Data Ingestion and Enrichment Pipeline

Studiengang: MAS Information Technology
Betreuer: Christian Sprecher
Expertin: Ursula Deriu (Tirsus GmbH)

Die «Data Ingestion and Enrichment Pipeline» ermöglicht es, Daten von Quellen in die Suchmaschine Elasticsearch zu integrieren. Die Pipeline wurde so aufgebaut, dass sie mittels einer zentralen Konfiguration spezifische Komponenten einbinden kann. So kann die Pipeline durch das Hinzufügen von spezifischen Implementationen der Komponenten an verschiedene Use Cases angepasst werden.

Problemstellung

Moderne Suchmaschinen bieten nicht nur die klassische Suchfunktionalität, sondern auch erweiterte Funktionen zur Filterung und Analyse des Datenbestandes. Um gezielte Informationen aus dem Datenbestand zu erhalten, müssen die Daten nicht nur von der Quelle bezogen, sondern auch spezifisch aufbereitet werden.

Ziel

Die Pipeline soll durch Konfiguration von spezifischen Schritten die Daten Use-Case-spezifisch aufbereiten können. Dabei sollen Veredelungsschritte nicht nur per Konfiguration, sondern auch automatisch von der Pipeline ausgewählt werden, wenn diese auf die Daten anwendbar sind. Der Anpassungsaufwand an einen Use Case soll möglichst tief gehalten werden. Die Pipeline muss zudem wechselnden Anforderungen bezüglich Datenmenge Rechnung tragen. Beispielsweise muss initial ein grosser Datenbestand verarbeitet werden können, obwohl im normalen Betrieb das Datenvolumen sehr klein ist.

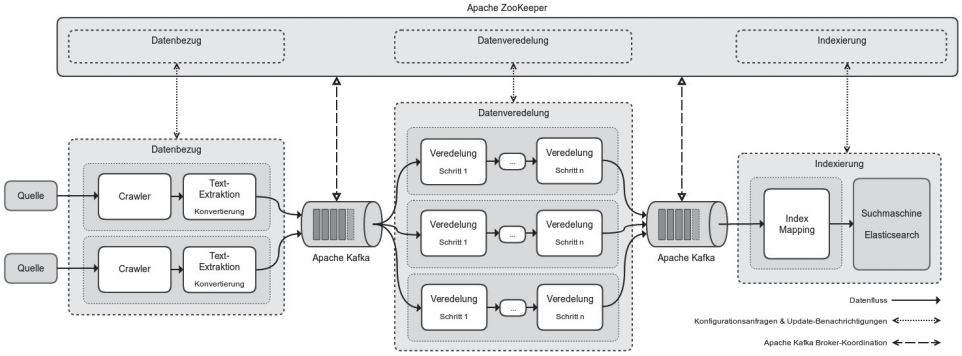
Lösung

Die Pipeline wurde in die drei Phasen «Datenbezug», «Datenveredelung» und «Indexierung» aufgeteilt. Die Architektur der Lösung sorgt für eine Entkopplung der einzelnen Phasen. Die Komplexität wird dadurch reduziert und die einzelne Phase ist unabhängig von den anderen. Die Zwischenresultate der Phasen werden in Apache Kafka Topics für die nächste Phase gespeichert.
Nach dem Bezug der Daten werden diese in der zweiten Phase durch die «Datenveredelung» aufbereitet und ergänzt. Dies geschieht anhand einzelner konfigurierter oder dynamisch hinzugefügter Veredelungsschritte. Die letzte Phase sorgt für die Integration der Daten in Elasticsearch. Ebenfalls wird sichergestellt, dass die in den Veredelungsschritten definierten Datentypen in Elasticsearch korrekt interpretiert werden.
Für die zentrale Konfiguration wird Apache ZooKeeper verwendet. Jede Komponente besitzt einen eigenen Bereich für ihre Konfiguration und kann bei Konfigurationsänderungen durch ZooKeeper benachrichtigt werden.
In allen Phasen wird ein neutrales Datenmodell verwendet. Dies sorgt dafür, dass einzelne Veredelungsschritte beliebig angeordnet und die Daten auch für andere Dienste aus den Kafka Topics abgegriffen werden können.

Resultat

Die ausgewählten Technologien und die strikte Trennung der Phasen haben sich bewährt. Durch das gewählte Design und das einheitliche Datenmodell ist eine Lösung entstanden, die den Anforderungen entspricht und einfach an einen Use Case angepasst werden kann.

Christof Lüthi



Schematische Darstellung der Pipeline