Yellowbrick

Modernizing Your Data Warehouse for Hybrid Cloud Analytics

A Yellowbrick Data White Paper

Yellowbrick

Contents

Yellowbrick

THE PROBLEM WITH LEGACY DATA WAREHOUSES

For most organizations, data warehouses are more critical than ever. Being able to consolidate information from many operational and transactional sources into a single resource for business-critical, responsive queries is key to responding in today's competitive market.

But all too often, existing data warehouses are no longer able to keep up to the task. They are simply too inflexible. They're too hard to scale. They're too expensive to scale. They require too many technical resources to manage and update. And they're too hard to manage in the face of modern requirements such as huge data volumes, growing numbers of users, increasingly complex queries, and real-time data.

As a result, organizations increasingly are looking for a more modern approach to data warehousing. Queries that return answers in seconds, not minutes or hours. A data warehouse that doesn't slow down as more users log on. A data warehouse that can natively ingest real-time data for up-to-the-minute queries. A data warehouse that doesn't require busloads of highly trained and expensive DBAs to manage. A data warehouse that can be deployed and accessed anywhere—from on premises to the public cloud.

While organizations continue to rely on existing data warehouses for business-critical insights, the problems with those systems are numerous. Legacy data warehouses:

- **Are extremely expensive to use and expand.** Organizations have seen an explosion of data over the past five years, from new online sources, real-time sources, self-service transactions, and more. Yet it's very expensive to increase the capacity of traditional data warehouses, and most warehouses can accommo date only a small section of total corporate data for immediate analysis or ad hoc queries. Upgrading data warehouses to make them agile and responsive for new needs can be a costly capital expense (CAPEX).
- **Are slow and inflexible for certain queries.** When data warehouses initially were rolled out in the 1980s and 1990s, corporate data was homogeneous and could be well defined and structured for specific, repeatable queries run by defined business analysts or users. But that's not the world we live in today. Not only has the volume of data exploded, but most businesses have moved to expand access to the data inside data warehouses to new populations of users, creating even more strain on systems. Many organizations find they have to limit the number of queries running against the data warehouse, or the number of concurrent users, to get adequate performance or reports in a reasonable time period.
- **Require an abundance of highly trained staff to maintain.** Legacy data warehouses were not built for ease of management. To maintain performance, they require a significant number of trained staff and DBAs to configure, maintain, tune, and upgrade. From building indexes to pruning data to tuning queries, standard data warehouses are filled with time-intensive management requirements, adding to ongoing costs and overhead while diverting resources from value-added IT projects.
- **Have platform, space, and cost limitations.** Legacy data warehouses aren't flexible and agile. And neither is their footprint. Traditional data warehouses can be deployed only on premises and don't have the flexibility for private or public cloud deployment. They require significant space in the data center, with all the associated costs and overhead. They also tend to be very expensive to upgrade.

While legacy data warehouses are critical for the effective operation of almost all large organizations, they're not the future. In fact, they're increasingly a roadblock to better business decisions, if not a complete dead end.

As Figure 1 shows, legacy data warehouses simply cannot keep up with new requirements.

- **Usage patterns** have grown from the original daily batch reports to interactive BI tools that exploded on the scene around the year 2000 to today's need for high-speed, AI-based solutions that consume huge amounts of data.
- At the same time, **data size** has grown as organizations shifted the focus of data warehouses from quarterly or daily data repositories, to repositories that contain hourly transactional data, to repositories for real-time data. Capturing ever-more data, and finer gradations of it, requires ever-increasing storage—from terabytes to petabytes.
- Of course, the **number of users** has increased at the same time, resulting in more people (and AI programs) running more queries against more data—a triple threat to the traditional data warehouse.
- Underlying all these changes have been **shifting consumption habits,** from internal, on-premises users of the original data warehouses, to private and public cloud deployments over the past 10 years, and now to mixed hybrid cloud environments, requiring flexibility that legacy data warehouses were not designed for. Furthermore, the traditional CAPEX pricing model is now out of step with enterprise requirements.
- And, if all that weren't enough, the basic **data warehouse input platforms have changed.** Originally, data flowed from disparate transactional and operational systems into data warehouses. From there, data warehouse sources were expanded to include data lakes, unstructured data, and semi-structured data. More recently, organizations have looked to "lakehouse" capabilities that would enable them to more reliably import data from data lakes and other sources.

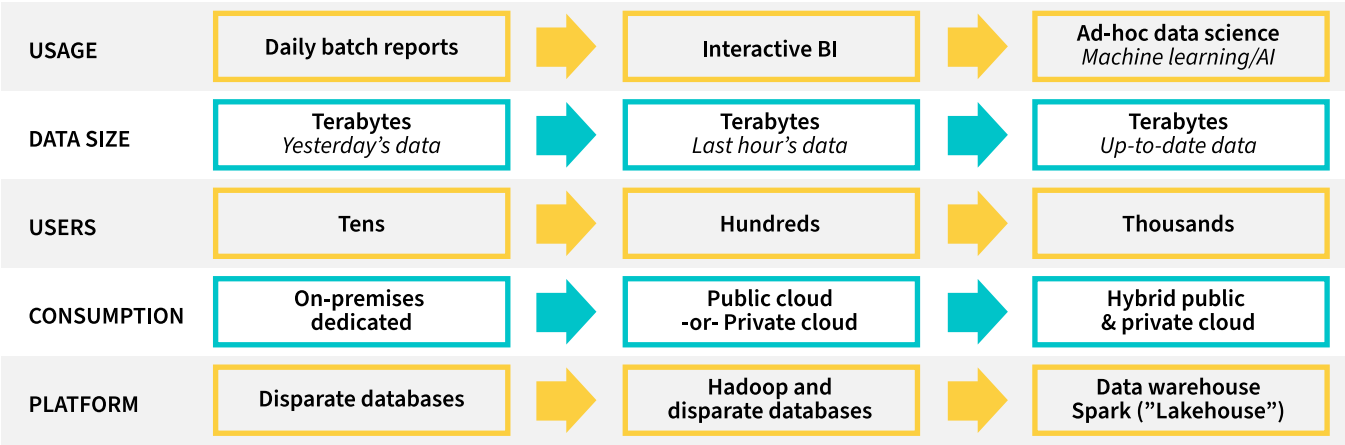

Figure 1: Legacy data warehouse can't keep up with new requirements

Legacy data warehouses simply can't keep up with all these changes. And even if they could, the cost in hardware, software, staff, and maintenance would be exorbitant.

DEFINING A MODERN DATA STRATEGY

What these organizations need is a modern data strategy. One that not only supports all of today's requirements, including superior price/performance regardless of data scale, but also provides a path to the future, with flexible deployment options and expand-as-you-grow architecture. The ideal platform for that strategy would support several key requirements:

- **Price/performance.** One big problem with traditional data warehouses is their high price. Legacy data warehouse vendors have struggled to refresh their platforms in a way that produces good price/performance as data volumes grow and concurrent users increase in numbers.
- **Scalability.** The one thing that's constant with data warehouse deployments is that the volume of data will continue grow, as will the number of users and types of queries. Therefore, when evaluating a more modern platform, it's critical to understand how easy it is to add more data or support more users.
- **Real-time data support.** The ability to ingest and query real-time data (for example, via Kafka) is now a critical requirement. The platform should support complicated analysis on real-time data.
- **Practical support for cloud migration without lock-in.** A modern platform should support a flexible range of deployment options, so that organizations can select the lowest-risk way of migrating to the cloud (e.g., to respect security and data gravity concerns). Some organizations will want to deploy workloads on premises or move to the cloud in a gradual way. A modern solution should run identically across all environments—from on-premises deployments, to a single public cloud deployment, to a hybrid cloud deployment. It shouldn't require an all-or-nothing move to the cloud, and just as important, it shouldn't lock users into a specific cloud platform or impose significant financial or time investments to move data off the platform in the future.
- **Streamlined, consistent management.** Legacy data warehouses typically involve lots of specialized tuning, indexing, workload management, and overall management. A modern platform should be easy to manage, with as few operational tasks as possible to ensure good performance. And whatever management is required should be consistent across all deployment platforms, from on premises to cloud.
- **Predictable OPEX pricing.** While most enterprises now avoid CAPEX as a general policy, their need for accurate forecasting is incompatible with the hidden and complex costs typical of cloud-only alternatives. A consumption model that solves for both needs is important.

A NEW APPROACH TO DATA ANALYTICS ARCHITECTURE

A modern data strategy such as the one described above requires a modern architecture. One like that of Yellowbrick.

Yellowbrick Data was founded in 2014 by experts in database and flash memory technologies who saw an opportunity to solve a huge challenge for data-driven organizations: their inability to get answers to the hardest questions with the speed, detail, and flexibility they need, regardless how much data is involved, while having the freedom to deploy on premises and/or in the cloud.

What was needed, they recognized, was not just an optimization of existing approaches but a complete re-architecture of them, with the critical component being a radical expansion of data bandwidth far beyond traditional boundaries.

With persistent innovation across every layer of the stack—including storage, memory architecture, networking, and OS and database design—Yellowbrick created a modern hybrid cloud data analytics platform. One that's built for today's challenges and tomorrow's opportunities.

Yellowbrick's new approach was a unique combination of Nonvolatile Memory Express (NVMe) storage, a multi-CPU architecture, flash memory in the hardware layer, an optimized OS kernel, drivers, file systems, schedulers, memory managers, loaders, and a Postgres-based SQL database layer (including an innovation called Native Flash Queries) in the software layer to take full advantage the new hardware approach. Just as important, that solution can be deployed as an always-on/single-tenant instance in a data center, private cloud, any major public cloud, or all of the above, with near-real-time replication occurring in the background.

The result is a hybrid cloud data analytics platform that radially expands data bandwidth to support lightning-fast queries on petabytes of data for thousands of concurrent users (see Figure 2).

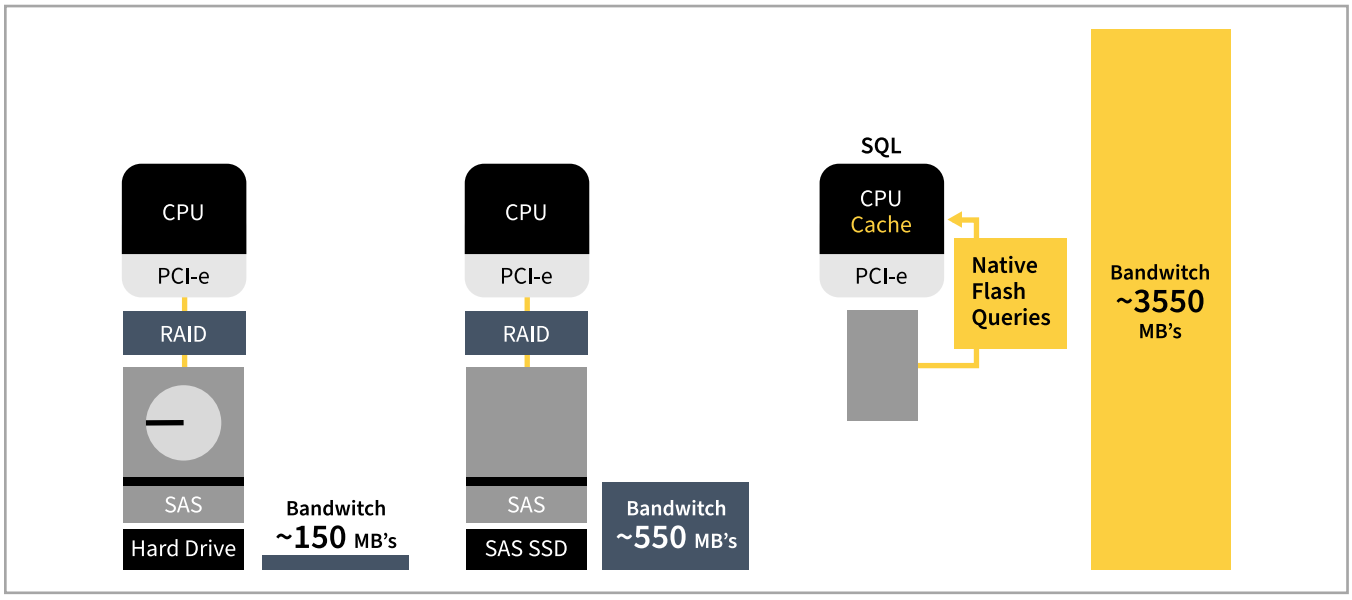

Figure 2: Yellowbrick radically expands data bandwidth to support lightning-fast queries on petabytes of data.

As a result, only Yellowbrick can:

- Rapidly import data at massive rates, in bulk (up to 10TB/hour) or as a stream (200K+ inserts/sec), with data immediately query-able and actionable—giving enterprises instant access to data for timely decisions
- Enable lightning-fast, sub-second ANSI SQL queries across petabyte-size data sets with latencies up to orders of magnitude faster than alternatives—increasing the richness (for example, spanning multiple months of historical data) and rate of insights
- Support parallel queries by hundreds or even thousands of users in familiar BI and data science tools including Tableau, SAS, MicroStrategy, R, and Python, preserving investments in existing tools
- Eliminate mundane tasks that consume valuable administrative time, such as tuning, creating indexes, and reclaiming storage space—streamlining and simplifying data management
- Ensure flexibility via support for hybrid and multicloud by letting organizations run mixed workloads wherever it makes the most sense: in on-premises data centers, private clouds, or any major public cloud platform (AWS, Microsoft Azure, and Google Cloud Platform)

For specific industry use cases, benefits include:

- Insurance and reinsurance companies can do much deeper and more thorough analyses, comprising multiple dimensions and months of historical data
- Financial services companies can assess fraud or quantitative risk much more quickly and accurately
- Retailers can get a true 360-degree view of their customers via real-time access to more transactions
- Manufacturers can identify potential points of failure across IoT fleets much faster, enabling more comprehensive proactive maintenance

BENEFITS OF YELLOWBRICK HYBRID CLOUD DATA ANALYTICS PLATFORM

Yellowbrick provides a scalable data analytics platform with superior price/performance for any data center, private cloud, or major public cloud. Yellowbrick allows organizations to build a real-time analytics capability that leverages existing tools and investments in the cloud or on premises.

Architected as a modern, hybrid cloud platform from the ground up, Yellowbrick allows organizations to consolidate all their corporate data, regardless of where it resides, into a single infrastructure for easy, fast access (see Figure 3). It accelerates standards-based access to business analytics and reporting data spread across public clouds, private clouds, or existing on-premises infrastructures.

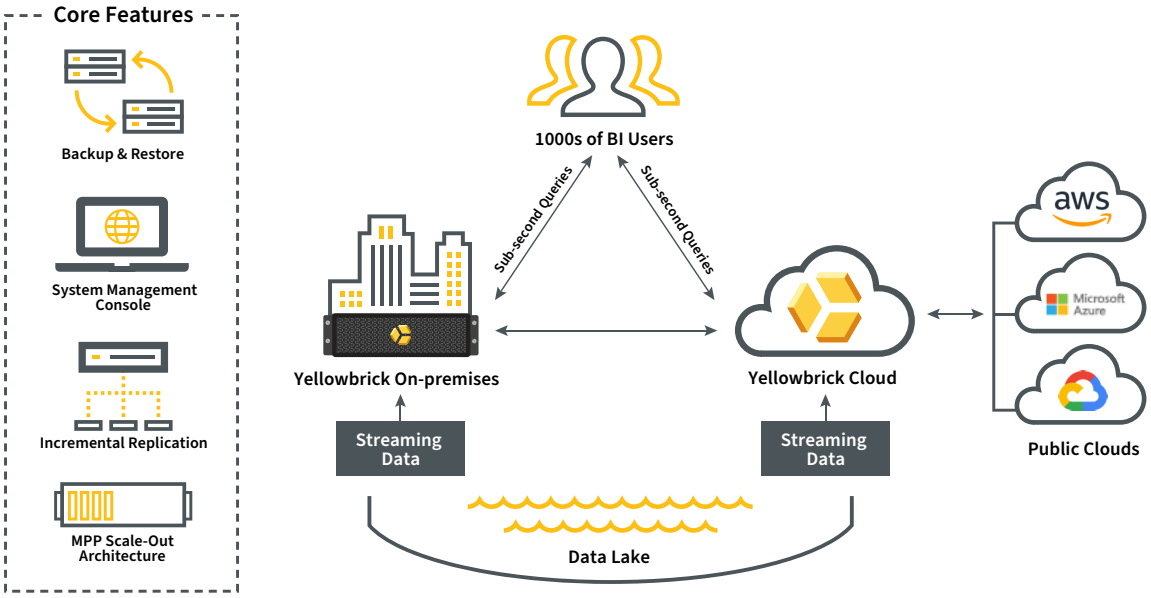

Figure 3: Yellowbrick's modern hybrid cloud data warehouse architecture

Here are a few of the key benefits involved:

- **Superior price/performance.** Yellowbrick is architected for superior price/performance over systems that operate only on premises or only in public clouds.
- **Extreme scale.** Yellowbrick can scale to meet the needs of organizations that have petabytes of data to manage or that need to allow thousands of concurrent users access to data. In addition, Yellowbrick's modular design allows customers to easily add compute nodes as needed.
- **Business continuity and availability.** Yellowbrick has no single point of failure and offers automatic failover for manager node, manager node storage, switches, worker blades, and worker storage. Yellowbrick Assisted Support provides advanced 24x7 predictive monitoring to ensure the health and availability of an organization's warehouse. Furthermore, the Yellowbrick Cloud Disaster Recovery service allows organizations to replicate all on-premises databases to the cloud, and should disaster strike, it can activate a cloud instance (failover). When it is safe to do so, the organization can switch back to its primary warehouse (failback).
- **Ease of Manageability.** Yellowbrick increases performance and flexibility without increasing management complexity. In fact, Yellowbrick simplifies management of hybrid cloud data warehousing. It starts with the single management interface across the entire hybrid environment for managing roles and privileges, multitenancy options, and more. It also provides an easy-to-use interface for managing workloads, so that critical queries, reports, or users can be prioritized. Workloads can be changed on the fly, so that slow jobs don't halt everything. In addition, Yellowbrick eliminates many of the traditional data warehouse management tasks typically performed by DBAs. For example, Yellowbrick automatically takes care of indexes, statistics, and grooming, and it automatically redistributes data when additional analytic blades are added.

- **Deployment flexibility.** Yellowbrick was designed from the ground up to support hybrid and multi-cloud deployments, enabling companies to run their analytics workloads wherever it makes the most sense. This flexibility lets them optimize the economics of their analytics workloads, and it also lets them minimize risk—and avoid taking an all-or-nothing leap—as they migrate to the cloud. For example, Yellowbrick can run entirely on premises, in a company's own data center, with a physical footprint that's a fraction of the size of other solutions. It can be managed via all leading public clouds including AWS, Microsoft Azure, and Google Cloud Platform, and accessible via sub-second latency connections. Or it can be deployed via multiple public clouds, where it works identically and can operate seamlessly across them, enabling it to support a combination of analytics-application, data-storage, and data-motion scenarios.
- Predictable, transparent pricing. All Yellowbrick products are available via a simple annual subscription for OPEX.

USE CASES

Here are some examples of how customers are using Yellowbrick to modernize their data warehouses with proven success.

- ThreatMetrix, a LexisNexis Risk Solutions company, was running a Greenplum analytic database integrated with an on-premises Hadoop-based data lake. After deploying Yellowbrick, the company was able to meet SLAs for running highly complex queries that were impossible before, with data streaming in at 1,500 transactions/sec via Apache Kafka—using just one-third of the nodes, 20 times less memory, and one-quarter of the compute cores of the legacy system.
- For more than 30 years, TEOCO has been a leading provider of analytics, assurance, and optimization solutions to the telecom industry. The company has more than 300 communication service providers (CSPs) and OEM customers worldwide to whom it provides actionable intelligence about network quality of service and customer experience. In late 2017, TEOCO recognized it needed a modern data warehouse after IBM announced the end-of-life for many of the Netezza data warehousing appliances the company relies on. A quick evaluation confirmed that Yellowbrick's flash memory architecture could handle the 30 or 40 billion records TEOCO loads each day without an impact on query performance. With its smaller data center footprint, the Yellowbrick solution is also expected to save TEOCO $5 million in data center costs over the next several years.
- One of the world's largest casino and resort operators wanted to collect and aggregate customer data across multiple touchpoints, but its legacy data warehouse running against Hadoop wasn't up to the task. Today, with Yellowbrick, nearly 100 concurrent users run complex queries against the same data set. Complex dashboards and other views into customer behavior now load in a few seconds instead of a few minutes, with some workloads exhibiting up to a 700x increase in query performance.

Yellowbrick

SUMMARY

If your data warehouse isn't keeping up with your business needs, it's time to consider modernizing it.

Yellowbrick is a modern hybrid cloud data analytics platform that radically expands data bandwidth to support lightning-fast queries on petabytes of data while supporting thousands of concurrent users. It also provides a path forward for data lakes, since Yellowbrick can overcome data lake access limitations and empower analysts with the insights needed to improve decision-making and drive real business transformation.

In addition, Yellowbrick was designed from the ground up to support hybrid and multi-cloud deployments, enabling organizations to run their analytics workloads wherever it makes the most sense and minimize risk as they migrate to cloud.

To learn more about Yellowbrick Data, call us at **877.492.3282** or visit **yellowbrick.com** to book a demo today.