MarkLogic®

DATENBLATT

MarkLogic und Hadoop

90 Prozent der weltweit vorhandenen Daten wurden in den vergangenen zwei Jahren erstellt. Unternehmen müssen riesige Mengen strukturierter und unstrukturierter Daten aus den unterschiedlichsten Quellen speichern und analysieren. Mit herkömmlichen relationalen Datenbanken lassen sich solche Mengen schon längst nicht mehr bewältigen. Hadoop* ist ein hervorragendes Tool für diese Aufgabe, und MarkLogic® ist die beste Datenbank für Hadoop.

Hadoop: HDFS und MapReduce

Die Popularität von Hadoop hat in den letzten Jahren rasant zugenommen, denn Hadoop ist gut dafür geeignet, günstig große Datenmengen in dem Hadoop File System abzulegen (HDFS). Hadoop kommt in vielen Anwendungen zum Einsatz. Es bietet die Möglichkeit, diese Daten kostengünstig im Hadoop* Distributed File System (HDFS*) zu speichern und umfangreiche MapReduce-Prozesse für die Batch-Analyse auszuführen.

- HDFS ist ein Java-basiertes Dateisystem zur flexiblen und zuverlässigen Speicherung von Daten auf Commodity-Server-Clustern. In Produktionsumgebungen wurde HDFS bereits erfolgreich auf 4.500 Server und 200 Petabyte skaliert, womit nahezu eine Milliarde Dateien verarbeitet werden konnten.
- MapReduce ist ein Verarbeitungs-Framework, das auf dem Teile-und-Herrsche-Ansatz („Divide and Conquer“) basiert, bei dem große Aufgabenstellungen in kleinere Aufgaben zerlegt („Map“) und die Ergebnisse der einzelnen Aufgaben zusammengeführt werden („Reduce“). Jede in Teilbereiche zerlegbare Aufgabe ist mit Hadoop kompatibel.

MarkLogic: Eine der besten Datenbanken für Hadoop

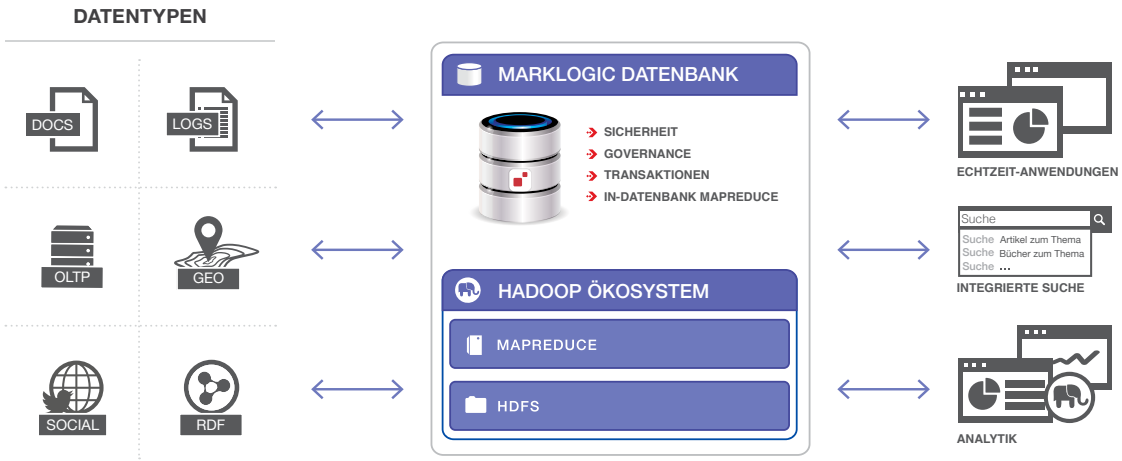
Hadoop ist perfekt für das Speichern und Analysieren von Daten geeignet. Um Transaktionen mit geringer Latenz für interaktive Echtzeit-Anwendungen oder Anwendungen mit Funktionen wie Hochverfügbarkeit oder Disaster Recovery zu ermöglichen, benötigt Hadoop jedoch eine Datenbank. Obwohl das Hadoop Ökosystem kontinuierlich weiterentwickelt wird, lässt sich das volle Potenzial von Hadoop nur dann ausschöpfen, wenn es mit einer leistungsfähigen Datenbank kombiniert wird.

MarkLogic ist eine der besten Datenbanken für Hadoop, da sie nahtlos in das Hadoop Ökosystem integriert werden kann und leistungsstarke transaktionale Echtzeit-Anwendungen ermöglicht. Darüber hinaus lassen sich mit MarkLogic die Vorteile von HDFS mit einem Tiered-Storage-Modell verknüpfen, sodass Daten beliebig zwischen HDFS, S3, SSD, SAN, NAS oder verschiedenen Datenträgern verschoben werden können. So erfüllen Sie spezifische SLAs und Kostenvorgaben, ohne entsprechende Änderungen am Code vornehmen zu müssen.

Durch die Kombination von MarkLogic und Hadoop profitieren Sie vom kostengünstigen Hadoop Speicher und den Funktionen von MarkLogic, wie z. B. ACID-Transaktionen, Hochverfügbarkeit, Disaster Recovery, den höchsten Sicherheitsstandards und Tools für die Leistungskontrolle.

MarkLogic®

MARKLOGIC	HADOOP
• Online-Anwendungen mit geringer Latenz • Echtzeit-Transaktionen • Integrierte Suchfunktionen	• Offline-Verarbeitung mit hoher Latenz • Umfangreiche Batch-Analyse • Verteilter, kostengünstiger Speicher



Moderne Hadoop Infrastruktur

Ähnlich wie Hadoop speichert auch MarkLogic unstrukturierte Daten in Clustern. Dies macht es besonders einfach, mithilfe des MarkLogic Connectors für Hadoop Datenpartitionen („Forests“) zwischen MarkLogic Hosts und dem Hadoop Ökosystem zu verschieben, wobei MarkLogic sowohl als Quelle als auch als Ziel agieren kann.

Anwendungsbeispiele

- **Umfassende ETL-Prozesse** – Verwenden Sie Hadoop für die Verarbeitung von Rohdaten, insbesondere für ressourcenintensive Prozesse. Sie können dafür auf hochspezialisierte Bibliotheken wie Gesichtserkennung in Bildern, maschinelles Lernen oder die komplexe Extraktion von Entitäten zurückgreifen. Übermitteln Sie dann das Ergebnis an MarkLogic für Ad-hoc-Abfragen anhand von MarkLogic-Indexen.

- **Archivierung** – Verwalten Sie Daten während des gesamten Lebenszykluses: Betriebsdaten werden in MarkLogic gespeichert, während HDFS als Tiered Storage zur Archivierung älterer Daten dient, die mit geringerer Wahrscheinlichkeit benötigt werden. Daten in HDFS sind weiterhin jederzeit in MarkLogic verfügbar. So können Sie operative Vorgaben zu geringeren Kosten erfüllen.

- **Datensicherung in Hadoop** – MarkLogic ist die sicherste NoSQL-Datenbank und bietet ein Modell, das auch für in HDFS gespeicherte Daten gilt und die Einschränkung analytischer Prozesse erlaubt. Diese Einstellung ist bereits vorkonfiguriert, sodass Sie keine anderen Komponenten wie Zookeeper oder Accumulo integrieren müssen.

MARKLOGIC GMBH
Theatinerstr. 11, 8. Etage, München 80333 | +49 897 10 42 2151
Skyper Villa, Taunusanlage 1, Frankfurt 60329 | +49 695 05 06 0588
de.marklogic.com | germany@marklogic.com