▸ DATA WAREHOUSING

Parallelkraftfeld.

Eng abgestimmte Kombinationen aus Hard- und Software, sogenannte Appliances, können den Nutzen von Big Data-Analysen steigern. Das zeigt die positive Erfahrung eines Mobilfunkanbieters.

› Von Stephan Köppen *

In unserer Projektarbeit der vergangenen zwei Jahre kristallisierte sich klar heraus, dass «Big Data-Appliances» im Kommen sind. Dies zeigt vor allem das Praxisbeispiel eines großen deutschen Mobilfunkanbieters: Er stand vor der Herausforderung, eine verlässliche, stabile und zukunftssichere Plattform aufzubauen, um künftig sowohl den analytischen Anforderungen als auch dem stetigen Datenwachstum nachkommen zu können.

Das Unternehmen entschied sich für das Parallel Data Warehouse (PDW) von Microsoft, eine Big Data-Appliance. Die Wahl war goldrichtig, denn das Unternehmen konnte sofort mit der Entwicklung von Datenmodellen und Beladungen beginnen. Ein aufwendiges Konfigurieren und Installieren von Treibern oder Software war nicht erforderlich.

Aber was genau ist das PDW? Kurz gefasst: Eine perfekt aufeinander abgestimmte und für das Hochladen in das Data Warehouse optimierte Kombination aus Hard- und Software, die durch den Zusammenschluss mehrerer SQL-Server verschiedenste Prozesse und Abfragen parallel verarbeitet. Vielen Lesern wird diese Funktion unter dem Begriff «Massive Parallel Processing» (MPP) sicher bekannt sein.

Analysen auf der gleichen Datenbasis.
In dem praktischen Anwendungsfall beschränkt der Mobilfunkanbieter die Administration auf ein Minimum. Die zugrundeliegende Architektur baut auf der bewährten und stabilen SQL Server 2012-Technologie auf. Durch ihre Flexibilität und Skalierbarkeit bietet sie hinsichtlich der Datenspeicherung kaum Grenzen und ist auch bei geringeren Datenmengen eine gute, ausbaubare Alternative. Die Arbeit beginnt mit dem sogenannten «Base Rack» (Systemserver, zwei Datenbankknoten inklusive Speicherung, rund 50 Terabyte Kapazität). Dieses kann mit bis zu 26 zusätzlichen «Scale Units» (zwei Datenbankknoten inklusive Speicherung) auf sechs Petabyte erweitert werden.

In dem Projekt zeigte sich einmal mehr, wie wichtig es ist, den unterschiedlichen Datennutzern eines Unternehmens eine «Single Version of Truth» bereitzustellen, um alle Analysen und Auswertungen auf der gleichen Datenbasis zu ermöglichen. In diesem Fall sollte konkret zum einen das auf Oracle bestehende Data Warehouse abgeschaltet, zum anderen sollten die Daten auf das SQL-Server basierte Business Warehouse migriert werden.

Laufzeit um 50 Prozent reduziert.
Die bestehende Fast-Track-Architektur – entwickelt von Microsoft und seinen Hardware-Partnern – stieß bei dem Mobilfunkprovider immer häufiger an ihre Grenzen, was sich durch Ressourcen-Engpässe sowie langwierige Abfragen und Verspätungen bei der Bereitstellung der Daten widerspiegelte.

Und die Anforderungen waren hoch: Während der täglichen Bewirtschaftung werden bei dem Mobilfunkbetreiber jeweils der aktuelle und der Vormonat neu geladen. Täglich starten bis zu 200 Prozesse der Extraktion, Transformation und des Ladens (ETL) verschiedenster Daten, die wiederum etwa eine Kapazität von 600 Gigabyte für Lese- und Schreibzugriffe erfordern. In Spitzenzeiten beträgt der Datendurchsatz bis zu 40 GBit pro Sekunde.

Dank der Umstellung auf die neue PDW-Technologie reduzierte das Telekommunikationsunternehmen die benötigte Laufzeit von sechs auf vier Stunden. Dieser Zeitgewinn eröffnet den dringend benötigten Spielraum für Erweiterungen. Darüber hinaus werden in der Praxis die Applikationen mit enormen Anforderungen an Geschwindigkeit und Rechenleistung durch das PDW und die skalierbare Anzahl der Knoten auf einfache, aber sehr effektive Weise unterstützt.

Die schnelle Bereitstellung der Daten erweist sich sowohl für die Controlling-Abteilung als auch die angeschlossenen Verkaufsstellen als entscheidender Vorteil bei ihrer täglichen Arbeit.

Neben den zahlreichen ETL-Prozessen – abgebildet durch «Microsoft Integration Services» (SSIS) – greifen bei dem Mobilfunkanbieter auch diverse weitere Anwendungen innerhalb der Organisation auf die Daten zu wie etwa die unternehmensweiten OLAP-Modelle (Cubes), die mithilfe von «Microsoft Analysis Service» (SSAS) umgesetzt wurden.

Das größte und zugleich komplexeste Modell hat eine Größe von mehr als zwei Terabyte. Es beinhaltet (voll anonymisierte) Daten der jüngsten fünf Jahre aller 25 Millionen Kunden, deren jeweiligen Umsatz und Deckungsbeitrag im Verlauf sowie deren Nutzungsklassifizierung.

Achtfache Datenmenge verarbeiten.
Die Appliance unterstützt auch einen Upgrade der Lösung. Mit dem kürzlich erschienenen Update auf die zweite Version der Appliance kommt neben den bekannten SQL Server-Funktionalitäten die Integration für Hadoop-Daten durch «Polybase» hinzu, eine Brücke zwischen den Technologien SQL und Hadoop.

Falls bereits ein Hadoop-Cluster im Unternehmen existiert, kann dieser an die Appliance angebunden und ohne zusätzlichen Hardwarekauf genutzt werden. Unterstützt werden dabei zurzeit die Lösungen Hortonworks, Windows Azure und Cloudera.

Diese Flexibilität eröffnet Anwendern die Möglichkeit, sowohl strukturierte als auch un- oder teilstrukturierte Daten gleichermaßen mit SQL abzufragen. Einer Kombination aus unterschiedlichen Quellen wie etwa Social Media und relationalen Daten steht somit nichts im Wege. Insgesamt ist der praktische betriebswirtschaftliche Nutzen bei dem Mobilfunkprovider sehr hoch: Trotz strikter Einhaltung aller Datenschutzanforderungen laufen die Analysen in bis zu 70 Dimensionen.

Foto: gyb311/iStock/Thinkstock

Autobahn für Informationen: Schnelle Weiterleitung riesiger Datenmengen.

Das stellt für das Controlling einen wichtigen Erfolgsfaktor dar: Früher ließ der enorme Datenbedarf bei der Bewirtschaftung des Cubes die bestehende Architektur teilweise bei schon vier parallelen Abfragen an ihre Grenzen stoßen. Mit Unterstützung des PDW kann jetzt die achtfache Menge parallel verarbeitet werden – ohne andere Prozesse auch nur ansatzweise zu beeinflussen.

Speicherkapazität flexibel anpassen.
Fazit: Die Erfahrung bei dem Telekommunikationsanbieter zeigt, dass das PDW in der Praxis als zukunftssichere Big Data-Plattform mit maximaler Skalierbarkeit und Performance dienen kann. Durch einfaches Hinzufügen von Rechenknoten können sowohl die Geschwindigkeit als auch die Speicherkapazität dem Bedarf angepasst werden – auch im späteren Projektverlauf. Die Lösung gliedert sich nahtlos in das Microsoft-Produktportfolio ein und reduziert Migrationsaufwände auf ein Minimum.

Neben der Kostenersparnis bietet es dank Hadoop-Integration eine hohe Flexibilität. Das Speichern riesiger Datenmengen ist erschwinglich. Damit stellt die PDW-Lösung bereits in kleineren Konstellationen eine sehr gute Alternative zu den herkömmlichen Architekturen für Data Warehousing dar. ■

› Stephan Köppen ist Berater beim Business Intelligence-Dienstleister ORAYLIS GmbH mit Sitz in Düsseldorf. In seinen acht Jahren Datenbank- und BI-Erfahrung hat er sich zunehmend auf das Produkt «Microsoft Parallel Datawarehouse» spezialisiert. In seiner Rolle als «PDW Technical Expert» unterstützt er seit rund anderthalb Jahren die unterschiedlichen Projekte und hält Vorträge auf diversen Konferenzen.