Kriterien für den Einsatz von Big Data Technologien

Peter Welker
Trivadis GmbH
Stuttgart

Schlüsselworte

Big Data, Business Intelligence, Kosten, Know-How, Performance, Kriterien, Auswahl

Einleitung

Wann lohnt sich der Einsatz von Big Data Technologien im Data Warehouse & Business Intelligence Umfeld und wann nicht? Der Vortrag betrachtet "klassische" und "neue" Anforderungen an die Datenanalyse und setzt diese mit technischen und weiteren Kriterien in Relation: Latenz, Durchsatz, Datenmenge, Informationsqualität und -struktur sind dabei nicht genug. Auch Anwenderfreundlichkeit, Reifegrad, Stabilität, Administrierbarkeit oder schlicht das verfügbare KnowHow und Vorlieben von Technikern und Anwendern sind zu berücksichtigen. Und natürlich spielen die Kosten eine wichtige Rolle. Der Artikel fasst die Einzelthemen knapp zusammen.

Einführung

Soll ich Big Data Technologien für meine Anforderungen einsetzen oder lieber doch bei meinen „Schusterleisten“ bleiben und mit RDBMS und klassischen Werkzeugen arbeiten? Diese Frage mögen sich mancher Datenbankentwickler- und Administrator genauso wie zahlreiche IT-Entscheider heute stellen – und sie ist nicht so ohne weiteres zu beantworten. Es gibt keine klaren, 100%igen Kriterien, die man hier einfach als Messlatte anlegen könnte. Vielmehr müssen zahlreiche Parameter gegeneinander abgewogen werden. Zuallererst aber sollte man seine fachlichen Anforderungen sehr gut kennen. Diese Präsentation versucht, wichtige Parameter herauszustellen und sie den Teilnehmern näher zu bringen.

Big Data?

Klar, eine kurze Einführung in Big Data darf natürlich nicht fehlen. Hier also ist sie:

Im Augenblick werden über das Internet jede Sekunde mehr als 22 TB an Daten versendet. Die bestehen beispielsweise aus über 2 Millionen Emails, 45000 Google-Suchanfragen, 85000 angeklickten Youtube Videos, 7500 tweets und sehr vielem mehr. Das meiste davon ist natürlich redundant, auch auf lange Sicht uninteressant oder für viele von uns schlicht irrelevant. Manches aber wirkt sich stark auf unsere Umgebung, uns persönlich oder unser Unternehmen/Arbeitgeber aus. Wenn beispielsweise ein „Shitstorm“ über ein Produkt unseres Unternehmens hereinbricht, kann es wichtig sein, innerhalb von Minuten oder Stunden gegenzusteuern – und nicht erst, wenn die Medien einen Tag später davon berichten. Solche Maßnahmen, man nennt sie Sentiment-Analyse, erfordern aber einen nicht unerheblichen Aufwand, mit nicht unerheblichem Know-How nicht unerhebliche Datenmengen in nicht unerheblich kurzer Zeit zu analysieren und zu bewerten.

Und da haben wir an einem einzigen Beispiel auch schon die wichtigsten Kriterien gelistet. Es geht um Datenmengen, Performance, Latenz, Analytik, Kosten, Know-How und noch ein paar andere Dinge.

Natürlich geht es nicht immer nur um's Internet, in unserer Umgebung produzieren wir auch so jede Menge Daten, wie Sensorinformationen bei Forschungs- und Produktionsstätten, Bewegungsprofile via GPS und RFID oder Bilder und Videos mit unseren Kameras und Handies.

Falls uns manche dieser Daten interessieren – und einige tun es definitiv – mit welcher Technik wollen wir uns der Aufgabe stellen, hieraus Nutzen zu ziehen? Wollen wir herkömmliche (RDBMS, App Server usw.) oder neue Technik (Big Data!?) dafür einsetzen. Schauen wir uns dafür mal an, was die Analysten dazu meinen:

- **McKinsey**
  "Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

- **Gartner**
  "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization."

- **BARC**
  "Big Data designates methods and technologies for the highly scalable acquisition, storage, and analysis of polystructured data"

Ich erlaube mir, das in Kürze so zusammenzufassen:

Für manche Aufgaben brauchen wir heute unkonventionelle Methoden und Technologien für „unlimitiertere" Datenverarbeitung

Unkonventionell, das sind die (teils gar nicht mehr so) neuen Plattformen wie Hadoop oder NoSQL Datenbanken mit großem Potential für bestimmte Fragestellungen, meist hoher Skalierbarkeit, eingeschränkter Konsistenz und limitierter Administierbarkeit. Konventionell sind die herkömmlichen RDBMS und Softwarearchitekturen mit gereifter und „allumfassender" Funktionalität, erprobtem Nutzen und Konsistenz, dafür aber Einschränkungen bei extremen Anforderungen und nicht selten hohen Lizenzkosten.

Was setzen wir wo ein und warum. Betrachten wir nun also die Kriterien im Detail.

Latenz & Performance

Latenz ist die Zeit von der Entstehung der Information bis zur gewünschten Reaktion. Typische Klassen gelten beispielsweise für

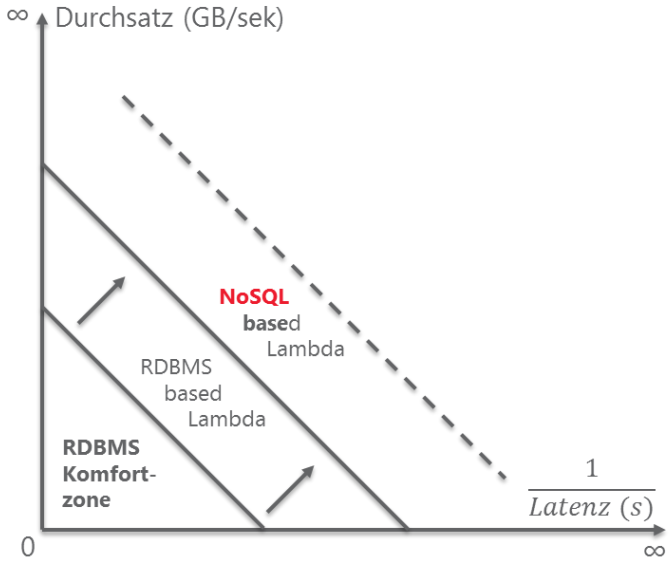
Fraud Detection oder Aktienhandel – Reaktion innerhalb **Millisekunden oder Sekunden**

- Sperrung von Kreditkarten oder Mobiltelefonen
- Entdecken und Unterbinden unbefugter Zugriffe auf IT Systeme

„Real Time" (Business) Intelligence – Verfügbarkeit **< Minuten bis Stunden**

- Fehleranalyse ungewöhnlicher Sensordaten in Stromversorgungsnetzen
- Erkennen sicherheitsrelevanter Ereignisse im Social Media Umfeld
- Sentimentanalyse von produktrelevanten Aussagen im Internet

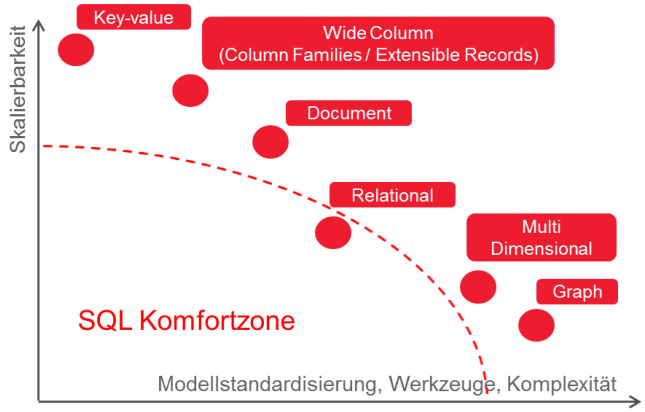
Klassische Business Intelligence (**täglich, wöchentlich, monatlich**)



Je kleiner die Transaktionen, desto geringer der Durchsatz und je kürzer die Latenzen, desto kleiner die Transaktionen. Somit gilt also auch, **je kürzer die Latenzen, desto geringer der Durchsatz**

Hier haben NoSQL Datenbanken mit geringen Anforderungen an Konsistenz und dadurch schnelleren Verarbeitungszeiten auch bei hohem Durchsatz einen klaren Vorteil. Im Zusammenhang mit speziellen Architekturen (bspw. Lambda) und neuen Features lassen sich die Grenzen der RDBMS zwar verschieben, aber für die Sekundennahe Analyse von hunderttausenden oder Millionen von Ereignissen ist der Einsatz von Speziallösungen meist unumgänglich

Abb. 1: Grenzen von klassischen RDBMS in Abhängigkeit von Durchsatz und Latenz



Bestimmte Klassen von NoSQL Datenbanken haben dabei bestimmte Stärken und Schwächen. Prinzipiell gilt aber, je einfacher in der grundlegenden Struktur und den Konsistenzbedingungen, desto mehr Transaktionen und Daten sind innerhalb eines Zeitraums durchführbar.

Mit batchorientierten Lösungen wir Hadoop und Map/Reduce steht dabei nicht die Latenz sondern der reine Durchsatz im Vordergrund – und die Möglichkeit, strukturierte wie unstrukturierte Daten zu verarbeiten.

Abb. 2: Einordnungsversuch von Technologien in Abhängigkeit von Komplexität und Skalierbarkeit

Strukturen, Modelle, Visualisierung und Analytik

Komplexe Analytik und Visualisierung benötigt im Allgemeinen speziell vorbereitete Datenstrukturen. Hier spielt die Auswahl der Basisplattform meist eine untergeordnete Rolle. Eine Ausnahme bilden die Standardreporting- und BI-Werkzeuge. Diese sind für die Nutzung des vollem Funktionsumfangs oft auf RDBMS ausgerichtet, öffnen sich aber mehr und mehr auch für die NoSQL- und Hadoop-Welt.

Anders sieht es bei spezialisierten Modellierungsvarianten aus. Für das Speichern von Dokumenten oder das Darstellen und Auffinden von Beziehungen zwischen Entitäten eignen sich RDBMS beispielweise ziemlich schlecht. Bei letzterem spielen Datenbanken für ontologische Modellierung (Semantic Web) die Hauptrolle. Allerdings bietet z.B. Oracle mit der Spatial Option bereits eine Lösung an, um Semantic Webs innerhalb der Oracle DB abzubilden – nur eben nicht relational.

Kosten

Auf den ersten Blick ist die Kostenfrage schnell entschieden. Selbst innerhalb der Oracle Welt liegt eine mit den üblichen DWH spezifischen Lizenzen (Enterprise Edition, Partitioning, RAC) ausgestattete Exadata gemäß der offiziellen Preisliste locker um eine Größenordnung (> Faktor 10) höher als eine vergleichbar große Big Data Appliance. Bei genauer Betrachtung spielen aber natürlich auch die Anforderungen, Know-How, Komplexität der darauf aufbauenden Applikation und viele weitere Kriterien eine Rolle. Bei typischen TCO Diskussionen ist daher kein dauerhafter Sieger auszumachen. Vielmehr sind die Kosten extrem abhängig davon, wie gut sich die jeweilige Plattform für die Umsetzung spezifischer Anforderungen eignet (s.o.).

Know-How

Bei der Know-How Frage gibt es eine triviale und eine eher „bunte" Antwort. Die triviale berücksichtigt Fragen nach Verfügbarkeit des Know-Hows auf dem Arbeitsmarkt, die Kosten für Ausbildung, die potentiell höheren Kosten bei Best-Of-Breed Ansätzen weil das Know-How meist auf mehr Köpfe verteilt ist, die anfängliche Ineffizienz bei der Entwicklung nach einer Neuausbildung usf.

Die psychologische Betrachtung stellt in erster Linie die Frage nach dem bevorzugten „Hammer" jedes Entwicklers oder Administrators – aber auch IT Entscheiders: Je nachdem aus welcher „IT-Kultur" wir stammen, fällt uns der Einsatz „neuer" Technik (Hadoop, NoSQL usw.) unvergleichlich viel leichter oder schwerer als der Einsatz von RDBMS und klassische Entwicklertools und wir neigen oft zum Einsatz des bekannten, wenn der Druck steigt. Dazu kommt aber noch der teils positive, teils negative Einfluss persönlicher Interessen (auch an Neuem) und Neigungen, der Kollegen, des Chefs usw. In Summe wird der direkte oder indirekte Einfluss von technischen Keyplayern auf Entscheidungen bei HW/SW/Architekturentscheidungen gerne unterschätzt.

Kontaktadresse:
Peter Welker
Trivadis GmbH
Industriestrasse 4
D-70565 Stuttgart

Telefon: +49 (0) 162-2969681
E-Mail peter.welker@trivadis.com
Internet: www.trivadis.com