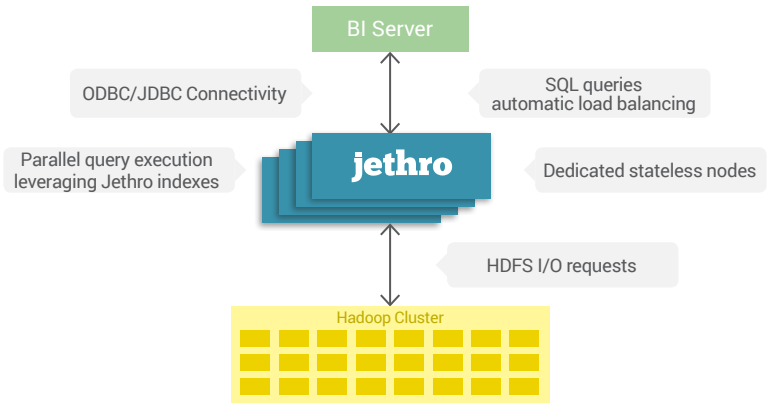
Accelerating Performance on **Hadoop**

jethro

Jethro Is the Only **SQL Engine** That Harnesses **Indexing** to Enable **Real-time BI** on Dashboards like Tableau or Qlik with BILLIONS of Rows of Big Data on **Hadoop**.

High Level Architecture

Jetho's architecture harnesses the power of indexes to deliver superior performance.



SQL Interface BI tools connect to Jethro using JDBC or ODBC and issue standard SQL queries. The driver automatically load-balances SQL statements across all Jethro hosts.

Query Processing Jethro runs on one or a few dedicated, higher-end hosts optimized for SQL processing complete with extra memory and CPU cores and local SSD for caching. The query hosts are statelessv and new ones can be dynamically added to support additional concurrent users.

Storage Layer Jethro stores its files (e.g. indexes) in an existing Hadoop cluster. It uses a standard HDFS client (libhdfs) and is compatible with all common Hadoop distributions. Jethro only generates a light I/O load on HDFS – offloading SQL processing from Hadoop and enabling sharing the cluster between online users and batch processing.

Data Loading and Indexes A loader service processes input files and creates query-optimized column and index files, which are encoded, compressed and then stored on HDF. This service can run on its own host or on one of the query processing hosts.

Jethro Technology Overview

Jethro is an innovative index-based SQL engine that enables interactive BI on Big Data. It fully indexes every single column and select datasets on Hadoop HDFS. Queries use the indexes to access only the data they need instead of performing a full scan, leading to a much faster response time and lower system resources utilization. Queries can leverage multiple indexes for better performance. The more a user drills down, the faster the query runs.

Jethro Highlights

Easy To Get Started Just install Jethro on one or few dedicated servers, point it to your Hadoop cluster, load some data and start querying. Jethro is safe and easy to implement because it only connects remotely as an HDFS client. It doesn't install new services or run resource-intensive computations inside the cluster (MapReduce, Spark and others).

Performance From a Unique Indexing Technology A unique index-based approach is the key to Jethro's superior performance. Jethro's indexes are sorted, multi-hierarchy, compressed bitmaps. Automatically created for every column, indexes are written in an efficient, append-only fashion, thus avoiding expensive random writes and locking. Queries use indexes to read only the data they need, instead of performing full scans, leading to faster response time.

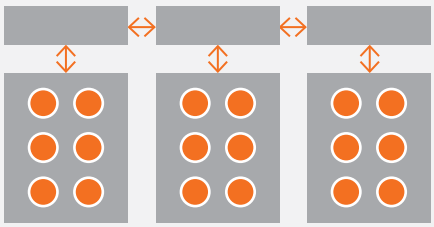
Scalability and High-Availability Jethro's nodes are stateless and highly elastic, allowing easy scale-out to meet concurrency requirements. Jethro's index and column files are stored as standard files on HDFS and benefit from their native scalability and high availability.

Minimal Hadoop Cluster Load Other SQL-on-Hadoop solutions use a brute-force method where each node of the cluster scans and processes its local data for every query. In contrast, Jethro leverages its indexes to surgically fetch only the relevant data for each query, which dramatically reduces the load on the shared Hadoop cluster and frees it for other computations and supporting more concurrent queries.

Hadoop Full Scan / Brute Force

All SQL-on-Hadoop Solutions

1. Reads entire dataset. Every time.



Impact on Hadoop Cluster

Massive number of unnecessary I/O

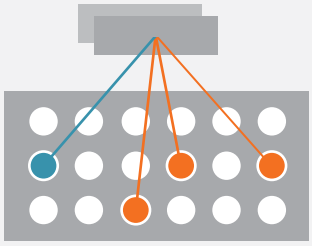
Extensive cluster I/O, CPU and memory usage

VS

Index Access

Jethro

1. Analyzes indexes
2. Fetches only relevant data



Impact on Hadoop Cluster

Drastically lower cluster load

Minimal cluster I/O, CPU and memory usage