Big, Fast & Furious

Marcel Amende, Michael Bräuer
ORACLE Deutschland B.V. & Co. KG

Schlüsselworte
Fast Data, Big Data, Complex Event Processing, Hadoop, IoT

Einleitung

Immer mehr Bereiche der Gesellschaft, des privaten Lebens und der Geschäftswelt werden digitalisiert und vernetzt. Informationsquellen senden kontinuierlich Daten. Dabei können solche Informationsquellen Geräte sein, die von Menschen bedient werden, oder Geräte, die für sich autonom, aber miteinander vernetzt arbeiten (Internet of Things - IoT).

Abb. 1: Digitale Vernetzung, das „Internet der Dinge" (IoT)

Durch die Innovation in den Bereichen Netzwerk-, Rechen- und Speichertechnologien der letzten Jahre können anfallende Umwelt-, Zustands und Bewegungsdaten mittlerweise kontinuierlich erfasst werden. Zum einen fallen dabei riesige Mengen von Daten an, deren Informationsgehalt zunächst unklar ist und sich erst aus der Kombination mit weiteren Datenquellen und Analysen ergibt. Zum anderen müssen Entscheidungen aus Mustern, die sich in den Datenströmen verbergen, in naher Echtzeit getroffen werden. In beiden Fällen muß Software helfen, diese Datenströme zu verarbeiten.

Mit diesen Herausforderungen der modernen Datenverarbeitung beschäftigen sich zwei Bereiche der IT, die unter den Akronymen „Big Data" und „Fast Data" bekannt sind.

Fast Data

Information verliert über die Zeit schnell an Wert, Einblicke in Unternehmensabläufe in Echtzeit sind für Unternehmen daher von großem Nutzen. Diese Fähigkeit, schnell auf Begebenheiten reagieren und möglicherweise sogar proaktiv handeln zu können, erhöht die „Operative Intelligenz" (OI) von Unternehmen.

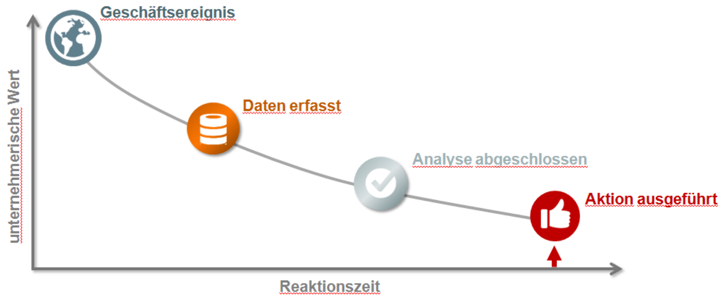

Abb. 2: „Ereignis-Aktions-Latenz“, abnehmender Wert der Infromation mit der Zeit

Dafür müssen hochvolumige (Ereignis-) Datenströme verschiedenster Quellen nahe ihrer Entstehung in Echtzeit verarbeitet, analysiert und mit historischen Daten verglichen werden. Man untersucht Beziehungen zwischen Ereignissen und wendet Verfahren zur Mustererkennung an, um „komplexe“ Ereignisse im Sinne eines Geschäftsprozesses zu erkennen: Läutende Kirchenglocken (1), ein Mann im Frack (2), eine Frau in einem weißen Kleid (3) und fliegender Reis (4) deuten auf das komplexe Ereignis einer Hochzeit hin.

Die Technologie zur Erkennung von Mustern in Ereignisdatenströmen wird als „Complex Event Processing“ bezeichnet. Im Gegensatz zur einfachen Ereignisverarbeitung z.B. mit einem Service Bus kann hier die zeitliche Abfolge einzelner Ereignisse in die Analyse einbezogen werden. Nachrichten aus verschiedensten Datenquellen und Datenströmen werden nahe der Ereignisgenerierung bzw. Entstehung erfasst. Sie werde miteinender in Beziehung gesetzt, um darin komplexe Muster zu erkennen und Geschäftseregnisse zu generieren, auf die nahezu in Echtzeit reagiert werden kann.

Die Regeln zur Mustererkennung werden dabei deklarativ, parametrisierbar und somit schnell änderbar in der „Continuous Query Language“ (CQL), einer Kombination aus SQL und regulären Ausdrücken, formuliert.

Big Data

Big Data beschreibt ein Konzept für das Sammeln und Verabeiten von großen Datenmengen, welche so groß und komplex sind, dass es schwer ist, die dabei entstehenden Herausforderungen mit herkömmlichen Strategien für die Datenverarbeitung zu lösen. Diese Herausforderungen umfassen das Erfassen, die Art und Weise der Speicherung, die Suche, die Analyse und Visualisierung der Daten, welche in verschiedensten Formaten vorliegen.

Oft ist bei der Erfassung der Daten noch nicht bekannt, wonach später gesucht werden soll. Man erfasst somit alle anfallenden Daten unter Zuhilfenahme kostengünstiger Speichermedien, um dann anschließend wiederkehrend beliebige Analysen auf den Rohdaten durchzuführen, die verschiedenste Fragestellungen beantworten können.

In den letzten Jahren haben sich Technologien, Algorithmen und Programmiermodelle entwickelt, die Lösungen für die Speicherung der Daten und deren effizienten Verarbeitung (u.a. Filterung, Sortierung, Aggregation) in verteilter und paralleler Weise ermöglichen. Der Treiber für diese Entwicklungen waren die enormen Anforderungen von Internetunternehmen, wie Google, Amazon und Facebook. Für die Umsetzung von Internetsuchindex, Vorschlagsystem und Social Graph müssen viele Petabytes von Rohinformationen immer wieder aufs Neue analysiert und verarbeitet werden. Die Rohdaten sind dabei oft unstrukturierte Daten, wie Webseiten, Einkaufslisten und Social Posts.

Die Verarbeitung der Daten muss dabei zeitlich unabhängig geschehen können. Die Verwendung von klassischer Datenbanktechnologien wäre bei großen Datenmengen ein sehr kostspieliger Luxus. Hier gibt es verschiedene Alternativen. Das Java-basierte Hadoop Dateisystem (HDFS) vernetzt hunderte kostengünstiger Server, die -mit einfachen Festplatten vollgepackt- Petabytes von Daten in großen Blöcken (mehrere Megabytes) unabhängig von ihrer Struktur speichern können. Ausfallsicherheit wird durch die Verteilung von Kopien der Datenblöcke im Hadoop Netzwerk erreicht. Diese Daten können nun von den Prozessorthreads der Server immer wieder und massiv-parallel nach neuen Schlüsselinformationen durchsucht (Map) und nach neuen Kriterien verdichtet (Reduce) werden.

Wie passen beide Konzepte zusammen?

Im Gegensatz zu Fast Data, wo Events (mit bekannter Struktur!) zur Analyse benutzt werden und die Auswertung eine Reaktion quasi in Echtzeit nach sich zieht, ist die Datensammlung bei Big Data vergleichsweise zeitintensiv und deren Analyse erfolgt in Jobs (Hadoop), also nicht in Echtzeit und meistens zeitversetzt! Die Fragestellung, die bei Fast Data analysiert werden soll ist vorher bekannt und wird entspechend modelliert (Abfragen, Datenstrukturen der Events). Bei Big Data werden u.U. erst nach Sammlung der Daten klar, welche Art der Fragestellungen/Analysen auf den Daten erfolgen.

Wenn Fast Data und Big Data miteinander kombiniert werden, erhält man ein kraftvolles Konzept, um Datenströme bestmöglich zu nutzen:

Fast Data kann als ein optionaler Schritt vor Big Data angesehen werden, um Daten zu filtern, zu Aggregieren und zu Geschäftsinformationen zu verdichten. Events und Ergebnisse in Fast Data werden so zu einer Quelle für Daten von Big Data Systemen.

Ebenso können Erkenntnisse aus Big Data Analysen genutzt werden, um Fast Data Algorithmen abzuleiten und zu parametrisieren. Big Data Systeme mit ihrer inhärenten Verarbeitungslatenz können so um Echtzeitalgorithmen und -Auswertungen ergänzt werden, um schnell reagieren zu können.

Im Idealfall entsteht so ein Kreislauf: Rohdaten werden in Big Data Systemen gesammelt. Wenn sich bestimmte Fragestellungen ergeben, können die Rohdaten immer wieder für spezifische Analyseprozesse herangezogen weden. Aus den Analyseergebnissen werden dann Algorithmen für die Echtzeitanalyse in Fast Data Systemen abgeleitet.

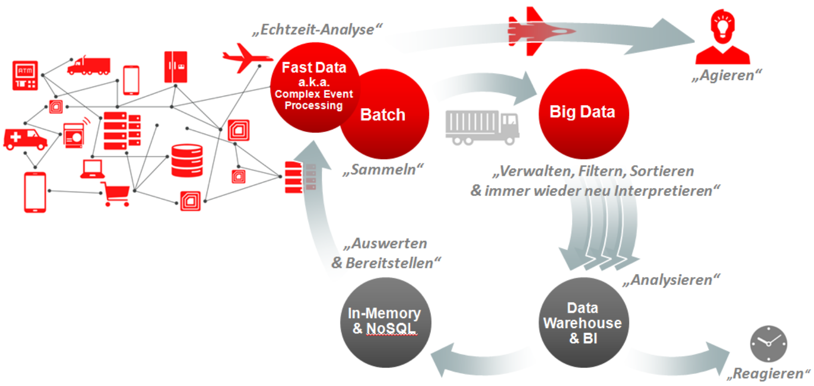

Abb. 2: Umfassender Datenanalysekreislauf mit Big Data und Fast Data

Ein Beispiel: Ein Maschinenbauer zeichnet sämtliche Kennzahlen (Drehzahlen, Temperaturen, Torsion, ...) seiner Systeme auf uns speichert sie unstrukturiert und kostengünstig in einer Big Data Lösung. Gibt es wiederholt Maschinenausfälle z.B. durch Lagerdefekte, wird in den aufgezeichneten Daten nach Auffälligkeiten schon vor Entstehung des fatalen Defekts gesucht. Gefunden wird z.B.

eine erhöhte Lagertemperatur bereits einige Minuten vor dem Totalausfall. Diese Erkenntnis kann in der Fast Data Echtzeitanalyse als Schwellwert genutzt werden, der auf einen bevorstehenden Defekt hindeutet. So wird es möglich, rechtzeitig die Maschine abzuschalten und eine kostengünstige Reparaturmaßnahme einzuleiten, bevor die komplette Maschine durch einen fatalen Defekt verloren geht.

Unser Vortrag

Im Vortrag werden die beiden Themengebiete Fast Data (Complex Event Processing) und Big Data (Hadoop Filesystem, Map/Reduce) konzeptionell vorgestellt und mit Blick auf ihr Zusammenspiel diskutiert. Es wird ein fachliches Beispiel eingegführt, das mit Oracle Software praktisch umgesetzt ist und die nahtlose Integration von Fast Data und Big Data Lösungen zeigt. In einer ausführlichen Demo wird der Mehrwert dieser integrativen und ganzheitlichen Lösung dargestellt.

Kontaktadresse:

Marcel Amende
ORACLE Deutschland B.V. & Co. KG
Hamborner Straße 51
D-40472 Düsseldorf

Telefon:	+49 (0)211 74839539
E-Mail:	marcel.amende@oracle.com
Internet:	www.oracle.com

Michael Bräuer
ORACLE Deutschland B.V. & Co. KG
Schiffbauergasse 14
D-14467 Potsdam

Telefon:	+49 (0)331 2007306
E-Mail:	michael.braeuer@oracle.com
Internet:	www.oracle.com