Distributed Database Architecture

Data Distribution

This document explains how ClustrixDB distributes data sets across a large number of independent nodes, as well as provides reasoning behind some of our architectural decisions.

Consistency, Fault Tolerance, and Availability

ClustrixDB provides a consistency model that can scale using a combination of intelligent data distribution, multi-version concurrency control (MVCC), and Paxos. Our approach enables ClustrixDB to scale writes, scale reads in the presence of write workloads, and provide strong ACID semantics.

Evaluation Model

ClustrixDB uses parallel query evaluation for simple queries and Massively Parallel Processing (MPP) for analytic queries (akin to columnar stores).

Concurrency Control

ClustrixDB uses a combination of Multi-Version Concurrency Control (MVCC) and 2 Phase Locking (2PL) to support mixed read-write workloads. In our system, readers enjoy lock-free snapshot isolation while writers use 2PL to manage conflict. The combination of concurrency controls means that readers never interfere with writers (or vice-versa), and writers use explicit locking to order updates.

Rebalancer

The Rebalancer is an automated system for maintaining a healthy distribution of data in the cluster. It's the Rebalancer's job to respond to an "unhealthy" cluster by modifying the distribution and placement of data. The Rebalancer is an online process that effects changes to the cluster with minimal interruption to user operations. It relieves the database administrator from the burden of manually manipulating data placement.

Query Optimizer

At the core of ClustrixDB Query Optimizer is the ability to execute one query with maximum parallelism and many simultaneous queries with maximum concurrency. This is achieved via a distributed query planner and compiler and a distributed shared-nothing execution engine.