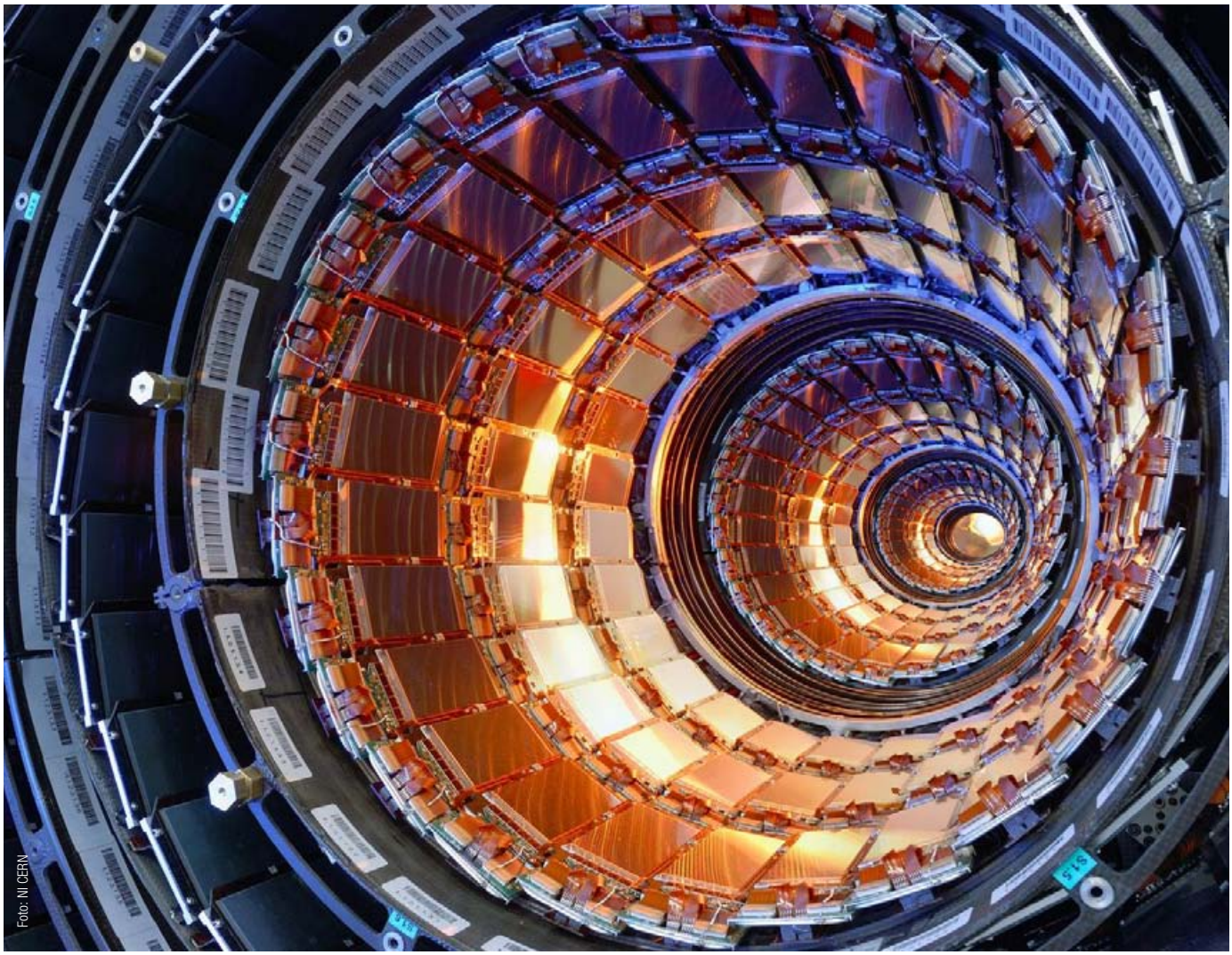

Teilchenbeschleuniger am Genfer Kernforschungszentrum CERN: Nachweis der Gottesteilchen mithilfe von Big Data-Analysen.

‣ HIGHEND-FORSCHUNG

Big Big Data.

Vorbild Wissenschaft: Die Analyse riesiger Datenmengen hilft als Instrument bei großen Durchbrüchen. Beispiele sind die Ergebnisse der CERN-Versuchsreihen oder die Forschungen zweier Nobelpreisträger. ›

› Von Daniel Liebhart

Die Motivation für Unternehmen, die Ausbreitung der innovativen Technologien und Methoden, die sich hinter dem Begriff «Big Data» verbergen, rasch vorabzutreiben, ist im Grunde relativ einfach. Ein vollständig informiertes Unternehmen trifft bessere Entscheidungen, kann effektiver, agiler und effizienter arbeiten und ist der Konkurrenz voraus. Vollständig informiert sein bedeutet, dass sämtliche Daten, die im Zusammenhang mit bestimmten Unternehmensaktivitäten wichtig sein könnten, zu handlungsrelevanten Informationen aufbereitet werden können: nicht nur wie bisher die strukturierten Daten aus den internen operativen Systemen, sondern darüber hinaus neu auch die schnell wachsende Menge der nichtstrukturierten Daten. Informationen, die in immer größerer Menge in immer höherer Geschwindigkeit erzeugt werden, in einer immer größeren Vielfalt vorliegen und zunehmend von unterschiedlichsten verifizierbaren Quellen bezogen werden. «Big Data» macht es möglich, die traditionelle Business Intelligence-Aufgabenstellung – die faktenbasierte Unternehmensführung – durch die Speicherung, Verarbeitung und Aufbereitung des gesamten Fundus ›

Chemienobelpreisträger Martin Karplus: Tiefere Erkenntnisse durch Computermodelle, die das reale Leben widerspiegeln.

an intern und extern vorhandenen Daten zu lösen. Aus diesem Grund sehen auch viele Hersteller, Analysten und Beratungshäuser den Aufbau der notwendigen Plattform als Erweiterung bestehender BI Infrastrukturen.

«Big Data»-Architektur als Erweiterung. In den vergangenen Jahren hat sich eine Vielzahl von Big Data-Referenzarchitekturen für die Verarbeitung von «Big Data» oder auch «Fast Data» etabliert. Jeder große Hersteller und jedes namhafte Beratungshaus hat seine eigene Referenz. Oracle empfiehlt die «Big Data & Analytics Reference Architecture»[1] als Erweiterung der «Information Management Reference Architecture»[2].

Der Anbieter IBM etwa entwickelt seine Lösungen der «Big Data & Analytics Reference Architecture»[3] als Vertiefung der «Business Analytics and Optimization Architecture»[4]. Microsoft hat die «Big Data Ecosystem Reference Architecture»[5] definiert, die in das NIST Big Data Programm eingeflossen ist. Teradata kombiniert seine beiden wichtigsten Produktlinien mit der «Big Data Reference Architecture»[6]. Die Lösungsansätze vieler Beratungshäuser basieren auf der «Logical Datawarehouse Architecture»[7] des Marktforschungsinstituts Gartner.

Die Liste ließe sich lange fortsetzen. Allen gemeinsam ist der nachvollziehbare Ansatz, der eine Erweiterung der klassischen Business Intelligence-Infrastruktur darstellt – bestehend aus den Datenquellen, Mechanismen zur Sammlung, der Aggregation und Aufbereitung dieser Informationen für das Data Warehouse, dem DWH selbst und einer Vielzahl von Analyseinstrumenten.

In der Umsetzung hat dieser Ansatz jedoch einen gravierenden Nachteil. Die Erweiterung bestehender, oftmals gut

› Daniel Liebhart ist Dozent für Informatik an der ZHAW (Züricher Hochschule für Angewandte Wissenschaften), Experte für Enterprise-Architekturen und Solution Manager der Trivadis AG. Er ist Autor und Coautor verschiedener Fachbücher.

funktionierender, aber über die Jahre gewachsener BI-Infrastrukturen ist für viele Unternehmen eine große Herausforderung. Nicht zuletzt angesichts der Produktvielfalt, zu deren Einsatz die Hersteller in vielen Fällen raten.

Eine Alternative zum «Big Data als BI-Erweiterung»-Ansatz kommt aus der Wissenschaft. Im Fokus steht die Gewinnung neuer Erkenntnisse durch den Einsatz geeigneter Instrumente. «Computermodelle, die das reale Leben widerspiegeln, sind entscheidend für die meisten Fortschritte, die heute in der Chemie gemacht werden»[8] sagte Staffan Normark, der Sekretär der Akademie anlässlich der Vergabe des Nobelpreises für Chemie an die Wissenschaftler Warshel, Karplus und Levitt im Jahr 2013.

«Wir haben in unseren Daten klare Anzeichen für ein neues Teilchen»[9] gab das CERN im Juli 2012 in einer sehr kurzen Pressemitteilung bekannt. Der mit großer Wahrscheinlichkeit verifizierte Nachweis des Gottesteilchens ist durch die Datenanalyse dieser CERN-Versuche gelungen, obwohl die Analysen immer noch im Gang sind.

In vielen Teilbereichen der Wissenschaft sind in den vergangenen Jahren signifikante Fortschritte durch die Analyse sehr grosser Datenmengen aus Experimenten oder Beobachtungen erreicht worden. Die Datenanalyse hat sich zu einem sehr wichtigen Instrument der modernen Forschungstätigkeit entwickelt.

Das Data Analytics Ecosystem.
Jack Dongarra und Daniel Reed, zwei renommierte Experten auf dem Gebiet der hochentwickelten Data-Processing-Technologie, leiteten aus der Entwicklung des Instrumentariums für die Datenanalyse ein einfaches generelles Modell ab: das «Data Analytics Ecosystem»[10].

Es basiert auf einem 4 Schichtenmodell, bestehend aus den Schichten Cluster, System, Datenveredelung (Middleware & Management) und Anwendung (Application) und ist für die verteilte Verarbeitung und die Analyse von sehr großen Datenmengen ausgelegt. Die Cluster-Schicht enthält die grundlegenden und eventuell virtualisierten Elemente Netzwerk, Speicher und Rechner. Die Systemschicht besteht aus dem Betriebssystem und virtuellen Maschinen.

Das Herzstück des Ecosystems aber sind die Bestandteile der Datenveredelungsschicht. Ein verteiltes Dateisystem und eine nichtrelationale Datenbank sind für die Speicherung sehr großer Datenmengen zuständig. Für das Laden, das Verschieben von Daten sowie für die Verarbeitung von gestreamten Daten (Sensordaten, Film, Ton und Daten aus sozialen Netzen) sind spezielle Instrumente definiert. Der Datenzugriff erfolgt mit Instrumenten für verteilte oder strukturierte Abfragen oder durch direkte Zugriffe auf Pipelines zur Datenverarbeitung. Darüber hinaus sind Komponenten für die Serialisierung von Daten und die Koordination von Datenströmen geplant. Die Anwendungsschicht sieht Mining, Statistik und spezialisierte Anwendungen vor. In der konkreten Ausprägung sehen Dongarra und Reed die in der Wissenschaft sehr oft eingesetzten Tools rund um das Apache Hadoop Framework vor.

Relevanz für Unternehmen.
Das Data Analytics Ecosystem ist für Unternehmen, die sich mit der oft unübersichtlichen Erweiterung der BI-Infrastrukturen schwertun, sehr nützlich. Die Einfachheit mit knapp zehn wesentlichen Komponenten und deren klarer Aufgabenteilung eignet sich sehr gut als Vorlage und Orientierungshilfe.

Gerade beim Aufbau einer Big Data-Plattform, also im Rahmen der Vorbereitungs- und Umsetzungsplanungsphase eines Big Data-Vorhabens lohnt es sich, den Werkzeugkasten der Wissenschaftler als Bereicherung zu nutzen. Auch wenn der konkrete Tooleinsatz für ein Unternehmen sich von demjenigen der Forscher natürlich unterscheiden wird. ■

---

[1] D. Chappelle: Big Data & Analytics Reference Architecture, Oracle White Paper, September 2013
[2] D. Cackett et Al: Information Management and Big Data, A Reference Architecture, Oracle White Paper, February 2013
[3] IBM: IBM Big Data & Analytics Reference Architecture V1, June 12, 2014 IBM Corporation
[4] IBM: BAO Reference Architecture, IBM GBS Business Analytics, & Optimization, 27 January 2011
[5] O. Levin: Big Data Ecosystem Reference Architecture, Microsoft Corporation, July 1, 2013
[6] http://thinkbig.teradata.com (letzter Aufruf 11.8.2015)
[7] M.A. Beyer, R. Edjlali: Understanding the Logical Datawarehouse: The Emerging Practice, Gartner, 21 June 2012
[8] http://www.nobelprize.org/mediaplayer/index.php?id=1956 (Letzter Aufruf 11.8.2015)
[9] CERN: CERN experiments observe particle consistent with long-sought Higgs boson, 04 July 2012
[10] D.A. Reed, J. Dongarra: Exascale Computing and Big Data: Communications of the ACM, July 2015, Vol. 38 / No. 7