WHITE PAPER

DataStax Enterprise and Apache Kafka™ for Modern Architectures

DATASTAX®

CONTENTS

INTRODUCTION

The modern digital landscape is peppered with disparate technologies, each designed for a specific role within the ecosystem.

These systems are often scattered across a variety of clouds and the data is pouring in with unprecedented velocity. This rapidly evolving environment comes with new business requirements that force enterprises to deliver an always-on, personalized experience with millisecond speed to their customers. The challenges are immense, and in order to achieve success in today's world companies must transform their technology stack with the innovative solutions that are best suited for these conditions.

These circumstances gave birth to platforms such as DataStax Enterprise (DSE) and Apache Kafka, which are designed specifically to fit the needs of modern, next-generation businesses. With DSE providing the blazing fast, highly available hybrid cloud data layer and Apache Kafka detangling the web of complex architectures by way of its distributed streaming attributes, these two form a perfect match for event-driven enterprise architectures.

This white paper details how these products solve the problems cited above and dives into the DataStax Apache Kafka® Connector, which brings these technologies together to form a data ecosystem fit for the future.

WHAT IS DATASTAX ENTERPRISE?

For an enterprise to successfully engage in the modern business environment, enterprises must start with a foundation that allows them to create real-time applications at massive scale in order to exceed expectations through consumer and enterprise applications that provide responsive and meaningful engagement of each customer wherever they go.

DSE is the always-on, active everywhere database built on Apache Cassandra™ and designed for hybrid cloud. DSE also gives businesses full data autonomy, allowing them to retain control and strategic ownership of their most valuable asset in a hybrid cloud world.

DSE provides advanced security features to protect sensitive data, management services that automatically perform key maintenance and tuning functions, advanced visual management and administration capabilities, and world-class around-the-clock support.

CARDS

DataStax identifies the new requirements for cloud applications with the CARDS acronym:

C. Contextual service to provide the ability to instantly serve relevant information and experiences with every interaction no matter the data format.

A. Always on: Not simply "high availability" but continuous availability without complexity or the cost of multiple systems for replication and failover.

R. Real time: Manage data "in the moment," no matter the workload demands or location.

D. Distributed: Achieve the highest workload requirements no matter the location, with the ability to span data centers, cloud regions, and hybrid cloud.

S. Scalable: Scale linearly and predictably. Scale out with commodity hardware or efficiently scale up with compute-intensive hardware.

Figure 1 – CARDS: The baseline requirements for successful hybrid and multi-cloud applications.

Enterprises demand immediately actionable insights with always-available data to drive enterprise applications. Applications demand the highest level of responsiveness, no matter the surge in user engagement. Lower latency means even more pleasant end-user experience. Higher throughput means the ability to handle more traffic. These applications must also be operationally simple to manage so that teams can focus on innovation and customer success.

DSE also solves the cloud and IoT application "mixed workload" problem by smartly integrating analytics and search functionality into a unified platform with multi-model capabilities. This makes it the perfect complement to Apache Kafka, which features a storage structure that fits a time-series model. DSE eliminates the requirement for multiple data management providers and application sharding, further reducing the volume of separate components in the modern data landscape.

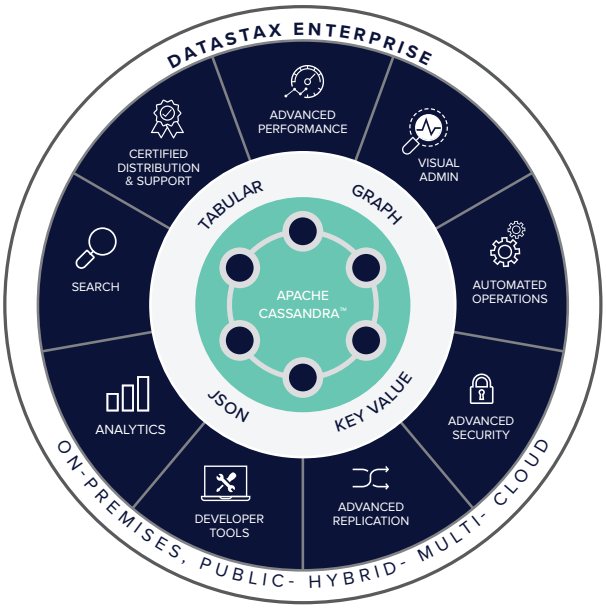

Figure 2 – DataStax Enterprise Components

WHAT IS APACHE KAFKA?

Apache Kafka is a distributed streaming platform and messaging queue that stores data in a log-based fashion, allowing the system to react to events as they enter the system.

This mechanism enables simplification of tangled architectures and provides the ability for lightweight transformation and filtering on the streams. Because the data is stored as a series of records, this plays nicely into event-driven use cases that span across industries.

Kafka as a Streaming Platform

The stream processing component of Kafka manifests via the Kafka Streams API. This paradigm is restricted to Kafka as the data source and store, where incoming data from Kafka can be manipulated and written back to Kafka. This is designed for low-touch alterations and aggregations of multiple streams and is typically used to enhance or join data before sending it off to another system. For more heavyweight analysis of large data sets, an advanced solution such as Spark, Hadoop, or DSE Analytics may be required.

Kafka as a Message Bus

Kafka caters to a modern enterprise landscape that incorporates several disparate technologies or individual microservices. Companies may write to Kafka first and then send the data off to other systems that serve their own purposes. In this sense, Kafka acts as the central ETL layer in conjunction with a group of publishers and consumers. These publishers and consumers can be custom applications that use the Kafka Client APIs or pre-built source/sink connectors that are deployed in Kafka Connect.

As an example, within a single architecture, a portion of the data may be sent from Kafka to DSE as the operational data layer, to S3 for long-term storage, and to Snowflake for the analytical data warehouse. DSE further simplifies this architecture by offering transactional, search, analytics, and graph workloads all in a single, masterless, highly available database.

WHAT IS THE DATASTAX APACHE KAFKA CONNECTOR?

The DataStax Apache Kafka Connector brings the strengths of Apache Kafka to the DataStax ecosystem by allowing data to seamlessly flow from Kafka to DSE. This allows users to simplify their complex architectures with Kafka while serving mission-critical applications with DSE in any cloud.

Built by the team that authors the DataStax Drivers for Apache Cassandra, the DataStax Apache Kafka Connector capitalizes on the best practices of ingesting to DSE while delivering enterprise-grade resiliency and security. This connector is the bridge that moves Apache Kafka records automatically to DSE without any need for a custom solution or a DSE Analytics deployment. Known in the Kafka Connect framework as a sink, the key features of this connector are its market-leading performance, flexibility, security, and visibility. All of this is offered with DataStax Basic, DSE, and DataStax Distribution of Apache Cassandra subscriptions at no additional cost.

Performance

As mentioned, the DataStax Apache Kafka Connector is engineered by the experts that develop and maintain Apache Cassandra's drivers. Without going into the weeds, the same techniques used in the DataStax Bulk Loader that proved to outperform all other bulk loading solutions for Cassandra are also leveraged in the connector.

Flexibility

The design of our Kafka sink considers the varying data structures found in Apache Kafka. Also, the selective mapping functionality in the connector allows the user to specify which Kafka fields should be written to DSE columns. This allows a single connector instance to read from multiple Apache Kafka topics and write to many DSE tables, thereby removing the burden of managing several connector instances. Whether the Apache Kafka data is Avro, JSON, or string format, the DataStax Apache Kafka Connector extends advanced parsing to account for the wide range of data inputs.

Security

One of the core value additions of DSE is DSE Advanced Security. With built-in SSL, LDAP/Active Directory, and Kerberos integration, DSE contains the tools needed to achieve strict compliance regulations for the connection from the client to the server. These security features are also included in the DataStax Apache Kafka Connector, ensuring that the connection between the connector and the data store is secure.

Visibility

In complex distributed environments, things are bound to hit points of failure. The engineering team at DataStax took special care to account for these error scenarios and all of the intelligence of the DataStax Drivers is applied in the DataStax Apache Kafka Connector. Additionally, there are metrics included that give the operator visibility into the failure rate and latency indicators as the messages pass from Kafka to DSE.

Supported Versions

The DataStax Apache Kafka Connector works with the following Kafka versions:

- **Apache Kafka 0.10.2 and higher**
- **Confluent 3.2 and higher**

Stream data using the connector to DSE database version 5.0 and higher.

A Closer Look at the Connector

FEATURES	DATASTAX	DESCRIPTION
Fully supported by DataStax	✓	DataStax fully supports and provides expert services for the connector.
Consume Kafka Primitive data format	✓	Connector accepts Kafka record data that is in primitive type form.
Consume Kafka JSON data format	✓	Connector accepts Kafka record data that is valid JSON form.
Consume Kafka Avro data format	✓	Connector accepts Kafka record data that is valid Avro form.
Pluggable Connect converters	✓	Connector works with StringConverter, JsonConverter, AvroConverter, ByteArrayConverter, and Numeric Converters, as well as custom data converters. Note that the producer of the data must use the same Converter as the connector.
Provides JMX metrics	✓	Connector exposes JMX metrics for record/failure count and latency recordings.
Runs within Connect Worker	✓	Connector is deployed in the Kafka Connect framework.
At least once guarantee	✓	Connector stores the offset in Kafka and will pick up where it left off if restarted. This minimizes the additional work but there are situations where writes to DSE will be retried if many records are in a single failed batch. The connector ensures that no records are missed.
Standalone mode support	✓	Connector is deployed in Kafka Connect framework and works in standalone mode (meant for dev/test).
Distributed mode / HA support	✓	Connector is deployed in Kafka Connect framework and works in distributed mode (meant for production).
Flexible Kafka topic => DSE table mapping	✓	Connector extends flexible mapping functionality to control the specific fields that are pulled from Kafka and written to DSE.
Single Kafka topic => multiple DSE tables	✓	Connector enables common denormalization patterns for DSE by allowing a single topic to be written to many DSE tables.
Connector throttling + parallelism	✓	Connector has built-in throttling to limit the max concurrent requests that can be sent by a single connector instance. Parallelism is delivered through the integration with the Kafka Connect distributed framework and asynchronous connector internals.
Flexible date/time/timestamp formats	✓	Connector accounts for the case that typically separate teams write to the same Kafka deployment and may use varying formats for date/time fields.
Configurable consistency level	✓	Connector allows configuring DSE consistency level on a per topic-table basis.

FEATURES	DATASTAX	DESCRIPTION
Row-level TTL	✓	Connector allows configuring DSE row-level TTL on a per topic-table basis.
Deletes	✓	Connector allows configuring DSE deletes on a per topic-table basis.
Handling of nulls	✓	Connector allows configuring DSE null handling on a per topic-table basis.
Error handling	✓	Connector has built-in error handling for various failure scenarios. These scenarios include bad mappings and DSE write issues.
Offset management	✓	Connector leverages the Kafka Connect framework to manage offsets by storing the offset in Kafka.
Connector => DSE SSL	✓	Connector allows configuring connection to DSE with SSL.
Connector => DSE username/password	✓	Connector allows configuring connection to DSE with username/password.
Connector => DSE LDAP/Active Directory	✓	Connector allows configuring connection to DSE with LDAP/Active Directory.
Connector => DSE Kerberos	✓	Connector allows configuring connection to DSE with Kerberos.
Configurable DSE write timeout	✓	Connector allows configuring write timeout to DSE.
Connector => DSE compression	✓	Connector allows configuring connection to DSE with compression strategies.

CONCLUSION

It is important in today's age of technological disruption to choose solutions that give your business an edge both now and in the future.

DSE and Apache Kafka provide powerful, data-centric innovation that grants enterprises the means to build personalized digital experiences for their customers. The DataStax Apache Kafka Connector brings these worlds into one and makes this stack the clear choice for the modern architecture.

Download the DataStax Apache Kafka Connector today or read the official connector documentation for more information.

ABOUT DATASTAX

DataStax delivers the only active everywhere hybrid cloud database built on Apache Cassandra™: DataStax Enterprise and DataStax Distribution of Apache Cassandra, a production-certified, 100% open source compatible distribution of Cassandra with expert support. The foundation for contextual, always-on, real-time, distributed applications at scale, DataStax makes it easy for enterprises to seamlessly build and deploy modern applications in hybrid cloud. DataStax also offers DataStax Managed Services, a fully managed, white-glove service with guaranteed uptime, end-to-end security, and 24x7x365 lights-out management provided by experts at handling enterprise applications at cloud scale. More than 400 of the world's leading brands like Capital One, Cisco, Comcast, Delta Airlines, eBay, Macy's, McDonald's, Safeway, Sony, and Walmart use DataStax to build modern applications that can work across any cloud. For more information, visit www.DataStax.com and follow us on Twitter @DataStax.

DATASTAX®