EVALUIERUNG VON EVENTBASIERTEN ECHTZEITSYSTEMEN

Alexander Christoph B.Sc

Josef Witt GmbH
CB1-CIO-AIN

Schillerstraße 4-12
92637 Weiden
alexander.christoph@witt-gruppe.eu

Sven Hornberg M.Sc

Josef Witt GmbH
CB1-CIO-AIN

Schillerstraße 4-12
92637 Weiden
sven.hornberg@witt-gruppe.eu

Professor Dr.
Frank Herrmann

OTH Regensburg
Innovationszentrum für Produktionslogistik und Fabrikplanung
Seybothstraße 2
93053 Regensburg
frank.herrmann@oth-regensburg.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Echtzeit Datenverarbeitung, Apache Flink, Apache Spark

Zusammenfassung

In Zeiten von Big Data und einer immer mehr vernetzten Infrastruktur gewinnt die Verarbeitung von Echtzeit Daten zunehmend an Bedeutung. Daten, die in dem Moment der Generierung noch aktuell sind, können innerhalb von wenigen Sekunden schon veraltet und unbrauchbar sein. Dieser Herausforderung muss sich auch die Josef Witt GmbH stellen, um konkurrenzfähig zu bleiben und den Kunden ein besseres Einkaufserlebnis zu bieten. Ein Beispiel für eine solche Echtzeit Datenverarbeitung ist die Analyse von Benutzeraktionen in dem Online Shop. Mithilfe dieser können Produktvorschläge noch individueller auf den jeweiligen Nutzer angepasst werden.

Um Echtzeit Daten zu transferieren und zu verarbeiten gibt es unterschiedliche Lösungsansätze. Insbesondere gibt es viele Stream-Processing Engines, die Programmierschnittstellen zur Verfügung stellen, welche die Verarbeitung von solchen kontinuierlichen Daten vereinfachen.

Diese Stream-Processing-Engines unterscheiden sich jedoch teilweise stark in ihrer Funktion und Performance. Außerdem bietet jede dieser Laufzeitbibliotheken eigene Programmierschnittstellen. Dies erschwert es den Entwicklern unterschiedliche Bibliotheken auf ihre Eignung für die jeweils zu lösende Aufgabe zu testen.

In dieser Arbeit wurde Apache BEAM entwickelt. BEAM bietet eine einheitliche Programmierschnittstelle für die Entwicklung von Stream-Verarbeitungs-Pipelines. Dies ermöglicht es entwickelte Pipelines mit angepasster Konfiguration auf unterschiedlichen Stream-Processing-Engines bereitzustellen.

Im Einzelnen wurde ein generischer Testablauf er schaffen, mit dem die Performance von Stream-Processing-Engines getestet werden kann. Anschließend wurde dieser auf Apache Spark und Apache Flink angewandt, um ihre Tauglichkeit zur Echtzeit Datenverarbeitung zu prüfen. Für Spark und Flink wurde sich aufgrund der hohen Verbreitung der beiden Frameworks, sowie deren Unterschied in der Stream Verarbeitung entschieden.

Als Erfolgskriterien wurde die folgenden Metriken verwendete: der durchschnittliche Datendurchsatz pro Sekunde (Verarbeitete Events pro Sekunde) und die Latenz (Dauer für die Verarbeitung von einzelnen Events).

Um einen kontinuierlichen Datenstrom zu simulieren wurde Apache Kafka als Messaging Broker verwendet. Kafka sendet die einzelnen Spalten eines Testdatensatzes[1] als kontinuierlichen Datenstrom an die Stream-Processing Engines. Diese verarbeiten die Daten und senden sie zurück an Kafka.

Um die Metriken zu berechnen, wurde der von Kafka generierte Zeitstempel für die Erstellung jeder Nachricht verwendet.

Die mit BEAM erstellten Pipelines sind folgende:

- Einfaches weiterleiten der Events.
- Filtern von ca. 45% aller Einträge des ursprünglichen Datensatzes.
- Hinzufügen einer neuen Spalte zu dem Event und weiterleiten des Events mit der neuen Spalte.

Die Ergebnisse der Arbeit belegen, dass der Performance Test zur Evaluierung unterschiedlicher Stream-Processing-Engines geeignet ist.

Bei dem Vergleich von Spark und Flink hat sich herausgestellt, dass der Datendurchsatz pro Sekunde für Spark im Durchschnitt höher war (teilw. +25,3%). Die Latenzen für die Verarbeitung von einzelnen Nachrichten waren jedoch bei Flink besser.

Hieraus lässt sich ableiten, dass Apache Flink für die Verarbeitung von zeitkritischen Echtzeitdaten eine bessere Wahl ist.

Wenn jedoch der Datendurchsatz im Fokus steht, sollte die Verwendung von Apache Spark ebenfalls in Betracht gezogen werden.

[1] https://www.kaggle.com/mkechinov/ecommerceevents-history-in-cosmetics-shop von: https://rees46.com/