S@PPORT
Entscheidungsgrundlagen für Auswahl, Installation und Betrieb von SAP*-Lösungen

Sonderdruck aus Heft 10/2014 vom 01. Oktober 2014 · www.sap-port.de

Notwendigkeit temperaturbasierter Datenhaltung in Zeiten von Big Data und In-Memory-Technologie

Adäquate und kostengünstige Datenhaltung

Ziel einer Big-Data-Strategie ist es, das riesige Datenaufkommen adäquat und kostengünstig zu speichern, um es für die spätere Datenanalyse bereitstellen zu können. Um dieses Ziel fachlich und technologisch zu erreichen, kommt der temperaturbasierten Datenhaltung oder Datenablage eine zentrale Bedeutung zu. Hierbei werden frühzeitig Kriterien für die Priorisierung der generierten und gesammelten Daten definiert und die Datenaufkommen schlussendlich klassifiziert.

Von Markus Fuchs,*
*Florian Moosmann***
*und Michael P. May****

Anforderungen an das Informationsmanagement steigen mit zunehmender Globalisierung und wachsendem Datenaufkommen rasant an. Laut der Digital-Universe-Studie der Firma EMC aus dem Jahre 2014 verdoppeln sich die weltweiten Datenbestände in etwa alle zwei Jahre. Der Studie zufolge betrug das weltweite Datenaufkommen im Jahre 2013 bereits 4.4 Zetabyte.
Um dieser Größenordnung Vorstellungskraft zu verleihen, anbei folgendes Exempel: Würde die soeben genannte Datenmenge auf iPads (Modell iPad air; Höhe 7.5 mm; Speicherkapazität 128 Gigabyte RAM) gespeichert, so wäre es beispielsweise möglich, die für die Speichermenge benötigten iPads auf einer Länge von knapp 256.000 Kilometern übereinander zu stapeln – das entspricht ungefähr 66 Prozent der Distanz von der Erde zum Mond. Unter Berücksichtigung der etwaigen Verdoppelung der Datenbestände bis zum Jahre 2020, könnten bereits 6.6 Stapel von der Erde zum Mond aufgebaut werden – das Datenvolumen betrüge dann in etwa 44 Zetabyte.
Während Unternehmen in der Vergangenheit nur einen Bruchteil der anfallenden Daten innerhalb der IT-Architekturen speicherten und ganz klar zwischen prioritären und weniger prioritären Daten unterschieden, geht der Trend heutzutage ganz klar Richtung „Big Data“. Was heißt das konkret? Es werden immer mehr Datenquellen angezapft. Zumeist wird die Vielfalt der zur Verfügung stehenden Daten unabhängig und losgelöst von zukünftigen analytischen Fragestellungen gesammelt und gespeichert.

Intelligentes Informationsmanagement und zukunftsfähige IT-Architektur

Um jedoch die Kosten für diese umfassende Datenablage nicht explodieren zu lassen, bedarf es zunächst eines intelligenten Informationsmanagements und einer zukunftsfähigen IT-Architektur. Es muss möglichst effizient zwischen unterschiedlich prioritären Datenablageformen differenziert werden können, um nachhaltig die analytischen Anforderungen optimal zu bedienen.

In den meisten Unternehmen, die Business-Intelligence-Plattformen betreiben, wird das klassische Setup einer IT-Architektur eingesetzt. Dieses besteht in der Regel aus

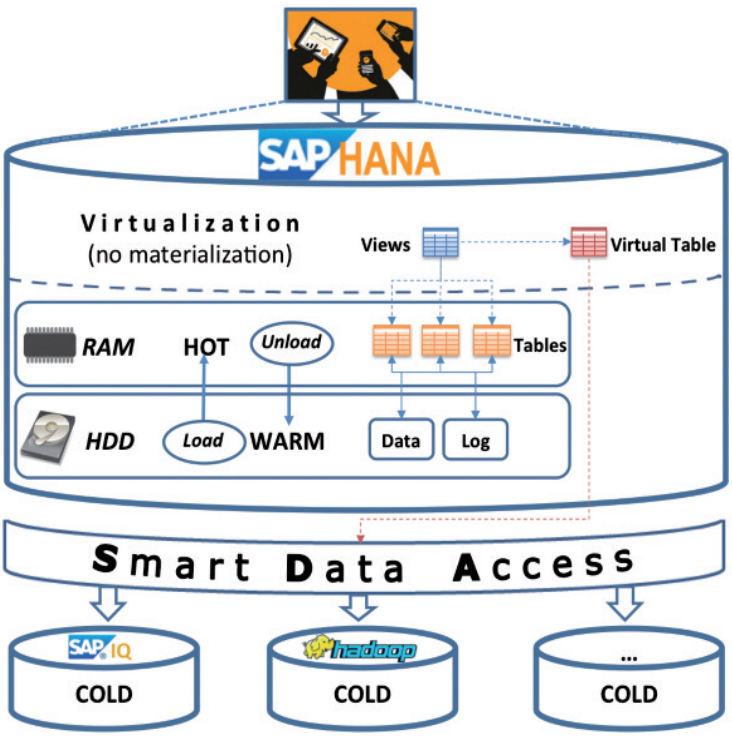
- einem Data Warehouse (zum Beispiel „SAP BW“) im Backend und
- diversen nachgelagerten Reporting Tools als Frontend-Clients (wie zum

**Markus Fuchs ist Consultant Information Management der thinkbetter AG.*
***Florian Moosmann ist Consultant Manager Information Management der thinkbetter AG.*
****Michael P. May ist Geschäftsführer der thinkbetter AG.*

Beispiel „SAP BEx“ oder „SAP BO Reporting Tools“).

Im Data Warehouse werden hierbei vor allem unternehmensinterne Systeme wie zum Beispiel ERP- oder CRM-Systeme als Datenquelle genutzt. Die prozessbezogenen Daten liegen meist in strukturierter Form vor, sind hochfrequentiert und geschäftskritisch. In vielen Unternehmen, in denen Big Data Einzug gehalten hat, werden zwar immer noch Data-Warehouse-Systeme eingesetzt, jedoch richtigerweise meist nur als ein technologischer Baustein im Kontext einer gesamtheitlichen Big-Data-Strategie. Denn Big Data bedeutet vor allem eines: Neben strukturierten Datenaufkommen werden auch semi- und unstrukturierte Daten gespeichert, verarbeitet, miteinander verknüpft und ausgewertet. Als Datenquellen kommen beispielsweise Web-Click-Logs, Sensordaten, Geo-Daten, Social-Media-Daten (wie zum Beispiel Facebook, Twitter und Xing), E-Mails, Dokumente und Streaming-Daten in Frage. Also Daten, die innerhalb und außerhalb eines Unternehmens generiert werden. Diese Daten sind vor allem durch ein rasantes und exponentielles Datenwachstum gekennzeichnet.

Ziel einer Big-Data-Strategie ist es nun, dieses riesige Datenaufkommen adäquat und vor allem kostengünstig zu speichern, um es für die spätere Datenanalyse überhaupt bereitstellen zu können. In Zeiten von Big Data entstehen daher auch neue Problemstellungen: Durch die Daten- und Informationsüberflutung können wichtige Informationen im Rahmen des analytischen Selektionsprozesses verloren gehen. Daher sollten zunächst die wesentlichen, unternehmerischen Fragestellungen frühzeitig eruiert werden, die im Rahmen einer intelligenten, vorausschauenden Geschäftsführung beantwortet werden sollen.



Mithilfe von Smart Data Access können extern persistierte Tabellen aus Quellen wie zum Beispiel SAP IQ oder Apache Hadoop als virtuelle Tabelle übernommen und in die HANA-basierte Modellierung einbezogen werden.

Temperaturbasierte Datenhaltung

Um dieses Ziel fachlich und technologisch zu erreichen, kommt der temperaturbasierten Datenhaltung oder Datenablage eine zentrale Bedeutung zu. Hierbei werden frühzeitig Kriterien für die Priorisierung der generierten und gesammelten Daten definiert und die Datenaufkommen schlussendlich klassifiziert. Kriterien für eine Klassifizierung der Daten sind unter anderem

- die Frequenz, mit der die Daten abgefragt werden,
- die Höhe der Priorität dieser Daten oder
- gesetzliche und regulatorische Anforderungen für bestimmte Arten von Daten.

Im Anschluss an die Definition der Kriterien für die Klassifizierung können die Daten letztlich in

- „Hot Data“ (hoch frequentierte Daten, hohe Priorität),
- „Warm Data“ (mittel frequentierte Daten, mittlere Priorität) und
- „Cold Data“ (niedrig frequentierte Daten, geringe Priorität)

eingeteilt werden.

Für die klassifizierten Daten gilt es nun in einem weiteren Schritt die jeweils geeignete Speichertechnologie innerhalb einer Unternehmensarchitektur zu definieren und die Daten in diese zu übertragen, um die temperaturbasierte Datenhaltung schlussendlich auch technologisch zu realisieren.

Hot Data

Gerade hochperformante Datenbanktechnologien wie beispielsweise „SAP HANA“ (High Performance Analytic Appliance) gewinnen im Zeitalter von Big Data immens an Bedeutung, da durch diese der performante Zugriff auf immens große Datenmengen ermöglicht wird. Diese In-Memory-Datenbanklösungen ermöglichen es, die als „Hot Data“ deklarierten Datenbestände innerhalb des RAM zur Verfügung zu stellen. Mithilfe dieses leistungsfähigen Speichers wie auch unter Einsatz von optimierten Datenablage- und Datenabfrageinstrumenten können Szenarien aus den Bereichen des Echtzeit-Reportings oder der Echtzeitanalyse gewährleistet werden. Die HANA-Datenbank besteht jedoch noch aus einer weiteren Speicherform.

Warm Data

Alle im RAM gehaltenen Datenbestände werden parallel noch in der sogenannten Persistenz festgehalten, da der RAM-Speicher architekturbedingt bei einem Neustart des Systems oder einer Unterbrechung der Stromversorgung geleert werden würde. Unter Persistenz sind nichtflüchtige Speicherformen wie Hard Disk Drives (HDD) und Solid State Drives (SSD) zu verstehen.

In diesen Speichern können die Daten im Gegensatz zum RAM auch über eine Unterbrechung der Stromversorgung hi-

naus dauerhaft gespeichert werden. Im Zuge eines Neustarts oder im Falle einer Unterbrechung der Stromversorgung werden die Daten aus der Persistenz zurück ins RAM geladen und stehen anschließend für die Verarbeitung wieder zur Verfügung.
Neben des Ansatzes, die Daten aus der Persistenz wieder zurück ins RAM zu schreiben, gibt es bereits auch Strategien wie zum Beispiel das „Non-Active Data Konzept“. Durch Anwendung dieses Konzeptes kann im Falle eines Speicherengpasses bestimmt werden, wie schnell die Daten ihren Zustand von „Hot Data“ in „Warm Data“ verändern und somit aus dem RAM verdrängt und nur noch innerhalb der Persistenz gehalten werden. Im Konzept der temperaturbasierten Datenhaltung bildet die Persistenz also am Beispiel der HANA-Datenbank den Bereich „Warm Data“.

Unterschied zwischen „Hot Data“, „Warm Data“ und „Cold Data“

Cold Data

Daten, welche als „Cold Data“ klassifiziert sind, könnten extern der HANA-Datenbank auf kostengünstigere Systeme wie beispielsweise „SAP IQ“ (als Near-Line Storage) oder auf ein „Apache Hadoop“-Cluster ausgelagert werden. Der Großteil der Daten, die im Rahmen einer Big-Data-Strategie erhoben werden, wird entsprechend der Klassifizierung der temperaturbasierten Datenhaltung „Cold Data“ zugeordnet. Demnach werden Daten unabhängig der initial definierten Klassifizierung in jedem Fall dauerhaft abgelegt und bei Bedarf in etwaige Analysen einbezogen. Gerade diese selten oder zumeist ad hoc frequentierten Datenbestände könnten auf diesem Weg „eingefroren“ und bei Bedarf zu einem späteren Zeitpunkt für Analysen wieder „aufgetaut“ werden. Bei Nutzung der SAP-HANA-Technologieplattform ist es möglich, auf ausgelagerte Datenquellen zuzugreifen und diese in Analysen einzubeziehen, ohne dass die Daten direkt auf der HANA-Datenbank innerhalb des RAMs oder Persistenz abgelegt sein müssen. Dies wird über die Smart-Data-Access-(SDA-)Technologie ermöglicht. Mithilfe von SDA können extern persistierte Tabellen aus Quellen wie zum Beispiel SAP IQ oder Apache Hadoop als virtuelle Tabelle übernommen und in die HANA-basierte Modellierung einbezogen werden. Somit sind Analysen über die verschiedensten Datenlieferanten innerhalb einer Unternehmensarchitektur hinweg möglich.
Smart Data Access erlaubt zudem den Push-down von Filtern, Aggregationen und anderen Funktionalitäten, sodass bereits der Großteil der Verarbeitungsschritte in den externen Datenquellen selbst stattfinden kann. Dies führt wiederum zu einer Entlastung des SAP-HANA-Systems und einem geringeren Datenvolumen. Am Beispiel der temperaturbasierten Datenhaltung würden jene über die SDA-Technologie angebundenen Tabellen „Cold Data“ repräsentieren.

Fazit

Die intelligente Klassifizierung und Einteilung von Daten in „Hot Data“, „Warm Data“ und „Cold Data“ ist fester Bestandteil der SAP-Strategie zu SAP HANA und findet sich in der SAP HANA Product Roadmap unter dem Begriff „Intelligent Data Tiering“ (intelligente Aufteilung der Daten) wieder. Durch ein strategisch klar ausgerichtetes Informationsmanagement und die richtige Umgebung werden das steigende Datenvolumen und die Vielfalt an Datenlieferanten sinnvoll und kostengünstig in die IT-Infrastruktur des Unternehmens integriert.

Um dies erfolgreich zu bewerkstelligen sollten jedoch folgende Fragen im Vorfeld geklärt werden:

- Wie werden BI-Informationssysteme optimal in strukturierte und unstrukturierte Datenquellen integriert?
- Wie lassen sich Daten hinsichtlich ihrer Relevanz für das tägliche Geschäft auch innerhalb der IT-Architektur unterscheiden und gleichzeitig auch entsprechend ihrer Priorität bereitstellen?

(ap) @