REPLY DATA

Data Governance

DATENKATALOGE MIT OPEN SOURCE TOOLS ERSTELLEN.

Information ist heutzutage der wertvollste Rohstoff und es ist fast unmöglich, sich ein Unternehmen vorzustellen, das nicht mit großen Mengen an Datensätzen arbeitet. Einkäufe, Warensendungen, Websitebesuche, Werbeanzeigen und viele andere Bereiche sind Quellen. Die Inhalte werden verarbeitet und analysiert, um die Unternehmens-Strategien anzupassen und erfolgreich am Markt zu konkurrieren.

Während die Anzahl der Datensätze steigt, fällt es Unternehmen immer schwerer die Datenmengen zu verwalten und Individuen sowie Teams zu identifizieren, die für einen konkreten Datensatz zuständig sind. Dies ist der Zeitpunkt, an dem nach Lösungen zur Bewältigung der Datenmengen gesucht wird.

DATENKATALOGE LÖSEN WEITVERBREITETE PROBLEME MIT DATENSÄTZEN

Es gibt viele Tools, die das einfache Ermitteln und Verwalten von Datensätzen unterstützen. Darunter fallen insbesondere Datenkataloge.

Ein Datenkatalog ist ein Bestandsverzeichnis von Datensätzen, der auf dem Metdatenlevel arbeitet und diverse Möglichkeiten, wie etwa die Datenermittlung, Data-Lineage oder Datenbesitz bietet. Datenkataloge können folgende Problemstellungen, die in Unternehmen immer wieder anzutreffen sind, lösen:

- **Information über verfügbare Datensätze fehlt.**
  Dies ist eine Situation die sich häufig bei größeren Organisationen mit mehreren Abteilungen, welche sich mit Daten beschäftigen, finden lässt. Jedes Team kennt nur die Datensätze, mit dem es direkt arbeitet und eine vollständige Liste aller Datensätze des Unternehmens ist nicht verfügbar. Der Mangel einer Übersicht aller verfügbarer Datensätze kann dazu führen, dass verschiedene Teams dieselbe Arbeit mehrfach erledigen und Duplikate vorhandener Datensätze erzeugen. Ebenso ist es möglich, dass Teams ihre Aufgaben nicht erfüllen können, da die Teammitglieder nicht wissen, dass die benötigten Datensätze überhaupt existieren.

- **Beschreibungen der Datensätze sind veraltet.**
  Dieses Problem entsteht in der Regel, wenn die Dokumentation der Daten die niedrigste Priorität hat. Selbst dann, wenn das Unternehmen eine Liste von Datensätzen hat, veraltet die Beschreibung der Daten sehr schnell. Dies führt dazu, dass sich niemand auf die Beschreibungen verlässt und diese nicht genutzt oder dass veralteten Beschreibungen eingesetzt werden, was wiederum zu fehlerhaften Lösungen führt.

- **Erforderliche Datensätze können nicht gefunden werden.**
  Eine Liste von Datensätzen zu haben ist wenig hilfreich, wenn dennoch ein konkreter Datensatz nicht ohne Weiteres gefunden werden kann. Bei einigen Unternehmen ist diese Information auf internen Wiki-Seiten verfügbar. Dennoch kann es Probleme bereiten, nach einem Datensatz zu suchen, wenn es davon Hunderte oder Tausende gibt.

- **Inhaber der Datensätze sind unbekannt.** Wenn ein Unternehmen über hunderte Datensätze und mehrere Teams verfügt, die mit diesen Datensätzen arbeiten, ist es eine Herausforderung, den Besitzer eines bestimmten Datensatzes zu identifizieren. Dies erschwert es, Informationen über Datenaktualität, Änderungsanfragen oder zum Datensatz zu erhalten.

- **Abhängigkeiten eines Datensatzes können nicht identifiziert werden.** Oft ist es für den Besitzer eines Datensatzes nicht einfach nachzuvollziehen, wer die Daten nutzt. Dies ist ein Problem, wenn der Datensatz geändert werden muss. Ohne die Abhängigkeiten zu kennen, ist es unmöglich festzustellen, wer von dieser Änderung betroffen sein wird.

DATENKATALOGE ERSTELLEN – EINE ÜBERSICHT ÜBER OPEN SOURCE TOOLS

Es gibt zahlreiche Plattformen und Dienste, welche die Möglichkeit bieten, Datenkataloge zu erstellen. Alle hier gezeigten Lösungen sind Open-Source-Projekte. Open-Source-Produkte zu nutzen bietet viele Vorteile, wie etwa den Community Support, Sicherheit, ein hohes Maß an Innovation und die Möglichkeit, den Code abzuändern und das Produkt den Bedürfnissen des Unternehmens anzupassen. Die untenstehende Tabelle beinhaltet

Projekt	Initial Commit	Merged PRs (1m)	Open PRs (1m)	Forks	Stars	Watches	Contributors
Apache Atlas	12/07/2014	0	2	375	605	55	93
LinkedIn DataHub	11/19/2019	30	5	640	2100	187	51
Lyft Amundsen	05/14/2019	26	1	217	1100	234	41
Marquez	06/06/2018	14	0	42	341	25	24

H Marquez ist in Teilprojekte unterteilt, die Tabelle enthält nur Statistiken zu den Kerndiensten. Deshalb können die Zahlen für dieses Projekt etwas niedriger ausfallen als für andere Projekte. In den folgenden Abschnitten werden die einzelnen Tools genauer betrachtet.

APACHE ATLAS

Die Zusammenstellung von Kerndiensten bietet Möglichkeiten zur Metadatenverwaltung und Governance. Das Projekt wurde 2015 als Apache Incubator-Projekt ins Leben gerufen. Seither hat das Projekt viele interessante Features erhalten, wie etwa hohe Verfügbarkeit und Benachrichtigungen.

Repositorium	https://github.com/apache/atlas	
Lizenz	Apache 2.0	
Tech Stack	Java, HBase, Solr, JanusGraph, Kafka	
Quellen	HBase, Hive, Sqoop, Storm, Kafka	
Integration	REST API, Kafka	
Features	Generische Typen, Data-Linage, Data Ownership, Benachrichtigungen, Tags, hohe Verfügbarkeit	

Intern verwaltet Apache Atlas Objekte mithilfe von JanusGraph. JanusGraph ist eine skalierbare Graphdatenbank, welche steckbare Datenspeicher unterstützt. Dies ermöglicht es, verschiedene Datenspeicherlösungen zu verwenden. Zum Beispiel können die Daten in HBase oder Solr gespeichert werden, wobei auch Apache Cassandra und ElasticSearch als Alternativen verwendet werden können. Apache Atlas unterstützt das Abfragen von Metadaten aus einer Vielzahl von Datenquellen. Zudem ist es möglich, die Integration anderer Datenquellen durch die REST API oder Kafka hinzuzufügen.

Die Dokumentation des Projekts ist lückenhaft, dennoch verleiht sie eine gute Übersicht über die Architektur des Projekts und die enthaltenen Features. Die Möglichkeit ein Modell für die Metadatenobjekte zu definieren und Instanzen dafür zu erstellen, sowie der fortschrittliche Autorisierungsmechanismus, der durch Apache Ranger bereitgestellt wird, machen Apache Atlas zu einem Produkt der Enterprise-Klasse.

LINKEDIN DATAHUB

Die Plattform zur Suche und Ermittlung von Metadaten wurde 2019 von LinkedIn als Open Source freigegeben. So wie viele Open Source Produkte wurde auch DataHub als internes Produkt des Unternehmens entwickelt.

Repositorium	https://github.com/linkedin/datahub
Lizenz	Apache 2.0
Tech Stack	Java, Ember, Play, MySQL, Kafka, Pegasus, Neo4j, ElasticSearch, etc.
Quellen	HBase, Hive, Sqoop, Storm, Kafka
Integration	REST API, Kafka
Features	Generische Typen, Data-Linage, Data Ownership, Benachrichtigungen, Tags, hohe Verfügbarkeit

Dies ist nicht das erste Verwaltungstool für Metadaten das von LinkedIn erschaffen wurde. Es war als Nachfolger von WhereHows gedacht, einem anderen Produkt des Unternehmens welches 2016 der Community vorgestellt wurde.

Dies ist nicht das erste Verwaltungstool für Metadaten das von LinkedIn erschaffen wurde. Es war als Nachfolger von WhereHows gedacht, einem Produkt des Unternehmens das 2016 der Community vorgestellt wurde.

Im Backend baut DataHub auf die Generalized Metadata Architecture (GMA) auf, welche durch mehrere Dienste vertreten ist. Dank der entkoppelten Architektur erlaubt die Plattform Teams ihre eigenen Metadatendienste zu verwalten und Daten zu sammeln. DataHub kann mit einer Vielfalt von Datenquellen integriert werden – wie etwa mit RDBMS, Hive, Kafka über REST-API-Abrufe oder Kafka Events. Um die Metadaten zu erhalten nutzt die Plattform verschiedene Speicherlösungen, wie etwa ElasticSearch und MySQL.

Das Produkt sollte wegen der vielen interessanten Features, wie etwa Data-Lineage, Data Lifecycle in Betracht gezogen werden. Die Dokumentation des Projekts ist sehr kurzgefasst. Aus diesem Grund sollten Teams, die sich für DataHub in ihrem Unternehmen entscheiden, darauf vorbereitet sein, Zeit für das Auffinden von Informationen zum Deployment und der Nutzung der Plattform in der Codebasis zu investieren. DataHub kann von jedem mithilfe der Docker Images aus dem GitHub Repositorium ausprobiert werden.

LYFT AMUNDSEN

Diese Datenermittlungs- und Metadatenplattform wurde von dem Ridesharing-Unternehmen Lyft entwickelt. Das Unternehmen gab das Projekt 2019 als Open Source frei und präsentierte es der Community. Die Plattform strebt an, das Datenermittlungsproblem zu lösen und bietet verschiedene Features, wie etwa Data Ownership und Data-Lineage.

Repositorium	https://github.com/lyft/amundsen
Lizenz	Apache 2.0
Tech Stack	Python, Node, Flask, React, Neo4j, ElasticSearch
Quellen	Hive, AWS Glue, Cassandra, BigQuery, Druid, Snowflake, RDBMS
Integration	Python library
Features	Data-Lineage, Data Ownership, Datenvorschau, Frequent Users, Tags

Amundsen bietet für das Sammeln von Metadaten eine eigene Herangehensweise. Hierfür wird eine Bibliothek zur Aufnahme von Metadaten, genannt DataBuilder, angeboten. DataBuilder hat eine modulare Struktur, welche zur Beschreibung von ETL Jobs zur Aufnahme von Metadaten verwendet werden kann. Die Bibliothek unterstützt eine Reihe von Datenquellen wie beispielsweise Hive oder Kafka. Eine möglicher Anwendungsfall besteht in der Nutzung von AirFlow Pipelines zur Ausführung der Datenaufnahme. Nach Angaben von Lyft verwendet das Unternehmen AirFlow dazu, die internen Metadaten täglich zu erneuern.

Wie auch bei einigen Konkurrenten wird Amundsens Persistence Layer von Neo4j betrieben und ElasticSearch für den Suchdienst verwendet. Interessanterweise lässt die Plattform zu, Apache Atlas als Backend für den Metadatendienst zu verwenden. Einer der Vorteile von Apache Atlas statt Neo4j als Metadatendienst ist, dass letzterer Plugins anbietet, welche push-basierte Updates erlauben. Dieser Ansatz ermöglicht es auch Apache Ranger Richtlinien zu verwenden, um festzustellen, welche Metadaten zum Betrachten und Bearbeiten verfügbar sind.

MARQUEZ

Marquez ist ein Open-Source-Metadatendienst zum Sammeln, Aggregieren und Visualisieren der Metadaten eines Datenökosystems. Marquez wurde 2018 von WeWork veröffentlicht und als Open Source freigegeben. Zusätzlich zum eigentlichen Dienst enthält das Projekt eine Vielzahl von Integrationsbibliotheken sowie eine Webapplikation mit einer einfachen Benutzeroberfläche.

Repositorium	https://github.com/MarquezProject/marquez
Lizenz	Apache 2.0
Tech Stack	MySQL, Postgres, RedShift, Snowflake, Kafka
Quellen	MySQL, Postgres, RedShift, Snowflake, Kafka
Integration	REST API, Client Bibliotheken für Java und Python, Bibliothek zur AirFlow Integration
Features	Data-Lineage, Data Health, Data Owndership, Versionierung von Datensätzen, Tags

Der Dienst ist zur Nutzung mit einem Workflow Scheduler, wie etwa AirFlow, gedacht und kann auch zur direkten Aufnahme von Metadaten aus Spark oder Flink Jobs verwendet werden. Die wesentlichen Instanzen von Marquez sind Quellen, Datensätze und Jobs. Sein Metadaten Repositorium speichert Daten über alle Jobs und Datensätze, einschließlich einer kompletten Verlaufshistorie der Jobs und Job-Level Statistiken. Die Ausführung eines Jobs ist mit versioniertem Code verbunden und produziert ein oder mehrere versionierte Outputs. Datensätze und Jobs haben Besitzer, welche durch Namensräume definiert sind.

Der Kern von Marquez ist in Java geschrieben. Unter der Haube wird Postgres, Caylay und ElasticSearch verwendet. Momentan sind die Aufnahmequellen noch auf MySQL, Postgres, RedShift, Snowflake und Kafka beschränkt. Quellen werden in generischer Weise beschrieben, sodass es kein großer Aufwand sein sollte, neue Quellen hinzuzufügen. Es sei denn, sie passen nicht in das stark meinungsbehaftete Datenmodell von Marquez, welches eine Schwachstelle der Architektur darstellt. Mithilfe der REST API kann Marquez auch mit anderen Plattformen und Bibliotheken, wie etwa Amundsen und Dagster, integriert werden.

ALTERNATIVE LÖSUNGEN

Neben den Datenkatalogen gibt es auch andere Tools und Frameworks für das Management von Datensätzen, die es wert sind, berücksichtigt zu werden, so wie etwa CKAN und Frictionless Data.

CKAN wird hauptsächlich von Regierungen und Stiftungen eingesetzt. Als Tool zum Veröffentlichen von Daten kann es nicht mit Datenquellen wie SQL oder Hive Datenbanken integriert werden. Um CKAN mit diesen Datenquellen zu verwenden, müssen Nutzer ihre Daten zuerst in einem der unterstützten Datenformaten einordnen. Andererseits hat CKAN eine recht einfache und benutzerfreundliche UI. Es besteht die Möglichkeit, Erweiterungen hinzuzufügen, welche die Funktionalität erheblich ausweiten können. Zu den enthaltenen Features zählen Zugriffskontrolle, Datenvisualisierung, Datenvorschau, Datensammlung (einige der Features können die Installation zusätzlicher Erweiterungen voraussetzen). CKAN ist in Python geschrieben und nutzt Postgres und Solr um Metadaten dauerhauft zu erhalten. Das Tool unterstützt sowohl lokalen Speicher als auch Cloudspeicher wie etwa AWS S3 oder Azure Storage. CKAN-Quellen sind auf GitHub verfügbar.

Frictionless Data ist ein Framework zum Erstellen von Dateninfrastrukturen und zielt auf Datenmanagement, Integration und Workflows ab. Es bietet eine Vielfalt an Tools zum Datenmanagement, zum Beispiel Data Packages oder Good Tables. Das Framework wird von datahub.io verwendet, welche Datenveröffentlichungsdienste anbieten. Der Quellcode befindet sich auf GitHub.

FAZIT

Die Datenmengen werden von Jahr zu Jahr größer. Daher steigt die Problematik des Datenmanagements und der Datenermittlung. Viele der genannten Produkte wurden anfänglich für den internen Gebrauch in Unternehmen erstellt, und ihr Code später veröffentlicht. Dieser Trend ist nicht nur für Nutzer, sondern auch für die Produkte positiv: Eine Folge ist, dass die entsprechenden Plattformen vielseitiger werden. Bereits jetzt ist offensichtlich, dass die meisten größeren Unternehmen die hier beschriebenen Tools verwenden oder eigene Tools entwickeln. Das zeigt, dass sich dieses Gebiet in Zukunft weiterentwickeln wird und in einiger Zeit neue Tools zur effektiven Bewältigung aufkommender Problemstellungen zur Verfügung stehen werden.

QUELLEN

http://atlas.apache.org

https://developer.ibm.com/articles/apache-atlas-and-janusgraph-graph-based-meta-data-management

https://www.ibmbigdatahub.com/blog/insightout-role-apache-atlas-open-metadata-ecosystem

https://www.youtube.com/watch?v=OB-O0Y6OYDE

https://engineering.linkedin.com/blog/2019/data-hub

https://eng.lyft.com/open-sourcing-amundsen-a-data-discovery-and-metadata-platform-2282bb436234

https://medium.com/wbaa/facilitating-data-discovery-with-apache-atlas-and-amundsen-631baa287c8b

https://marquezproject.github.io/marquez/

https://www.youtube.com/watch?v=BIVUXruv5io

https://www.youtube.com/watch?v=dRaRKob-IRQ

https://frictionlessdata.io/

https://ckan.org/