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.

Contents

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

8 Beispiele

Im Kapitel Beispiele (siehe chapter 8) werden die möglichen Funktionen und Möglichkeiten dies LaTeX-Dokuments demonstriert.

8.1 Quelltext

Nachfolgend der Codeauszug 8.1.

```
1 /**
2 * The HelloWorldApp class implements an application that
3 * simply prints "Hello World!" to standard output.
4 */
5 class HelloWorldApp {
6   public static void main(String[] args) {
7     System.out.println("Hello World!"); // Display the string.
8   }
9 }
```

Codeauszug 8.1: Hello World

8.2 Bild

Die rechts zu sehende Grafik demonstriert die Möglichkeiten des Paketes „wrapfig“. Grafiken innerhalb einer „wrapfigure“ können entweder links oder rechts von Text umlaufen werden.

Die nachfolgende Figure 8.2 demonstriert die Darstellung eines „*.jpg“ Bildes innerhalb des Textes (beim Einfügen kann auf die Endung verzichtet werden, solange der Name einzigartig ist). Zusätzlich enthält dieses einen Untertitel der über das bereits verwendete Label verlinkt werden kann. Der Untertitel erscheint im Abbildungsverzeichnis (Abbvz.).

Figure 8.1: Beispielbild [PEX]

8.3 Text Formatierungen und sonstiges

Dieser Text enthält eine Fußnote[1].

8.3.1 Listen

Listen könne sowohl mit Bullet points als auch mit Zahlen erstellt werden

- Eine Liste mit Bullet points
- Ein weiteres Element

1. Eine Liste mit Zahlen
2. Ein weiteres Element

[1] Fußnoten sind Anmerkungen, die im Druck-Layout aus dem Fließtext ausgelagert werden, um den Text flüssig lesbar zu gestalten.

8.3.2 Text Hervorhebungen

> The problem with internet quotes is that you can't always depend on their accuracy
>
> — Abraham Lincoln, 1864

"Inspirierende Zitate können mit epigraph eingefügt werden

> The problem with internet quotes is that you can't always depend on their accuracy
>
> Abraham Lincoln, 1864

Seitenumbrüche können nur direkt nach Text geschrieben werden, sonst lässt sich das Latex nicht mehr compilieren.

Figure 8.2: Beispielbild [PEX]

8.4 Tabelle

Nachfolgend Table 8.1.

Inhaber:
Alice
Peer (Ersteller):
Bob
Öffentlicher Schlüssel des Inhabers:
F2 D2 0E ED FA 4E 9E 0A F2 DD 23 8A 32 44 F3 E9
Gültigkeit:
2015-07-01 – 2016-06-30

Table 8.1: Digitales Zertifikat

8.5 Long-Table

Die „Long-Table“kann über definierte Header und Footer über Seitenumbrüche hinweg angezeigt werden.

Version	Codename	API	Verteilung
2.2	Froyo	8	0.1%
2.3.3 - 2.3.7	Gingerbread	10	2.7%
4.0.3 - 4.0.4	Ice Cream Sandwich	15	2.5%
4.1.x	Jelly Bean	16	8.8%
4.2.x		17	11.7%
4.3		18	3.4%
4.4	KitKat	19	35.5%

Fortsetzung auf nachfolgender Seite

Fortsetzung - Verteilung der Androidversionen (Stand 01.02.2016)

Version	Codename	API	Verteilung
5.0	Lollipop	21	17.0%
5.1		22	17.1%
6.0	Marshmallow	23	1.2%

Table 8.2: Verteilung der Androidversionen (Stand: 01.02.2016)

8.6 Literaturverweis

Weil für die alte und die neue Rechtschreibung verschiedene Trennregeln gelten, sind Deutsch mit alter Rechtschreibung und Deutsch mit neuer Rechtschreibung zwei verschiedene Sprachen ([Kna09], S. 192).

8.7 Onlineverweise

Siehe Google.de [Goo].

8.8 Glossar

Der Glossar enthält die Beschreibung verwendeter Begriffe für das bessere Verständnis gegenüber dem Leser. Beispiele sind: Berlin, Outsourcing, Application Service Providing, Policy und PCI Express.

8.9 Abkürzungsverzeichnis

Das Abkürzungsverzeichnis listet alle verwendeten Abkürzungen auf. Einige Beispiele sind Serial Attached SCSI (SAS), Compact Disk (CD), Local Area Network (LAN) und

Internationale Organisation für Normung (ISO). Die erneute Verwendung zeigt nur noch die Abkürzung: SAS, CD, LAN und ISO.

List of Figures

List of Tables

List of Source Codes

Index

Glossar

Application Service Providing Der Application Service Provider (Abkürzung ASP) bzw. Anwendungsdienstleister ist ein Dienstleister, der eine Anwendung (z. B. ein ERP-System) zum Informationsaustausch über ein öffentliches Netz (z. B. Internet) oder über ein privates Datennetz anbietet. Der ASP kümmert sich um die gesamte Administration, wie Datensicherung, das Einspielen von Patches usw. Anders als beim Applikations-Hosting ist Teil der ASP-Dienstleistung auch ein Service (z.B. Benutzerbetreuung) um die Anwendung herum. 12

Berlin Berlin ist die Bundeshauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. Die Stadt Berlin ist mit über 3,4 Millionen Einwohnern die bevölkerungsreichste und mit 892 Quadratkilometern die flächengrößte Gemeinde Deutschlands sowie nach Einwohnern die zweitgrößte der Europäischen Union. Sie bildet das Zentrum der Metropolregion Berlin/Brandenburg (6 Millionen Einw.) und der Agglomeration Berlin (4,4 Millionen Einw.). Der Stadtstaat unterteilt sich in zwölf Bezirke. Neben den Flüssen Spree und Havel befinden sich im Stadtgebiet kleinere Fließgewässer sowie zahlreiche Seen und Wälder. 12

Outsourcing Outsourcing bzw. Auslagerung bezeichnet in der Ökonomie die Abgabe von Unternehmensaufgaben und -strukturen an externe oder interne Dienstleister. Es ist eine spezielle Form des Fremdbezugs von bisher intern erbrachter Leistung, wobei Verträge die Dauer und den Gegenstand der Leistung fixieren. Das grenzt Outsourcing von sonstigen Partnerschaften ab. 12

PCI Express PCI Express („Peripheral Component Interconnect Express“, abgekürzt PCIe oder PCI-E) ist ein Standard zur Verbindung von Peripheriegeräten mit dem

Chipsatz eines Hauptprozessors. PCIe ist der Nachfolger von PCI, PCI-X und AGP und bietet im Vergleich zu seinen Vorgängern eine höhere Datenübertragungsrate pro Pin.. 12

Policy Im geschäftlichen Bereich bezeichnet Policy eine interne Leit- bzw. Richtlinie, die formal durch das Unternehmen dokumentiert und über ihr Management verantwortet wird. 12

List of abbreviations

Abbvz. *Abbildungsverzeichnis* 9

CD *Compact Disk* 12, 13

ISO *Internationale Organisation für Normung* 13

LAN *Local Area Network* 12, 13

SAS *Serial Attached SCSI* 12, 13

Bibliography

[Kna09] Joerg Knappen. *Schnell ans Ziel mit LATEX 2e* -. ueberarbeitete und erweiterte Auflage. Muenchen: Oldenbourg Verlag, 2009. ISBN: 978-3-486-59015-9.

Online resources

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

[Goo] Google. *Google*. URL: http://www.google.de (visited on 10/06/2015).

Image resources

[PEX] PEXELS. *Black and white branches tree*. URL: https://www.pexels.com/photo/black-and-white-branches-tree-high-279/.

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