Hausarbeit

Cloud

Apache Airflow

|  |  |
| --- | --- |
| Verfasser: | Albrecht, Florian |
| Arbeit eingereicht am: | 17.07.2023 |
| Prüfer: | Prof. Dr. Blochinger, Wolfgang |
| Fach: | Cloud & Big Data Technologien |
| Studiengang: | Wirtschaftsinformatik (M.Sc.) |
| Fachsemester: | 1 |

Inhaltsverzeichnis

[1 Einleitung 1](#_Toc140178303)

[1.1 Motivation 1](#_Toc140178304)

[1.2 Einführung Apache Airflow 1](#_Toc140178305)

[1.3 Zielsetzung & Struktur 2](#_Toc140178306)

[2 Einführung in Apache Airflow 3](#_Toc140178307)

[3 Verwendung von Apache Airflow 4](#_Toc140178308)

[4 Anwendungsbeispiele und Best Practices 5](#_Toc140178309)

[5 Vergleich mit anderen WfMS wie Luigi, Azkaban, etc. 6](#_Toc140178310)

[6 Herausforderungen und zukünftige Entwicklungen 7](#_Toc140178311)

[7 Fazit 8](#_Toc140178312)

[Literaturverzeichnis III](#_Toc140178313)

# Einleitung

Dieses Kapitel hat den Zweck, diese Hausarbeit einzuführen. Zu Beginn wird der Verwendungszweck von Workflow-Management-Systemen motiviert.

## Motivation

In der Regel teilen traditionelle Anwendungssysteme betriebliche Funktionen voneinander ab und ermöglichen somit keine ausreichende Unterstützung von Geschäftsprozessen, bei denen häufig mehrere Funktionen berücksichtigt werden müssen. Workflow-Management-Systeme (WfMS) hingegen ermöglichen es, Geschäftsprozesse kontinuierlich abzuwickeln und eventuell zu steuern. Es wird immer wieder versucht, WfMS im Controlling-Bereich zu nutzen, um stark standardisierte Prozesse zu unterstützen. WfMS sind Softwaresysteme, welche die Durchführung von Aufgaben durch verschiedene Bearbeiter koordinieren. Dies machen sie, indem sie die benötigten Personen über anstehende Aufgaben im Geschäftsprozess informieren, wichtige Informationen zur Verfügung stellen und entsprechende Werkzeuge bzw. Anwendungssysteme zur Unterstützung der einzelnen Bearbeiter zur Verfügung stellen (vgl. Workflow-Management-Systeme 2006, S. 199).

## Einführung Apache Airflow

In der Wirtschaft bieten mehrere Unternehmen diverse WfMS an. Eine Softwarelösung wird von der Apache Software Foundation bereitgestellt: Apache Airflow. Apache Airflow ist eine Open-Source-Lösung, welche verwendet werden kan, um Workflows zu erstellen, zu verwalten und zu überwachen. Airflow hat eine grafische Benutzeroberfläche und wurde in Python geschrieben. Die Software hat sich zu einem Standard für Workflow-Management entwickelt (vgl. Luber und Litzel 2020).

Im Rahmen dieser Hausarbeit werden wir uns eingehend mit Apache Airflow befassen und seine Funktionen, Architektur und Verwendung untersuchen. Wir werden auch die Vor- und Nachteile von Apache Airflow im Vergleich zu anderen Workflow-Management-Systemen analysieren und prüfen, wie es in verschiedenen Anwendungsfällen eingesetzt wird.

## Zielsetzung & Struktur

Das Ziel dieser Ausarbeitung ist es, Apache Airflow als WfMS näher zu bringen. Um dies zu erreichen wird die Verwendung von Airflow erläutert anhand von verschiedenen Branchen. Um dieses Ziel zu erreichen wird diese Hausarbeit in der folgenden Struktur aufbereitet.

Kapitel 2 führt Apache Airflow ein. Hier wird zuerst auf den historischen Ursprung eingegangen und anschließend werden die wichtigsten Komponenten von Airflows Architektur vorgestellt und erläutert. Abgeschlossen wird dieses Kapitel mit der Vorstellung der DAGs (Directed Acyclic Graph) und den dazugehörigen Tasks.

Kapitel 3 hingegen beschäftigt sich mit der initialen Installation & Konfiguration von Airflow. Ebenfalls werden hier wichtige Funktionen, wie z. B. die Überwachung von Workflows aufgezählt und dargestellt. Abschließend werden mögliche Anbindung bzgl. anderen Programmiersprachen, Cloud-Services etc. aufgezählt.

Das Kapitel 4 beschäftigt sich mit praktischen Anwendungsfällen. Hier werden Use Cases von verschiedenen Branchen aufgezählt. Ebenso werden Best Practices für DAGs aufgestellt, um die Entwicklung effizienter gestalten zu können.

Kapitel 5 befasst sich im Kern mit dem Vergleich von anderen WfMS. Hier wird Airflow u. a. mit Luigi uns Azkaban verglichen.

Im sechsten Kapitel werden mögliche Herausforderungen bei einer Nutzung von Airflow evaluiert. Zusätzlich werden mögliche Verbesserungen für dieses System aufgezählt.

Abgeschlossen wird diese Hausarbeit mit einem Fazit. Dieses fasst die Hausarbeit kurz zusammen. Ebenfalls werden die großen Vor- und Nachteile explizit aufgezählt.

# Einführung in Apache Airflow

Ursprung und Entwicklung von Apache Airflow

Architektur und Komponenten von Apache Airflow (Scheduler, Executor, Webserver, etc.)

Konzepte wie Directed Acyclic Graphs (DAGs), Operators und Tasks

# Verwendung von Apache Airflow

Installation und Konfiguration von Apache Airflow

Erstellung und Ausführung von DAGs

Überwachung, Fehlerbehandlung und Wiederherstellung von Workflows

+ Cloud anbindungen und so

# Anwendungsbeispiele und Best Practices

Praktische Anwendungsfälle von Apache Airflow in verschiedenen Branchen (Data Engineering, Data Science, etc.)

Best Practices für die effiziente Gestaltung und Organisation von DAGs

Performance-Optimierung und Fehlervermeidung in Apache Airflow

# Vergleich mit anderen WfMS wie Luigi, Azkaban, etc.

# Herausforderungen und zukünftige Entwicklungen

Herausforderungen bei der Verwendung von Apache Airflow (Skalierbarkeit, Monitoring, etc.)

Aktuelle Trends und zukünftige Entwicklungen von Apache Airflow

Potenzielle Erweiterungen und Verbesserungen der Plattform

# Fazit

Zusammenfassung der wichtigsten Erkenntnisse und Ergebnisse der Hausarbeit

Bewertung der Vor- und Nachteile von Apache Airflow

Ausblick auf zukünftige Anwendungen und Relevanz von Apache Airflow

Literaturverzeichnis

Luber, Stefan; Litzel, Nico (2020): Was ist Apache Airflow? Online verfügbar unter https://www.bigdata-insider.de/was-ist-apache-airflow-a-948609/, zuletzt aktualisiert am 17.07.2020, zuletzt geprüft am 13.07.2023.

Workflow-Management-Systeme (2006). In: *Z Control Manag* 50 (4), S. 199–200.