WHITEPAPER

Continuous Real-Time Integration for Real-Time Analytics

Table of Contents

Executive Summary

Because companies once made business decisions based on reports of historical data, they could only address events after the fact. Today, business leaders are increasingly looking to improve responsiveness by using real-time analytics to understand and address business challenges in real-time.

To achieve real-time analytics, businesses need to integrate data from a variety of sources in real-time. This white paper describes why traditional extract, transform, and load (ETL) data integration approaches are insufficient to meet the demands of real-time analytics. It then describes how organizations can achieve business agility through real-time analytics by taking advantage of solutions that integrate data in a highly flexible manner—with superior performance, reliability, and security.

Real-Time Information For the Real World

1 An eCommerce retailer might track how customers are interacting with online touch points and use real-time analytics to respond instantly in a personalized manner. If an online customer demonstrates clear purchase intent by viewing a product video and webpage, real-time analytics might indicate that offering a 10% off coupon will likely drive the customer to purchase the item.

2 A manufacturer might use real-time analytics to inform workers on the shop floor of current inventory levels, status of projects, and so on at the beginning of each shift. With this knowledge, shop floor workers are better able to manage the assembly process.

How Real-Time Analytics Improves Business Agility

Business leaders are increasingly implementing real-time analytics to support faster, more dynamic and agile business decisions. According to a survey by TDWI, 76 percent of respondents consider real-time data, business intelligence, and analytics essential to the success of their organization[1].

Real-time analytics can improve decision making for virtually every role in an organization—from operations to customer service and support, sales, IT systems/network management, marketing, line of business managers, executives, finance, eCommerce, and research and development.

Users' priorities for real-time analytics are quite diverse. They include the desire to:

React quickly to customer needs

Differentiate from the competition

Maximize the customer experience

Improve customer loyalty

Identify problems early

Address changes in demand

Respond instantly to unplanned outages

[1] https://tdwi.org/research/2014/09/best-practices-report-real-time-data-bi-and-analytics.aspx?tc=page0

IoT Pushes the Limits

Real-time analytics are particularly important as companies increasingly adopt the Internet of Things (IoT) to meet customer demand or open new opportunities.

For example, the John Deere Company[2] now puts IoT sensors on tractors that measure the pressure exerted on each seed as it's planted. Based on analysis of the readings, managers can adjust pressure and spacing to set seeds at the optimal depth and distance apart for successful growth.

Networked sensors and smart devices produce data every second or even in fractions of a second. This stream of data requires continual analysis to be of value.

Real-Time Analytics Starts with Real-Time Data Integration

To perform real-time analytics, organizations need to gather all the necessary data together in one place in real-time. Of course, the term "real-time" is subjective and lacks an agreed upon definition. For financial services, real-time may mean sub-millisecond updates. In industries where reporting is performed monthly, real-time could mean hourly data processing and analysis. The point is that the solution needs to integrate each data source at the time it is updated.

How Organizations Integrate Data Today

Until recently, organizations have used traditional extract, transform, and load (ETL) approaches to consolidate all the data they need to perform analysis. These solutions are designed to bulk load data from enterprise applications into a data warehouse or data lake for analysis.

In the past, this batch data integration served enterprises well by incrementally updating historical data sources during quiet periods—such as at the end of the business day. Today, many organizations perform these batch updates between one and four times a day, and are looking to reduce the update window to less than an hour.

Removing Obstacles to Real-Time Data Integration

Unfortunately, older ETL solutions were never designed to support real-time analysis. These solutions have several characteristics that make them unsuitable for this use case. The following sections describe challenges organizations using traditional ETL solutions face and the capabilities necessary for a real-time integration solution to support real-time analytics.

[2] http://www.networkworld.com/article/3071340/internet-of-things/john-deere-is-plowing-iot-into-its-farm-equipment.html

Keeping up with Growing Data Volumes
And Shrinking Processing Windows

With traditional ETL solutions, organizations typically run their business on applications that use data managed by a transaction-oriented RDBMS system such as Oracle or Ingres. At pre-determined intervals—often overnight—the organization updates the data warehouse with data processed the previous day from one or more applications.

These batch processes present several challenges to modern organizations:

Growing volumes of data are making it increasingly difficult to load all data within the available quiet period. This means data available for analysis can be out of synch with operational data.

Decisions have to be made based on the analysis of historical data alone.

As organizations increasingly serve customers 24x7, time windows available to update data on a batch basis have decreased. Many organizations have reached a point where they can no longer stop their operations for any amount of time to update the data warehouse.

SOLUTION—LOG-BASED CHANGE DATA CAPTURE DELIVERS CONTINUAL INTEGRATION

A purpose-built real-time data integration solution provides continuous integration through log-based change data capture. This capability allows the solution to transfer and integrate changes to the data incrementally as they occur, rather than making larger updates all at once. As a result, it:

Reduces the risk that data in production databases and data warehouses will be out of synch

Ensures that decisions can be made based on analysis of the latest data

Supports 24x7 operations

Support for Disparate Systems

Today's organizations want to integrate any data from any source, whether it's stored on premises, in the cloud, or is generated as a stream.

While most organizations continue to use on-premise applications, they are also increasingly adopting software as a service (SaaS) applications in the cloud. And they're looking to integrate these systems. Some of this need for integration is strategic, and some is driven by shadow IT. Organizations need to be able to integrate these data sources where it makes sense and to manage that integration in a hybrid fashion. But while it is possible to use traditional ETL solutions to integrate data from cloud-based solutions, they were not designed for this purpose. Similarly, traditional ETL systems are unable to take advantage of native functionality in streaming environments.

SOLUTION—SUPPORT FOR MULTIPLE HETEROGENEOUS DATA SOURCES

A solution designed from the ground up to support the environments the organization needs to integrate will be more streamlined and require less manual coding than one that bolts this functionality onto a traditional data integration product. A real-time data integration solution should support the most popular relational, columnar, document storage and streaming data sources, Hadoop targets and file locations. It should also bridge the gap between on-premise applications and the cloud as well as between clouds from different service providers.

Improving Data Integration Flexibility

Organizations have long employed data management and data integration professionals to manage data integration across the enterprise. These professionals were responsible for defining the data schema that could address a specific set of questions—a process that could take months.

Because organizations had these data management/ integration resources available, they often did not design their data warehouses strategically. Despite the fact that data warehouse technology promised to provide a single data repository across the organization, they purpose-built data warehouses to answer specific questions. The result was siloed data warehouses across the organization.

Increasingly, developers and business users want to integrate data from multiple sources very quickly to drive new applications or make business decisions. They don't want to wait for a new data processing pipeline, which can lead to bottlenecks in responding to change.

Organizations using real-time analytics to improve business agility need to think strategically about the way they build their data warehouses. And they need more strategic data integration tools that make it easy to quickly change data integration pipelines so they can ask different questions or incorporate new data sources quickly.

SOLUTION—EASIER DATA INTEGRATION

Organizations looking to improve business agility need a solution that makes new integrations quick and easy. The solution should provide a single place to design, manage and monitor real-time data integrations. An intuitive, point and click interface that requires no coding can improve productivity as data transformations are performed.

Automated capabilities should be available to save time when performing the integration. Look for a solution that allows the organization to automatically map tables and columns to avoid getting bogged down in details and reduce the likelihood of configuration errors. Such a solution enables developers to apply transformations and other built-in logic, such as soft deletes, to one, several, or all tables in just a few clicks. At the same time, developers can maintain control by easily overriding automatic mapping if necessary.

With the right solution, organizations can specify mappings and transformations to create missing target tables and columns or perform parallel heterogeneous data validations and repair only once. These mappings and transformations can then be automatically included in parallel heterogeneous data loads. Such capabilities streamline the development lifecycle by allowing developers to specify transformations one time, then deploy them many times—to test, QA and production environments.

Additional Critical Elements of a Real-Time Data Integration Solution

In addition to providing instant integration of multiple data sources in an easy-to-manage fashion, a real-time data integration solution should deliver high performance, reliability, security, and monitoring capabilities.

1. FAST DATA CAPTURE AND TRANSMISSION

Organizations need a solution that delivers high-speed data capture and transmission locally and globally.
Solutions can deliver excellent throughput through the use of:

- **Log-based change data capture.** Rather than replicating all of the data, the solution should replicate changes to the data to minimize impact on network traffic and performance.
  Log-based change data capture also has no impact on the performance of the source system.

- **Data compression.** Advanced compression algorithms can dramatically reduce the traffic flowing over the network, minimizing network latency.

- **Smart network utilization.**

- **Use of native technology wherever possible.**

2. RELIABILITY

The solution should be extremely reliable, preventing failure in the event of a host crash or power outage. Mechanisms that enhance reliability include backing up metadata in a backup database as well as sharing a disk between the primary hub and the backup hub to make transaction and state files available on the failover server.

3. SECURITY

Security for data in motion and data at rest is critical for compliance with numerous regulations. The system should deliver security measures that include encrypting sources and targets, as well as providing the option to use SSL to securely transmit data across the wire. Some solutions include a utility to generate private key/public certificate pairs (but any pair can be used). Because every source and target can use a different pair, organizations can take advantage of granular security. For example, if an organization has different groups managing the source and the target systems, it can specify one login for replication from the source to the hub and another from the hub to the target.

4. MONITORING

Database administrators should have the option to use a management console to monitor replications and view email and SNMP alerts or replicate statistics to a third-party monitoring system.

Conclusion

As organizations look to become more agile, they need to make better decisions as events transpire. A real-time data integration solution makes real-time analytics a reality by ensuring these solutions can work with up-to-the-minute information, without having to wait hours for batch updates. Solutions that are built specifically for this purpose allow organizations to quickly add any data source or answer new questions as they arise without waiting for weeks or months for specialists to build new data pipelines. At the same time, offer the pre-requisite high performance, reliability, and secure communications.

HVR

For more information on HVR and to request a trial and consultation please refer to the provided links:

hvr-software.com

info@hvr-software.com

hvr-software.com/resources/videos

twitter.com/hvr_software