requirements
- Telecom, logistics, manufacturing, and retail companies with emerging use cases for IoT analytics that are challenged by data gravity

Distributed clouds can be seen as an inevitable consequence of the hybrid cloud deployment model that incorporates public cloud, private cloud, and cloud stacks at the network edge, and provides seamless integration of resources, data, and analytics across multi-cloud boundaries, orchestrated from a single control plane.

Distributed clouds can be seen as an inevitable consequence of the hybrid cloud deployment model

Defining the distributed cloud

To take full advantage of the benefits that distributed clouds offer, we must rethink our approach to how data is managed in such a homogeneous, geographically separated and logically interconnected environment. Distributed clouds will provide the common foundational hardware and software infrastructure on which new applications that are federated across clouds will be deployed. In order to support these applications, the underlying data management services higher up in the stack must also be federated.

Consider the various components of a modern enterprise data architecture today. We can decompose an analytical ecosystem into a set of vertical services that span from the producers of data through to the consumers of data and analytics, and the common horizontal services needed to support each vertical service. (See Figure 1.)

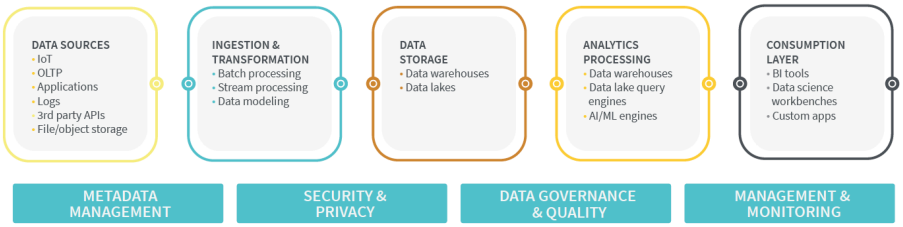


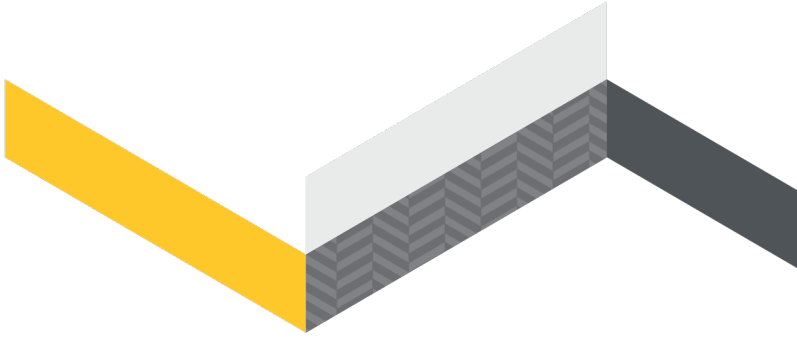
Figure 1. Modern enterprise data analytics architecture

For many enterprises today, these services are largely localized to work within the same data center or public cloud. With the rise of distributed cloud comes the need to consider these components in the context of geographically separated, but integrated, resources.

We believe that the localized analytical ecosystems in many enterprises today will give way to delocalized ecosystems that deploy services based on data gravity, or because of latency or governance needs. The set of federated data management services running on distributed cloud infrastructure will form a distributed cloud.

Data warehouse platforms will have to evolve to meet those requirements. Managed from a single control plane, they will enable analytic applications to be provisioned at the point of need on a right-sized blend of physical and virtualized infrastructure, based on data gravity, data sovereignty, data governance, and latency requirements.

Localized analytical ecosystems will give way to delocalized ones that deploy services based on data gravity, latency, or governance



Such systems will also require:

- Components with the ability to run anywhere with consistently good price/performance: in public, private, and hybrid clouds, and at the network edge
- A unified control plane for provisioning and managing the lifecycle of services
- Federated data management and analytics services that seamlessly integrate across cloud boundaries utilizing common API standards and providing self-service capabilities
- Code and workloads should be portable across deployments, along with the ability to govern the movement of data across clouds and locations in line with regulatory and data sovereignty requirements
- Comprehensive security and privacy capabilities to protect data wherever it resides in a distributed cloud

The distributed cloud model seeks to reverse this centralized management of data and leave the data where it belongs: with the producers

Data as a first-class citizen

Distributed clouds should provide seamless access to the right data, at the right place and at the right time, with guarantees around data quality and provenance. In current ecosystem architectures, the focus is too often on the infrastructure components, such as data lakes and ETL pipelines, and not on the business problems and data that the ecosystem is there to support. Data flows from around the business into centralized data lakes and warehouses and away from the data producers and consumers in the lines of business. When detached from the lines of business owners and managed by a centralized IT organization, this data can start to lose meaning and value.

The distributed cloud model seeks to reverse this centralized management of data and leave the data where it belongs: with the producers. The producers make data available to consumers elsewhere in the business directly, through APIs and self-service means. The data becomes a first-class citizen in the distributed cloud, served up through a common infrastructure available to every business domain, and is discoverable through-out the enterprise. Automated data movement, federation, cataloging, and access policies ensure that the right data is made available to consumers in the most optimal way. Within this model, the infrastructure itself is centrally managed and secured, but each business domain manages and secures access to its own data products.

The central benefit of such an approach is that the data remains with the subject matter experts in a business domain who can utilize common data management and analytics services to ingest, transform, analyze, describe, and serve up data to other business users. Infrastructure components such as data lakes and data warehouses simply be-come services within the distributed cloud, and the data itself becomes a first-class citizen.

Kubernetes is the key that unlocks distributed clouds

The vendors that will be most successful in supporting the distributed cloud will be those who have adapted their data management and analytic applications to be cloud-native, such that they integrate with the core public cloud services that support provisioning, scaling, security, storage, and networking. Furthermore, those vendors that adapt their applications to take advantage of micro-services architectures and containerized deployment will have an additional advantage because of the benefits these approaches provide in terms of faster deployment of new features, resilience, and scalability.

Kubernetes has become the de facto standard choice for orchestrating the lifecycle of cloud-native applications. It provides a uniform environment for deploying and managing those applications in private clouds and public clouds, and at the network edge. The most successful data management and analytic applications for distributed clouds will be Kubernetes-orchestrated to take advantage of the “run anywhere” capabilities that Kubernetes provides.

In the context of data warehousing, building a best-of-breed distributed cloud implies the need for a rich ecosystem that preserves customer choice. The distributed cloud model will feature a wide range of federated services and applications to support data cataloging, governance, lineage, security, data virtualization, data ingest, data warehousing, data lakes, streaming analytics, workflow management, AI/ML engines, and many other components. The implementation of a distributed cloud will require a ecosystem effort, where distributed cloud-ready ISVs work together with SIs to put in place the technology, people, and processes needed to ensure success. (See Figure 2.)

Kubernetes has become the de facto standard choice for orchestrating the lifecycle of cloud-native applications

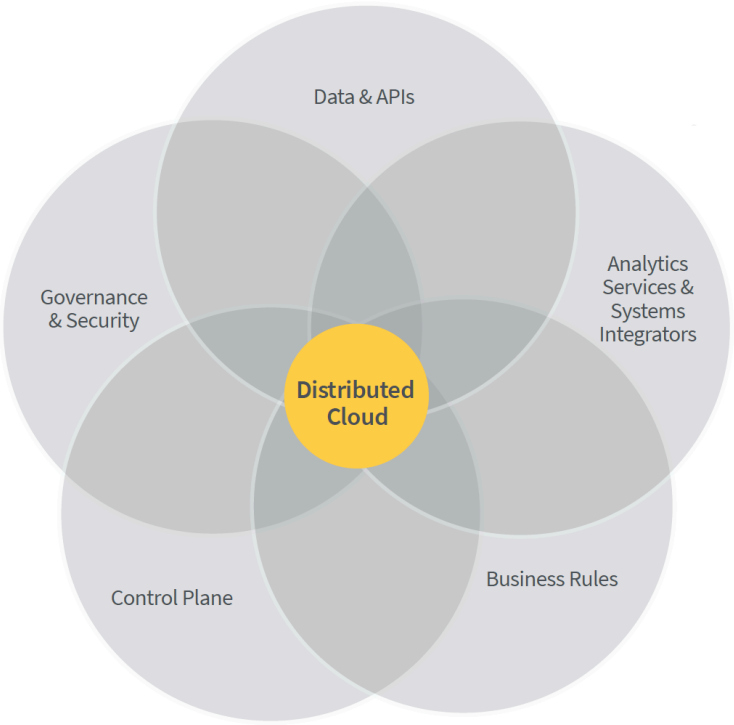


Figure 2. Distributed clouds require an ecosystem for success.

Yellowbrick is the first data warehouse capable of integrating into the new distributed cloud environment

Yellowbrick: The data warehouse for distributed clouds

Yellowbrick has been preparing for the arrival of the distributed cloud, and Yellowbrick is the first data warehouse capable of integrating into this new environment. Our approach may help to illustrate the steps that technology vendors will need to pursue to be ready to build distributed clouds for customers.

Yellowbrick has embraced Kubernetes as core, cloud-native architecture to help customers deploy, manage, and orchestrate data warehouse workloads across private cloud and public cloud environments, as well as at the network edge for future use cases like IoT analytics. And Yellowbrick's unique adaptive "cut-through" architecture ensures excellent price/performance in any environment, whether on bare metal (e.g. Andromeda optimized instances) or virtualized infrastructure (VMs or Kubernetes stacks).

Yellowbrick will have applications at the IoT edge, as well as in the public cloud data center. Services such as AWS Outpost and Wavelength support

EKS (Azure and Google Cloud have their own equivalents) and are equipped to run the Yellowbrick database outside of the cloud data center. (See Figure 3.)

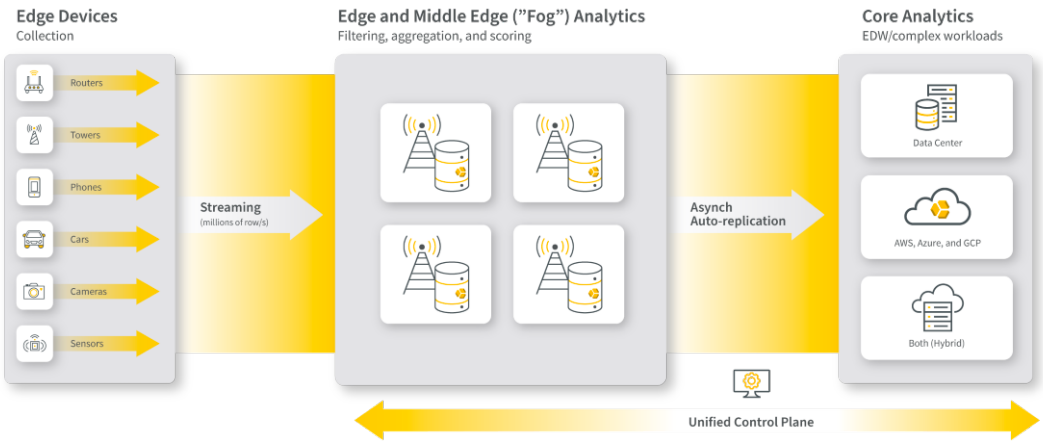


Figure 3. Yellowbrick in distributed clouds

Furthermore, our existing replication and near-real time streaming ingest capability opens new horizontal technical use cases where data could be collected, filtered, and aggregated by Yellowbrick instances at the edge or middle edge, and sent to one or more Yellowbrick instances or multiple at the cloud center.

We have developed a control plane for controlling the provisioning and management of resources across different clouds, supporting the true instantiation of a distributed cloud for data warehousing. It provides a single point of management for Yellowbrick across distributed clouds, and allows instances to be created, configured, integrated, monitored, and decommissioned from a single web console.

Conclusion

At Yellowbrick, we believe that our strategy to invest in the distributed cloud space means we are well placed to satisfy not only the hybrid cloud deployments of today, but also the data warehousing needs of emerging distributed clouds.

Try our free 7-day Test Drive: yellowbrick.com/test-drive

250 Cambridge Avenue, Suite 300, Palo Alto, California 94131 I USA I 1.650.687.0896