Master Thesis
Computer Science

UIMA, Docker and Kafka

Buzzwords oder doch interessant?

Simon Gehring

Am Jesuitenhof 3
53117 Bonn
simon.gehring@fkie.fraunhofer.de
Matriculation Number 2553262

At the
Rheinische Friedrich-Wilhelms-Universität Bonn
in cooperation with the
Fraunhofer-Institut für Kommunikation,
Informationsverarbeitung und Ergonomie

supervised by
Prof. Dr. Heiko Röglin and Daniel Töws

February 22, 2018

Contents

Chapter 1

Introduction

Natural language is part of everyone's everyday life and is most commonly used to transmit information human-to-human. While most of this interaction takes place orally or written on paper, the digital revolution and the rise of social media increased the amount of digitally stored natural language tremendously. Gantz and Reinsel predicted 2012 that the amount of digital data stored globally will double about every two years until at least the year 2020 [GR12].

Many opportunities arise from this amount of digital data, specifically in the field of machine learning. In 2011, IBM's QA (Question Answering) system "Watson" famously outmatched professional players in the quiz show "Jeopardy!" [Fer12, ESI+12]. Kudesia et al. proposed 2012 an algorithm to detect so called CAUTIs (Catheter-associated Urinary Tract Infections), common hospital-acquired infections, by utilizing a NLP (Natural Language Processing) analysis on the medical records of patients [KSDG12].

1.1 Motivation

1.2 Basics

1.2.1 UIMA

1.2.2 Docker

1.2.3 Hadoop

1.2.4 Spark

1.2.5 Kafka

1.3 Problem

1.3.1 Scaling UIMA

UIMA-AS

UIMA-CPM

1.3.2 Implementation Requirements

1.4 Related Work

1.4.1 Watson

1.4.2 Something else that warrants another subsection

1.4.3 Maybe something not UIMA related?

1.5 Outline

Chapter 2

Implementation

2.1 Concrete Application

2.2 Documents -> Kafka

2.3 Kafka -> Spark

2.4 Spark -> UIMA

2.5 UIMA -> Java

2.6 Kafka -> Output

2.7 Bottlenecks

Chapter 3

Evaluation

3.1 Computation Speed

3.2 Memory Usage

3.3 Extensibility

3.4 Maintainability

Chapter 4

Summary

4.1 The Judgement

Chapter 5

Future Work

Glossary

Catheter-associated Urinary Tract Infection

A urinary tract infection (UTI) is an infection involving any part of the urinary system, including urethra, bladder, ureters, and kidney. UTIs are the most common type of healthcare-associated infection reported to the National Healthcare Safety Network (NHSN). Among UTIs acquired in the hospital, approximately 75% are associated with a urinary catheter, which is a tube inserted into the bladder through the urethra to drain urine. Between 15-25% of hospitalized patients receive urinary catheters during their hospital stay. The most important risk factor for developing a catheter-associated UTI (CAUTI) is prolonged use of the urinary catheter. Therefore, catheters should only be used for appropriate indications and should be removed as soon as they are no longer needed. 3

Natural Language Processing

Natural-language processing (NLP) is a field of computer science, artificial intelligence concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language data. Challenges in natural-language processing frequently involve speech recognition, natural-language understanding, and natural-language generation. 3

Question Answering

Question answering (QA) is a computer science discipline within the fields of information retrieval and NLP, which is concerned with building systems that automatically answer questions posed by humans in a natural language. 3

Bibliography

[ESI+12] Edward A Epstein, Marshall I Schor, BS Iyer, Adam Lally, Eric W Brown, and Jaroslaw Cwiklik. Making watson fast. *IBM Journal of Research and Development*, 56(3.4):15–1, 2012.

[Fer12] David A Ferrucci. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1–1, 2012.

[GR12] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16, 2012.

[KSDG12] Valmeek Kudesia, Judith Strymish, Leonard D'Avolio, and Kalpana Gupta. Natural language processing to identify foley catheter–days. *Infection control and hospital epidemiology*, 33(12):1270–1272, 2012.

Eidesstattliche Erklärung

Hiermit versichere ich, Simon Gehring, dass ich die vorliegende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinne nach anderen Werken und Quellen, einschließlich Quellen aus dem Internet, entnommen sind, habe ich in jedem Fall unter Angabe der Quelle deutlich als Entlehnung kenntlich gemacht. Dasselbe gilt sinngemäß für Tabellen, Karten und Abbildungen.

Unterschrift: _____

Simon Gehring, Student
Universität Bonn