Yellowbrick

TECHNICAL BRIEF

Yellowbrick Versus Apache Impala

The Apache Hadoop technology stack is designed to process massive data sets on commodity servers, using parallelism of disks, processors, and memory across a distributed file system. Apache Impala (available commercially as Cloudera Data Warehouse) is a SQL-on-Hadoop solution that claims performant SQL operations on large Hadoop data sets.

Hadoop achieves optimal rotational (HDD) disk performance by avoiding random access and processing large blocks of data in batches. This makes it a good solution for workloads, such as recurring reports, that commonly run in batches and have few or no runtime updates. However, performance degrades rapidly when organizations start to add modern enterprise data warehouse workloads, such as:

- Ad hoc, interactive queries for investigation or fine-grained insight
- Supporting more concurrent users and more-diverse job and query types
- Analytics for IoT and other real-time data

Customers have migrated more and more of their data to Hadoop over time and have found a much greater need to support these workloads. SQL-on-Hadoop technologies such as Apache Hive and Impala were developed to provide this support. The problem is that these technologies are a SQL abstraction layer and do not operate optimally: while they allow users to execute familiar SQL statements, they do not provide high performance.

For example, classic database optimizations for storage and data layout that are common in the MPP warehouse world have not been applied in the SQL-on-Hadoop world. Although Impala has optimizations to enhance performance and capabilities over Hive, it must read data from flat files on the Hadoop Distributed File System (HDFS), which limits its effectiveness.

Architecture comparison: Impala versus Yellowbrick

While Impala is a SQL layer on top of HDFS, the Yellowbrick hybrid cloud data warehouse is an analytic MPP database designed from the ground up to support modern enterprise workloads in hybrid and multi-cloud environments. Key architectural differences include:

- **Storage and data access.** Impala reads data in large blocks from HDFS over racks of disks in connected servers. The Yellowbrick hybrid cloud data warehouse reads data in small blocks from NVMe flash storage connected via PCI-e.

600TB Usable Capacity	Impala	Yellowbrick	Improvement
Number of Nodes	50	15	3.3x
System Size	1.5 racks (100 rack units)	6 rack units	16.6x
Total Memory	37,500GB	1,920GB	19.5X
Total Cores	2,000	540	3.7x

Figure 1. Yellowbrick architecture proves far more efficient than Impala in customer testing.

Query Response Time %	Yellowbrick (ms)	Impala (ms)	Improvement
99	5,597	1,6791	3x
98	4,826	13,513	2.8x
95	3,793	9,103	2.39x
90	3,070	7,675	2.5x
75	2,044	4,497	2.2x
50	1,070	2,675	2.5x
25	592	1,184	2x
Average Response	**2,999**	**7,920**	**Average Improvement: 2.64x**

Figure 2. Response time percentile for 200,000 queries in customer testing.

- **Data Distribution.** The most common transport for Hadoop and Impala is 10G Ethernet. The Yellowbrick hybrid cloud data warehouse employs a dedicated, self-managed InfiniBand backplane that utilizes RDMA to move large data sets very quickly, at more than 5x the speed of 10G Ethernet.
- **Memory.** Impala must continually aggregate batched, large block data pulled into main memory across many servers. The Yellowbrick system flows small block I/O directly to L3 cache on the CPU at PCI-e speeds. In addition, it uses memory in tandem with Intel Data Direct I/O processing to achieve performance that is orders of magnitude faster than Hadoop-based solutions. A single Yellowbrick MPP node can perform joins at more than 20GB/sec.
- **Data protection.** Hadoop relies on three-way redundancy via replication to protect data, which means the Hadoop architecture requires triple the hardware. The Yellowbrick hybrid cloud data warehouse employs erasure encoding to tolerate failure of 2 out of 15 nodes before data loss occurs, without introducing the massive cost of 3x replication.

Customer example

One Impala customer turned to a Yellowbrick solution when increasing performance demands presented the company with a costly and complex Hadoop upgrade. The customer was already paying $500,000 per year in licensing and maintenance fees, and its Hadoop vendor's recommendation was to double the size of the cluster, which

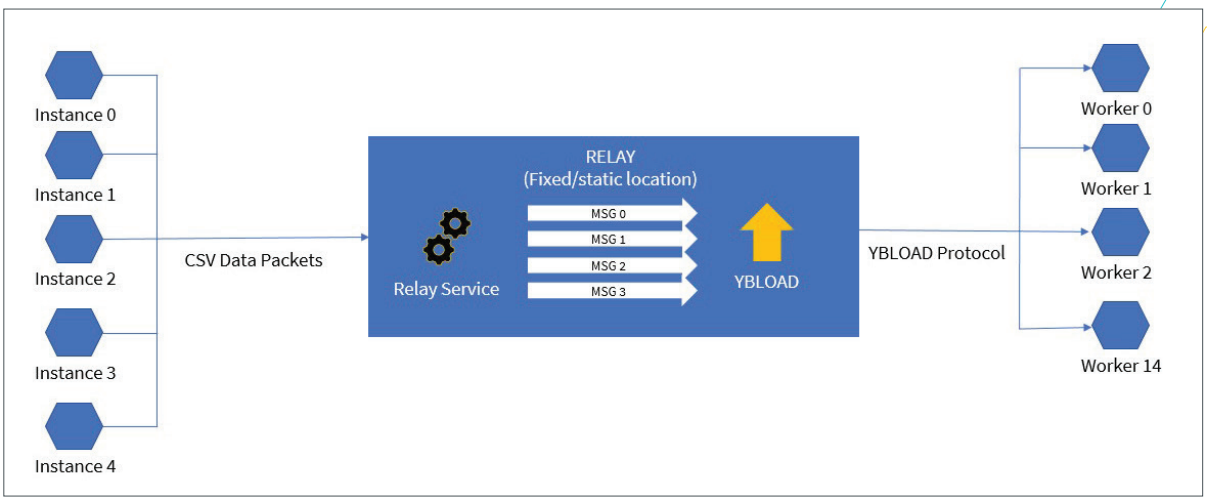

Figure 3. ybrelay listens for Spark or Sqoop jobs and sends data to the Yellowbrick system via ybload.

would lead to more than $1 million per year in licensing and maintenance costs.

The Yellowbrick hybrid cloud data warehouse delivered far greater performance at a far lower cost (see figures 1 and 2).

As Figure 1 shows, to deliver the same 600TB of usable capacity, the Yellowbrick hybrid cloud data warehouse requires 16x less space, which reduces tile costs, and far less memory, which reduces power and cooling costs.

The Yellowbrick system also dramatically improves performance, enabling organizations to achieve faster, higher-quality insights. The example customer runs hundreds of thousands of unique queries per day, looking for anomalies. To compare Impala to Yellowbrick technology, the customer generated ~200,000 queries and measured performance as a response time percentile, in milliseconds.

Importantly, the superior price-to-performance will continue well into the future: with the Yellowbrick hybrid cloud data warehouse, the customer can increase performance 2.5x again simply by adding 15 nodes (for 30, total) in just 4U of additional rack space.

End users noticed performance improvements immediately when the customer replaced Impala with a Yellowbrick solution. Complex queries that used to take 30 seconds now execute in fewer than 2 seconds, and short queries that executed in several seconds are now measured in milliseconds—all while the Yellowbrick system constantly ingests terabytes of new data.

Simple, fast migration

Customers can migrate data from a Hadoop environment into a Yellowbrick system by using any number of known Hadoop-based tools, including Apache Sqoop or Apache Spark. Some customers also use message queues such as Apache Kafka. With these tools, the Yellowbrick system looks and acts just like a PostgreSQL database, utilizing the standard ODBC and JDBC drivers. In fact, the Yellowbrick system uses PostgreSQL ODBC and JDBC drivers for any tooling connectivity.

Customers seeking a more robust feature set and faster data delivery can use the Yellowbrick high-speed tools ybload and ybrelay for loading data. ybload is a Java-based tool that can parallel load directly to the Yellowbrick system nodes. ybrelay is a Java-based listener that can spawn parallel ybload jobs for the fastest loads from Hadoop. ybrelay performs optimally with Spark by using Yellowbrick's Spark SDK. Yellowbrick systems can

access all data from Hadoop via Spark to ingest JSON, Parquet, Avro, or any other file type; perform any data enrichment; and format the data into CSV for loading via ybrelay.

ybrelay is installed and configured on "edge nodes" and listens for a Spark or Sqoop job (see Figure 3). Once a job completes and is ready to write, it requests a ybload job from ybrelay. ybrelay configures four pipes for data, which feed into four ybload jobs, handle the job termination, and commit signaling to ybload. With this parallel functionality, bulk data movement into Yellowbrick systems can happen at wire speed—in excess of 10TB/hour.

Migrating Apache NiFi data feeds to Yellowbrick

Apache NiFi is a tool for data routing and data flow that enables data to move globally across IT ecosystems. Data flows have numerous sources and targets, and Yellowbrick hybrid cloud data warehouses can easily be configured into NiFi pipelines. The Yellowbrick system can be a powerful tool for ingesting, cleaning, storing, and archiving data as well as a powerful SQL-based analytics tool for reporting and data discovery.

NiFi incorporates the Yellowbrick hybrid cloud data warehouse either as a source or a target. As a source, the Yellowbrick system receives data from NiFi via the PutSQL command. It uses the same JDBC driver as PostgreSQL and is configured the same as a PostgreSQL database. The system uses two storage engines: a row store and a column store. JDBC writes flow to the row store, which is perfect for trickle feeds of data less than 1GB. Bulk writes are written to the column store, using the Yellowbrick ybload tool. Customers who need to write bulk data volumes to their Yellowbrick cloud data warehouse should use ybload.

NiFi can source data from a Yellowbrick system by using three common statements: ExecuteSQL, QueryDatabaseTable, and GenerateTableFetch. The ExecuteSQL statement returns the result set of a SQL Select statement executed on the Yellowbrick system. When combined with the ability to scan hundreds of trillions of rows in seconds, a Yellowbrick hybrid cloud data warehouse can provide blazing access speeds to petabyte-scale data sets. This opens the door to deep insights from more data, making Yellowbrick faster in NiFi environments than any other solution available.

Conclusion

The Yellowbrick hybrid cloud data warehouse offers a simple, scalable, high-performance, cost-effective MPP alternative to Impala and other SQL-on-Hadoop offerings. In addition, it automates many common tuning and maintenance tasks, allowing your staff to put more focus on delivering business value, and is easy to install and use. You can deploy today and have results tomorrow.

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. The Yellowbrick system 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