striim

Get More Value from Apache Kafka

with Enterprise-Grade Streaming Integration and SQL-Based Stream Processing

Striim White Paper

February 2018

Get More Value from Apache Kafka

Table of Contents

Executive Summary

Apache Kafka has proven itself as a fast, scalable, fault-tolerant messaging system, and has been chosen by many leading organizations as the standard for moving data around in a reliable way. Striim provides a non-intrusive, end-to-end solution for moving real-time data from enterprise databases, logs, message queues, and sensors to Kafka, as well as other data targets. In addition, Striim enables users to query and analyze data in Kafka with a SQL-like language.

If organizations are adopting, or are considering adopting, Apache Kafka, they need to determine:

- How does one get data into Kafka from their enterprise sources?
- How does one deliver data from Kafka to targets like Hadoop, databases, or cloud storage?
- How does one process and prepare data?
- How does one perform analytics?
- How does one ensure all the pieces work together and are enterprise grade?
- How does one do all this in a fast and productive way without involving an army of developers?

This white paper will outline how to make the most of Kafka when building streaming integration or analytics applications. It will also highlight critical considerations when including Kafka as part of an overall enterprise data architecture.

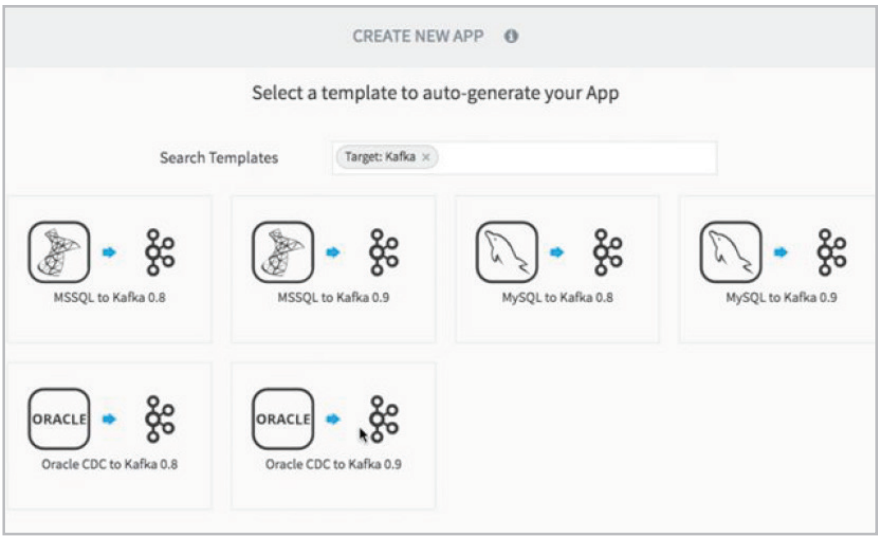
Make Kafka Easy

Kafka is extremely developer-oriented. It has been created by developers, for developers. This implies that if organizations are adopting Kafka they will need a team of developers to build, deploy, and maintain any stream processing or analytics applications that use it. The documentation for Kafka talks about APIs and gives lots of code examples, but if organizations want to do real-time web analytics, build location tracking applications, or monitor and manage security threats, the documentation quickly becomes inadequate.

At Striim, things are thought about differently. All of the things organizations need to do to make the best use of Kafka should be easy, and people shouldn't have to write Java code to create business solutions.

striim

Striim provides a non-intrusive, end-to-end solution for moving change data from transactional systems to Kafka, as well as other data targets.



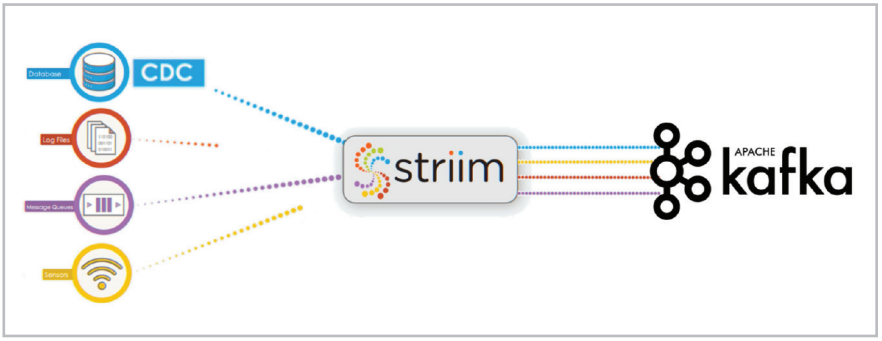
Striim makes it easy to deliver end-to-end streaming integrations and analytic applications involving Kafka

Striim not only integrates Kafka as a source and target, but also ships with Kafka built-in. One can optionally start a Kafka cluster when spinning up a Striim cluster, and easily switch between Striim's high speed in-memory messaging and Kafka using a keyword (or a toggle in the Striim UI). Kafka can become transparent and its capabilities harnessed without having to code to a bunch of APIs.

Striim has offered SQL-query-based processing and analytics for Kafka since 2015. This, combined with Striim's drag-and-drop UI, pre-built wizards for configuring ingestion into Kafka, and custom utilities, makes Striim the easiest platform to deliver end-to-end streaming integration and analytics applications involving Kafka.

Ingestion into Kafka

Striim enables real-time data ingestion into Kafka from a wide variety of enterprise sources.



Real-time data ingestion into Kafka from many Enterprise data sources

GETTING DATA INTO KAFKA

When users are considering how to get data into Kafka, they need to determine how to collect source data in a streaming fashion, and how to "massage" and transform that data into the required format on Kafka. Neither of these steps should require any coding, yet should be flexible enough to cover a wide range of use cases.

STREAMING DATA COLLECTION

The Striim platform ingests real-time streaming data from a variety of sources out-of-the box, including databases, files, message queues and devices. All of these are wrapped in a simple easy to use construct — a data source — that is configured through a set of properties. This can be done through Striim's SQL-based, TQL scripting language, or the UI. The platform also provides wizards to simplify creating data flows from popular sources to Kafka.

The way Striim collects data varies depending on the source, and databases require special treatment.

Most people think of databases as a record of what has happened in the past, with access to that data through querying. However, they can change that paradigm using a technology known as change data capture (CDC). This method non-intrusively listens to the transaction log of the database and sees each insert, update and delete as they happen. Striim makes use of CDC for the database sources, and through configuration can stream out each database operation in real time.

Striim takes a similar approach with files. The file reader does not wait for files to be complete before processing them in a batch-oriented fashion. Instead, the reader waits at the end of the file and streams out new data as it is written to the file. As such, it can turn any set of log files into a real-time streaming data source.

A range of other sources are also available, including support for IoT and device data through TCP/UDP/HTTP/MQTT/AMQP, network information through NetFlow and PCAP, and other message buses like JMS, MQ Series, and Flume.

The data from all these sources can be delivered "as-is," or go through a series of transformations and enrichments to create exactly the data structure and content needed. Data can even be correlated and joined across sources, before delivery to Kafka.

To enable easy scalability and ensure high performance in high-volume environments, Striim uses multi-threaded apply with automatic thread management. Striim supports developers with deep visibility into Kafka integration adapters. Organizations can monitor various integration and processing metrics in real time to easily pinpoint any performance bottlenecks.

When writing to Kafka in Striim, users can choose the data format through a simple drop down and optional configuration properties, without a single line of code.

Kafka is not a destination. Data must ultimately be delivered to enterprise targets. This should not be difficult.

DATA FORMATTING

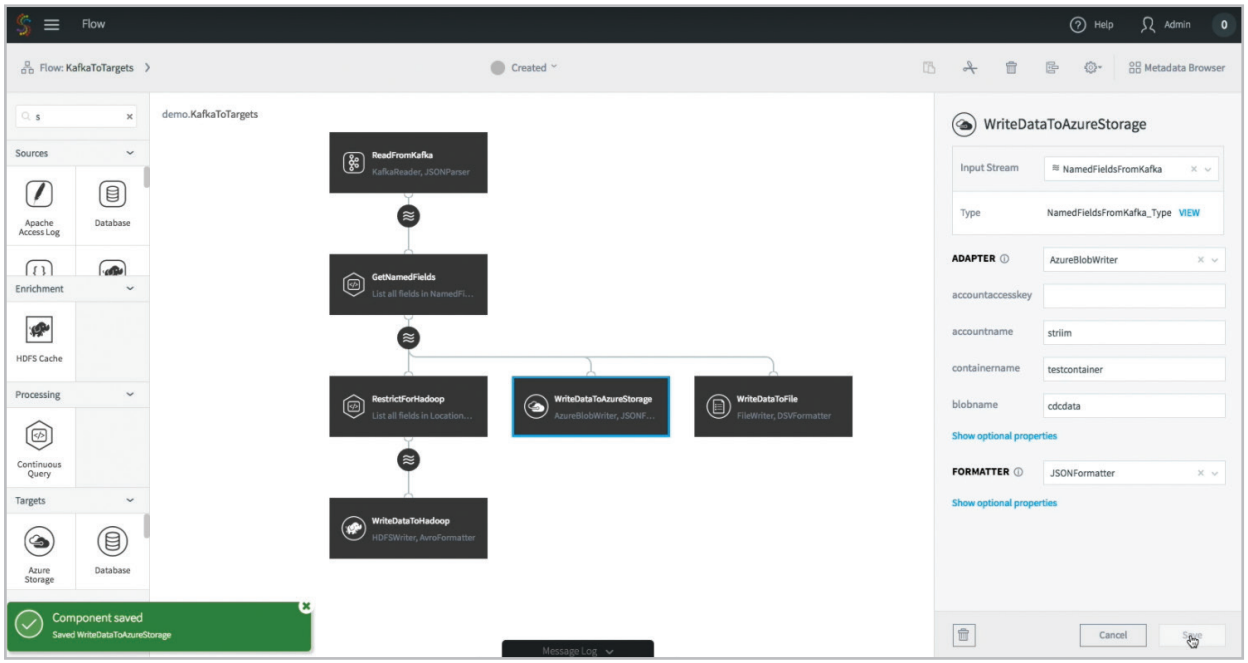
Different Kafka consumers may have different requirements for the data format. Since Kafka deals with data at the byte level, it is not aware of the format of data at all, but consumers may need a specific representation. This could range from plain text, or delimited data (think CSVs), to structured XML, JSON or Avro formats.

When writing to Kafka in Striim, users can choose the data format through a simple drop down and optional configuration properties, without a single line of code.

Delivering Kafka Data to Enterprise Targets

Kafka is not a destination. It may be used as the underpinning of stream processing and analytics, or as a data distribution hub. Whatever the use case, sooner or later organizations will want to deliver the data from Kafka to somewhere else. As with sourcing data, this should not be difficult and should also not require coding.

The Striim platform can write continuously to a broad range of data targets, including databases, files, message queues, Hadoop environments, and cloud data stores like Azure Blob Storage, Azure SQL Database, Amazon Redshift, and Google BigQuery (to name just a few cloud targets that Striim supports). The written data format is again configurable and can be applied to the raw Kafka data, or to the results of processing and analytics. This is all achieved through the drag-and-drop UI or scripting language, and is simply a matter of choosing the target and configuring properties. When reading from Kafka queues, Striim offers automated mapping of partitions to increase development productivity and accelerate time to market.

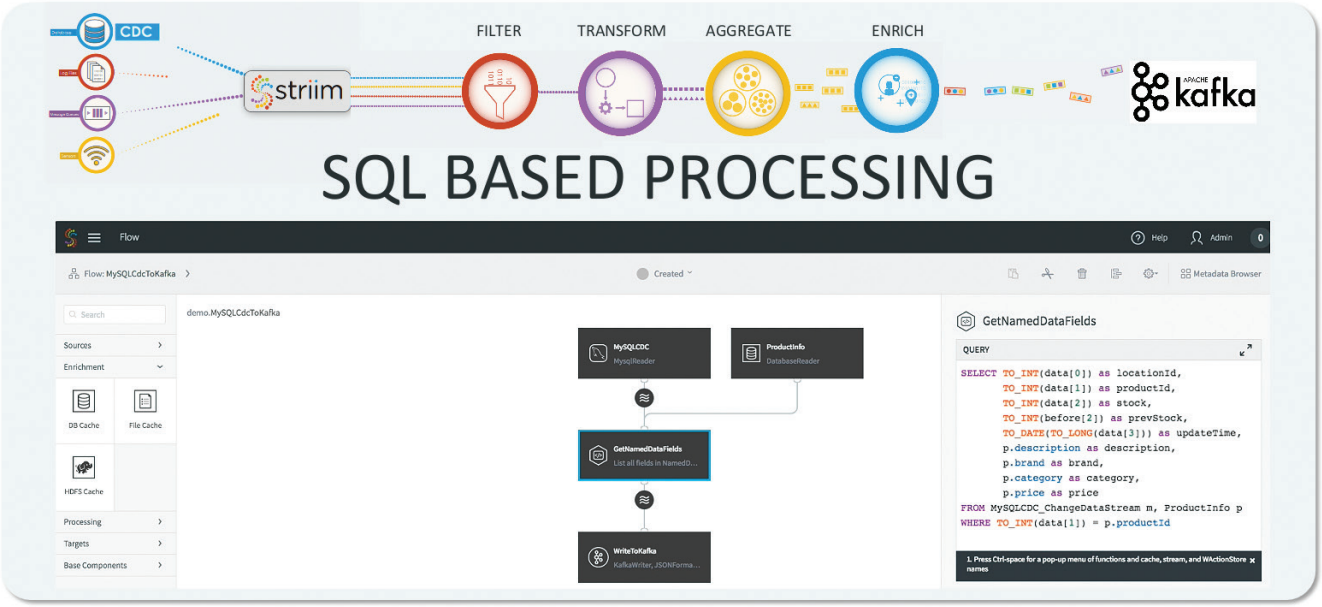


Writing data in real-time to multiple targets simultaneously

What's more, a single data flow can write to multiple targets at the same time in real time, with rules encoded as queries in between. This means companies can source data from Kafka and write some of it — or all of it — to HDFS, Azure SQL Database, and the enterprise data warehouse simultaneously.

Data Processing and Preparation for Kafka

When delivering data to Kafka, or writing Kafka data to a downstream target like HDFS, it is essential to consider the structure and content of the data that is being written. Based on the use case, organizations may not require all of the data, only that which matches certain criteria. Users may also need to transform the data through string manipulation or data conversion, or only send aggregates to prevent data overload.



In-stream data processing and preparation for Kafka

Most importantly, users may need to add additional context to the data. A lot of raw data may need to be joined with additional data to make it useful.

Imagine using CDC to stream changes from a normalized database. If the database has been designed properly, most of the data fields will be in the form of IDs. This is very efficient for the database, but not very useful for downstream queries or analytics. IoT data can present a similar situation, with device data consisting of a device ID and a few values, without any meaning or context. In both cases, users may want to enrich the raw data with reference data, correlated by the IDs, to produce a denormalized record with sufficient information.

The key tenets of streaming integration and stream processing — filtering, transformation, aggregation and enrichment — are essential to any data architecture, and should be easy to apply to Kafka data without any need for developers or complex APIs.

The key tenets of streaming integration and stream processing – filtering, transformation, aggregation and enrichment – are essential to any data architecture, and should be easy to apply to Kafka data.

The Striim platform simplifies this by using a uniform approach utilizing in-memory continuous queries, with all of the stream processing expressed in a SQL-like language. Anyone with any data background understands SQL, so the constructs are incredibly familiar. Transformations are simple and can utilize both built-in and Java functions, CASE statements and other mechanisms. Filtering is just a WHERE clause.

Aggregations can utilize flexible windows that turn unbounded infinite data streams into continuously changing bounded sets of data. The queries can reference these windows and output data continuously as the windows change. This means a one-minute moving average is just an average function over a one-minute sliding window.

Enrichment requires external data, which is introduced into the Striim platform through the use of distributed caches (otherwise known as a Data Grid). Caches can be loaded with large amounts of reference data, which is stored in-memory across the cluster. Queries can reference caches in a FROM clause the same way as they reference streams or windows, so joining against a cache is simply a JOIN in a query.

Multiple stream sources, windows and caches can be used and combined together in a single query, and queries can be chained together in directed graphs, known as data flows. All of this can be built through the UI or Striim's scripting language, and can be easily deployed and scaled across a Striim cluster, without having to write any code.

Streaming Analytics and Data Visualization for Kafka

To perform analytics on the streaming data running in Kafka, delivering the data to Hadoop or a database for analytics is not effective, as it introduces latency and diminishes the value users get in time-sensitive use cases. Users need to do the analytics in-memory, as the data is flowing through, and be able to surface the results of the analytics through visualizations in a dashboard.

Analytics can involve correlation of data across data streams, looking for patterns or anomalies, making predictions, understanding behavior, or simply visualizing data in a way that makes it interactive and interrogable.

The Striim platform enables organizations to perform analytics in-memory, in the same way as they do processing — through SQL-based continuous queries. These queries can join data streams together to perform correlation, and look for patterns (or specific sequences of events over time) across one or more data streams utilizing an extensive pattern-matching syntax.

Continuous statistical functions and conditional logic enable anomaly detection, while built-in regression algorithms enable predictions into the future based on current events.

Surface the results of the analytics through visualizations in a dashboard

Of course, analytics can also be rooted in understanding large datasets. Striim customers have integrated machine learning into data flows to perform real-time inference and scoring based on existing models. This utilizes Striim in two ways:

Firstly, as mentioned previously, users can prepare and deliver data from Kafka (and other sources) into storage in their desired format. This enables the real-time population of raw data used to generate machine learning models.

Secondly, once a model has been constructed and exported, users can easily call the model from SQL, passing real-time data into it, to infer outcomes continuously. The end result is a model that can be frequently updated from current data, and a real-time data flow that matches new data to the model, spots anomalies or unusual behavior, and enables proactive responses.

The final piece of analytics is visualizing and interacting with data. The Striim platform UI includes a complete Dashboard Builder that enables rapid development of custom, use-case-specific dashboards and effectively highlights real-time data and the results of analytics. With a rich set of visualizations, and simple query-based integration with analytics results, dashboards can be configured to continually update and enable drill-down and in-page filtering.

Via the interactive, live dashboards, Kafka users can compare live data to historical averages or to a specific date and time in the past, without having to write code. Users can view live data with detailed field and time-based filtering at the page or chart level. In addition, users can search streaming

With a rich set of visualizations, and simple query-based integration with analytics results, dashboards can be configured to continually update and enable drill-down and in-page filtering.

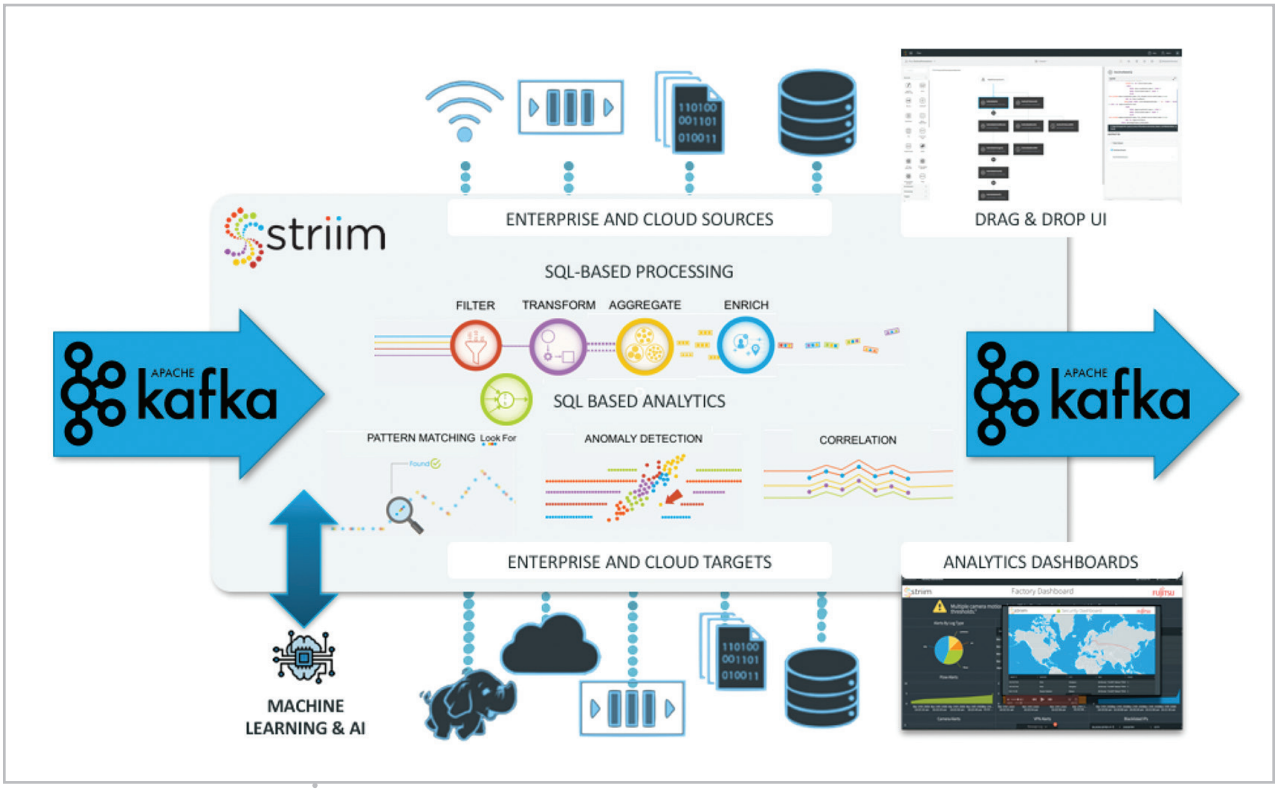
The Striim platform enables Data Scientists, Business Analysts and other IT and data professionals to get the most value out of Kafka without having to learn, and code to, APIs.

data directly on the dashboard and drill down to detail pages. Striim's charts can be embedded to any custom dashboard or web page to support broad collaboration and distribution of real-time insights.

Conclusion

Building a Platform that makes the most of Kafka by enabling true stream processing and analytics is not easy. There are multiple major pieces of in-memory technology that have to be integrated seamlessly and tuned in order to be enterprise grade. This means organizations have to consider the scalability, reliability and security of the complete end-to-end architecture, not just a single piece.

Joining streaming data with data cached in an in-memory data grid, for example, requires careful architectural consideration to ensure all pieces run in the same memory space, and joins can be performed without expensive and time-consuming remote calls. Continually processing and analyzing hundreds of thousands, or millions, of events per second across a cluster in a reliable fashion is not a simple task, and can take many years of development time.



The Striim platform has been architected from the ground up to scale, and Striim clusters are inherently reliable with failover, recovery and exactly-once processing guaranteed end-to-end, not just in one slice of the architecture.

Security is also treated holistically, with a single role-based security model protecting everything from individual data streams to complete end-user dashboards.

If organizations want to make the most of Kafka, they shouldn't have to architect and build a massive infrastructure, nor should they need an army of developers to craft their required processing and analytics. The Striim platform enables Data Scientists, Business Analysts and other IT and data professionals to get the most value out of Kafka without having to learn, and code to, APIs.





striim

Connect with us:

www.striim.com/blog/

www.linkedin.com/company/striim

www.facebook.com/striim

www.twitter.com/striimteam

www.striim.com/youtube

For more information, or to schedule a free trial, please contact us at **info@striim.com** or at **+1 (650) 241-0680**

www.striim.com