**DMAS Demo-Konzept**

SAP Datasphere und SAP Analytics Cloud

|  |  |
| --- | --- |
| Version: | 1.0 |
| Datum: | 08.01.2025 |
| Status: | **in Arbeit** / vorgelegt / freigegeben |
| Autor: | Irina Khutornova, Robert Dunst |
| Ablage: |  |

Inhalt

[1. Einleitung 3](#_Toc187400937)

[2. Systemarchitektur 3](#_Toc187400938)

[3. Datensätze 3](#_Toc187400939)

[4. Datenquellen 4](#_Toc187400940)

[5. Datenverarbeitung und Speicher (ETL - Prozess) 5](#_Toc187400941)

[6. Reporting und Frontend 5](#_Toc187400942)

[7. Machine Learning 6](#_Toc187400943)

1. Einleitung

Das vorliegende IT-Konzept beschreibt die Entwicklung einer zentralisierten Datenplattform auf Basis von SAP Datasphere und SAP Analytics Cloud. Ziel ist es, eine nahtlose Integration, Transformation und Bereitstellung von Daten aus verschiedenen Quellen zu ermöglichen. Dies soll Unternehmen dabei unterstützen:

* Heterogene Datensätze zu konsolidieren und einheitlich aufzubereiten.
* Echtzeitanalysen und -vorhersagen zu erstellen
* Effiziente und fundierte Entscheidungen durch interaktive Dashboards, Berichte und Machine Learning-gestützte Analysen zu treffen.

1. Systemarchitektur

Die Systemarchitektur von SAP Datasphere ist so konzipiert, dass sie eine leistungsstarke, skalierbare und flexible Datenmanagementplattform bereitstellt, die auf moderne Unternehmensanforderungen ausgerichtet ist. Die Architektur basiert auf einer Cloud-nativen Infrastruktur, die vollständig in die SAP Business Technology Platform (BTP) integriert ist. Dies ermöglicht eine nahtlose Anbindung an SAP- und Nicht-SAP-Systeme sowie die Nutzung von Daten in Echtzeit

Die Architektur besteht aus mehreren Schichten:

1. **Datenintegration:** Diese Schicht ermöglicht die Anbindung und Integration verschiedener Datenquellen, einschließlich On-Premises-Systeme, Cloud-Datenbanken und Drittanbieter-APIs. Sie unterstützt sowohl ETL- als auch Datenvirtualisierungsprozesse, um maximale Flexibilität zu gewährleisten.
2. **Datenmodellierung und -verwaltung:** Die Plattform ermöglicht die Erstellung und Verwaltung von Datensätzen (Datasets), die als Grundlage für Analysen, Vorhersagen und Berichte dienen. Daten können angereichert, transformiert und modelliert werden, um spezifische Geschäftsanforderungen zu erfüllen.
3. **Analysen und maschinelles Lernen:** Die Systemarchitektur unterstützt die Integration von Analyse- und Prognosefunktionen, wie sie in SAP Analytics Cloud (SAC) und maschinellen Lernmodellen (z. B. FedML und SAP HANA ML) realisiert werden.

Ein Architekturdiagramm zeigt die Datenquellen, Layer-Struktur (in SAP Datasphere), ETL-Prozesse und die Verbindung zu SAP Analytics Cloud.

Ein Bild, das Text, Diagramm, Plan, technische Zeichnung enthält.

Automatisch generierte Beschreibung

Bild 1.

Bild soll aktualisiert werden. Ergänzung durch ML-Tools und Daten

1. Datensätze

In SAP Datasphere stellen Datensätze die Grundlage für Analysen, Berichte und datengetriebene Entscheidungen dar. Ein Datensatz ist eine strukturierte Sammlung von Informationen, die aus einer oder mehreren Datenquellen stammen und in Datasphere integriert werden. Diese Datensätze können aus tabellarischen Daten, Zeitreihen oder hierarchischen Strukturen bestehen, die spezifisch für die Geschäftsanforderungen aufbereitet sind.

Mit Hilfe von Datenmodellen in Datasphere können Datensätze definiert, transformiert und angereichert werden. Nutzer können Datensätze aus verschiedenen Quellen miteinander verknüpfen, um einheitliche und umfassende Ansichten zu erstellen, beispielsweise für Finanzberichte, Vertriebsanalysen oder Produktionskennzahlen. Zusätzlich können Datensätze in Echtzeit aktualisiert werden, was ihre Relevanz und Genauigkeit in dynamischen Geschäftsumfeldern sicherstellt.

Eine zentrale Funktion ist die Möglichkeit, Datensätze mithilfe von Business-Semantik zu erweitern, sodass sie intuitiv und verständlich für Endnutzer sind. Dank leistungsstarker Werkzeuge für die Visualisierung und Analyse können Datensätze direkt in Dashboards oder Berichten verwendet werden, um wertvolle Einblicke zu gewinnen und datenbasierte Entscheidungen zu treffen.

Die folgenden Datensätze sollen verwendet werden:

**Beispiel:**

* Verkaufsdaten: Umsätze, Produkte, Regionen, Zeitpunkte (aus internen ERP-Systemen).
* Kundendaten: Demografie, Historie, Präferenzen (aus CRM-Systemen).
* Wetterdaten: Externe API-basierte Datenquellen (z. B. Temperatur, Niederschlag).
* CSV-Daten: Einmalig importierte Daten wie historische Verkaufszahlen oder externe Benchmark-Daten.

1. Datenquellen

SAP Datasphere unterstützt eine Vielzahl von Datenquellen, um eine umfassende Datenintegration und -analyse zu ermöglichen.

**Daten aus SAP-Systemen:**

* SAP S/4HANA: Extraktion von Finanz- oder Logistikdaten, wie z. B. Umsatzberichte oder Bestandsdaten, über vorkonfigurierte Konnektoren.
* SAP BW/4HANA: Zugriff auf vorhandene analytische Modelle oder InfoProvider, um Daten direkt in SAP Datasphere zu importieren.

**Daten aus Nicht-SAP-Systemen:**

* SQL-Datenbanken: Abrufen von Kundendaten aus einer PostgreSQL- oder Oracle-Datenbank für eine Customer-360-Analyse.
* Cloud-Datenbanken: Verbindung zu Amazon Redshift oder Google BigQuery, um Marketing- oder Vertriebsdaten für globale Analysen zu extrahieren.

**Dateibasierte Quellen:**

* CSV- oder Excel-Dateien: Hochladen von Verkaufszahlen oder Budgetberichten, die in lokal gespeicherten Dateien vorliegen.
* JSON- oder XML-Dateien: Extraktion von Daten aus APIs oder Webservices, z. B. zur Integration von externen Marktdaten.

**APIs und Webservices**:

* Abruf von Echtzeitdaten aus Drittanbieter-APIs, wie z. B. Wetterdaten oder Börsenkursen, um sie in unternehmensspezifische Modelle zu integrieren.
* Integration von IoT-Daten über REST- oder OData-Schnittstellen, z. B. Sensordaten aus Produktionsanlagen.

**Cloud-Services:**

* Salesforce: Extraktion von Kundendaten wie Leads, Opportunities und Kontakten zur weiteren Analyse.
* Google Analytics: Abruf von Web-Traffic-Daten zur Integration in Marketing-Reports.

**Streaming-Daten:**

* IoT-Datenströme: Erfassen von Produktionskennzahlen in Echtzeit, z. B. von Maschinen oder Sensoren.
* Event-Daten: Extraktion von Log- oder Transaktionsdaten aus Streaming-Plattformen wie Apache Kafka.

**Manuelle Eingabe:**

* Integrieren von Daten, die durch Benutzer manuell in SAP Datasphere eingegeben oder hochgeladen werden, wie z. B. Ad-hoc-Reports.

1. Datenverarbeitung und Speicher (ETL - Prozess)

Der ETL-Prozess (Extract, Transform, Load) in SAP Datasphere funktioniert in drei klar definierten Phasen, die nahtlos ineinandergreifen, um eine effiziente und zuverlässige Datenintegration und -verarbeitung zu gewährleisten:

**1. Extract (Datenextraktion):**

In dieser Phase werden Daten aus verschiedenen Quellen extrahiert, darunter SAP- und Nicht-SAP-Systeme, Cloud-Datenbanken, On-Premises-Datenbanken und APIs. SAP Datasphere verwendet dafür vorkonfigurierte Konnektoren, die einen sicheren und schnellen Zugriff auf strukturierten und unstrukturierten Daten ermöglichen. Daten können in Echtzeit oder als Batch-Prozess extrahiert werden, abhängig von den Anforderungen des Unternehmens.

**2. Transform (Datenaufbereitung):**

Die extrahierten Rohdaten werden im nächsten Schritt aufbereitet und transformiert, um sie für analytische oder operative Zwecke nutzbar zu machen. SAP Datasphere bietet leistungsstarke Werkzeuge zur Datenbereinigung (z. B. Duplikaterkennung und Anomalieprüfung), zur Anreicherung von Daten durch Verknüpfung mit anderen Quellen und zur Anwendung von Regeln und Geschäftslogik. Mit grafischen Modellierungstools und der Unterstützung von Code-freien Transformationen wird der Prozess vereinfacht. Für komplexere Anforderungen können SQL-basierte Transformationen oder benutzerdefinierte Skripte eingesetzt werden.

**3. Load (Datenladung):**

Die aufbereiteten Daten werden schließlich in SAP Datasphere gespeichert. Hierbei stehen zwei Ansätze zur Verfügung: **physische Speicherung**, bei der die Daten direkt in die Speicherinfrastruktur von Datasphere geladen werden, oder **Datenvirtualisierung**, bei der die Daten in ihrer ursprünglichen Quelle verbleiben, jedoch für Analysen und Berichte bereitgestellt werden. Beide Ansätze gewährleisten die Konsistenz und Aktualität der Daten. Die Daten können sofort für Echtzeitanalysen, Berichte oder maschinelles Lernen genutzt werden.

**Erweiterte Funktionen:**

Zusätzlich zu den Kernschritten bietet SAP Datasphere Funktionen wie Automatisierung von ETL-Workflows, Echtzeit-Synchronisierung, Datenqualitätsprüfungen und die Möglichkeit, komplexe Datenpipelines zu überwachen und zu optimieren. Mit diesen erweiterten Funktionen wird sichergestellt, dass der ETL-Prozess nicht nur zuverlässig, sondern auch flexibel und skalierbar bleibt, um den Anforderungen moderner datengetriebener Unternehmen gerecht zu werden.

1. Reporting und Frontend

SAP Analytics Cloud (SAC) ist nahtlos in SAP Datasphere integriert und ermöglicht es Nutzern, Daten effizient zu analysieren, Visualisierungen zu erstellen und fundierte Geschäftsentscheidungen zu treffen. Diese Integration bietet einen direkten Zugriff auf Echtzeitdaten, die in Datasphere gespeichert sind, und nutzt dabei leistungsstarke Analysewerkzeuge von SAC. Über die Live Data Connection können sich Anwender mit Datasphere verbinden, wodurch Analysen ohne Duplizierung oder Verschiebung von Daten durchgeführt werden können, was die Datenintegrität und Aktualität sicherstellt. Mit SAC können interaktive Dashboards, Berichte und Prognosemodelle basierend auf den in Datasphere aufbereiteten Daten erstellt werden, wodurch Geschäftsplanungs- und Vorhersageszenarien unterstützt werden. Zudem ermöglichen die integrierten Machine-Learning-Funktionen in SAC eine automatisierte Datenanalyse, das Erkennen von Trends und die Erstellung von Vorhersagen. Diese Integration vereinfacht die Arbeit mit großen Datenmengen, beschleunigt Entscheidungsprozesse und sorgt für eine höhere Transparenz in Geschäftsabläufen.

Funktionen:

* Erstellung von interaktiven Dashboards und Berichten.
* Ad-hoc-Analysen basierend auf den bereitgestellten Datenmodellen.
* Darstellung von KPIs, Trends und Vorhersagen.

**Reporting-Inhalte:**

**KPI-Berichte**:

* Umsatz nach Region, Produktkategorie und Zeitraum.
* Top-Kunden und -Produkte.
* Einfluss externer Faktoren (z. B. Wetter) auf die Verkaufszahlen.

**Trendanalysen**:

* Saisonale Schwankungen und langfristige Trends

**Dashboards**:

* Interaktive Visualisierungen mit Filter- und Drilldown-Funktionen.

1. Machine Learning

Die **Machine-Learning-Funktionalität in SAP Datasphere** ermöglicht es Unternehmen, ihre Daten durch den Einsatz fortschrittlicher Algorithmen und KI-gestützter Analysen effektiv zu nutzen. Mit der Integration von Machine Learning können Benutzer Vorhersagemodelle erstellen, um Muster in großen Datenmengen zu erkennen und fundierte Entscheidungen zu treffen. SAP Datasphere bietet eine intuitive Umgebung, die es erlaubt, Machine-Learning-Modelle direkt auf die konsolidierten Daten anzuwenden. Die Plattform unterstützt den gesamten Lebenszyklus eines Modells, von der Datenaufbereitung und Modellbildung bis hin zur Implementierung und kontinuierlichen Optimierung. Durch den Einsatz von Technologien wie neuronalen Netzen und Entscheidungsbäumen können Unternehmen Prognosen, Anomalieerkennungen und personalisierte Empfehlungen generieren, wodurch operative Prozesse effizienter gestaltet werden. Die Plattform selbst unterstützt die Integration von Python-Skripten und Jupyter-Notebooks, was die Entwicklung und den Einsatz maßgeschneiderter ML-Modelle erleichtert.

Der **Python Machine Learning Client für SAP HANA** (hana-ml) erweitert diese Funktionalität, indem er Datenwissenschaftlern ermöglicht, direkt auf SAP HANA-Daten zuzugreifen und verschiedene Machine-Learning-Modelle in Python zu entwickeln. Dieses Paket umfasst zwei Hauptkomponenten: den SAP HANA DataFrame, der Methoden zum Zugriff und zur Abfrage von Daten in SAP HANA bereitstellt, ohne die Daten zum Client zu übertragen, sowie eine Sammlung von Machine-Learning-APIs für die Modellentwicklung. Insbesondere bietet das hana-ml-Paket Zugriff auf die Predictive Analysis Library (PAL) und die Automated Predictive Library (APL) von SAP HANA, die eine Vielzahl von Algorithmen für maschinelles Lernen und automatisierte Analysen bereitstellen. Durch die Nutzung von hana-ml können Unternehmen ihre Daten in SAP Datasphere nahtlos mit Python-basierten Machine-Learning-Workflows integrieren und so die Entwicklung und Implementierung von Modellen beschleunigen.

Eine Abbildung der Architektur ist unten dargestellt:

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

Automatisch generierte Beschreibung

Mit SAP Datasphere können Prognosemodelle durch die Integration mit **SAP Data Intelligence** oder durch die Nutzung eingebauter Machine-Learning-Funktionen erstellt werden. Über SAP Data Intelligence können Daten aus verschiedenen Quellen zusammengeführt, bereinigt und für maschinelles Lernen vorbereitet werden. Die Plattform bietet umfangreiche Tools für die Datenaufbereitung, Feature-Engineering und Modellentwicklung.

Zudem bietet SAP mit **FedML** (Federated Machine Learning) eine Python-Bibliothek an, die es ermöglicht, Machine-Learning-Experimente auf verschiedenen Plattformen durchzuführen, ohne die Daten aus ihren ursprünglichen Quellen replizieren oder migrieren zu müssen. Durch die Nutzung der Datenföderationsarchitektur von SAP Datasphere können Unternehmen Modelle auf Hyperscaler-Plattformen wie AWS SageMaker, Azure Machine Learning und Google Vertex AI trainieren und bereitstellen, was den Bedarf an aufwändigen ETL-Prozessen reduziert und die Datenkonsistenz gewährleistet

**Beispiele:**

* Zeitreihenanalysen: Vorhersage von Verkaufszahlen basierend auf historischen Daten und Wetterdaten.
* Kundensegmentierung: Clustering von Kunden basierend auf Kaufverhalten und Demografie.
* Umsatzprognosen: Regression zur Vorhersage der Verkaufsentwicklung.

**Die Integration von SAP Datasphere mit Databricks.**

Die Integration von SAP Datasphere und Databricks erleichtert es Unternehmen, ihre Datenstrategien zu realisieren und gleichzeitig die Effizienz ihrer Prozesse zu steigern. Mit der Fähigkeit, Echtzeitdaten zu analysieren, Machine-Learning-Algorithmen zu integrieren und präzise Vorhersagen zu treffen, können operative Prozesse entscheidend optimiert werden.

**Mit der FedML-Python-Bibliothek** in Databricks können Machine-Learning-Modelle direkt auf Daten angewendet werden, die in SAP Datasphere gespeichert sind, ohne dass die Daten physisch verschoben werden.

Vorteile:

* Erhöhte Datensicherheit durch Verbleib der Daten innerhalb der SAP-Architektur.
* Direkte Rückspeicherung der Analyseergebnisse in SAP Datasphere zur Weiterverarbeitung und Visualisierung in SAP Analytics Cloud.

**Vorgehensweise:**

SAP Datasphere bietet vorgefertigte Konnektoren, die es ermöglichen, Daten aus SAP-Quellen wie SAP HANA, S/4HANA oder BW/4HANA sowie aus Nicht-SAP-Systemen zu integrieren.

Databricks stellt APIs und Konnektoren bereit, um diese Daten in das Lakehouse-Modell zu integrieren.

Über die Open-Dataset-Schnittstelle von SAP Datasphere können Daten direkt von Databricks genutzt werden.

Der Datenaustausch erfolgt über standardisierte Schnittstellen (z. B. JDBC/ODBC, REST APIs) oder Delta Lake, ein Databricks-Format für optimierte Abfragen.

Ein Bild, das Text, Screenshot, Schrift, Webseite enthält.

Automatisch generierte Beschreibung