WHITEPAPER

Streaming and Data Lakehouses:
A Comprehensive Market Guide

Introduction

How much does a 12–hour delay cost you?

In the age of real-time financial markets, customers demanding immediate gratification, and fast-moving external threats, insights that arrive late quickly lose their business impact. That's why the most successful companies today don't just react fast – they predict, adapt, and act in real time.

But that level of responsiveness requires more than just good data. It requires immediate data.

Organizations that achieve real-time data insight capabilities secure an incredible advantage today. But for the time being, that capability isn't universal, with some industries embracing and leveraging it while others lag behind. As the business world begins to recognize the vast potential associated with real-time streaming, what currently presents as a competitive advantage for some will become standard practice in the not-so-distant future.

This market guide explores the opportunity of real-time streaming and data lakehouses: why they matter, what's holding enterprises back, and how data lakehouses can eliminate latency altogether, making your organization more competitive, responsive, and resilient to today's most pressing challenges.

Table of Contents

Why Real-Time Streaming Matters: Batch vs Stream Processing 4

Benefits of Real-Time Streaming: Real-Time Streaming in Practice 5

Roadblocks On the Path to Real-Time Streaming Data 5

Achieve Data Immediacy with Cloudera 6

Make it Happen: Real-Time Streaming Solutions 6

The Apache Impact 7

What's Next? Combining the Power of Streaming and Data Lakehouses 8

About Cloudera 9

Why Real-Time Streaming Matters: Batch vs Stream Processing

Before we fully dig in, let's examine the current state of real-time data and enterprise insights. For many organizations, access to real-time insights still feels like a pipe dream — a distant goal made challenging by high volumes of data stored across siloed systems using outdated, batch-based thinking.

The primary challenge with batch-based processing is that it collects data in groups and then holds it to process at once before forwarding the results on to their intended destination. While this strategy worked in the past, with data volume skyrocketing and pressure for real-time insights growing, batch processing is revealing its limitations.

What are those limitations? In short, batch processing delays insights until it reaches the pre-set threshold required for delivery. Each component adds latency, ultimately forcing companies to make today's decisions based on yesterday's data in markets and conditions that have already changed.

Compare this to the growing interest in real-time streaming. This method, in contrast, ingests, processes, and analyzes data as it is created, eliminating that unnecessary source of latency. When paired with the power of a data lakehouse, real-time streaming can take vast quantities of data — structured, unstructured, semi-structured — and generate impactful results at scale.

This architecture combines the performance and structure of data warehouses with the flexibility of data lakes, into something new, called data lakehouses — like Cloudera's Open Data Lakehouse — to support real-time streaming.

Benefits of Real-Time Streaming: Real-Time Streaming in Practice

Unified data accessible in any location offers significant benefits, including:

- Access to faster or dynamic decision-making
- Better customer experiences
- Reduced risk of fraud and downtime
- Lower total data processing costs
- Improved operational resilience

Here's what some of the benefits of real-time streaming look like in practice:

- **Financial Services.** Banking and finance truly pioneered real-time streaming, with some firms employing the technology for more than 20 years. The ability to move financial information, process transactions, catch fraud, and gain stock information in real-time is critical in modern business markets. Today, real-time transactions are the norm for the industry and its customers alike, and many industries are sure to soon follow suit.
- **Insurance.** Insurance companies use real-time streaming in various risk-management-related capacities, which allows them to process claims faster, predict disaster zones and seasons, and analyze data to forecast and move funds.
- **Healthcare.** Healthcare use cases are vast and cutting-edge, allowing opportunities for patient data monitoring for earlier interventions, medical record analysis and pattern recognition, diagnostic and anomaly detection, streamlined billing, and more.
- **Energy.** Energy suppliers can use real-time streaming to gather data from sensors and meters that help them balance supply and demand dynamically. With instant alerts and monitoring, faster compliance, fire/outage disaster mitigation, demand-based distribution, and more all become possible.

The ability to analyze data and forecast trends and challenges to make better decisions for the future is invaluable. In all industries, real-time data streaming allows for faster, more accurate customer and market analysis, early pattern and fraud detection, improved forecasting, predictive maintenance, smarter prioritization, and avoids costly downtime.

Roadblocks On the Path to Real-Time Streaming Data

While the benefits are clear, there are also a few roadblocks that make real-time streaming difficult to achieve at the enterprise level.

- **Monolithic and Centralized Data Architectures.** Most enterprise systems are not designed for real-time data streaming. Many of these architectures still rely on centralized data management processes that are optimized for high-volume, structured batch workloads rather than low-latency, high-velocity data streams.

  Records held in traditional storage systems require time to ingest, store, and later retrieve records, adding latency to every stage of the process. While third-party storage systems can be used to temporarily hold this data, they can be very expensive, especially at scale.
- **Siloed Systems and Network Limitations.** For many organizations, streaming data is gathered via edge locations before it's moved to a central processing location. Moving data from one location to another adds latency, especially over slower or more remote networks.

  To fix this issue, many organizations turn to high-bandwidth, low-latency pipelines, which tend to be quite costly.
- **Processing Power.** Streaming vast amounts of data requires a lot of computing power. Organizations can try to solve for this by adding more hardware, but that endeavor quickly becomes cost prohibitive.

Achieve Data Immediacy with Cloudera

So, how can organizations overcome these challenges and realize the benefits of real-time streaming? The answer is to embrace a modern data architecture, and the best way to do that is with a "**true hybrid**" approach.

What is True Hybrid? Hybrid is the ability to manage all manner of data–structured, semi-structured, and unstructured–to perform analytics seamlessly and consistently across all clouds and data centers, achieving consistent operations, portability, and governance regardless of where your data lives. Hybrid by itself involves using different ecosystems in cloud and data center, working together in a delicate balance. TRUE Hybrid is doing the SAME thing in the data center and across all clouds, providing a single seamless ecosystem wherever your data is born, processed, analysed, moved or stored.

It's not just splitting workloads between cloud and on-premises environments and connecting them in hopes that they somehow find balance.

A true hybrid strategy focuses on creating a highly flexible, scalable, and adaptable data architecture so that data and workloads can move freely between environments.

Open data lakehouses are an essential component of a true hybrid strategy. They can manage large swaths of data (in all its forms) and make it available in the right environments for analysis or AI modeling.

As data immediacy needs grow, stream processing capabilities that can be easily integrated within data lakehouses are critical. **Cloudera's Open Data Lakehouse**'s flexibility enables these organizations to support the increased workload.

Combining data lakehouses with streaming in a unified system empowers organizations to effectively process, store, and analyze both historical batch data and real-time streaming data. Cloudera offers flexible deployment, including public cloud stream processing for easier scaling. Cloudera Streaming allows organizations to turn streams of data into data products by providing capabilities to analyze streaming data for complex patterns and gain actionable intel. It also includes faster data processing and analytics, which makes data available for those organizations in near real-time.

Make it Happen: Real-Time Streaming Solutions

As organizations across industries look to solve for real-time streaming, there are several solutions on the market today:

- **Snowflake.** Snowflake is a familiar face in data offerings, but its solutions rely on a network of several partners to deliver critically needed streaming tools. Snowflake also has very limited support for data in the data center, where many data streams originate.
- **Databricks.** Like Snowflake, Databricks' approach with data lakehouses and streaming involves working with multiple partner organizations to deliver those critical tools. Databricks also has limited data center support, where we often find data streams originating.
- **Cloudera.** Cloudera does things differently. With its deep integration of multiple, open-source and unique tools, Cloudera's open data lakehouse is an out-of-the-box real-time streaming solution, meaning you have just one vendor to worry about, offering a lower risk, seamless experience from day one. And Cloudera's true hybrid approach means you run the same services in your data center and in any cloud, with a common unified control center and single view of data security, governance, lineage and metadata.

Workflow Map: Cloudera delivers a seamless workflow to better process streaming data in real time

The Apache Impact

Cloudera's platform taps into a vast ecosystem of open-source projects and technologies. That integration, particularly with Apache open source projects, helps set the foundation for a true hybrid platform that can fuel data, analytics, and AI. But how deep does that integration go?

Let's dive in.

From Apache NiFi to Flink, Kafka, and Iceberg, Cloudera's platform offers robust, deep, and frictionless integrations that enhance organizations' functionality and ease of use.

Apache NiFi brings no-code and low-code data ingest, as well as efficient pre-processing of data streams to lower the time and cost of processing and analysis under Flink and Kafka. This further reduces latency and shortens time to insight on any large and wide data stream.

Apache MiniFi adds additional pre-processing at the edge, allowing for local analysis and immediate response while also supporting a global view of the data stream for further real-time insight. This lowers the impact of edge data on stressed infrastructure supporting traditional Kafka and Flink solutions alone.

Apache Iceberg, an open table format purpose-built for large-scale analytics, makes data accessible to multiple compute engines concurrently while guaranteeing data reliability and consistency. Iceberg enriches and reduces the cost and complexity of data storage needed in stream processing while offering key technical differentiators like schema evolution and time travel.

With the added power of Iceberg, Cloudera's Open Data Lakehouse helps organizations maximize the value of their streaming data at scale. This means data professionals can work with the same data sets, avoiding duplication while leveraging strong data governance and faster processing.

The Iceberg open table format allows organizations to avoid vendor lock-in because it makes the data lakehouse vendor agnostic. This means the data itself remains fully owned by the organization and freely accessible by any Iceberg-compatible analytics solution.

Cloudera's streaming capability is powered by **Apache Flink**, an open-source framework that processes both real-time data and historical batches, and **Kafka**, a high-performance, highly available, and redundant streaming message platform.

NiFi and **MiniFi** complement Flink and Kafka by offering greater opportunities for optimization and cost reduction while reducing overall latency between data creation and insight, right up to edge devices.

What's Next? Combining the Power of Streaming and Data Lakehouses

As organizations look to integrate streaming capabilities with the scalability of data lakehouses, working with the right partner can simplify the process.

Cloudera is the only true hybrid platform that works wherever your data lives, across any cloud or on-premises data center. With a single view of the entire data suite and consistent technology across all environments, it brings the compute to the data and not the other way around, reducing costs and simplifying management.

Cloudera's combination of real-time streaming and open data lakehouse capabilities empowers organizations to harness the full potential of their data so they can improve data accuracy, achieve real-time analytics, stand up AI and ML models, and fuel better and faster decision-making. Cloudera Data Flow and Cloudera Stream Processing make that integration easier and help generate more valuable real-time insights from streaming data.

Ready to elevate your streaming data capabilities? Dive in and learn how Cloudera's open data lakehouse works seamlessly to support the most complex streaming needs.

About Cloudera

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible — today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X. Cloudera and associated marks are trademarks or registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com