Softwarelösung für Datenkreise

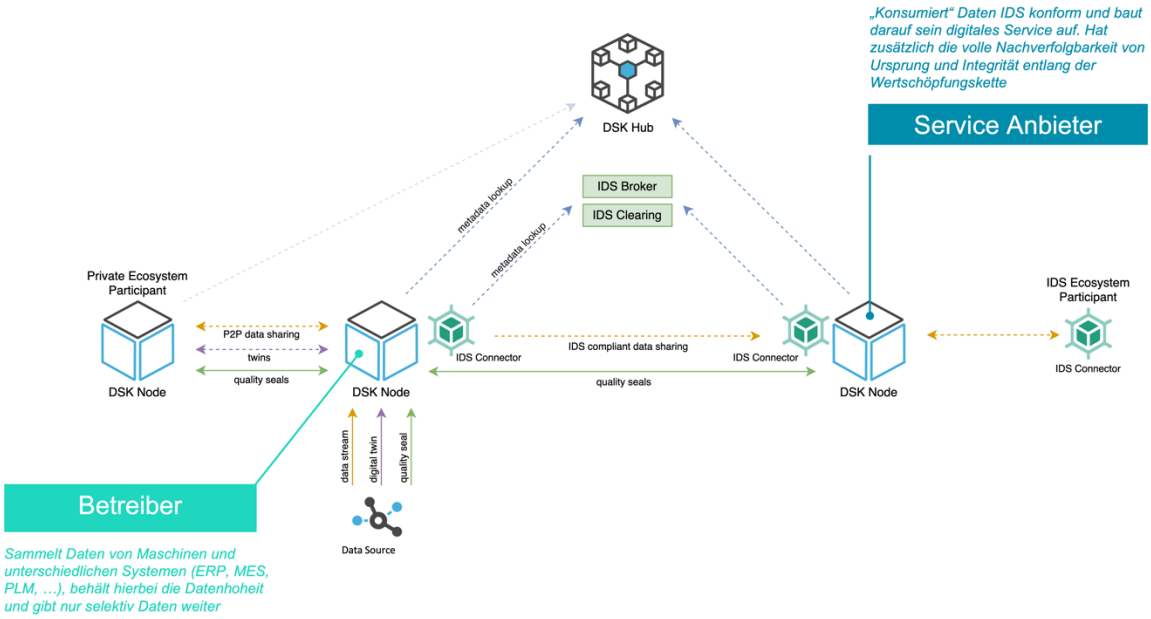
IÖB Challenge

Einleitung

Die Technologie der Tributech Solutions GmbH (kurz Tributech) ermöglicht es Unternehmen, Daten unternehmens- oder prozessübergreifend auf selektive, manipulationssichere Weise und unter Wahrung der Datenhoheit zu sichern und weiterzugeben. Die Technologie wurde nach der Referenzarchitektur gemäß DIN Spec 27070 entwickelt und ist somit mit den Guidelines der Initiativen Gaia-X, Catena-X sowie der International Data Space Association kompatibel. Die integrierte Blockchain-Technologie stellt sicher, dass den Daten vertraut werden kann und niemand die Daten manipuliert hat.
Um den kompletten Daten-Life-Cycle abzubilden, ergänzt die RISC Software GmbH (kurz RISC) ihre Expertise in der Datenhaltung sowie Data Engineering, in der Datennachverarbeitung (Data Analytics) einzelner Unternehmen als auch relevanter Use-Cases.

Lösungsansatz und Softwarearchitektur

Die Umsetzung und der Betrieb von Datenkreisen bzw. Dataspaces für den unternehmensübergreifenden Datenaustausch sind in der Praxis mit hoher Komplexität und hohem Aufwand verbunden. Neben Security, Trust und Datenhoheit sind der sogenannte „erste und letzte Meter“ und Daten zu integrieren sowie zu konsumieren meist die größte Herausforderung. Genau dort kann die Lösung von Tributech ihre Stärken ausspielen. In den folgenden Absätzen werden die relevanten Komponenten und Aspekte beschrieben, die für einen erfolgreichen Betrieb von Datenkreisen erforderlich sind.



Broker & Clearing House - DataSpace Hub

Bevor Daten zwischen zwei Teilnehmer*innen ausgetauscht werden können ist ein entsprechendes Match-Making sowie anschließendes Clearing notwendig, um zu vereinbaren welche Daten unter welchen Bedingungen ausgetauscht, synchronisiert und dokumentiert werden sollen. Der DataSpace Hub von Tributech bietet dazu einen Metadatenbroker sowie eine Clearing House Funktionalität, um diese Aufgabe mittels Match-Making zwischen Datenanbieter*innen und Datenkonsument*innen durchzuführen. Verfügbare Datenquellen werden dabei als Datensets beschrieben und über den Metadatenbroker als Datenkataloge indexiert. Über das Match-Making können die Bedingungen zwischen Datenanbieter*innen und Datenkonsument*innen vereinbart werden.

Über Audit Logs wird in Clearing House Komponente dokumentiert, dass diese Daten auch tatsächlich unter den vereinbarten Bedingungen ausgetauscht werden.

Connector - DataSpace Node

Damit Daten zwischen den Systemen zweier Teilnehmer*innen ausgetauscht werden kann, wird ein „Connector" auf beiden Seiten benötigt. Dieser verbindet die Datenanbieter*innen direkt mit dem Datenkonsument*innen und die Daten werden unter den vereinbarten Bedingungen synchronisiert. Eine wichtige Anforderung an diese Komponente ist die Möglichkeit der plattformunabhängigen Installation und Integration, damit diese für alle potenziellen Teilnehmer*innen in einem Datenkreis nutzbar ist. Die Tributech DataSpace Node bietet dazu ein Data-Sharing Gateway mit dem Datenquellen einfach in jeder Cloud-, Hybrid- oder On-Premises-Datenplattform installiert und integriert werden kann. Damit ein vertrauenswürdiger, nachvollziehbarer und sicherer Datenaustausch zwischen den Teilnehmer*innen erfolgen kann beinhaltet diese Komponente folgende Core-Services.

Data Governance Service

Um Daten auszutauschen, wird – wie bereits erwähnt – ein entsprechendes Match-Making zwischen Datenanbieter*innen und Datenkonsument*innen benötigt. Das Governance Service bietet dazu eine einfach zu bedienende Weboberfläche worüber sogenannte „Data Contracts" erstellt werden können. Ein Data Contract definiert dabei welche Datenpunkte bzw. Datenströme über welchen Zeitraum und zu welchen Bedingungen ausgetauscht werden können. Die Informationen der Data Contracts sowie der Nachweis der tatsächlichen Durchführung werden in entsprechenden Audit Logs dokumentiert. Damit können Daten nachweislich unter Wahrung von Datenhoheit bzw. Souveränität zwischen Anbieter*innen und Konsument*innen ausgetauscht werden.

Storage Service

Das Storage Service bietet einen integrierten Datenspeicher und intelligenten Cache, der dazu dient Daten zu puffern bzw. zwischenzuspeichern. Durch die flexiblen Konfigurationsmöglichkeiten kann der Lebenszyklus je Datenstrom eingestellt werden und einfach zwischen internen und externen Datenbanken für ein effizientes Datenmanagement synchronisiert werden. Zusätzlich bietet der interne Datenspeicher den Vorteil das dieser für die Komplexität der Verarbeitung von geteilten Daten stark reduziert und auch als primärer Datenspeicher genutzt werden kann, wenn dieser bei den Datenanbieter*innen oder Datenkonsument*innen nicht vorhanden ist.

P2P Sync Service

Für die Datensynchronisierung zwischen den Datenanbieter*innen und Datenkonsument*innen übernimmt das P2P Sync Service die Datenübertragung mittels Streamingtechnologie. Die Daten werden dabei auf Basis der vereinbarten Data Contracts ausgetauscht.

Trust Layer Service

Ein zentraler Aspekt im plattform- und unternehmensübergreifenden Datenaustausch ist die Vertrauenswürdigkeit und Echtheit der Daten. Dazu bietet die Technologie die Möglichkeit Datenqualitätssiegel in Form von signierten Hashes je Datenpunkt bzw. Datensatz zu erstellen und diese in einem auf Blockchain Technologie basierenden Trust Layer zu speichern. Damit können zwischen Anbieter*innen und Konsument*innen bzw. Datenquelle und Applikation auf Integrität, Herkunft und Ownership geprüft werden. Dies bietet die Möglichkeit Daten in verifizierbare Assets zu transformieren welche Datenkonsumenten ein überprüfbares Gütesiegel bieten.

Datenhaltung

Bei der Datenhaltung werden Daten in drei unterschiedliche Kategorien unterteilt:

Daten und Datenströme

Diese beinhalten die tatsächlichen Daten einer Datenquelle wie z.B. Sensordaten eines IoT Devices oder Dokumente aus einem ERP System. Diese Daten sind immer in der Hoheit ihrer Inhaber*innen und werden

dezentral bzw. in der Plattform des jeweiligen Stakeholders gespeichert. Bei erfolgreichem Abschluss eines Data Contracts werden die ausgewählten Daten direkt peer-to-peer in das Zielsystem der Datenkonsument*innen übertragen.

Metadata
Metadaten umfassen Beschreibungen von Datensets und Datenquellen sowie Einträge für Data Contracts, Status der Datensynchronisierung sowie weitere Stamm- und Userdaten. Metadaten werden im DataSpace Hub bzw. in der DataSpace Node gespeichert. Zukünftig ist geplant das Metadatenmanagement (wo sinnvoll) weiter zu dezentralisieren.

Qualitätssiegel
Wie bereits beschrieben bestehen Qualitätssiegel für Datenströme und Datensätze aus signierten Hashes. Diese werden in einer Blockchain-basierten Datenbank manipulationssicher gespeichert und auf die entsprechenden Daten, die damit geprüft werden können, referenziert. Je Datenraum kann ein eigenes Blockchain Netzwerk betrieben oder ein bestehendes (z.B. Public Blockchains) integriert werden. Die Technologie stellt dabei sicher das auch hohe Datenmengen skalierbar und effizient verarbeitet werden können.

Integration von Daten & Systemen

Für die Datenintegration stehen Open API-Schnittstellen für Daten und Metadaten sowie ein MQTT Message Bus zur Verfügung. Alle komplexen Aufgaben für das Datenmanagement werden dabei bereits durch die Lösung abgedeckt und Datenquellen können in den meisten Fällen innerhalb von Stunden integriert werden. Zusätzlich bietet ein System für Data-Pipelines und Workflows die Möglichkeit eigene Schnittstellenkataloge zu entwickeln. Darüber hinaus können eine Vielzahl an unterstützte Partnerlösungen im Cloud- und IoT-Bereich genutzt werden. Damit ist es möglich dieselben Basisdaten verschieden aufzubereiten oder je nach Anforderung und Kund*in verschiedene Untermengen zu bilden. Die Daten verbleiben dabei unter der Kontrolle der Datenanbieter*innen, womit die Anbindung an den Datenkreis zum Datenaustausch mit Kund*innen genutzt wird.

Unterstützte Standards

Um Daten standardisiert und plattformübergreifend in Datenkreisen auszutauschen, benötigt es offene Standards welche Interoperabilität und einfache Erweiterbarkeit des Systems gewährleisten. Das Tributech DataSpace Kit bietet dazu offene API-Schnittstellen nach dem OAS3 Standard für die Datenintegration sowie Automatisierung des Systems. Folgende ausgewählte Standards, welche speziell für Datenkreise große Vorteile mitbringen, werden hier hervorgehoben:

IDS Reference Architecture nach DIN Spec 27070
Die Referenzarchitektur für den sicheren und souveränen Datenaustausch wurde von der International Dataspace Association und ihren Mitgliedern erarbeitet. Das Tributech Dataspace Kit wurde nach der Referenzarchitektur gemäß DIN Spec 27070 entwickelt und befindet sich in Vorbereitung zur Zertifizierung, sobald diese erstmals verfügbar ist. Als Mitglied der IDSA ist Tributech zusätzlich in relevanten Arbeitsgruppen für die Architektur, Zertifizierung und Anwendungsfälle vertreten.

Standardisierte Beschreibung von Datenquellen
Um Datenquellen und Datensets standardisiert beschreiben zu können, kommt die Standards Digital Twin Definition Language (DTDL) bzw. Web of Things (WoT) zum Einsatz, um eine Beschreibung in Form eines Digitalen Zwillings zu erstellen. Dies ermöglicht eine einheitliche und maschinenlesbare Beschreibung von Datenquellen und Datensets inkl. aller erforderlichen Kontextinformationen, welche für eine automatisierte Integration und Nutzung der Daten erforderlich sind.

Datenaustausch vs. Datenhandel über Marktplätze

Ein vielfach diskutiertes Thema im Kontext von Datenkreisen und Dataspaces ist der Handel bzw. Verkauf von Daten. In vielen Bereichen oder Industrien ist es jedoch nicht möglich (oder noch nicht möglich) einen Datenpunkt oder Datensatz direkt zu bewerten. Ein messbarer Wert wird oft erst durch den Einsatz eines Datenservices geschaffen. Wie auch in anderen Branchen sehen wir einen Bedarf für domainspezifische Datenmarktplätze wie

z.B. für Finanzdaten. Aus diesem Grund empfehlen wir im Kontext von Datenkreisen zwischen zwei Bereichen zu unterscheiden:

Datenaustausch in der Wertschöpfungskette
Dabei werden Daten in Datenkreisen unabhängig vom Preis mit Kund*innen, Lieferant*innen oder Partner*innen entlang der Wertschöpfungskette bzw. im eigenen Ökosystem für die Nutzung durch Digitale Services und Algorithmen ausgetauscht. Die Verrechnung erfolgt dabei auf einer anderen Ebene durch das Preis- bzw. Geschäftsmodell der entwickelten Services. Da auf das bestehende Netzwerk der Wertschöpfungskette zurückgegriffen werden kann, stellt dies für die meisten Unternehmen einen einfachen Einstiegspunkt für eine Teilnahme an Datenkreisen dar.

Datenaustausch über domainspezifische Marktplätze
Für den Verkauf von Daten können bestehende Marktplätze genutzt werden bzw. neue etabliert werden. Die Herausforderung ist dabei einerseits die Bewertung der Daten sowie die Generierung von Angebot und Nachfrage in neuen Marktplätzen. Der Bedarf an Trainingsdaten für Modelle der Künstlichen Intelligenz (KI) könnte dabei ein entscheidender Treiber sein, der solche Marktplätze ermöglicht.

Damit Daten innerhalb der Wertschöpfungskette genutzt werden können sowie zukünftig auch auf domainspezifischen Marktplätzen angeboten werden können, bietet das Tributech DataSpace Kit Unternehmen die Möglichkeit an mehreren Datenkreisen und Marktplätzen teilzunehmen bzw. diese zu integrieren. Datenmarktplätze sowie Pricing Systeme können dabei auf Basis der verfügbaren API-Schnittstellen umgesetzt bzw. durch Drittanbieter*innen angeboten werden.

Mehrwerte für Stakeholder (inkl. Use-Cases)

Datenanbieter*innen

Für Datenanbieter*innen ergeben sich aus der Nutzung von Datenkreisen mehrere Vorteile. Diese umfassen ein einheitliches Management ihrer Daten, welches sowohl die Steuerung der Zugriffsberechtigungen als auch die technische Umsetzung des Datenzugriffs umfasst. Damit erhalten sich die Datenanbieter*innen die Datenhoheit und -souveränität, da die Daten nur unter den Bedingungen des vereinbarten Data Contracts geteilt werden.
Beispiele von Stakeholdern, die von einer datenkreisbasierten Lösung besonders profitieren, sind Firmen, deren Geschäftsmodell den Verkauf von Daten einschließt, wie beispielsweise Wetterdatenanbieter*innen. Diese können ohne zusätzlichen technischen Aufwand ihre Daten neuen Kund*innen zur Verfügung stellen oder das Datenangebot für Bestandskund*innen anpassen.
Weitere Schlüsselvorteile für den Einsatz von standardisierten Datenkreislösungen ist die Gewährleistung einer sicheren Datenbereitstellung, damit unberechtigte Dritte keinen Zugriff auf die Daten erhalten können.
Seitens der Datenanbieter*innen wird bei der Datenbereitstellung bisher oft mit ad-hoc Lösungen gearbeitet, wie beispielsweise die Daten auf einem Web- oder FTP-Server zur Verfügung zu stellen. Einerseits können diese Systeme zwar grundsätzlich gegen den Zugriff unberechtigter Dritter abgesichert werden, andererseits ist die Flexibilität der Zugriffssteuerung eingeschränkt. Ein gutes Beispiel dafür ist die zeitliche Beschränkung der Zugriffsberechtigung.

Datenkonsument*innen

Für Datenkonsument*innen ergibt sich aus der Nutzung von Datenkreisen, die Möglichkeit standardisiert auf verschiedene Datenquellen zugreifen zu können beziehungsweise auch zwischen Datenanbieter*innen zu wechseln. Einheitliche Schnittstellen vermeiden für die Datenkonsument*innen den Vendor-Lock-In und ermöglichen so einen Wettbewerb der Datenanbieter*innen.
Darüber hinaus ermöglichen Datenkreise eine strukturierte Zugriffsmethodik, die die Wiederverwendbarkeit von kundenseitigen Anwendungen für die Dienste von verschiedenen Datenanbieter*innen gewährleistet.
Die Vertrauenswürdigkeit der bezogenen Datenströme und -sätze kann kundenseitig jederzeit durch das entsprechende Gütesiegel überprüft werden, womit die Vertrauenswürdigkeit der Daten für die Datenkonsument*innen gegeben ist.

BMK

Seitens des BMK (Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie) ergibt sich der Vorteil, dass die hier beschriebene Datenkreislösung sich auch für die Zusammenarbeit im Rahmen von BMK betreuten Forschungsprojekten einsetzen lässt. Beispielsweise kann der Datenaustausch im Rahmen der Beantragung sowie der Berichtslegung so abgewickelt werden. Der Einsatz einer Datenkreis-basierten Lösung im Rahmen von Forschungsprojekten kann auch zu einem breiteren Einsatz solcher Lösungen in der wirtschaftlichen Interaktion von nicht-öffentlichen Akteur*innen führen.
Ein weiterer Aspekt, welcher für das BMK von Interesse sein dürfte, ist die europäische Sichtbarkeit einer solchen Lösung. Durch die Konformität mit europäischen Standards wird auch ein Datenaustausch auf europäischer Ebene ermöglicht. Unter Einsatz der hier vorgeschlagenen Lösung kann das BMK unter Berücksichtigung des europäischen Data Governance Acts aktiv Open Data Strategien verfolgen.

Use Cases

Grundsätzlich ist der Einsatz von Datenkreisen in allen Bereichen, wo Datenaustausch über Organisationsgrenzen hinweg stattfindet, interessant. Im Besonderen gilt das, sofern der Datenaustausch im Rahmen einer wirtschaftlichen Transaktion stattfindet.
Einen typischen Anwendungsfall stellen beispielsweise die Anbieter*innen von Wetterdateninformation dar, die ihren Kund*innen Wetterinformationen für bestimmte Orte und Zeitbereiche anbieten. Über eine Datenkreislösung können die Datenanbieter*innen selektiv Daten für verschiedene Kund*innen freischalten, ohne die Daten duplizieren zu müssen. Im Besonderen müssen verschiedenen Kund*innen verschiedene Untermengen der Daten angeboten werden, im Falle von Wetterdaten, zum Beispiel, unterschiedliche Mengen an Orten, Zeitbereichen oder Genauigkeiten.

Ein anderer Anwendungsfall betrifft den Datenaustausch zwischen Maschinenhersteller*innen und ihren Kund*innen. Durch diesen erhalten die Maschinenhersteller*innen Feedback über die Nutzung ihrer Maschinen und sie sind damit in der Lage die Maschinen gemäß den Kundenanforderungen zu verbessern. Da die Datenweitergabe an die Maschinenhersteller*innen aber oft kundenseitig umstritten ist, könnte in diesem Kontext auch eine Vergütung für die Kund*innen angedacht werden, wie beispielsweise Preisreduktionen oder Unterstützung bei der laufenden Wartung. Auch diese Transaktion könnte im Rahmen des Datenkreises abgebildet werden. Dieser Anwendungsfall hat die Charakteristik, das Daten potenziell in beide Richtungen fließen können, und damit sowohl für die Anbieter*innen als auch den Kund*innen Themen des sicheren Datenzugriffs relevant sind. Ein weiterer Vorteil des Einsatzes von Datenkreisen in diesem Kontext ist es, dass die Datenkreisplattform auch zur gegenseitigen Verrechnung der weitergegebenen Daten eingesetzt werden kann.

Ein dritter Use Case betrifft Datenkreise für Logistikdienstleister*innen, welche vor der Herausforderung stehen, die Ware pünktlich, kosteneffizient und emissionsarm an den Zielort zu transportieren. Auf die Resilienz der geplanten multimodalen (Straße, Schiene und Schiff) Wege hat nicht nur die Pünktlichkeit und Kapazität der einzelnen Verkehrsmodi einen Einfluss, sondern auch Verkehrsbedingungen, Unfälle, Baustellen, Umwelteinflüsse wie Wetter, Wasserstände und Krisen. Durch Erfassung und Verknüpfung dieser bereits verfügbaren Daten entlang der Logistikkette können Erkenntnisse und Prognosen gewonnen und darauf aufbauend KI-basierte Systeme entworfen werden, welche auf Änderungen autonom reagieren. Es lässt sich damit die Resilienz der Logistikketten für Unternehmen steigern, positive Auswirkungen auf die Umwelt erzielen und auch zukünftige Initiativen wie das autonome Fahren können auf diesen Daten aufbauen.

Fazit

Die in unterschiedlichen Branchen eingesetzte Technologie von Tributech bietet einen standardisierten und vertrauensvollen Unterbau für Datenservices, AI/ML-Modelle oder Plattformen. Das selektive Teilen der Daten sowie die Nachvollziehbarkeit und Auditierbarkeit bietet sowohl dem Datenprovider als auch dem Datenkonsumenten ein hohes Maß an Vertrauen und Datenhoheit.
Dieser technologische Unterbau ermöglicht zB. RISC mit ihren Partner*innen eine deutliche Reduzierung des Aufwandes beim Aufsetzen von Data Pipelines sowie das Umsetzen neuer Use Cases und Applikationen.