Yellowbrick

Transform your data lake into a "data lakehouse"

Get insights from your data lake (finally) with interactive, ad hoc SQL analytics

The original promise of the data lake was to provide low-cost storage for all unstructured enterprise data—originally in HDFS but also more recently in object repositories such as Amazon S3—where it could be subsequently analyzed in place. In reality, even with a wide selection of data lake query engines (Apache Hive, Apache Impala, Presto, et al) available for the analytics layer, most data lakes still don't fulfill the promise of large-scale, real-time analytics, including support for hundreds or thousands of concurrent BI tool users, sophisticated ad hoc queries by data scientists, or data-intensive reports.

Best Price/Performance

Yellowbrick Data Warehouse is now the fastest and most modern solution available for data lake analytics, letting thousands of concurrent users explore all your data at the speed at a fraction of the cost of alternatives. Even conventional cloud data warehouses, which rely on VMs tied to general purpose hardware, fail to match Yellowbrick's breakthrough price/performance.

With Yellowbrick, you can continue to use your data lake for what it does really well while meeting the original goal of enabling real-time analytics at scale, and much more.

Only Yellowbrick lets you:

- **Run sub-second, ad hoc ANSI SQL queries** on billions of rows at 100x speed and beyond—increasing the richness (for example, spanning multiple months or even years of historical data) and rate of insights from your data lake

- **Support thousands of users for mixed workloads,** not the tens of users typically supported by data lake query engines and cloud-only data warehouses

- **Rapidly ingest data from HDFS and object stores at massive rates,** in bulk (up to 10TB/hour) and as a real-time stream from Kafka at millions of rows/ second, with no impact on performance and data immediately query-able and actionable

- **Eliminate mundane tasks** such as tuning, creating indexes, repartitioning data, and reclaiming storage space—streamlining and simplifying data management

- **Deploy your data warehouse anywhere:** in data centers, in multiple public clouds, and both (hybrid)
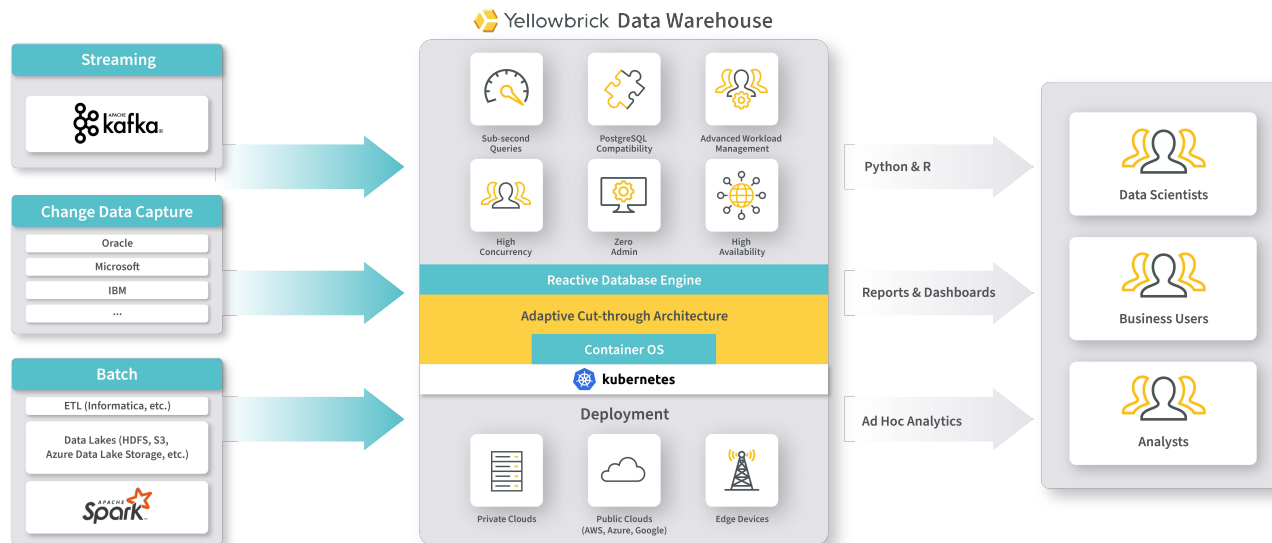
Built for ad hoc queries

Yellowbrick is built for a world where most queries are ad hoc and the data warehouse isn't running a predefined, repeatable workload day in and day out. This requires the following characteristics:

Rich workload management

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

Yellowbrick doesn't rely on inverted indexing or partitioning strategies for peak performance. Forward indexes and statistics are gathered at import and kept up to date automatically, and data is reformatted into the most optimal columnar form for fast querying.

Ease of management

Yellowbrick requires virtually 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 has fault tolerance suitable for "Tier 1" applications.

Integration with the modern big data ecosystem

Yellowbrick reads numerous Hadoop file formats and integrates with R, Python, SAS, Kafka, and Spark, as well as with common BI and visualization tools like SAS, Tableau, PowerBI, and MicroStrategy.

Fast & easy migrations

Migrations are fast and easy from any legacy platform, and we'll work with you to validate your use cases and success metrics along the way.

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

TOP 10
Mobile Operator

TELECOM CASE STUDY

One of the top 10 mobile operators in the world depends on Yellowbrick to safeguard millions in revenue and increase time-to-insight by 20X.

Previously, the company used a legacy data warehouse on top of a data lake to analyze customer data for use by multiple departments--for example, to reconcile revenue from prepaid SIM cards sold by retailers. But with an 800% growth in data volume over time, that platform could no longer keep up with business needs.

Results of a painless migration to Yellowbrick Data Warehouse include:

- **Ad hoc queries complete** 20X faster, with hundreds of concurrent users accessing six months of transactions for deeper, more accurate insights to support up-sell and cross-sell

- **Operational reports now** update in real time, instead of in hours or even days, contributing to more efficient infrastructure utilization 8X more data can be ingested for immediate analysis (up to 1TB per day), enabling analysis of fresher data for use cases like fraud detection

- **Revenue reconciliation now** happens in real time instead batch, safeguarding millions in monthly revenue that was at risk

FRAUD DETECTION CASE STUDY

ThreatMetrix (part of LexisNexis Risk Solutions) is a global anti-fraud SaaS product accessed by thousands of end-users. With Yellowbrick Data Warehouse adding a high-speed SQL analytics layer to its Hadoop data lake, ThreatMetrix helps 5,000 brands validate 5 billion online transactions per month.

- **Even with real-time ingestion** happening in the background, LexisNexis can deliver richer insights to its customers, more quickly, and with fresher data than its previous solution.

- **Yellowbrick automatically reallocates** resources to respond to spikes or unusual usage patterns, and performance tuning is no longer needed.

- **With Yellowbrick instances** located in different global regions, workloads can shift seamlessly between instances when needed.