h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Hochschule Darmstadt
-Fachbereich Informatik-

Evaluierung der Qualität von Open Source Stream Processing Frameworks

zur Erlangung des akademischen Grades
Master of Science (M. Sc.)

vorgelegt von

Simon Kern
737708

Referentin:	Prof. Dr. Uta Störl
Korreferentin:	Prof. Dr. Inge Schestag
Ausgabedatum:	16.03.2015
Abgabedatum:	16.09.2015

Kurzfassung

Evaluierung der Qualität von Open Source Stream Processing Frameworks

Big Data ist im Bereich der Informationstechnologie aktuell eines der großen Schlagworte. Technologisch bezieht sich der Begriff meist auf die Stapelverarbeitung von gesammelten Daten. Im Gegensatz dazu verarbeiten Stream Processing Frameworks Datenströme direkt. Innerhalb der Open-Source-Community sind mehrere konkurrierende Systeme für diesen Aufgabenbereich entstanden. Bei der Neuentwicklung einer Anwendung muss jedoch ein Framework als Basis ausgewählt werden. Ziel der Arbeit ist es daher, aktuelle Open Source Stream Processing Frameworks hinsichtlich ihrer Qualität zu evaluieren, um diese Auswahl zu unterstützen.

Hierzu stellt diese Arbeit eine Referenzarchitektur vor, anhand derer die Komponenten eines Stream Processing Frameworks zur Verbesserung der Vergleichbarkeit zugeordnet werden können. Mithilfe des Evaluierungsprozesses nach ISO 25000 und insbesondere des Softwarequalitätsmodells nach ISO 25010 werden Kriterien zur Evaluierung von Stream Processing Frameworks entwickelt. Zur Erhöhung der praktischen Relevanz werden die Methoden Cognitive Walkthrough und Cognitive Dimensions zur Feststellung der Benutzerfreundlichkeit genutzt. Dafür werden Aufgaben definiert, die mithilfe der ausgewählten Frameworks implementiert werden. Sie werden zusätzlich zur Durchsatz- und Latenzmessung herangezogen. Als Rahmen der Aufgaben dient die Beobachtung von Wertpapieren, die an der New York Stock Exchange gelistet sind.

Für die Evaluierung werden Apache Flink, Apache Samza, Apache Spark und Apache Storm als Kandidaten identifiziert. Sie werden anhand der entwickelten Kriterien verglichen. Dabei zeigt sich, dass kein Framework alle Anforderungen vollständig abdeckt. Vielmehr müssen die Ergebnisse dieser Arbeit nach den konkreten Anforderungen einer Neuentwicklung gewichtet werden, um eine Auswahlentscheidung zu treffen.

Abstract

Evaluation of the Quality of Open Source Stream Processing Frameworks

Big Data is a huge trend in current information technology. This term mostly refers to software that does batch processing of stored data. But there is also software, namely stream processing frameworks, which operates directly on data streams. Several competing projects have emerged in the open source community, with the goal to provide software that addresses this issue. When developing a new application, one has to choose between these alternatives. Therefore, the goal of this thesis is to evaluate the quality of open source stream processing frameworks, in order to support decision making.

For this purpose, a reference architecture is proposed. Stream processing frameworks are composed of components that can be mapped onto this reference architecture. This facilitates an easy comparison of multiple frameworks and establishes a common understanding of terminology. The evaluation criteria are developed according to the process described in ISO 25000 and especially derived from the software quality model specified in ISO 25010. To increase the practical relevance of this evaluation, the usability inspection methods cognitive walkthrough and cognitive dimensions are used. These methods are based on predefined tasks that are implemented utilizing the frameworks under test. This implementation is later used to benchmark the frameworks in terms of throughput and latency. The tasks are built around the observation of stock pricing at the New York Stock Exchange.

The candidates for the evaluation are Apache Flink, Apache Samza, Apache Spark and Apache Storm. They are evaluated according to the developed criteria. The results show that no candidate meets all requirements. In fact, the results of this thesis have to be weighted to match the specific requirements of the application to be implemented, to support the decision making.