



Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Daniel Kirchner

Skalierbare Datenanalyse mit Apache Spark

*Fakultät für Technik und Informatik
Studiendepartment Informatik*

*Faculty of Engineering and Computer Science
Department of Computer Science*

Daniel Kirchner

Skalierbare Datenanalyse mit Apache Spark

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

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

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

Eingereicht am: 1. Januar 2345

Daniel Kirchner

Thema der Arbeit

Skalierbare Datenanalyse mit Apache Spark

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

1	Einführung	1
1.1	Motivation	1
1.2	Kontextabgrenzung	2
2	Vorstellung von Apache Spark	4
2.1	Überblick	5
2.2	Kernkonzepte	7
2.2.1	Resilient Distributed Datasets	7
2.2.2	Scheduling/Shuffling	10
2.2.3	Anwendungsdeployment und -lebenszyklus	10
2.2.4	Zusammenfassung und Bewertung	12
2.3	Standardbibliotheken	12
2.3.1	Dataframes/Spark SQL	13
2.3.2	MLlib	13
2.3.3	Streaming	13
2.3.4	GraphX	13
2.4	Betrieb und Security	13
2.5	Spark im Kontext von Parallelisierungspattern	13
2.6	Entwicklergemeinschaft	13
2.7	Verwandte Produkte	14
2.7.1	YARN	15
2.7.2	Mesos	15
2.7.3	Flink	15
2.7.4	MPI	15
2.7.5	Kafka	15
2.7.6	HBase	15
2.7.7	Lustre	15
3	Beispielanwendungen von Spark zur Datenanalyse	16
3.1	Anwendung zur Echtzeitanalyse eines Datenstroms anhand eines veränderlichen Stammdatenbestandes	16
3.1.1	Anforderungen	17
3.1.2	Hardwarekontext und Performance-Basisdaten	17
3.1.3	Lösungsskizze	18
3.1.4	Ergebnisse und Bewertung	20

3.2	Spark-basierte Implementation von Operatoren aus der Klimaforschung . . .	20
3.2.1	Beschreibung des Problems	20
3.2.2	Hardwarekontext und Performance-Basisdaten	21
3.2.3	Architekturübersicht	21
3.2.4	Detaillierte Lösungsbeschreibung	21
3.2.5	Ergebnisse	22
4	Schlussbetrachtung	23
4.1	Kritische Betrachtung der Ergebnisse	23
4.2	Ausblick und offene Punkte	23
	Acronyme	24
	Glossar	25
	Anhang	26
1	Installation der Plattform	27
2	Quellcode (Auszüge)	27
3	Sonstiges	27
3.1	Einschätzung des theoretischen Spitzendurchsatzes von Mittelklasse- Servern	27

Abbildungsverzeichnis

1.1	Google Trends	2
2.1	Verteilungsdiagramm einer typischen Sparkinstallation	6
2.2	Resilient Distributed Datasets aus Verteilungssicht	8
2.3	RDD Lineage vor Aktion	8
2.4	RDD Lineage nach Aktion	9
2.5	RDD Lineage nach Aktion und mit Persist()	9
2.6	Resilient Distributed Datasets mit Datenquelle aus Verteilungssicht	9
2.7	Application Deployment im Client Modus	11
2.8	Application Deployment im Cluster Modus	11
2.9	Application Deployment Prozess im Client Modus (vereinfacht)	12
2.10	Aktivität auf den offiziellen Spark Mailinglisten	14
3.1	Hardwareumgebung des Programms zur Tweetanalyse	17
3.2	Datenzentrierte Sicht auf die Komponenten	20
3.3	Verteilungssicht auf die Demo App	21
3.4	Komponentendiagramm des Demo App Packages	22

Tabellenverzeichnis

2.1	Theoretische Spitzenleistungen bei Mittelklasse-Servern	7
3.1	—DUMMY— Netzwerkdurchsatz	18
3.2	Übersicht ausgewählter Datenquellen für Spark	19
3.3	Übersicht verfügbarer Clustermanager für Spark	19
1	Theoretische Spitzenleistungen bei Mehrzweck-Servern der 2000 Euro Klasse	27

Listings

2.1	Word Count in der Spark Konsole	4
-----	---	---

1 Einführung

1.1 Motivation

Die Entwicklung und Verbesserung von Frameworks zur Verarbeitung großer Datenmengen ist zur Zeit hochaktuell und zunehmend im Fokus von Medien und Unternehmen angekommen [JON14]. Verschiedene Programme und Paradigmen konkurrieren um die schnellste, bequemste und stabilste Art großen Datenmengen einen geschäftsfördernden Nutzen abzurufen [SR14].

Mit dem Begriff „große Datenmengen“ oder „Big Data“ werden in dieser Arbeit solche Datenmengen zusammengefasst, die die Kriterien Volume, Velocity, Variety [Lan01] erfüllen oder „Datenmengen, die nicht mehr unter Auflage bestimmter Service Level Agreements auf einzelnen Maschinen verarbeitet werden können“ (Vgl. [SW14]).

Als Unternehmen, das früh mit zeitkritischen Aufgaben (u.a. Indizierung von Webseiten und PageRank [Pag01]) auf solchen Datenmengen konfrontiert war implementierte Google das Map-Reduce Paradigma [DG04] als Framework zur Ausnutzung vieler kostengünstiger Rechner für verschiedene Aufgaben.

In Folge der Veröffentlichung dieser Idee im Jahr 2004 wurde Map-Reduce in Form der OpenSource Implementation Hadoop (gemeinsam mit einer Implementation des Google File Systems GFS, u.a.) [GGL03] zum de-facto Standard für Big-Data-Analyseaufgaben.

Reines Map-Reduce (nach Art von Hadoop) als Programmierparadigma zur Verarbeitung großer Datenmengen zeigt jedoch in vielen Anwendungsfällen Schwächen:

- Daten, die in hoher Frequenz entstehen und schnell verarbeitet werden sollen erfordern häufiges Neustarten von Map-Reduce-Jobs. Die Folge ist kostspieliger Overhead durch Verwaltung/Scheduling der Jobs und gegebenenfalls wiederholtem Einlesen von Daten.
- Algorithmen die während ihrer Ausführung iterativ Zwischenergebnisse erzeugen und auf vorherige angewiesen sind (häufig bei Maschinenlernalgorithmen) können nur durch

persistentes Speichern der Daten und wiederholtes Einlesen zwischen allen Iterationsschritten implementiert werden.

- Anfragen an ein solches Map-Reduce-System erfolgen imperativ in Form von kleinen Programmen. Dieses Verfahren ist offensichtlich nicht so intuitiv und leicht erlernbar wie deklarative Abfragesprachen klassischer Datenbanken (z.B. SQL).

In der Folge dieser Probleme entstanden viele Ansätze dieses Paradigma zu ersetzen, zu ergänzen oder durch übergeordnete Ebenen und High-Level-APIs zu vereinfachen [SR14].

Eine der Alternativen zu der klassischen Map-Reduce-Komponente in Hadoop ist die „general engine for large-scale data processing“ Apache Spark.

Ein Indiz für das steigende Interesse an diesem Produkt liefert unter anderem ein Vergleich des Interesses an Hadoop und Spark auf Google:

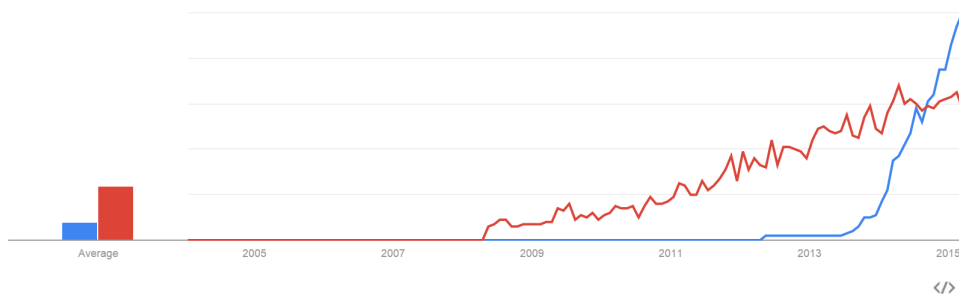


Abbildung 1.1: Suchanfragen zu „Apache Spark“ (*blau*) und „Apache Hadoop“ (*rot*), Stand 24.03.2015 [Goo]

1.2 Kontextabgrenzung

Das Ziel dieser Arbeit ist es die grundlegenden Konzepte und Möglichkeiten von Apache Spark zu untersuchen und ausgewählte Aspekte im Rahmen konkreter Anwendungen zu betrachten.

Für ein tieferes Verständnis werden Installation, Cluster-Betrieb und die Modellierung und Implementation von Treiberprogrammen beispielhaft durchgeführt, dokumentiert und bewer-

tet. Hierbei kommt Apache Spark in der Version 1.3.0 zum Einsatz.

Nur am Rande wird betrachtet:

- Der Vergleich mit ähnlichen Produkten
- Die empirische Untersuchung des Skalierungsverhaltens
- Sämtliche Features der Spark Standardbibliotheken

Apache Spark ist überwiegend in der Programmiersprache Scala geschrieben. Die Beispiele in dieser Arbeit werden ebenfalls in Scala verfasst um

1. einen einheitlichen Stil und Vergleichbarkeit zwischen Quellcode-Auszügen und eigenen Beispielen zu gewährleisten.
2. Ausdrücke in kurzer, prägnanter Form darzustellen.

2 Vorstellung von Apache Spark

Aus Sicht eines Nutzers ist Apache Spark eine API zum Zugriff auf Daten und deren Verarbeitung.

Diese API (wahlweise für die Programmiersprachen Scala, Java und Python verfügbar), kann im einfachsten Fall über eine eigene Spark Konsole mit **Read Evaluate Print Loop**[HKK99] verwendet werden.

Die Zählung von Wortvorkommen in einem Text - das „Hello World“ der Big Data Szene - lässt sich dort mit zwei Befehlen realisieren (Listing 2.1).

```
1 $ ./spark-shell
2 [...]
3   /  __/___  ____  ____/  /___
4   _\  \/_  _\/_  _ ' /  __/  '/_
5   /___/  .__/\_,_/_/_/  /_/\_\  version 1.3.0
6   /_/_
7 Using Scala version 2.10.4 (OpenJDK 64-Bit Server VM, Java 1.7.0_75)
8 Type in expressions to have them evaluated.
9 [...]
10 scala> val text = sc.textFile("../Heinrich Heine - Der Ex-Lebendige")
11 [...]
12 scala> :paste
13 text.flatMap(line => line.split(" "))
14 .map(word => (word, 1))
15 .reduceByKey(_ + _)
16 .collect()
17 [...]
18 res0: Array[(String, Int)] = Array((Tyrann,,1), (im,2), (Doch,1) ...)
```

Listing 2.1: Word Count in der Spark Konsole

Aus Sicht eines Administrators oder Softwarearchitekten ist Apache Spark eine Applikation auf einem **Rechnercluster**, die sich in der Anwendungsschicht befindet und charakteristische

Anforderungen insbesondere an Lokalität des Storages und die Netzwerkperformance stellt.

Was das konkret bedeutet, welche Mechanismen und Konzepte dahinterstehen und in welchem Ökosystem von Anwendungen sich Apache Spark bewegt wird in den folgenden Abschnitten dieses Kapitels beleuchtet.

2.1 Überblick

Im Allgemeinen Fall läuft eine Spark-Anwendung auf drei Arten von Rechnern (s. Abb. 2.1):

1. Clientknoten

Auf Nutzerseite greift die Anwendung auf die API eines lokalen Spark-Kontextes zu, der die Kontaktdaten eines Clustermanagers sowie verschiedene Konfigurationseinstellungen enthält.

2. Masterknoten

Der Masterknoten betreibt den *Clustermanager*, läuft auf einem entfernten Rechner und ist der Einstiegspunkt in den *Rechnercluster*. Hier werden Aufträge des Anwenders an die Arbeitsknoten verteilt und Ergebnisse eingesammelt und zurückgereicht.

3. Workerknoten

Die Workerknoten beherbergen die Spark *Executors* und sind die ausführenden Elemente der Aktionen und Transformationen. Die *Executors* können untereinander Zwischenergebnisse austauschen und melden ihre Ressourcenverfügbarkeit an den *Clustermanager*.

Um die Architektur und Optimierungskonzepte eines verteilten Systems bewerten zu können ist es offensichtlich wichtig, welche Eigenschaften der unterliegenden Hardware angenommen werden.

Weil Spark explizit für den Betrieb innerhalb eines Hadoop/YARN [VERWEIS auf Abschnitt Scheduling] geeignet ist und YARN wiederum für den Betrieb auf einem Rechnercluster auf Mittelklasse-Mehrzweckmaschinen (Commodity Hardware) optimiert ist[AM14], kann für Spark von einer vergleichbaren Hardwarekonfiguration ausgegangen werden.

Der Vergleich von drei aktuellen Rack Servern der 2000-Euro-Klasse (in der Grundausstattung) - hier als Mittelklasse-Geräte bezeichnet - liefert die folgenden Verhältnisse der wesentlichen Schnittstellen zueinander (Siehe Anhang 3.1).

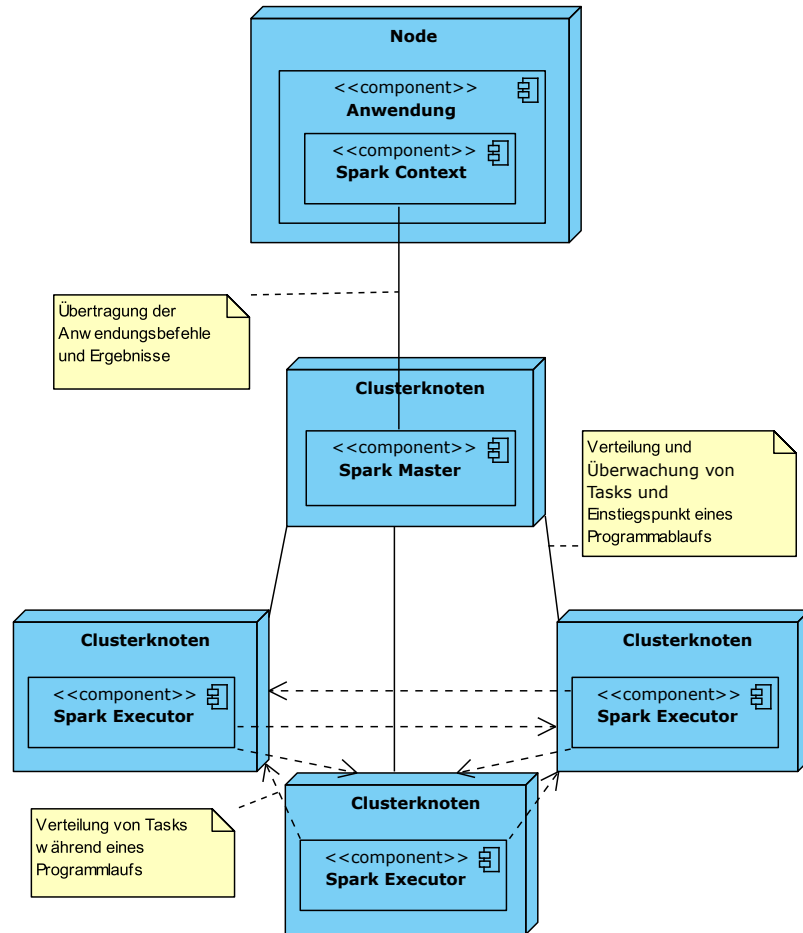


Abbildung 2.1: Verteilungsdiagramm einer typischen Spark-Installation

Auf eine detaillierte Analyse des Zugriffsverhaltens wird im Rahmen dieser Arbeit verzichtet. Bei den folgenden Bewertungen der Kernkonzepte ist es wichtig sich die aus Tabelle 2.1 abgeleiteten Größenordnungen des Durchsatzes (D) der verschiedenen Datenkanäle zu vergegenwärtigen:

$$D_{\text{Netzwerk}} < D_{\text{Festspeicher}} < D_{\text{Arbeitsspeicher}}$$

Für eine effiziente Verarbeitung von Daten ist es - ganz allgemein - also wünschenswert den größten Anteil des Datentransfers im Arbeitsspeicher zu haben, einen kleineren Anteil auf der Festplatte und einen noch kleineren Anteil auf Netzwerkverbindungen.

Netzwerk	Festspeicher	Arbeitsspeicher
0,125 GB/s	1 GB/s	17 GB/s

Tabelle 2.1: Theoretische Spitzenleistungen bei Mittelklasse-Servern

Es ist das wichtigste Ziel der folgenden Kernkonzepte von Apache Spark unter diesen Bedingungen die effiziente und stabile Verarbeitung *großer Datenmengen* [SW14] zu gewährleisten.

2.2 Kernkonzepte

2.2.1 Resilient Distributed Datasets

Die universelle Einheit mit der ein Datenelement auf Spark repräsentiert wird ist ein sogenanntes **Resilient Distributed Dataset (RDD)** [ZC+12].

Ein Beispiel für ein solches **RDD** wurde bereits erwähnt, nämlich das in Listing 2.1 erzeugte Objekt `text`:

```
1 val text = sc.textFile("../Heinrich Heine - Der Ex-Lebendige")
```

RDDs können auch explizit erzeugt werden, ohne dass dazu vorhandene Daten genutzt werden:

```
1 val list = sc.parallelize(List(1,2,3,4,5,6))
```

Die gesamte operative Kern-**Application Programming Interface (API)** dreht sich um die Steuerung dieser Dateneinheiten. Insbesondere sind auch die in den Standardbibliotheken verfügbaren „höheren“ **APIs** auf diesen **RDDs** implementiert.

Sie sind damit die wichtigste Abstraktion des Applikationskerns.

In erster Näherung können **RDDs** als eine Variante von **Distributed Shared Memory (DSM)** [NL91] [ZC+12] verstanden werden, haben allerdings sehr charakteristische Einschränkungen und Erweiterungen, die in diesem Kapitel erläutert werden.

Verteilungssicht Aus Verteilungssicht ist ein **RDD** ein Datensatz, der über den Arbeitsspeicher mehrerer Maschinen partitioniert ist (Abb. 2.2).



Abbildung 2.2: Resilient Distributed Datasets aus Verteilungssicht

Laufzeitsicht **RDDs** sind nicht veränderbar. Anstatt durch feingranulare Schreiboperationen modifiziert zu werden ist es nur möglich ein einmal definiertes **RDD** durch globale Anwendung von Operationen in ein anderes zu überführen.

Diese Folge von Operationen $op_1 op_2 op_3 \dots$ wird als *Lineage* eines **RDD** bezeichnet. Die *Lineage* kann als das „Rezept“ zur Erstellung eines Datensatzes verstanden werden.

Dabei gibt es zwei grundsätzlich verschiedene Operationen, nämlich *Transformationen* und *Aktionen*.

Transformationen sind ...

Aktionen sind ...

Solange nur *Transformationen* auf einem **RDD** ausgeführt werden, ist dieses noch ein bloßes „Rezept“ zur Erstellung eines Datensatzes. Tatsächlich wurde noch kein Speicher reserviert und der Cluster wurde noch nicht aktiv[ZC+12]:



Abbildung 2.3: RDD Lineage vor Aktion

Sobald die erste *Aktion* aufgerufen wird, werden die Transformationen nach der vorgegebenen Reihenfolge ausgeführt und die geforderte *Aktion* ausgeführt:

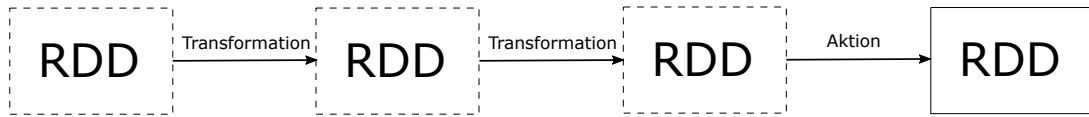


Abbildung 2.4: RDD Lineage nach Aktion

Wie in Abb. 2.4 zu erkennen, werden während der Transformationsvorgänge keine Zwischenergebnisse gespeichert. Möchte Zwischenergebnisse zu einem späteren Zeitpunkt oder in anderem Zusammenhang wiederverwenden, kann man dies explizit über das Kommando `persist()` anweisen:

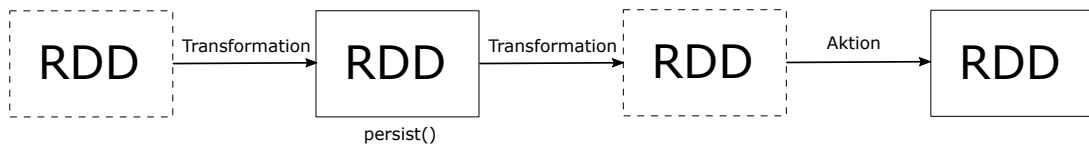


Abbildung 2.5: RDD Lineage nach Aktion und mit Persist()

Ein möglicher Vorteil dieser Art von Arbeitsdatensatz wird sofort deutlich: Im optimalen Fall sind die zu ladenden Daten von jedem der **Worker** auf unabhängigen Kanälen erreichbar (z.B. auf dem lokalen Festspeicher) und gleichmäßig auf diesen Kanälen partitioniert.

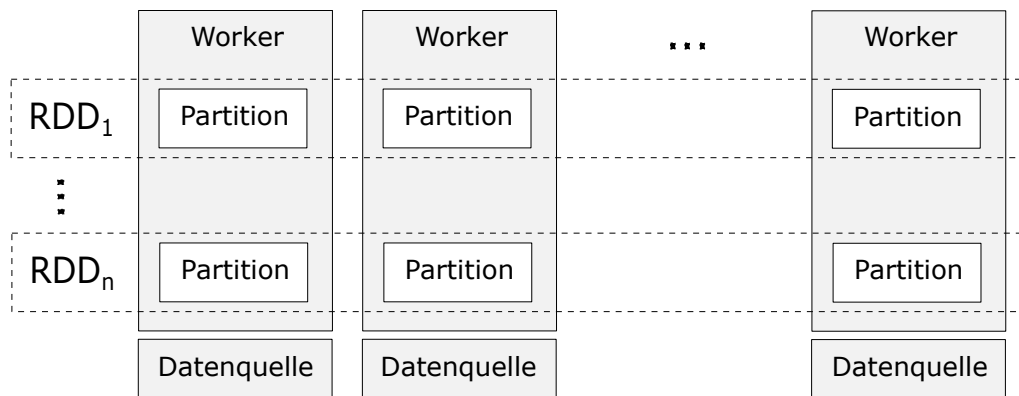


Abbildung 2.6: Resilient Distributed Datasets mit Datenquelle aus Verteilungssicht

Im diesem Fall ergäbe sich mit einer Anzahl **Worker** n und einem Durchsatz δ zu der jeweiligen Datenquelle also ein Gesamtdurchsatz beim Einlesen von Daten von:

$$\sum_{i=1}^n \delta_i \quad (2.1)$$

2.2.2 Scheduling/Shuffling

2.2.3 Anwendungsdeployment und -lebenszyklus

Anwendungsdeployment Die Komponente einer Anwendung, die von der Spark API Gebrauch macht wird in der Spark-Terminologie als *Treiberprogramm* bezeichnet.

Es gibt verschiedene Szenarien des Deployments. Das Treiberprogramm kann grundsätzlich auf zwei verschiedene Arten gestartet werden:

1. Übermittlung des Treibers als kompiliertes Package (z.B. als `jar`) mit statischer Verlinkung aller erforderlichen Bibliotheken (Ausnahmen sind Bibliotheken die auf allen Knoten bereits verfügbar sind, z.B. Spark, Hadoop, etc.). Standardbibliotheken von Spark und Konfigurationseinstellungen wie z.B. die Angabe des *Clustermanagers* können zur Startzeit des Treibers durch das Spark *Submission-Skript* erfolgen.
2. Start eines eigenständig lauffähigen Treibers mit vollständig konfigurierter und verlinkter Spark-Umgebung und expliziter Angabe eines *Clustermanagers*.

Der zweite Fall ist eher exotisch, weil eine derart enge Kopplung zwischen dem Treiber und der Konfiguration des Clusters offensichtlich aus Wartungsgründen nicht wünschenswert ist.

Im ersten Fall ergeben sich zwei weitere Möglichkeiten zum Ort der Ausführung des Treibers:

1. **Client-Modus** Der Treiber wird direkt auf dem Host (Gateway-Rechner) ausgeführt auf dem der Treiber übermittelt wurde (Abb. 2.7). Tatsächlich wird er sogar innerhalb des Submission-Skript-Prozesses gestartet (**VERWEIS**).
2. **Cluster-Modus** Der Treiber wird von dem Gateway-Rechner an **Worker** des Clusters übertragen und dort ausgeführt (Abb. 2.8).

Offensichtlich kann die Lokalität des Treibers (der mit den Executors auf den **Workern** kommunizieren muss) Einfluss auf Laufzeit und Latenzverhalten des Programms haben.

Im Fall eines clusterfernen Gateway-Rechners kann also Treiberdeployment im Clustermodus sinnvoll sein, die Standardeinstellung ist jedoch der Clientmodus (siehe auch [Spa]).

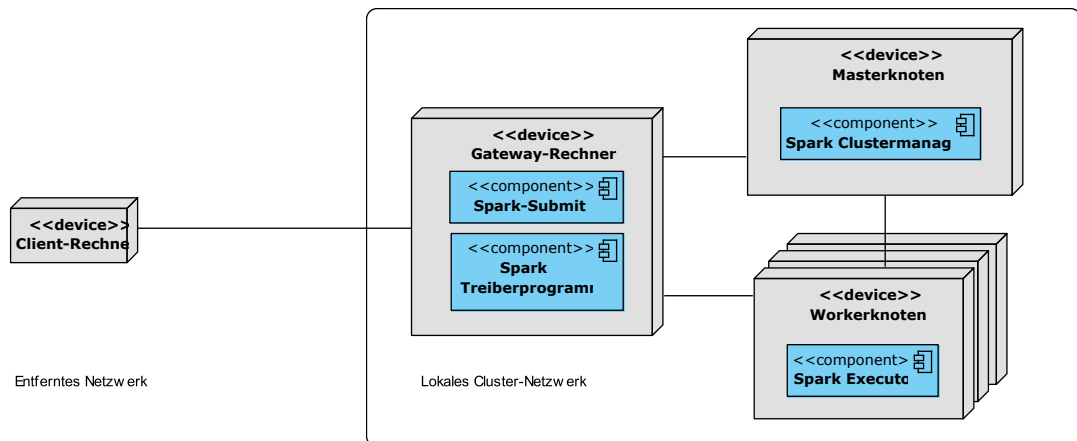


Abbildung 2.7: Application Deployment im Client Modus

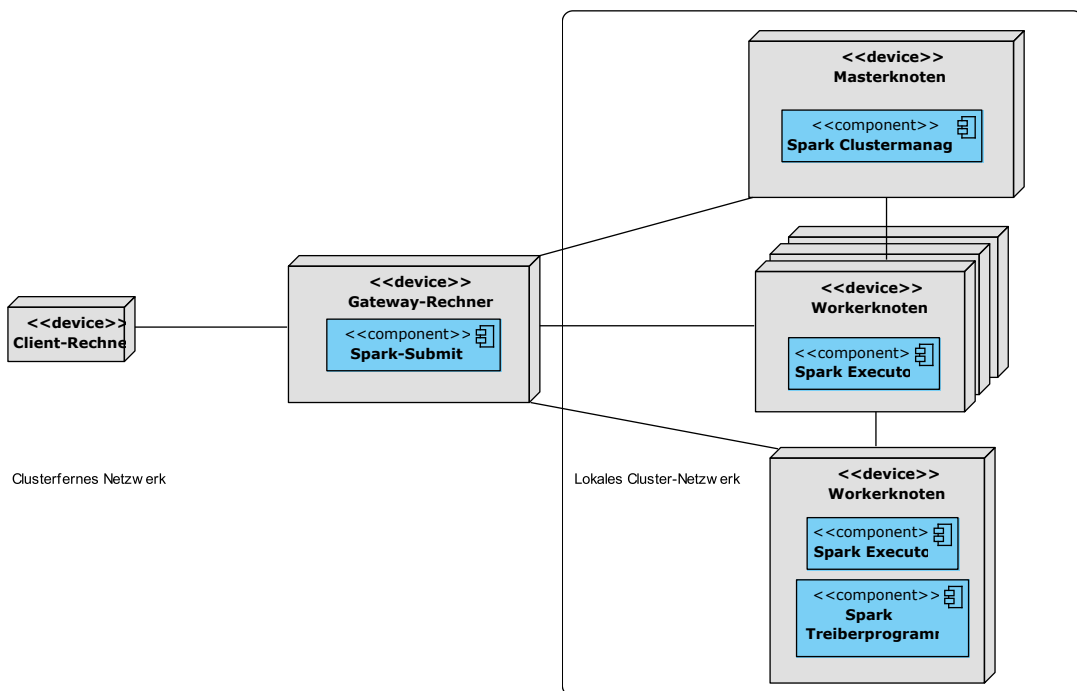


Abbildung 2.8: Application Deployment im Cluster Modus

Abb. 2.9 zeigt das Sequenzdiagramm eines typischen Deploymentprozesses, wie er auch im praktischen Teil dieser Arbeit zum Einsatz kommt.

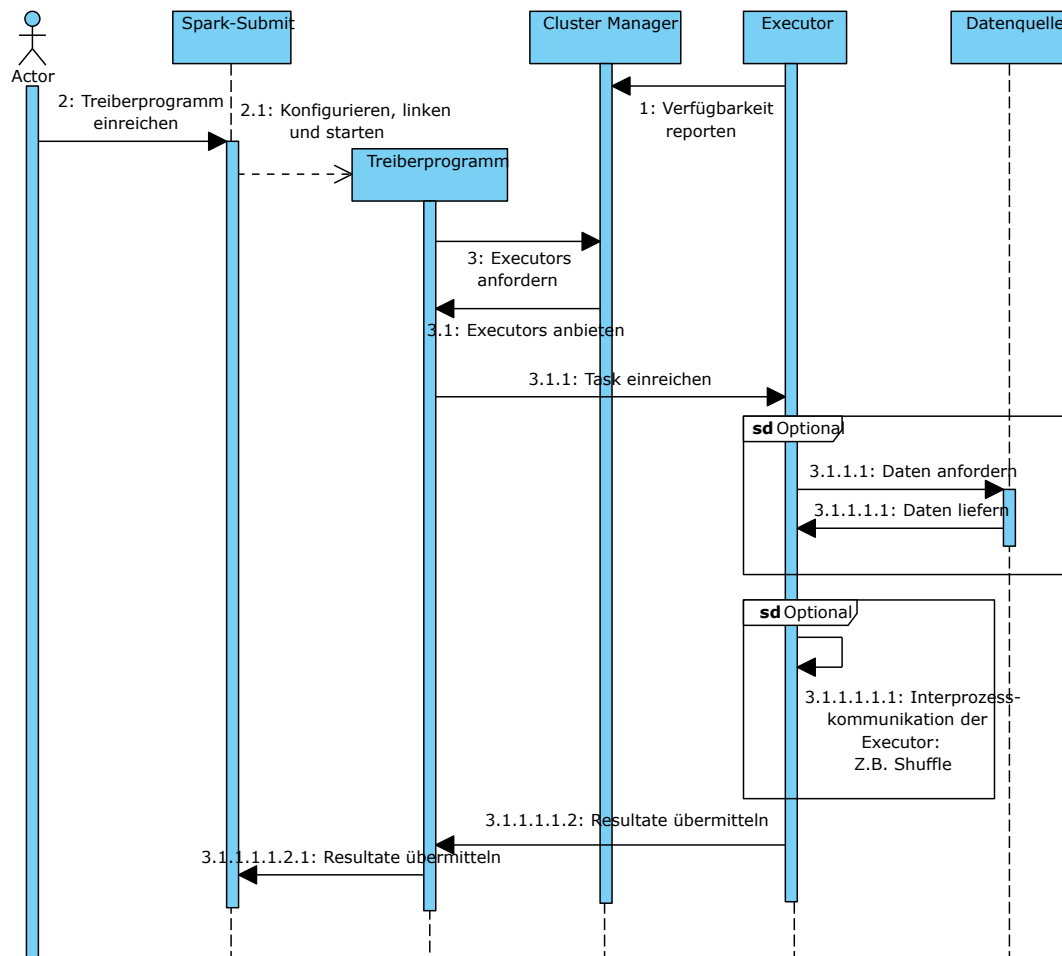


Abbildung 2.9: Application Deployment Prozess im Client Modus (vereinfacht)

Lebenszyklus einer Spark-Applikation Beginn und End eines Treiberprogramms. Beginn und Ende eines Tasks / Executorallokation?

2.2.4 Zusammenfassung und Bewertung

2.3 Standardbibliotheken

— Warum ist Spark so einfach (und wo vielleicht nicht)? —

Die vier Standardbibliotheken erweitern die Kern-API für bestimmte, häufig genutzte Aufgaben aus Bereichen der Datenanalyse.

Die bedienten Bereiche

- Deklaratives Abfragen auf strukturierten Datensätzen (*Spark SQL*)
- Maschinenlernverfahren (*MLlib*)
- Echtzeitbehandlung von eingehenden Daten (*Streaming*)
- Operationen auf Graph-Strukturen (*GraphX*)

werden in diesem Abschnitt erläutert.

2.3.1 Dataframes/Spark SQL

2.3.2 MLlib

2.3.3 Streaming

2.3.4 GraphX

2.4 Betrieb und Security

2.5 Spark im Kontext von Parallelisierungspattern

— Buch: Algorithms and Parallel Computing —

2.6 Entwicklergemeinschaft

— Herkunft, Apache Foundation, Entwicklungsphilosophien, Anzahl Entwickler, ... —

Apache Spark begann als Entwicklung einer Gruppe von Forschern der University of California, Berkely. Spark ist eine Implementation der von dieser Gruppe untersuchten **RDDs**[**ZC+12**]. Als wesentlicher Meilenstein der Entwicklung von Apache Spark, kann die Veröffentlichung eines gemeinsamen Papers der Forschungsgruppe um Matei Zaharia im Jahr 2012 gelten.

Seit dem 27. Februar 2014[**apa**] ist Spark ein Top-Level Projekt der Apache Software Foundation[**Apaa**] und wird dort unter der Apache License 2.0[**Apab**] weiterentwickelt.

Eine Übersicht der verantwortlichen Committer kann unter [Com] eingesehen werden. Zum Zeitpunkt dieser Arbeit gehören u.a. Entwickler von Intel, Yahoo! und Alibaba zu den Stammentwicklern.

Die Kommunikation innerhalb der Entwickler- und Anwendergemeinschaft findet wesentlich in den offiziellen Mailinglisten (Abb. 2.10) und dem Issue-Tracker[Iss] der Apache Software Foundation statt.

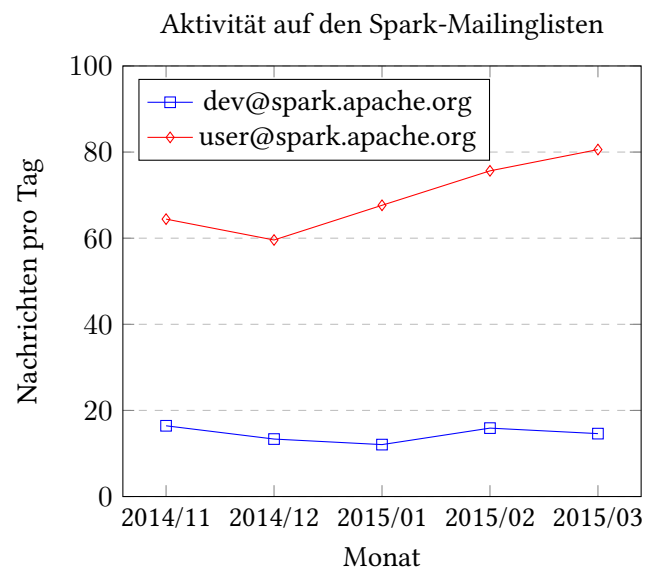


Abbildung 2.10: Aktivität auf den offiziellen Spark Mailinglisten

2.7 Verwandte Produkte

— Ergänzende oder konkurrierende Produkte —

2.7.1 YARN

2.7.2 Mesos

2.7.3 Flink

2.7.4 MPI

2.7.5 Kafka

2.7.6 HBase

2.7.7 Lustre

3 Beispielanwendungen von Spark zur Datenanalyse

Im Folgenden wird Apache Spark im Rahmen zweier grundsätzlich verschiedener Anwendungsfälle betrachtet.

Beispiel 1: Eine typische Anwendung mit verteiltem lokalem Storage (HDFS) und Spark als „Client“ eines bestehenden Yarn Clustermanagers. – Commodity Hardware (Raspberry Pi Cluster). –

Beispiel 2: Eine untypische Anwendung mit verteiltem entfernten Storage und dem Spark Standalone Clustermanager. – HPC Hardware („Thunder“ des Hamburger KlimaCampus). –

3.1 Anwendung zur Echtzeitanalyse eines Datenstroms anhand eines veränderlichen Stammdatenbestandes

Es sollen aktuelle Informationen bereitgestellt werden, die besonders für Nutzer und Entwickler des Softwareproduktes Apache Spark relevant sind.

Dazu werden die beiden Mailinglisten `user@spark.apache.org` und `dev@spark.apache.org` als Textquellen herangezogen. Aus diesen Quellen sollen Informationen über aktuelle Themen innerhalb der Nutzer- und Entwicklergemeinschaft gewonnen werden.

Mit diesen Informationen über aktuelle Themen, sollen dann in einem Echtzeitverfahren Nachrichten aus dem Twitter-Datenstrom nach der gemessenen Relevanz bewertet werden.

3.1.1 Anforderungen

Für die Software soll folgende lose Sammlung funktionaler und nicht-funktionaler Anforderungen gelten. Mit *Information* ist jeweils eine Auflistung von Twitter-Meldungen gemeint.

- **A1: Zugriff auf die Information**

Der Zugriff soll über eine grafische Benutzerschnittstelle erfolgen und keine Konfigurationen benötigen.

- **A2: Aktualität der Information**

Es sollen stets Informationen dargestellt werden, die unmittelbar zuvor entstanden sind und in Quasi-Echtzeit¹ verarbeitet wurden.

- **A3: Relevanz der Information**

Die Relevanz soll an aktuellen Themen der Entwicklergemeinschaft gemessen werden.

3.1.2 Hardwarekontext und Performance-Basisdaten

Als Testumgebung dient ein **Rechnercluster** aus vier identischen **Workern** und einem speziellen **Masterknoten** (Abb. 3.1).

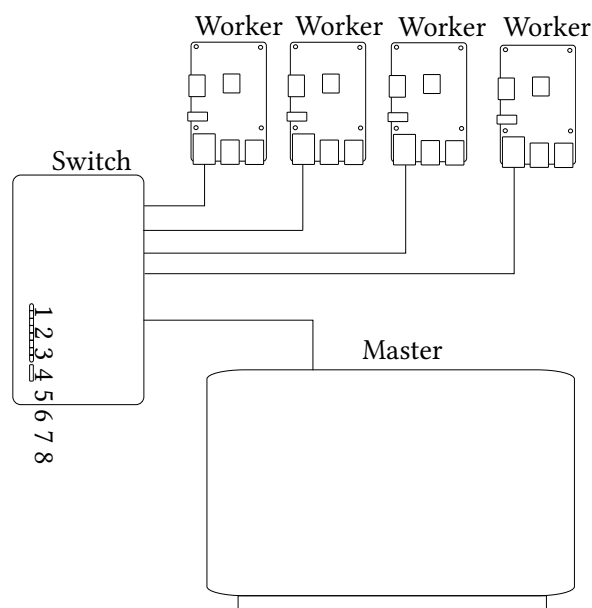


Abbildung 3.1: Hardwareumgebung des Programms zur Tweetanalyse

¹Eine Latenz von unter einer Minute sei hier tolerabel

Worker Raspberry Pi 2

- CPU: 900MHz Quad-Core ARM Cortex A7
- RAM: 1GB SDRAM
- Ethernet: 100MBit/s
- Festspeicher: SDHC Class 4 Speicherkarte 16GB

Als Betriebssystem kommt das Debian-Derivat Raspbian[Ras] 32-Bit zum Einsatz.

Master Dell d420

- CPU: 1,2 GHz Core2 Duo U2500
- RAM: 2GB DDR2 SDRAM
- Ethernet: 100MBit/s
- Festspeicher: 60GB 4200RPM Hard Drive

Als Betriebssystem kommt Ubuntu[Ubu] 14.04 32-Bit zum Einsatz.

Netzwerk Vernetzt sind die Rechner mit RJ45 über einen TP-Link TL-SF1008D Switch mit maximalem Durchsatz von 100MBit/s.

Tabelle 3.1: –DUMMY– Netzwerkdurchsatz

Nachrichtengröße	Worker → Worker	Master → Worker	Worker → Master
1kB	50ms	837ms	970ms
64kB	47ms	877ms	230ms
1MB	31ms	25ms	415ms
64MB	35ms	144ms	2356ms

3.1.3 Lösungsskizze

Wahl des Dateisystems Als Quelle der persistenten Daten (Nachrichtenkorpus der Mailinglisten) kommen verschiedene Technologien in Frage:

Für diesen Versuch wird HDFS zum Verwalten der Textdatei gewählt. Für das Einlesen des Textkorpus wird keine Echtzeitfunktionalität benötigt. Weil an den Textdateien nichts geändert

Tabelle 3.2: Übersicht ausgewählter Datenquellen für Spark

Name	Typ	Beschreibung
Cassandra	Datenbank	...
HBase	Datenbank	...
HDFS	Verteiltes Dateisystem	...
Kafka	??	...

wird ist Versionierung ebenso unnötig wie Verknüpfungen zu anderen Datensätzen.

Die „In-Memory“-Funktionen (VERWEIS Glossar?) anderer Systeme, sind hier eher hinderlich, weil der lokale Arbeitsspeicher der Arbeitsknoten im Versuchsaufbau stark beschränkt ist und eine leicht erhöhte Latenz beim Einlesen in Kauf genommen werden kann. Das Erstellen des Feature Vektors (VERWEIS) ist nicht zeitkritisch für die Echtzeitkomponente.

HDFS als Komponente von Apache Hadoop ist in den Versionen 2.x deutlich bezüglich der Verfügbarkeit und Skalierbarkeit verbessert worden (VERWEIS), allerdings auch aufwändiger zu installieren. Für den Zweck dieses Versuchs wird Version 1.2.1 gewählt.

- item
- item

Tabelle 3.3: Übersicht verfügbarer Clustermanager für Spark

Name	Typ	Beschreibung
Standalone	Apache Spark	...
Mesos	Datenbank	...
YARN	Apache Hadoop 2.x	...

Wahl des Cluster-Managers Spark läuft in diesem Versuch als alleinige Computeanwendung auf dem Cluster. Es ist also nicht nötig Konkurrenz um Ressourcen zu berücksichtigen. Für diesen Versuch wird daher der Standalone Clustermanager gewählt.

Verteilungsdiagramm

Komponentendiagramm

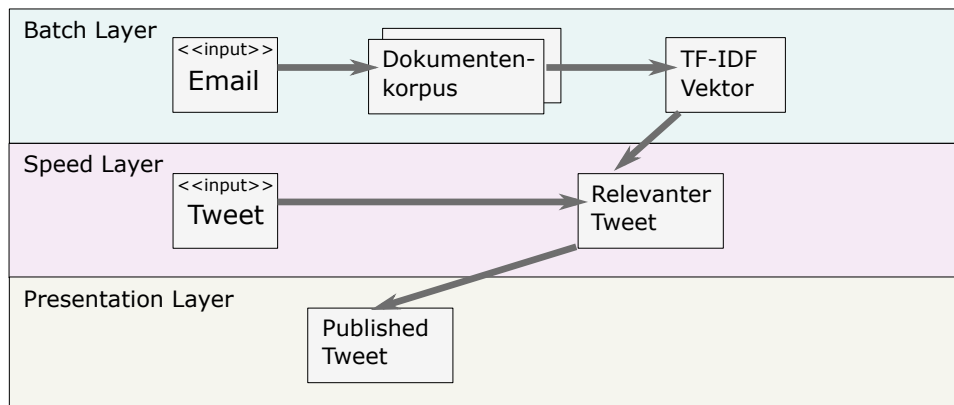


Abbildung 3.2: Datenzentrierte Sicht auf die Komponenten

Laufzeitdiagramme

3.1.4 Ergebnisse und Bewertung

Dashboard

Laufzeitverhalten

Bewertung und Probleme

3.2 Spark-basierte Implementation von Operatoren aus der Klimaforschung

— Implementation ausgewählter CDOs (sehr wenige, möglicherweise nur 1-2) mit der Core-API von Spark. Testlauf auf einem HPC Cluster mit nicht-lokalem, allerdings per Infiniband angeschlossenen Storage. Insbesondere Betrachtung des Skalierungsverhaltens und der „Sinnhaftigkeit“. —

3.2.1 Beschreibung des Problems

— Erläuterung von CDOs (Climate Data Operators). —

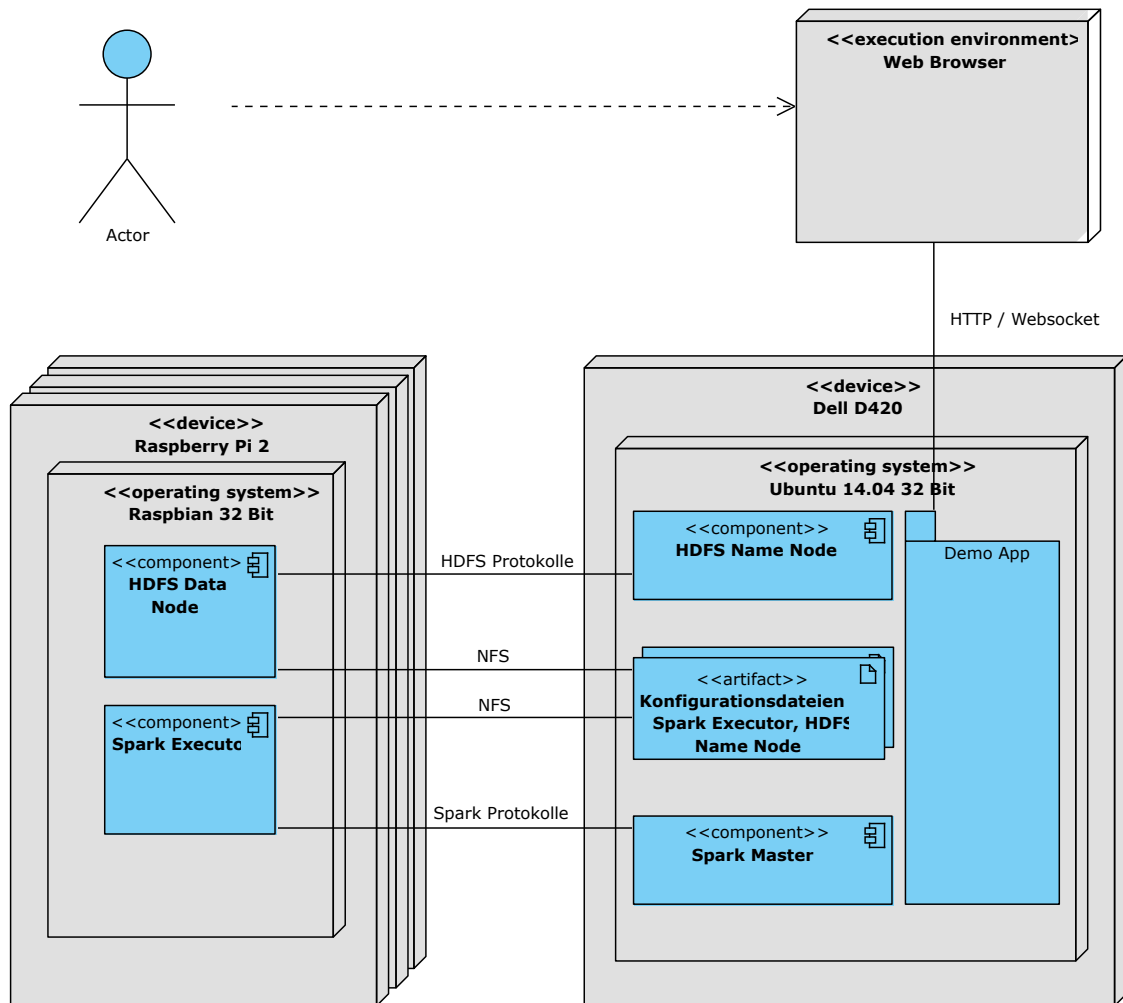


Abbildung 3.3: Verteilungssicht auf die Demo App

3.2.2 Hardwarekontext und Performance-Basisdaten

— hier kommen die eingesetzten Systeme, und relevante Laufzeitmessungen (netzwerk, storage, cpu) hin —

3.2.3 Architekturübersicht

— hier kommen Verteilungs- und Komponentendiagramm hin —

3.2.4 Detaillierte Lösungsbeschreibung

— hier kommen laufzeitdiagramme und codeschnipsel hin —

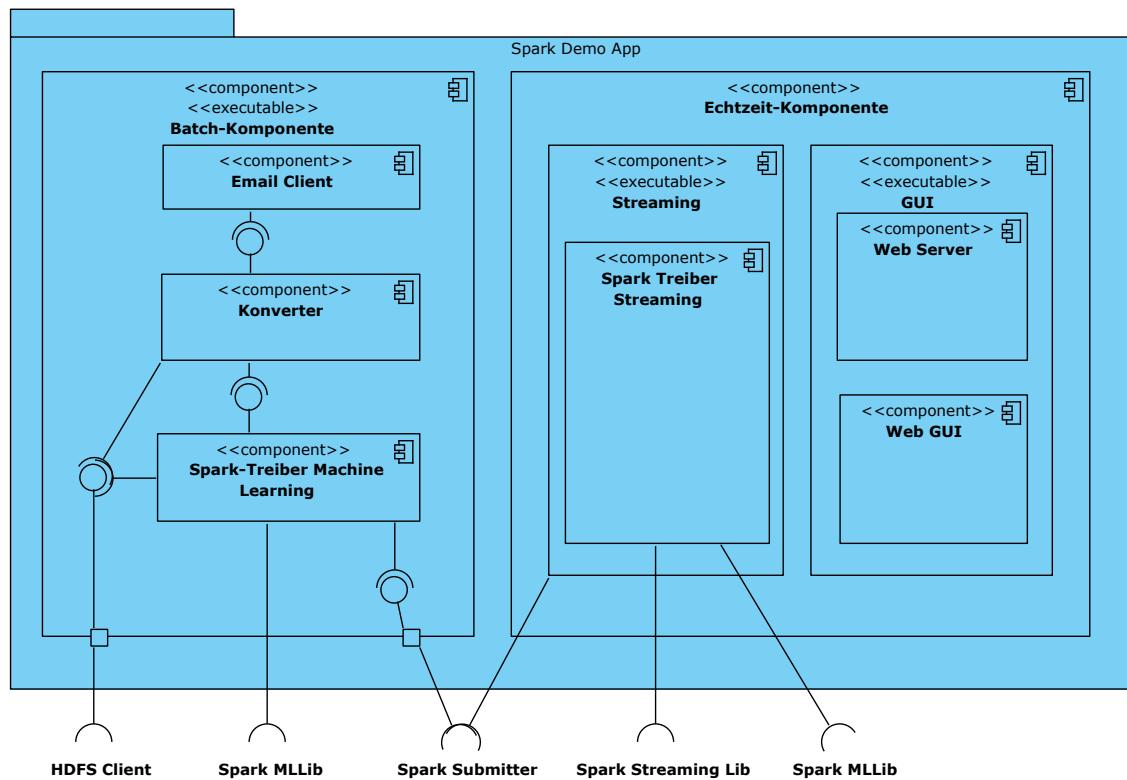


Abbildung 3.4: Komponentendiagramm des Demo App Packages

3.2.5 Ergebnisse

— Tabellen und Diagramme Ergebnissen, evt. Skalierungsverhalten — — Bewertung —

4 Schlussbetrachtung

4.1 Kritische Betrachtung der Ergebnisse

4.2 Ausblick und offene Punkte

Acronyme

API Application Programming Interface. 7

DSM Distributed Shared Memory. 7

RDD Resilient Distributed Dataset. 7, 8

Glossar

Master Host, der Verwaltungsaufgaben innerhalb eines Rechnerclusters übernimmt und dazu mit hierarchisch untergeordneten Rechnern kommuniziert. Zu den Aufgaben kann insbesondere das Verteilen von Arbeitsaufträgen oder Speicherblocks gehören. [5](#), [12](#)

Read Evaluate Print Loop Pattern zum Erzeugen einer Konsole, die in einer Endlosschleife Eingaben liest, die auswertet und das Ergebnis wieder ausgibt. [4](#)

Rechnercluster Vernetzter Verbund aus vollwertigen Rechnern. [5](#), [12](#)

RJ45 Achtpolige Modularsteckverbindung zur Datenübertragung. [13](#)

Service Level Agreement Übereinkunft zwischen dem Anbieter und dem Nutzer eines Dienstes über dessen Qualität (z.B. Antwortzeiten, Durchsatz, Verfügbarkeit, etc.). [1](#)

Worker Host, der als Arbeitsknoten in einem Rechnercluster dient. Falls nicht anders beschrieben ist hier ein Rechner gemeint, der seine Ressourcen einer Spark-Anwendung zur Verfügung stellt und mit seinem Festspeicher Teil eines verteilten Dateisystems ist. [5](#), [8](#), [12](#)

Anhang

1 Installation der Plattform

2 Quellcode (Auszüge)

3 Sonstiges

3.1 Einschätzung des theoretischen Spitzendurchsatzes von Mittelklasse-Servern

Um zu einer groben Einschätzung des möglichen Datendurchsatzes verschiedener Schnittstellen bei „Commodity Servern“ zu gelangen, wurden drei Systeme von großen Herstellern ausgewählt.

In der Grundkonfiguration kosten diese Systeme (zum Zeitpunkt dieser Arbeit) um die € 2000,- und lassen damit auf die Größenordnungen bei dem Datendurchsatz bestimmter Schnittstellen bei preisgünstigen Mehrzweck-Rechenknoten schließen.

Modell	Netzwerkschnittstelle	Interner Bus für Festspeicher	Arbeitsspeicher
Dell PowerEdge R530	1Gb/s Ethernet	PCIe 3.0	DDR4
HP Proliant DL160 Gen8	1Gb/s Ethernet	PCIe 3.0	DDR3
System x3650 M5	1Gb/s Ethernet	PCIe 3.0	DDR4

Tabelle 1: Theoretische Spitzenleistungen bei Mehrzweck-Servern der 2000 Euro Klasse

Mit [Law14] und [Fuj] lassen sich grobe obere Abschätzungen errechnen, die in Tabelle 2.1 angegeben sind.

Literatur

- [AM14] Vinod Kumar Vavilapalli Arun Merthy. *Apache Hadoop YARN*. 2014, S. 42.
- [apa] apache. *Apache Blog*. Abgerufen am 11.04.2015. URL: https://blogs.apache.org/foundation/entry/the_apache_software_foundation_announces50.
- [Apaa] Apache. *Apache Software Foundation*. URL: <http://apache.org>.
- [Apab] *Apache License, Version 2.0*. 2004. URL: <https://www.apache.org/licenses/LICENSE-2.0>.
- [Com] *Apache Spark Confluence*. Abgerufen am 11.04.2015. URL: <https://cwiki.apache.org/confluence/display/SPARK/Committers>.
- [DG04] Jeffrey Dean und Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". In: *OSDI* (2004).
- [Fuj] *Fujitsu PRIMERGY SERVER - Basics of Disk I/O Performance*. 2011. URL: <http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-basics-of-disk-io-performance-wv-en.pdf>.
- [GGL03] Sanjay Ghemawat, Howard Gobioff und Shun-Tak Leung. *The Google File System*. Techn. Ber. Google, 2003.
- [Goo] Google. *Google Trends*. URL: <https://www.google.com/trends>.
- [HKK99] Max Hailperin, Barbara Kaiser und Karl Knight. *Concrete Abstractions: An Introduction to Computer Science Using Scheme*. 1999, 278ff.
- [Iss] *Apache Spark Issue Tracker*. Abgerufen am 11.04.2015. URL: <https://issues.apache.org/jira/browse/SPARK/?selectedTab=com.atlassian.jira.jira-projects-plugin:summary-panel>.
- [JON14] KAREN SPARCK JONES. *Big-Data-Technologien – Wissen für Entscheider*. 1914. URL: http://www.bitkom.org/files/documents/BITKOM_Leitfaden_Big-Data-Technologien-Wissen_fuer_Entscheider_Febr_2014.pdf.

- [Lan01] Doug Laney. “3D Data Management: Controlling Data Volume, Velocity and Variety”. In: *Application Delivery Strategies* (2001).
- [Law14] Jason Lawley. *Understanding Performance of PCI Express Systems*. 2014.
- [NL91] Bill Nitzberg und Virginia Lo. “Distributed Shared Memory: A Survey of Issues and Algorithms”. In: *Computer* 24.8 (Aug. 1991), S. 52–60. ISSN: 0018-9162. DOI: [10.1109/2.84877](https://doi.org/10.1109/2.84877). URL: <http://dx.doi.org/10.1109/2.84877>.
- [Pag01] Lawrence Page. *Method for node ranking in a linked database*. US Patent 6,285,999. 2001. URL: <http://www.google.com/patents/US6285999>.
- [Ras] Raspbian. *Raspbian Operating System*. URL: <http://www.raspbian.org>.
- [Spa] *Spark Submission Guide*. abgerufen am 12.04.2015. URL: <https://spark.apache.org/docs/1.3.0/submitting-applications.html>.
- [SR14] Dilpreet Singh und Chandan Reddy. “A survey on platforms for big data analytics”. In: *Journal of Big Data* (2014).
- [SW14] Jasson Venner Sameer Wadkar Madhu Siddalingaiah. *Pro Apache Hadoop*. 2014, S. 1.
- [Ubu] Ubuntu. *Ubuntu Operating System*. URL: <http://www.ubuntu.com>.
- [ZC+12] Matei Zaharia, Mosharaf Chowdhury u. a. “Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing”. In: *NSDI* (2012).

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
