HANSER

Vorwort

Stefan Edlich, Achim Friedland, Jens Hampe, Benjamin Brauer, Markus Brückner

NoSQL

Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken

ISBN: 978-3-446-42753-2

Weitere Informationen oder Bestellungen unter

http://www.hanser.de/978-3-446-42753-2

sowie im Buchhandel.

Geleitwort

Im Internetzeitalter werden Begriffe auf unkonventionelle Weise „definiert“. So etablierte Tim O’Reilly den Begriff Web 2.0 durch Beispiele. Diese Vorgehensweise lässt in unserer schnelllebigen, marketingorientierten Welt Spielräume, um Begriffe bei Bedarf „umzudefinieren“ und zu erweitern. Ähnlich verhält es sich mit „NoSQL“. Dieser Begriff wird zunächst durch das definiert, was er nicht ist, also kein SQL. Natürlich sind Negativdefinitionen nicht zielführend – insbesondere nicht, um aktuelle Trends auszudrücken, denn natürlich sind alle Datenbanksysteme von IMS über hierarchische Systeme und sogar einige relationale Systeme NoSQL-Datenbanken im Sinne dieser Definition. Allerdings ist das nicht von den NoSQL-Protagonisten beabsichtigt. NoSQL hat nichts mit der strukturierten Anfragesprache SQL zu tun, sondern ist nur als provokative Phrase zu verstehen, um Aufmerksamkeit zu erregen. Dies zeigt sich auch daran, dass heutzutage Teile der Community NoSQL als „not only SQL“ auffassen und somit die strikte Abgrenzung von SQL aufweichen.

Worum geht es der NoSQL-Community also wirklich? NoSQL will neue Alternativen zum allgegenwärtigen relationalen Datenmodell und zu üblichen Datenbanktechnologien wie Transaktionsmanagement herausstellen, die für bestimmte Anwendungsklassen hinsichtlich der Betriebskosten, Anwendungsentwicklung oder Skalierbarkeit der eierlegenden Wollmilchsau „relationales Datenbanksystem“ überlegen sind. Dabei spielt die Demokratisierung der Softwareentwicklung durch die Open Source-Community eine große Rolle für NoSQL. Heutzutage werden Standards wie z.B. SQL, XML oder XQuery nicht mehr topdown von Standardisierungsgremien entwickelt. Stattdessen entstehen De-facto-Standards wie das Map/Reduce-Programmiermodell durch die breite Anwendung von Open Source-Projekten wie Hadoop. Ferner liefert die Open Source-Community eine große Zahl von anwendungsspezifischen Lösungen für bestimmte Datenverarbeitungsprobleme wie Graphanalyse, Data Mining oder Textanalyse. Diese Lösungen sind zunächst kostengünstiger als kommerzielle Lösungen und erfahren daher insbesondere bei softwareentwicklungsaffinen Internetunternehmen großen Zuspruch. Ein weiterer Aspekt ist die Anwendungsadäquatheit der Systeme und der Datenmodelle. So eignen sich Key/Value Stores hervorragend zur Logfile-Analyse. Map/Reduce-Programme sind aufgrund ihrer funktionalen Programmierweise für viele Entwickler leichter zu entwickeln als deklarative SQL-Anfragen. Ferner skaliert Map/Reduce bei großen Datenmengen auf großen Rechnerclustern durch spezielle Protokolle zur parallelen, fehlertoleranten Verarbeitung. Andere Systeme wie RDF-Datenbanken eignen sich gut zur Verarbeitung von Ontologien oder anderen graphisch strukturierten Daten mit der Anfragesprache SparQL.

Es bleibt abzuwarten, wie sich die NoSQL-Systeme in Zukunft entwickeln werden. Aufgrund ihrer Anwendungsunabhängigkeit haben sich relationale Datenbanken nachhaltig am Markt behauptet und Strömungen wie Objektorientierung und XML erfolgreich integriert. Ferner verringern relationale Datenbanken den Impedance Mismatch zwischen Programmiersprache und Datenbanksystem durch Technologien wie LinQ. Da NoSQL allerdings nicht nur auf das Datenmodell abzielt, ist eine evolutionäre Integration von Technologien aus diesem Bereich in relationale Datenbanken schwieriger. Die nächsten Jahre werden zeigen, welche der vielen NoSQL-Systeme sich am Markt behaupten werden, und ferner, ob NoSQL-Systeme spezifische Lösungen für Internetunternehmen bleiben oder den Sprung in den Mainstream der Anwendungsentwicklung schaffen.

Prof. Dr. Volker Markl

TU-Berlin, Fachgebiet Datenbanksysteme und Informationsmanagement

Vorwort

Es ist schon ein besonderes Glück, eine Zeit mitzuerleben, in der sich die Datenbankwelt stark reorganisiert und neu erfindet. Und obwohl sie das schon seit der Web 2.0-Datenexplosion tat, fiel es den neuen Datenbanken schwer, sich zu formieren und auf sich aufmerksam zu machen. Dies änderte sich radikal mit der Vereinigung fast aller nichtrelationalen Datenbanken unter dem Begriff NoSQL. Täglich schrieben Datenbankhersteller aller Couleur die Autoren an, um ebenfalls auf *http://nosql-databases.org* gelistet zu werden. Dazu gehörten beispielsweise alle XML-Datenbankhersteller, aber auch Firmen wie IBM oder Oracle, die selbst unbedingt mit Lotus Notes/Domino oder BerkeleyDB im NoSQL-Boot sein wollten. Nichtrelationale Datenbanken und die Vermeidung von SQL fingen an, hoffähig zu werden. Die Botschaft, dass die Welt aus mehr als relationalen Beziehungen besteht, begann sich langsam auch in den Köpfen einiger Entscheider festzusetzen.

Dabei ist die Trennung der relationalen SQL- und der NoSQL-Welten nicht trivial, zumal ja die Vermeidung von SQL nicht unbedingt heißen muss, dass auf ein relationales Modell verzichtet wird. Dennoch gibt es in der NoSQL-Welt eine Vielzahl von Systemen wie InfoGrid, HyperGraph, Riak oder memcached, die intern auf bewährt relationale Datenbanken aufsetzen oder aufsetzen können. Sie bieten nur nach oben hin andere Datenmodelle, Abfragesprachen und ggf. sogar Konsistenzzusicherungen an. Die Zahl der Hybridlösungen wie HadoopDB oder GenieDB steigt ebenfalls an und erschwert eine scharfe Trennung der Welten.

Ein weiteres Kennzeichen für schnelles Wachstum des NoSQL-Bereichs ist auch, dass bereits viele NoSQL-Systeme wiederum andere NoSQL-Systeme nutzen. Dafür lassen sich viele Beispiele finden. So nutzt OrientDB die Hazelcast-Bibliotheken, um ein verteiltes System zu realisieren. Oder Scalaris kann auch auf Erlangs ETS-Tabellenspeicher oder auf Tokyo Cabinet aufsetzen. Wie ein Netzwerk nutzt hier eine passende Lösung die andere – auch übergreifend zwischen der SQL – und der NoSQL-Welt.

NoSQL dürfte sicherlich einer der Bereiche der Informatik sein, der sich derzeit am schnellsten entwickelt. Dies macht es doppelt schwer, ein Buch zu dieser Thematik zu schreiben. Den Autoren erging es fast immer so, dass sich die API eines Produktes nach Abschluss eines Kapitels zu 25 % vergrößert oder verändert hatte. Aus diesem Grunde war es uns wichtig, auch vorab die theoretischen Grundlagen hinter NoSQL zu beschreiben, um wenigstens einen relativ stabilen Teil im Buch zu haben. Dies kann in diesem weltweit ersten NoSQL-Werk natürlich nicht in der Ausführlichkeit und Tiefe geschehen, wie dies sicherlich einige Experten aus den Universitäten durchaus hätten schreiben können. Da wahrscheinlich der Großteil der Leser jedoch sicherlich nicht übermäßig an den formal theore-

tischen und mathematischen Grundlagen interessiert ist, haben wir mit den Themen in Kapitel 2 - Map/Reduce, Hashing, CAP, Eventually Consistent und den Basisalgorithmen – hoffentlich einen guten Kompromiss für alle Leser gefunden. Wirklich vertieft konnten wir auf die Theorie hier jedoch nicht eingehen. Durchaus wichtige Themen wie verschiedene Consensus-Protokolle, Replikations- oder Routing-Strategien sind daher bewusst nicht enthalten, da unser Ziel ein praxisorientiertes Werk war. Sie sollten sich aber auch nicht von der Theorie zu Beginn des Buches abschrecken lassen. Sie können in jedes Kapitel einsteigen, ohne unbedingt das Vorwissen von Kapitel 2 mitzubringen.

Der praktische Teil, also die Kapitel zu den bekanntesten NoSQL-Werkzeugen, dient in erster Linie dazu, ein Gefühl für die jeweilige Datenbank zu bekommen. Aus diesem Grunde war es uns einerseits besonders wichtig, zunächst einen kurzen Steckbrief zu jedem Werkzeug anzugeben. Andererseits sind die Bewertungen, Vor- und Nachteile sowie Einsatzgebiete der Datenbanken hoffentlich hilfreich für die Einschätzung der Werkzeuge. Das Gleiche gilt für das letzte Kapitel, welches vorsichtig versucht, einen Leitfaden für die gesamte Datenbankwelt zu entwickeln. Denn auffällig ist derzeit, dass es zwar viele Schriften und Werkzeuge zu Datenbanken gibt, aber kaum Bücher, die Anwender auf dem Weg durch dieses DB-Universum an die Hand nehmen und Orientierung bieten. Dies ist dann auch die wesentliche Zielrichtung des Buches: die Breite des theoretischen und praktischen Spektrums darzustellen und mit Bewertungsrichtlinien zu helfen. Es war nicht das Ziel, ein NoSQL-Monogramm zu erstellen, welches alle Bereiche des NoSQL-Feldes abdeckt. Wir bezweifeln, dass es in diesem expandierenden Universum überhaupt möglich wäre, ein solch umfassendes Werk zu erstellen.

Abschließend möchten die Autoren den Verlagen danken. Zunächst dem Hanser Verlag, der immer sehr gut ansprechbar war und uns mit Frau Margarete Metzger eine kompetente und leidenschaftliche Begleiterin war. Ferner der Neuen Mediengesellschaft Ulm, die mit Erik Franz für dieses Buch gute Synergieeffekte entwickelt hat. Wir hoffen, es ist nicht das letzte Werk in dieser Serie. Unser Dank gilt natürlich auch Marko Röder, der sich sehr viel Zeit nahm, unsere Texte ausführlich mit Kommentaren zu versehen und zu verbessern. Außerdem sei gedankt: Peter Neubauer für das Neo4j-Kapitel, Marc Boeker für das MongoDB Review, Henrik und Reidar Hörning für praktische Hilfe, Kamoliddin Mavlonov. Und schließlich möchten wir uns bei den folgenden Firmen für die Unterstützung bedanken: Objectivity / Infinite Graph, Versant, NeoTechnologies und der Sones GmbH.

Berlin im August 2010

Stefan Edlich, Achim Friedland, Jens Hampe, Benjamin Brauer

Vorwort zur 2.Auflage

Für die Autoren war es eine angenehme Überraschung, nur sechs Monate nach Erscheinen der ersten Auflage des NoSQL-Buches die Arbeit für die zweite Auflage aufnehmen zu dürfen. Dies spiegelt sehr stark die unglaubliche Dynamik des NoSQL-Bereiches wider. Jedoch ist eine NoSQL-Neuauflage – selbst nach einer so kurzen Zeit wie einem oder einem halben Jahr – wirklich kein leichtes Unterfangen. Denn die meisten Datenbanken entwickeln sich rasant weiter und täglich kommen neue Vertreter hinzu. Die wichtigsten Neuerungen stellen wir in diesem Buch in drei neuen Kapiteln vor:

- REST für NoSQL
- Membase und
- OrientDB

Darüber hinaus sind fast alle Datenbank-Kapitel der ersten Auflage überarbeitet worden. Grund dafür sind auch drei weitere wichtige Entwicklungen, die bisher das Jahr 2011 geprägt haben.

Die erste betrifft natürlich die NoSQL-Unternehmensnachricht des Jahres: den Zusammenschluss von Membase und CouchDB. Auf den ersten Blick erscheint die Fusion auch sinnvoll: die Skalierbarkeit und Performance von Membase mit den Vorteilen der Replikation und dem mobilen Einsatzbereich der CouchDB zu kombinieren. Dennoch wird die Fusion sicher schwieriger umzusetzen sein als erwartet und noch einige Fragen – wie das Angebot leistungsfähiger Ad-hoc-Queries – offenlassen.

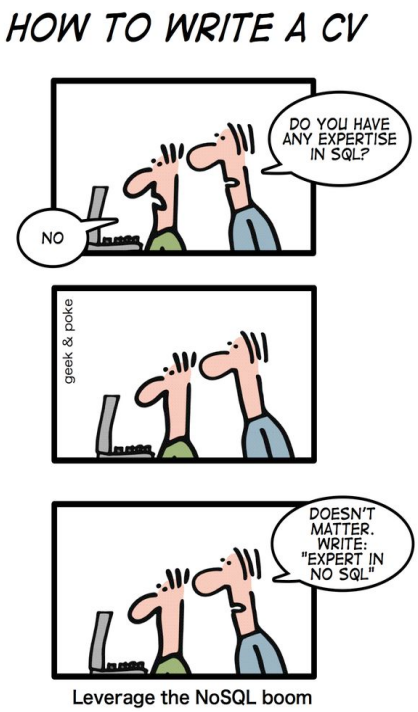
Die zweite Entwicklung ist die explosionsartige Ausdehnung des Bereiches der Graph Datenbanken. Selbst Microsoft spielt hier mit. Auch diesem Bereich wurde mit einem noch umfangreicheren Kapitel Rechnung getragen.

Die letzte Entwicklung betrifft einen weiteren Datenbankbereich, der ebenfalls mit einem „Buzzword“ auf sich aufmerksam macht: „NewSQL“. NewSQL ist ein Begriff, der stark von der 451group – eine Firma für Analysen ähnlich wie Gartner – geprägt worden ist. Damit gemeint ist der Bereich der skalierbaren und hochperformanten klassischen, relationalen Datenbanken, die oftmals auch als Cloud-Service angeboten werden. Dazu gehören beispielsweise Systeme wie VoltDB oder database.com in der Cloud. Diese Systeme versprechen, bei voller Skalierbarkeit die Transaktionalitätsanforderungen nicht zu vernachlässigen, und hoffen, auf diese Weise Mehrwert gegenüber vielen NoSQL-Systemen liefern zu können.

Über diesen NewSQL-Bereich ließe sich trefflich ein komplett neues Buch schreiben. Unserer Ansicht nach hat aber NoSQL entscheidend dazu beigetragen, die Schwächen der klassischen Datenbanken aufzuzeigen und die NewSQL-Bewegung zu fördern, was wiederum ein Vorteil für den gesamten Datenbankraum – und damit für uns alle – ist.

NoSQL ist derzeit immer häufiger in weltweiten IT-Stellenausschreibungen zu finden. Aufstrebende Technologien wie Facebook und Cloud Computing halten jedoch nicht nur dort Einzug, sondern auch in heiteren Portalen wie Geek&Poke (http://geekandpoke.typepad.

com). Und so wundert es nicht, dass auch NoSQL im Januar 2011 mit einem wunderbar selbstironischen Sketch vertreten war.



Quelle: http://geekandpoke.typepad.com/geekandpoke/2011/01/nosql.html

Wir hoffen, dass NoSQL auch selbst dazu beiträgt, die Welt der Daten unbeschwert und mit frischen Technologien zu betrachten, und so diesem Bereich neuen Schwung gibt.

Für die tatkräftige Unterstützung bei der zweiten Auflage möchten wir uns bei Alex Vollmar und Manuela Brückner bedanken.

Berlin, im Juli 2011

Stefan Edlich, Achim Friedland, Jens Hampe, Benjamin Brauer, Markus Brückner