HTWK

Hochschule für Technik,
Wirtschaft und Kultur Leipzig

Exposé für eine Masterarbeit

zur Erlangung des akademischen Grades

Master of Science

im Studiengang Informatik
der Fakultät Informatik und Medien
der Hochschule für Technik, Wirtschaft und Kultur Leipzig

Performance-Evaluation von Scheduling-Tools für den Einsatz im Data-Warehousing

Vorgelegt von:	Raphael Drechsler
Anschrift:	Kieler Str. 34, 04357 Leipzig
Kontaktdaten:	Tel.: +49 1525 4194262 E-Mail: raphael.drechsler@googlemail.com
Matrikelnummer:	69872
Fachsemester:	4
Erstgutachter:	Prof. Dr.-Ing. Thomas Kudrass (HTWK)
Kontaktdaten:	Tel.: +49 341 3076-6420 E-Mail: thomas.kudrass@htwk-leipzig.de
Zweitgutachter:	B.Sc. Torsten Böttcher (integration-factory GmbH & Co. KG)
Kontaktdaten:	Tel.: +49 69 25669269-0 E-Mail: boettcher@integration-factory.de
Vorauss. Abgabedatum	Dezember 2019
Datum	3. Juli 2019

Inhaltsverzeichnis

1 Problemstellung

Das Unternehmen *integration-factory GmbH & Co. KG* (im Folgenden *integration-factory* genannt) ist ein Consulting-Unternehmen und bietet für seine Kunden Lösungen in den Bereichen Business Intelligence und insbesondere Integration von Unternehmensdaten an. Gegenwärtig besteht dabei in zwei Kunden-Projekten eine Situation, in der durch *integration-factory* eine DWH-Lösung bereitgestellt wurde, welche von einer Marktanalyse von Scheduling-Tools sowie einer Betrachtung von ETL-Worklflow-Optimierung profitieren könnte.
Im Folgenden sollen diese zwei Kundenszenarios beschrieben werden.

Kundenszenario A In Kunden-Solution A ist bereits das kommerzielle ETL-Scheduling-Tool *Control-M* von *BMC Software* im Einsatz. Mit den derzeit im Projekt umgesetzten ETL-Workflows treten Probleme mit unnötigen Wartezeiten bei Statusübergängen im Workflow auf. Es besteht der Bedarf an einer performanteren Lösung. Durch die Umsetzung dieser Arbeit sollen für diese Kundensituation entsprechende Handlungsoptionen evaluiert werden. Dies soll zum einen durch die genauere Betrachtung der in *Control-M* umgesetzten Workflows und zum anderen durch einen Vergleich der *Control-M*-Lösung mit weiteren Scheduler-Alternativen geschehen.

Kundenszenario B In Kunden-Solution B ist gegenwärtig kein dediziertes Scheduling-Tool im Einsatz. Eingesetzt wird das ETL-Tool *Informatica PowerCenter* von *Informatica*, welches Funktionalitäten für ein rudimentäres Scheduling von ETL-Workflows bereitstellt. Über die Integration eines Open-Source Scheduling-Tools in die Kunden-Solution besteht die Möglichkeit, komplexere ETL-Workflows abzubilden.
Durch die Umsetzung dieser Arbeit sollen für diese Kundensituation infrage kommende Open-Source-Kandidaten gefunden und verglichen werden.

2 Zielsetzung

Aus der Problemstellung ergeben sich im Wesentlichen die folgenden Punkte für die Zielsetzung.

Hauptziel 1: Marktanalyse von ETL-Schedulern Es soll eine Marktanalyse für ETL-Scheduler durchgeführt werden, um einen Überblick über die zur Verfügung stehenden Optionen zu gewinnen. Da die Suche nach einem adäquaten Scheduling-Tool primär durch Szenario B getrieben ist, liegt der Fokus dabei vorrangig auf Open-Source-Tools. Um einordnen zu können, was ggf. ein Austausch des Schedulers in Szenario A bedeuten würde, sollen jedoch auch *Control-M* und ggf. weitere kommerzielle Scheduler-Tools in die Analyse mit einbezogen werden. Ebenso sind die rudimentären Scheduling-Funktionalitäten von *Informatica* in den Vergleich mit einzubezeihen, um den Gewinn an Features durch eine Integration eines dedizierten Scheduling-Tools in

die Kunden-Solution B bemessen zu können.

Hauptziel 2: Erstellen einer prototyischen Solution Nach Abschluss der Evaluation soll eine prototypische Solution umgesetzt werden, welche in das Lösungs-Portfolio von *integration-factory* aufgenommen und zur Überzeugung der Kunden von einer entsprechenden Anpassung der bestehenden Kunden-Solution verwendet werden kann.

Untersuchen von Optimierungsmöglichkeiten für bestehende Workflows Ggf. können bestehende, komplexere Workflows auf Optimierungsmöglichkeiten bezüglich Parallelisierung und Sequentialisierung untersucht werden.

3 Vorgehensweise

3.1 Vorbereitung Evaluation

Zunächst sollen die theoretischen Grundlagen besprochen werden. Dabei soll auf die Themen ETL-Workflows, deren Einordnung in den DWH-Prozess eingegangen und die Anforderungen an einen ETL-Scheduler zusammengetragen werden. Anschließend sollen die speziellen Kundensituationen betrachtet und sich daraus ergebende, zusätzliche Anforderungen aufgenommen werden.
Anhand der gesammelten Anforderungen erfolgt das Aufstellen von Bewertungskriterien, Mindestanforderungen und Bewertungsmaßstäben, sowie die Gewichtung der Kriterien nach dem Scoring-Modell.
Für die durchzuführenden Performance-Messungen sollen im Rahmen der Vorbereitung die in den Kundenszenarios implementierten Workflows betrachtet und repräsentative Referenz-Workflows abgeleitet werden. Ggf. werden bei der genauen Betrachtung der bestehenden Workflows Optimierungsmöglichkeiten bezüglich Parallelisierung und Sequentialisierung sichtbar.
Für die Einrichtung des Systems, auf welchem die Evaluation durchgeführt werden soll, sollen aus den Beschaffenheiten der Kunden-Systeme die entsprechenden System-Anforderungen abgeleitet werden.

3.2 Kandidaten suchen

Im Anschluss an die Vorbereitung erfolgt die Suche von kommerziellen und Open-Source Scheduler-Tool-Kandidaten für die Evaluation. Diese sollen in der schriftlichen Ausarbeitung vorgestellt und bereits an dieser Stelle gegen die zuvor definierten Mindestanforderungen geprüft werden.

3.3 Durchführung Evaluation

In der Evaluation werden für alle Kandidaten die vorbereiteten Bewertungskriterien anhand der definierten Bewertungsmaßstäbe bepunktet. Die Teil-Ergebnisse werden in einer Bewertungsmatrix festgehalten.
Die Evaluation wird primär auf die Performance der Tools ausgerichtet sein. Die zuvor definierten Referenz-Workflows werden dabei mit den jeweiligen Tools umgesetzt und anschließend Laufzeitmessungen vorgenommen. Bei der Umsetzung der Workflows in den jeweiligen Tools können dabei voraussichtlich weitere Kriterien, wie z.B. intuitive Benutzerführung und Bedienbarkeit, bewertet werden. Sind alle Kriterien für alle Kandidaten bewertet, erfolgt die Auswertung der Bewertungsmatrix und somit das Feststellen des Gewinner-Tools.

3.4 Aufbau der prototypischen Solution

Da nicht sichergestellt werden kann, dass die betreffenden Kunden eine Anpassung ihrer Solution wünschen, soll das Ergebnis der Evaluation im Rahmen dieser Arbeit zunächst dazu verwendet werden, eine prototypische Solution zu erstellen. Diese kann in das Lösungs-Portfolio von *integration-factory* aufgenommen und anschließend dazu verwendet werden, die Kunden von einer entsprechenden Anpassung zu überzeugen.
Für den Aufbau der prototypischen Solution sollen zunächst die System-Anforderungen und die umzusetzende Tool-Landschaft in der schriftlichen Ausarbeitung festgehalten werden (Wobei sich voraussichtlich die System-Anforderungen mit denen der Evaluations-Umgebung decken werden).
Für die Verwendung des Prototyps als Demo-System ist ggf. die Umsetzung weiterer Workflows sinnvoll, damit sich der Kunde im angepassten System schnell wiederfindet. Für das Szenario B können neue Workflows erstellt werden, welche den Feature-Gewinn durch Einsatz des dedizierten Scheduling-Tools aufzeigen.
Abschließend soll eine Einordnung der Eigenschaften des Prototyps in das Kano-Modell erfolgen. Ggf. ist es hierbei sinnvoll, die Umsetzung weiterer Workflows und die Betrachtung der voraussichtlichen Kunden-Zufriedenheit mittels Kano-Modell als iterativen Prozess erfolgen zu lassen.

4 Voraussichtliche Gliederung

Aus der geschilderten Vorgehensweise lässt sich die folgende voraussichtliche Gliederung ableiten.

1. Einleitung
   1.1. Ausgangssituation, Problemstellung und Ziel
   1.2. Vorgehensweise
2. Vorbereiten Evaluation

   2.1. Einordnung ETL-Workflows in DWH-Prozess
   2.2. Anforderungen an einen ETL-Scheduler
   2.3. Anforderungen aus kundenspezifischen Begebenheiten
   2.4. Bewertungskriterien und Bewertungsmaßstäbe
   2.5. Referenz-Workflows
   2.6. Anforderungen an das Evaluations-System
3. Scheduling-Tool-Kandidaten
4. Durchführung Evaluation
   4.1. Kandidat 1
   4.2. Kandidat 2
   4.3. ...
   4.4. Control-M
   4.5. ETL-Tool: Informatica PowerCenter
5. Auswertung der Evaluation
6. Aufbau der prototypischen Solution
   6.1. System und Tool-Landschaft
   6.2. Umsetzung weiterer Workflows
   6.3. Eigenschaften der prototypischen Solution im Kano-Modell
7. Fazit

5 Erste Literaturverweise

Generelle Literatur zu DWH: z.B. *Schnider, Dani, Jordan, Claus, Welker, Peter, and Wehner, Joachim. Data Warehouse Blueprints : Business Intelligence in Der Praxis. 2016. Web.* [1]

Literatur zur Optimierung von ETL-Workflows (Parallelisierung und Sequentialisierung): *Karagiannis, Vassiliadis, and Simitsis. SScheduling Strategies for Efficient ETL Execution.Ïnformation Systems 38.6 (2012): 927-945. Web.* [2]

Vergleich von Scheduling-Tool-Features über Informationsmaterial der Anbieter: z.B. *Cisco Tidal Enterprise Scheduler Data Sheet*[3]

Nutzwertanalyse nach Scoring-Modell: z.B. *Nöllke, Matthias. Entscheidungen Treffen : Schnell, Sicher, Richtig. 5. Aktualisierte Auflage. ed. München: Haufe-Lexware GmbH & KG, 2011. Web.*[4]

6 Zeitplan

Grobe Eckdaten:

15.07.19	Anmeldung, Start Einarbeitung
15.08.19	Abschluss Einarbeitung, Start Evaluation
15.11.19	Abschluss Evaluation, Start Erstellung Prototyp
15.12.19	Abschluss Erstellung der Prototypischen Solution
15.01.20	Abschluss schriftlicher Ausarbeitung und Einreichen der Arbeit (spätester Termin)

Der 15.01.2020 stellt den spätesten Termin für das Einreichen der Arbeit beim Prüfungsamt dar. Ein früherer Termin für Abschluss der ist angestrebt. Eine genauere Terminierung der Abgabe erfolgt mit der Erstellung eines detaillierteren Zeitplanes.

Literatur

[1] D. Schnider, *Data Warehouse Blueprints : Business Intelligence in der Praxis*. 2016.

[2] A. Karagiannis, P. Vassiliadis, and A. Simitsis, "Scheduling strategies for efficient etl execution," *Information Systems*, vol. 38, no. 6, 2012.

[3] I. Cisco Systems, "Cisco tidal enterprise scheduler data sheet." https://www.cisco.com/c/en/us/products/collateral/cloud-systems-management/tidal-enterprise-scheduler/c78-636900-00_cte_scheduler.html, 2019. [Online; Stand 26. Juni 2019].

[4] M. Nöllke, *Entscheidungen treffen : Schnell, sicher, richtig*. München: Haufe-Lexware GmbH & Co. KG, 5. aktualisierte auflage. ed., 2011.