WHITE PAPER

Moderne Architekturen auf Basis von DataStax Enterprise und Apache Kafka™

DATASTAX®

INHALT

EINFÜHRUNG

Die moderne digitale Infrastruktur ist geprägt von spezialisierten Technologien, die ganz bestimmte Aufgaben übernehmen. Diese Systeme liegen oft in verschiedenen Clouds und müssen enorme Datenmengen verarbeiten. Immer neue Anforderungen strömen auf die Unternehmen ein, die sich in immer kürzeren Zyklen anpassen müssen.

Kunden erwarten durchgängige Erreichbarkeit, personalisierte Angebote und Interaktionen im Millisekundentakt. Die Aufgabe ist gewaltig. Unternehmen, die heute auf dem Markt bestehen möchten, müssen ihre Technologien mit den besten innovativenLösungen transformieren, die sie finden können. Aus dieser Aufgabenstellung heraus sind Plattformen wie DataStax Enterprise und Apache Kafka entstanden, die speziell für die Anforderungen moderner Next-Generation-Unternehmen entwickelt wurden. Beide ergänzen sich perfekt und bringen alles mit, worauf es bei ereignisgetriebenen Unternehmensarchitekturen ankommt. DataStax Enterprise (DSE) stellt einen schnellen, hochverfügbaren Hybrid-Cloud-Datenlayer bereit, während Apache Kafka™ komplexe Architekturen durch verteiltes Streaming vereinfacht. In diesem White Paper erfahren Sie, wie DataStax Enterprise und Apache Kafka die oben beschriebenen Probleme lösen. Darüber hinaus werfen wir einen Blick auf den DataStax Apache Kafka® Connector, der diese Technologien zu einer zukunftsfähigen Dateninfrastruktur zusammenfügt.

WAS IST DATASTAX ENTERPRISE?

Unternehmen, die in der digitalen Welt bestehen und die Erwartungen ihrer Kunden langfristig übertreffen möchten, benötigen ein starkes Fundament, um massiv skalierbare Echtzeit-Anwendungen zu erstellen, die ein schnelles und sinnvolles Kundenengagement ermöglichen – egal wo der Kontakt stattfindet.

DataStax Enterprise (DSE) ist eine Always-on-Active-Everywhere-Datenbank, die auf Apache Cassandra basiert und für die Hybrid Cloud konzipiert wurde. Mit DSE behalten Unternehmen die strategische Kontrolle über ihre Daten und bleiben so selbst in einer Hybrid-Cloud-Welt Eigentümer ihrer wertvollsten Ressource.

DSE hat viele Vorteile: leistungsfähige Sicherheit zum Schutz vertraulicher Daten, Verwaltungsdienste zur automatischen Wartung und Optimierung, visuelle Verwaltungs- und Administrationsfunktionen sowie erstklassiger Support rund um die Uhr.

CARDS

Die neuen Anforderungen für Cloudanwendungen fasst DataStax in dem Kürzel CARDS zusammen:

C (Contextual) – Kontextbezogenheit: Bei jeder Interaktion werden unabhängig vom Datenformat relevante Informationen und Dienste bereitgestellt.

A (Always on) – Immer verfügbar: Nicht einfach nur Hochverfügbarkeit, sondern dauerhafte Verfügbarkeit abseits der Komplexität und Kosten, die typischerweise mit einer großen Zahl von Replikations- und Failover-Systemen einhergehen.

R (Real time) – Echtzeit: Datenverwaltung in Echtzeit, unabhängig von Umfang oder Speicherort.

D (Distributed) – Verteilung: Erledigen Sie die größten Workloads über verschiedene Rechenzentren, Cloudregionen und Hybrid-Cloud-Umgebungen hinweg.

S (Scalable) – Skalierbarkeit: Vorhersagbare, lineare Skalierbarkeit. Horizontale Skalierung mit handelsüblicher Hardware oder vertikale Skalierung mit Hochleistungshardware.

Abb. 1 – CARDS: Die grundlegenden Anforderungen für erfolgreiche Hybrid- und Multi-Cloud-Anwendungen.

Unternehmen benötigen sofort verwertbare Erkenntnisse mit aktuellen, ständig verfügbaren Daten für ihre Enterprise-Anwendungen. Anwendungen erfordern höchste Reaktionsfähigkeit, auch während Lastspitzen. Geringe Latenz ist gleichbedeutend mit einer ansprechenden Benutzererfahrung. Ein höherer Durchsatz bedeutet, dass mehr Datenverkehr verarbeitet werden kann. Gleichzeitig müssen sich die Anwendungen möglichst einfach verwalten lassen, damit sich die Mitarbeiter auf innovative Tätigkeiten und die Kundenzufriedenheit konzentrieren können.

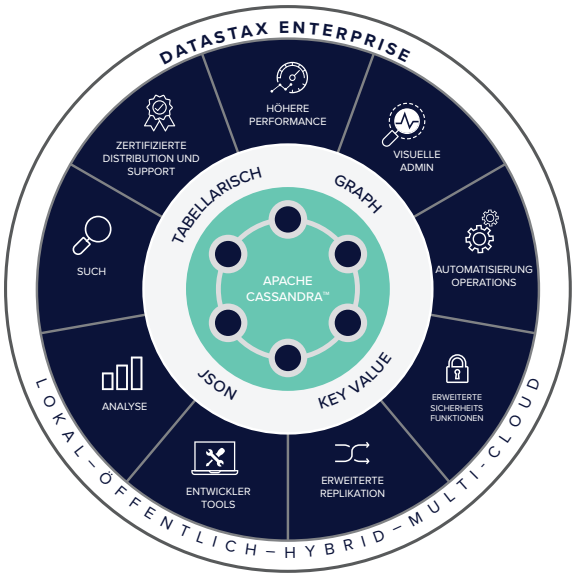

Abbildung 2 – Komponenten von DataStax Enterprise

DSE vereint Analyse- und Suchfunktionen in einer Unified-Plattform mit Multimodellfunktionalität und bietet so eine Antwort auf das „Mixed-Workload“-Problem von Cloud- und IoT-Anwendungen. Deshalb ist es auch die perfekte Ergänzung zu Apache Kafka, dessen Speicherstruktur zeitreihenbasierte Datenmodelle unterstützt. Mit DSE benötigt man weder mehrere Datenmanagement-Provider noch die Verwendung von Sharding, wodurch zahlreiche isolierte Komponenten entfallen und eine moderne, schlanke Datenumgebung möglich wird.

WAS IST APACHE KAFKA?

Apache Kafka ist eine verteilte Streaming-Plattform und Messaging Queue, die Daten in Protokollform speichert und einem System die unmittelbare Reaktion auf Ereignisse ermöglicht.

Der Mechanismus ermöglicht einerseits die Vereinfachung komplexer Architekturen, andererseits lassen sich die Datenströme ohne großen Aufwand umwandeln und filtern. Die Daten werden als Datensatzreihen gespeichert, was gut zu den ereignisbasierten Anwendungsfällen passt, die in allen Branchen vorzufinden sind.

Kafka als Streaming-Plattform

Die Kafka-Komponente, die für die Verarbeitung von Streams zuständig ist, wird über die Kafka Streams API angesprochen. Datenquelle und Datenspeicher sind in diesem Fall auf Kafka beschränkt. Von Kafka kommende Daten können verändert und wieder in Kafka gespeichert werden. Dieses Modell eignet sich zur effizienten Änderung und Bündelung mehrerer Datenströme und wird normalerweise verwendet, um Daten vor dem Transfer an ein anderes System aufzubereiten oder zusammenzuführen. Für aufwendigere Datenanalysen von großen Datasets wird möglicherweise eine leistungsfähigere Lösung wie Spark, Hadoop oder DSE Analytics benötigt.

Kafka als Message Bus

Kafka erfüllt die Anforderungen einer modernen Unternehmensinfrastruktur mit vielen verschiedenen Technologien und Mikrodiensten. Unternehmen können ihre Daten zum Beispiel zuerst in Kafka speichern, um sie von dort an andere Systeme weiterzuleiten. In diesem Fall agiert Kafka als zentraler ETL-Layer und dient als Schnittstelle zwischen Publisher und Verbrauchern. Diese können entweder Anwendungen sein, die von den Kafka Client APIs Gebrauch machen, oder vorgefertigte Quellen/Senk-Konnektoren, die in Kafka Connect implementiert sind. So kann innerhalb einer Architektur zum Beispiel ein Teil der Daten von Kafka an DSE gesendet werden, das als Datenlayer fungiert, ein Teil an S3 zur Langzeitspeicherung, und ein anderer Teil an Snowflake zur Analyse im Data Warehouse. DSE vereinfacht diese Architektur noch weiter, indem es Transaktions-,Such-, Analyse-, und Visualisierungs-Workloads in einer einzigen hochverfügbaren Masterless-Datenbank bereitstellt.

DER DATASTAX APACHE KAFKA CONNECTOR

Der DataStax Apache Kafka Connector sorgt für den reibungslosen Datenaustausch zwischen Kafka und DSE und macht so die Stärken von Apache Kafka in der DataStax-Welt nutzbar. Benutzer können mit Kafka komplexeArchitekturen vereinfachen und mit DSE in jeder Cloud erfolgskritische Anwendungen unterstützen.

Der DataStax Apache Kafka Connector wurde vom selben Team entwickelt, das auch die DataStax-Treiber für Apache Cassandra™ geschrieben hat. Der Connector verlässt sich bei der Datenaufnahme in DSE auf bewährte Best Practices und überzeugt durch eine hohe Fehlertoleranz und Sicherheit der Enterprise-Klasse. Als eine Art Brücke verschiebt er Datensätze automatisch von Apache Kafka nach DSE. Dafür wird weder eine speziell entwickelte Lösung noch DSE Analytics benötigt. Der Connector, der im Kafka Connect-Framework als Datensenke (Sink) dient, zeichnet sich durch hohe Leistung, Flexibilität, Sicherheit und Kontrolle aus. Er ist kostenloser Bestandteil von DataStax Enterprise und DDAC (DataStax Distribution of Apache Cassandra).

Performance

Wie erwähnt, stammt der DataStax Apache Kafka Connector von den Programmierern, die auch die Treiber von Apache Cassandra entwickelt haben. Dabei kamen dieselben Techniken zur Anwendung, die auch schon den DataStax Bulk Loader zu einer der schnellsten Bulk-Loading-Lösungen für Cassandra gemacht haben.

Flexibilität

Die von uns entwickelte Kafka-Datensenke berücksichtigt die verschiedenen Datenstrukturen, die in Apache Kafka vorkommen können. Benutzer können mithilfe einer selektiven Zuordnungsfunktion im Connector angeben, welche Kafka-Felder in DSE-Spalten geschrieben werden sollen. Eine einzelne Connector-Instanz kann Daten aus mehreren Apache Kafka-Themen lesen und in viele DSE-Tabellen gleichzeitig schreiben, sodass nicht mehrere Connector-Instanzen verwaltet werden müssen. Apache Kafka-Daten können in verschiedenen Formaten wie Avro, JSON oder als Strings vorliegen, weshalb der DataStax Apache Kafka Connector leistungsfähige Parsing-Funktionen zur Verarbeitung dieser Formate bietet.

Sicherheit

Eine der zentralen Neuerungen von DSE ist DSE Advanced Security. DSE unterstützt SSL, LDAP/Active Directory und Kerberos und erfüllt damit strengste Compliance-Anforderungen für Client/Server-Verbindungen. Der DataStax Apache Kafka Connector bietet dieselben Sicherheitsmerkmale und gewährleistet somit einen sicheren Datenaustausch zwischen Connector und Datenspeicher.

Kontrolle

In großen verteilten Umgebungen kommt es immer wieder zu Fehlern. Deshalb haben die Entwickler bei DataStax besondere Sorgfalt darauf verwendet, alle möglichen Fehlerszenarien abzudecken. Der DataStax Apache Kafka Connector wurde mit denselben intelligenten Funktionen ausgestattet, die auch die DataStax-Treiber auszeichnen. Verschiedene Metriken erfassen zudem Informationen wie die Fehlerquote und die Latenz bei der Übertragung von Nachrichten zwischen Kafka und DSE.

Unterstützte Versionen

Der DataStax Apache Kafka Connector funktioniert mit den folgenden Versionen von Kafka:

- **(Apache Kafka ab Version 0.10.2)**
- **Confluent ab Version 3.2**

Das Streamen von Daten über den Connector funktioniert mit der DSE-Datenbank ab Version 5.0.

DER CONNECTOR IM DETAIL

FEATURES	DATASTAX	BESCHREIBUNG
Von DataStax vollständig unterstützt	✓	DataStax bietet vollständige Unterstützung für den Connector sowie Expertendienstleistungen.
Verarbeitung des Primitive Datenformats von Kafka	✓	Der Connector verarbeitet Kafka-Datensätze, die im primitive Format vorliegen.
Verarbeitung des Kafka JSON-Datenformats	✓	Der Connector verarbeitet Kafka-Datensätze, die in einem gültigen JSON-Format vorliegen.
Verarbeitung des Kafka Avro-Datenformats	✓	Der Connector verarbeitet Kafka-Datensätze, die in einem gültigen Avro-Format vorliegen.
Plug-in-fähige Connect-Konverter	✓	Der Connector unterstützt StringConverter, JsonConverter, AvroConverter, ByteArrayConverter und Numeric Converters sowie benutzerdefinierte Datenkonverter. Die Daten müssen vom selben Konverter erzeugt werden, den auch der Connector verwendet.
Bereitstellung von JMX-Metriken	✓	Der Connector liefert JMX-Metriken zur Anzahl der Datensätze/Fehler und Informationen zur Latenz.
Funktioniert innerhalb des Connect Workers	✓	Der Connector wird im Kafka Connect-Framework bereitgestellt.
„At Least Once“-Garantie	✓	Der Connector speichert den Offset in Kafka und macht da weiter, wo er zuletzt aufgehört hat. Das vermeidet zusätzlichen Aufwand. Es kann jedoch auch vorkommen, dass Schreibversuche auf DSE wiederholt werden, wenn ein einzelner fehlgeschlagener Batch zu viele Datensätze enthält. Der Connector sorgt dafür, dass keine Datensätze verloren gehen.

FEATURES	DATASTAX	BESCHREIBUNG
Standalone-Modus	✓	Der Connector wird im Kafka Connect Framework bereitgestellt und arbeitet im Standalone-Modus (für Dev/Test).
Distributed-Modus/HA	✓	Der Connector wird im Kafka Connect Framework bereitgestellt und arbeitet im Distributed-Modus (für Produktionsumgebungen).
Flexibles Kafka-Topic => DSE-Tabellenzuordnung	✓	Der Connector nutzt die Funktion zur flexiblen Zuordnung, um einzelne Felder zu kontrollieren, die aus Kafka exportiert und in DSE importiert werden.
Einzelnes Kafka-Topic => mehrere DSE-Tabellen	✓	Einzelne Topics können in mehrere DSE-Tabellen geschrieben werden. Auf diese Weise ermöglicht der Connector die Denormalisierung in DSE.
Connector-Drosselung + Parallelisierung	✓	Der Connector kann die Anzahl der gleichzeitig möglichen Anfragen, die von einer einzelnen Connector-Instanz gesendet werden können, begrenzen. Die Parallelisierung ergibt sich aus der Integration mit dem Kafka Connect Distributed Framework und asynchronen Connector-Funktionen.
Flexible Datums-/Zeit-/Zeitstempel-Formate	✓	Der Connector unterstützt den Fall, dass verschiedene Teams in dieselbe Kafka-Bereitstellung schreiben und dabei abweichende Formate für die Datums-/Zeitfelder verwenden.
Konfigurierbarer Konsistenzgrad	✓	Mit dem Connector kann der DSE-Konsistenzgrad für einzelne Topic-Tabelle-Kombinationen konfiguriert werden.
TTL auf Zeilenebene	✓	Mit dem Connector kann die TTL auf Zeilenebene in DSE für einzelne Topic-Tabelle-Kombinationen konfiguriert werden.
Löschvorgänge	✓	Mit dem Connector können Löschvorgänge in DSE für einzelne Topic-Tabelle-Kombinationen konfiguriert werden.
Nullen-Verarbeitung	✓	Mit dem Connector kann die Verarbeitung von Null-Werten in DSE für einzelne Topic-Tabelle-Kombinationen konfiguriert werden.
Fehlerverarbeitung	✓	Die Fehlerverarbeitung des Connectors unterstützt verschiedene Fehlerszenarien, zum Beispiel falsche Zuordnungen und Probleme beim Schreiben in DSE.
Offset-Management	✓	Der Connector verwaltet Offsets mithilfe des Kafka Connect Frameworks und speichert diese in Kafka.
Connector => DSE SSL	✓	Der Connector unterstützt die SSL-Verbindung mit DSE.
Connector => DSE Benutzername/Kennwort	✓	DSE-Verbindungen lassen sich mit einem Benutzernamen/Kennwort schützen.
Connector => DSE LDAP/Active Directory	✓	Der Connector unterstützt DSE-Verbindungen über LDAP/Active Directory.
Connector => DSE Kerberos	✓	Der Connector unterstützt mit Kerberos gesicherte DSE-Verbindungen.
Konfigurierbares Timeout für DSE-Schreibvorgänge	✓	Für DSE-Schreibvorgänge kann ein Timeout-Wert angegeben werden.
Connector => DSE-Komprimierung	✓	Der Connector unterstützt komprimierte DSE-Verbindungen.

FAZIT

In einer Zeit des schnellen technischen Wandels kommt es darauf an, die Lösungen auszuwählen, die ein Unternehmen auf lange Sicht zukunftsfähig machen. DSE und Apache Kafka sind leistungsstarke, datenorientierte Lösungen, auf deren Basis Unternehmen Kunden hochpersonalisierte digitale Erfahrungen bereitstellen können. Der DataStax Apache Kafka Connector ist das Zwischenstück, das diese Welten verbindet. Gemeinsam bilden diese Technologien die Grundlage für moderne Architekturen.

Laden Sie den DataStax Apache Kafka Connector noch heute herunter oder lesen Sie für weitere Informationen die offizielle Dokumentation zum Connector.

ÜBER DATASTAX

DataStax bietet eine ständig verfügbare, verteilte Active Everywhere Database für die Hybrid Cloud auf Basis von Apache Cassandra™. DataStax Enterprise bildet die Basis für personalisierte, extrem skalierbare Echtzeitanwendungen und macht es Unternehmen leicht, Hybrid- und Multi-Cloud-Umgebungen über einen nahtlos integrierten Data Layer zu nutzen. Mit DataStax gehören Probleme, die normalerweise bei der Implementierung von Anwendungen über mehrere lokale Datencenter und/oder Public Clouds hinweg auftreten, endgültig der Vergangenheit an. Unternehmen behalten die Kontrolle über ihre Daten, profitieren von umfassender Datentransparenz und -übertragbarkeit und bleiben selbst in einer Hybrid-/Multi-Cloud-Welt Eigentümer ihrer wertvollsten Ressource. DataStax unterstützt mehr als 400 weltweit führende Marken aller Branchen bei der Transformation ihrer Geschäfte mithilfe eines Data Layers, der Silos und Vendor Lock-in beseitigt und den Einsatz bahnbrechender Applikationen vorantreibt. Für weitere Informationen besuchen Sie www.DataStax.com und folgen Sie uns auf Twitter @DataStax.

© 2019 DataStax, All Rights Reserved. DataStax, Titan, and TitanDB are registered trademarks of DataStax, Inc. and its subsidiaries in the United States and/or other countries.

Apache, Apache Cassandra, and Cassandra are either registered trademarks or trademarks of the Apache Software Foundation or its subsidiaries in Canada, the United States, and/or other countries.

DATASTAX®