Design and Implementation of a Tool to Collect Execution- and Service-Data of Big Data Analytics Applications

Bachelor's Thesis

for obtaining the academic degree
Bachlor of Science (B.Sc.)

at

Beuth Hochschule für Technik Berlin
Department Informatics and Media VI
Degree Program Mediainformatics

1. Examiner and Supervisor:	Prof. Dr. Stefan Edlich
2. Examiner:	Prof. Dr. Elmar Böhler

Submitted by:	Markus Lamm
Matriculation number:	s786694
Date of submission:	06.09.2016

Ackknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Inhaltsverzeichnis

1 Introduction

1.1 Motivation

According to a survey in Germany, nine out of ten companies (89 percent) analyze large volumes of data for operational decision-making processes using modern Big Data Analytics Architectures, where 48 percent of respondents see the greatest potential of Big Data [Bar14]. The analysis of continuous data streams is taking up a growing importance for companies and therefore constitutes an important factor for business success.

Collecting, storing and analyzing system and operational data of Big Data Architectures is therefore an essential tool in order to ensure successful operation. By analyzing execution and service data, problems can be tracked and potential sources of error identified as early as possible.

1.2 Objective

The main goal of the thesis is the design and implementation of a software system to ingest and store system and operational data of Big Data Analytics Applications. It should be investigated which data is available for Apache Flink and Apache Kafka, what data is relevant and shall be collected, how to collect from source systems and how the data can be stored in a centralized persistence system.

1.3 Structure of thesis

2 Theoretical Foundations

2.1 Big-Data-Analytics-Architectures

2.1.1 Definition Big Data

2.2 Stream-Processing

Stream processing is the real-time processing of data continuously, concurrently, and in a record-by-record fashion. in which data is treated not as static tables or files, but as a continuous infinite stream of data integrated from both live and historical sources.

Benefits:

- Accessibility: live data can be used while still in motion, before being stored.
- Completeness: historical data can be streamed and integrated with live data for more context.
- High throughput: high-velocity, high-volume data can be processed with minimal latency.

2.2.1 Apache Flink

2.2.2 Apache Kafka

3 Data Analysis

3.1 System data

3.2 Java Management Extensions (JMX)

3.3 Representational State Transfer (REST)

3.4 Data Quality

4 Requirements

4.1 Collection

4.2 Transport

4.3 Persistence

5 Architecture and Implementation

5.1 Collected data as time-series based stream

5.2 Microservices and Service-Discovery

5.3 System components

5.3.1 CollectorClient

5.3.2 CollectorManager

5.3.3 Message-Broker

5.3.4 Indexer

5.3.5 Persistence

6 Evaluation

7 Conclusion

Abbildungsverzeichnis

Tabellenverzeichnis

Quelltextverzeichnis

Onlinequellen

[Bar14] Jörg Bartel. *Big Data Technologien*. 2014. URL: https://www.bitkom.org (besucht am 06. 08. 2016).

A

A.1 Diagramm

A.2 Tabelle

A.3 Screenshot

A.4 Graph

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Stadt, den xx.xx.xxxx

Max Mustermann