Ein Vorgehensmodell zur Einführung eines Data Lakes

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftsingenieur der Fakultät für Elektrotechnik und Informatik, Fakultät für Maschinenbau und der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Sievering Vorname: Leon

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, den 16.09.2019

Inhaltsverzeichnis

1. Einleitung

Volumina und Diversität von Daten steigen stetig an. Auch die Geschwindigkeit, mit der die Daten erzeugt und übermittelt werden. Die damit einhergehenden Chancen und Herausforderungen werden unter dem Begriff Big Data zusammengefasst. Laut einer Studie (Reinsel, Gantz und Rydning, 2018, S.3) steigt die Menge weltweit generierter Daten von 33 ZB[1] in 2018 auf 175 ZB im Jahr 2025. Das Sammeln und Auswerten von Daten aller Art möglichst in Echtzeit übersteigt die Kapazitäten traditioneller Informationstechnologien. Die Fähigkeiten von Data Warehouses als zentrale Datenbasis kommen bei den hohen Datenvolumina an ihre Grenzen. Zudem können bei dem Einsatz einer kommerziellen Datenbanksoftware hohe Lizenzkosten anfallen.

Das neuartige Konzept des Data Lakes bietet eine skalierbare Plattform, um große Datenmengen bei geringen Kosten zu speichern und zu verarbeiten. Open-Source Tools in der Cloud mit einem Pay-per-Use Konzept unterstützen die Kosteneffizienz und Skalierbarkeit. Vor allem in der Fertigungsindustrie werden im Zuge der digitalen Transformation riesige Datenmengen generiert und analysiert. Industrie 4.0 und die neuen Cloud-Services wie von Microsoft, Google oder Amazon sind die Paten dieser Entwicklung. Data Lakes speichern Rohdaten unabhängig von ihrer Struktur und bieten auf diese Weise einen flexiblen Rahmen für Advanced Analytics und Big Data.

Das Laden von Rohdaten, ohne diese zu bereinigen oder zu transformieren, schafft Flexibilität und Agilität, jedoch sind auch neue Herausforderungen damit verbunden. Daten zur Entscheidungsunterstützung zu nutzen, setzt voraus, dass die Daten vertrauenswürdig und von geforderter Qualität sind. Zudem ist es einerseits das Ziel, die Vielzahl an Daten im Unternehmen für umfassende Analysen bereitzustellen, aber andererseits die Datenschutz- und Datensicherheits-Richtlinien nicht zu verletzen.

Die vorliegende Arbeit beschäftigt sich mit der Identifikation der neuen Herausforderungen, die das Data Lake Konzept mit sich bringt. Zudem werden Anforderungen und Aktivitäten zur Überwindung dieser Herausforderungen definiert. Ein generisches Vorgehensmodell zur Einführung eines Data Lakes fasst die Anforderungen und Aktivitäten zusammen, um den interessierten Unternehmen einen Leitfaden zur Unterstützung bei der Implementierung zu bieten.

Für die Entwicklung des Vorgehensmodells wird der Design Science Research Ansatz nach Hevner et al. (2004) und Peffers et al. (2007) verwendet. Als

---

[1] ZB = Zettabyte = 10^{12} Gigabyte

Informationsgrundlage dienen eine Literaturanalyse nach Webster und Watson (2002) und drei Experteninterviews.

Zunächst werden in Kapitel 2 die Begriffe Metadaten, Big Data, Data Warehouse und Data Lake definiert. Kapitel 3 beschreibt die verwendete Forschungsmethodik. Anschließend werden die Herausforderungen bei der Einführung eines Data Lakes anhand der Literaturanalyse und der Experteninterviews identifiziert (Kapitel 4). Darauf aufbauend werden in Kapitel 5 Aktivitäten und Anforderungen abgeleitet und in einem Vorgehensmodell zusammengefasst, mit dem Ziel Unternehmen bei der erfolgreichen Einführung eines Data Lakes zu unterstützen. Die anschließende Diskussion (Kapitel 6) befasst sich mit der Umsetzung und Realisierung der entwickelten Vorgehensweise. Darauf folgen die Limitationen und Implikationen. Ein Fazit und ein Ausblick schließen die Arbeit ab.

2. Fazit und Ausblick

In der vorliegenden Arbeit wurde ein allgemeingültiges Vorgehensmodell zur Einführung eines Data Lakes entwickelt. Die Phasen beschreiben Anforderungen und Aktivitäten für eine zielgerichtete Implementierung. Zudem wurden die Ideen und ihre Realisierungsmöglichkeiten diskutiert. Die Ergebnisse basieren dabei auf einer Literaturanalyse und Experteninterviews.

Rasant wachsende Datenmengen und wachsende Ansprüche der End-User nach Analysen und Auswertungen dieser neuen Daten erhöhen den Kostendruck auf die ITAbteilungen. Teuer sind nicht nur die Speicherkapazitäten, sondern auch die SoftwareLizenzen. Das quantitative Problem findet seinen Ausweg in der Cloud mit neuen Bezahlsystemen. Daneben ist ein qualitatives Problem zu lösen, nämlich der Umgang mit unstrukturierten Datenformaten, die gleichwohl gespeichert und ausgewertet werden sollen. Die Antwort auf diese Herausforderung ist der Data Lake mit neuen Mechanismen zur Beherrschung der Komplexität.

Vor allem die Gefahr, dass durch unkontrolliertes Laden großer Mengen heterogener Daten ein Datensumpf entsteht, wird in Theorie und Praxis gesehen. Ein schlecht dokumentierter und unorganisierter Data Lake erschwert die Datenentdeckung und verringert das Vertrauen in die Daten. Das entwickelte Vorgehensmodell dient der Vermeidung dieser Gefahr. Durch die Anpassung und Erweiterung der Data Governance Richtlinien, das Definieren von Datenzonen, das Integrieren eines ganzheitlichen Metadatenmanagements, die Identifikation von Anwendungsfällen und die Auswahl von Data Stewards wird ein Rahmen geschaffen, um einen möglichst hohen Geschäftsnutzen aus dem Data Lake zu ziehen. Dabei werden die Agilität und Flexibilität des Data Lakes gewährleistet, um skalierbar und kosteneffizient große Mengen an Rohdaten zu verarbeiten.

Die Identifikation von Anwendungsfällen mit hohem Geschäftsnutzen unterstützt die Vermeidung eines Datensumpfes. Dabei kann mithilfe prototypischer Erstellung von integrierten Datenbeständen der Nutzen der Datenintegration mit geringem Aufwand überprüft werden. Bei der Auswahl und Priorisierung der Anwendungsfälle kann es zudem sinnvoll sein, die Komplexität der Integration zu Beginn zu berücksichtigen, um erste Erfahrungen mit dem Data Lake zu sammeln.

Self-Service Funktionen, um Business Usern die Datenabfrage und -analyse zu vereinfachen, können auf Grundlage von Metadaten und Zonen aufbereiteter, den Bedürfnissen der Nutzer angepassten Daten umgesetzt werden. Business Usern das Entdecken und Analysieren relevanter Daten möglich zu machen, fördert den Nutzen eines Data Lakes und verhindert Engpässe bei IT- und Analytics-Ressourcen.

Data Governance Richtlinien sollten für den Data Lake überprüft und angepasst werden. Entgegen dem strikten Governance Konzept im Data Warehouse werden, unterstützt vom Zonenmodell, verschiedene Governance-Stufen abgedeckt, um dabei das richtige Maß zwischen Kontrolle und Flexibilität zu gewährleisten. Die Datenstrategie sollte darauf ausgerichtet sein, die Daten als Asset zu betrachten und die Vorteile des Data Lakes zu nutzen.

Das Metadatenmanagement im Data Lake ist so auszulegen, dass einerseits die Nutzergruppen bei ihrer Arbeit mit den Daten unterstützt werden und andererseits die Richtlinien und Verantwortlichen der Data Governance dokumentiert werden. Dabei wird zur Verwaltung der Metadaten sowohl in der Literatur als auch in einem Experteninterview ein Datenkatalog als geeignet angesehen.

Die erfolgreiche Umsetzung eines Data Lake Konzepts wird nicht nur von einer zielgerichteten Einführung beeinflusst, sondern auch von der kontinuierlichen Pflege und Kontrolle der Daten im Data Lake. Zu diesem Zweck werden für alle Datensätze Data Stewards definiert, die verantwortlich für die Verwaltung der Daten über ihre Lebensdauer sind.

Die Fortsetzung dieses Forschungsansatzes ist die praktische Anwendung des Vorgehensmodells, um die Umsetzbarkeit und den Nutzen zu überprüfen. Auf diese Weise können Verbesserungsmaßnahmen identifiziert und integriert werden. Da das Modell einen generischen Ansatz darstellt, ist es für die individuellen Bedürfnisse und Eigenschaften der Unternehmen und die Ziele der Data Lake Projekte anzupassen. Zukünftige Forschungen könnten sich mit Einflussfaktoren bei der Umsetzung und mit Modifikationen des Vorgehensmodells beschäftigen.

Da bei der Einführung eines Data Lakes, die technologische Umsetzung und die Architektur eine entscheidende Rolle spielen, ist es sinnvoll, eine Referenzarchitektur zu konzipieren, die die wichtigen Aspekte zur erfolgreichen Einführung eines Data Lakes aus dieser Arbeit integriert und eine zielgerichtete Realisierung unterstützt.