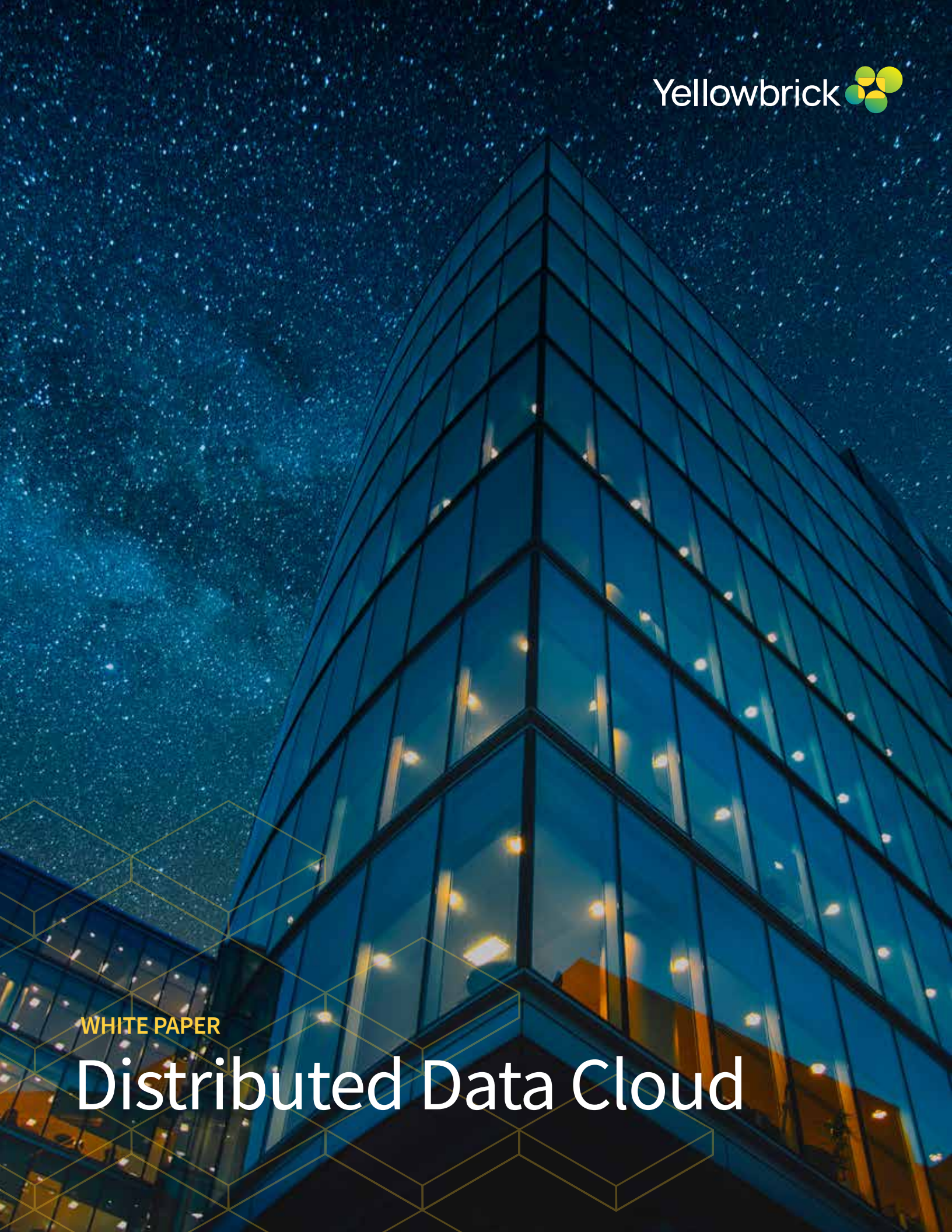
Yellowbrick
WHITE PAPER
Distributed Data Cloud

Distributed Data Cloud

The range of options available to purchasers of data management and analytics solutions is immense. Whether it's business intelligence applications, data warehouses, data integration products, or tools to support governance, security, management, monitoring, machine learning, etc., the number of competing products in the market can often overwhelm the buyer. The degree of choice is multiplied when you consider the native services in these areas offered by the major public cloud providers, in addition to the ones offered by ISVs.

However, are the choices real or an illusion? Selecting a particular cloud provider or a particular data warehouse vendor, for example, can in fact limit a buyer's choices moving forward and increase business risk. The agony of choice can simply become agony, ultimately constraining data-driven procurement decisions. In this white paper, we examine the business challenges caused by the current trend of data management and analytics buying decisions – particularly in the context of enterprise data warehousing – and discuss an emerging approach to reducing risk, avoiding vendor lock-in, eliminating integration overhead, and controlling spend, known as Distributed Data Cloud.

The current data and analytics market limits choice

Within the largest enterprises we see three recurring areas of focus within digital transformation initiatives, aimed at:

1. **Eliminating cloud concentration risk**
2. **Increasing efficiency**
3. **Modernizing the analytic ecosystem**

Without dedicated investments in these three areas, business risks are introduced, budgetary planning is difficult, and the full potential of the company's data remains untapped. We discuss these issues in more detail below, and some of the consequences that only become apparent over longer periods of time in production and at larger scale and utilization levels:

1. Cloud concentration risk

Concentration risk refers to an overreliance on technology from a single cloud provider or SaaS vendor. Deploying a solution in one cloud can lead to single points of failure and reduced system availability. All three major CSPs, as well the data warehouse provider Snowflake, experienced serious and widespread outages in 2021, which had a significant impact on their customers' ability to do business. In the case of Snowflake, it was the single point failure of its authentication mechanism that caused a global outage.

Cloud concentration risk is a real concern in heavily regulated industries. Regulators in the financial services and insurance industries have highlighted the operational resilience and systemic risks that arise when banks place their

critical IT infrastructure in a single cloud. A bank that places all its IT services in one cloud now depends on a 3rd party infrastructure provider, which can lead to operational resiliency issues due to the shared responsibility for services. The CSP may not be willing to open its technology implementation, controls or processes to be scrutinized by regulators in the same way they can be if the financial institution owns the entire stack in their own data center.

Cloud concentration risk becomes even more concerning if multiple institutions and functions use the same CSP. In this situation, we face systemic risks to the broader financial system if critical clearing, settlement, and payment systems are deployed on the same public cloud infrastructure and are subject to a CSP-wide outage at the same time.

The notion of data ownership is called into question when companies use a SaaS provider running in a public cloud to manage their data. In this case, ownership of the data, as well as the infrastructure, is delegated to a 3rd party, and the business does not retain the ability to secure and govern their data in the same way they do if the data resides within their own CSP account or on-premises data center.

The buildup of data over time on a single cloud platform also presents a challenge if the decision came to move to a different provider or to repatriate workloads back to an on-premises data center. Data gravity and native cloud service integrations create a vendor lock-in, which is compounded by the data egress fees levied by the major three CSPs. As more data is generated and more use cases deployed, we enter a vicious cycle where the ecosystem gets increasingly entrenched within a particular cloud.

2. Efficiency

Why do enterprises move to the cloud in the first place? It is usually to chase efficiencies, economies of scale and cost reduction versus their own data center operations. However, moving data and running analytics in the cloud can be far from efficient, and companies are often disappointed when they don't realize the cost savings anticipated when they get there.

With the introduction of cloud data warehousing came the widespread adoption of on-demand billing. The ability to spin up data warehouses on the fly and only pay for what you use is compelling. It's led to the democratization of data warehousing and self-service access to data and analytics to anyone in the business. This agility speeds up the development and operationalization of new analytic applications and use cases and is perhaps the most important benefit delivered by modern cloud data warehouses. However, the ability to pull out a credit card and pay-as-you-go leads to budgetary planning challenges. One senior data and analytics executive explained to us that one of their monthly bills for cloud data warehousing was 7X larger than expected, and he was struggling to generate a credible cloud spending plan that he could put in front of his CFO.

The high and unpredictable spend with cloud data warehouses today stems from the inherent inefficiency of the solutions on the market. The answer to scale and concurrency challenges from vendors such as Snowflake, Google and AWS is to throw more credits, slots or virtual machines at the problem, leading to a higher overall cost per query. Traditional on-premises systems had a different design goal compared to most of today's cloud data warehouses. The legacy enterprise

data warehouses focused on efficiency, high concurrency and high availability. Due to limited data center space, data warehouse vendors in the past concentrated on squeezing resilience and performance out of the hardware at the expense of elasticity and agility. When companies choose to migrate their existing data warehouse to a new cloud platform, they often find what they gained in elasticity, they lost in efficiency, cost control and performance at scale.

3. Modernization

Due to the complexity and scale of their analytic ecosystems or the type of industry they serve, many larger enterprises have been slower to modernize and move their analytics to the cloud. One of the largest US financial institutions has over 15,000 siloed data systems, and so planning and realizing a wholesale move to the cloud in this instance will take them years. In some heavily regulated industries, it may not be desirable to move everything to the cloud either. For regulatory or security reasons, a company may choose to keep some of its business-critical applications on-premises in their own data centers.

Larger companies that are early on in their cloud adoption cycle typically choose to migrate certain applications to the cloud, or mandate that new applications should be born in the cloud. They are forced to adopt an inefficient hybrid cloud stance, where they must maintain the technologies and skillsets related to their legacy ecosystem, while at the same time recruiting cloud skills to support new applications, incurring additional cost. In addition to the different technology stacks across on-premises data centers and in the public cloud, there's also no common control plane spanning these environments. This means that managing, monitoring, securing and governing hybrid cloud solutions is non-trivial.

Migrating a data warehouse to the cloud can be a huge undertaking due to the complexity of the upstream data feeds and number of downstream applications it supports. One of our customers had over 5,000 Informatica ETL jobs pointing at its legacy data warehouse, as well as thousands of downstream applications and users. While it's easy to say "just lift-and-shift" your data warehouse into the cloud, the reality is that you must pull in – or rearchitect – all the connective ecosystem tissue too. Re-platforming as part of a modernization initiative is made easier if the data warehouses involved are based on a common open standard, such as PostgreSQL. When moving between two PostgreSQL-compatible data warehouses, many of the data integration and business intelligence tools used will continue to work, or will require a lower level of migration effort compared to switching between two data warehouses with their own SQL dialects.

When considering data warehouse modernization, it is crucial to ensure that the chosen platform will be capable of scaling as your business grows. Cloud data warehouses such as Snowflake operate well at relatively low data volumes and concurrency levels, but scaling out to accommodate more users and more workloads can be expensive. Look for modern data warehouses that combine the best-of-breed approaches from the on-premises world, such as workload management and high concurrency and availability, with the cloud characteristics of elasticity and separate compute and storage. Also, future-proof your investments by considering platforms that not only handle traditional bulk data management processes, but also are ready to support analytics for emerging near real-time streaming use cases.

Yellowbrick

An alternative approach: Distributed Data Cloud

What if there could be a different way? What if enterprises could consume data warehousing in an environment that delivers the same user experience, security, features, and performance anywhere: on-prem, public cloud, or even at the network edge? What if data warehousing could be provisioned at the point of need, based on data gravity, latency, sovereignty, and governance considerations? What if you could eliminate concentration risk, vendor lock-in and unpredictable cloud spend? Why should an enterprise be forced to commit to a single public cloud vendor, or forced to delay a move to the cloud and remain in their own data center? Why should they care about the underlying infrastructure they run their analytics on, provided it delivers the required business outcomes at the desired cost?

There is an alternative to the status quo that the largest enterprises find themselves trapped in today – one that addresses the challenges identified above. This alternative is known as the Distributed Data Cloud, and it's deliberately designed to break down the barriers to data integration across on-premises, cloud and network edge deployments.

Distributed Data Cloud hides the details of the underlying platform that data management and analytics tools run on and provides the same user experience everywhere. In a Distributed Data Cloud, integration and migration become easier, the risks from vendor lock-in are reduced, and cloud spend becomes more predictable.

A Distributed Data Cloud displays five traits:

1. A platform agnostic runtime allowing the provisioning of data and analytics anywhere

2. A common user experience anywhere

3. Common security and governance features on any deployment target

4. Cost and technology efficiency anywhere – minimize resources and allow for strong cost management (FinOps) and spend guardrails

5. A single control plane, tying all deployments together, public cloud, on-premises and at the network edge

The Distributed Data Cloud is a data management and analytics architecture that abstracts away the details of the cloud, on-premises and network edge infrastructure from the end user. The consumer remains free to focus on generating value from analytics, managing data and controlling costs, rather than having to concern themselves with infrastructure details.

Yellowbrick

Summary

Cloud complexity and diversity are the enemies of business productivity and the friends of project risk. We believe that data management and analytics platforms should be deployable everywhere, at the point of need. A new category is emerging in large-scale data management and analytics that addresses the business challenges faced by the largest enterprises, particularly those in highly regulated industries. This is called Distributed Data Cloud. The Distributed Data Cloud is an architectural pattern designed to address cloud concentration risk, cloud integration and migration challenges, vendor lock-in and the unpredictability of cloud spending.