Aufbau und Inbetriebnahme einer Big Data Infrastruktur bei der BEBK

Studiengang: MAS Data Science

Kann man Kafka als Datenbanksystem einsetzen? Kann es ein Datenbanksystem ersetzen? Kann man Kafka als zentralen Datenhub einer Unternehmung nutzen? Kafka ist doch ein Messagingsystem, das nicht mit einem Datenbanksystem verglichen werden kann. Diese Fragen waren Grund genug das Thema dieser Arbeit zu stellen.

Boris Bötzel
boris.boetzel@bluewin.ch

Die Arbeit verfolgt neben der strategischen Positionierung der Streaming Plattform zwei technische Ziele mit denen auf Basis bestehender Metadaten die aktuelle ETL-Verarbeitung auf der Streaming Plattform automatisiert konfiguriert und migriert werden kann.

Die technischen Voraussetzungen für die Konfiguration der Streaming Plattform waren sehr hoch. So musste eine nahezu produktionsreife Cluster-Umgebung mit zookeeper, kafka, schema registry und ksqlDB automatisiert aufgebaut werden, die zudem im Minimum sämtliche Netzverbindungen verschlüsselt. Die Applikationen wurden als application-container mittels ansible-Scripts in Betrieb genommen. Die grössten Herausforderungen bestanden in der Verschlüsselung der Daten. Vor allem die korrekte Konfiguration von TLS bei intra-cluster- wie auch bei client-server-Verbindungen konnte wegen unvollständiger Dokumentation nur in endlosen Testreihen erfolgreich in Betrieb genommen werden. Die starke Resilienz der Cluster der Streaming Plattform muss hier explizit hervorgehoben werden. Der in der frühen Phase der Arbeit einmal angebotene Service eines frisch gestarteten Clusters wurde dank nutzbaren rolling-upgrade-Verfahren trotz etlichen Anpassungen der Konfiguration wie auch Software-Upgrades nie wieder unterbrochen. „Hut ab!“

Mit den Metadaten über die ETL-Verarbeitung werden im Rahmen dieser Arbeit die ersten Data Pipelines der Streaming Plattform generiert und in Betrieb genommen. Es wurden generierte Verarbeitungsschritte direkt mit Kafka-Topics sowie auch Verarbeitungsschritte mit ksqlDB-Streams und ksqlDB-Tables erfolgreich getestet. Einfache Extract, Transform Load Verarbeitungsschritte konnten relativ einfach implementiert werden. Bei der Implementierung von sog. Slow Changing Dimensions (SCD) vom Typ 2 nach Kimball konnte eine erste Version, die ausschliesslich auf ksqlDB-Streams basiert nicht erfolgreich umgesetzt werden. Für eine solche Data Pipeline braucht es Status-Informationen, die nachgeführt werden können. Das kann mit Streams nicht abgebbildet werden. Die zweite Version inkl. ksql-Tables konnte nahezu erfolgreich realisiert werden.

Das „nahezu“ hat einen bestimmten Grund. Klassische ETL-Verarbeitungen setzen auf Datenstände, die als Batch verarbeitet werden. Der folgende Verarbeitungsschritt wird erst gestartet, wenn der vorhergehende fertig ist. Obwohl dies bei einer Streaming Plattform genau gleich implementiert werden könnte, ist doch der Sinn der Übung in dieser Arbeit ETL-Verarbeitungsschritte zu „streamen“ anstatt diese als „batch“ abzuarbeiten. Die SCD Typ 2 zu „streamen“ ist nach den Erkenntnissen im Rahmen dieser Arbeit durchaus möglich, konnte jedoch noch nicht abschliessend implementiert und getestet werden.

Kommen wir zu den eingangs gestellten Fragen zurück. Kann man Kafka als Datenbanksystem einsetzen? Ja, man kann es sehr wohl als Datenbanksystem einsetzen. Kann es ein Datenbanksystem ersetzen? Ja, Kafka und ksqlDB könnten ein Datenbanksystem ersetzen. Und nein, es macht nicht in allen Fällen Sinn ein Use Case mit Kafka und ksqlDB zu realisieren. Es gibt Datenbanksysteme, die in einigen Fällen wesentlich adäquater sind, sei es nur schon wegen vielen sehr ausgeklügelten SQL-Funktionalitäten, die ein ksqlDB nicht bietet.

Kann man Kafka als zentralen Datenhub einer Unternehmung nutzen? Das ist nun eine zentrale Frage, die sich die BEKB intensiv stellen muss. Ein zentraler Datenhub dient als Datenaustauschplattform für alle Applikationen der Unternehmung. Der Datenhub kann nicht durch ein klassisches Datenbanksystem implementiert werden, das Daten in seiner als Grundfunktionalität speichert und diese nur in Form von Batches verarbeiten kann. Der Datenhub muss als Streaming Plattform implementiert werden, die Daten sowohl near-realtime zwischen Applikationen streamen kann, aber auch in anderen Anwendungsfällen speichern kann.