Datenqualität

Die Güte von Big Data steuert der Business Case

In Transaktionssystemen gibt es klare Standards für die Datenqualität. Bei Predictive Analytics und Big Data entscheidet der Business Case darüber, wie genau, vollständig oder aktuell die Datenbasis sein muss.

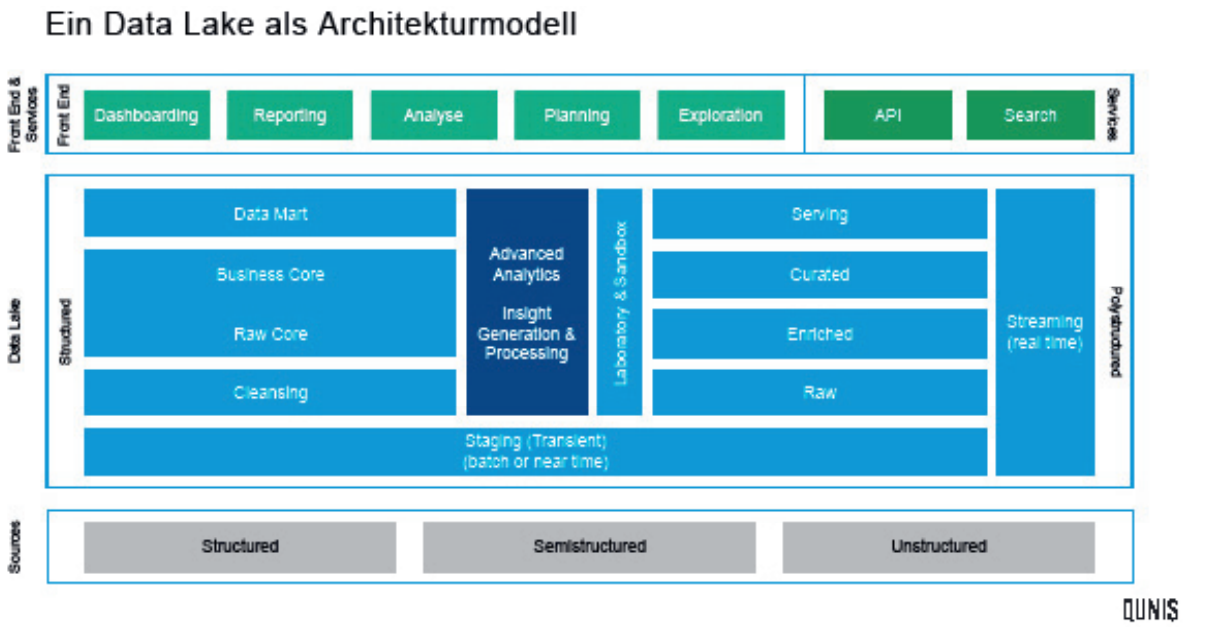
DER DURCH KLASSISCHE Business-Intelligence-Systeme (BI) erzeugte „Blick in den Rückspiegel" reicht Firmenlenkern nicht mehr aus. Unternehmen interessieren sich zunehmend für die auf fortgeschrittenen Analytics-Methoden beruhenden Vorhersagemodelle der Predictive Analytics. Laut der Studie Predictive Analytics 2018 der International Data Group gehen 66 Prozent der befragten Unternehmen in Deutschland, Österreich und der Schweiz davon aus, dass Predictive Analytics binnen drei Jahren für sie wichtig wird. Einige Unternehmen haben bereits derartige Projekte gestartet. Als eine der größten Herausforderungen gilt dabei die mangelnde Datenqualität. Das ist nicht verwunderlich, denn schon im klassischen BI-System sind Datenmängel ein Dauerthema.

In einer klassischen BI-Umgebung lagern typischerweise strukturierte Daten aus internen Vorsystemen wie Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) oder der Buchhaltung. Für die Qualitätssicherung gibt es Best Practices und erprobte Technologien - man weiß genau, wie und wo man bei der Optimierung ansetzen kann, wenn der Bedarf da ist. Unklar ist hingegen die Qualitätssicherung bei den für Predictive Analytics nötigen Big-Data-Quellen. Nutzen und Wertschöpfung der anvisierten Vorhersagemodelle hängt auch hier maßgeblich von der Qualität der zugrundeliegenden Daten ab. Systemarchitekten diskutieren nun, wie sie die Qualität von riesigen semi- und polystrukturierten Daten bewerten und sichern, welche Systemarchitekturen dabei ins Spiel kommen und wie das Datenmanagement funktioniert. Die Projekterfahrungen von Vorreiter-Unternehmen helfen bei der Klärung dieser Fragen.

Das Data Warehouse bleibt der Single Point of Truth

Zunächst ist festzuhalten: Das klassische Data Warehouse hat als Kern einer typischen BI-Umgebung auch in der Welt der fortgeschrittenen Analysen seine Daseinsberechtigung. Es ist die beste Grundlage für standardisierte Berichts- und Analyseprozesse mit den unverzichtbaren Finanz- und Steuerungskennzahlen. Geht es um vorausschauende Unternehmenssteuerung, so sind diese Berichtsstandards um Trendanalysen und Prognosen aus der Big-Data-Welt zu ergänzen. In der Praxis erweist sich der sogenannte Data Lake als pragmatischer Implementierungsansatz, um hochstrukturierte Daten aus Transaktionssystemen und wenig strukturierte Big Data zusammenzuführen. Ausgehend von der bewährten BI-Architektur mit offenen Schnittstellen lassen sich damit Big-Data-Komponenten in eine vorhandene Informationsplattform integrieren.

Wichtig dabei ist, dass das zentrale Data Warehouse seinen Anspruch als Single Point of Truth im Unternehmen behält. Will man das Datenmanagement und die Qualität der Datenbasis für Advanced Analytics optimieren,



Über einen Data Lake lassen sich die hochstrukturierten Daten aus Transaktionssystemen mit wenig strukturierten Big Data pragmatisch zusammenführen.

ist es daher eine gute Idee, mit dem meist vorhandenen Data Warehouse zu starten. Wie die Projektpraxis zeigt, besteht hier nämlich oft noch Handlungsbedarf. Abgesehen von Qualitätsmängeln und inkonsistenten Datenstrukturen wird mit dem Trend zu Self-Service-BI auch das bekannte Problem von Insellösungen und Datensilos wieder akut, das früher durch diverse Excel-Lösungen der Fachabteilungen verursacht wurde.

Self-Service im Fachbereich ist praktisch und hat seine Berechtigung, aber das zentrale Business Intelligence Competence Center oder der BI-Verantwortliche müssen die Datenströme unter Kontrolle halten und darauf achten, dass der Single Point of Truth nicht ausgehebelt wird.

Data Warehouse Automation sichert die Datenqualität

Sind die Datenströme gut modelliert, bestehen große Chancen für eine dauerhaft hohe Datenqualität im BI-System. Durchweg strukturierte Daten von der operativen bis zur dispositiven Ebene, standardisierte Auswertungsverfahren und mächtige ETL-Werkzeuge (Extraktion, Transformation, Laden) mit integrierten Prüffunktionen ermöglichen eine hohe Automatisierung der Datenauswertung. Für den effizienten Aufbau, die Anpassung und die Optimierung von Data Warehouses gibt es inzwischen ausgereifte Verfahren, die Standardisierung und Automatisierung erhöhen und damit die Fehlerrisiken auf ein Minimum senken.

Diese Data Warehouse Automation beruht auf Frameworks, die bereits Best Practices für ETL nach etablierten Verfahren sowie Prüflogiken zur Sicherung der Datenqualität enthalten beziehungsweise deren Modellierung auf Meta-Ebene unterstützen. Neben einer effizienten Entwicklung und Administration vermeidet dieser lösungsorientierte Ansatz Konstruktionsfehler und sorgt dafür, dass für eine saubere Datenverarbeitung Best Practices zum Einsatz kommen.

Da Business Intelligence die Business-Realität möglichst genau abbilden will, ist ein hohe Datenqualität unverzichtbar. Klassische Kriterien wie Exaktheit und Vollständigkeit sind dabei zentrale Anforderungen. Schon ein Datenfehler oder eine Lücke im operativen Bestand kann das Ergebnis einer aggregierten Kennzahl verfälschen. Im Rahmen des internen und externen Berichtswesens stehen damit schnell falsche Entscheidungen oder Compliance-Verstöße im Raum.

Bei Big Data steuert der Business Case die Governance

In der Big-Data-Welt gestaltet sich die Datenqualität anders. Hier geht es zunächst darum, die relevanten Datenquellen zu bestimmen, die Daten abzuholen und zu speichern. Das ist nicht immer trivial angesichts einer Bandbreite von Daten aus dem Internet of Things, unstrukturierten Informationen aus Blogs und Social Networks, Sensordaten aus Kassensystemen und Produktionsanlagen, Messdaten aus Leitungsnetzen bis zu Datensätzen aus Navigationssystemen. Im Gegensatz zur BI-Welt bestehen hier für die interne Datenarbeit keine allgemeingültigen Geschäftsregeln und Standards. Da es um die statistische Auswertung von Massendaten geht, sind die BI-typischen Qualitätskriterien Vollständigkeit und Exaktheit weniger wichtig. Im Rahmen der statistischen Verfahren fallen

einzelne Fehler und Lücken nicht ins Gewicht, und Ausreißer lassen sich regelbasiert eliminieren.

Wie groß die kritische Masse für belastbare Ergebnisse ist, wie genau, vollständig oder aktuell die Datenbasis sein muss und in welcher Form Informationen nutzbar gemacht werden, das ist für Big-Data-Analysen fallbezogen zu klären. Die Vielfalt der Einsatzbereiche und damit die Rahmenbedingungen für die Bewertung und Bearbeitung von Daten sind nahezu unbegrenzt.

Geht es etwa beim Internet of Things um die grobe Ressourcenplanung von Wartungsarbeiten für angebundene Geräte, sind Ausfälle einzelner Geräte-Meldesysteme irrelevant, da die Ermittlung von Peaks ausreicht. Im Rahmen von Predictive Maintenance ist dagegen jede konkrete Ausfallmeldung eines Gerätes wichtig. Für Kundenzufriedenheits-Indizes auf Basis von Weblog-Analysen kommt es nicht auf jeden Beitrag an. Vielmehr geht es darum, Trends abzuleiten und diese in sinnvoll definierte Kennzahlen zu überführen.

Bei Big-Data-Anwendungen fallen also Datenqualitätsmanagement und Governance ebenso individuell aus wie das Analyseszenario des jeweiligen Business Case. In hoch automatisierten Anwendungen wie Autonomes Fahren oder Predictive Maintanance, in denen ausschließlich Maschinen über die Ergebnisse und Auswirkungen von Datenanalyen entscheiden, ist die Data Governance besonders wichtig.

Die Quellen von Big Data liegen häufig außerhalb des Einflussbereichs der internen Prozesse: Maschinen-Output, Nutzereingaben oder Internet-Datenströme lassen sich nicht über interne organisatorische Maßnahmen kontrollieren. Bei permanent fließenden, unstrukturierten Datenquellen wie Chatforen greifen auch die klassischen ETL-Methoden nicht, und Störungen wie etwa eine Leitungsunterbrechung können nicht durch Wiederholung oder das Wiederherstellen des Datenbestands ausgeglichen werden.

Die Experten

Foto: Qunis

Foto: Qunis

Ilona Tag ist Head of Unit Big Data & Advanced Analytics bei der QUNIS GmbH. Sie hat die heute als QUNIS Datawarehouse Framework bekannte Methodik für die Modellierung, Architektur, Datenintegration und Implementierung von Data Warehouses entwickelt. Hermann Hebben ist Gründer und Geschäftsführer von Qunis.

Data Scientists spielen künftig eine Schlüsselrolle

Angesichts spezifischer und kaum standardisierbarer Anforderungen kommt es bei Big Data besonders auf die Qualität der individuell definierten Maßnahmen zur Datenerhebung und Qualitätssicherung an. Hier kommt der Data Scientist ins Spiel. Er formuliert für die Datenarbeit fallbezogen Regeln, Ziele und Qualitätsmaßstäbe. Dieses Regelwerk wird auf Basis permanenter Aufgaben wie der Prüfung von Cleansing-Protokollen, der Bewertung der Daten-Relevanz oder dem Monitoring von Anomalien ständig optimiert.

Der Data Scientist nimmt im Bereich Big Data und Advanced Analytics eine Schlüsselrolle ein. Das ist eine hohe Hürde für die Entwicklung und Umsetzung von Predictive-Anwendungen, denn Fachkräfte mit den nötigen Skills sind rar gesät. Der Data Scientist braucht tiefe mathematisch-statistische Kenntnisse, er muss programmieren können, sich mit Datenschutz und Compliance-Regeln auskennen und er sollte ein umfangreiches Business-Know-how und viel Praxiserfahrung mitbringen. Mit dieser umfassenden Kompetenz ist der Data Scientist maßgeblich an der Entwicklung digitalisierter Geschäftsprozesse bis hin zur Konzeption neuer Geschäftsmodelle beteiligt. Solche Alleskönner mit viel Spezialwissen und Erfahrung sind sehr gesucht, und es ist nicht absehbar, dass sich der Personal-Engpass in naher Zukunft lösen wird.

Eine profunde Konzeption sichert den Projekterfolg

Das Potenzial von Predictive Analytics ist riesig, und viele Unternehmen erschließen sich gerade neue Dimensionen der Informationsgewinnung. Durch Cloud-Betriebsmodelle lassen sich neue Anwendungen schnell und kosteneffizient umsetzen. Voraussetzung dafür ist eine profunde Konzeption, die den kompletten Wertschöpfungsprozess der Daten mit Blick auf ein präzise formuliertes Projektziel abdeckt.

Für ein erfolgreiches Projekt müssen anspruchsvolle Fragen der Fachlichkeit, Technik und Organisation geklärt werden. Hier empfiehlt es sich, die Erfahrung eines ganzheitlich orientierten Beratungsunternehmens hinzuzuziehen, um sich zeitraubende Umwege und schmerzhafte Lernzyklen zu ersparen. *jf*