Automatisierung des Aufbaus flexibler Infrastrukturen analytischer Systeme in Cloud-Umgebungen

Viktor Schneider

Technische Hochschule Mittelhesen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
viktor.schneider@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhesen

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

Sebastian Becker

INFOMOTION GmbH

BU Duheric
Westhafenplatz 1
60327 Frankfurt am Main
sebastian.becker@infomotion.de

Kategorie

Masterarbeit

Schlüsselwörter

Amazon Web Services (AWS), Analytische Systeme, Big Data, Cloud Computing, Data Lake, Data Warehouse (DWH), Infrastructure-as-Code (IaC), Konfigurationsmanagement, Microsoft Azure, Provisionierung.

Zusammenfassung

Die in den Unternehmen vorkommenden Anwendungsfälle von Datenauswertungen weisen eine große Vielfalt auf. Als Konsequenz wachsen kontinuierlich die Anzahl und die Funktionsvariabilität von Softwaresystemen, die diese Aufgaben unterstützen. Der gemeinsame Einsatz dieser Softwareteile führt zu komplexen und verschiedenartigen Architekturen, deren Realisierung einen immensen Aufwand erfordert. Außer klassischen Data-Warehouse-Systemen (DWH) finden sich seit einiger Zeit auch Big-Data-Anwendungen in der analysebezogenen IT-Landschaft von Unternehmen. Die Vielfalt und Komplexität dieser Architekturen behindert, verlangsamt und verteuert die Erprobung neuer und innovativer Zusammensetzungen aus analystischen Systemen im Rahmen von Proof-of-Concept-Studien.

Cloud-Computing-Dienste vereinfachen die Bereitstellung solcher Systemverbünde durch die Automatisierung der Ressourcenbereitstellung und -betriebs. Der enorme Zeitaufwand, die Fehleranfälligkeit, die evtl. zu einem mühsamen Neuaufbau des ganzen Systemverbunds führen kann, sowie die eingeschränkte Wiederverwendbarkeit eines manuell durchgeführten Aufbaus lassen sich dadurch aber nicht vermeiden. Eine Alternative zur manuellen Vorgehensweise bietet der Infrastructure-as-Code-Ansatz (IaC). Die Bereitstellung von Softwaresystemen und der darunter liegenden Infrastruktur lässt sich damit als Code einer universellen Programmiersprache oder einer domänenspezifischen Sprache definieren. Dies ermöglicht die Verwendung von Methoden aus dem Software Engineering, wie die Versionsverwaltung von Code, den Einsatz wiederverwend- und parametrisierbarer Komponenten, die Definition und Durchführung von Tests u.v.a.m.

Im Rahmen dieser Arbeit ist es trotz des breiten Spektrums an Architekturvarianten analytischer Systeme gelungen, mit Hilfe des IaC-Ansatzes ein gesamtheitliches Konzept für die Aufbauautomatisierung in mehreren Cloud-Umgebungen zu entwicklen und seine Realisierbarkeit anhand eines Prototyps nachzuweisen. Sowohl die klassischen DWH als auch die unterschiedlichen Arten von Big-Data-Anwendungen lassen sich zu einer sie umfassenden mehrschichtigen Referenzarchitektur vereinen. Die Hauptbestandteile dieser Architektur sind ein Data Lake für die Rohdaten, SQL- und NoSQL-DBMS, ETL-Werkzeuge, Big-Data-Processing-Engines sowie Datenanalyse- und Visualisierungswerkzeuge.

Alle Systeme aus dieser Referenzarchitektur können auf virtuellen Maschinen eigenadministriert, in Form einer SaaS-Lösung oder eines PaaS-Dienstes betrieben werden. Obwohl sich die Cloud-Umgebungen in ihrem Angebot von Diensten unterscheiden, ist ihre Architektur ähnlich gestaltet. Die in dieser Arbeit analysierten Cloud-Umgebungen Amazon Web Services (AWS) und Microsoft Azure unterscheiden sich in Bezug auf IaaS-Dienste nur marginal. Aber auch im Fall von Big-Data-Diensten greifen die beiden Cloudanbieter auf dieselben Processing Engines und Machine-Learning-Frameworks zurück. Insgesamt entstand der Eindruck, dass beide Cloudanbieter den Fokus mehr auf Big-Data-Technologien als auf komplette DWH-Lösungen legen.

Jedes Softwaresystem kann – unabhängig vom Clouddienst, auf dem es läuft – von diesem Automatisierungskonzept erfasst werden, solange es die Kriterien einer Ressource erfüllt. Eine Ressource hat zugreifbare Eigenschaften, einen überwachbaren Zustand und kann erzeugt, modifiziert und zerstört werden. Ressourcen lassen sich zu wiederverwendbaren Modulen zusammenstellen. Durch die Parametrisierung von Modulen können die Softwaresysteme schnell und flexibel zu vielfältigen Systemverbünden kombiniert werden. Für die Erfassung

von neuen Systemen in die Automatisierungslösung wurde außerdem ein Prozessmodell erstellt.

Die Implementierung des Konzepts erfolgt unter Einsatz von speziellen IaC-Werkzeugen. Sie verwenden die Programmierschnittstellen von Clouddiensten oder den Softwaresystemen selbst, um die Systeme in der Cloud-Umgebung zu provisionieren und zu konfigurieren. Im Rahmen einer Vorauswahl wurden aus im Internet aufgefundenen IaC-Werkzeugen vier Instrumente mit dem nötigen Funktionsumfang bestimmt. Für die Endauswahl fand eine Evaluierung anhand von in der Literatur ermittelten und anforderungsgerechten Kriterien statt.

Für den Prototyp wurde ein DWH-System bestehend aus einem Data Lake, einem RDBMS, einem ETL-Werkzeug sowie einer Analyse- und Visualisierungsanwendung entwickelt. Die drei Clouddienstarten IaaS, PaaS und SaaS sind jeweils von mindestens einem DWH-Systemteil repräsentiert. Für jedes Systemteil wurden Ressourcen und Module sowohl für die AWS- als auch die Azure-Cloud erstellt, erfolgreich getestet und damit die cloudübergreifende Eigenschaft des Konzepts nachgewiesen. Weiterhin wurde in der AWS-Cloud eine RDBMS-Ressource auf einer virtuelle Maschine durch eine AWS-PaaS-Ressource ersetzt, um die flexible Anwendung der Lösung zu zeigen.

Literatur

Baars, Henning; Kemper, Hans-Georg: Business Intelligence & Analytics – Grundlagen und praktischen Anwendungen. 4. überarb. und erw. Aufl., Wiesbaden: Springer Vieweg Verl., 2021.

Bauer, Andreas; Günzel, Holger (Hrsg.): Data-Warehouse-Systeme – Architektur Entwicklung Anwendung. 4. überarb. und erw. Aufl., Heidelberg: dpunkt.verlag, 2013.

Brikman, Yevgeniy: Terraform: Up & Running. 2. Aufl., Sebastopol: O'Reilly Media, Inc., 2019.

Erl, Thomas; Mahmood, Zaigham; Puttini, Ricardo; Wise-Martinez, Pamela J.: Cloud computing – Concepts, technology & architecture, Upper Saddle River usw.: Prentice-Hall/Pearson, 2013.

Freiknecht, Jonas; Papp, Stefan: Big Data in der Praxis – Lösungen mit Hadoop, Spark, HBase und Hive – Daten speichern, aufbereiten, visualisieren. 2. erw. Aufl., München: Carl Hanser Verl., 2018.

Gorelik, Alex: The Enterprise Big Data Lake – Delivering the Promise of Big Data and Data Science, Sebastopol usw.: O'Reilly Media, Inc. Verl., 2019.

Heap, Michael: Ansible - From Beginner to Pro, New York: Apress Verlag, 2016.

Morris, Kief: Infrastructure as code – Managing servers in the cloud, Sebastopol: O'Reilly Media, Inc., 2016.

o.V.: Übersicht über Amazon Web Services – Whitepaper zu AWS, o.O., 2021, online im Internet: URL: https://docs.aws.amazon.com/de_de/whitepapers/latest/aws-overview/aws-overview.pdf. (Abruf: 08.03.22)

o.V.: Übersicht über die Dienste der Azure-Plattform, o.O., 2022, online im Internet: URL: https://docs.microsoft.com/de-de/azure/?product=all. (Abruf: 08.03.22)

Savill, John: Mastering Microsoft Azure Infrastructure Services, Indianapolis: Syber Verl., 2015.

Wittig, Michael; Wittig, Andreas: Amazon Web Services in Action, New York: Manning Publications Co. LLC, 2018.