‣ PROTOTYPING

Small Big Data.

Sind Investitionen in Big Data ein zu großes Risiko? Durch vertikales Prototyping lassen sie sich besser kontrollieren. Und die technischen Einstiegshürden können «spielend» überwunden werden.

› Von Eike Straehler-Pohl und Timo Linde

Erinnerung an das WM-Sommermärchen 2006. Elfmeterschießen im Viertelfinalkrimi Deutschland gegen Argentinien. Deutschland gewinnt 5:3. Und ein kleiner Spickzettel geht in die Geschichte des Fußballs ein. Eben jener Zettel, auf dem Torwarttrainer Andreas Köpke die Elfmeter-Statistiken der gegnerischen Schützen notiert und dem deutschen Keeper Jens Lehmann kurz vor Beginn des Schießens zugesteckt hatte.

Der tatsächliche Einfluss dieses kleinen Zettels auf das Spielergebnis wurde im Nachhinein leidenschaftlich diskutiert. Nicht wegzudiskutieren ist hingegen die rasant wachsende Bedeutung von Analytics im Profisport der letzten 10 Jahren. Mithilfe von GPS-Trackern werden Laufwege und Spielmuster analysiert und optimiert. Kontinuierlich erfasste Herzfrequenzdaten der Spieler erlauben individuell abgestimmte Fitness-Einheiten. Aktuell werden Körpersensoren entwickelt, mit deren Hilfe sich beispielsweise Frühwarnsysteme für Verletzungen realisieren lassen – ganz ähnlich wie es Predictive Analytics bereits für den Betrieb von Anlagen leistet. Big Data ist im Profisport angekommen und wird dieses Jahr für die Teams der EM ein wichtiger Erfolgsfaktor sein.

Und die europäischen Unternehmen? Hier ist die Entwicklung nicht ganz so «sportiv». Die Mehrzahl der Unternehmen hat die Nutzung von Big Data zwar als möglichen Wettbewerbsvorteil erkannt, jedoch nur ein kleiner Teil von ihnen hat die dafür notwendigen Systeme bereits in die Unternehmensprozesse integriert. Als die größten Einstiegshürden gelten gemeinhin eine mangelnde Erfahrung mit der entsprechenden Technologie, Unsicherheit in Bezug auf den erzielbaren Nutzen und nicht zuletzt das Investitionsrisiko. Big Data gleich Big Investment gleich Big Risk, das befürchten viele Unternehmen. Doch diese Hürden lassen sich mithilfe von geeignetem Prototyping überwinden.

Bewegungsdigitalisierung.

In der Produktentwicklung gehören Modellbau und Prototyping nach wie vor zum Standardvorgehen. Mit dem Ziel, das Risiko und die Kosten zu senken, wird die Realität auf entscheidende Kernprobleme reduziert. Eine pragmatische Nachbildung, häufig im verkleinerten Maßstab, hilft dabei, potentielle Probleme aufzudecken, mögliche Lösungen zu testen und die Funktionsweise im Vorfeld besser zu verstehen. Die grundsätzliche Machbarkeit eines Produkts und die zu erwartenden Entwicklungskosten lassen sich frühzeitig einschätzen. So kann das Investitionsrisiko vor Beginn der eigentlichen Umsetzung erheblich gesenkt werden.

Dieser Ansatz lässt sich auch auf Big Data-Projekte übertragen. Mit einem vertikalen Prototyp, der möglichst viele Bausteine einer Big Data-Architektur nutzt, um einen konkreten Anwendungsfall zu lösen, kann die Funktionsweise des Systems aus unterschiedlichen Blickwinkeln beurteilt werden: Problemstellungen der Umsetzung aus Entwicklersicht, Fragestellungen der zukünftigen Nutzung aus Anwendersicht, Chancen- und Risikobewertung des Managements.

Zur anschaulichen Demonstration, wie ein solcher Prototyp dazu beitragen kann, ein geplantes Big Data-Projekt hinsichtlich seiner Realisierbarkeit besser zu beurteilen, haben wir unseren eigenen Kickertisch im Büro digitalisiert. Dieses Beispiel ist so branchenunabhängig wie möglich und es passt inhaltlich gut zur anstehenden EM 2016.

Das definierte Ziel war, den Tisch mit entsprechender Technik so auszustatten, dass bereits während eines Spiels aufschlussreiche Statistiken wie der Ballbesitz, maximale Schussgeschwindigkeit und der Spielstand in einem Live Dashboard angezeigt werden. So kann ein Spieler noch während der laufenden Partie reagieren und seine Taktik anpassen.

Darüber hinaus sollte das System Langzeitanalysen ermöglichen und beispielsweise Auswertungen aller bereits absolvierten Spiele liefern. So können unter anderem strategische Trainingsparameter erarbeitet werden. Diese Aufgabe ›

› Eike Straehler-Pohl ist Managing Consultant bei blueforte und berät Führungskräfte bei der strategischen Planung von BI-Systemen mit innovativem Schwerpunkt. Darüber hinaus verantwortet er bei blueforte den Bereich Business Development & Innovation. Straehler-Pohl studierte Wirtschaftsinformatik an der Nordakademie. eike.straehler-pohl@blueforte.com

› Timo Linde ist als Management Consultant im Geschäftsfeld Visual Business Analytics tätig und berät Mitarbeiter aller Ebenen bis hin zum Management bei der Umsetzung von Visualisierungskonzepten für moderne Reporting- und Analyseanwendungen. Linde studierte im Hochschulübergreifenden Studiengang Wirtschaftsingenieurwesen Hamburg (HWI). timo.linde@blueforte.com

Musterablauf: Digitalisierter Kickertisch.

erfordert die Verwendung einer Videokamera, einiger Sensoren, eines Micro-Controllers, eines Test-Server und einer Reihe von Softwarekomponenten.

Die Videokamera zeichnet oberhalb des Tisches sämtliche Bewegungen auf dem Spielfeld auf. Eine Video-Analytics-Engine transformiert das entstandene Bildmaterial in statistisch auswertbare Daten. Zusätzlich im Tor eingesetzte Lichtschranken verifizieren die optischen Daten bei kritischen Torentscheidungen. Die von der Kamera und den Sensoren aufgezeichneten Daten werden sowohl als Live-Stream verarbeitet als auch für Langzeitanalysen dauerhaft gespeichert.

Dabei repräsentiert die hier eingesetzte Technik zahlreiche Anwendungsfälle aus der Praxis: Video Analytics etwa wird eingesetzt, um Qualitätsmängel in der Produktion zu identifizieren, Bewegungsmuster von Systemen zu erkennen oder Logistikprozesse wie das Sortieren von Gütern zu unterstützten.

Durch die Nutzung von Sensordaten werden immer mehr Geräte des Alltags «intelligent» – Stichwort Internet of Things (IoT). Überdies lässt sich die Funktionsweise von Big Data gut anhand der Kriterien Velocity, Volume und Variety mit Video- und Sensordaten demonstrieren.

Velocity, Volume, Variety.

Big Data wird häufig durch die Merkmale Velocity (Verarbeitungsgeschwindigkeit), Volume (Datenmenge) und Variety (Vielfältigkeit der Daten) charakterisiert. Mithilfe des Prototyps können diese Merkmale sehr gut analysiert werden – es lässt sich genau beobachten, welche Herausforderungen sich im Vergleich zu klassischen Data-Warehouse-Konzepten ergeben. Die am Prototyp gemessenen Performance-Werte, wie die benötigte Speicherkapazität oder Verarbeitungsgeschwindigkeit, lassen sich auf das spätere Zielsystem hochrechnen. Sie erlauben somit eine realistische Planung der benötigten Systemkonfiguration.

Was bedeutet Velocity? Im Prototyp wird eine Webcam verwendet, die etwa 30 Bilder pro Sekunde in HD-Qualität aufzeichnet. So werden pro Sekunde ca. 200 Mio. Datenwerte erzeugt. Bei Profi-Tischfußballern erreicht der Ball Geschwindigkeiten bis zu 30 km/h. Wollte man ihn lückenlos erfassen, wäre sogar eine Spezialkamera mit einer Leistung von mehr als 200 Bildern pro Sekunde nötig. Das entspräche einer Verarbeitungsgeschwindigkeit von 1,3 Mrd. Datenwerten pro Sekunde – ein enormer Datenstrom, aus dem in Echtzeit Live-Informationen für ein Dashboard gewonnen werden. Der Prototyp setzt dafür einen sogenannten Speed-Layer ein, der eine Realtime-Verarbeitung des Bildmaterials ermöglicht.

Was bedeutet Volumen? Neben der absoluten Menge an Daten verdeutlicht das Verhältnis von Rohdaten zu Nutzdaten den Paradigmenwechsel bei Big Data und den daraus wachsenden Speicherbedarf sehr anschaulich. Während im klassischen Data Warehouse nutzungsoptimierte Speicherkonzepte verfolgt werden, basieren Big Data-Architekturen auf langfristig persistierten Rohdaten, sodass stets alle Ursprungsinformationen erhalten bleiben. Diese Speicherform wird oft als Data Lake bezeichnet.

Im Beispiel wird dieser Data Lake genutzt, um das gesamte Videomaterial sowie die Sensor- und Stammdaten zu sichern. Der Prototyp erzeugt selbst bei einer starken Video-Komprimierung etwa zwei Megabyte Daten pro Sekunde. Um die Ballposition im gleichen Zeitabschnitt zu bestimmen, sind lediglich 0,1 Kilobyte unkomprimierter Daten notwendig.

Das Rohdaten-Nutzdaten-Verhältnis beträgt somit etwa 1 : 20'000. Da der Prototyp aktuell auch die Position sämtlicher Spieler und deren Neigungswinkel liefert, wozu 1,5 Kilobyte an Nutzdaten pro Sekunde nötig sind, ergibt sich bei gleicher Rohdatenmenge ein Verhältnis von etwa 1 : 1'300. Umgekehrt bedeutet das, dass die Speicherung der Rohdaten im Data Lake ungefähr 1'300-mal mehr Speicher benötigt als in klassischen DWHs.

Was heißt Variety? Die Vielfältigkeit der Daten beim Kicker ergibt sich aus der Mischung sehr unterschiedlicher Quellen: Stamm-, Sensor- und Videodaten. Gerade Letztere sind hochgradig unstrukturiert. Um von Farbwerten und einzelnen Pixeln, über das zusammengesetzte Bild bis hin zur Information, an welcher Position sich der Ball befindet, zu schließen, sind hochkomplexe mathematische Transformationen aus dem Gebiet der Computer Vision (CV) notwendig.

Diese Berechnungen gehen weit über das hinaus, was klassische ETL-Prozesse leisten. Ebenso ist das Transformationsergebnis stets mit Unsicherheit behaftet. Ähnlich dem menschlichen Auge lassen sich die meisten CV-Algorithmen täuschen und oft reichen die Bildinformationen nicht aus, um ein Objekt fehlerfrei zu erkennen.

So liegt die Erkennungsquote des Balls im Speed-Layer je nach Lichtverhältnissen bei rund 95 Prozent. Mit komplexeren Verfahren, zum Beispiel unter Verwendung von lernenden Algorithmen, lassen sich jedoch deutlich genauere Ergebnisse erreichen. Dafür nutzt der Prototyp einen sogenannten Batch-Layer. Mit mehr Rechenzeit und der Möglichkeit zur parallelen Verarbeitung werden dabei die Berechnungen als Batch-Prozesse auf den gespeicherten Rohdaten ausgeführt. Auch eignet sich der Batch-Layer hervorragend für den Einsatz von Advanced-Analytics-Werkzeugen, sodass sich sogar Spielmuster vorhersagen lassen.

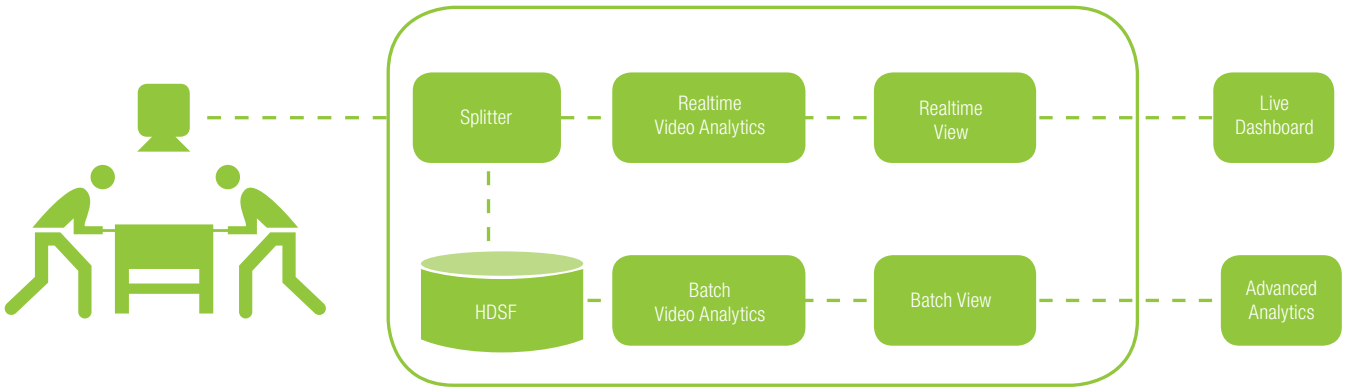
Lambda, Hadoop, Cloud.

Die Fachbegriffe Speed-Layer, Batch-Layer oder auch Data Lake sind nicht zufällig gewählt. Tatsächlich orientiert sich die Umsetzung des Prototyps an der von Nathan Marz vorgeschlagenen Lambda-Architektur, die sich aus diesen Bausteinen zusammensetzt. Die Lambda-Architektur eignet sich gerade für Anwendungsfälle mit sehr heterogenen Anforderungen an die Verarbeitungsgeschwindigkeit und Informationsqualität bei ein und denselben Rohdaten. In der Praxis basieren immer mehr Big Data-Projekte auf dieser Lambda-Architektur.

Beim Prototyp wurde der Speed-Layer als proprietäre Lösung (C++) umgesetzt, die das Videosignal zunächst für den Batch- und Speed-Layer splittet, um dann in Echtzeit die Ball- und Spielerposition zu berechnen. Die ermittelten Werte werden fortlaufend in eine Datenbank geschrieben und stehen der Visualisierungskomponente unmittelbar zur Darstellung im Dashboard zur Verfügung. Der Endanwender erhält so ständig aktuelle Informationen zum laufenden Spiel in Echtzeit.

Parallel zu diesem Verarbeitungsprozess werden die Videodaten als Rohdaten in einem Data Lake gespeichert, der als Teil des Batch-Layers auf Hadoop/HDFS basiert. Die analytischen Berechnungen werden anschließend in einem Batch-Pro-

Die Small Big Data-Architektur.
Die schlanke Struktur kann Unternehmen in der Praxis viel Aufwand und Zeit ersparen.



Quelle: blueforte

zess unter der Verwendung des Spark-Frameworks durchgeführt. Mithilfe von Spark lassen sich Bild- und Videodaten massiv parallel verarbeiten, sodass auch sehr große Videobestände bei entsprechender Dimensionierung des Hadoop-Clusters in kurzer Zeit berechenbar sind.

Diese Bausteine wurden in einer Hadoop-Sandbox umgesetzt. Die bekannten Anbieter von Hadoop-Distributionen (zum Beispiel Cloudera, Hortenworks, MapR) bieten in der Regel kostenlose vorkonfigurierte Systeme als virtuelle Maschinen an. Mit diesen lässt sich ein Prototyp hervorragend umsetzen und zeitlich meist unbegrenzt testen.

Alternativ oder sogar ergänzend kann das System in eine Cloudumgebung migriert werden. Mittlerweile gibt es verschiedene Serviceanbieter, die geeignete Lösungen anbieten. So lassen sich die Skalierbarkeit des Systems testen und Aussagen über die tatsächlich zu erwartende Performance bei einem überschaubaren Kostenrahmen treffen.

Nutzen, Kosten, Fazit.
Die Erstellung des Prototyps erlaubt es dem Team, zahlreiche Bausteine einer typischen Big Data-Architektur umzusetzen, real zu testen und wertvolles Wissen aufzubauen. So wurde beispielsweise die notwendige Realtime-Fähigkeit des eingesetzten Analytics-Tools arcplan geprüft. Der Test ergab, dass mit einer geringen Anpassung der standardmäßig erstellten Applikation eine automatische Aktualisierung des Dashboards im Sekundentakt möglich ist – eine wichtige Erkenntnis, wollte man ein solches Werkzeug für den späteren Praxiseinsatz nutzen.

Die Kosten für den Prototyp sind recht überschaubar geblieben. Mit einer Hard- und Softwareausstattung zum Preis von weniger als 1'000 Euro und einem Umsetzungsaufwand von weniger als zwei Mannmonaten machen die Kosten nur einen Bruchteil eines realen Big Data-Projekts aus.

Obwohl die Chancen von Big Data vielen Unternehmen bewusst sind, werden die vermeintlichen Einstiegshürden oft gescheut. Durch Prototyping können bei einem geringen Kosten- und Zeitaufwand fehlendes Wissen aufgebaut sowie die Umsetzbarkeit und der Nutzen eines Projekts aufgezeigt werden. Bei geringem Risiko werden Erfahrungen gesammelt und mögliche Fallstricke früh erkannt.

Der hier beschriebene Prototyp eignet sich aufgrund seines einfachen und nachvollziehbaren Aufbaus sehr gut, ein Big Data-Anwendungsszenario «spielend» zu erleben. Die eingesetzten Technikkomponenten sowie ihr Zusammenwirken erlauben es, die Funktionsweise auch technisch weniger versierten Anwendern einfach zu erklären.

Dieser Aspekt macht den Aufbau eines Prototyps besonders interessant für Unternehmen: Dank des «Live-Erlebnisses» lässt sich die Investitionsentscheidung leichter begründen. Steht der Prototyp, lässt sich die Technologie schnell hochskalieren und auf konkrete Anwendungsfälle in der Praxis übertragen. ■

Literatur:

Miller, T. W.: Sport Analytics and Data Science – Winning the Game with Methods and Models, Pearson Education, 2015

Marz, N.: Big Data – Principles and best practices of scalable real-time data systems, Manning Publications Co., 2015

Priese, L.: Computer Vision – Einführung in die Verarbeitung und Analyse digitaler Bilder, Springer Vieweg, 2015

Schmaker, R .P.: Sports Data Mining, Springer, 2010