Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Daniel Kirchner

Skalierbare Datenanalyse mit Apache Spark

Beispielimplementation eines Influenza-Frühwarnsystems

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Daniel Kirchner

Skalierbare Datenanalyse mit Apache Spark

Beispielimplementation eines Influenza-Frühwarnsystems

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kahlbrandt
Zweitgutachter: Prof. Dr. Zweitprüfer

Eingereicht am: 1. Januar 2345

Daniel Kirchner

Thema der Arbeit

Skalierbare Datenanalyse mit Apache Spark Beispielimplementation eines Influenza-Frühwarnsystems

Stichworte

Schlüsselwort 1, Schlüsselwort 2

Kurzzusammenfassung

Dieses Dokument …

Daniel Kirchner

Title of the paper

Scalable Data Analysis with Apache Spark

Keywords

keyword 1, keyword 2

Abstract

This document …

Inhaltsverzeichnis

Listings

1 Einführung

1.1 Motivation

Der Bedarf auf großen Datenmengen zu operieren ist nicht neu. Spätestens seit in den späten Neunzigerjahren Suchmaschinenanbieter mit Mengen von Daten und Anfragen konfrontiert wurden, die eine nicht mehr wirtschaftlich durch einzelne Rechner zu bewältigen waren, wurden neue Verfahren benötigt. Algorithmen wurden nun auf die Eigenschaft optimiert möglichst effizient und fehlertolerant auf verschiedene Maschinen verteilbar zu sein.

Inzwischen ist die Analyse großer Datenmengen auch für Unternehmen und Einrichtungen interessant geworden, deren Kerngeschäft nicht die Daten selbst sind. Regierungen, Wissenschaftler, Industrieunternehmen, Militärs, Handelssoftware und viele andere treffen Entscheidungen auf Grundlage von Daten die die Kapazitäten einzelner Systeme weit übeschreiten.

Ständige Veränderung und Unvorhersehbarkeit der Anforderungen sind altägliche Praxis. Daten denen man in dem Moment keine Bedeutung beimisst, können sich in Zukunft als sehr kritisch erweisen und in einem anderen Kontext eine wichtige Rolle spielen. Dabei hat sich ein Paradigma als besonders wirksam herausgestellt: Daten werden erst durch die Abfrage in eine höhere Struktur gebracht, während bei der Speicherung nur ein Minimum an Struktur eingefordert wird.

Ein weiteres Problem ist die Vielfalt der möglichen Abfragen. Für bestimmte Aufgaben genügen Anfragen, wie an eine klassische Datenbank, für andere Probleme sind möglicherweise komplexere Graph-Analysen, das Anlernen von Maschinenlernalgorithmen oder einen Quasi-Echtzeit-Auswertung von Datenströmen gefordert. An dieser Stelle kommt Apache Spark ins Spiel, dass einen Versuch macht alle bisher genannten Probleme zu lösen.

1.2 Kontextabgrenzung

1.3 Relevante Produkte und Meilensteine

1.3.1 Überblick

1.3.2 Big Table

1.3.3 Map/Reduce

1.3.4 Hadoop

2 Vorstellung von Apache Spark

2.1 Übersicht

2.1.1 Architekturübersicht

2.1.2 Standardbibliotheken

Spark SQL

MLlib

Streaming

GraphX

2.2 Wesentliche Konzepte

2.2.1 Abgrenzung zu Hadoop

2.2.2 Resilient Distributed Datasets

3 Vorstellung des Beispiels

3.1 Aufgabenbeschreibung

3.2 Lösungsidee

3.2.1 1. Schritt: Ähnlichkeitsmaß für Wörter erzeugen

3.2.2 2. Schritt: Echtzeitbewertung von Textnachrichten aus einem Datenstrom

4 Implementation und Bewertung

4.1 Technischer Rahmen

4.1.1 OpenStack

4.2 Architekturübersicht

4.3 Architekturdetails

4.3.1 Modell für Ähnlichkeit von Wörtern mit MLlib erzeugen

4.3.2 Einlesen von Nachrichten aus dem Twitter Livestream

4.3.3 Verarbeiten und Bewerten der Nachrichten

4.4 Bewertung der Verfahren

5 Schlussbetrachtung

5.1 Kritische Würdigung der Ergebnisse

5.2 Ausblick und offene Punkte

See also One und Two (2010).

Literaturverzeichnis

[One und Two 2010] One, Author ; Two, Author: A Sample Publication. (2010)

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 1. Januar 2345 Daniel Kirchner