Dezentrale und Microservice-orientierte Datenintegration am Beispiel externer Datenquellen

Christoph Schröer
Volkswagen AG, Corporate Foresight, Wolfsburg, Germany;Carl von Ossietzky Universität Oldenburg, Very Large Business Applications (VLBA), Oldenburg, Germany

Jonas Frischkorn
Volkswagen AG, Corporate Foresight, Wolfsburg, Germany

Dezentrale und Microservice-orientierte Datenintegration am Beispiel externer Datenquellen

Christoph Schröer[1,2], Jonas Frischkorn[1]

[1] Volkswagen AG, Corporate Foresight, Wolfsburg, Germany
{christoph.schroeer,jonas.frischkorn}@volkswagen.de
[2] Carl von Ossietzky Universität Oldenburg, Very Large Business Applications (VLBA), Oldenburg, Germany

Zusammenfassung. Data Lakes bieten gute Möglichkeiten, um heterogene Daten für analytische Fragestellungen in Unternehmen zentral zu nutzen. Allerdings gibt es auch Herausforderungen und Risiken wie fehlende Referenzarchitekturen bei der Zugänglichkeit oder Nutzbarkeit. Durch den Einsatz von modernen Architekturmustern wie Microservices können alternativ Daten technisch und organisatorisch dezentral verwaltet und in den fachlich passenden Organisationseinheiten verantwortet werden. Durch offen zugreifbare Schnittstellen und Microservice-Architekturmuster können wichtige Data Lake Charakteristika beibehalten werden, wie z. B. die Zugänglichkeit und die Bereitstellung von Metadaten. Dadurch können Kosten eingespart, Daten in den jeweiligen Domänen verantwortet, und gleichzeitig Schnittstellen für analytische Fragestellungen bereitgestellt werden. Der Beitrag zeigt die Idee in Form eines Work-In-Progress-Papers am Beispiel der Integration von externen Unternehmensdaten auf.

Keywords: Data Lake, Microservice, externe Datenquellen

1 Einleitung

Die Digitalisierung und die damit verbundene digitale Transformation eröffnen Zugänge zu neuen Informationen, können Prozesse neu organisieren, Kosten senken und Marktchancen entstehen lassen [1]. Im Kontext digitaler, strategisch relevanter Entscheidungsprozesse stehen unternehmensexterne, heterogene Datenquellen im Fokus, um einen ganzheitlichen Überblick über das Marktgeschehen, Innovationen und den Wettbewerb zu erzielen. Die Integration unterschiedlicher Datenquellen kann in zentralen, schemalosen Data Lakes resultieren [2, 3], die in Unternehmen zu monolithischen Datenarchitekturen anwachsen [4]. Um schlagkräftig und kontextabhängig Daten und Informationen zu sammeln, abzulegen und zweckbezogen zu verarbeiten, braucht es Ansätze zu einer effizienten Informationsverarbeitung.

Die Forschungsfrage, die im vorliegenden Positionspapier beantwortet werden soll, lautet: Wie können Microservices die Datenintegration hinsichtlich der beschriebenen Herausforderungen unterstützen?

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

Wir zeigen zunächst im zweiten Abschnitt die Idee, die Konzeption und den weiteren Forschungsbedarf auf, wie Unternehmen externe Datenquellen in dezentralisierte, domänenspezifische Data Lake-Microservices integrieren und damit Plattformlösungen effizienter gestalten können. Entscheidend ist dabei die Idee, einen zentralen Data Lake zu vermeiden und direkt auf den Fachdaten basierend analytische Fragestellungen in der Architektur zu berücksichtigen. Im dritten Abschnitt veranschaulichen wir neben positiven Effekten für Unternehmen, ebenfalls wissenschaftliche Implikationen von Microservice-Architekturen. Dieses Positionspaper endet mit einem Ausblick für die weitere Forschung.

2 Problemstellung und Herausforderungen

Sawadogo (2020) definiert Data Lakes als eine skalierbare Speicher- und Analyseplattform auf Basis von Rohdaten für Statistiker, Data Scientists oder Analysten. Wichtige Bestandteile sind neben Speichermöglichkeiten auch die Implementierung eines Metadatensystems, Datenintegrationskomponenten sowie Data Governance. Charakteristika sind die Zugänglichkeit, die logische und physische Organisation und die Skalierbarkeit bezüglich der Speicher- und Berechnungskapazitäten [5]. Ein Data Lake repräsentiert den Zustand der Daten zu jedem Zeitpunkt [5–8]. Der weitere, aktuelle State-of-the-Art von Data Lakes wird zum Beispiel in dem Beitrag von [1] weiter beschrieben.

Herausforderungen bei der Implementierung von Data Lakes sind zum Beispiel fehlende Referenzarchitekturen [4, 5]. Spezifische Charakteristika wie Metadatenmanagement und Data Governance sind nicht trivial erfüllbar. Ein Data Lake bildet nicht nur technisch einen zentralen Datenspeicher ab, sondern wird auch organisatorisch durch ein zentrales Datenteam verantwortet. Dadurch wird der Datenfluss über mehrere organisatorische Einheiten hinweg durchbrochen, wodurch Data Lineage erschwert wird [9]. Hadoop als Technologie wird häufig mit dem Begriff des Data Lakes zusammen verwendet [6]. Allerdings bestehen Data Lakes aus verschiedenen Speichertechnologien, die für strukturierte, semi-strukturierte und unstrukturierte Daten geeignet sind. Die Herausforderung ist hier, die Datenabfrage über heterogene Speichersysteme zu ermöglichen [5].

Sollen weiterhin externe Datenquellen für analytische Fragestellungen integriert werden, ergeben sich zu treffende Architekturentscheidungen. Aufgrund der externen Generierung können diese nicht direkt einer organisatorischen, operationellen Einheit zugeordnet werden. Es gibt verschiedene Möglichkeiten für die Festlegung des fachlichen Verantwortungsbereichs:

- bei der anfordernden Fachabteilung, die den Bedarf für eine externe Datenquelle erkennt,
- bei einem zentralen Datenteam, das externe Daten für strategisch relevante Projekte heranziehen möchte,
- bei einer Fachabteilung derjenigen Domäne, die auch der externen Datenquellen zugeordnet werden kann. Beispielsweise können Nachrichtendaten in den Verantwortungsbereich der Kommunikationsabteilung fallen.

Die Zuordnung von externen Daten zu internen Experten kann den Domänenbezug und das Verständnis härten. Gleichzeitig können diese auch von Analyseergebnissen profitieren. Die Integration der externen Daten innerhalb des Fachkontextes (s. 3. Abschnitt) stellt einen Erfolgsfaktor dar, den wir weiter untersuchen.

3 Microservice-orientierte Datenintegration

Dieser Beitrag zeigt eine alternative Herangehensweise zur technisch und organisatorisch zentralen Data Lake Architektur. Seit 2003 etabliert sich der Ansatz von Domain Driven Design (DDD) in der Softwareentwicklung. Dadurch wird Software entlang der Fachlichkeit der Domäne modelliert und entwickelt [10]. Daten sind stets einer Domäne und organisatorisch einer Fachabteilung zuzuordnen, in denen ein fachspezifisches Verständnis der Daten vorliegt. Ein zentrales Datenteam müsste dieses Domänenwissen aufbauen, um für die Bearbeitung analytischer Fragestellungen die richtigen Daten zu selektieren und zu bewerten. Um die Charakteristika eines Data Lakes bezüglich der Datenintegration zu erfüllen, braucht es nicht zwangsläufig eine zentrale Speicherplattform.

Microservices etablieren sich seit 2014 als Architekturansatz, um Applikationen als eine Menge von kohärenten, lose gekoppelten Services zu entwickeln [11]. Abbildung 1 gibt einen ersten Architekturvorschlag für eine Microservice-orientierte Datenintegration, der in den folgenden Absätzen näher beschrieben wird. Dabei werden technische Aspekte von Microservices mit funktionalen Aspekten von Data Lakes kombiniert. Wir nutzen technisch diese Form der Architektur prototypisch für die Integration und Verarbeitung externer Datenquellen aus den Nachrichten-, Unternehmens-, Politik- und Patentdomänen. Dieser Architekturvorschlag wurde mittels DDD-Ansätzen hergeleitet, indem die externen Datenquellen jeweils einen eigenen Bounded Context bilden.

Die Architektur sieht dedizierte Bounded Contexts und somit resp. **Microservices** für die Integration und Speicherung externer Daten je Datenquelle und Domäne vor. Abbildung 1 zeigt dazu beispielhaft eine *Datenquelle A aus Domäne 1*. Eine weitere Datenquelle könnte ebenfalls aus Domäne 1 stammen, sieht aber andere Konzepte und ggf. eine eigene Fachsprache vor. Beispielsweise kann die Entität Unternehmen durch die Begriffe Company oder Organization definiert werden.

Um Zustände über die Zeit nachvollziehen zu können, werden *Event-Datenbanken* genutzt, die Datenveränderungen in Form von Events abbilden. Dazu nutzen wir das Microservice-Architekturmuster des Event-Sourcings [12]. Neben dieser domänenspezifischen Bereitstellung der externen Daten, können ein oder mehrere *Microservices für Abfrage-optimierte Zugriffe* für analytische Fragestellungen implementiert und über Schnittstellen bereitgestellt werden. Hierfür erproben wir das Microservice-Architekturmuster Command Query Responsible Separation (CQRS) [12], um Daten über Event Sourcing zu aktualisieren.

Schnittstellen (API) der Microservices folgen einer einheitlichen Konvention. Die Schnittstellenbeschreibungen werden in dem Metadatensystem resp. *Datenquellenkatalog* abgelegt. In Tabelle 1 sind wichtige Charakteristika der Schnittstellenmethoden

dargestellt. Diese wurden aus internen Workshops und auf Basis der Integration vierer externer Datenquellen sowie der Bearbeitung einer praktischen Fallstudie abgeleitet.

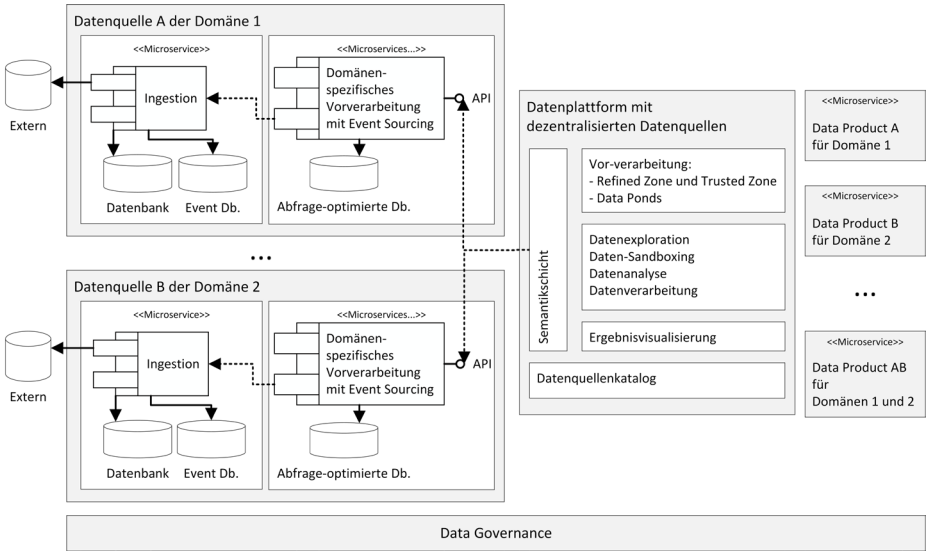

Abbildung 1. Dezentrale Datenbereitstellung auf Basis von Microservices für zentrale Datenteams zur Bearbeitung analytischer Fragestellungen.

Die in Abbildung 1 dargestellte *Datenplattform* implementiert unter anderem bekannte Data Lake Architekturmuster und Komponenten [1] zur Datenverarbeitung. Ein Metadatensystem und ein *Datenquellenkatalog* helfen, Informationen über Daten und Datenquellen abfragen zu können. Das Metadatensystem ist nicht neu in Bezug auf den Microservice-orientierten Ansatz, muss aber für diesen adaptiert werden. Über Schnittstellen können Metadaten der einzelnen Datenquellen abgefragt und in dem Metadatensystem oder *Datenquellenkatalog* hinterlegt werden [13]. Auch Schnittstellenbeschreibungen und -dokumentationen können das Metadatensystem ergänzen. Semantische Beschreibungen oder auch sogenannte Knowledge Graphen können in der *Semantikschicht* innerhalb der *Datenplattform* ergänzen. Knowledge Graphen können genutzt werden, um Microservice-übergreifende und somit auch datenübergreifende Abfragen zu tätigen, ohne genau zu wissen, welche Microservices technisch angesprochen werden müssen [14, 15].

Der Microservice-orientierte Ansatz grenzt sich von Data Marts dahingehend ab, dass dedizierte Microservices über Schnittstellen den Vollzugriff und somit nicht nur einen Zugriff in aggregierter Form auf den Datenbestand ermöglichen. Aggregierte Daten und Analyseergebnisse finden sich in Abbildung 1 in den *Data Products* wieder.

Data Governance kann durch die organisatorische Verantwortung der Microservices in den Fachabteilungen gestärkt werden. Diese kann prüfen, welche Daten beispielsweise im Rahmen der EU-Datenschutzgrundverordnung herausgegeben werden dürfen. Weiterhin können auch hier Datenqualitätsaspekte insbesondere bei externen

Datenquellen durch Domänenexperten bewertet werden. In Abbildung 1 ist dieser Aspekt dargestellt, indem *Data Governance* bereits bei der Datenintegration beginnt.

Tabelle 1. Charakteristika der Schnittstellenmethoden.

Charakteristik	Beschreibung
Accessability (Zugänglichkeit)	Durch eine API-Methode kann der Zugriff für jeden Nutzer angefragt werden.
Meta Data (Metadaten)	Über eine API-Methode sollen Metadaten bereitgestellt werden, die z.B. deskriptive Statistiken beinhalten.
Semantics (Semantik)	Durch eine API-Methode sollen semantische Konzepte, die zum Beispiel in Form eines Datenmodells repräsentiert werden, abgefragt werden können.
Data Retrieval (Datenabfrage)	Durch ein Datenabfrage sollen Rohdaten abgefragt werden können, die dann für analytische Fragestellungen genutzt werden können.
Scalability (Skalierbarkeit)	Nicht immer können alle Daten aufgrund der Menge über eine Schnittstelle bereitgestellt werden. Da Microservices technologie-offen sind, können abfrageoptimierte Technologien sowie direkter Datenbankzugriff und Hadoop genutzt werden.

4 Zusammenfassung und Ausblick

Zusammenfassend zeigt der Work-In-Progress-Beitrag die Idee auf, Daten mittels DDD-Prinzipien und dezentralen Microservices zu integrieren. Bestehende Microservice-Architekturmuster und semantische Konzepte der Datenintegration werden zu einem neuen, technisch und organisatorisch dezentralen Integrationsansatz verknüpft.

Weiterer Forschungsbedarf liegt in der Evaluation, inwieweit weitere Qualitätsmerkmale, zum Beispiel Performance, und rechtliche Rahmenbedingungen, etwa das Thema Datenschutz, berücksichtigt werden können. Zudem besteht weiterer Forschungsbedarf in den Fragen, ob die Charakteristika von Data Lakes beibehalten werden können und, ob der Domänenbezug tatsächlich die Beantwortung analytischer Fragestellungen unterstützten kann.

Eine aktuell bekannte Limitation ist der Microservice-Architektur selbst geschuldet, da sie initial eine erhöhte Komplexität mit sich bringt. Ist in Unternehmen sowohl organisatorisch als auch technisch die Microservice-Architektur weitestgehend etabliert, kann unser vorgestellter Ansatz die Datenintegration mittels Microservices und somit analytische Fragestellungen erweitern und unterstützten.

Disclaimer. Ergebnisse, Meinungen und Schlüsse dieser Veröffentlichung sind nicht notwendigerweise die der Volkswagen Aktiengesellschaft.

Referenzen

1. Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., Mitschang, B.: Data Lakes auf den Grund gegangen. Datenbank Spektrum (2020). https://doi.org/10.1007/s13222-020-00332-0
2. H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes, J. Riekki: Implementing big data lake for heterogeneous data sources. 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW) (2019). https://doi.org/10.1109/ICDEW.2019.00-3
3. Dixon, J.: Pentaho, Hadoop, and Data Lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (2010). Accessed 14 July 2020
4. Hai, R., Geisler, S., Quix, C.: Constance. An Intelligent Data Lake System. Proceedings of the 2016 International Conference on Management of Data (SIGMOD) (2016). https://doi.org/10.1145/2882903.2899389
5. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. J Intell Inf Syst (2020). https://doi.org/10.1007/s10844-020-00608-7
6. Gupta, S., Giri, V.: Practical Enterprise Data Lake Insights. Handle Data-Driven Challenges in an Enterprise Big Data Lake. Apress, Berkeley, CA (2018)
7. Kim, J., Ha, H., Chun, B., Yoon, S., Cha, K.: Collaborative analytics for data silos. International Conference on Data Engineering (ICDE) (2016). https://doi.org/10.1109/ICDE.2016.7498286
8. Terrizzano, I., Schwarz, P., Roth, M., Colino, J.E.: Data Wrangling: The Challenging Journey from the Wild to the Lake. 7th Biennial Conference on Innovative Data Systems Research (2015)
9. Janssen, M., Brous, P., Estevez, E., Barbosa, L.S., Janowski, T.: Data governance. Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly (2020). https://doi.org/10.1016/j.giq.2020.101493
10. Evans, E.: Domain-driven design. Tackling complexity in the heart of software. Addison-Wesley, Upper Saddle River, NJ (2011)
11. Lewis, J., Fowler, M.: Microservices. a definition of this new architectural term. https://martinfowler.com/articles/microservices.html (2014). Accessed 30 April 2018
12. Richardson, C.: Microservice Patterns. With examples in Java. Manning, Shelter Island, NY (2019)
13. Samourkasidis, A., Athanasiadis, I.N.: A semantic approach for timeseries data fusion. Computers and Electronics in Agriculture (2020). https://doi.org/10.1016/j.compag.2019.105171
14. Schmid, S., Henson, C., Tran, T.: Using Knowledge Graphs to Search an Enterprise Data Lake. In: Hitzler, P., Kirrane, S., Hartig, O., Boer, V. de, Vidal, M.-E., Maleshkova, M., Schlobach, S., Hammar, K., Lasierra, K., Stadtmüller, S., Hose, K., Verborg, R., Maleshkova, M., Lasierra, N., Verborgh, R. (eds.) The Semantic Web: ESWC 2019 Satellite Events. ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers. Lecture Notes in Computer Science book series, vol. 11762. Springer International Publishing (2019)
15. Galkin, M., Auer, S., Vidal, M.-E., Scerri, S.: Enterprise Knowledge Graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems (2017). https://doi.org/10.5220/0006325200880098