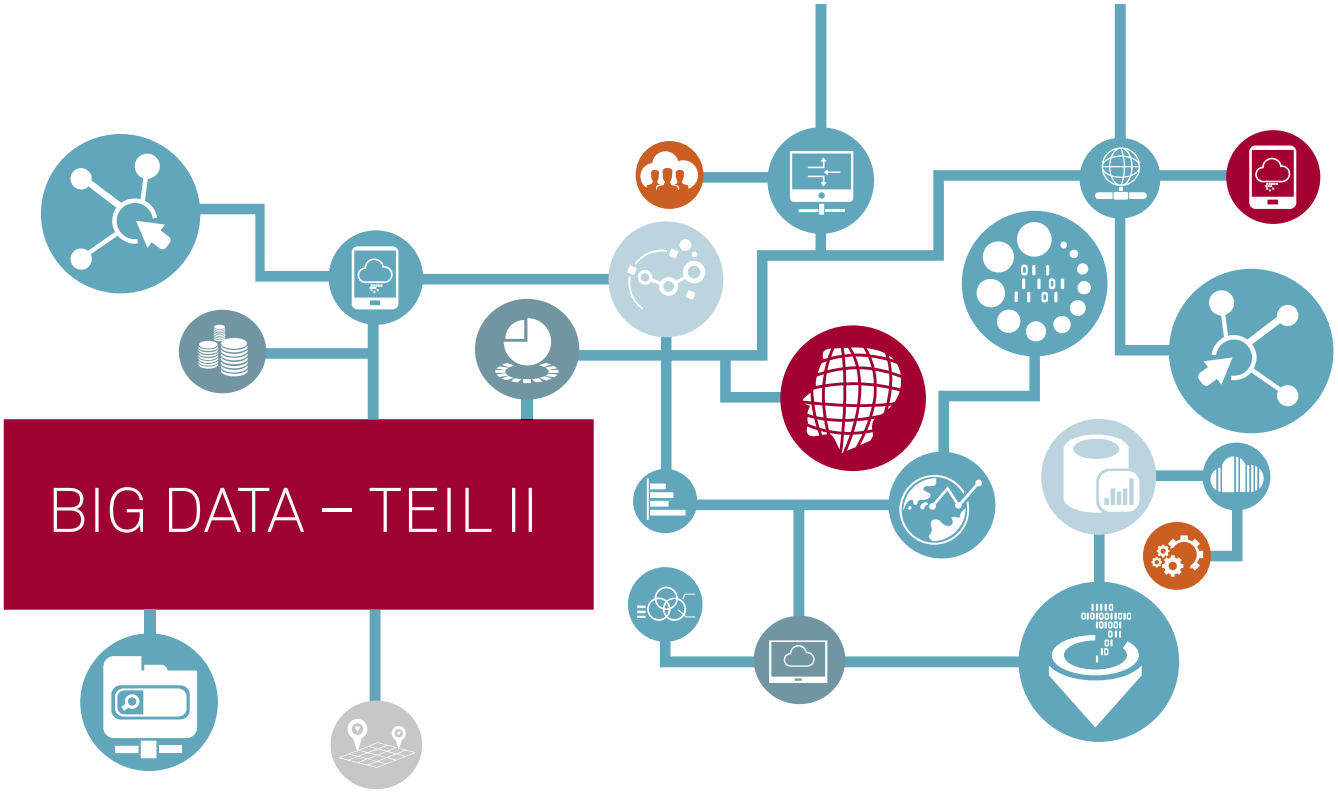

BIG DATA – TEIL II

Big Data ist sicher einer der meistdiskutierten IT-Trends der letzten Jahre und bewegt nicht nur die Businesswelt – große Internetkonzerne wie Google, Yahoo oder Facebook bauen ganze Geschäftsmodelle auf Big Data auf –, sondern zunehmend auch den öffentlichen Bereich. Die Impulse und Erfahrungen aus der Privatwirtschaft werden in Zukunft auch verstärkt die Erwartungen der Bürger an das digitale Angebot der öffentlichen Verwaltung beeinflussen.

| von DR. DIRK JÄGER und NEDISLAV NEDYALKOV

Im ersten Teil unserer Artikelserie Big Data[1] haben wir die spezifischen Rahmenbedingungen im öffentlichen Bereich untersucht und in Hinblick auf Einsatzmöglichkeiten von Big Data bewertet sowie, darauf aufbauend, die wesentlichen Potenziale und Risiken aufgelistet und deren Auswirkungen betrachtet.[2]

In diesem Teil gehen wir detailliert auf zwei Anwendungsbeispiele ein, zeigen die kritischen Erfolgsfaktoren für eine Big-Data-Lösung und beschreiben, wie der Weg zum ersten Big-Data-Projekt in der öffentlichen Verwaltung aussehen kann.

ANWENDUNGSBEISPIELE

Auch wenn der Ursprung von Big Data in der privaten Wirtschaft liegt, gibt es schon Anwendungsfälle aus dem öffentlichen Bereich. Ideen, die in anderen Ländern bereits erfolgreich umgesetzt wurden, können gegebenenfalls übernommen werden, wenn die spezifischen deutschen Rahmenbedingungen beachtet werden.

SCHWEIZER ARBEITSMARKTSTATISTIK

Seit 2007 läuft beim Staatssekretariat für Wirtschaft in der Schweiz das Projekt „Labor Market Data Analysis“ (LAMDA X). Im Rahmen des Projekts wurden BI-Anwendungen, wie zum Beispiel die offizielle Arbeitsmarktstatistik, mit teils komplexen Berechnungen auf Basis von Big Data realisiert.

Vor der Einführung der Big-Data-Lösung hatte sich herausgestellt, dass die bestehende technologische Basis und die Infrastruktur nicht ausreichend waren, um die Komplexität der nötigen Auswertungen abzudecken. Erhebliche Performan-

1 Man spricht von „Big Data“, wenn die Menge der zu verarbeitenden Daten, die Verschiedenheit der Daten oder die notwendige Geschwindigkeit der Verarbeitung zu groß werden, um sie mit klassischen IT-Architekturen zu handhaben.

2 .public 01-2015

ce-Probleme der eingesetzten Systeme waren häufig auf die herkömmliche Datenbankarchitektur zurückzuführen. Außerdem führte die gemeinsame Nutzung der Netzwerkinfrastruktur mit anderen Teilnehmern zu hohen netzwerkbedingten Latenzzeiten. Vor allem aus Performance-Gründen wurde daher nach einer Lösung im Big-Data-Umfeld gesucht, die unter anderem folgende Verbesserungen bieten sollte:

- erhöhte Flexibilität und Skalierbarkeit auch nach unten, da für Big Data relativ kleine Datenmengen von 500 GB verarbeitet wurden (darunter Arbeitsmarktstatistiken, Auszahlungsstatistiken der Arbeitslosenversicherungen, Führungskennzahlen und Arbeitsloseninformationen für die Öffentlichkeit),
- einen Betrieb auf preisgünstiger Hardware und die Anbindung heterogener Anwendergruppen mit unterschiedlichen Ansprüchen und Zielen

Die nun durch das Staatssekretariat gewählte Big-Data-Lösung besteht aus einer Hauptkomponente, die Arbeitsaufträge auf mehrere getrennte Rechenknoten mit Zugriff auf einen Teil der Daten verteilt (siehe Grafik). Die Proof-Of-Concept-Lösung wurde 2011 umgesetzt. Das bisherige Datenbanksystem wurde durch eine EMC-Greenplum-Database ersetzt. Dieses neue Datenbanksystem arbeitet mit massiv paralleler Verarbeitung. Dabei verteilt und steuert ein Masterserver die Verarbeitung auf beliebig vielen Datenbankknoten, die in der Arbeiterebene wiederzufinden sind. Diese Konstellation ermöglicht Flexibilität und Skalierbarkeit und ist nicht an bestimmte Hardwarekomponenten gebunden.

Der Hauptnutzen entsteht durch die gesteigerte Performance, die mit deutlich schnelleren Ergebnissen für die Nutzer eindeutig verbessert wurde. Außerdem wurde eine sehr viel schnellere und einfachere Duplizierung der Datenbestände für besondere Test- und Auswertungszwecke ermöglicht.

Von Bedeutung für die erfolgreiche Projektdurchführung war auch die evolutionäre Einführung von Big Data – eine attraktive Möglichkeit, in die Welt von Big Data einzusteigen. In diesem Fall bedeutet das, dass die bestehenden Anwendungen und Auswertungen beibehalten wurden und auf der Basis verteilter Datenbankkonzepte, also der neuen Konstellation mit dem Masterserver und den Arbeiterknoten, beschleunigt wurden. Dies öffnet Perspektiven für neue Auswertungen und Algorithmen, die mit der abgelösten technologischen Basis in diesem Umfang nicht möglich waren.

BIG DATA IN NEW YORK

Ein weiteres Beispiel, wie Big Data die Effizienz der öffentlichen Verwaltung steigern kann, stammt aus New York. Da Wohnraum in New York knapp und teuer ist, besteht für Vermieter ein großer Anreiz, Wohnungen ungenehmigt in kleinere Einheiten aufzuteilen und diese einzeln zu vermieten. Durch viele Menschen auf engem Raum drohen aber nun hygienische Probleme, Lärmbelästigung und erhöhte Brandgefahr durch überlastete elektrische Leitungen und verbaute Fluchtwege. Auch stadtplanerische und soziale Maßnahmen wie die Planung von Schulen werden durch eine „fehlerhafte" Bevölkerungsstatistik erschwert.

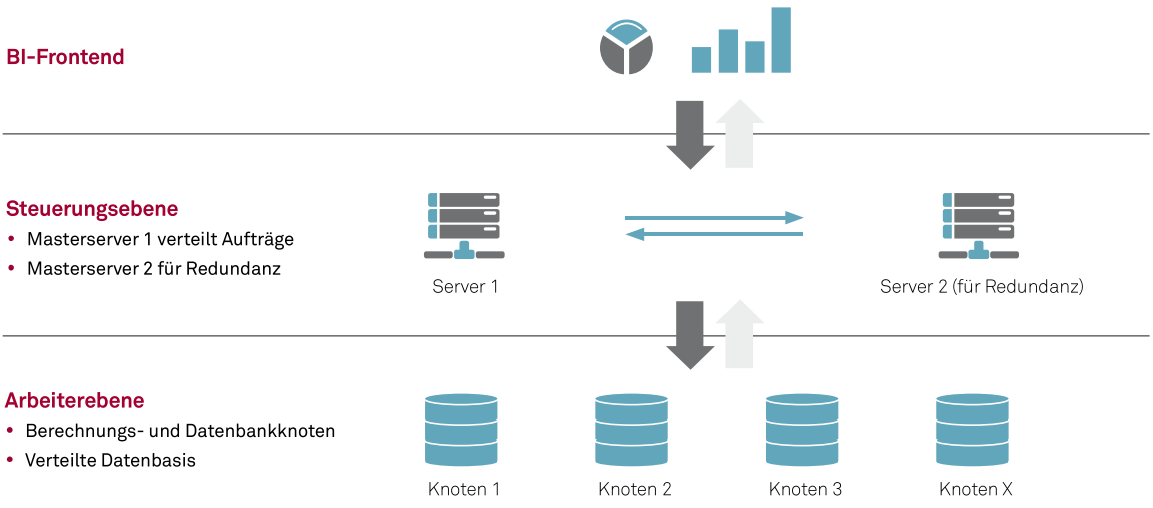

Abbildung 1: Beim Staatssekretariat bevorzugte Big-Data-Lösung

Die Stadt New York geht Beschwerden über illegale Wohnungsaufteilungen mit städtischen Inspektoren nach. Täglich trifft eine Vielzahl von Beschwerden ein. Die wenigen Inspektoren müssen entscheiden, welchen davon sie vordringlich nachgehen. Finden sie vor Ort dann tatsächlich eine illegal aufgeteilte Wohnung vor, ordnen sie die Räumung an. Bis zum Jahr 2011 geschah dies in 14 Prozent der durch Inspektoren vor Ort überprüften Wohnungen.

Im Jahr 2009 startete ein Big-Data-Projekt, das nach besonderen Merkmalen dieser Fälle suchte. Dazu wurden die Daten von zurückliegenden Räumungen mit Daten aus 19 verschiedenen Quellen der Stadt verknüpft, wie zum Beispiel die Lage der Wohnung, Alter der Bausubstanz, Verbrauchsdaten der Energieversorgung, Fälle von Schädlingsbefall und Häufigkeit von Polizeieinsätzen. Aus der Verknüpfung dieser Daten entwickelte ein Team von Datenanalysten ein Vorhersagemodell, das die Wahrscheinlichkeit berechnet, dass eine eintreffende Beschwerde von Nachbarn berechtigt ist. Seit 2011 wird auf Basis dieses Modells entschieden, welchen Beschwerden vorrangig nachgegangen wird. Die Trefferquote der Inspektionen stieg dadurch von 14 Prozent auf 70 Prozent.

ERMITTLUNG DES BESTEN VORHERSAGEMODELLS

Um das beste Vorhersagemodell zu ermitteln, rechneten die Datenanalysten in New York sehr viele Modellvarianten für alle bereits zurückliegenden Fälle durch. Jede Modellvariante verknüpfte dabei auf eine bestimmte Art die Daten des Falls mit allen verfügbaren Daten aus den anderen 19 Datenquellen. Durch Ausprobieren einer Vielzahl von Varianten wurde so das Modell mit der größten Vorhersagekraft identifiziert.

Das Durchrechnen vieler Modelvarianten zur Ermittlung des besten Vorhersagemodells ist charakteristisch für Big Data. Es erfordert, dass Berechnungen hochgradig parallel und auf großen, unterschiedlich strukturierten Datenmengen aus verschiedenen Quellen ausgeführt werden. Big-Data-Technologien, wie zum Beispiel NoSQL-Datenbanken, bilden hierfür die technische Grundlage. Die spätere Anwendung des Vorhersagemodells erfordert dagegen deutlich weniger Rechenleistung, da nur noch die Daten eines konkreten Falls mit dem einmal gefundenen Modell durchgerechnet werden.

DER WEG ZU BIG DATA

Die Konzepte und Technologien von Big Data ermöglichen heute die Entwicklung völlig neuer Anwendungen und deren Skalierung auf immer größere Datenvolumina. Damit ist jedoch auch ein Paradigmenwechsel verbunden, der gleichermaßen Anwender wie auch Entwickler und Betreiber von IT-Systemen betrifft: Vor allem müssen Organisationen die Hürden für den Zugriff auf Daten anderer Fachbereiche und Nutzer (Bürger) abbauen. Fachanwender müssen ihre Datenbestände als Quelle neuer Erkenntnisse wahrnehmen und diese mit Datenanalysten und anderen Fachanwendern teilen. Außerdem müssen Software-Entwickler neue Technologien, Architekturen und Systeme entwickeln und beherrschen, um Big-Data-Anwendungen zu realisieren. Der IT-Betrieb muss eine Infrastruktur für diese Anwendungen aufbauen, die stark von der klassischen Systemlandschaft aus Applikationsservern und relationalen Datenbanken abweicht.

Um allen Beteiligten Zeit für die erforderlichen Lern- und Veränderungsprozesse zu geben, ist es sinnvoll, in kleinen Schritten vorzugehen. Das erste Big-Data-Projekt sollte ein überschaubares und konkretes Ziel besitzen, das einen unmittelbaren Nutzen verspricht. Das Ziel kann technisch motiviert sein, beispielsweise, wenn die Kapazität eines bestehenden Verfahrens nicht mehr ausreicht, um die anfallenden Daten zu verarbeiten. Oder es kann sich um einen fachlichen Anwendungsfall handeln, in dem die (neue) Verknüpfung bestehender Daten einen erheblichen Mehrwert bringt und sich ein entsprechender Nutzen rechnen lässt.

Sofern es um die Verknüpfung von fachlichen Daten geht, ist die Analyse der bestehenden Datenbestände der erste Schritt, um herauszufinden, welche Daten überhaupt vorliegen und wie sie in Beziehung gesetzt werden können. Dabei sind auch die rechtlichen Rahmenbedingungen zu klären. Dieser erste Schritt ist unerlässlich, um zu vermeiden, dass ein Software-Projekt begonnen und entsprechende Hard- und Software sowie eine Infrastruktur beschafft werden, ohne dass im Vorfeld geklärt wurde, dass die entsprechenden Big-Data-Ziele mit dem vorliegenden (oder einem zu erhebenden beziehungsweise erhebbaren) Datenbestand überhaupt erreicht werden können.

Da Big-Data-Projekte aufgrund ihrer sehr speziellen mathematischen, algorithmischen und technischen Verfahren nicht einfach nur „Datenbank-Projekte mit vielen Daten“ sind, ist es sinnvoll, für ein erstes Big-Data-Projekt einen Beratungspartner auszuwählen, der entsprechende Erfahrungen nachweisen kann.

Ist man nicht an einen bestimmten Anbieter gebunden, können Open-Source-Lösungen wie zum Beispiel Apache Hadoop einen leichtgewichtigen Einstieg in das Thema Big Data ermöglichen. In der Pilot-Phase fallen damit zunächst keine Lizenzkosten an. Um die erste technische Hürde zu überwinden, sollte der Beratungspartner bereits Erfahrungen mit der jeweiligen Plattform besitzen.

Der Einsatz kommerzieller Systeme hingegen bietet sich für ein erstes Projekt an, wenn bereits eine Systemlandschaft dieses Herstellers in Betrieb ist. Entsprechende Lösungen haben alle großen Hersteller von Datenbanken und ERP-Systemen in ihrem Portfolio. Die unmittelbare Integration mit den vorhandenen Systemen ist dabei der große Vorteil.

FAZIT – MEHR EFFIZIENZ IM ÖFFENTLICHEN BEREICH

Big Data schafft nicht nur für Unternehmen, sondern auch für den öffentlichen Bereich einen Mehrwert. Dies geschieht z. B. durch die Erweiterung existierender Verfahren auf größere Datenmengen, die nur mit Big-Data-Technologien verarbeitet werden können, oder durch eine neuartige Verknüpfung vorhandener Daten. Auch wenn die Big-Data-Geschäftsmodelle kommerzieller Unternehmen nicht ohne Weiteres auf öffentliche Stellen übertragbar sind, so sind es doch die Verfahren und Technologien, die von diesen Unternehmen im Bereich Big Data entwickelt wurden.

Für innovative Lösungen müssen zunächst die vorhandenen Datenbestände verfügbar gemacht und durch geeignete Experten analysiert werden. Dabei sind die spezifischen Rahmenbedingungen des öffentlichen Bereichs, wie zum Beispiel Zweckgebundenheit, Datenschutz und das Vertrauensverhältnis zu den Bürgern, zu beachten.

Durch Big-Data-Anwendungen lassen sich dann im öffentlichen Bereich erhebliche Effizienzsteigerungen und Kosteneinsparungen erzielen. Beispielprojekte dazu sind vorhanden. Der Einstieg in neue Technologien, die Big Data mit sich bringt, ist zunächst eine Hürde. Realistische Zielsetzungen und die Auswahl eines Pilotprojekts, das ein konkretes Problem löst, helfen, diese Hürde zu überwinden. ●

ANSPRECHPARTNER – NEDISLAV NEDYALKOV

IT Consultant

Public Sector

- +49 151 14863087
- nedislav.nedyalkov@msg-systems.com

WEITERFÜHRENDE LITERATURHINWEISE

BITKOM (Hrsg.). „Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte". BITKOM Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V., 2012

Klaus-Peter Eckert, Lutz Henckel, Petra Hoepner. „Big Data – Ungehobene Schätze oder digitaler Albtraum". Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT, 1. Auflage, März 2014

Viktor Mayer-Schönberger und Kenneth Cukier. „Big Data – Die Revolution, die unser Leben verändert". Redline Verlag, München, 2013

Dr. Andreas Schäfer, Dr. Melanie Knapp, Dr. Michael May, Dr. Angelika Voß. „Big Data – Vorsprung durch Wissen. Innovationspotenzialanalyse". Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, 2012