Yellowbrick

WHITE PAPER

Build a lightning-fast data hub on Yellowbrick

See what real-time at scale analytics looks like

Contents

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.

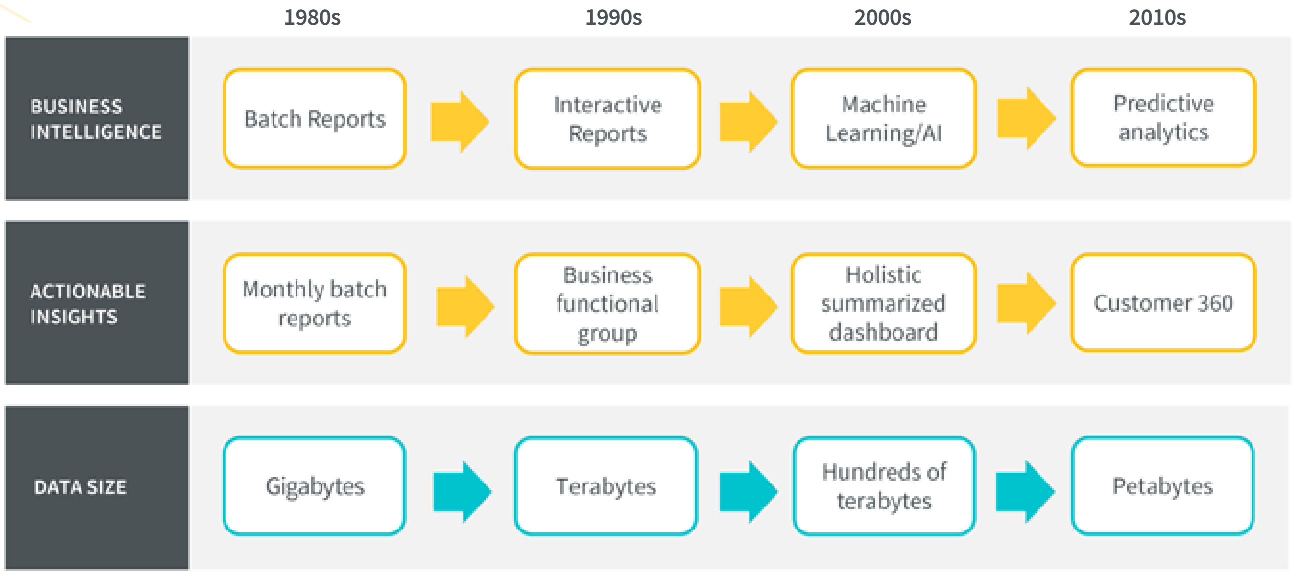
But all too often, existing data warehouses are no longer able to keep up to the task. They are simply too inflexible. They're too hard to scale. They're too expensive to scale. They require too many technical resources to manage and update. And they're too hard to manage in the face of modern requirements 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 accommodate 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.
- 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 warehouse platforms like IBM Netezza, Teradata, Oracle Database, and Microsoft SQL Server 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.



	1980s	1990s	2000s	2010s
BUSINESS INTELLIGENCE	Batch Reports	Interactive Reports	Machine Learning/AI	Predictive analytics
ACTIONABLE INSIGHTS	Monthly batch reports	Business functional group	Holistic summarized dashboard	Customer 360
DATA SIZE	Gigabytes	Terabytes	Hundreds of terabytes	Petabytes

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

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.

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.

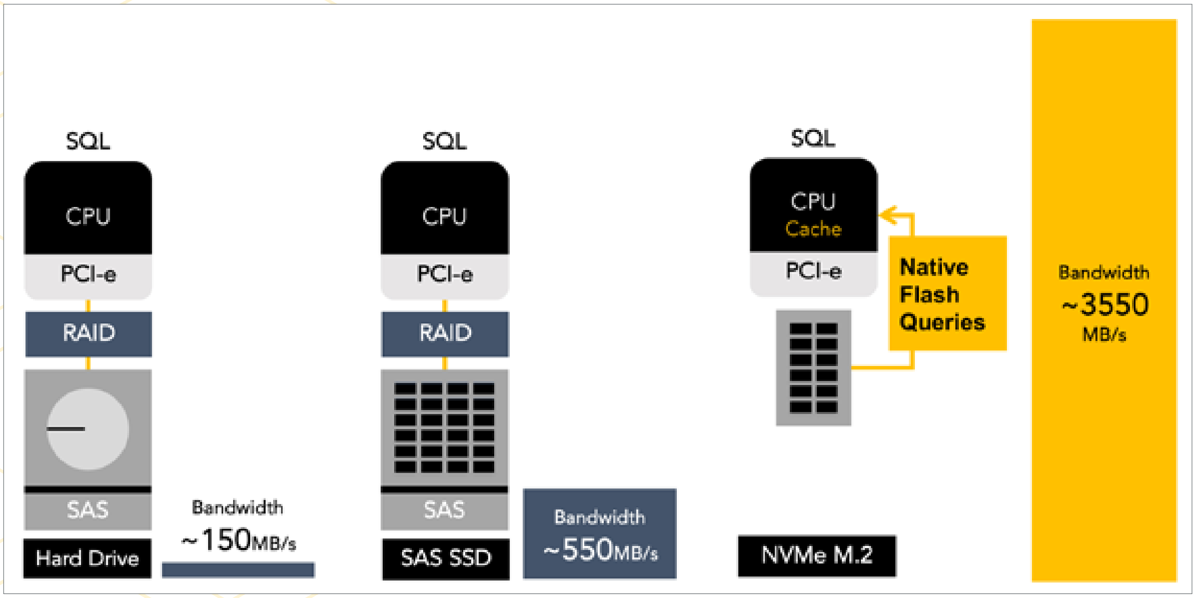


Figure 2: Yellowbrick radically expands data bandwidth to support lightning-fast quieries on petabytes of data.

- **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 Apache 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 subscription 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 predictable pricing model that solves for both needs is important.

A new approach to data analytics

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

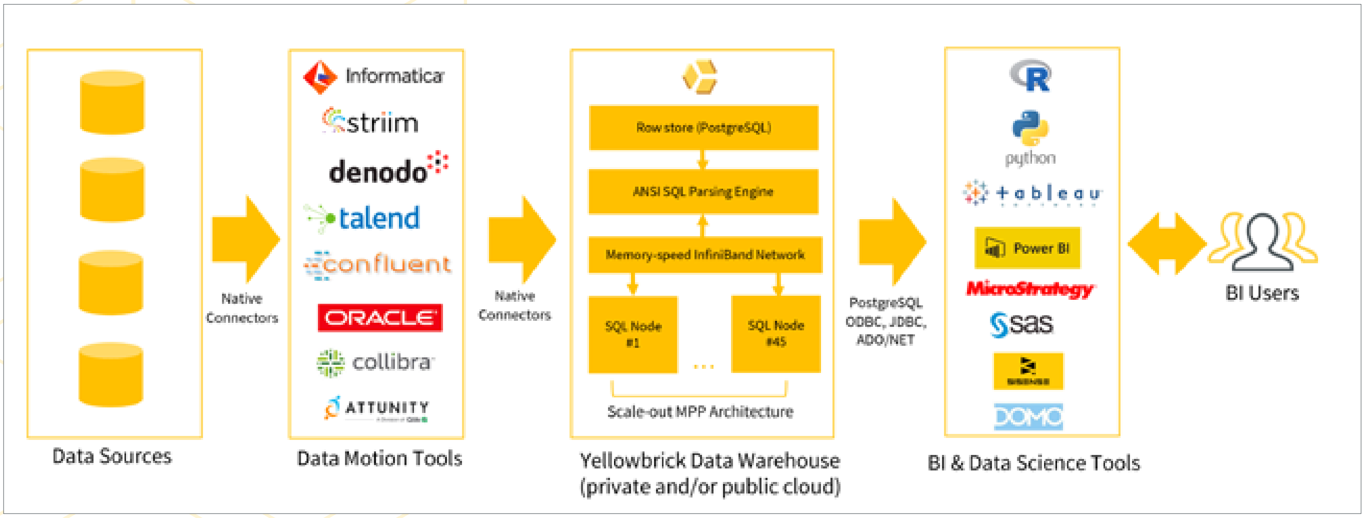


Figure 3: Yellowbrick Data Warehouse and ecosystem

re-architecture of them, with the critical component being a radical expansion of data bandwidth far beyond traditional boundaries.

With deep, 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).

As a result, only Yellowbrick can:

- Enable lightning-fast, sub-second ANSI SQL queries across multi-petabyte data sets at 100x speed and beyond—increasing the richness (for example, spanning multiple months of historical data) and rate of insights
- Support parallel queries by hundreds or thousands of users in familiar BI and data science tools
- Rapidly import data at massive rates, in bulk (up to 10TB/hour) via ETL tools, as a real-time stream from Kafka or via CDC (continuous data capture) from OLTP systems, with data immediately query-able and actionable
- Eliminate mundane tasks that consume valuable admin time, such as tuning, creating indexes, repartitioning data, and reclaiming storage space—streamlining and simplifying data management
- Let users consume analytics from anywhere, whether inside your firewall, from a major public cloud (AWS, Microsoft Azure, Google Cloud Platform), or both

Built for ad hoc workloads

Furthermore, the Yellowbrick solution is built for a world where most queries are ad hoc and the

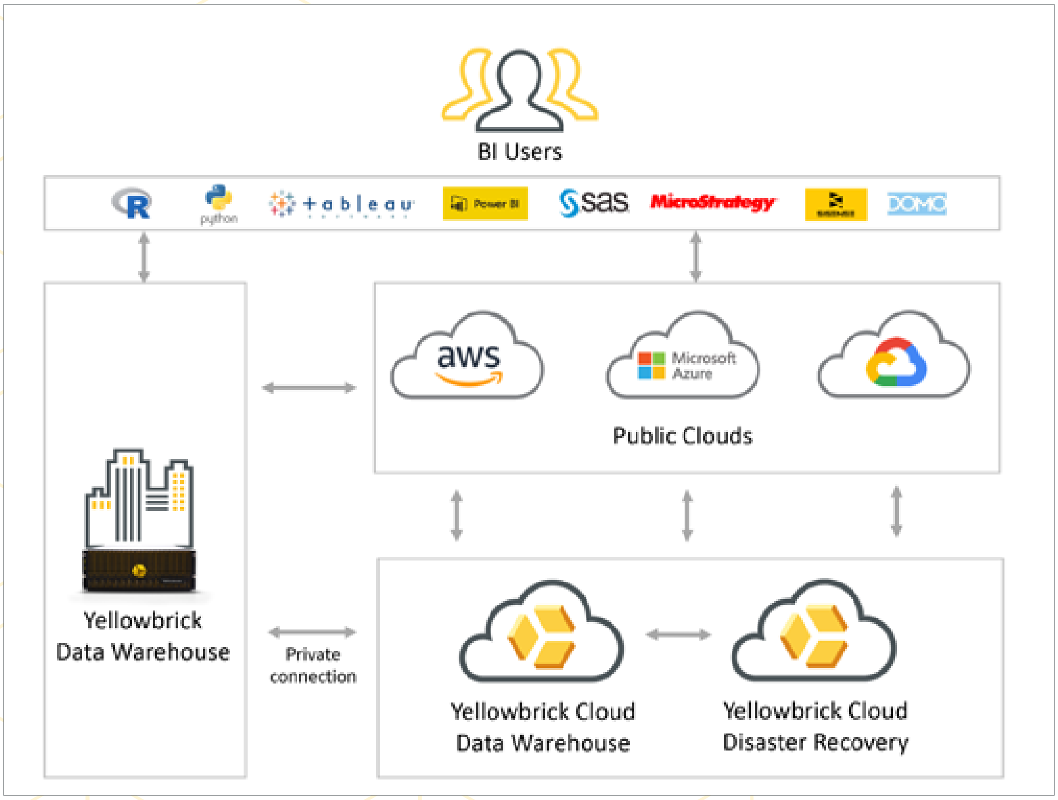


Figure 4: Yellowbrick's modern hybrid cloud data warehouse architecture

data warehouse isn't running a predefined, repeatable workload day in and day out. This requires the following characteristics:

- **Workload management for bad and long-running queries:** Ad hoc users make mistakes and submit poorly coded queries that either return too much data, produce incredibly complex cross-products, or sometimes just are really complicated. In Yellowbrick, such queries can be run-time reprioritized and placed into a "penalty box" to ensure that shorter, interactive queries still complete and resources aren't tied up.
- **Brute-force computation:** The Yellowbrick Data Warehouse is a brute-force query engine that does not rely on inverted indexing or partitioning strategies to achieve good performance. Forward indexes and statistics are automatically gathered on data as it is imported and are kept up to date automatically, and data is reformatted into the most optimal columnar form for fast querying.
- **Ease of management:** Yellowbrick requires basically no management of space. Data partitioning, while supported, typically is unnecessary, and issues with storage space utilization due to skewed partitioned data don't exist.
- **Stability and predictability:** Yellowbrick is highly available, with no single point of failure, and is fault tolerant, suitable as a back-end for 24x7x365 SaaS applications
- **Integration with the modern big data ecosystem:** Yellowbrick interoperates seamlessly with R, Python, SAS, Kafka, and Spark via open APIs, as well as traditional business intelligence and data mining tools. By leveraging the PostgreSQL interface, the user and developer experience feels just like you're building for, and working with, the most advanced open source database in the world.

Benefits of Yellowbrick hybrid cloud architecture

Yellowbrick provides a scalable data analytics platform with industry-leading 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 4).

Here are a few of the key benefits involved:

- **Superior price/performance.** Yellowbrick is architected for superior performance versus aging legacy systems that operate only on premises or only in public clouds on virtualized commodity hardware, and provides identical performance regardless of how it's deployed.
- **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, Yellowbrick offers a full spectrum of business continuity options, from backup/restore, to database replication, to a Cloud Disaster Recovery service that allows organizations to replicate all on-premises databases to the cloud.
- **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 re-distributes data when additional analytic blades are added.
- **Deployment flexibility.** Yellowbrick was designed rom the ground up to support hybrid and multi-cloud deployments, enabling companies to consume their analytics workloads however 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.
- **Radically reduced footprint.** Yellowbrick's amazingly compact all-Flash memory designed uses 1/30th the space of legacy options, putting millions in data center cost savings in reach.
- **Predictable, transparent pricing.** All Yellowbrick products are available via a simple fixed-cost annual subscription.

Use cases

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

- Catalina Marketing is the market leader in shopper intelligence and targeted in-store and digital media. The company delivers $6.1 billion in consumer value annually, combining the richest buyer history database in the world with its own deep analytics and insights to help retailers, CPG brands, and agencies optimize every stage of media planning, execution, and measurement. The company's legacy IBM Netezza system lacked the capacity to support growing workloads, with analysts having to wait 20 minutes or more for their queries to run—leaving only a small window of time (25% of the day) for complex analysis. Now, a single 10U, 30-node Yellowbrick system delivers up to 182x better performance than an 8-rack, 56-node Netezza Mako system, and queries that used to take up to 30 minutes—if they weren't killed first—are now completed in a few seconds to a few minutes.

- 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 Data Warehouse 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 over SQL Server.

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 enables 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.

About Yellowbrick Data

Yellowbrick Data provides the world's fastest data warehouse for hybrid and multi-cloud environments. Enterprises rely on the Yellowbrick hybrid cloud data warehouse to do the impossible in data analytics: get answers to the hardest business questions for improved profitability, better customer loyalty, and faster innovation in near real time, and at a fraction of the cost of alternatives. Yellowbrick offers superior price/performance for thousands of concurrent users on petabytes of data, along with the unique ability to run analytic workloads on premises, in a private cloud, and/or in any public cloud and manage them in a simple, consistent way—all with predictable pricing via annual subscription.

Learn more at **yellowbrick.com**.

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