Aufbau einer Datenpipeline mit No-Code-Applikationen in der Cloud als SaaS-Anwendung

Steven Art Johnston (B.Sc.)

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Datenverarbeitung
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
steven.art.johnston@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Informatik
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Armin Wagenknecht

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Informatik
Wiesenstraße 14
35390 Gießen
E-Mail:
armin.wagenknecht@mni.thm.de

Kategorie

Abschlussarbeit

Schlüsselwörter

Datenpipeline, No-Code Development, Datenreplikation, Datenintegration, Change Data Capture

Zusammenfassung

Ein wichtiger Bestandteil der modernen Dateninfrastruktur ist die Vielfalt der Datenquellen, von denen Unternehmen über Hunderte verfügen. Wie und in welcher Form die Daten aus den Quellsystemen zur Verfügung gestellt werden, ist von zentraler Bedeutung, wobei jede Schnittstelle ihre eigenen Herausforderungen mit sich bringt. Bei der Modernisierung der Dateninfrastruktur sehen sich Unternehmen häufig mit unflexiblen Prozessen und Architekturen konfrontiert. Die Extraktion von Daten aus verschiedenen Quellen zur Unterstützung der Entscheidungsfindung stellt in diesem Zusammenhang nach wie vor eine Herausforderung dar. Daten müssen bereinigt, verarbeitet und kombiniert werden, um Informationen, Wissen und Erkenntnisse zu generieren. Der Aufbau effizienter Datenpipelines ist notwendig, damit die Daten die verschiedenen Stufen der Wertschöpfungskette durchlaufen können. Der zeitaufwändige und komplexe Aufbau ist nicht nur kostenintensiv, sondern erfordert auch den Einsatz qualifizierter Entwickler. Der Mangel an diesen Spezialisten behindert zunehmend die schnelle Umsetzung von Datenpipeline-Projekten und erfordert einen Paradigmenwechsel.

In dieser Arbeit wird mit der Technologie des No-Code Developments eine Lösung angeboten, die den Prozess der Entwicklung von Datenpipelines automatisiert. Das Ziel dieser Masterarbeit ist der Aufbau einer Datenpipeline als SaaS-Anwendung unter Verwendung von No-Code-Applikationen, die von der Qlik-Cloud-Plattform bereitgestellt werden. Dabei werden die Unterschiede in der Nutzung der kollaborativen Datenexploration sowie die Unterschiede in der Performance im Vergleich zu einem vorangegangenen Projekt mit einer clientbasierten Datenpipeline-Architektur untersucht. Um dieses Ziel zu erreichen, werden die Nutzungsunterschiede bei der Konzeption der Datenpipelines analysiert. Darüber hinaus werden Qlik Cloud Data Integration für die Replikation der Daten und den Aufbau der Datenpipeline sowie Qlik Cloud Analytics für die Visualisierung der Daten im Rahmen eines End-to-End-Prozesses eingesetzt. Die Echtzeitübertragung wird durch die Change-Data-Capture(CDC)-Technologie realisiert. CDC ist ein Verfahren zur Identifizierung von Änderungen (Deltas) in einem Datensatz.

Transaktionale Datenbanken speichern Datenänderungen für Wiederherstellungsprozesse im Transaction Log (TLOG). Durch Auslesen des TLOG können die Änderungen abgefragt und verarbeitet werden. Eine Herausforderung liegt im Abrufen der TLOGs, da Datenbankhersteller nicht notwendigerweise eine Schnittstelle zu diesem Protokoll anbieten. Ein Lösungsansatz in dieser Arbeit ist die Aktivierung von Änderungstabellen auf der Datenbank, die die Datenänderungen in Metadatentabellen speichern.

Für den Performancevergleich werden die Datenübertragungsgeschwindigkeiten von zwei Extraktionsmechanismen, der Full-Load-Replikation und der Delta-Load-Replikation untersucht. Für den Performancevergleich der Full-Load-Replikation werden die Logfiles der No-Code-Applikationen ausgewertet. Der Performancevergleich der Delta-Load-Replikation (CDC-Replikation) wird durch das CDC-Verfahren durchgeführt. Hierbei werden die Zeitpunkte der Datenübertragung anhand von Zeitstempeln aus den durch den CDC-Prozess erzeugten Änderungstabellen vom Quellsystem über das Zielsystem bis hin zu den

Data-Warehouse- und Data-Mart-Tabellen nachvollzogen und ausgewertet.

Um die Hintergrundprozesse des No-Code-Developments zu verstehen, werden die Verfahren des Schema Matching und des Schema Mapping behandelt. Basis beim Schema Mapping sind Metadaten, insbesondere die Metadaten der lokalen Schemata. In diesem Projekt erfolgt eine Zuordnung von Attributwertpaaren, die zur Darstellung der semantischen Verbindung zwischen dem Quell- und dem Zielschema verwendet werden. Die No-Code-Technologie automasiert das Mapping welches eine Menge von Attributwertpaaren darstellt, die alle gemeinsamen Attribute des Quellschemas mit Attributen des Zielschemas verbindet.

Die Ergebnisse zeigen, dass die Technologie des No-Code Developments einen ganzheitlichen End-to-End-Prozess zum Aufbau einer Cloud-Datenpipeline ohne herkömmliche Programmierung ermöglicht. Dieses Ziel wird durch den Einsatz der iPaaS-Lösung Qlik Cloud Data Integration erreicht. Die integrierte CDC-Technologie ermöglicht zudem eine Echtzeit-Datenanalyse. Die Schnittstellen- und Datenübertragungsprozesse werden nach der Einrichtung der notwendigen Systemvoraussetzungen automatisiert, während die Erstellung des Datenmodells manuelle Eingriffe über Drag-and-Drop-Eigenschaften und andere No-Code-Funktionen erfordert. Die Untersuchung kann weiterhin eine Lösung für das Problem des Fachkräftemangels aufzeigen, wobei die (Teil-) Automatisierung Ressourcen einspart und somit die Anzahl der am Entwicklungsprozess beteiligten Mitglieder reduziert.

Die Forschung zeigt weiterhin die kollaborative Datenexploration in zwei unterschiedlichen Bereitstellungsvarianten einer Datenpipeline. Die Vorteile des Cloud Computings ermöglichen ein kontrolliertes Vorgehen im Rahmen der kollaborativen Zusammenarbeit. Durch die Zuweisung von Rollen innerhalb des Cloud-Mandanten werden automatisch Benutzerrechte vergeben und definierte Grenzen zwischen den Benutzern etabliert. Zudem bietet die eingesetzte Cloud-Plattform einen dezentralen Zugriff sowie eine hohe Benutzerfreundlichkeit. Die Zusammenarbeit wird für jeden Block der Datenpipeline in einem vordefinierten gemeinsamen Bereich mit Möglichkeiten zur Versionskontrolle unterstützt. Änderungen an der SaaS-Anwendung werden für jede Rolle automatisch übernommen.

Die vorliegenden Ergebnisse des Performancevergleichs zeigen, dass eine signifikante Überlegenheit der Cloud-Datenpipeline hinsichtlich der Übertragungsgeschwindigkeit im Kontext der durchgeführten Full-Load-Replikation besteht. Die CDC-Replikation hat gezeigt, dass die Übertragungsgeschwindigkeit zwischen den Projekten sehr ähnlich ist. Es wurde weiterhin gezeigt, dass die Daten zwischen den Blöcken vom Quellsystem zum Data Warehouse innerhalb der clientbasierten Datenpipeline mit höherer Geschwindigkeit übertragen werden. Bei näherer Betrachtung der gesamten Blöcke wurde ermittelt, dass die Übertragungsgeschwindigkeit der cloudbasierten Datenpipeline schneller ist. Es ist festzuhalten, dass eine ereignisgesteuerte Datenübertragung, wie sie im cloudbasierten Projekt eingesetzt wird, eine effizientere Datenübertragung darstellt als eine zeitgesteuerte Intervallübertragung.

Literatur

Amghar, Souad; Cherdal, Safae; Mouline, Salma (2019) Data Integration and NoSQL Systems: A State of the Art, in: The 4th International Conference On Big Data and Internet of Things, S. 1-6, [Online]. DOI: https://doi.org/10.1145/3372938.3372954

Chandra, Harry (2020) Experimental results on change data capture methods implementation in different data structures to support real-time data warehouse, in: Int. J. Business Information Systems, Vol. 34, Nr. 3, S. 373–402 [Online]. DOI: https://doi.org/10.1504/IJBIS.2020.108651

Densmore, James (2021) Data Pipelines Pocket Reference (- Moving and Processing Data for Analytics), 1. Auflage, O'Reilly Media, Sebastopol.
Leser, Ulf; Naumann, Felix (2007) Informationsintegration (-Architekturen und Methoden zur Intergration verteilter und heterogener Datenquellen), 1. Auflage, dpunkt, Heidelberg.

Ma, Chuangtao; Molnár, Balint; Tarcsi, Ádám; Benczúr, András (2020) Knowledge Enriched Schema Matching Framework for Heterogeneous Data Integration, in: IEEE 2nd Conference on Information Technology and Data Science, S. 183-188, [Online]. DOI: https://doi.org/10.1109/CITDS54976.2022.9914350

Moskal, Monika (2021) NO-CODE APPLICATION DEVELOPMENT ON THE EXAMPLE OF LOGOTEC APP STUDIO PLATFORM, in: Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska, Vol. 11, Nr. 1, S. 54–57, [Online]. DOI: https://doi.org/10.35784/iapgos.2429

Munappy, Aiswarya; Bosch, Jan; Holmström Olsson, Helena (2020) Data Pipeline Management in Practice (- Challenges and Opportunities), in: Lecture Notes in Computer Science, S. 168-184, [Online]. DOI: https://link.springer.com/ chapter/10.1007/978-3-030-64148-1_11