Bewertung von Werkzeugen zur Datenintegration in die Cloud am Beispiel von HVR, dbt und Snowflake

Lenard Damm

Technische Hochschule Mittelhessen

Fachbereich Mathematik, Naturwissenschaften und Datenverarbeitung
Wilhelm-Leuschner-Str. 13
61169 Friedberg
lenard.damm@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich Mathematik, Naturwissenschaften und Informatik
Wiesenstr. 14
35390 Gießen
harald.ritz@mni.thm.de

Azmat Ahmad (M.Sc.)

INFOMOTION GmbH

BU Sayar
Westhafenplatz 1
60327 Frankfurt
azmat.ahmad@infomotion.de

Kategorie

Masterarbeit

Schlüsselwörter

Datenintegration, ETL, ELT, Business Intelligence, Cloud Computing, Data Warehouse, Cloud Data Warehouse, HVR, dbt, Snowflake

Zusammenfassung

Unternehmen müssen mit den großen Datenmengen, die sie sammeln, umgehen können. Häufig werden hierzu die Daten aus den verschiedenen Datenquellen in ein Data Warehouse (DWH) integriert. Dieses bildet einen zentralen Ort für verlässliche Daten (Single Point of Truth), die der Entscheidungsunterstützung dienen.

Zunehmend setzen Unternehmen Cloud Data Warehouses (CDW) ein. Diese bieten viele Vorteile der Cloud, wie eine einfachere Zugänglichkeit und Bereitstellung, eine höhere Skalierbarkeit und Elastizität, eine höhere Rechenleistung, eine verbesserte Integration sowie Notfallwiederherstellung.

Im Rahmen der Umstellung von einem DWH zu einem CDW muss ein passendes Datenintegrationswerkzeug ausgewählt werden. Es stellt sich somit die Frage, welche Anforderungen ein Werkzeug für die Datenintegration in die Cloud erfüllen muss.

Zwei solcher Datenintegrationswerkzeuge sind HVR und dbt Cloud. HVR bietet die Möglichkeit Daten zwischen Quell- und Zielsystemen, zu denen auch CDWs zählen, zu replizieren und deckt somit die Schritte der Extraktion und des Ladens ab. Dbt Cloud dient hingegen der Transformation und führt diese direkt auf dem zugrundeliegenden CDW aus. Gemeinsam können diese Werkzeuge genutzt werden, um ELT-Prozesse umzusetzen. Snowflake ist ein CDW das sowohl von HVR als auch von dbt Cloud unterstützt wird.

Um festzustellen, inwiefern sich HVR und dbt Cloud als Werkzeuge für die Datenintegration eignen, wurden die Anforderungen an solche Werkzeuge aus verschiedenen Quellen gesammelt. Anschließend wurden Experteninterviews durchgeführt, um die Relevanz, Gewichtung und Kritikalität der gesammelten Kriterien zu bestimmen und weitere Kriterien ergänzen zu können. Daraus wurde anschließend ein Kriterienkatalog erstellt. In diesem wurden die Kriterien gruppiert in Kriterien für die Extraktion und das Laden und Kriterien für die Transformation. Die enthaltenen Kriterien sind: Konnektivität, Datentransformationen, Produktivität und Zusammenarbeit, Logging, Jobsteuerung, Data Governance, Unterstützung aktiver Metadaten, Formen der Datenintegration, Interoperabilität, Performanz, Skalierbarkeit, Einfache Nutzung und Self-Service, Bereitstellung bei verschiedenen Cloud-Anbietern, Integration Portability, Open-Source, Support und Dokumentation und Kosten und Lizenzmodell.

Weiterhin wurde die Umsetzung eines Datenmodells auf Snowflake, unter Verwendung der beiden Datenintegrationswerkzeuge, beschrieben, um einen Einblick in diese zu bieten.

Anhand des Kriterienkatalogs erfolgte dann die Bewertung von HVR und dbt Cloud. HVR wurde anhand der Kriterien für die Extraktion und das Laden bewertet, dbt Cloud anhand der Kriterien für die Transformation. Insgesamt konnte HVR ein Ergebnis von 87% erzielen. Verbesserungen wären vor allem im Bereich der Produktivität und Zusammenarbeit, bspw. durch eine Git-Integration, im Bereich der Unterstützung aktiver Metadaten und durch eine Bereitstellung in der Cloud möglich. Dbt Cloud erzielte 90%. Dieses bietet keine Low-Code- oder No-Code-Funktionen für eine einfachere Nutzung und Self-Service. Außerdem kann es nicht bei verschiedenen Cloud-Anbietern bereitgestellt werden. Weder HVR noch dbt Cloud sind Open-Source-Produkte. Weiterhin eignen sie sich nur für die Durchführung von ELT-Prozessen, nicht für ETL.

Durch die Bewertung der Werkzeuge wurde ersichtlich, für welche Anwendungszwecke sich diese jeweils eignen. Dadurch können die Ergebnisse bei der Auswahl einer geeigneten Lösung unterstützen.

Literatur

Dageville, Benoit; Cruanes, Thierry; Zukowski, Marcin; Antonov, Vadim; Avanes, Artin; Bock, Jon et al.: The snowflake elastic warehouse. In: Proceedings of the 2016 International Conference on Management of Data, 2016, S. 215–226.

dbt Dokumentation: What is dbt?, o.O., o.J., online im Internet: https://docs.getdbt.com/docs/introduction [Stand 20.10.2022].

Finger, Ralf (Hg.): BI & Analytics in der Cloud - Architektur, Vorgehen und Praxis, dpunkt.verlag GmbH, Heidelberg, 2018.

HVR 6 Dokumentation: Getting Started, o.O., o.J., online im Internet: https://fivetran.com/docs/hvr6/getting-started [Stand 17.10.2022].

Franzke, Georg: Hybride Datenarchitekturen - als Grundlage für ein modernes Data Warehouse, SIGS DATACOM GmbH, Troisdorf, 2019.

Halper, Fern: TDWI Cloud Data Warehouse Readiness Guide. In: TDWI Research, 2020.

Halper, Fern: Five Must-Have Data Integration Capabilites for Your Cloud Data Warehouse. In: TDWI Insight Accelerator, 2022.

Russom, Philip: The Cloud Data Integration Primer - Get Started Integrating Your Data in the Cloud. In: TDWI Checklist Report, 2018.

Thanaraj, Robert; Zaidi, Ehtisham; Menon, Sharat; Thoo, Eric; Showell, Nina: Critical Capabilities for Data Integration Tools, Stamford, 2021.