Konzeption und Entwicklung einer Datenpipeline zur automatisierten Validierung und Verarbeitung von Bankdaten am Beispiel der Mittelstand.ai

Stani Lennart Schlegel

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstr. 14
35390 Gießen
stani.lennart.schlegel@mni.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstr. 14
35390 Gießen
harald.ritz@mni.thm.de

Dr. Michel Becker

Mittelstand.ai GmbH & Co. KG

Data Science
Schiffenberger Weg 110
35394 Gießen
michel.becker@mittelstand.ai

Kategorie

Bachelorarbeit

Schlüsselwörter

Data Engineering, Data Warehousing, Data Science, Apache Spark, Data Governance

Zusammenfassung

Mit dem Voranschreiten der Digitalisierung im Bankensektor und der stetig wachsenden Menge an Daten, die durch Banken erfasst und gespeichert werden können, ergeben sich viele verschiedene Möglichkeiten der Verwertung dieser unterschiedlichen Arten von Daten.

Ein Beispiel der Verarbeitung dieser Daten ist die Vertriebssteuerung anhand von Data-Science-Analysen und mittels Machine-Learning-Modellen. Die Aussagekraft und Qualität der Analysen und Machine-Learning-Modellen hängt stark von der Qualität der Daten ab, die als Grundlage für die Analysen und Modelle dienen.

Eine Herausforderung in diesem Prozess der Datenverarbeitung besteht darin, die Datenqualität automatisiert sicherzustellen. Stammen Daten aus externen Quellsystemen, so sind diese Systeme nicht direkt kontrollierbar und können durch verschiedene Arten von Fehlern die Qualität der bereitgestellten Daten negativ beeinflussen.

Diese Thesis beschäftigt sich damit, eine Datenpipeline anhand von fest definierten Datenqualitätsrichtlinien zu konzipieren und zu implementieren. Mit der Implementierung soll eine automatisierte Validierung der Qualität und Konsistenz von Bankdaten anhand von einem konkreten Anwendungsfall umgesetzt werden.

Die Datenpipeline basiert auf dem Apache Spark Framework und wird auf einem Dataproc Cluster in der Google Cloud ausgeführt. Das Cluster wird nicht dauerhaft betrieben, sondern dynamisch gestartet, sobald neue Daten zur Verarbeitung bereitstehen. Die Speicherung der validierten Daten findet in einem BigQuery Data Warehouse statt.

Diese Arbeit geht auf einige theoretische Grundlagen hinter den Konzepten von Data Warehousing und ETL-Prozessen, sowie den Systemen Apache Spark und BigQuery, ein. Es wird anhand eines Prototyps einer cloudbasierten Datenpipeline gezeigt, wie diese theoretischen Aspekte in der Praxis umgesetzt werden können.

In der Arbeit wird untersucht, ob durch den Einsatz der prototypischen Implementierung die Datenqualitätsanforderungen der Eindeutigkeit, Vollständigkeit, Konsistenz und Gültigkeit in einer sequenziellen bzw. parallelen Verarbeitung der Daten erfüllt werden können. Außerdem wird das Erfüllen von weiteren nicht funktionalen Anforderungen überprüft.

Die Evaluation stellt heraus, dass sämtliche Datenqualitätsanforderungen durch die Implementierung der Datenpipeline bei einer sequenziellen Abarbeitung der extrahierten Daten sichergestellt werden können. Im Falle der parallelen Verarbeitung wird ein Szenario gefunden, in dem die Implementierung die funktionalen Anforderungen nur teilweise erfüllen kann.

Es stellt sich weiterhin heraus, dass dynamisch erstellte Dataproc Cluster nicht für zeitkritische ETL-Prozesse geeignet sind, da die Prozesse des dynamischen Startens und Beendens des Clusters per Dataproc API lange Wartezeiten nach sich ziehen. Daher eignen sich für diese Art von Anforderung dauerhaft ausgeführte Dataproc Cluster.

Literatur

Cai, Li; Zhu, Yangyong: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Data Science Journal, 2015

Gluchowski, Peter: Data Governance: Grundlagen, Konzepte und Anwendungen, Heidelberg: dpunkt.verlag, 2020

Provost, Foster; Fawcett, Tom: Data Science for Business, Sebastopol: O`Reilly Media Verlag, 2013

Salloum, Salman; Dautov, Ruslan; Chen, Xiaojun; Xiaogang Peng, Patrick; Zhexue Huang, Joshua: Big data analytics on Apache Spark, 2016, URL: https://link.springer.com/content/pdf/10.1007/s41060-016-0027-9.pdf (besucht am 05.04.2022)