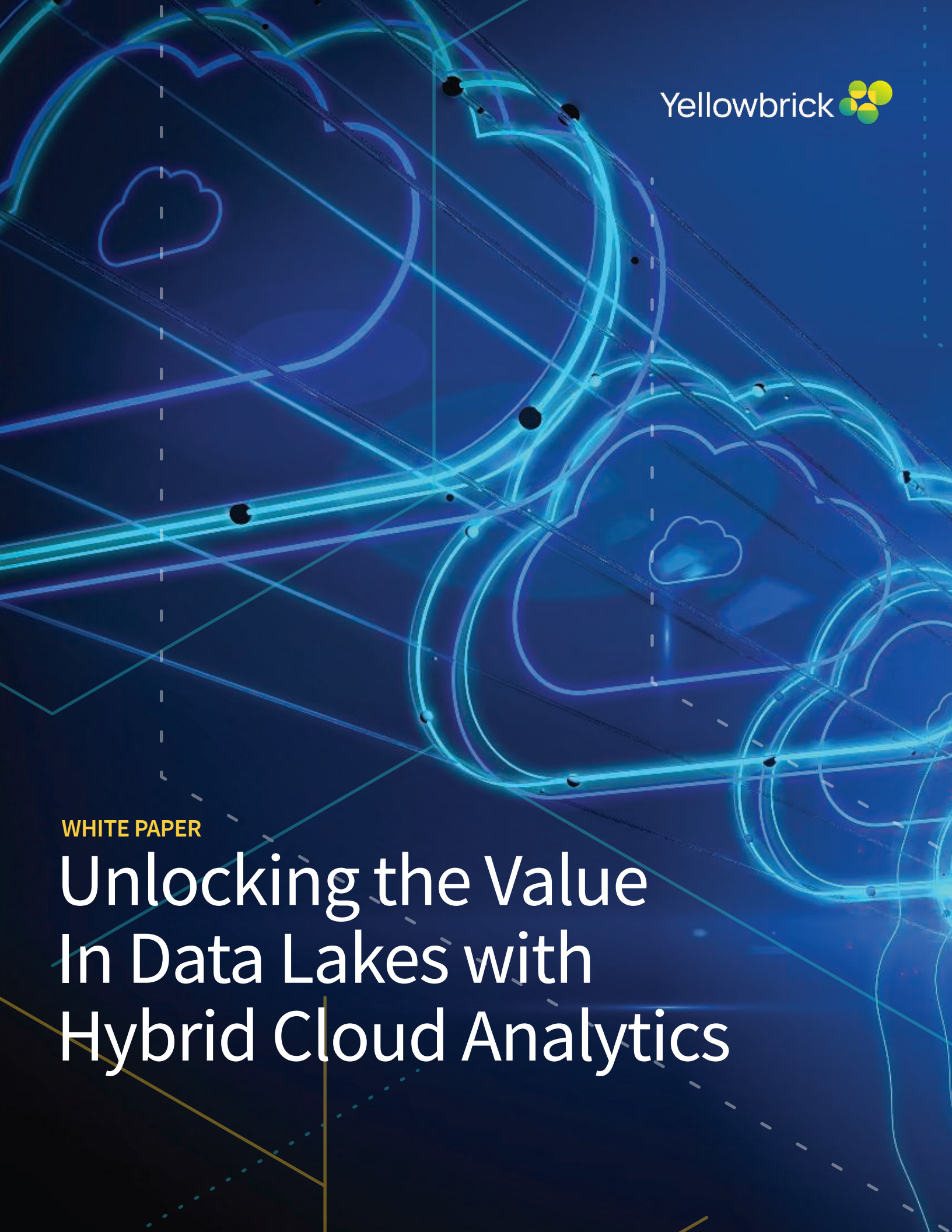


WHITE PAPER

Unlocking the Value In Data Lakes with Hybrid Cloud Analytics

Yellowbrick

Contents

Business Challenge of Data Lakes

Many large companies today have invested in some form of data lake, whether it's based on a Hadoop cluster running on-premises; stored in a cloud object repository like Amazon S3, Azure Blob storage, or Google Cloud Storage; or a combination of the above, with the promise of the ability to get value from unstructured/semi-structured enterprise data via a unified architecture in the form of real-time data analytics. Unfortunately, that promise, despite repeated attempts via open source and commercial solutions, such as Apache Impala, Apache Hive, Presto, Spark SQL, Vertica, and Greenplum, has failed to materialize.

As a result, most data lake owners have discovered that while their investments are very effective as low-cost repositories for storing vast amounts of raw data cheaply, most fall short for meeting the main requirement for doing real-time, large-scale analytics: enabling large numbers of users to run ad-hoc, interactive, and complex SQL queries in parallel on very large data sets in any format. Instead, operationalizing a useful data lake analytics pipeline is difficult, expensive, and time-consuming, requires special data engineering skills, and often fails to meet SLAs for latency.

The solution is to continue using a data lake for what it does well—cost-effectively landing and storing all raw enterprise data—and augmenting it with a modern, real-time enterprise analytics environment that is purpose-built for performance at scale, and for enabling hundreds, or even thousands, of analysts and data scientists to answer the hardest questions, accurately, using their favorite tools. That environment should also:

- Read all required data formats,
- Present a single view of the data for users,
- Automatically optimize data layout to align with query patterns,
- Work with existing tools and skill sets, and
- Offer the flexibility to consume workloads inside the firewall and/or on the cloud

In this paper, we'll explain how Yellowbrick Data has re-thought Massively Parallel Processing (MPP) analytic database architecture across storage, CPU, networking, and software to create a purpose-built platform that finally delivers on the data lake promise via:

1. Industry-leading performance at scale (no more latency/scale trade-offs)—100X faster and beyond for answering the hardest questions,
2. The ability to query data in polyglot formats immediately as it arrives in batch or in a real-time stream,
3. Compatibility with existing BI, data science, and data motion tooling ecosystems as well as existing skill sets,
4. The ability to deploy in an on-premises data center, in a private cloud, or in any major public cloud (or a combination of the above)

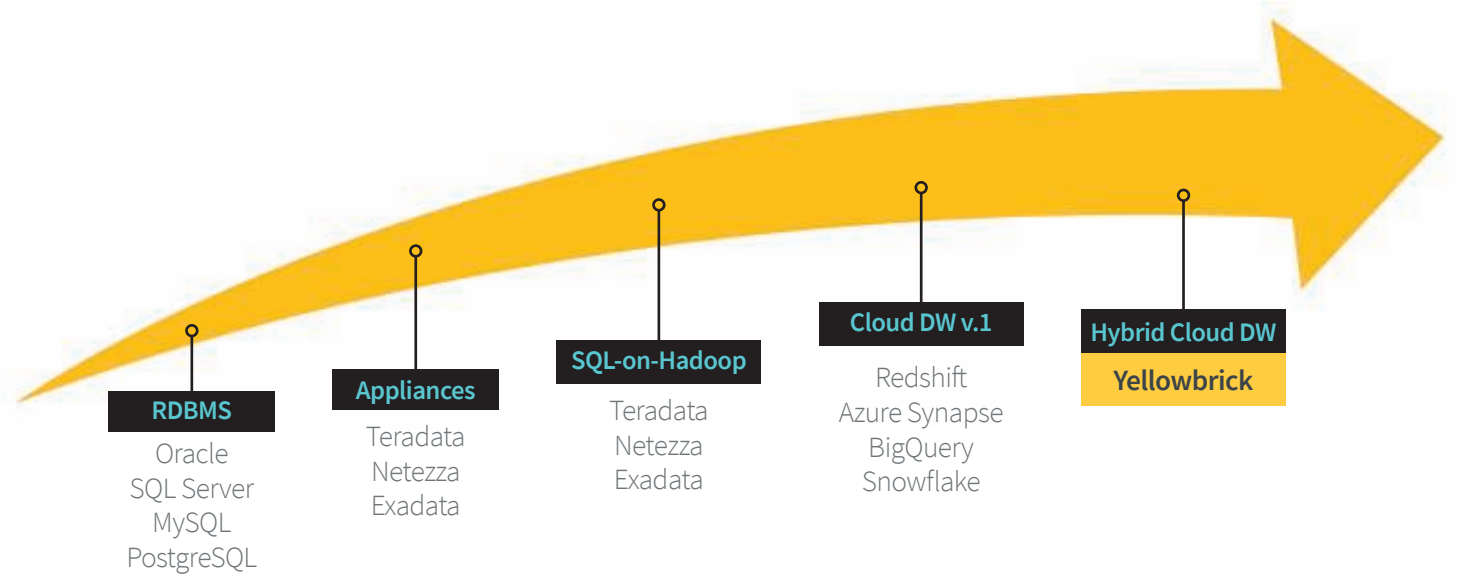
Evolution of Data Analytics

The Yellowbrick hybrid cloud data analytics platform is the next step in the evolution of modern data analytics, providing unparalleled price/performance at massive scale. It enables analysts and data scientists to ask the hardest "what if" questions and get immediate answers via SQL queries, giving them the insights needed to work smarter, deliver better products, delight customers in new ways, optimize operations, save money, and fuel business transformation.

This next step requires deep, persistent innovation in MPP database architecture design, spanning optimizations in both hardware and software, as well as a refreshingly flexible approach to deployment. The result is an approach that combines the performance advantages of purpose-built on-premises technology with the scale and flexibility of the cloud.

Traditional Limitations

Traditional approaches to MPP database design—in-



Evolution of Data Warehouse Platforms

cluding open source and commercial approaches for SQL-on-Hadoop--have been hamstrung by the inherent limitations in hardware. The consequences include high costs (continual upgrades), limited flexibility (only certain queries can be run), poor/unpredictable performance (leading to blown SLAs), and risky forced migrations to cloud-only options.

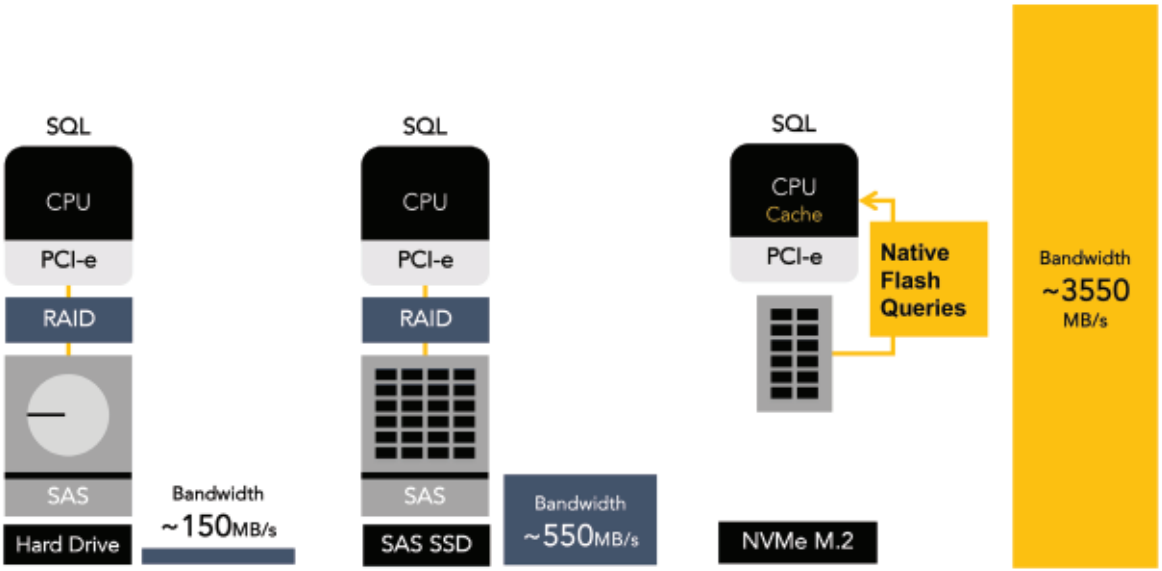
Specifically, bottlenecks between disk, memory, and CPU throttle internal bandwidth, severely restricting the amount of "hot" (query-able) data. To maintain acceptable performance, the on-premises workaround has always been to either invest in increasingly powerful and higher-capacity hardware (scaling up—in the case of data warehouses), or to add more resources/nodes via commodity hardware (scaling out—in the case of Hadoop). In both cases, capital and operating expenses quickly spiral out of control, and acceptable latency and accuracy are not always guaranteed.

Although cloud-native data warehouses abstract away many of these problems by separating storage from compute and virtualizing hardware resources, they fail to address the underlying limitation (and in some cases, add to it through network latency). Consequently, query latency and reliability (not to mention IT budgets) can suffer when too much complexity, or too many concurrent users, are involved.

Rethinking MPP Analytic Database Architecture

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 and detail they need no matter how much data is involved, while having the freedom to deploy on-premises and/or in the cloud.

Almost immediately, they recognized the goal on a technical level to be not just the elimination of existing bottlenecks inside current architectures, but a radical expansion of data bandwidth far beyond the current boundaries. Persistent innovation across every layer of the stack—including storage, memory, networking, and OS and database—was required to meet that goal. The most successful recipe proved out to be a combination of Nonvolatile Memory Express (NVMe) storage, multi-CPU architecture, and flash memory in the hardware layer, and optimized OS kernel, drivers, filesystems, schedulers, memory managers, loaders, and Postgres-based SQL database layer (including an innovation called Native Flash Queries) in the software layer to take full advantage it.



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

As a result, only Yellowbrick can:

- Enable lightning-fast, subsecond 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 from the data lake
- Support parallel queries by hundreds or thousands of users in familiar BI and data science tools such as Tableau, SAS, MicroStrategy, R, and Python—preserving investments in existing tools
- Rapidly import data at massive rates, in bulk (up to 10TB/hour) via Spark or legacy ETL tools, as a real-time stream from Kafka or via CDC (continuous data capture) from OLTP systems, with data immediately queryable 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

To augment the data lake as a low-cost repository of unstructured or semi-structured enterprise data,

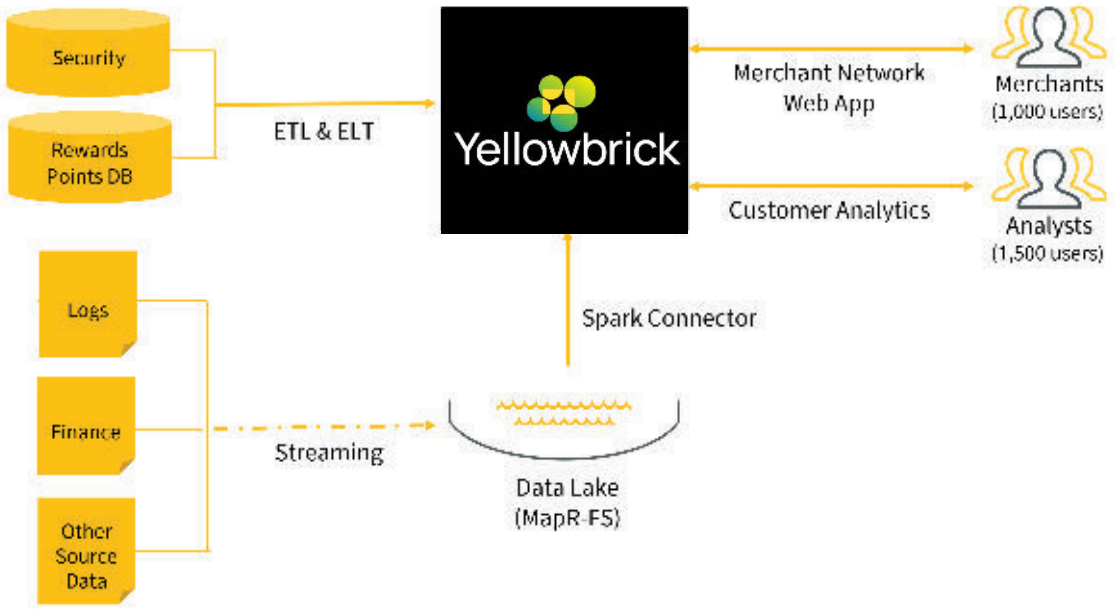
Yellowbrick also supports:

- Any data lake repository, including Cloudera Hortonworks HDFS, Apache HDFS, Mapr-FS, and Amazon S3
- Polyglot Hadoop file formats (ORC, Parquet, Avro, JSON, and so on)
- Bulk ingestion of Hadoop data via an Apache Spark connector as well motion tools such as Informatica, Pentaho, Talend, Syncsort, and Denodo
- Data streaming via Apache Kafka

Built for Ad hoc Queries

The Yellowbrick solution 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:

- **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.



Data Lake Augmentation with Yellowbrick - Reference Architecture

- **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.

For specific industry use cases, that means (for example):

- Insurance and re-insurance 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 the customers via real-time access to more transactions than ever before
- Manufacturers can identify potential points of failure across IoT fleets much faster, enabling more comprehensive proactive maintenance

Benefits of Yellowbrick Hybrid Cloud Analytics

Performance at scale is not the only thing at the top of the enterprise wish list for data analytics, however. Cloud computing has transformed IT requirements ubiquitously because of its ability to give enterprises flexibility like never before, along with easily consumable pricing.

As described previously, due to its reliance on

virtualized commodity hardware, cloud computing alone doesn't solve for the performance/scale trade-off in data analytics. But combining its pricing model, flexibility, and resource-management abstractions with the performance advantages of specialized hardware/software solutions offers the proverbial best of both worlds.

To that end, 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.

Example Use Cases

Here are some examples of how customers are augmenting their data lakes with the Yellowbrick solution, with proven success.

- ThreatMetrix, a Lexis Nexis Risk Solutions company, was running an open-source query engine against its 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—with just one-third of the nodes, 20 times less memory, and one-quarter of the compute cores of the legacy system.
- 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.
- Another customer turned to Yellowbrick when faced with a costly SQL-on-Hadoop upgrade that would push licensing and maintenance fees from $500,000/year to $1 million/year. Testing showed that Yellowbrick could deliver the same 600 TB of usable capacity in a single 6U appliance, with 70% fewer nodes, about 5% as much memory, and 27% as many processor cores—all while consuming about 1/16th the rack space and delivering query response times 3x faster than previously.

Summary

While data lakes provide a cost-effective way to store vast amounts of raw data, they just weren't designed to meet the analytics needs of today's modern enterprise. By augmenting your data lake with Yellowbrick, you can immediately overcome this limitation and empower analysts with the insights needed to improve decision-making and drive real business transformation.

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**.

660 W. Dana Street Mountain View, CA 94041 | USA | 1.650.687.0896