Big Data

Schatztauchen im Datenmeer

Big Data beschreibt das Phänomen des Datenwachstums und bietet neue Methoden und Produkte zur Verarbeitung größter Datenmengen. Materna Monitor erläutert, wie Unternehmen ihre Datenschätze heben und gewinnbringend einsetzen können.

Im Jahr 2012 wurden weltweit erstmals 1,8 Zettabyte an Daten produziert und Prognosen zufolge verdoppelt sich diese Menge künftig alle zwei Jahre. Diese Datenflut entspricht der Speicherkapazität von 57,5 Milliarden Apple iPads mit 32 Gigabyte Speicher. Mit diesen Geräten ließe sich eine 31 Meter hohe Mauer quer durch Europa von Moskau nach Lissabon bauen. Die Fläche der aneinandergelegten iPads wäre so groß wie die Stadtflächen von München und Berlin zusammen.
Doch nicht nur die Menge, auch die Vielfalt der Daten nimmt zu. Dies liegt an der zunehmenden Verbreitung des Internets – Stichwort Internet der Dinge – sowie intelligenten netzwerkfähigen mobilen Geräten. Darüber hinaus sind Unternehmen durch Regularien oder Gesetze gezwungen, immer mehr Daten aufzubewahren. Zahlreiche Manager haben bereits erkannt, dass die Daten aus dem operativen Geschäft zum Erreichen der Unternehmensziele vielfältig nutzbar sind.
Die Datenquellen sprudeln also reichlich. Täglich entstehen mehrere Exabytes an neuen Daten: Facebook hat ein tägliches Datenaufkommen von mehr als 500 Terabytes, bei Google sind es 20 Petabytes und WalMart verarbeitet stündlich 2,5 Petabytes an Benutzertransaktionen.

Schatzsuche: Wer benötigt Big Data?

Eine zentrale Frage für Unternehmen lautet daher, ob es sich lohnt, den Schatz in den Datenbergen zu heben. Folgende Kriterien helfen bei der Analyse, ob Big Data für eine Organisation relevant ist:

- Speichert Ihr Unternehmen mehr Daten als ausgewertet werden?
- Müssen regelmäßig neue Datenquellen und Formate hinzugefügt werden?
- Beklagen sich Fachbereiche über Datensilos und fehlende aktuelle Informationen?
- Könnten Geschäftsprozesse verbessert werden, wenn Informationen schneller und detaillierter vorliegen würden?

- Wo stoßen bisherige Verfahren zur Datenauswertung an ihre Grenzen?

Diese und weitere Fragen können bei der Entscheidungsfindung helfen, ob Big Data für eine effizientere Unternehmensführung, eine Individualisierung von Dienstleistungen oder auch bei der Entwicklung intelligenterer Produkte nützlich sein kann. Viele Firmen setzen bereits unbewusst Big Data-Technologien ein, wenn sie beispielsweise große Datenmengen indizieren, wie bei der Volltextsuche mit Apache Solr oder beim Auswerten von Weblog-Dateien in Realzeit.
Die Relevanz von Big Data lässt sich nach dem V-Schema unterteilen:

- Velocity (Änderungshäufigkeit): Wie schnell ändern sich die Daten?
- Volume (Größe): Wie groß ist die zu verarbeitende Datenmenge?
- Variety (Datentypen): In welchen Formaten, Quellen oder Strukturen liegen die Daten vor?
- Viscosity (Datenfluss): Wie verändern sich die Daten im Laufe der Zeit oder innerhalb eines Geschäftsprozesses?
- Virability (Ausbreitung): Wo werden die Daten genutzt oder benötigt?
- Value (Geschäftswert): Welchen Wert haben die Daten für Geschäftsprozesse?

Welche Aspekte relevant sind, müssen Unternehmen anhand ihrer Prioritäten entscheiden, bevor das Big Data-Öl durch den V-Motor fließen kann. Die Professoren Erik Brynjolfsson und Andrew McAfee des Massachusetts Institute of Technology (MIT) nennen in ihrem Artikel „Wie Big Data das Management revolutioniert“ im Harvard Business Magazin zahlreiche anschauliche Beispiele für Big Data. Sie kommen zu der Erkenntnis, dass Unternehmen, die Big Data einsetzen, klügere Entscheidungen treffen.

Der kleine Elefant und die großen Daten

Eine Idee von Big Data besteht darin, Daten direkt an ihrem Entstehungsort zu verarbeiten, anstatt verteile Daten aufwendig an zentraler Stelle zu konsolidieren. Nur die für die weitere Verarbeitung relevanten Ergebnisse werden anschließend weitergeleitet. Somit müssen nur die Programme zu den Daten gebracht werden und nicht umgekehrt. Die Erstellung und Verwaltung solcher parallelisierten und verteilten Anwendungen ist jedoch keineswegs trivial und verlangt den Einsatz ausgereifter Technologien.
Eine bereits vielfach genutzte Lösung für verteilt arbeitende Software ist Apache Hadoop, benannt nach einem Spielzeugelefanten. Der Systemarchitekt Doug Cutting aus den USA hat sein Java-Framework Hadoop im Jahr 2003 ursprünglich entwickelt, um die Indizierung und Suche von Internet-Seiten zu optimieren. In dem System kommen zahlreiche von Google verwendete Algorithmen und Verfahren zum Einsatz, die teilweise auch veröffentlicht wurden. Damit hat sich Hadoop mit seinem über die Jahre entstandenen Ökosystem als eine Art Linux der Datenverarbeitung etabliert.
Hadoop verhält sich anders als eine service-orientierte Architektur (SOA), bei der zuerst teure Konzepte entwickelt und Infrastrukturkomponenten aufgebaut werden müssen, bevor ein wirtschaftlicher Nutzen erzielbar ist. Mit den Big Data-Technologien von Hadoop gelingt der rasche Aufbau eines Testprototypen, so dass IT-Abteilungen mit vergleichsweise geringem Aufwand erste verwertbare Ergebnisse erzielen können.

Breiter statt höher

Die bei Hadoop genutzte Architektur verwendet eine horizontale Skalierung (scale out vs. up). Bei herkömmlichen Datenbanksystemen erfolgt die Skalierung dagegen mit immer größeren und aufwändiger zu kontrollierenden Datenbank-Clustern. Zum Betrieb von Hadoop können Unternehmen kostengünstige Standard-Server verwenden und das Gesamtsystem bei Bedarf sehr einfach ergänzen, ohne die Gesamtarchitektur negativ zu beeinflussen.
Kern der bei Hadoop genutzten Big Data-Technologie ist der sogenannte Map-Reduce Algorithmus. Die weite Verbreitung der Lösung hat dazu geführt, dass zur Implementierung von Map-Reduce kaum noch manuelle Programmierarbeiten notwendig sind: Es gibt sowohl deklarative QL-Hochsprachen als auch grafische ETL-Werkzeuge zur Generierung von Programmen, die Map-Reduce verwenden.
Eine weitere Neuerung bei Hadoop ist die Art der Datenspeicherung. Anstatt Daten nur in einer Datei oder in einer Zeile einer Datenbank abzulegen, kann Hadoop Daten transformieren und intern bereits so speichern, dass sich diese direkt weiterverarbeiten lassen. Zusätzlich werden sowohl die Speicherung beschleunigt als auch der Speicherplatz durch geeignete Komprimierungsverfahren optimiert.
Die Leistungsfähigkeit von Hadoop zeigt sich eindrucksvoll in den Ergebnissen des TeraSort-Benchmarks der vergangenen Jahre. Vor fünf Jahren dauerte es mit traditionellen Verfahren noch Stunden, um die Häufigkeit der Wortvorkommen in einem ein Terabyte großem Text zu berechnen. Hadoop schaffte dies zu Beginn in drei Minuten. Heutige Hadoop-Cluster können bereits 100 Terabytes in nur zehn Sekunden sortieren. Diese Beispiele zeigen, welche enormen Leistungssprünge durch die neuen Technologien möglich sind. Neben der Geschwindigkeit ist aber auch die Kosteneffizienz ein wesentliches Argument für diese Technologie.

Ein Blick in die Glaskugel

Die Herausforderung des Datenwachstums ist so alt wie die IT selbst. Daher existieren hierfür auch keine Allheilmittel, sondern nur Medikamente, die die (Kopf-)Schmerzen der IT-Verantwortlichen lindern. Frameworks wie Apache Hadoop bieten Lösungen an, um neue Antworten auf alte Fragen zu finden. Als Defacto-Standard optimiert Hadoop einerseits die

Speicherung und Verarbeitung großer unstrukturierter Datenmengen in Echtzeit. Dabei kann es flexibel auf Datenformatänderungen und stark wachsende Datenvolumen reagieren. Außerdem lässt es sich sehr einfach in bestehende Datenbank- und Business Intelligence-Lösungen integrieren. Big Data ist daher keine Frage von entweder NoSQL oder SQL, sondern verbindet beide Welten.

Mit Hadoop werden die bei Google im Internet etablierten Algorithmen und Verfahren dank Open Source für alle verfügbar. Die auf Hadoop aufbauenden Distributionen wie Cloudera oder Hortonworks aber auch von den etablierten Herstellern, wie IBM oder Oracle, erleichtern den Einstieg und optimieren den Einsatz von Big Data-Lösungen. Aus diesen Gründen hat Hadoop in einer recht kurzen Zeit enorme funktionale Fortschritte gemacht und bereits eine große Verbreitung erreicht.

Die Schätze liegen im Keller

Präzise und aktuelle Informationen werden für Unternehmen immer wichtiger und in vielen Rechenzentren schlummern ungehobene Schätze in den Datenspeichern. Mit Big Data existiert eine Schatzkarte, um große Datenmengen zu analysieren und daraus konkreten Mehrwert zu erzielen.

Wer die Implementierung einer Big Data-Lösung plant, sollte im ersten Schritt bestehende Geschäftsprozesse analysieren und den Mehrwert ermitteln, der sich durch Big Data-Analysen ergibt, um nicht Nadeln in Heuhaufen zu suchen, in denen keine sind. Anschließend kann schrittweise das benötigte Expertenwissen im eigenen Haus aufgebaut werden. Noch fristet der Datenwissenschaftler gegenüber den traditionellen Rollen des Datenbankadministrators oder Business Intelligence-Spezialisten ein Schattendasein. Es ist jedoch nur eine Frage der Zeit, bis sich auch Big Data-Experten etablieren. Gleiches gilt für das Wissen und den Aufwand, optimal konfigurierte Hadoop-Umgebungen aufzubauen. Hier helfen vorkonfigurierte Appliances oder virtualisierte Umgebungen, die sogar in der Cloud verfügbar sind. So lassen sich die ersten Hürden schnell nehmen und Big Data wird ein nützliches Werkzeug für den Schatztaucher im Datenmeer. ■

Frank Pientka, Software-Architekt, Materna