



VOLTDDB AND HADOOP: HIGH-PERFORMANCE DATA PIPELINING



Hadoop holds great promise for a wide range of historical and exploratory analytics. Its parallel architecture allows data-intensive workloads to be partitioned across large, inexpensive server farms, exploiting massively parallel compute capacity. However, there are stages in the data pipeline where Hadoop processing can be significantly improved. Challenges in typical Hadoop data pipelines include:

- Hadoop is not designed to directly ingest high velocity data.
- Data arriving from source systems must be cleaned, enriched and reorganized before it can be efficiently processed by Hadoop.
- Access to data is only available after batch processes have run, introducing significant reporting latencies.

Due to these challenges, many organizations report unanticipated costs and operational complexities when implementing Hadoop systems.

Hadoop's strength is deep data analytics, allowing insights to be derived from mountains of data. However, many Hadoop uses involve high-speed data capture, cleaning, enrichment and preparation for downstream processing by analytic applications. By combining VoltDB and Hadoop, organizations gain immediate, measurable benefits.

BIG DATA PIPELINE

In combination, VoltDB and Hadoop create a Big Data pipeline that enables real-time analytics and dashboarding, while also conditioning data for delivery to Hadoop. When used as a universal Hadoop loader, VoltDB makes time-critical data available immediately, significantly reducing time-to-value.

VoltDB's relational model and familiar SQL language allow incoming data to be quickly deduped, aggregated, enriched and denormalized—all of which reduce the time needed by Hadoop to produce actionable insights. And because data is scrubbed, improved and right-sized before it reaches Hadoop, system resource costs are often dramatically reduced.

1. Data streams into VoltDB from high-velocity sources. VoltDB is an in-memory RDBMS offering read/write throughput that can exceed millions of operations per second and can scale out across CPUs in a distributed cluster.
2. Pre-processing of data bound for Hadoop results in higher quality Hadoop inputs and more cost-effective, predictable execution of Hadoop jobs. VoltDB supports a wide variety of data transformations including de-duplication, aggregation and denormalization.

3. As soon as data is in VoltDB, it is instantly accessible for reporting and dashboards. Additionally, VoltDB integrates easily with popular BI tools such as JasperReports, BIRT and Tableau.
4. VoltDB's high-performance export module allows data to be delivered to Hadoop efficiently and reliably. Selected data can be streamed from VoltDB directly to target HDFS volumes or indirectly via loaders such as Sqoop.

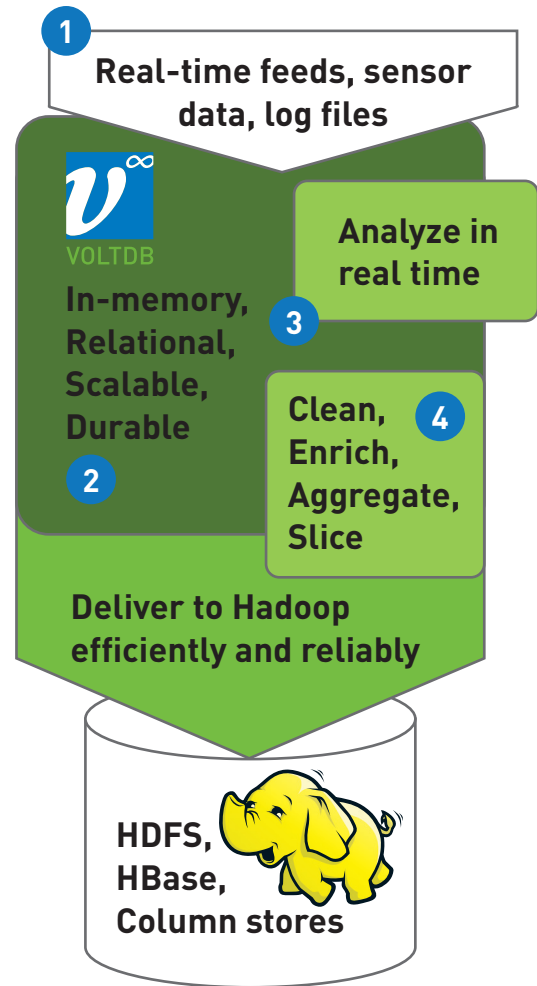
A leading Web search firm updates indexes in four-hour increments while their competitor indexes dynamically in real time.

A digital advertising vendor gauges campaign effectiveness via end-of-day reports, but campaign performance fluctuates continually over the course of the day.

A service provider collects click stream logs to a file system, loads them into a Hadoop cluster to prepare the data, then exports and reloads for data analysis — essentially doubling Hadoop costs and operational complexities.

GET STARTED TODAY

VoltDB enables you to get the best ROI from your Hadoop investments. Deliver real-time analytics that support faster decision making. Transform and enrich data while reducing Hadoop hardware costs. Simplify operations to get the best productivity from your team. Contact VoltDB today, and put your Hadoop systems into overdrive.



VoltDB and Hadoop – the Ideal Big Data Pipeline

209 Burlington Road, Suite 203
 Bedford, MA 01730
 Phone: +1.978.528.4660
 Fax: +1.978.528.0568
<http://voltdb.com>

