SMART DATA FAST.™ VOLTDB

HADOOP DATABASE INTEGRATION

DATA SHEET

Without understanding customers, businesses might as well lock their doors. Large or small, businesses need to have a pulse on what their customers want, and that means having as much real-time information about them as possible. But it doesn't end there: Companies also need to have current data that breaks down inventory, operations, finances and every component of business management to best serve those customers.

Many businesses try to make sense of all this information. It's within their reach, but they simply can't handle the massive amount of data. Their systems are too slow to handle the rapid speed of data ingestion, analysis and decisioning. They can't handle the velocity that accompanies multiple data inputs and thus can't act on data in real time.

VoltDB shatters this critical velocity barrier with an in-memory relational database that combines high-velocity data ingestion, massive scalability and real-time analytics and decisioning. A wide range of industries use VoltDB to close the data ingestion-to-decision gap from minutes, or even hours, to milliseconds. They include advertisers, financial services, national defense departments, online media outlets, telecommunications companies and public utilities.

VoltDB's clients realize that effectively handling high velocity data ingestion, analysis and decisioning reaps big rewards. Data has tremendous value as soon as it's created. Making the most of that information in real time gives businesses a tremendous advantage over competitors that helplessly watch as the value of their data diminishes.

Organizations can leap even further ahead of the competition by integrating a real-time database solution such as VoltDB to a back-end analytics database. Joining these two database engines allows companies to mine historical data for deeper analytical insights while still benefitting from real-time data – delivering new value from a previously untapped and underused class of information.

Hadoop is an open-source framework that's capable of managing large datasets of historical information. Because of its distributed processing and distributed file systems, Hadoop can handle terabytes – even petabytes – of data. This is possible because Hadoop separates the physical storage of data from the application interface for reading and writing. But that is also Hadoop's shortcoming: It is not designed to process and act against data in real time as it is being ingested.

Organizations that attempt to analyze long-term data without the integration of a front-end system such as VoltDB will spend much more time and money trying to understand and act on the information, squandering the opportunity to capture the highest value from data.

For one, the process becomes more complex. Data scrubbers are needed to pull data out of the Hadoop Distributed File System (HDFS), clean it, organize it, and then write it back. This has a few undesirable side-effects: The process adds latency to the ability to analyze data; it puts unnecessary processing requirements on Hadoop, reducing the time that could be spent on valuable insight and data science analytics; and it requires custom written code as well as the coordination of "new data processing" for a continual process as more data arrives.

Also, an organization will see higher costs because this roundabout method doesn't offer operational insight or a holistic view of the data entering the system. The data being pushed to HDFS is opaque to an operational team. The team has to perform unnatural and costly acts – data sharding, expensive input-output cards and other measures – to gain insight into the real time of the correctness, veracity and quality of service of the incoming data.

This all amounts to a lost opportunity. As a business uses Hadoop analytics, it will want to act on the patterns and opportunities it discovers. If all those actions have to be repeated for each event, a business will have to build or integrate a separate decision platform. Using VoltDB's capable and scalable decision platform, an organization can immediately begin with data cleaning and export to Hadoop and not lose out on the opportunities of today and tomorrow.

Integrating VoltDB and Hadoop closes the real-time and historical long-term loop, joining the front and back ends of Big Data. This closed-loop system merges Hadoop's deep analysis of troves of historical data with the in-the-moment decisioning and analytics of VoltDB technology.

Businesses can leverage the built-in Hadoop integration found in VoltDB to combine these technologies. Applications and dashboards can effortlessly interact with VoltDB via SQL and merge data to ultimately present a complete picture of data – both historical and the "now."

VoltDB's relational model and familiar SQL language allow incoming data to be quickly de-duped, aggregated, enriched and denormalized, reducing the time Hadoop needs to produce actionable insights. Because the data is scrubbed, improved and right-sized before it reaches Hadoop, the drain on system resources is often dramatically reduced.

VoltDB's Hadoop Export plugin processes transactional data from VoltDB and writes it, in batches, to HDFS. Configuring this behavior is easy, and requires no programming. Users automate the export process by identifying the specific VoltDB tables in the schema as sources for export data. At runtime, any data written to the specified tables is automatically sent to the VoltDB export connector, which manages the exchange of the updated information to the Hadoop destination. The VoltDB export process queues export data to the connector automatically. The export client runs within the VoltDB cluster, so it, like VoltDB, is highly available.

VoltDB's export processing has other capabilities: No matter what the target for exporting data is – another database, a repository such as a sequential log file, or a system monitor – users don't have to worry. Our technology automates the process.

With its simplicity and quickness, the VoltDB and Hadoop closed-loop system delivers a full picture of information, allowing organizations to not only read data as events happen but also to combine current information with historical trends and complex analytics to make the best business decisions.

This all-encompassing perspective of data in real time empowers businesses to quickly analyze and make bottom-line decisions on vital customer information that can mean the difference between thriving and shutting their doors.