Übergreifende Suche mit dem ELK-Stack

Gunther Thielemann
SLIX Gesellschaft für Computersysteme mbH
Au i.d. Hallertau

Schlüsselworte
Elasticsearch, Logstash, Kibana, NOSQL, Suche, Enterprise Search, REST API, JSON

Einleitung

Kennen Sie das?

Die Information muss vorhanden sein, aber momentan können Sie sie nicht finden. Bezog sie sich auf einen Kunden, einen kürzlich gelesenen Artikel, eine Email, eine anderweitige Notiz? Jede Quelle einzeln zu durchsuchen kostet wertvolle Zeit, bleibt jedoch manchmal der einzige Weg. Eine Suche über alle verschiedenen Datentöpfe, die entweder das präzise Ergebnis liefert oder zumindest Hinweise zur Unterstützung weiterer Recherchen, könnte den Aufwand drastisch reduzieren.

Mit diesem Problem sehen sich viele Unternehmen konfrontiert. Es fallen immer mehr Daten an. Einige liegen strukturiert in relationalen Datenbanken vor. In der Mehrzahl handelt es sich jedoch um unstrukturierte Dokumente. Nach Schätzungen von Gartner beträgt der Anteil am Gesamtvolumen ca. 85%. Die Herausforderung besteht neben der Speicherung zunehmend darin, in den Datenmengen aus unterschiedlichen Quellen mit vertretbarem Aufwand präzise Informationen zu finden.

Gegenüber einer Reihe von etablierten kommerziellen Systemen zur unternehmensweiten Suche gibt es mittlerweile auch äußerst erfolgreiche Alternativen im Open Source Bereich. Eine davon ist Elasticsearch. Zusammen mit den Komponenten Logstash und Kibana hat die Firma Elastic ein optimal aufeinander abgestimmtes „Ökosystem“ geschaffen, welches anlehnend an die Kernkomponenten mit „ELK Stack“ bezeichnet wird. Anfangs für die performante, hochgradig skalierbare Sammlung, Indizierung und Analyse von Logfiles entwickelt, wird dieser Stack zunehmend auch für den Aufbau einer unternehmensweiten Suche verwendet.

ELK Stack

Den Kern bildet Elasticsearch, eine dokumentenorientierte NoSQL Datenbank. Intern werden die Inhalte in auf Apache Lucene basierenden invertierten Indices gespeichert. Gepaart mit der konsequenten Ausrichtung auf horizontale Skalierung und Ausfallsicherheit besitzt dieses System hervorragende Eigenschaften einer Suchmaschine. Die Software läuft auf vielen Plattformen, wie Linux, Windows und Mac OS. Vorausgesetzt wird nur ein Java JDK. Die Kommunikation erfolgt über eine REST- API. Für Ein- und Ausgaben wird JSON verwendet.

Kibana liefert die UI. Dieses Tool eignet sich für einfache Übersicht, bietet Funktionen zum Durchsuchen der Datenbestände bis zur Unterstützung bei der Entwicklung von komplexen Statements und beinhaltet umfangreiche Möglichkeiten zur Analyse und Visualisierung.

Mit Logstash steht ein flexibles, leistungsfähiges Werkzeug zum Extrahieren, Transformieren und Laden von Inhalten aus den unterschiedlichsten Quellen in verschiedene Zielsysteme, vor allem aber nicht nur Elasticsearch zur Verfügung. Die Eingabekanäle beschränken sich nicht nur auf Dateien, sondern erstrecken sich über ein Spektrum, zu dem unter anderem ein leistungsfähiges JDBC- Plug-in

gehört. Die Ladeprozesse bilden aus den Stufen Input, Filter und Output eine Pipeline. Plug-ins führen die einzelnen Operationen aus. Viele stellt der Hersteller zu Verfügung. Das Angebot wird die aktive

Community ergänzt. Da der Quellcode verfügbar und die Schnittstellen dokumentiert sind, lassen sich bei Bedarf auch eigene Plug-ins erstellen.

Das Portfolio von Elastic umfasst eine Reihe weiterer Produkte. Neben sogenannten Beats, leichtgewichtigen, hoch spezialisierten Datensammlern, werden mit dem „X Pack“ kommerzielle Erweiterungen angeboten. Dazu gehören Benutzerverwaltung, Authentifizierung, Monitoring, erweiterte Analyse und Visualisierung bis hinein in den Bereich der künstlichen Intelligenz.

Erste Schritte

Elastic bietet einen dem Entwickler einen leichten Einstieg. Die Software lässt sich einfach installieren, entweder durch Auspacken eines Zip Archives, die Installation mit dem Packetmanager oder via Puppet. Alternativ stehen Docker Images oder die Cloud zur Verfügung.

Mit den vorhandenen Voreinstellungen lassen sich Indizes erstellen. Datentypen werden weitestgehend selbständig erkannt. Der Anfang gelingt auch ohne ein vordefiniertes Schema. Diese anfängliche Leichtigkeit täuscht jedoch nicht darüber hinweg, dass Suche ein komplexes Problem ist. Elasticsearch bietet für dessen Lösung eine hervorragende Infrastruktur und gestattet das komfortable Management der Indices. Auf Tuchfühlung gekommen, steigt die Lernkurve schnell an. Anspruchsvolle Implementierungen erfordern neben der Auseinandersetzung mit den theoretischen Grundlagen ein ausgereiftes fachliches Konzept und ein Objektmodell.

Dokumentation

Auf seiner Website bietet der Hersteller eine aktuelle, umfangreiche Referenz, auch für zurückliegende Versionen. Videos demonstrieren Leistungsmerkmale und Verwendung der Komponenten. Ausgewählte Schwerpunktthemen werden in Blogs ausführlicher beleuchtet. Der angebotene „Definitive Guide“ beschreibt die grundlegenden Prinzipien. Bei Problemen hilft die aktive Community. Einige der Im Netz kursierenden Tutorials funktionieren mit dem aktuellen Release nicht mehr, da die Entwicklung der Software rasch voranschreitet.

Anforderungen

Ein aus mehreren Modulen bestehendes System sollte um eine übergreifende Suche erweitert werden, vorzugsweise mit Standard Komponenten unter Verwendung offener Protokollen ohne Anpassungen des vorhandenen Codes. Wegen der Skalierbarkeit mit moderaten Anforderungen an die Hardware sowie ein optimales Ineinandergreifen der im ELK Stack verfügbaren Tools fiel die Wahl auf Elasticsearch.

Die vorhandenen Anwendungen bieten ausgefeilten Funktionen zur Suche. Wie aus CRM oder ERP Systemen bekannt, können beispielsweise zu einem Haushalt alle Kunden selektiert oder Verträge nach unterschiedlichen Merkmalen aufgelistet werden. Über alle Anwendungsbereiche hinweg, soll die Suche nach Informationen, einschließlich der Textfeldern, wie Bemerkungen, Notizen ermöglicht werden. Mit den Treffern soll der direkte Sprung zur Bearbeitung der Ressourcen angeboten werden.

Umsetzung

Die Suche erweitert die Anwendung parallel zu den existierenden Modulen ohne Änderungen am vorhandenen Code. Suchanfragen werden von einem eigenen Server beantwortet. Diese Komponente übernimmt Authentifizierung und Autorisierung und kommuniziert ihrerseits mit dem Elasticsearch Backend.

Beim Laden der Daten hat sich Logstash bewährt. Das JDBC- Plug-in ist äußerst flexibel und leistungsfähig. Der Durchsatz bei der initialen Indizierung entspricht den Erwartungen. Für die Synchronisierung wird das gleiche Tool verwendet.

Die Benutzer Oberfläche stellt der Suchserver zur Verfügung. Kibana ist für den Betrieb nicht unbedingt erforderlich. Entwicklung und Test erleichtert es ungemein. Analysen liefern den nötigen Überblick. Die Oberfläche unterstützt die Erstellung komplexer Statements.

Kontaktadresse:
Gunther Thielemann
SLIX Gesellschaft für Computersysteme mbH
Nandlstädter Weg 6
D-84072 Au i.d. Hallertau

Telefon: +49 (0) 8752-219308
Fax: +49 (0) 8752-85034
E-Mail gunther.thielemann@slix.de
Internet: www.slix.de