Zukunftsorientierte Analytische Plattformen

Handlungsempfehlung zur Ausrichtung der Analytics-Strategie

Autoren

Name	Firma
Andreas Wilmsmeier (Arbeitsgruppen-Sprecher)	TekLink International
Georg Schukat (Arbeitsgruppen-Sprecher)	Schukat Electronics
Marco Koppik	Arvato Systems
Dr. Michael Metzner	MHP
Steffen Weitz	Enercity

Vorwort

Der historische Antrieb für die Entwicklung des Data Warehouse war die Reduktion von Kosten und Laufzeiten für die Ausführung von Berichten, Auswertungen und Analysen. Operative Datenbanken waren in der Regel für operative Zwecke optimiert, die Auswertungen waren aufwändig und beeinträchtigten die Performance der operativen Anwendungen.

Für Auswertungen optimierte Datenmodelle auf separierten Data-Warehouse-Systemen sorgten für Abhilfe, erforderten aber eine Transformation der Daten in diese optimierten Datenmodelle. Der Erfolg dieses Ansatzes, kombiniert mit einer vielfältiger werdenden Systemlandschaft, hat dazu geführt, dass Daten aus verschiedenen Systemen zusammengeführt wurden, um übergreifende Auswertungen ausführen zu können – der zweite wesentliche Treiber für die Entwicklung des Data Warehouse.

Dank der rasanten technischen Entwicklung, hauptsächlich in den vergangenen zehn Jahren, sind die ursprünglichen Treiber der Data-Warehouse-Entwicklung heute nicht mehr relevant, Ressourcen stehen im Überfluss und zu geringen Kosten zur Verfügung. Das hat auch dazu geführt, dass heute praktisch jedes (operative) System über mehr oder weniger ausgefeilte analytische Funktionalitäten verfügt.

Gleichzeitig haben die Vielfalt und Heterogenität von Systemen und Daten in den letzten Jahren deutlich zugenommen und wird beschleunigt durch den Trend zu Cloud-Anwendungen, durch das Wachstum der erfassten Datenmengen und durch die hohe Geschwindigkeit dieses Wachstums.

Eine zentrale Aufgabe für das Data Warehouse, die Integration von Daten und die Koordination von Datenflüssen, gewinnt dadurch an Relevanz – und wenn man den Begriff des Data Warehouse großzügiger auslegt, umfasst das auch unstrukturierte Daten und semi-strukturierte Daten und damit auch das, was man heute allgemein unter den Begriffen „Big Data“, „Data Lake“ etc. zusammenfasst.

In der bisherigen Praxis implizierte die Datenintegration immer auch das Erzeugen (teils mehrfach) redundanter Kopien von Daten mit entsprechendem Aufwand für Entwicklung, Anpassung, Monitoring, Speicherplatz etc. Immer damit im Zusammenhang steht der Anspruch, eine möglichst hohe Qualität der Daten zu erreichen.

Als Grundlage für die Entwicklung einer zukunftsfähigen analytischen Architektur betrachten wir den Begriff der Integration neu entlang der Dimensionen „Integrationstiefe“ und „Qualität“.

1. Physische Integration der Daten an (ggf. mehreren) zentralen Orten – das entspricht dem klassischen Data Warehouse mit hoher Integrationstiefe und einem in der Regel hohen Anspruch an die Datenqualität. Data Lakes dagegen speichern zwar Daten aus verschiedenen Quellen an einem zentralen Ort, überlassen die Integration der Daten aber mehrheitlich den auswertenden Prozessen.

2. Logische, virtuelle Integration der Daten mit einem geeigneten Management von Metadaten, d. h. Daten werden – ähnlich wie bei Data Lakes – bei Bedarf gelesen, interpretiert und in geeigneter Form zusammengeführt, bleiben aber physisch dezentral gespeichert. Die Sicherung der Datenqualität wird dabei meist an die Datenquelle delegiert und durch die Integrationslogik abgebildet.

3. Integrierte Präsentation der Daten an der Benutzerschnittstelle – letztlich werden Daten aus verschiedenen Quellen so aufbereitet und z. B. in Dashboards nebeneinander dargestellt, dass sich für den Anwender ein Mehrwert gegenüber einer isolierten Präsentation ergibt. Integration oder Qualität der Daten werden in den Datenbankabfragen oder im Dashboard selbst adressiert.

Bezüglich der Qualität von Daten gibt es den klassischen Trade-Off zwischen Qualität, Kosten und Zeit. Die Halbwertszeit von Daten sinkt, die Kosten für die Integration steigen mit den Qualitätsanforderungen, gleichzeitig ist es für bestimmte Anwendungen nicht unbedingt erforderlich, eine sehr hohe Datenqualität zu erreichen: Finanzdaten müssen in der Regel zu 100 Prozent korrekt sein, die Ergebnisse von Kundenanalysen oft nicht – es kommt auf den Verwendungszweck an. Wichtig ist, diesen Trade-Off zu bewerten und auf dieser Basis eine bewusste Entscheidung über die wirtschaftlich erreichbare Datenqualität zu treffen.

Durch den Trend zu höheren Anforderungen an die Aktualität von Auswertungen zu Realtime-Szenarien und den weiteren technischen Fortschritt nimmt die Bedeutung der virtuellen Integration zu: Operative Auswertungen werden zunehmend wieder in die operativen Systeme zurückgeführt, externe Services (Cloud, Social Media etc.) bieten ausgereifte analytische Funktionalität an usw. – die Frage nach der Integration stellt sich für viele Betreiber von Data Warehouses (und übrigens auch Data Lakes, Data Oceans etc.) unabhängig von der verwendeten Technologie neu.

Losgelöst davon gibt es weiterhin einen Bedarf für die Kernfunktionen des Data Warehouse, nämlich die Bereitstellung integrierter, historischer, qualitätsgesicherter Daten. Die Anforderungen daran sind sogar gewachsen:

- Mehr Daten
- Mehr Agilität/Flexibilität
- Schnellere, komplexere Auswertungen
- Reduktion von Kosten
- Höherer Nutzen

Innerhalb des SAP-Portfolios (und natürlich darüber hinaus) gibt es aktuell eine ganze Reihe von Entwicklungen und Ansätzen, die in Richtung einer verteilten analytischen Architektur gehen:

- Offene Schnittstellen und virtuellen Zugriff gibt es bereits heute (SDA, Data Hub etc.), integriert in die HANA-Plattform.

- Mit BW/4HANA steht die nächste, voll mit der HANA-Plattform integrierte Generation eines klassischen Data Warehouse mit ausgereiften Modellierungswerkzeugen und administrativen Werkzeugen zur Verfügung.
- Die HANA-Plattform bietet Unterstützung für Modellierung und Betrieb nativer, SQL-basierter Data Warehouses.
- Der SAP Data Hub bietet vielversprechende Ansätze, etwa die automatisierte Erfassung von Metadaten (Datenbankschemata, APIs, Data Lineage etc.) oder die Möglichkeit, komplexe, mehrstufige Auswertungen zu orchestrieren und deren Ausführung zu überwachen, ohne dabei die Flexibilität bei der Auswahl der verwendeten Tools einzuschränken.
- Entwicklungen am Frontend, wie die SAP Analytics Cloud, machen Anwender unabhängiger von der IT und von den technischen Randbedingungen einzelner Systeme – es wird einfacher und schneller möglich sein, Analysen zu entwickeln
- Zukünftig könnten Nutzungsstatistiken, in Kombination mit Performance-Daten, genutzt werden, um einen Teil der Persistierung von Daten im Data Warehouse zu automatisieren. Ähnlich wie bei der – schon lange üblichen – Optimierung von Ausführungsplänen für SQL-Abfragen in heutigen Datenbanken könnte in vielen Fällen ein intelligenter Algorithmus entscheiden, ob Daten persistent (und damit redundant) gehalten werden, oder ob ein Durchgriff auf die Originaldaten effizienter ist.
- Die jüngste Ankündigung der SAP, das SAP Data Warehouse Cloud (kurz auch „DWC“ genannt), konsistent mit der SAP Cloud Strategy, bietet vielversprechende Ansätze. Die DSAG AK BI & Analytics wird an den Entwicklungen rund um die DWC dranbleiben und regelmäßig darüber informieren.

Insgesamt erwarten wir eine tiefere Integration der heutigen Lösungsansätze rund um die HANA-Plattform, wie BW/4HANA, Data Hub, Native SQL Data Warehousing, der SAC oder eben des DWC – insbesondere mit Fokus auf das Management und die gemeinsame Nutzung von Metadaten.

Vor diesem Hintergrund stellen wir in diesem Positionspapier einen Entwurf für eine zukunftsorientierte Referenzarchitektur für eine verteilte analytische Plattform mit ihren Komponenten und Anwendungsszenarien zur Diskussion. Das Ziel dieses Papiers ist es,

- DSAG-Mitgliedern aufzuzeigen, wohin die Reise der analytischen Anwendungen und des Data Warehousing geht, und ihnen Impulse zu geben, wie sie diese Entwicklungen in ihrer aktuellen oder künftigen Strategie berücksichtigen können, sowie
- SAP (und anderen Herstellern) aufzuzeigen, wo künftige Bedarfe liegen, und zur Weiterentwicklung bestehender Produkte beizutragen.

Inhaltsverzeichnis

1. Anforderungen an eine analytische Plattform 6
   1.1. Agilität 7
   1.2. Self-Service 7
   1.3. Daten 7
   1.4. Performance 8
   1.5. Kosten 8
2. Paradigmenwechsel 9
   2.1. Data as a Service 9
   2.2. Net of Truth 11
3. Komponenten der analytischen Plattform 13
   3.1. Datenintegration 13
   3.2. Physische Integration 13
   3.3. Virtuelle Integration 14
   3.4. Integration auf Präsentationsebene 16
   3.5. Orchestrierung 16
   3.6. Persistenz 16
   3.7. Analytisches Labor 17
   3.8. Harmonisierte & konsolidierte Business-Sicht 17
   3.9. Repository – Metadaten/Data-Lineage 17
4. Szenarien – ein Beispiel 19
   4.1. IT-Stack – SAP ERP, Hadoop & Fieldglass 19
   4.2. Datenvirtualisierung – SAP Data Hub 19
   4.3. Persistenz – SAP BW/4HANA und das Cloud Data Warehouse 20
   4.4. Datenorchestrierung – SAP Data Hub 20
   4.5. Harmonisierte und konsolidierte Business-Sicht – SAP Analytics Cloud 20
   4.6. Analytisches Labor 20
5. Zusammenfassung & Ausblick 21
6. Impressum 23

1. Anforderungen an eine analytische Plattform

Das Thema Business Analytics hat weiterhin einen hohen Stellenwert im Unternehmen, sieht sich aber im Vergleich zu klassischen BI-Architekturen aufgrund von hohem Innovationsdruck und neuen Anforderungen an die Analytik einem starken Wandel unterlegen.

Die Treiber dieses Wandels sind vielfältig. Das beginnt schon damit, dass Plattformen wie Google oder Facebook einen Zugang zu IT-Lösungen geschaffen haben, der – intuitiv, einfach und von jedem zu bedienen – einen nahezu ungehinderten Zugang zu einer großen Bandbreite von Information ermöglicht. Warum sollte das nicht auch in einem Unternehmen möglich sein?

Gleichzeitig hat die technische Weiterentwicklung im Hardwarebereich dazu geführt, dass heute Kriterien wie Rechnerleistung, Datenübertragungsraten oder Datenvolumen für die meisten Anwendungen eine deutlich geringere Rolle spielen, da die Kosten für leistungsfähige Komponenten drastisch gesunken sind. Die Verfügbarkeit von Cloud-Lösungen (z. B. Platform as a Service - PaaS) senkt die Hemmschwelle für den Zugang zu Hochleistungssystemen. Entsprechend steigt die Verfügbarkeit von Daten verschiedenster Herkunft und Formate an – das Internet of Things (IoT) ist ein prominentes Beispiel dafür.

Hinzu kommen die Fortschritte in der Software: Die Verwaltung großer Datenmengen und komplexer Datentransformationen ist grundsätzlich kein Problem mehr. Bis hin zu komplexen neuronalen Netzen ist fast alles als Open Source verfügbar und ermöglicht komplexe Analysen, die teils sogar in Echtzeit ausgeführt werden können. Analytische Frontends sind leistungsfähiger geworden und einfacher zu bedienen als noch vor wenigen Jahren. Das folgende Schaubild gibt einen Überblick über die Entwicklung.

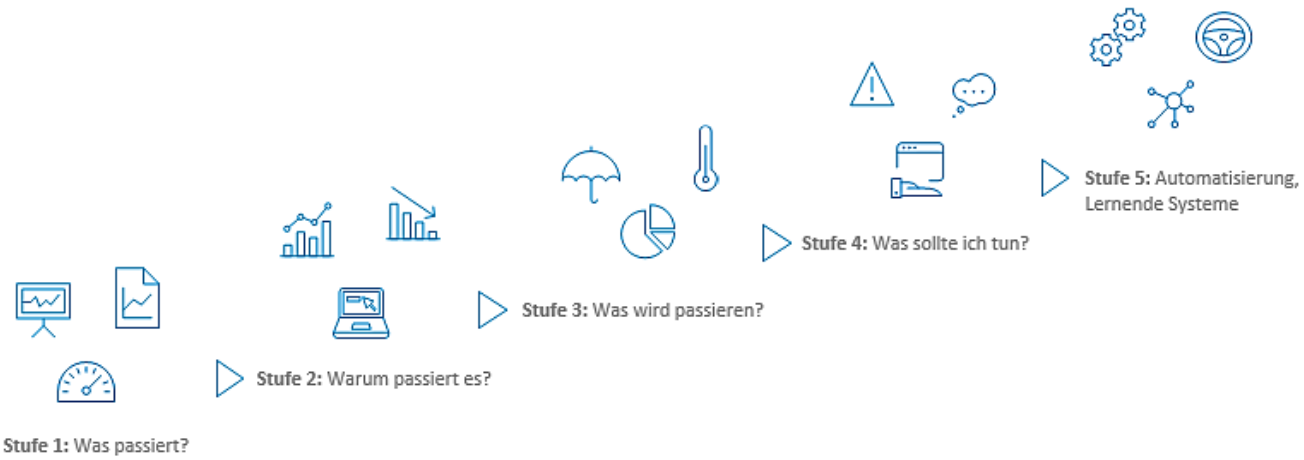

Bild 1 – Evolution analytischer Systeme (eigene Adaption)

Die harten, noch zu knackenden Nüsse kommen daher eher aus einer anderen Richtung: Welche Geschäftsmodelle versprechen einen tatsächlich messbaren Erfolg? Wie kann sich ein Unternehmen so aufstellen, dass es effektiv neue Herausforderungen bewältigen kann?

Unabhängig davon, wie diese Fragen im Einzelfall konkret beantwortet werden: Analytics auf großen, heterogenen und vielfältig strukturierten Datenmengen wird dabei in den meisten Fällen eine zentrale Rolle spielen und neue Anforderungen an die Architektur und Umsetzung analytischer Systeme stellen.

1.1. Agilität

Mehr denn je ist der Einsatz von Informationstechnologie in Unternehmen im Wandel. Unternehmen reduzieren die Komplexität ihrer IT-Lösungen durch Konsolidierung, gekoppelt mit einer Harmonisierung von Geschäftsprozessen. Gleichzeitig werden Prozesse, die nicht als Kernprozesse angesehen werden, ausgelagert – zumeist in moderne Cloud-Lösungen. Speziell für analytische Lösungen entstehen daraus hohe Anforderungen an deren Flexibilität und Agilität. Überall entstehen neue Daten, die mit anderen kombiniert und ausgewertet werden müssen. Neue Fragestellungen und Geschäftsmodelle erfordern neue, zum Teil komplexe oder automatisierte Analysen und Berichte.

1.2. Self-Service

Wie oben schon erwähnt, sind Anwender heute daran gewöhnt, sich individuell, schnell und unkompliziert Zugang zu Informationen zu verschaffen: Der Begriff „Self-Service Analytics“ ist in aller Munde. Viele solcher Konzepte wurden bereits umgesetzt, und doch hat die Anforderung angesichts steigender Heterogenität von Daten nichts an Aktualität eingebüßt – insbesondere, wenn es darum geht, Daten verschiedenen Inhalts mit verschiedenen Formaten zu kombinieren. Gleichzeitig steigen die Anforderungen an den Datenschutz – nicht alle Daten können geteilt und frei ausgewertet werden. Die Datenschutz-Grundverordnung (DSGVO) ist ein aktuelles Beispiel dafür.

1.3. Daten

Der traditionelle zeitliche Horizont für Datenanalysen im betriebswirtschaftlichen Bereich liegt beim aktuellen Jahr und einer Historie von zwei bis drei Jahren. Ältere Daten werden meist archiviert und nur selten wieder genutzt. Längere Analysezeiträume gibt es in einigen Branchen (z. B. Versicherungen) oder für spezielle Analysen, die einen längeren Zeithorizont erfordern.

Anders sieht es in den Bereichen Internet of Things (IoT) und Big Data aus. Daten sind tendenziell kurzlebiger und schneller veraltet. Erkenntnisse werden automatisiert zur Steuerung verarbeitet, aggregiert und nur im Fall der Eskalation schnell an die richtigen Entscheider kommuniziert. Die Datenqualität sinkt tendenziell und verlangt fehlertolerante Analyseverfahren. Aufgrund von fehlender Strukturierung steigt die Datenkomplexität an.

1.4. Performance

Performance ist immer weniger ein Problem, solange man Daten in einem zentralen Data Warehouse (oder einem zentralen Data Lake) zusammenführt und dort auswertet. Bei verteilten Architekturen gibt es hier jedoch noch einige Herausforderungen und viel Optimierungspotenzial.

1.5. Kosten

Kosten sind und bleiben ein wesentlicher Faktor. Gerade die Umsetzung neuer analytischer Lösungen (wie z. B. Machine Learning etc.) scheitern oft daran, dass der Nutzen erst erkennbar wird, nachdem man die Analysen durchgeführt hat. In Unternehmen mit einer stark kostengetriebenen Kultur ist diese Art von Innovation daher schwer durchzusetzen.

2. Paradigmenwechsel

Betrachtet man die aktuellen Herausforderungen und die Entwicklungen der letzten Jahre mit ein wenig Abstand, so wird erkennbar, dass künftige Lösungen, Architekturen und Infrastrukturen möglichst modular, verteilt und austauschbar konzipiert sein müssen. Schon vor Längerem wurden serviceorientierte Ansätze entwickelt, die diesen Anforderungen Rechnung tragen. Cloud-basierte Lösungen (SaaS) sind eine derzeit sehr erfolgreiche Ausprägung, aber auch aktuelle On-Premise-Lösungen profitieren davon, dass eigentlich alle wesentlichen Funktionen auch als Services oder als Micro-Services zur Verfügung gestellt werden können; analytische Services sind meist im Serviceangebot enthalten.

Diese neuen Strukturen haben einen großen Einfluss auf die Konzeption künftiger analytischer Plattformen, die sich mit diesen unterschiedlichen Datenressourcen integrieren müssen. Und natürlich ist eine integrierte Analytik über eine verteilte, dezentrale IT-Landschaft nicht trivial und nicht immer wirtschaftlich. Klassische Data-Warehouse-Lösungen stoßen da an ihre Grenzen, und auch Hadoop-basierte Data Lakes werden diesen Anforderungen bisher nicht wirklich gerecht. Im Kern der hier vorgeschlagenen Lösungsansätze steht daher auch die Abkehr vom rein monolithischen, zentralisierten Data Warehouse hin zu einer virtuellen analytischen Plattform.

Diese analytische Plattform greift auf die Services der verschiedenen Datenquellen zu, verfügt über ein API-Management und bietet gleichzeitig – wo notwendig – auch die Möglichkeit, Daten zu persistieren. Auf dieser Plattform setzen dann Prozesse auf, die Daten integrieren, kombinieren und analysieren. Grundvoraussetzungen für die erfolgreiche Umsetzung einer solchen Plattform sind eine solide Basis an Metadaten, eine leistungsfähige Orchestrierung von Datenflüssen über Systemgrenzen hinweg und eine hohe Agilität in der Integration von neuen Services/Technologien.

Aus Sicht der DSAG bedarf es einiger Paradigmenwechsel in der Architektur und Entwicklung analytischer Anwendungen. Auf diese gehen wir im Folgenden ein.

2.1. Data as a Service

Eines der Kernprobleme für analytische Anwendungen ist der Zugriff auf Daten. Hierbei spielen die technischen Fragen mittlerweile eher eine untergeordnete Rolle, da praktisch alle aktuellen Softwarepakete über offene Schnittstellen zum Austausch von Daten und über Services (oder APIs) zum Zugriff auf bestimmte Funktionen verfügen. Eine größere Rolle spielen organisatorische Probleme:

- Wo finde ich die benötigten Daten?
- Wer darf auf welche Daten zu welchem Zweck zugreifen?
- Wie werden die Kosten für Datenmanagement und -austausch geteilt?
- Welche organisatorischen oder politischen Probleme stehen dem Datenaustausch im Weg?

Viele der Fragen basieren letztlich darauf, dass diejenigen im Unternehmen, denen die Daten „gehören“, d. h. die die Anwendungen betreiben, in der Regel kein ausgeprägtes Interesse daran haben, diese mit anderen im Unternehmen oder auch außerhalb des Unternehmens zu teilen.

Ein radikaler Ansatz wäre daher, diesen Zustand umzukehren und alle, die innerhalb eines Unternehmens Anwendungen betreiben, grundsätzlich dazu zu verpflichten, die daraus entstehenden Daten mit anderen zu teilen. Das Konzept, Daten unabhängig von der organisatorischen Aufteilung von Systemen und Nutzern für Jedermann bereitzustellen, wird als „Data as a Service“ bezeichnet.

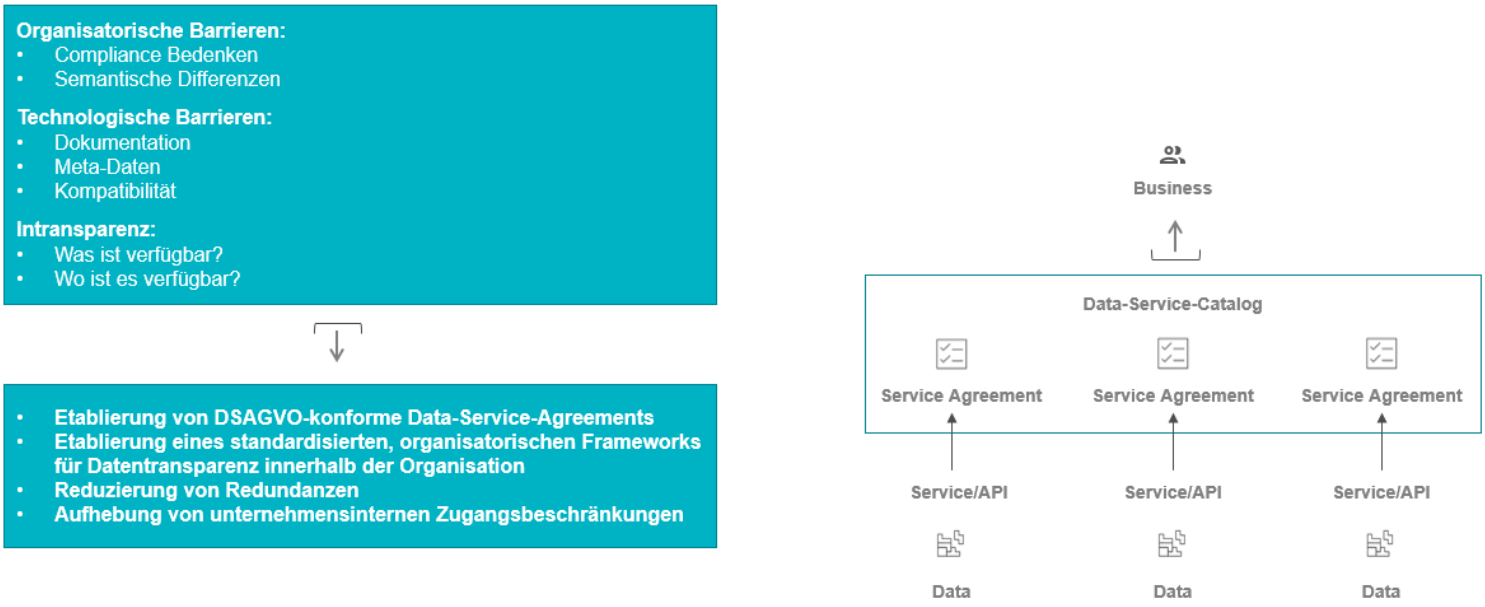

Bild 2 – „Data as a Service“-Ansatz (eigene Adaption)

Um nicht nur den Anforderungen des Datenschutzes Rechnung zu tragen, sondern auch die gegenseitigen Erwartungen an die Verfügbarkeit und die Qualität des Datenservices zu regeln, werden „Data Service Agreements“ (DSAs) eingeführt, die die verschiedenen Aspekte des Datenaustauschs regeln – ähnlich den bekannten Geschäftsbedingungen von Plattformen wie z. B. „weather.com“, „twitter.com“ oder „facebook.com“:

- Welche Daten sind in welcher Qualität und Aktualität verfügbar?
- Wie können welche Daten genutzt werden?
- Welche Kosten entstehen beim Abruf der Daten und wie werden diese abgerechnet?
- ...

Bei den großen Internetplattformen können Nutzer einfach die AGB akzeptieren und die Daten abonnieren – warum sollte das nicht grundsätzlich auch im Unternehmen funktionieren?

Kombiniert mit einem Verzeichnis von verfügbaren Services („API Repository“), ließen sich aus Sicht der DSAG so eine Reihe der üblichen Probleme beim Zugriff auf Daten lösen. Es ließen sich effektiv die Anzahl und Komplexität der Schnittstellen und damit Kosten reduzieren und gleichzeitig eine bessere Datentransparenz erzielen sowie schneller und flexibler neue Analysen vornehmen.

Tools wie der SAP Data Hub können helfen, die Logistik für Daten und Metadaten zu organisieren, transparent zu machen, wer wann die Daten wo nutzt, und ein API Repository zu betreiben. Die Bereitstellung und das Management von Data Service Agreements ist dagegen aktuell nicht durch bestehende Komponenten abgedeckt.

2.2. Net of Truth

Das Paradigma des monolithischen Data Warehouse als „Single Point of Truth“ – oft auch „One DWH“ oder „One BI“ genannt – ist überholt, da es auf die dargelegten neuen Anforderungen nicht schnell und flexibel genug reagieren kann.

Ersetzt wird es durch das Paradigma eines „Net of Truth“, eines dezentralen, virtuellen Data Warehouse, das das Datenkapital eines Unternehmens über alle relevanten Anwendungen, Services und APIs hinweg mit einem „Single Point of Entry“ transparent verfügbar macht. Daten können, falls notwendig (z. B. zur Performance-Optimierung), persistent gespeichert werden. Für Anwender erfolgt der Zugriff auch weiterhin über die bereits vorhandenen, üblichen Frontend-Technologien.

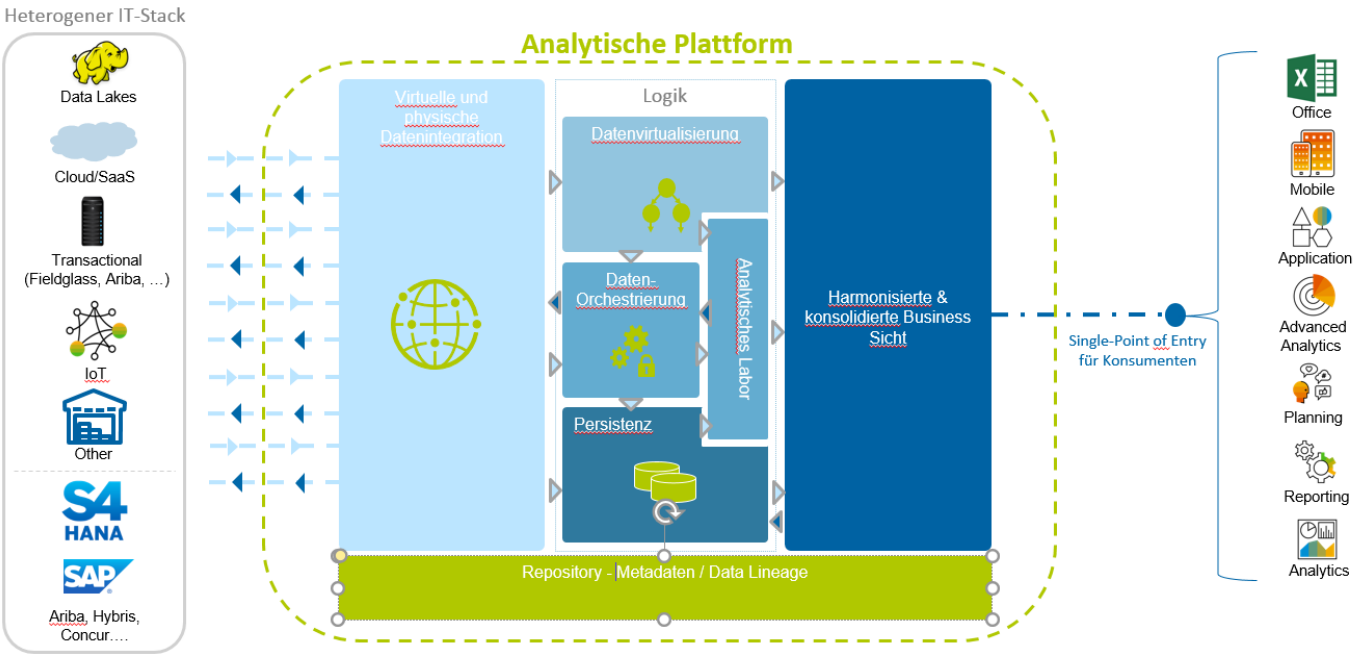

Bild 3 – Dezentrale analytische Plattform (SAP, eigene Adaption)

Dieses Konzept einer dezentralen analytischen Plattform ermöglicht eine schnelle Integration innerhalb des heterogenen IT-Stacks, mit einer klaren Trennung von Frontend und Backend, und erlaubt eine agile Umsetzung neuer analytischer Anforderungen.

Innerhalb dieser dezentralen Plattform ist außerdem eine smarte Trennung zwischen Fehlertoleranz und Stabilität abzuwägen. Es gibt Anforderungen, wie z. B. im Finanzbereich, die auf hohe Datenqualität und operative Stabilität angewiesen sind. Gleichzeitig ist es in anderen Bereichen wichtiger, schnell (bis hin zur Echtzeit) Muster zu erkennen und darauf reagieren zu können. Bei solchen Anwendungen sind Unschärfen oder Fehler in der Menge von meist unstrukturierten Datenbeständen bis zu einem gewissen Grad akzeptabel, da diese das Ergebnis nicht signifikant verändern. Um maximalen Nutzen zu erzielen, muss die Governance für Entwicklung und Betrieb der Plattform beide Varianten sowie Mischformen optimal unterstützen.

3. Komponenten der analytischen Plattform

Für den im vorherigen Kapitel aufgezeigten Paradigmenwechsel von einem klassischen Data Warehouse hin zu einem Net of Truth, verteilt über diverse Legacy-Systeme, Services, APIs oder Datenbanken, ist die Datenintegration der zentrale Schlüssel. Wenn man also davon ausgeht, dass eine künftige analytische Plattform verteilt auf Daten, auf Funktionen und auf Services zugreifen kann, dann müssen die folgenden, sich aus diesem Ansatz ergebenden Fragestellungen genauer betrachtet werden:

- Wie gut und auf welche technische Weise können die Daten virtuell oder physisch integriert werden?
- Wie kann sichergestellt werden, dass eine bestimmte Funktion bzw. ein Service auf den Daten, die die Datenquelle selbst anbietet, auch von der Business-Analytics-Plattform konsumiert werden kann?
- Wie kann die virtuelle oder physische Kombination von Daten einheitlich orchestriert und überwacht werden?
- Wie kann mit einem geeigneten Metadaten-Management sichergestellt werden, dass die abgerufenen Daten einen bestimmten semantischen Inhalt und eine gewünschte Aktualität und Qualität haben?

3.1. Datenintegration

Wie in der Einleitung beschrieben, kann die Integration von Daten konzeptionell auf drei verschiedenen Ebenen erfolgen: physisch, virtuell oder lediglich auf Präsentationsebene integriert. Die Bedeutung dieser unterschiedlichen Ebenen in der Vergangenheit und in der Zukunft wird in den nächsten Abschnitten erörtert.

Die bereits etablierten Tools für die physische Datenintegration müssen um Fähigkeiten zur virtuellen Integration ergänzt werden, so dass fließende Übergänge zwischen physischer und virtueller Datenintegration möglich werden. Der Fokus wird daher in Zukunft mehr auf der semantischen Modellierung der Zusammenhänge von Daten liegen, die Wahl der technischen Anbindung wird eher in den Hintergrund treten. Im Idealfall könnte es einer Optimierungskomponente überlassen werden, ob ein virtueller Durchgriff oder eine zusätzliche Persistenz für einen bestimmten analytischen Anwendungsfall am besten geeignet ist.

3.2. Physische Integration

Klassische Business-Intelligence-Architekturen wie ein Data Warehouse basieren auf dem Grundgedanken, dass alle relevanten Daten aus den verschiedensten Datenquellen extrahiert und in einem zentralen System für analytische Auswertungen gesammelt werden. Auch wenn – wie im nächsten Abschnitt beschrieben – die Virtualisierung der Datenzugriffe in analytischen Applikationen weiter voranschreitet, wird die physische Integration der Daten auch in Zukunft eine wichtige Rolle spielen.

In den klassischen Architekturen wird diese Form der Datenintegration über ETL-Prozesse erreicht. Dabei werden die Daten in periodischen Ladeprozessen aus der Quelle in den analytischen Speicher übertragen, traditionell entweder vollständig oder über ein Delta-Verfahren.

In den letzten Jahren hat die Nutzung von Daten aus kontinuierlichen Datenströmen stark an Bedeutung gewonnen: Eine Datenquelle liefert einen kontinuierlichen Strom an Daten, der in vielen Fällen direkt ausgewertet wird (Event Stream Processing, Streaming Analytics etc.), aber für die weitere Verwendung in Analysen ggf. auch mit anderen Daten integriert und dauerhaft persistiert werden kann.

Für die physische Integration von Daten ist der Komplexitätsgrad aufgrund der gewachsenen Anforderungen an die Aktualität der Daten, an diversifizierte Datenquellen, aufgrund der Abhängigkeiten zwischen den Daten und aufgrund der Heterogenität der Infrastruktur mit On-Premise- und Cloud-basierten Systemen in den letzten Jahren stark gestiegen – insbesondere mit Bezug auf das Monitoring und die Sicherstellung von Datenqualität und -aktualität.

Die Funktionalitäten für die physische Integration der Daten wandern daher auch zunehmend in spezialisierte Komponenten (HANA-Plattform, Data Hub, SDI, SDA etc.). Betrachtet man etwa das klassische SAP BW und vergleicht es mit einem modernen SAP BW/4HANA, so ist dieser Trend deutlich zu erkennen: Hatte das klassische BW noch proprietäre Schnittstellen zur Anbindung von Quellsystemen, so ist diese Funktionalität in einem BW/4HANA an die technischen Spezialisten zur Datenintegration verlagert worden: HANA EIM, SAP SLT oder SAP Data Services.

Es ist klar, dass die physische Integration von Daten zwar durch immer mächtigere Tools unterstützt wird, umgekehrt aber auch wartungsintensiv ist. In einer modernen analytischen Plattform muss also zu jeder Zeit hinterfragt werden, ob und wann eine physische Integration der Daten weiterhin notwendig und richtig ist.

3.3. Virtuelle Integration

Mit fortschreitender technischer Entwicklung spielt die virtuelle Integration von Daten eine zunehmend wichtige Rolle. Bei diesem Ansatz bleiben die Daten in ihrer jeweiligen Quelle persistiert und werden über offene APIs von der analytischen Applikation abgerufen. Man spricht auch von Datenföderation oder im Data-Warehousing-Kontext von einem logischen oder dezentralen Data Warehouse.

Die Anforderungen an eine virtuelle Integration sind vielfältig, so dass entsprechende Ansätze in der Vergangenheit immer wieder an ihre (Performance-) Grenzen stießen:

- Ein wichtiges Kriterium ist die Geschwindigkeit von Abfragen. Die Performance ist umso besser, umso mehr Operationen direkt in der Quelle ausgeführt werden können („Push Down“-Prinzip), ohne große Datenmengen im Netzwerk zu übertragen. Dies ist aber nicht trivial, wenn Daten aus mehreren Quellen integriert werden sollen.

Die Integrationsplattform benötigt hierfür einen Optimierer, der z. B. bei einem (semantisch) einfachen Join von Daten aus zwei unterschiedlichen Quellen entscheiden kann, wo und in welcher Form diese Operation optimal ausgeführt werden kann. Es ist klar, dass die Optimierung aufgrund der Besonderheiten jeder einzelnen Datenquelle und der kundenspezifischen Netzwerktopologie sehr komplex ist.

- Die Integrationsplattform muss in der Lage sein, heterogene Umgebungen auf eine einheitliche Art und Weise einzubinden. Dabei darf es keine Rolle spielen, ob die Daten in einem kundeneigenen On-Premise-System oder in einer Cloud-Umgebung gespeichert werden. Die virtuelle Integrationsplattform benötigt daher den Zugriff auf (Datenbank-) Schnittstellen, APIs und Services.
- Jegliche Business-Logik und -Harmonisierung findet „on-the-fly“ zur Laufzeit einer Abfrage statt. Die Redundanz an Daten in einem Data Warehouse wird also durch eine Redundanz an (teuren) Abfragen ersetzt, die wiederum über geeignete Caching-Algorithmen der Virtualisierungsplattform zu verbessern sind.
- Je mehr die Daten eines virtuellen Datenmodells verteilt sind, umso wichtiger wird es, dass es eine einheitliche Sicht auf Metadaten gibt und darüber hinaus die Herkunft der Daten („Data Lineage“) einfach nachvollzogen werden kann.

Die Vorteile einer Virtualisierungsplattform sind offensichtlich: Sie ermöglicht konsolidierte Sichten auf verteilte Datenquellen, ohne die Notwendigkeit, ressourcen-, wartungs- und betreuungsintensive Datentransferprozesse aufzubauen. Darüber hinaus bleiben die unterschiedlichen Daten je nach Verwendungszweck physisch getrennt.

Für eine analytische Plattform muss die Virtualisierung neben der rein technischen Datenintegration weitere wichtige Funktionen bieten:

- Eine Cache-Optimierung dient zur Laufzeitverbesserung bei mehrfachen Abfragen auf identische Daten eines virtuellen Datenmodells.
- Transformations- und Mapping-Logik muss abbildbar sein.
- Für ein Data Warehouse sind hierarchische Speicherkonzepte wichtig, in denen ältere oder weniger häufig benutzte Daten in günstigere Speicherarchive (Hadoop, NLS) ausgelagert werden können.

Aufgrund dieser Funktionen bietet die Virtualisierung der Datenintegration sehr viel Potenzial für eine höhere Agilität in einer künftigen analytischen Plattform und wird somit eine zentrale Rolle spielen. Im SAP-Portfolio übernimmt die HANA-Plattform mit Smart Data Access, Smart Data Integration etc. weitgehend die technischen Zugriffe auf entfernte Daten, der Data Hub mit seinen Metadaten, der Orchestrierung und den administrativen Funktionen wird die zentralen Aufgaben einer virtuellen Integrationsplattform übernehmen.

3.4. Integration auf Präsentationsebene

Die einfachste Form der Integration von Daten findet auf Präsentationsebene statt. Hier bedient sich ein Frontend diverser Quellen, die in einem gemeinsamen Bericht dargestellt werden, ohne für eine echte Kopplung der Daten zu sorgen. Im einfachsten Falle handelt es sich hierbei um Dashboards, bei denen mehrere Grafiken und Tabellen aus unterschiedlichen Quellen bedient werden. In allen SAP-Analytics-Produkten (SAP Lumira, SAP Analytics Cloud) und den Produkten anderer Hersteller ist diese Form der Datenintegration möglich.

3.5. Orchestrierung

Schon in der Welt des klassischen Data Warehouse erforderten die verschiedensten funktionalen und logischen Abhängigkeiten zwischen Anwendungen und Daten ein ausgefeiltes Konzept für die Koordination der Datenlogistik.

Umso wichtiger wird auf dem Weg hin zum intelligenten Unternehmen die Orchestrierung von Daten, verteilten Datenzugriffen und kontinuierlichen Datenströmen in einer diversifizierten Datenlandschaft. ETL, EAI und Streaming bleiben wichtige Bausteine einer Orchestrierungsstrategie, werden aber ergänzt durch übergreifende Steuerung und übergreifendes Monitoring. Im Portfolio von SAP ist für diesen Anwendungsfall der SAP Data Hub vorgesehen.

3.6. Persistenz

In letzter Konsequenz ist jede zusätzliche Persistenz von Daten eine Optimierung mit dem Ziel, eine System- oder Netzwerklast zu reduzieren. Die Performance von Berichten kann durch Vorberechnung von Daten oder durch eine optimierte Datenablage gesteigert werden.

Im Zuge der technischen Innovationen kann stetig jede zusätzliche Persistenz zur Optimierung mit dem Ziel hinterfragt werden, komplexe Ladeprozesse und aufwändige Modellierungen zu vereinfachen. Wo vor ein paar Jahren rechenintensive operative Berichte noch zwingend aus dem ERP in ein SAP BW ausgelagert werden mussten, findet heute ein operatives Reporting in einem SAP-S/4HANA-System kaum noch technische Grenzen. Auch innerhalb des SAP BW selbst kann das über Jahre bewährte LSA-Schichtenmodell nach Einführung von SAP HANA vereinfacht werden. In vielen Fällen ist die Performance der Datenbank so gut, dass die Anzahl der Persistenzschichten innerhalb des SAP BW reduziert werden kann. Ein offensichtliches Beispiel hierfür ist der Wegfall von InfoCubes.

Auch wenn die Persistenz von Daten in Zukunft weiter reduziert werden kann, wird diese nicht komplett hinfällig. Die folgenden Szenarien werden weiterhin eine Persistenz und damit eine Redundanz an Daten erfordern:

- Komplexe Datenmodelle/Berechnungen
  Besonders bei hochkomplexen Datenmodellen oder bei komplexer Berechnungslogik ist oft weiterhin eine redundante Speicherung der benötigten Basisdaten sinnvoll.

- Datenqualität
  Wie bereits oben erwähnt, haben bestimmte Prozesse (z. B. im Finanzwesen) besonders hohe Anforderungen an die Datenqualität, bis hin zum detaillierten Auditing von Daten.
- Snapshots von Daten
  Klassische Bestandsberichte, aber auch wiederum Auditing, erfordern Snapshots von Daten zu bestimmten Zeitpunkten, die dann jeweils durch Bewegungsdaten und entsprechende Berechnungen ergänzt werden.
- Historische Daten
  Darüber hinaus erfordern Auswertungen über weiter zurückreichende Datenhistorien einen Zugriff auf Daten, die in dieser Form im ERP-System üblicherweise nicht oder nicht mehr vorhanden ist.

Für solche Anwendungsfälle ist ein separates Data Warehouse weiterhin sehr sinnvoll. Allerdings nimmt die Anzahl der für ein Data Warehouse relevanten Anwendungsfälle deutlich ab.

3.7. Analytisches Labor

Das analytische Labor ist der Bereich, in dem der Data Scientist neue analytische Modelle erstellt und testet. Idealtypisch greifen diese Modelle virtuell auf Live-Daten zu, ohne eine Datenreplikation anzufordern, oder sie greifen auf temporäre Kopien dieser Datenbestände zurück.

Das analytische Labor kann im klassischen Data-Warehouse-Kontext genutzt werden, um im Rahmen eines Rapid-Prototyping-Ansatzes neue KPIs zu erstellen und die Datenqualität zu analysieren, bevor aus dem Prototyp ein standardisiertes Modell für eine harmonisierte Business-Sicht generiert wird.

Darüber hinaus werden im analytischen Labor auf Basis von Technologien wie beispielsweise Keras, Tensorflow, Jupyter Notebooks Predictive oder Machine Learning Modelle aufgebaut.

3.8. Harmonisierte & konsolidierte Business-Sicht

In der harmonisierten Business-Sicht werden den Fachanwendern alle Datenmodelle zur Verfügung gestellt. Diese Komponente ist der Single Point of Entry auf alle Datenquellen. In dieser Komponente können virtuelle oder temporär persistierte Views erzeugt werden, die beispielsweise in Fachbereiche untergliedert sind, aber auch Self-Service-Analysen ermöglichen. Auf diese Komponente greifen die Frontend-Werkzeuge zu, um die unterschiedlichen Bedarfe der Fachbereiche wie Reporting, Dashboards, Planung oder Analytics in eine geeignete Visualisierung zu bringen.

3.9. Repository – Metadaten/Data-Lineage

Übergreifend zu allen bisher beschriebenen Komponenten sind ein zentrales Repository sowie weitere zentrale Funktionen notwendig, um die gestiegene Komplexität in der verteilten und heterogenen Datenwelt beherrschbar zu machen:

- Metadaten sind ganz entscheidend notwendig, um alle Daten der analytischen Plattform in ihren richtigen Kontext einordnen und somit die inhaltlichen Bezüge zwischen den einzelnen Datenquellen herstellen zu können.
- Data-Lineage zeigt die Kette an Operationen an einem Merkmal oder einer Kennzahl von der originären Quelle bis zur Business-Sicht. Diese Funktion unterstützt den Fachbereich und die IT in der Fehleranalyse bzw. bei der Erstellung von Berichten.

In einem klassischen SAP-BW-System sind Funktionen wie z. B. die Prozessketten-Implementierung und das zugehörige Monitoring vorhanden. Neben der Integration mit SAP ERP hat SAP BW in der Vergangenheit immer mit diesen eher administrativen Verwaltungsfunktionen gepunktet; allerdings sind diese zentralen Funktionen nur innerhalb des BW nutzbar. Eine ähnliche administrative Oberfläche über Systemgrenzen hinweg unter Einbindung diverser logischer Systeme, Datenbanken und (Cloud) Services ist für eine zukunftsfähige analytische Plattform notwendig – der SAP Data Hub könnte hier zukünftig eine Lösung bieten.

4. Szenarien – ein Beispiel

Die Abbildung unten zeigt ein mögliches Szenario für eine künftige verteilte, analytische Plattform und ordnet die aktuellen SAP-Produkte in die beschriebene Architektur ein. Natürlich gibt es Überschneidungen in den Produkten, so ist etwa das BW/4HANA weit mehr als nur ein Tool zur Persistenz.

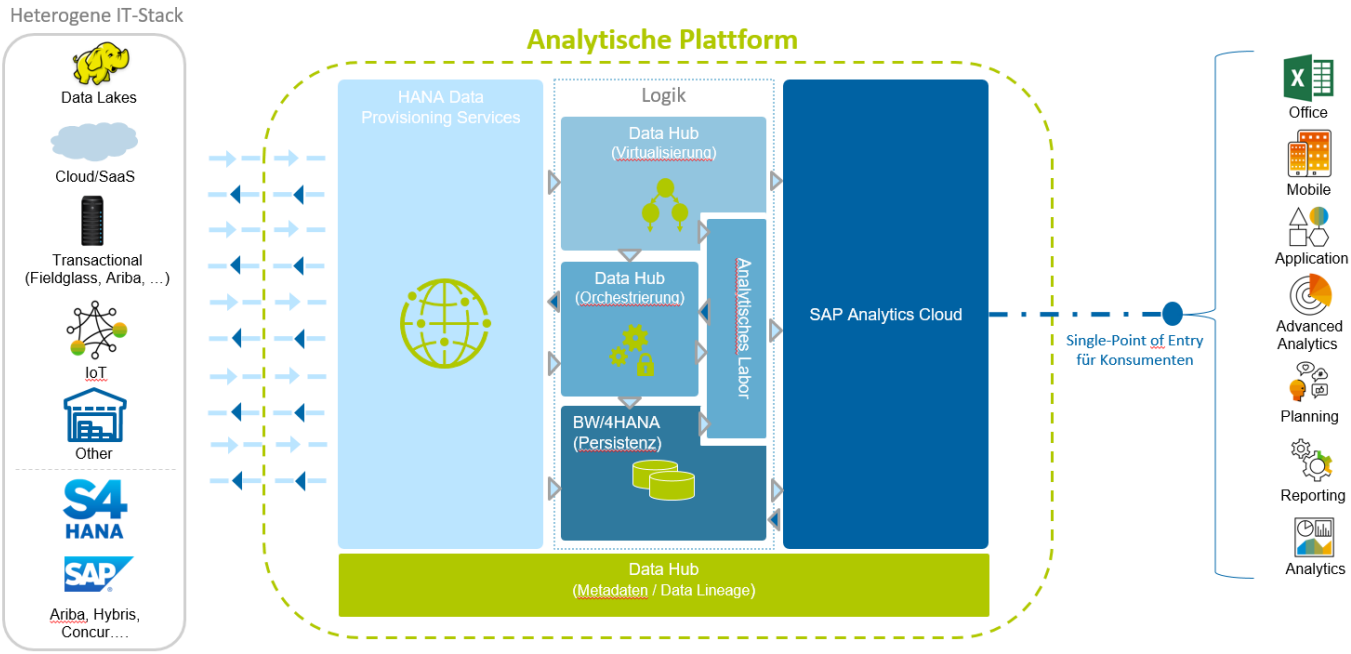

Bild 4 – Exemplarisches Szenario einer dezentralen analytischen Plattform mit SAP-Produkten (SAP, eigene Adaption)

4.1. IT-Stack – SAP ERP, Hadoop & Fieldglass

Zum einem gibt es hier die klassischen operativen Systeme, wie z. B. S/4HANA, Ariba und Fieldglass, oder auch Systeme außerhalb des SAP-Ökosystems, wie Workday. Innerhalb dieser transaktionalen Systeme wird das klassische operative Reporting in Echtzeit abgebildet. Dies ist möglich, solange die Komplexität der Analyse-Views überschaubar ist und keine Datenredundanzen erfordert.

Zum anderen gibt es neben den transaktionalen Systemen die Welten des Internet of Things (IoT), der Data Lakes oder des Big Data mit Sensordaten, Bildern, Videos und anderen mehr oder weniger strukturierten Daten.

4.2. Datenvirtualisierung – SAP Data Hub

Im bestehenden IT-Stack entsteht eine heterogene IT-Landschaft mit diversen Datensilos und diversen Technologien im Kern – die Grundlage für das Net of Truth. Um diese komplexen Datensilos wieder zusammenzuführen, empfiehlt sich eine Plattform zur Datenvirtualisierung wie z. B. der SAP Data Hub, der wiederum zur Integration von Daten auf die Data-Provisioning-Komponenten der SAP-HANA-Plattform wie SDI, SDA etc. zugreift. So lassen sich in dieser Plattform beliebige Datentöpfe konsolidieren und virtuell verknüpfen.

4.3. Persistenz – SAP BW/4HANA und das Cloud Data Warehouse

Für die Anwendungsfälle, die nach wie vor Datenredundanz erfordern, wird ein Data Warehouse betrieben, wie z. B. SAP BW/4HANA oder künftig auch das SAP Data Warehouse Cloud (kurz SAP DWC). Alternativen wären ein (HANA-basiertes) SQL DWH, AWS Redshift oder Google Big-Query. Auch in den IoT- und Big-Data-Welten gibt es Anwendungsfälle, in denen eine Vor-Aggregation bzw. Datenredundanz notwendig ist. Aufgrund der technischen Limitierung im BW und der enormen Mengen an Daten in dieser Umgebung werden Datenredundanzen dort eher auf Hadoop-Clustern gespeichert.

4.4. Datenorchestrierung – SAP Data Hub

Der SAP Data Hub dient zwischen den zwei Welten als Werkzeug zur Datenorchestrierung und verknüpft alle möglichen Datenquellen und datengetriebenen Prozesse zu einem quasi-intelligenten „Datenorganismus“, der eigenständig auf bestimmte Muster reagieren kann. Beispiel: Maschine A besitzt eine Vielzahl an Sensoren, die Daten in Echtzeit liefern. Wenn ein Schwellwert für z. B. Stromverbrauch überschritten wird, kann der Data Hub automatisiert und datengetrieben einen Reparatur- oder Serviceauftrag im S/4HANA-System anlegen.

4.5. Harmonisierte und konsolidierte Business-Sicht – SAP Analytics Cloud

In der SAP Analytics Cloud werden damit neben klassischen Berichten und Analysen auch Echtzeitberichte oder -vorhersagen möglich. Die SAP Analytics Cloud wäre damit auch der zentrale Einstiegspunkt in das heterogene Net of Truth.

4.6. Analytisches Labor

Das analytische Labor ist sowohl Datenempfänger als auch -sender. Orchestrierte Daten lassen sich dort verarbeiten und „intelligent“ für bestimmte, tiefer gehende Analysezwecke wieder zurückspielen. Hier können diverse Technologien für die Zielgruppe „Data Scientist“ Verwendung finden, wie z. B. Keras, Tensorflow, R, Python etc.

5. Zusammenfassung & Ausblick

Das bestehende und weit verbreitete Dogma „Single Point of Truth in einem großen monolithischen ONE-BI-System“ ist aus unserer Sicht überholt und wird mittelfristig von einem Service-orientierten Ansatz abgelöst. Denkt man die neuen Paradigmen „Data as a Service“ und „Net of Truth“ mit einem Single Point of Entry konsequent zu Ende, ergeben sich radikale Änderungen für die Analytik im Unternehmen.

- Die wenigsten Unternehmen haben es in der Vergangenheit geschafft, das eine, zentrale monolithische Data Warehouse für alle Daten im Unternehmen aufzubauen und zu betreiben. Das hatte nicht nur technische oder organisatorische Ursachen, sondern zumeist war dieser Ansatz schlicht wirtschaftlich nicht sinnvoll.
- Die Fokussierung auf Service-orientierte Architekturen wurde bereits vor mehr als zehn Jahren als die Lösung für komplexe Systeme verkauft. Erst mit dem heutigen Trend zur Cloud sind jetzt letztlich alle Softwarehersteller gezwungen, offene APIs für ihre Lösungen anzubieten. Das, kombiniert mit dem technischen Fortschritt der letzten Jahre, macht nun eine effektive Umsetzung von Service-orientierten Architekturen (SOA) möglich.
- Die Verfügbarkeit von standardisierten Schnittstellen erhöht die Wahlfreiheit der Anwenderunternehmen und bietet einen potenziellen Ausweg aus der Abhängigkeit von SAP oder anderen Unternehmen – und sei es nur punktuell für bestimmte Nischenanwendungen, die nun eben auch mit den zentralen Lösungen verknüpft werden können.

Die Herausforderung liegt also nicht so sehr darin, ein für jeden Zweck geeignetes Werkzeug zu finden, sondern vielmehr darin, den Überblick über die vorhandenen Werkzeuge (= Lösungen) und das vorhandene Material (= Daten) zu behalten und eine koordinierte Nutzung für ganz verschiedene Zwecke sicherzustellen.

Unternehmen wollen und müssen in der Lage sein, Anwendungen, Produkte oder Systeme je nach Anforderung bzw. Prozessen frei über die Grenzen der Portfolios einzelner Hersteller hinweg zu orchestrieren. Voraussetzung dafür ist ein gutes API- und Metadaten-Management. An dieser Stelle kommen Lösungen wie der Data Hub ins Spiel, die – wenn auch nicht unbedingt in der aktuell vorliegenden Version 2.4, so doch potenziell in der Zukunft – einen Schlüsselbeitrag zum Erfolg einer verteilten, virtuellen analytischen Plattform leisten können. Aus Sicht von SAP Kunden ist dafür auch eine enge Integration/Verzahnung mit der HANA-Plattform und mit Lösungen wie S/4HANA, BW/4HANA, C/4HANA oder Leonardo ebenso erforderlich wie die Integration mit nativen HANA-SQL-basierten Data-Warehouse-Lösungen und natürlich mit Lösungen, die heute (oft irreführend) unter Begriffen wie „Big Data“ oder „Hadoop“ zusammengefasst werden.

Idealerweise sollte die Entwicklung und Umsetzung einer solchen Architektur in die globale Enterprise-Architektur eingebettet sein, die auf flexible, Cloud-orientierte Architekturen ausgerichtet ist. Oft haben Unternehmen bereits eine solche Architektur für ihre E-Commerce-, B2C- oder B2B-Lösungen entwickelt und damit einige der wesentlichen Aspekte (Entkopplung, Serviceorientierung etc.) bereits umgesetzt. Die Architektur einer analytischen Plattform sollte darauf aufbauen und die eigenen, spezifischen Belange einbringen.

Und wie bereits eingangs erwähnt, wird es spannend sein zu beobachten und zu begleiten, ob und wie sich das gerade angekündigte SAP Data Warehouse Cloud in den nächsten Jahren in Richtung einer umfassenden analytischen Plattform entwickelt.

6. Impressum

Wir weisen ausdrücklich darauf hin, dass das vorliegende Dokument nicht jeglichen Regelungsbedarf sämtlicher DSAG-Mitglieder in allen Geschäftsszenarien antizipieren und abdecken kann. Insofern müssen die angesprochenen Themen und Anregungen naturgemäß unvollständig bleiben. Die DSAG und die beteiligten Autoren können bezüglich der Vollständigkeit und Erfolgsgeeignetheit der Anregungen keine Verantwortung übernehmen.

Deutschsprachige SAP® Anwendergruppe e.V.
Altrottstraße 34 a
69190 Walldorf | Deutschland
Telefon +49 6227 35809-58
Telefax +49 6227 35809-59
E-Mail info@dsag.de

dsag.de