

Big Data Processing mit Apache Spark

vorgelegt von

Sascha P. Lorenz

Matrikel-Nr.: 501 63 21

dem Fachbereich Technik
der Hochschule Emden-Leer
und der Beuth Hochschule für Technik Berlin
vorgelegte Masterarbeit
zur Erlangung des akademischen Grades
Master of Science (M.Sc.)
im Studiengang
Medieninformatik (Master)

Tag der Abgabe 07. März 2015

- | | | |
|--------------------|----------------------------|-------------------------------------|
| 1. Betreuer | Prof. Dr. Stefan Edlich | Beuth Hochschule für Technik Berlin |
| 2. Betreuer | Prof. Dr. Schiemann-Lillie | Hochschule Emden-Leer |
-

Kurzfassung

Gegenstand dieser Arbeit sind die Grundlagen der Verarbeitung und Analyse großer Datenmengen (Big Data) am konkreten Beispiel von Apache Spark. Zunächst sollen verschiedene Ansätze mit Ihren Funktionsweisen sowie den Vor- und Nachteilen diskutiert werden. Hier werden zuerst allgemeine Grundlagen zu Big Data erarbeitet. Was ist Big Data, was unterscheidet die Verarbeitung von strukturierten und unstrukturierten Daten, Relationale Datenbanken vs. noSQL, wie müssen die Quelldaten für die jeweiligen Verarbeitungen beschaffen sein, welche besonderen Herausforderungen stellen gestreamte Daten an die Verarbeitung. Besonders wird hier auf Hadoop und den Map/Reduce-Algorithmus eingegangen, um das bisher etablierte Vorgehen zu beschreiben und ein grundsätzliches Verständnis für die Domäne "Big Data Processing" zu schaffen. In diesem Kontext wird das gesamte Ökosystem rund um Hadoop vorgestellt.

Nachdem eine Einführung in das Thema "Big Data Processing" erfolgt ist und ein entsprechend quantitativ und qualitativ brauchbarer Datensatz zur Verfügung steht, werden die Next-Generation Data-Processing Technologien betrachtet. Kernthema ist hier Apache Spark und der gesamte BDAS (Berkeley Data Analytics Stack), der von den den AMP-Labs innerhalb von Apache-Projekten um Spark herum aufgebaut wurde. Zu praktisch jeder "offiziellen" BDAS-Implementierung existieren noch Alternativen. Besonders Apache flink wird hier als Alternative näher untersucht. Auch Applikationen, die auf dem eigentlichen Stack aufsetzen, werden näher betrachtet und entsprechenden Praxistests unterzogen (beispielsweise H2O für statistische Analysen).

Danach wird die API von Spark und deren Möglichkeiten mit Scala, Java und Clojure näher betrachtet und durch jeweils eigene Implementierungen untersucht.

Die Arbeit schließt mit durch verschiedene Versuchsreihen fundierte Empfehlungen für die unterschiedlichen Anforderungen im Bereich des Big-Data-Processing.

Abstract

Inhaltsverzeichnis

1	Einführung	1
1.1	Was versteht man unter Big Data?	1
1.2	Ansätze für Big Data Analytics	2
1.3	Motivation für Apache Hadoop/Spark	4
1.4	Ziel und Aufbau dieser Arbeit	5
2	Allgemeine Grundlagen	7
2.1	Cluster Computing	7
2.2	Anwendungen für Big Data Analytics	8
2.3	Machine Learning	10
2.4	Das MapReduce-Paradigma	19
2.5	Streaming Frameworks	20
2.6	Anwendungen von Graphen	20
3	Der Berkeley Data Analytics Stack (BDAS)	21
3.1	Die Schichten des BDAS	21
3.2	Apache Mesos	23
3.3	Hadoop Distributed File System (HDFS) und Tachyon	24
3.4	Apache Spark	25
3.5	Spark Streaming	25
3.6	GraphX	26

3.7	MLbase/MMLib	26
3.8	Spark SQL	27
4	Alternative Implementierungen der Bibliotheken und Frameworks des BDAS	29
4.1	Alternative zu Spark: Apache Flink	29
4.2	Alternative zu Spark Streaming: Storm	29
4.3	Alternative zu MLLibs: H2O - Sparkling Water	30
4.4	Alternative zu MLLibs: Dato GraphLab Create™	30
5	Funktionsweise von Spark	31
5.1	Spark im Cluster	31
5.2	Das Konzept der Resilient Distributed Datasets	33
5.3	In-Memory-Computing mit Spark	33
6	Exemplarischer Architekturaufbau einer Apache Spark Infrastruktur	35
6.1	Ausführungscontainer: Docker	35
6.2	Cluster Management: Mesos und Yarn	35
6.3	Caching-Framework: Tachyon	35
6.4	Der eigentliche Kern: Apache Spark	35
6.5	Streaming-Framework: Spark Streaming	35
6.6	Abfrageschicht: Spark SQL	36
6.7	Machine Learning Algorithmen: MMLib	36
6.8	Graphenanwendungen: GraphX	36
7	Implementierung der Prototypen	37
7.1	Prototyp: Spark	37
7.2	Prototyp: MMLib	37
7.3	Prototyp: Spark Streaming	37
7.4	Prototyp: GraphX	38

8	Evaluierung der Komponenten und Alternativen	39
8.1	Definition von Metriken für die Bibliotheken des BDAS	39
8.2	Beschreibung der Messverfahren	39
8.3	Beschreibung der Messumgebungen	39
8.4	Ergebnisse	40
9	Schlussbetrachtung	41
9.1	Zusammenfassung	41
9.2	Ausblick	41
10	Verzeichnisse	43
	Literaturverzeichnis	46
	Internetquellen	48
	Abbildungsverzeichnis	49
	Tabellenverzeichnis	51
	Quellenverzeichnis	53
A	Zusätze	55
A.1	Quelltext	55

Kapitel 1

Einführung

Big Data ist insbesondere in den letzten Jahren immer stärker in den verschiedensten Zusammenhängen in den allgemeinen Sprachgebrauch vorgedrungen und ist hier einem ständigen Bedeutungswandel ausgesetzt. Besonders in letzter Zeit wird dieses Thema auch verstärkt kontrovers diskutiert.

Im ersten Kapitel soll der Begriff *Big Data* jenseits von Management-Hype und Skepsis rational definiert werden. Des Weiteren werden einige grundlegende Konzepte des Umgangs mit sehr großen und unstrukturierten Datensätzen diskutiert und im Speziellen die Motivation hinter den Apache Frameworks Hadoop und Spark vorgestellt.

Das zweite Kapitel beschäftigt sich mit dem Berkeley Data Analytics Stack (BDAS), mit dem von der UCLA Berkeley rund um Hadoop ein leistungsfähiger Infrastruktur-Stack für die Einsatzbereiche von Big Data Analytics geschaffen wurde.

Innerhalb vom BDAS etabliert sich langsam auch eine schnellere und flexiblere Alternative zu Hadoop: Apache Spark. Im dritten Kapitel wird diese neue Kerntechnologie vorgestellt, die zugleich auch den Hauptteil dieses Wissenschaftlichen Projektes darstellt.

Im vierten Teil dieser Ausarbeitung wird Spark in der praktischen Anwendung gezeigt inklusive Installation und ersten kleineren Beispielen sowohl direkt in Spark, als auch aus darüber liegenden Schichten aus dem Stack.

1.1 Was versteht man unter Big Data?

Der Begriff *Big Data* wurde vermutlich zum ersten Mal Ende des 20. Jahrhunderts von John R. Marshey, damals Chefwissenschaftler bei Silicon Graphics, im Rahmen einer Usenix-Konferenz öffentlich erwähnt [Tur10]. Mittlerweile zielt dieser Begriff gefühlt jedes zweite Cover von IT-Zeitschriften mit Business-Fokus und auch Manager und *Sales-Professionals* werten Ihre Produktpräsentationen gerne mit diesem Buzzword auf. Aber dieser Begriff ist nicht nur positiv

assoziiert. Besonders seit Bekanntwerden der Tätigkeiten des Amerikanischen Auslandsgeheimdienstes weckt die Vorstellung des Datensammelns in großen Dimensionen auch Misstrauen.

Im Rahmen dieser Arbeit soll jedoch ausschließlich die technische Betrachtung und die exemplarische Darstellung von möglichen Anwendungsgebieten diskutiert werden.

Wie lässt sich der Begriff *Big Data* abgrenzen? Es existiert keine abschließend eindeutige Definition, jedoch gibt es einige Attribute, die sich in einem Großteil der Fachliteratur etabliert haben. Der Artikel aus dem O'Reilly Radar zum Thema [Dum11] fasst dies folgendermaßen zusammen:

„Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures.“

Neben der reinen Menge spielt also offensichtlich auch die mangelnde oder fehlende Strukturierung und unter Umständen die Flüchtigkeit der Daten eine nicht unerhebliche Rolle. Dies können beispielsweise Daten aus Social-Media-Quellen sein, die aus allen möglichen verschiedenen Einzeldaten bestehen, Daten von Sensoren, die permanent überwacht werden müssen, oder Datenströme (Video, Audio, Bilder, Text), die nach einheitlichen Kriterien gefiltert werden sollen, um hier nur einige Beispiele zu nennen. Auch die temporäre Komponente ist ein Einsatzgebiet für *Big Data*, und auch hier ist wieder das Beispiel der Datenströme heranzuziehen.

Bei der Definition von *Big Data* werden laut des BITKOM-Ratgebers zum Thema *Big Data* [BIT14] auch immer wieder die „Three Vs“ angeführt. Dies sind *Volume*, also die Datenmenge, *Variety*, die Datenvielfalt und *Velocity*, die Geschwindigkeit der Auswertung.

Die sinnvolle Analyse dieser Daten kann Unternehmen oder anderen Organisationen wichtige Informationen z.B. über Marktentwicklungen, bestimmte Kundenbedürfnisse, Epidemie-Ausbreitungen oder andere wichtige Sachverhalte liefern. Diese Analyse inklusive der dazu verwendeten Werkzeuge wird allgemein *Big Data Analytics* genannt.

1.2 Ansätze für Big Data Analytics

Die Disziplin *Big Data Analytics* umfasst Methoden und Werkzeuge zur automatisierten oder interaktiven Erkennung und daraufhin auch Verwendung von bestimmten Mustern und Assoziationen. Dies sind unter anderem:

- Prediction-Models zur Vorhersage bestimmter Sachverhalte
- statistische Verfahren, wie beispielsweise *Logistic Regression* oder *k-means-Algorithmen*
- Optimierungs- und Filteralgorithmen

- Werkzeuge zum Datamining
- Textanalyse
- Bild- und Tonanalyse
- Datenstromanalysen

Nach dem BITKOM-Leitfaden [BIT14] besteht die Taxonomie der Big-Data-Technologien grundsätzlich aus vier Schichten:

- Daten-Haltung
- Daten-Zugriff
- Analytische Verarbeitung
- Visualisierung

Diese werden durch *Daten-Integration* und *Daten-Governance*, sowie Daten-Sicherheit flankiert, um den Weg von Rohdaten bis zu nutzbaren Erkenntnissen in existierende Standards einzubetten.

Zahlreiche Hersteller herkömmlicher relationaler Datenbanksysteme versuchen derzeit, ihre bestehenden Lösungen mit dem Label *Big Data* zu versehen und diese so weiterhin in diesen sich verändernden Marktsegmenten zu positionieren. Wenn *Big Data* jedoch jenseits der Datengröße definiert wird und auch unstrukturierte und temporäre Daten-Stacks oder –ströme zu verarbeiten oder zu analysieren sind, stoßen RDBMS ¹ sehr schnell an ihre Grenzen. Doch auch was die Skalierbarkeit angeht, sind relationale Datenbanken meist nicht hinreichend flexibel [Lou10].

Für die Anforderungen an dedizierte Aufgaben im Bereich *Big-Data-Analytics* sind seit einigen Jahren einige *Frameworks* auf dem Markt, die in allen drei oben genannten Aspekten besser geeignet sind, als RDBMS. Der Ansatz ist hier primär, die Verarbeitung zu dezentralisieren, also auf unabhängige Knoten in einem Rechner-Cluster zu verteilen und nur Referenzen auf die Clusterknoten zentral zu verwalten.

Es existieren mittlerweile Lösungen am Markt, die speziell diese Aufgaben für derartige Aufgaben entwickelt wurden. Hier wären unter anderem Hadoop, Spark, HPCC, GPMR, Mincmeat, Sphere, Bashreduce und R3 zu nennen. Bis auf HPCC setzen alle eben genannten Implementierungen generell oder in Teilen auf das Programmiermodell MapReduce.

¹ RDBMS = Relational Database Management System, also ein relationales Datenbanksystem (im Gegensatz zu Objekt- oder Graphdatenbanken).

Der zweifellose De-facto-Standard in diesen Bereichen ist bereits seit einiger Zeit das Open-Source-Framework Apache Hadoop. Auf Hadoop basierend existieren etliche Derivate. Unter anderem sind hier Cloudera, Amazon Elastic MapReduce, Apache BigTop, Datameer, Apache Mahout, MapR und IBM PureData System zu nennen.

1.3 Motivation für Apache Hadoop/Spark

Anfang des 21. Jahrhunderts wurde das Bedürfnis für Möglichkeiten, sehr große Datenmengen effizient verarbeiten zu können, stetig größer. Nicht zuletzt durch die zu dieser Zeit exponentiell steigende Menge von Inhalten im World Wide Web und deren Indexierung durch Suchmaschinen wie Google. Davon motiviert wurde 2002 das Projekt *Nutch* mit dem Ziel gestartet, ein geeignetes *Such- und Crawlersystem* frei verfügbar zu machen. Die ersten Versuche skalierten sehr schlecht, bis Google 2003 die Funktionsweise ihres verteilten Dateisystems GFS (Google File System) veröffentlichte. Somit konnten die sehr großen Dateien, die durch die Indexierung entstanden, effizient auf verschiedene Knoten verteilt gespeichert werden und die Verwaltung dieser Knoten und Dateien aus dem eigentlichen Indexierungs- und Suchprozess ausgelagert werden.

Im Jahre 2004 publizierte Google den *MapReduce-Algorithmus*, der unter anderem die Indexierungs- und Analysefunktionen parallelisieren, delegieren und sinnvoll bündeln kann. In Nutch wurden daraufhin sämtliche wichtige Algorithmen auf MapReduce umgestellt, nachdem zuvor auch GFS unter dem Namen NDFS (Nutch Distributed File System) integriert wurde. Die möglichen Anwendungsgebiete von Nutch waren damit auch weit über das reine Suchen und Indexieren von Webseiten hinaus gewachsen. 2006 wurde aus Nutch ein Unterprojekt mit dem Namen Hadoop ausgegliedert, das im Jahre 2008 zum *Apache Top-Level-Project* ernannt wurde. Zu dieser Zeit nutzten bereits Firmen wie Yahoo!, Facebook oder die New York Times Hadoop. Ein exemplarischer Anwendungsfall bei der NY Times war, mit Hilfe der Hadoop-basierten EC2-Cloud von Amazon ca. vier Terabyte gescannter Archivdateien in PDF-Dateien umzuwandeln und dies in weniger als 24 Stunden auf 100 Knoten. Auch beim Sortieren von sehr großen Datenmengen stellten Hadoop-basierte Systeme nach und nach sämtliche Rekorde ein [Whi13].

Hadoop und Hadoop-basierte Systeme gelten mittlerweile als Industriestandard für Big-Data-Analytics-Anwendungen. Jedoch ist Hadoop nicht für alle Anwendungsgebiete gleichermaßen geeignet. Aufgrund der Charakterisierung der Paradigmen für Big Data Analytics im Paper „Frontiers in Massive Data Analysis“ der National Academic Press [Cou13], lassen sich die Einsatzgebiete und Schwächen für Hadoop ermitteln [Agn14].

So lassen sich mit Hadoop einfachere statistische Aufgabenstellungen sehr gut umsetzen. Dazu gehören Mittelwert, Median, Varianz und allgemein abzählende sowie ordnende statistische Aufgaben. Dies sind in der Regel Anwendungen mit einer Laufzeitkomplexität von $O(n)$ für n

Betrachtungswerte. Sie sind meist auch sehr gut parallelisierbar und somit sehr gut für Hadoop geeignet.

Für linear-algebraische Berechnungen (lineare Regression, Eigenwertproblem, Hauptkomponentenanalyse), generalisierte n -Körper-Probleme (mit einer Komplexität von $O(n^2)$ oder $O(n^3)$), Graphentheorie, Optimierungsprobleme (Verlust-, Kosten- oder Energiefunktionen, sowie Integrations- und Ausrichtungsfunktionen) ist *Hadoop* nur in jeweils einfacher Problemausprägung einsetzbar. Auch für Interaktive Abfragen ist *Hadoop* nur bedingt geeignet, da es ursprünglich für die *Batch-Verarbeitung* entwickelt wurde.

Aus diesem Grund wurde am *AMPLab* der University of California in Berkeley nach Alternativen geforscht, die auch für komplexe linear-algebraische Probleme, generalisierte n -Körper-Probleme und diverse Optimierungsprobleme geeignet sind. Das Ergebnis ist *Spark*, mittlerweile *Apache Top-Level-Projekt* und dazu geeignet, die Nachfolge von Hadoop als *Big-Data-Analytics-Framework* anzutreten.

1.4 Ziel und Aufbau dieser Arbeit

Die vorliegende Arbeit beschäftigt sich mit *Apache Spark* und dem dazugehörigen Ökosystem bestehend aus Schichten mit verschiedenen Bibliotheken, dem *Berkeley Data Analytics Stack (BDAS)*. Auf dem Markt befindet sich bislang noch verhältnismäßig wenig Literatur zu diesem Thema und wenn, dann werden zumeist Teilaspekte für bestimmte Anwendungsbereiche gekapselt betrachtet. Diese Arbeit soll, nachdem einige Grundlagen zum Thema *Big Data Analytics* im Allgemeinen diskutiert werden, zunächst einen ganzheitlichen Überblick über den BDAS bieten. Hier werden die einzelnen Schichten des Stacks kurz beschrieben und gegebenenfalls Alternativlösungen zu den jeweiligen Implementierungen vorgestellt.

Insbesondere werden in den darauffolgenden Kapiteln die Statistik- und Vorhersagebibliotheken *MLLib* betrachtet, in einigen Punkten mit der Alternativbibliothek *H2O* verglichen und die Kernimplementierung von Spark mit dem neueren Framework *Apache Flink* ² verglichen. Auch die übrigen Elemente des BDAS wie das Caching-Frameworks Tachyon, sowie die Streaming-Bibliothek Spark Streaming und die Graphenanwendung GraphX werden vorgestellt.

Im letzten Teil dieser Arbeit werden jeweils praktische Anwendungsbeispiele gezeigt und im Zuge dessen die APIs ³ von Spark und MLLib insbesondere vorgestellt.

² Entwickelt von der Technischen Universität Berlin zunächst unter dem Namen *Stratosphere* und mittlerweile (Stand Ende 2014) *Apache Incubator* Projekt

³ API = Application Programming Interface. Schnittstelle, die von einem Softwaresystem für die Einbindung in andere Softwaresysteme zur Verfügung gestellt wird.

Kapitel 2

Allgemeine Grundlagen

Das nachfolgende Kapitel behandelt die Grundlagen, die für ein Verständnis der Anwendungsbereiche von Apache Spark, dem Berkeley Data Analytics Stack und im Allgemeinen des Themenkomplexes Big Data Analytics und insbesondere für Machine-Learning-Anwendungen nötig sind. Im ersten Unterkapitel werden die grundsätzlichen Eigenschaften eines verteilten Systems beschrieben um die Basis für die in der Arbeit beschriebenen Besonderheiten von Verarbeitungen im Clusterbetrieb zu legen. Hier wird ein exemplarischer Clusteraufbau skizziert, Probleme mit Concurrency und Netzwerkverkehr beschrieben und welche Möglichkeiten es hier gibt. Im darauf folgenden Unterkapitel werden grundlegende Problemstellungen und Technologien beschrieben, die im Rahmen von Big Data Analytics im Allgemeinen vorkommen. Unter anderem werden hier Grundlagen und Begriffe aus den Themengebieten Anwendungen von Big Data Analytics, Machine Learning, Klassifikation, Vorhersagen, statistische Analysen, Graph-Suchen und Streaming-Frameworks in erklärt. Besonders die Algorithmen, die in den Machine-Learning-Implementierungen MLlib und H2O zum Einsatz kommen, werden hier detaillierter vorgestellt. In einer Zusammenfassung werde diese Grundlagen nochmals auf einen Blick dargestellt.

2.1 Cluster Computing

Die Nachfrage nach immer mehr Rechenleistung hat in den letzten Jahren dazu geführt, dass verstärkt Rechnercluster eingesetzt werden. Alternativ gibt es den Ansatz, Mainframes¹ mit immer mehr Rechenleistung auszustatten, diese jedoch ausdrücklich autonom zu betreiben². Je nach Aufgabenspektrum ist die eine oder andere Infrastruktur besser geeignet. In der Regel wird ein geclustertes System dort eingesetzt, wo hohe Verfügbarkeit oder gut parallelisierbare

¹ Unter Mainframe wird hier ein sehr leistungsfähiges Rechnersystem verstanden, das einen oder beliebig viele Prozessoren in einer physischen Einheit, also einem logischen Mainboard verbindet.

² In diesem Kontext kann durchaus ein Failover-Cluster vorhanden sein, also eine Mainframe wird zur Ausfallsicherheit repliziert. Dies wird an dieser Stelle jedoch nicht als Cluster im eigentlichen Sinn bezeichnet.

Aufgaben vorherrschen. Bei netzwerkintensiven Aufgaben, wie z.B. als Webserver oder Datenbanksystem sollten in der Regel besser Installationen auf einem autonomen System eingesetzt werden [TR10].

Ein Rechner-Cluster besteht in der Regel aus mehr oder weniger eng miteinander verbundenen Computern, wobei hier im Gegensatz zu Mainframes jeder Rechner über eigene Ressourcen wie Hauptspeicher, Massenspeicher, etc. verfügt. Ein Cluster, bzw. ein Verteiltes System ist nach Andrew S. Tanenbaum [AST07] folgendermaßen definiert:

„A distributed system is a collection of independent computers that appears to its users as a single coherent system.“

Bei der Verwendung eines Clusters sind einige Besonderheiten zu beachten, die bei der Ausführung auf gewöhnlichen Systemen nicht ins Gewicht fallen [AST07]. Unter Anderem sind die Tasks so gestalten, dass möglichst wenig Wartezeit durch Abhängigkeiten entsteht und diese möglichst autonom verarbeitet werden können. Außerdem muss beachtet werden, dass die einzelnen Knoten eines Clusters über Messaging-Mechanismen miteinander kommunizieren und dies insbesondere hohe Anforderungen an die Netzwerkinfrastruktur stellt. Eine typische Clustertopologie besteht aus mehreren Worker-Knoten und einem Masterknoten. Der Masterknoten delegiert die Tasks an die einzelnen Worker-Knoten und stellt das gesamte Cluster nach außen hin als ein geschlossenes System dar. Sämtliche Kommunikation mit dem Cluster findet grundsätzlich nur über den Masterknoten statt.

2.2 Anwendungen für Big Data Analytics

Im folgenden Unterkapitel werden exemplarisch einige Anwendungsfälle für *Big Data Analytics* (auch *Data Mining*) dargestellt, um zu klären, für welche Einsatzbereiche *Frameworks* wie *Apache Spark* und die darauf aufbauenden Bibliotheken in der Praxis benötigt werden.

Laut Arvind Sathi [Sat12] zeichnet sich *Big Data* unter Anderem durch ein mögliches Vorkommen von unstrukturierten Daten aus. Die Autoren Chakraborty und Pagolu gehen in ihrem Artikel [DGC14] davon aus, dass mittlerweile mehr als 80% der gesamten Daten im digitalen Raum in unstrukturierter Form vorliegen. In der Vergangenheit mussten für Analysetätigkeiten in aller Regel strukturierte Datensätze vorliegen. Grundsätzlich ist es mit den gängigen *Data Analytics Frameworks* nach wie vor möglich, beispielsweise quantitative Analysen auf strukturierten Datensätzen durchzuführen oder unstrukturierte Datensätze nachträglich zu strukturieren, um wiederum quantitative Analysen darauf anwenden zu können.

Das Potential dieser *Frameworks* zeigt sich jedoch dann in vollem Umfang, wenn auf unstrukturierten Daten diverse Analysemethoden oder Verarbeitungen angewendet werden. In jüngerer Vergangenheit ist das Aufkommen unstrukturierter Daten, wie bereits erwähnt, erheblich gestiegen. Dies wird nicht zuletzt durch die massive Verbreitung von Sensoren aller

Art verursacht. Dies können *Logdaten*, Bewegungsdaten, Sensorwerte zur Überwachung von technischen Einrichtungen, Messwerte aus Wetterstationen und unzähligen weiteren Quellen sein. Auch viele Internetanwendungen, besonders wenn es sich um laufende Datenströme handelt, verursachen erhebliche Datenmengen, die entweder *persistiert* oder sogar zur Laufzeit analysiert werden können.

*Data Mining*³ ist laut [JC14] ein analytischer Prozess mit dem Zweck, große Datenmengen nach konsistenten Mustern oder systematischen Beziehungen zu untersuchen. Die Ergebnisse werden in der Regel validiert, in dem gefundene Muster oder Ähnlichkeiten auf einer Teilmenge der ermittelten Daten angewendet werden. Ein weiteres Ziel von *Data-Mining-Prozessen* sind Vorhersagen von Ereignissen mittels geeigneter Algorithmen (Vergleich [JC14] und 2.3).

Der Prozess des Data Mining setzt sich nach [UF96] im Wesentlichen aus einem oder mehreren der folgenden Aufgabenbereiche zusammen:

- **Klassenbeschreibung:** Eine knappe Beschreibung der Charakterisierung der Datensätze, um sie eindeutig von anderen Daten unterscheiden zu können.
- **Assoziation:** Die Untersuchung der Daten nach assoziativen Verbindungen oder Korrelationen zwischen einzelnen Daten oder Datengruppen.
- **Klassifizierung:** Hier wird ein definierter Satz von Trainingsdaten⁴ analysiert und anhand deren Beschaffenheit ein Modell generiert. Durch die Klassifizierung werden Entscheidungsbäume (siehe Kapitel 2.3) oder Klassifizierungsregeln generiert, die schließlich für die Klassifizierung folgender Daten verwendet werden [SW97].
- **Vorhersage:** Die Vorhersage bezieht sich auf mögliche Werte von nicht-vorhandenen Daten oder Datenspektren, die wiederum durch *Approximation* einer Funktion mittels Beispielen durchgeführt wird [JC14]. Dies sind ebenfalls Trainingsdaten, die aus Datensätzen mit den dazugehörigen berechneten Funktionswerten bestehen.
- **Cluster-Analyse:** Diese dient dazu, *Cluster* innerhalb von Datensätzen zu ermitteln. Dies sind Daten, die definierte Ähnlichkeiten zueinander aufweisen.
- **Zeitreihenanalysen:** Hier werden in der Regel große Mengen an Zeitreihendaten analysiert, um nach Ähnlichkeiten oder Mustern innerhalb der Daten zu suchen.

³ Ein Großteil der Literatur verwendet die Begriffe *Big Data Analytics* und *Data Mining* synonym. *Machine Learning* wird jedoch von *Data Mining* abgegrenzt, da letzteres eine explorative Datenanalyse darstellt.

⁴ Trainingsdaten können beispielsweise Daten sein, die bereits im Vorfeld manuell klassifiziert wurden, deren Klassenzugehörigkeit also bekannt ist.

2.3 Machine Learning

„Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task (or tasks drawn from a population of similar tasks) more effectively the next time.“ [HS83]

Unter *Machine Learning* wird ein interdisziplinärer Teilbereich der Informatik und der Statistik verstanden. Ziel ist die Erstellung von Algorithmen, die in der Lage sind, selbstständig auf Grund von Daten iterativ zu lernen gemäß der oben zitierten Definition. Um dies zu erreichen, erstellen diese Algorithmen basierend auf den jeweiligen Eingabedaten Modelle, die Entscheidungen oder Vorhersagen treffen können [GJ13]. In den Abbildungen 2.1 und 2.2 wird dies veranschaulicht.

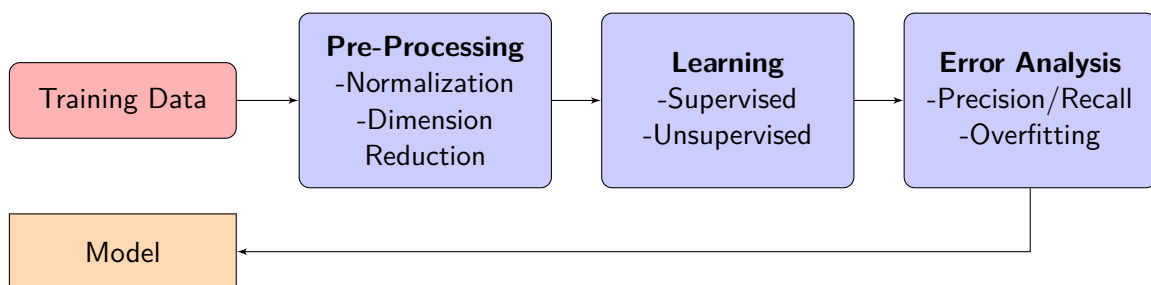


Abbildung 2.1: Der Machine-Learning-Prozess: Phase 1 - Lernphase

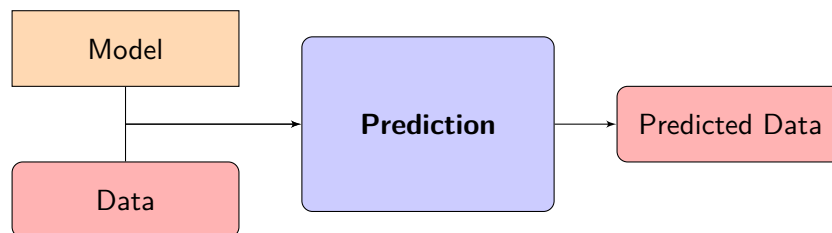


Abbildung 2.2: Der Machine-Learning-Prozess: Phase 2 - Prediction-Phase (Vorhersage)

Das erste Diagramm 2.1 zeigt den Lernprozess. Zunächst werden definierte Trainingsdaten einem *Pre-Processing*⁵ zugeführt (Vergleich Unterkapitel 2.3.1).

Danach folgt der eigentliche Lernprozess, der mit oder ohne *Überwachung*⁶, als *Minimalisierungsfunktion* oder mit anderen Lernalgorithmen durchgeführt wird (siehe Unterpunkte Clusterverfahren und Klassifikationsalgorithmen). Nach dem Lernprozess werden das erzeugte Modell einem Validierungsprozess zugeführt. Dies können *Precision/Recall*⁷, *Overfitting*⁸, oder andere

⁵ Das Pre-Processing kann aus *Normalisierung*, *Dimensionsreduktion*, *Bildverarbeitung* oder anderen Vorarbeiten bestehen.

⁶ Beim überwachten Lernen (Supervised Learning) liegen der Analyse vorher angelegte Trainingsdaten zugrunde, das unüberwachte Lernen (Unsupervised Learning) erzeugt vorher gänzlich unbekannte Modelle aus gefundenen Mustern [Gha04]

⁷ *Precision*: Wie viele der ausgewählten Datensätze sind relevant. *Recall*: Wieviele relevante Datensätze wurden ausgewählt.

⁸ *Overfitting* bezeichnet einen Zustand, in dem bekannte Daten (Trainingsdaten) von einem Algorithmus generell akkurater erkannt werden, als Daten, die von dem Algorithmus erkannt werden sollen (Predictions).

Validierungsfunktionen sein. Am Ende der Prozesse entsteht ein Modell, das für verschiedenste Aufgaben eingesetzt werden kann.

In Abbildung 2.2 wird das erzeugte Modell zusammen mit produktiven Daten genutzt, um mittels geeigneter Vorhersagealgorithmen (siehe Unterkapitel 2.3.2) bisher noch nicht vorhandene Datensätze erzeugen zu können.

Die Hauptanwendungsgebiete für *Machine Learning* sind sämtliche Bereiche, in denen eine Anwendung von strikten, regelbasierten Algorithmen nicht in Frage kommt [SW97]. Beispiele für diese Bereiche sind laut [Bis06] Suchmaschinen, Sprach- und Musikererkennung, Handschrifterkennung, Spamfilter, Umgebungserkennungen und viele mehr.

2.3.1 Klassifizierung von Daten

Im Gegensatz zu herkömmlichen *Business-Intelligence-Anwendungen*⁹ zeichnen sich Big-Data-Anwendungen in erster Linie durch Unterschiede in der Herangehensweise und in der Formulierung der primären Fragestellungen aus. In der Vergangenheit wurden zu Beginn einer Analyse zunächst Problemstellungen formuliert und aufgrund dessen Prozesse und vor allen Dingen die zu sammelnden Daten definiert [NG12]. Mit der steigenden Etablierung von Big-Data-Anwendungen hat sich dieser Ansatz gewandelt. Hier werden zunächst sämtliche anfallenden Daten gespeichert. Auf diese Datensätze, deren Größe in relativ kurzer Zeit massiv anwachsen kann, werden erweiterte Analyseprozesse¹⁰ angewendet, die unter anderem auch Vorhersagen erlauben [Rus11]. So liegen also bei Big-Data-Analytics große Datenmengen vor, ohne dass das Ziel der Analyse im Vorfeld bekannt ist.

Clusterverfahren

Nun gibt es verschiedene Herangehensweisen, wie sich relevante Fragestellungen aus den vorliegenden Daten ableiten lassen. Ein sinnvoller Ansatz besteht darin, sogenannte Cluster innerhalb der Daten zu ermitteln, also ähnliche Datensätze zu entsprechenden Gruppen zusammenzufassen. Aus diesen Gruppen werden Klassen gebildet, in dem ihnen aussagekräftige Namen zugeordnet werden [JC14]. Cluster werden beispielsweise benötigt, um Kunden nach bestimmten Interessen zu gruppieren, um Zeichen auf Ähnlichkeiten für die automatische Zeichenerkennung zu untersuchen, oder Bilder auf vergleichbare Bildpunktanordnungen oder Farbspektren, also generell überall, wo nach Ähnlichkeiten gesucht wird.

⁹ Laut [NG12] versteht man unter *Business Intelligence (BI)* die Sammlung, Auswertung und Darstellung von Daten anhand definierter Prozesse.

¹⁰ Unter dem Begriff *Advanced Analytics* sind verschiedene Werkzeugtypen aus der vorhersagenden Analyse (*Predictive Analytics*), aus dem Data Mining, aus der Statistik, der Künstlichen Intelligenz, der Sprachverarbeitung und weiteren Disziplinen zusammengefasst [Rus11].

Ein Cluster zeichnet sich dadurch aus, dass die Objekte innerhalb eines Clusters möglichst ähnlich sind, die Cluster untereinander jedoch möglichst unähnliche Inhalte besitzen. Um Ähnlichkeiten zwischen Datensätzen oder Clustern feststellen zu können, bedarf es einer Distanzfunktion, zur Ermittlung und Sicherstellung der Clustergüte einer Qualitätsfunktion (Vergleich [JC14]).

Laut [JC14] sind für Distanz- und Qualitätsfunktionen folgende Annahmen zu treffen:

Vorbedingungen

X (Instanzenraum)

$E \subseteq X$ (Instanzenmenge)

$X \times X \rightarrow \mathbb{R}^+$ (Abstandsfunktion)

$2^{2^x} \rightarrow \mathbb{R}$ (Qualitätsfunktion)

Gesucht wird eine Clustermenge $C = \{C_1, \dots, C_k\}$ mit folgenden Eigenschaften:

$$C_i \subseteq E$$

$$quality(C) \rightarrow max$$

$$C_i \cap C_j = \emptyset$$

$$C_1 \cup \dots \cup C_k = E$$

Die Abstandsfunktion $dist$, die gegebene Objekte so in Teilmengen zerlegt, dass der Abstand der Objekte innerhalb einer Teilmenge (Cluster) kleiner ist, als der Abstand zu Objekten anderer Teilmengen:

$$\forall C_i, C_j \in C (i \neq j) : \forall x_k, x_l \in C_i, x_m \in C_j : dist(x_k, x_l) < dist(x_k, x_m)$$

Die Qualitätsfunktion $quality$ beschreibt die Qualität des Clusterings basierend auf der Abstandsfunktion $dist$:

$$quality(C = \{C_1 \subseteq E, \dots, C_k \subseteq E\}) \rightarrow \mathbb{R}$$

Exemplarisch werden im Folgenden einige Distanzfunktionen dargestellt, die für die Clusteranalyse wichtig sind (Vergleich [CV14], [Bro98]):

- **Hamming-Distanz:** Hier werden die Vektorelemente der Position i zweier Vektoren miteinander verglichen. Wenn sich diese unterscheiden, wird die Distanz auf den Wert 1 gesetzt, bei gleichen Werten 0. Nachdem alle Elemente der Vektoren miteinander verglichen wurden, werden diese summiert. Ein Vorteil der Hamming-Distanz ist, dass sich Abstände zwischen metrischen, ordinalen und sogar nominalen Daten berechnen lassen. Der Nachteil dieser Distanzfunktion ist, dass Abstandsverhältnisse nicht erkannt werden, also ist beispielsweise der Abstand zwischen 2 und 3 für diese Funktion gleichwertig zum Abstand von 2000 zu 2100.

$$\text{dist}_H(x, y) = \text{count}_i(x_i \neq y_i)$$

- **Manhattan-Distanz:** Diese Distanzfunktion hat ihren Namen aufgrund der blockweisen Anordnung der Straßenzüge in Manhattan. Distanzen werden, wie auf einem Schachbrett, durch vertikale und horizontale Bewegungen und Richtungsänderungen beschrieben. Um die Ergebnisse der verschiedenen Distanzfunktionen untereinander vergleichbar zu machen, wird der Betrag der Distanzwerte verwendet. Die Manhattan-Distanz ist nur auf metrische Daten anwendbar. Diese Distanzfunktion ist auch unter dem Namen *Block-Distanz* bekannt.

$$\text{dist}_M(x, y) = \sum_i |x_i - y_i|$$

- **Euklidische Distanz:** Hier wird der direkte Abstand zwischen zwei Vektoren beschrieben, also zweier Punkte im n -dimensionalen Raum. Diese Funktion lässt sich nur auf metrische Daten anwenden. Auch hier wird durch Quadrieren und der Quadratwurze der Differenzen eine Normalisierung erreicht, die die Distanzwerte vergleichbar macht.

$$\text{dist}_E(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- **Tschebyscheff-Distanz:** Auch hier wird der Abstand zwischen zwei Punkten im n -dimensionalen Raum betrachtet. Im Unterschied zu den beiden vorhergehenden Verfahren wird hier jedoch der größtmögliche absolute Abstand zwischen allen Attributen ermittelt. Deshalb wird diese Distanzfunktion auch als *Maximum-Distanz* bezeichnet.

$$\text{dist}_T(x, y) = \max_i(|x_i - y_i|)$$

- **Minkowski-Distanz:** Diese Distanz-Funktion ist ähnlich zur Euklidischen Distanz. Allerdings entspricht hier der Abstand der n -ten Wurzel der Summe der n -ten Potenzen der Differenzen. Auch diese Funktion lässt sich nur auf metrische Daten anwenden.

$$\text{dist}_K(x, y) = \sqrt[n]{\sum_i (x_i - y_i)^n}$$

Darüber hinaus existieren noch zahlreiche weitere Distanzfunktionen, die je nach Einsatzzweck besser oder schlechter geeignet sein können, als die dargestellten.

Die Clusteranalyse zählt zu den *unüberwachten Lernalgorithmen* (*Unsupervised Learning*). Zoubin Ghahramani beschreibt dies in seinem Artikel *Unsupervised Learning* [Gha04] wie folgt:

„[...]in unsupervised learning the machine simply receives inputs x_1, x_2, \dots , but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise.“

Prinzipiell wird zwischen vier verschiedenen Clusterarten unterschieden [JC14]:

- **Partitionierendes Clustering:** Eine Menge von Datenobjekten wird in k Cluster, die um einen Medoid¹¹ oder einen Centroid¹² gebildet werden, zerlegt (Vergleich [Mir11]).
- **Hierarchisches Clustering:** Hier werden die Cluster hierarchisch aufgebaut, indem jeweils die Cluster mit der größten Ähnlichkeit, also mit der geringsten Distanz, miteinander vereinigt werden. Aus diesen vereinigten Clustern können in der Hierarchie übergeordnete Ebenen entstehen. Beim *Agglomerativen Clustering* wird mit den einzelnen Datenobjekten selbst begonnen, in dem jedes Cluster aus genau einem Datenobjekt besteht. Die jeweils ähnlichsten Cluster werden vereinigt und somit zu einer neuen Hierarchieebene. Dies geschieht so lange, bis in der obersten Ebene ein Cluster mit allen Datenobjekten entstanden ist. Beim *Divisiven Clustering* wird die Hierarchie in umgekehrter Reihenfolge aufgebaut, also zunächst werden alle Datenobjekte zu einem großen Cluster zusammengenommen und dieser dann iterativ anhand der Ähnlichkeiten der Datenobjekte untereinander geteilt.
- **Dichtebasiertes Clustering:** Die Cluster werden so gebildet, dass in der Umgebung eines Datenobjektes innerhalb eines Clusters möglichst viele weitere Datenobjekte liegen, die Dichte also besonders hoch ist. Die geforderte Dichte wird über Schwellenwerte definiert.
- **Clustering mit Neuronalen Netzen:** Auch über neuronale Netze können Cluster gebildet werden. Die Netze werden in diesem Fall nicht überwacht trainiert (Vergleich Einleitung Kapitel 2.3) und nutzen das Verfahren des *Wettbewerbslernens* (Vergleich [JC14], beispielsweise *Voronoi-Mengen*). Das Neuron, welches dem Eingabewert (dem Datenobjekt) am ähnlichsten ist, ist im jeweiligen Wettbewerb im Vorteil.

¹¹ Medoide sind Stellvertreterobjekte eines Clusters, deren durchschnittliche Ähnlichkeit zu allen Datenobjekten des Clusters maximal ist. Ein Medoid ist immer ein Datensatz aus dem Cluster.

¹² Centroide sind Vektoren der Mittelwerte der Attribute aus den Datenobjekten. Im zweidimensionalen Raum entspricht dies der Mittelwerte der x - und y -Koordinaten aller Datenobjekte.

Zu den Cluster-Verfahren zählen laut [WKH12] und [JC14] unter Anderem folgende Algorithmen:

- **klassischer k-Means-Algorithmus (partitionierendes Verfahren):** Wie bei allen partitionierenden Verfahren werden zunächst die Clusterzahl k und je Cluster ein Zentrum definiert. Die Zentrumsdefinition, und somit die Clusterbildung findet zufällig statt. Die Repräsentation des Clusters erfolgt durch den Centroid, also den Schwerpunkt. Eine Fehlerfunktion wird iterativ minimiert, indem die euklidischen quadrierten Abstände der Inhaltsobjekte zu den Centroiden hin verschoben werden. Nach jedem Iterationsschritt werden durch Mittelwertbildung die Centroiden neu definiert. Dieses Verfahren wird so lange iterativ durchgeführt, bis kein Datenobjekt mehr sein Cluster wechselt. Der k-Means-Algorithmus setzt metrische Werte¹³ voraus. Die Umwandlung von ordinalen Daten in metrische ist meist einfach umzusetzen, nominale Daten können mittels *Binarisierung* in numerische Werte gewandelt werden. Alternativ kann auch ein modifiziertes k-Means-Verfahren eingesetzt werden (Vergleich [JC14], Kapitel 8).
- **k-Medoid-Verfahren (partitionierende Verfahren):** Im Unterschied zum k-Means-Algorithmus wird hier nicht der Centroid, sondern der Medoid als Stellvertreter für ein Cluster definiert. Ein Medoid muss immer ein Element der Datenobjekte sein. Durch Vertauschung wird nun iterativ nach qualitativ hochwertigeren Medoiden gesucht, also solchen, deren mittlere Distanz zu den anderen Datenobjekten im Cluster geringer ist. Dies wird so lange durchgeführt, bis kein Tausch der Medoiden mehr stattfindet, also jedes Cluster seinen qualitativ hochwertigsten Medoid besitzt. Die beiden wichtigsten k-Medoid-Verfahren sind PAM (Partitioning Around Medoids) und CLARANS (Clustering Large Applications based on RANdomized Search) (Vergleich [Ng02]). Bei PAM wird jeweils nach dem besten neuen Medoid mit Hilfe einer Qualitätsverbesserung durch graphische Repräsentation der Datenobjektverteilung und -häufigkeit gesucht. Allerdings ist PAM nur für kleine Datenmengen zu verwenden. CLARANS ist für große Datenmengen besser geeignet, da hier nicht sämtliche Datenobjekte durchsucht werden. Statt dessen beschränkt sich die Suche auf Teilmengen.
- **k-Median-Algorithmus (partitionierendes Verfahren)** Im Gegensatz zu den k-Means-Verfahren wird beim k-Median nicht mit der Euklidischen Distanz der Abstand zu den Zentren ermittelt, sondern die Summen der Manhattan-Distanzen (1-Norm-Distanzen) werden iterativ minimiert [BJA06].

¹³ *Metrisch*: Merkmal, das aus einer Zahl besteht und Dimension, sowie Nullpunkt besitzt (z.B. Geschwindigkeit, Einkommen, Alter), *ordinal*: Merkmal mit natürlicher Ordnung (z.B. Schulnoten sehr gut, gut, befriedigend,...), *nominal*: Merkmal ohne natürliche Ordnung (z.B. Geschlecht, Name)

- **EM-Clustering (partitionierendes Verfahren):** Dieses Verfahren basiert auf dem Expectation-Maximization-Algorithmus. Dieser arbeitet nach dem Prinzip, zunächst mit einem zufällig gewählten Modell zu starten. In der Expectation-Phase werden die Zuteilungen der Datenobjekte zum Modell verbessert, in Maximization-Phase werden die Modellparameter entsprechend der aktuellen Zuteilung verbessert, also das Modell wird analog der Datenobjekte angepasst [CBD08]. Das EM-Clustering ist im Gegensatz zum k-Means-Algorithmus in seiner Zuordnung unscharf, da prinzipiell jedes Datenobjekt jedem Cluster zugehörig sein könnte, sowie jedes Datenobjekt jeden Parameter verändern könnte.
- **Fuzzy C-Means (partitionierendes Verfahren):** Die Entfernung vom Clusterzentrum bestimmt über den Zugehörigkeitsgrad eines Datenobjekts. Der Zugehörigkeitsgrad wird mittels eines Intervalls zwischen 0 und 1 angegeben, wobei 1 eine totale Zugehörigkeit zum jeweiligen Clusterzentrum bedeutet. Die Verschiebung der Zentren findet im Gegensatz zum k-Means-Algorithmus jedoch abhängig vom Zuordnungsgrad ab. So werden Datenobjekte mit dem Zuordnungsgrad 1 vollständig bei der Berechnung berücksichtigt, während kleinere Zugehörigkeitsgrade den Einfluss entsprechen minimieren [NRP05].
- **divisive Clusterverfahren (hierarchisches Verfahren):** Wie bei der Vorstellung der prinzipiellen Clusterverfahren beschrieben, wird beim divisiven Clustering zunächst mit einem Cluster begonnen, das sämtliche Datenobjekte beinhaltet. In diesem wird nun nach genau dem Datenobjekt gesucht, bei dem die Ähnlichkeit zu den anderen Datenobjekten im Cluster minimal, also die mittlere Distanz maximal ist. In jeder Iteration wird um das so ermittelte Zentrum ein neues Cluster aus den Datenobjekten gebildet, die eine geringe Distanz zum jeweiligen Clusterzentrum aufweisen. Die Iteration läuft genau so lange, bis jedes Cluster nur noch ein Datenobjekt enthält.
- **agglomerative Clusterverfahren (hierarchisches Verfahren):** Dieses ist analog zu den *divisiven Clusterverfahren*, allerdings umgekehrt. Zunächst ist jedes Datenobjekt zugleich ein Cluster. Die Cluster mit einer geringen Distanz werden so lange iterativ zusammengefasst, bis alle Datenobjekte in einem Cluster vorhanden sind. In einem Alternativansatz werden die Cluster zusammengefasst, deren *totale Varianz* (Streuung) die geringste Steigerung aufweist.
- **DBSCAN (dichtebasiertes Clusterverfahren):** Cluster werden hier als Areale behandelt, in denen Datenobjekte nah voneinander entfernt lokalisiert sind. Die Cluster werden getrennt durch Areale, in denen die Datenobjekte eine größere Entfernung aufweisen. Ein Cluster wird gebildet, sobald die Dichte der Datenobjekte einen definierten Schwellenwert überschreitet. Kernobjekte sind Datenobjekte, die innerhalb eines definierten Abstandes mindestens k Nachbardatenobjekte besitzen. Randobjekte sind Datenobjekte, die einem Cluster nahe sind und werden in dieses eingeschlossen, alle anderen Datenobjekte werden als *Rauschen* nicht berücksichtigt [JA03].

Klassifikations- und Regressionsalgorithmen

Die Klassifikations- und Regressionsalgorithmen haben zum Ziel, die Daten in Klassen einzuteilen (Vergleich Abbildung 2.1). Beides sind gebräuchliche Formen des überwachten Lernens (supervised Learning) [JWS90]. Das Training findet also anhand von Beispielen statt, bei denen die Einordnung der Daten in Klassen bereits manuell vorgenommen wurde. Der Unterschied zwischen Klassifikation und Regression liegt in erster Linie in der Art der Variablen, die bestimmt wird. Bei der Regression werden kontinuierliche Werte für die Variablen bestimmt, bei der Klassifikation werden dagegen diskrete Werte erwartet.

Grundsätzlich existieren zwei Arten der Klassifikation (Vergleich [JC14]). Die *instanzenbasierten* Klassifikationsverfahren führen eine direkte Klassifizierung auf Grund von Beispieldaten auf den noch zu klassifizierenden Daten durch. Diese Verfahren sind verhältnismäßig einfach aufgebaut, da ein zu untersuchendes Datenobjekt mit den vorhandenen Testobjekten anhand einer Distanzfunktion verglichen und der Klasse mit der größten Ähnlichkeit zugeordnet wird.

Im Gegensatz dazu erstellt das *modellbasierte* Klassifikationsverfahren aus den zur Verfügung gestellten Testdaten ein Modell als Metaebene. Die Testdaten werden nach Erstellung und Verifizierung des Modells für die Durchführung des Klassifikationsprozesses nicht mehr benötigt.

Im folgenden Abschnitt werden einige wichtige Klassifikations- und Regressionsverfahren vorgestellt, die auch Teil der Implementierungen der Machine-Learning-Bibliotheken MLlib und H2O sind.

- **Linear Regression:** Bei den Linear-Regression-Algorithmen handelt es sich um eine der gebräuchlichsten Regressionsmethoden. Hier wird eine statistisch abhängige Variable durch 1 bis n unabhängige Variablen beschrieben [DML13]. Die Regressionskoeffizienten werden in erster Potenz im Modell berücksichtigt. Deshalb wird dieses Verfahren als *Linear Regression* bezeichnet.

Zusammengefasst durch die *Simple Linear Regression* lässt sich das Verfahren wie folgt darstellen [Enz14]:

Die Ausprägung einer Variablen ist bekannt, es soll die Ausprägung einer anderen Variablen vorhergesagt werden. Die vorherzusagende Variable (abhängige Variable) ist das *Kriterium*. Die zur Vorhersage genutzte Variable (unabhängige Variable) ist der *Prädiktor*. Ziel der Regression ist nun, die lineare Funktion zu finden, welche die Abhängigkeit zwischen Prädiktor und Kriterium beschreibt.

Als einfachste Vorhersagefunktion wird hier ein lineares Gleichungssystem betrachtet. Hier werden die Werte von Y geschätzt, indem die Werte von X linear transformiert werden:

$$\hat{y} = a_{yx} + b_{yx} * x$$

mit \hat{y} = Schätzwert des Kriteriums

a_{yx} = Schätzwert des Kriteriums wenn der Prädiktor den Wert 0 hat (*Intercept*)

b_{yx} = Regressionskoeffizient, Steigung der Linie zum vorhergesagten Wert des Kriteriums (*Slope*)

x = Wert des Kriteriums zur Schätzung des Prädiktors.

So kann durch die Koeffizienten die Änderung in den abhängigen Variablen durch die Änderung um eine Einheit in der unabhängigen Variablen vorausgesagt werden.

Nun wird noch ein Maß für die Schätzfehler benötigt. Hierzu wird bei den *Linear-Regression-Algorithmen* auf die *Least-Squares* zurückgegriffen, also die summierten, quadrierten Abweichungen der tatsächlichen von den vorhergesagten Werten. *Intercept* und *Slope* werden so bestimmt, dass die *Least-Squares* minimal sind.

$$\sum (y - \hat{y})^2 = \min \text{ (Ordinary Least Squares - OLS)}$$

So wird die Linie für das Modell bestimmt, die aufgrund der geringen Schätzfehler am Besten passt.

- **Logistic Regression:** Die *Logistic Regression* ist in Ihrem prinzipiellen Zweck der *Linear Regression* sehr ähnlich. Auch hier wird die Beziehung zwischen einer oder mehreren unabhängigen und einer abhängigen Variable modelliert. Im Gegensatz zur *Linear Regression* wird bei der *Logistic Regression* jedoch die Eintrittswahrscheinlichkeit eines Ereignisses berechnet. Somit wird also kein linearer Zusammenhang zwischen den abhängigen und den unabhängigen Variablen hergestellt, in dem die abhängige Variable einen präzisen numerischen Wert erhält. Statt dessen wird mittels der *Logistic Regression* ein Wahrscheinlichkeitswert zwischen 0 und 1 erwartet [BG12]. Die entsprechende Funktion stellt sich folgendermaßen dar:

$$P = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

mit P als Eintrittswahrscheinlichkeit für das Ereignis

α, β als Koeffizienten des Modells (Vergleich *Linear Regression*), die von x abhängig sind. Wenn x den Wert 0 hat, erreicht α den Wert von P . β zeigt die Wahrscheinlichkeit, dass sich eine 1 ändert, wenn x sich um eine Einheit ändert.

Ein weiterer Bestandteil der Logistic Regression sind die *Odds*, also die Verhältnisse der Wahrscheinlichkeit zu ihrer jeweiligen Gegenwahrscheinlichkeit. Diese werden im Logistic-Regression-Algorithmus als *Logit* (Logarithmus eines Odds) bezeichnet und stellen hier die abhängige Variable dar. Ein Logit ist folgendermaßen definiert:

$$\text{logit}(p) = \ln\left(\frac{P}{1-P}\right)$$

- **Support Vector Machines:**
- **Naive Bayes:** Hierbei handelt es sich um einen wahrscheinlichkeitsbasierten Algorithmus, dessen Zweck die Vorhersage der Klasse mit der höchsten Wahrscheinlichkeit ist. Da es sich um ein instanzbasiertes Klassifikationsverfahren handelt, ist hier keine Modellentwicklung durch die Trainingsdaten nötig. Beim Naive-Bayes-Verfahren wird von der Annahme ausgegangen, dass sämtliche Datenattribute untereinander statistisch unabhängig sind. Dieser Algorithmus basiert auf der Bayesschen Formel (Vergleich: [Mit15]):

$$P(X | Y) = \frac{P(Y|X) * P(X)}{P(Y)}$$

Laut [DJF06] lässt sich die Naive-Bayes-Klassifikation für mittlere bis große Trainingsmengen für Diagnose oder Klassifikation für Dokumente wie folgt anwenden:

Das Ziel besteht darin, jede Instanz X in der Funktion $f : X \rightarrow V$ durch die Attributwerte $[a_1, a_2 \dots a_n]$ zu beschreiben. Hierzu wird der wahrscheinlichste Wert von $f(x)$ ermittelt durch:

$$f(x) = \operatorname{argmax} P(a_1, a_2 \dots a_n | v_i) P(v_i) \rightarrow v_i \in V$$

Beim Naive-Bayes-Algorithmus wird folgende Annahme getroffen:

$$f(x) = P(a_1, a_2 \dots a_n | v_i) = \prod_i P(a_i | v_i)$$

Somit wird die Naive-Bayes-Klassifikation zu:

$$v_{NB} = \operatorname{argmax} P(v_i) \prod_i P(a_i | v_i) \rightarrow v_i \in V$$

Also wird für jeden der Klassikationswerte v_i die Wahrscheinlichkeit $P(v_i)$ berechnet und für jeden Attributwert a_i die von v_i abhängige Wahrscheinlichkeit $P(a_i | v_i)$.

- **Decision Trees:**
- **Random Forests:**

2.3.2 Vorhersagealgorithmen

Blablba

2.4 Das MapReduce-Paradigma

Alternativ zu Spark lässt sich im BDAS auch Hadoop als Kernimplementierung für die Datenanalyse oder -verarbeitung im Cluster einsetzen. Dies könnte beispielsweise sinnvoll sein, wenn die Infrastruktur des Clusters über relativ wenig Hauptspeicher verfügt, so dass Spark seinen In-Memory-Vorteil nicht ausspielen kann und die Aufgaben ohnehin den im ersten Kapitel beschriebenen einfachen statistischen Problemen entsprechen.

Hadoop besteht in erster Linie aus dem Hadoop File System (HDFS) und dem MapReduce-Programmiermodell. Dieses wurde von Google speziell entwickelt, um große Datensätze parallel in Clustersystemen verarbeiten und generieren zu können.

In Abbildung 2.3 wird das MapReduce-Modell an einem praktischen Beispiel zur Wortzählung gezeigt. Bei MapReduce wird zunächst eine problemspezifische Map-Funktion definiert. Diese splittet die Ursprungsmenge in gleichgroße Teile, versieht unstrukturierte Worte in einem ersten Schritt mit Standardwerten, um so aus jedem unstrukturierten Datensatz ein Key-Value-Pair zu generieren. Im gezeigten Beispiel wird hier jedem Wert eine Eins als Anzahl zugeordnet. Im

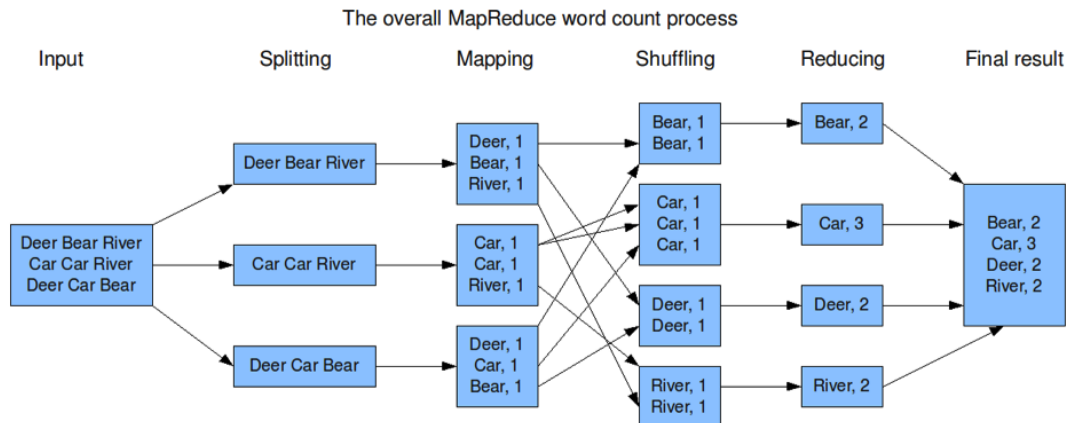


Abbildung 2.3: Ein konkretes Word-Count-Beispiel für MapReduce [Xia13]

nächsten Schritt werden diese so gewonnenen Zwischen-Key-Value-Paare in einem Shuffling-Prozess klassifiziert und die so homogenisierten Pakete auf die Knoten im Cluster verteilt. Nun enthält jede Ausführungseinheit nur noch gleiche Wörter. Im Reduce-Schritt schließlich, werden die gleichartigen Wörter zu Gesamtmengen zusammengefasst, addiert und im Endergebnis aggregiert.

An dem gezeigten Beispiel lässt sich gut erkennen, dass die einzelnen Ausführungsschritte jeweils parallelisiert und auf separate Knoten im Cluster verteilt werden können. Das Laufzeitsystem kümmert sich um die Details der Partitionierung der Eingabedaten, die Ressourcenverwaltung innerhalb des Clusters, das Behandeln von Fehlern und die Kommunikation zwischen den einzelnen Knoten.

2.5 Streaming Frameworks

Blablba

2.6 Anwendungen von Graphen

Blablba

Kapitel 3

Der Berkeley Data Analytics Stack (BDAS)

Rund um *Hadoop* beziehungsweise *Spark* wurde in Berkeley ein ganzer Infrastruktur-Stack für Big-Data-Analytics aufgebaut, der BDAS. Im folgenden Kapitel wird dieser Stack und die Bibliotheken, aus dem dieser besteht, vorgestellt. Im ersten Unterkapitel wird zunächst ein kurzer Überblick über den gesamten Stack gegeben. In den darauffolgenden Unterkapiteln werden die einzelnen Bestandteile im Einzelnen oberflächlich betrachtet, um dem Leser einen Einblick in die Nutzungsmöglichkeiten zu bieten. Eine Detailbetrachtung der jeweiligen Grundlagen, der praktischen Anwendung, Messungen und Vergleichsbetrachtungen der Alternativimplementierungen folgen im weiteren Verlauf dieser Ausarbeitung. Das Kapitel schließt mit einer Zusammenfassung, die auf einen Blick die Zusammensetzung des BDAS rund um Apache Spark zeigt.

3.1 Die Schichten des BDAS

Im Folgenden wird der *Big Data Analytics Stack (BDAS)* näher vorgestellt, der wie in der Einführung bereits erwähnt um Hadoop, bzw. Spark als Hauptbestandteile herum aufgebaut ist. Der BDAS wurde von den *AMPLabs*¹ von der University of California in Berkeley aufgrund von Forschungsergebnissen im Bereich der Analyse sehr großer Datenmengen ins Leben gerufen.

Der Einsatz des BDAS kann laut Vijay Agneeswaran [Agn14] dabei helfen, unter anderem Fragen wie die folgenden zu beantworten:

- Wie segmentiert man am besten eine Menge von Nutzern und kann herausfinden, welche Nutzersegmente an bestimmten Kampagnen interessiert sein könnten?

¹ kurz für „algorithms, machines and people“

- Wie kann man richtige Metriken für Nutzer-Engagement in Social-Media-Applikationen herausfinden?
- Wie kann ein Video-Streaming-Dienst für jeden Nutzer dynamisch ein optimales *Content-Delivery-Network (CDN)* ² basierend auf Daten wie Bandbreite, Auslastung, Pufferrate, etc auswählen?

Prinzipiell sind die in der Einführung in Kürze beschriebenen Einschränkungen von Hadoop und die damit verbundene Motivation für Spark auch die Motivation für den BDAS.

In Abbildung 3.1 wird eine Übersicht über die Hauptschichten des BDAS gezeigt, welche die jeweils von den AMPLabs empfohlene Implementierung zeigt.

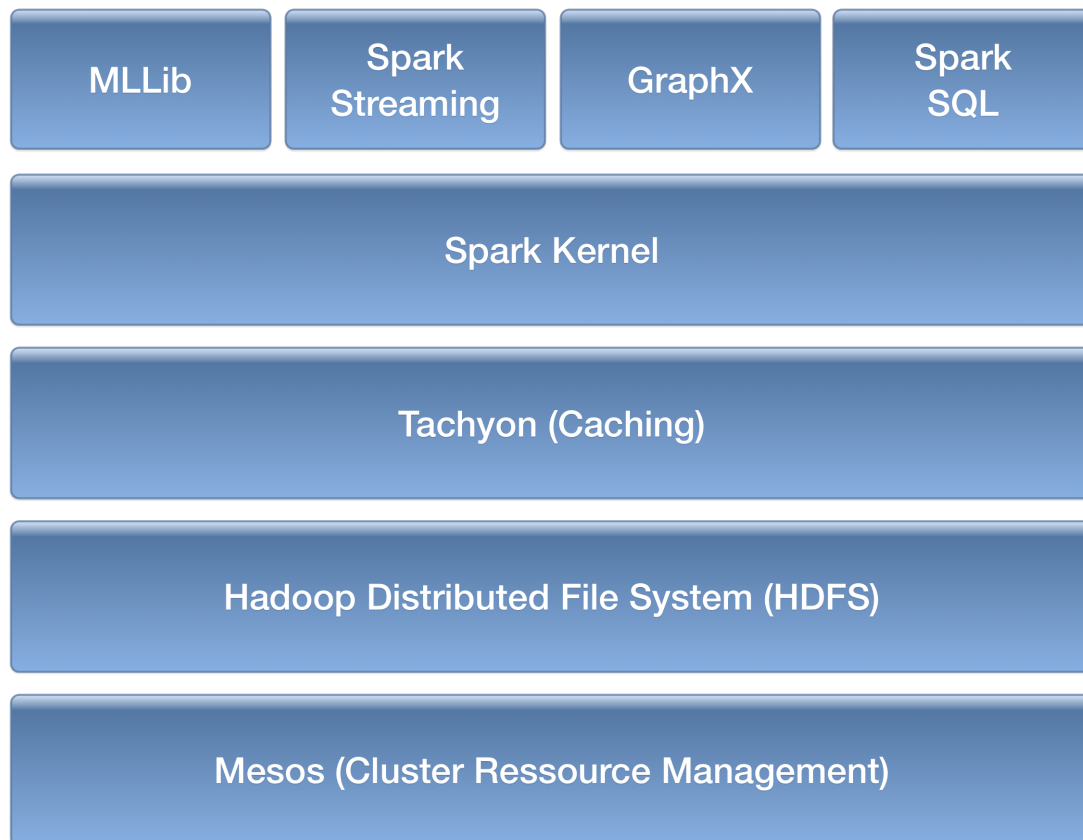


Abbildung 3.1: Übersicht des BDAS mit den vom AMPLab empfohlenen Bibliotheken.

In Abbildung 3.2 ist der Aufbau des BDAS nochmals schematisch dargestellt. Die grün hinterlegten Elemente markieren die Bestandteile des aktuellen BDAS, die violett hinterlegten zeigen alternative Implementierungen auf der jeweiligen Schicht. Grün schraffiert ist die Applikations-schicht, wo Applikationen oberhalb von Spark und den direkten Anwendungen angesiedelt sind.

² Ein Content Delivery Network (auch Content Distribution Network) ist ein Netzwerk verteilter und über das World Wide Web verbundener Server, das den Zweck hat, grosse Dateien bereitzustellen und auszuliefern.

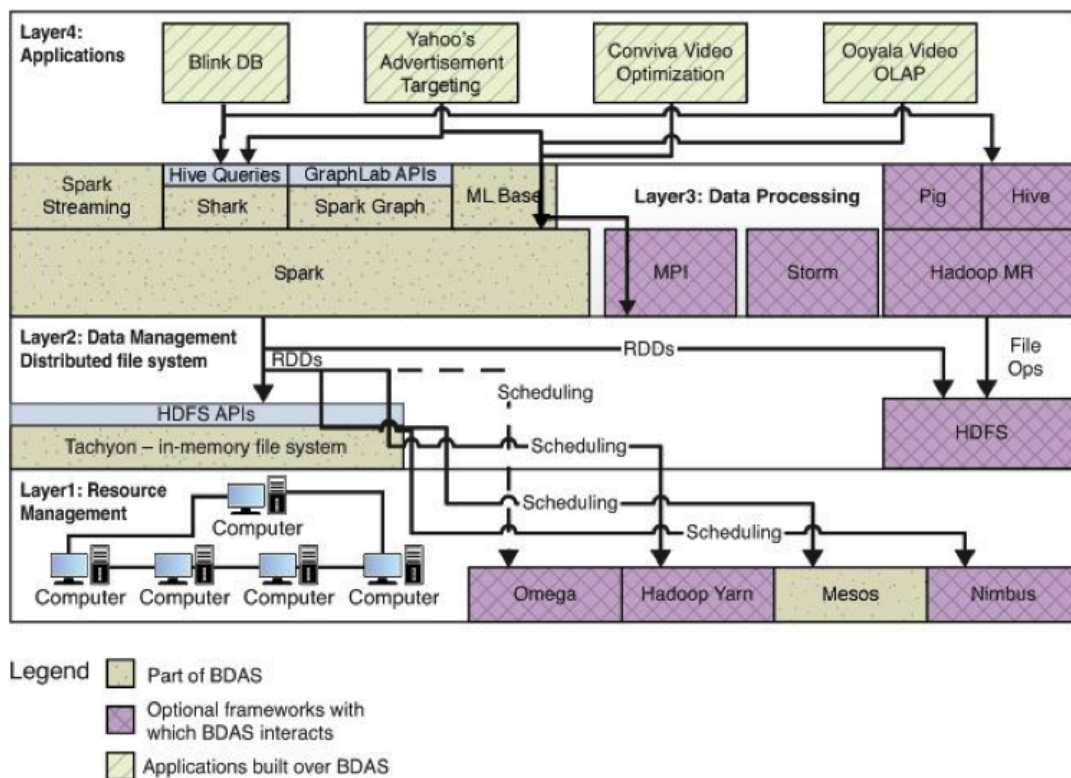


Abbildung 3.2: Der BDAS. Abbildung aus „Big Data Analytics Beyond Hadoop“, S 15 [Agn14]

3.2 Apache Mesos

Bei *Apache Mesos* handelt es sich um ein *Cluster-Management-Framework* für Anwendungen, die in verteilten Serverpools laufen sollen. Bestandteil von Mesos ist wiederum *Apache ZooKeeper*, das für Konfigurationsinformationen, Naming-Services und die Synchronisation von verteilten Anwendungen zuständig ist.

Mesos wird im BDAS eingesetzt, um die Prozesse von Hadoop/Spark effizient auf die einzelnen Knoten im Cluster zu verteilen. Besonders das Ressourcen-Management und –Monitoring innerhalb des Clusters ist ein wichtiger Faktor, um Jobs performant auf verteilten Systemen ausführen zu können. Auch das Fehlerhandling für Knoten, Prozesse und im Netzwerk wird im Berkeley-Stack von Mesos übernommen.

Ein besonderer Vorteil von Mesos gegenüber Yarn oder anderen Alternativen, wie dem Cloudera Cluster Manager oder Ambari von Hortonworks ist die Möglichkeit, verschiedene Frameworks gleichzeitig und isoliert in einem Cluster betreiben zu können. So kann beispielsweise Hadoop mit Spark in einer gemeinsamen Infrastruktur koexistieren.

3.3 Hadoop Distributed File System (HDFS) und Tachyon

Das Hadoop Distributed File System basiert ideologisch auf dem GoogleFileSystem (GFS) und hat zum Zweck, zuverlässig und fehlertolerant sehr große Dateien über verschiedene Maschinen hinweg in verteilten Umgebungen zu speichern. In entsprechenden Veröffentlichungen von Hortonworks [Hor14] wird von Produktivsystemen berichtet, die bis zu 200 PetaByte an Datenvolumen in einem Cluster von 4500 Servern basierend auf HDFS verwalten.

HDFS wurde speziell für den Einsatz mit MapReduce entwickelt, ist also auf geringe Datenbewegungen ausgelegt, da MR die Berechnungsprozesse jeweils zu den physischen Datensätzen selbst bringt und nicht, wie herkömmlich, die Daten zu den Prozessen geliefert werden müssen. So wird massiv Netzwerkverkehr innerhalb des Clusters eingespart und letztlich werden nur Prozesse und Prozessergebnisse verschickt.

Die Hauptbestandteile von HDFS sind der sogenannte NameNode, der die Metadaten des Clusters verwaltet und die DataNodes, die die eigentlichen Daten halten. Dateien und Verzeichnisse werden vom NameNode durch inodes repräsentiert. Diese wiederum enthalten Informationen über Zugriffsrechte, Zugriffszeiten oder Größenangaben der Dateien.

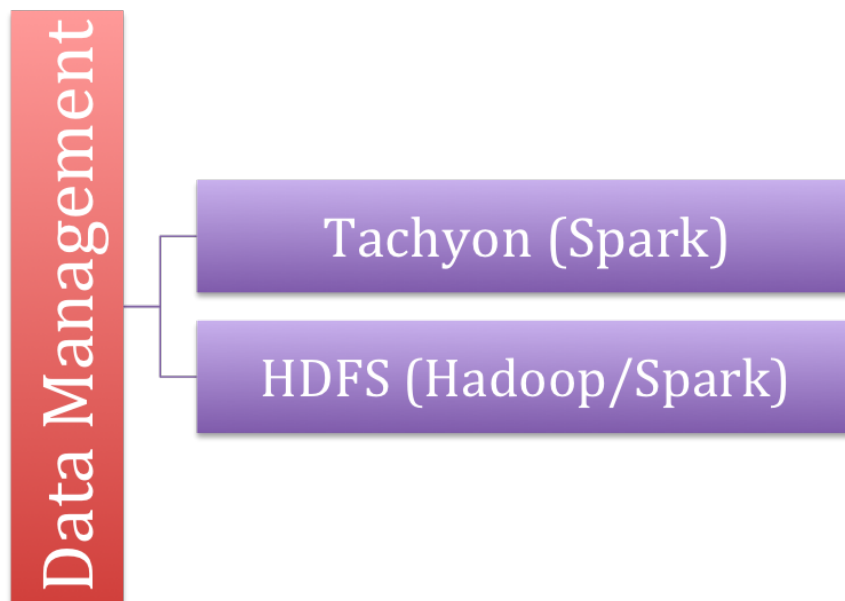


Abbildung 3.3: Der Datamanagement-Layer im BDAS mit HDFS und Tachyon

In Abbildung 3.3 wird die Datenmanagementschicht des BDAS nochmals detaillierter dargestellt. Hier ist erkennbar, dass reine Hadoop-Implementierungen direkt auf dem HDFS aufsetzen, da das HDFS für *MapReduce* optimiert ist.

Kommt hingegen Spark zum Einsatz, lässt sich wahlweise direkt das *HDFS* ansprechen oder alternativ eine Zwischenschicht nutzen, die auf das In-Memory-Modell von Spark zugeschnitten ist. Dies ist innerhalb des BDAS das verteilte Dateisystem Tachyon. Hier werden die zu verarbeitenden oder zu analysierenden Datensätze direkt in den Hauptspeicher des jeweiligen

Knoten in Form eines Cache gehalten. Somit werden Lade- und Speicheroperationen auf Massenspeicher minimiert und eine massiv höhere Ausführungsgeschwindigkeit erreicht. Unterhalb von Tachyon ist nach wie vor ein HDFS für die persistente Datenhaltung notwendig. Alternativ kann auch das Amazon S3-File-System eingesetzt werden. Tachyon wurde direkt innerhalb der AMPLabs entwickelt und ist mittlerweile fester Bestandteil des BDAS.

3.4 Apache Spark

Spark ist das Herzstück des BDAS. Bei Spark handelt es sich um ein open-source Data-Analytics-Framework, das, wie Hadoop, speziell für die Bedürfnisse im Rechner-Cluster konzipiert ist. Auch Spark nutzt das HDFS entweder direkt, oder indirekt über Tachyon. Im Gegensatz zu Hadoop bietet Spark jedoch Funktionen für In-Memory-Cluster-Berechnungen und ist nicht zwingend an MapReduce gebunden. Besonders interaktive Analyse oder Verarbeitung der Daten, Abfragen über verteilte Dateien und iterative

Lernalgorithmen erfahren so laut AMPLab eine bis zu hundertfache Ausführungsgeschwindigkeit im Gegensatz zu Hadoop. Auch die im ersten Kapitel angesprochenen Schwächen von Hadoop bei Berechnungen von komplexen linear-algebraischen Problemen, generalisierten n -Körper-Problemen, diversen Optimierungsproblemen und diversen anderen Aufgaben, treten bei Spark auf Grund der offenen Architektur und der Zerlegung von Datensätzen in die sogenannten Resilient Distributed Datasets (RDD) nicht mehr auf.

Spark wurde komplett in Scala entwickelt und bietet APIs für Scala, Java (inklusive Lambda-Expressions von Java 8) und Python. Im Labor existieren bereits Spark-Installationen mit bis zu 2000 Knoten, in Produktsystemen sind bisher Systeme mit bis zu 1000 Knoten im Einsatz [Met13]. Durch die Möglichkeit, die Datensätze im Speicher für interaktive Analyseaufgaben zu cachen und iterativ abzufragen, ist eine direkte Kommandozeileninteraktion über das integrierte Scala REPL (alternativ auch in Python) möglich.

Für Spark existieren dedizierte Bibliotheken für Verarbeitung von Datenströmen, Machine-Learning und Graphenverarbeitung. Ähnliche Artefakte existieren auch für Hadoop (Mahout, Vowpal Wabbit, etc.), jedoch ist die Architektur von Spark wesentlich besser für derartige Anwendungsbereiche zugeschnitten.

3.5 Spark Streaming

Spark Streaming ist eine der oben genannten Bibliotheken, die Spark um dedizierte Anwendungsbereich erweitert. Hierbei handelt es sich um eine Erweiterung, um die integrierte API von Spark für Anwendungen auf Datenströmen nutzen zu können. Das Programmiermodell

unterscheidet nicht zwischen Batch- und Streaming-Anwendungen. So lassen sich beispielsweise Datenströme zur Laufzeit mit Archivdaten vergleichen und direkt Ad-hoc-Abfragen auf die Ströme formulieren. Im Fehlerfall ermöglicht Streaming zahlreiche Wiederherstellungsoptionen, sowohl von verlorenen Datenströmen, als auch von Einstellungen. Ein Anwendungsbeispiel ist die Echtzeitanalyse von Twitter-Meldungen.

3.6 GraphX

GraphX ist eine Erweiterung für Spark, die verteilte, flexible Graphen-Anwendungen in einem Spark-Cluster ermöglicht [Xin13]. Besonders in den Disziplinen „Machine Learning“ und „Data Mining“ ist die Anwendung komplexer Graphen unerlässlich. Graph-datenbanken kommen immer dann zum Einsatz, wenn stark vernetzte Informationen und ihre Beziehungen zueinander interessant sind. Hier werden die Entitäten als Knoten behandelt, die Beziehungsart definiert die Kanten. Die Kanten können auch gewichtet

sein. Ein konkretes Beispiel sind die Mitglieder eines sozialen Netzwerks mit ihrem jeweiligen Beziehungsgeflecht. Je nach Kontaktintensität können diese Beziehungen auch priorisiert werden, was hier dem Kantengewicht entspricht.

GraphX nutzt hier die Vorteile der darunterliegenden Spark-Infrastruktur, in dem durch eine tabellarische Anordnung der Datenstrukturen eine massive Parallelisierung möglich ist und auch der Verarbeitung in RDDs voll unterstützt wird. So sind auch interaktive Operationen auf den Graphen jederzeit über REPL möglich.

3.7 MLbase/MLLib

MLbase ist eine Sammlung von Bibliotheken und Werkzeugen für Machine-Learning-Anwendungen mit Spark. Sie besteht grundsätzlich aus den drei Teilen MLLib, MLI und ML-Optimizer und ist oberhalb der Spark-Installation angesiedelt, wie auf Abbildung 3.4 zu erkennen ist.

Die MLLib ist eine verteilte Machine-Learning-Bibliothek die für die Spark-Laufzeitumgebung entwickelt wurde und die bekannten Algorithmen für Probleme wie Klassifikation, Regression, Clustering und kollaboratives Filtern enthält.

Bei MLI handelt es sich um eine API, die es ermöglicht, selbst ML-Features zu entwickeln und in erster Linie für komplexere Problemstellungen geeignet ist. Mit MLI lassen sich die Funktionen direkt gegen Spark entwickeln, gegebenenfalls unter Zuhilfenahme der Bibliotheken der MLLib.

Der ML-Optimizer soll ML-Probleme für Endnutzer vereinfachen, in dem Modellauswahlen automatisiert werden. Hierzu werden Features aus der MLLib und der MLI extrahiert und zur Hilfe genommen.

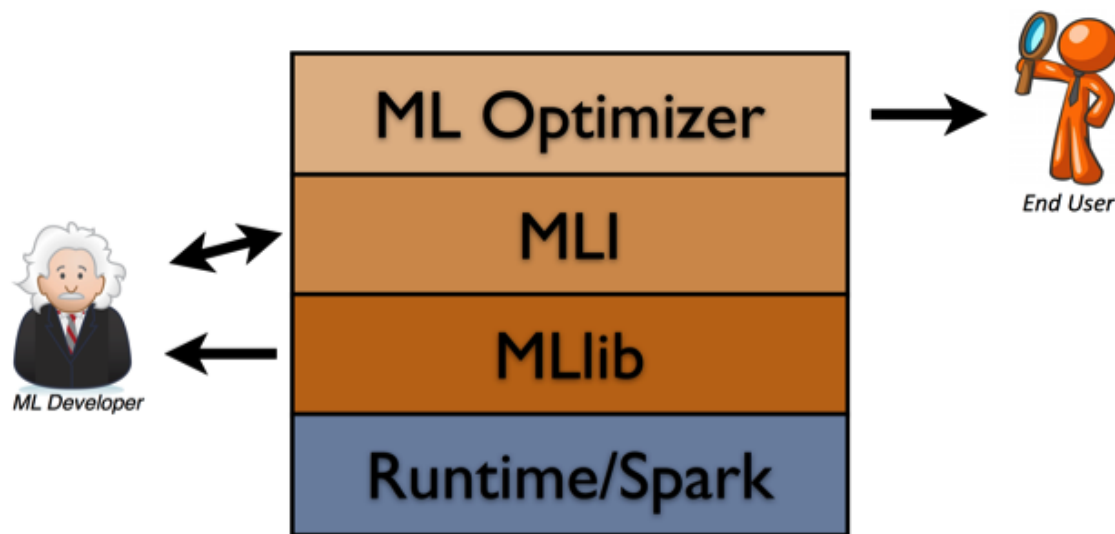


Abbildung 3.4: Die Bestandteile der MLbase [Lou10]

3.8 Spark SQL

Hive ist eine *SQL-Query-Engine* für Hadoop, die sich großer Beliebtheit in der Hadoop-Community erfreut. Spark SQL³ ist eine Portierung dieser Engine für Spark, um alle Vorteile der BDAS-Architektur nutzen zu können und ist kompatibel mit sämtlichen Hive-Daten, -Metastores und -Queries. Im Gegensatz zu Hive, das aus Datensätzen zur Laufzeit Java-Objekte generiert, nutzt Spark SQL eine zeilenorientierte Speicherung mittels Arrays primitiver Datentypen und ist somit selbst in einer Hadoop-Infrastruktur im Mittel bis zu fünfmal schneller als Hive.

Eine Besonderheit von Spark SQL ist neben seinem SQL-Interface die Möglichkeit, auch Machine-Learning-Funktionen als Abfragen formulieren zu können.

Für die Anwendung von Spark SQL hat sich die Architektur von Spark mit seinen RDDs als sehr vorteilhaft erwiesen, da Abfragen auf Fehlerhaften RDDs nach dem Neuaufbau des entsprechenden Datasets direkt erneut ausgeführt werden können.

Ein weiterer Unterschied zu Hive ist die sogenannte Partial-DAG-Execution (PDE). Dies bedeutet, dass logische Abfragepläne in Spark SQL aufgrund gesammelter Statistiken zur Laufzeit flexibel erstellt werden im Gegensatz zu Hive oder herkömmlichen relationalen Datenbanksystemen, wo bereits zur Kompilierungszeit starre physische Abfragepläne generiert werden. Besonders die Machine-Learning- und Failover-Funktionen wären mit einer Planerstellung zu Kompilierzeit nicht umsetzbar.

³ Ursprünglich war die für den BDAS-Stack empfohlene Implementierung einer SQL-Query-Engine unter dem Namen *Shark* bekannt. Im Juli 2014 wurde jedoch bekanntgegeben, dass die Entwicklung von Shark zugunsten von Spark SQL eingestellt wurde und die vorhandenen Shark-Implementierungen voll in Spark SQL integriert werden. Deshalb zeigen Schaubilder des BDAS von vor Juli 2014 Shark als Query-Engine.

3.8.1 Zusammenfassung

Kapitel 4

Alternative Implementierungen der Bibliotheken und Frameworks des BDAS

Wie in den Abbildungen 3.1 und 3.2 ersichtlich ist, existieren auf jeder Ebene des BDAS auch alternative Implementierungen. Einige davon werden im Folgenden kurz vorgestellt.

4.1 Alternative zu Spark: Apache Flink

lkjasdlfkjsaödlkfj aöksdjfölkajsajdfj äöaskdjfölkaskdfj

4.2 Alternative zu Spark Streaming: Storm

Storm ist, wie Apache Streaming, ein Framework für Hadoop, bzw. Spark für verteilte Streaming-Anwendungen. Wo Spark ganz klar eine Verbesserung gegenüber Hadoop darstellt und Shark dementsprechend für Hive, ist die Situation bei Storm und Apache Streaming dagegen nicht so klar determinierbar.

Storm und Spark Streaming unterscheiden sich fundamental in ihren Verarbeitungsmodellen [Agn14]. Das erstgenannte Framework verarbeitet eintreffende Events nacheinander, immer genau eines pro Zeitraum. Spark Streaming sammelt im Gegensatz dazu die Events in Mini-Batch-Jobs und verarbeitet sie paketweise zu definierten Zeiträumen nach wenigen Sekunden. Deshalb kann Storm Latenzzeiten von deutlich unter einer Sekunde erreichen, während Spark Streaming eine Latenzzeit von einigen Sekunden aufweist. Diesen Nachteil macht Spark Streaming durch eine sehr gute Fehlertoleranz wett, da die Mini-Batches nach aufgetretenen Fehlern

einfach nochmals bearbeitet werden können und die zuvor fehlerhaft ausgeführte Verarbeitung verworfen wird. Treten hingegen bei Storm Fehler auf, wird genau dieser Datensatz nochmals verarbeitet. Dies bedeutet, dass dieser auch mehrfach verarbeitet werden kann. Durch dieses Verhalten lassen sich die beiden Frameworks grob in zwei Einsatzgebiete verteilen:

Storm ist das Framework der Wahl, wenn Wert auf sehr kurze Latenzzeiten gelegt werden muss, hingegen ist es für statusbehaftete Anwendungen durch die Möglichkeit der Mehrfachverarbeitung ungeeignet. Im Umkehrschluss ist Spark Streaming eine gute Wahl, wenn aufgrund der gestreamten Daten eine Statusmaschine aufgebaut werden soll. Dafür müssen hier höhere Latenzzeiten in Kauf genommen werden.

4.3 Alternative zu MLLibs: H2O - Sparkling Water

BluBlaBlubb

4.4 Alternative zu MLLibs: Dato GraphLab Create™

BluBlaBlubb

4.4.1 Zusammenfassung

Kapitel 5

Funktionsweise von Spark

Im vorhergehenden Kapitel wurde der Berkeley Data Analytics Stack vorgestellt. Es wurde gezeigt, dass dieser aus einer Reihe von Bibliotheken, Infrastrukturkomponenten und dem eigentlich Kern, Apache Spark, besteht.

In diesem Kapitel werden die grundlegenden Konzepte von Spark vorgestellt und dessen Funktionsweise betrachtet. Einleitend wird gezeigt, wie eine Spark-Infrastruktur aufgebaut sein kann, wie diese intern Abfragen und eigene Spark-Programme verarbeitet und wie der *Spark-Context* sich als Cluster-Repräsentant gegenüber dem Anwender und der API exponiert. Im nächsten Unterkapitel wird die eigentliche Basis von Apache Spark vorgestellt. Spark basiert im Wesentlichen auf einer verteilten Datenstruktur, den *Resilient Distributed Datasets*. Deren Konzept wird sowohl theoretisch, als auch im Anwendungskontext dargestellt.

Ein weiteres Kernelement der Spark-Implementierung bildet das *In-Memory-Processing* der Daten. Spark ist in der Lage, je nach Konfiguration des Host-Systems, große Teile der Analysen und Verarbeitungen äußerst flexibel im Hauptspeicher durchzuführen und so massive Performanceverbesserungen gegenüber festspeicherbasierter Verarbeitung zu generieren. Hierzu bietet Spark spezielle *In-Memory-Primitives* an. In einem weiteren Unterkapitel werden dieses detailliert vorgestellt.

5.1 Spark im Cluster

Eine große Herausforderung im Umfeld verteilter und nebenläufiger Analyse und Verarbeitung großer Datenmengen stellt der Netzwerkverkehr da. Der klassische Aufbau einer verteilten Anwendung hält die Daten auf einer dafür vorgesehen Plattform im Netzwerk. Häufig ist dies ein dedizierter File- oder Datenbankserver, der mit möglichst großer Bandbreite mit dem Applikationsserver verbunden ist (Vergleich Oracle InfiniBand).

In diesem Aufbau entsteht in der Regel eine sehr hohe Netzwerklast, da die für die Applikation benötigten Daten dieser zunächst zur Verfügung gestellt werden müssen.

Spark geht hier einen anderen Weg. Ein Spark-Cluster besteht typischerweise aus einem zentralen *Master* und *n Worker-Nodes*. Diese können aus einfachen Servern bestehen, aber auch aus Clustern von Großrechnern (beispielsweise IBM Z, Oracle Exa). Das Hadoop Distributed File System und Spark skalieren über Cluster beliebiger Größenordnung. Über ein verteiltes Dateisystem werden die Daten auf dem Cluster gehalten und sowohl dem *Master*, als auch den *Worker-Nodes* so zur Verfügung gestellt.

Spark-Anwendungen laufen als unabhängiges Set von Prozessen auf Cluster-Infrastrukturen. Das Hauptprogramm, der sogenannte *Spark Driver*, instanziiert das *SparkContext-Objekt*, das die einzelnen Prozesse koordiniert. Auf Clustersystemen hält der *SparkContext* die Verbindung zum jeweiligen *Cluster-Ressource-Manager* (Mesos, Yarn), im Standalone-Betrieb instanziiert der Context selbst einen Dummy-Manager und allokiert in beiden Fällen die für die Anwendung nötigen Hardware-Ressourcen. Die Cluster-Manager liefern ihren aktuellen Status an Spark zurück und melden Auslastung und Gesundheitszustand der einzelnen Knoten. Über interne *Load-Balancing-Systeme*¹ wird ermittelt, welche Worker-Nodes die jeweiligen Tasks aus dem Spark-Kontext zugewiesen bekommen. Der SparkContext repräsentiert sowohl für die Spark-Konsole REPL, als auch in eigenen Spark-Programmen innerhalb der APIs das gesamte Cluster. Dem SparkContext wird bei der Initialisierung über ein Konfigurationsobjekt mitgeteilt, welche Ressourcen ihm für das aktuelle Programm zur Verfügung stehen. Die Entscheidung, welche der initial zur Verfügung gestellten Nodes oder Ressourcen des Clusters von Spark wann in Anspruch genommen werden, obliegt der Kombination aus Spark und *Cluster-Ressource-Manager*.

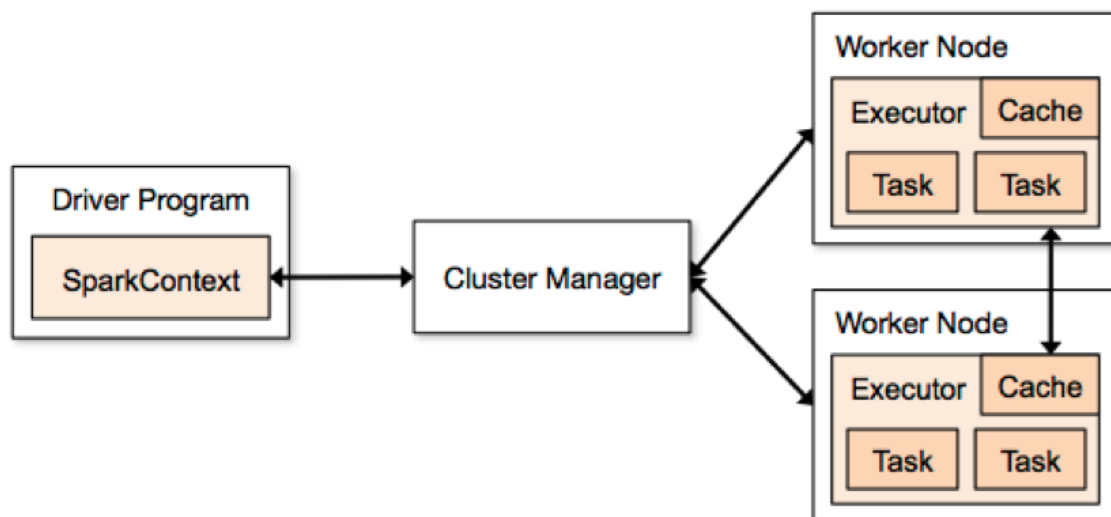


Abbildung 5.1: Clusteraufbau mit Spark [SP14]

¹ Load-Balancing-Systeme sind Überwachungsmechanismen in verteilten Systemen. Jedes Teilsystem meldet seine eigene Auslastung und seine Verfügbarkeit an den Load-Balancer. Dieser verteilt anstehenden Aufgaben so auf die Ressourcen, dass eine möglichst gleichmäßige Verteilung über die gesamte Infrastruktur möglich ist.

Wenn ein SparkContext initialisiert wurde, installiert der Spark sogenannte *Executors* auf sämtlichen Worker-Nodes des Clusters. Der Applikationscode wird nun als JAR² direkt an die Executors verteilt und dieser anschließend durch entsprechende Tasks ausgeführt.

5.2 Das Konzept der Resilient Distributed Datasets

Bei den Resilient Distributed Datasets (RDD) in Spark handelt es sich um verteilte Objekt-Collections, die im Allgemeinen im Speicher der Cluster-Knoten gecached und zwischen den Knoten bewegt werden. Sie können durch verschiedenste parallele Operatoren manipuliert und im Fehlerfall automatisch neu aufgebaut werden. Deshalb merken sich die RDDs die Transformationen, die zu ihrem Aufbau geführt haben und können so verlorene Daten schnell rekonstruieren. Da es sich bei RDDs prinzipiell um Scala-Collections handelt, können diese auch direkt in Scala-Code eingebunden und verarbeitet werden, oder interaktiv über die Scala-Konsole REPL genutzt werden. RDDs können nur durch

grobgranulare, deterministische Transformationen, wie beispielsweise map, filter, join, etc. erstellt werden.

RDDs können prinzipiell auf drei Arten gespeichert werden [SP14]:

- Als deserialisiertes Java-Objekt im Speicher der JVM – dieses Variante bietet die beste Performance, da die Objekte sich direkt im JVM-Heap befinden
- Als serialisiertes Java-Objekt direkt im Speicher – dieses Verfahren ist speicher-effizienter, aber schlechter in der Zugriffsgeschwindigkeit
- Im Dateisystem – diese Variante ist erwartungsgemäß die langsamste, jedoch nötig, wenn die RDDs zu groß für die Haltung im RAM sind.

Wie in Abbildung 5.3 dargestellt, werden die Daten bei einer Verarbeitung durch Spark zunächst aus dem HDFS geladen, in Resilient Distributed Datasets (RDD) verpackt, und dann im Hauptspeicher für Verarbeitung- oder Analysefunktionen zur Verfügung gestellt. Abfragen werden direkt entweder via Scala REPL oder SQL-artige Abfragen zur Laufzeit, über Batch-Jobs oder via Spark Streaming/Storm an die im RAM befindlichen RDDs geleitet.

5.3 In-Memory-Computing mit Spark

Bjsdfshdjhlslhfsajhlkjhas asdfhasdfjh asdlöfjhasdljfhagjskdjaskdf jasöldkfjaöl

² Ein JAR (Java ARchive) ist ein gepacktes und auf einer Java Virtual Machine ausführbares (Java, Scala, Clojure) Programmpaket, häufig inklusive der benötigten Bibliotheken.

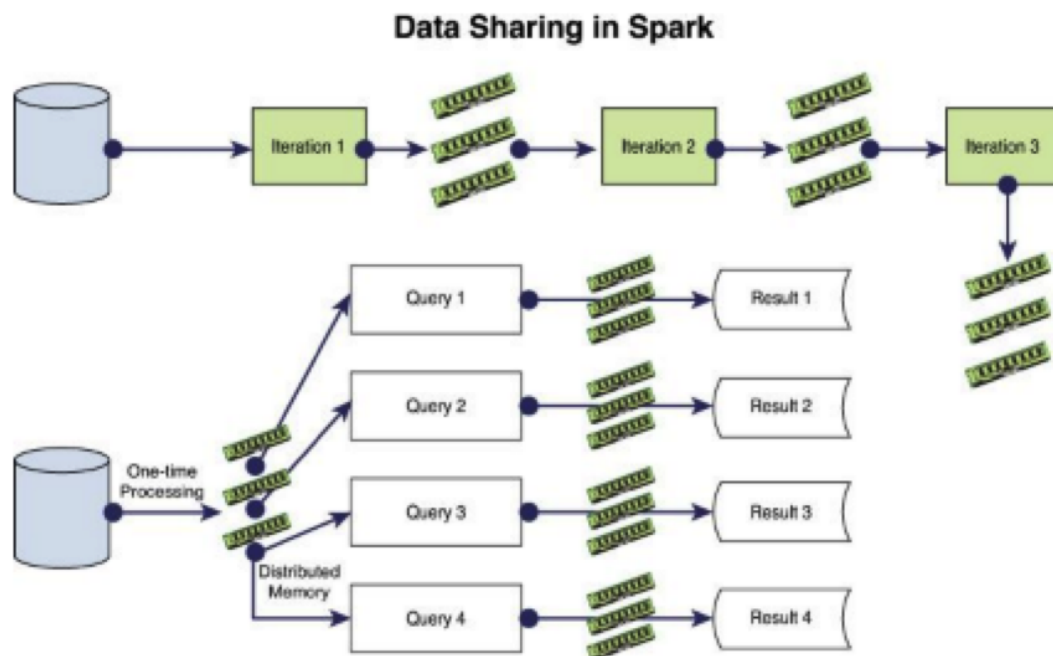


Abbildung 5.2: Schematische Darstellung der Funktionsweise von Spark va14

Kapitel 6

Exemplarischer Architekturaufbau einer Apache Spark Infrastruktur

Blablba

6.1 Ausführungscontainer: Docker

6.2 Cluster Management: Mesos und Yarn

Blablba

6.3 Caching-Framework: Tachyon

Blablba

6.4 Der eigentliche Kern: Apache Spark

Blablba

6.5 Streaming-Framework: Spark Streaming

Blablba

6.6 Abfrageschicht: Spark SQL

Blablba

6.7 Machine Learning Algorithmen: MLlib

Blablba

6.8 Graphenanwendungen: GraphX

Blablba

Kapitel 7

Implementierung der Prototypen

Blablba

7.1 Prototyp: Spark

Blablabla

7.1.1 Prototyp: Vergleich Prototyp Apache Flink

Blablba

7.2 Prototyp: MLLib

Blablba

7.2.1 Prototyp: Vergleich Prototyp H2O

Blablba

7.3 Prototyp: Spark Streaming

Blabla

7.4 Prototyp: GraphX

Blablba

Kapitel 8

Evaluierung der Komponenten und Alternativen

Blablba

8.1 Definition von Metriken für die Bibliotheken des BDAS

Blabla

8.2 Beschreibung der Messverfahren

Blabla

8.3 Beschreibung der Messumgebungen

8.3.1 Lokales Single Node Cluster

Blablba

8.3.2 Lokales Multi Node Cluster

Blablba

8.3.3 Remote Cluster an der Beuth Hochschule

Blablba

8.4 Ergebnisse

Blablba

8.4.1 Messergebnisse Apache Spark

Blablba

8.4.2 Messergebnisse Apache Flink

Blablba

8.4.3 Messergebnisse MLlib

Blablba

8.4.4 Messergebnisse H2O

Blablba

Kapitel 9

Schlussbetrachtung

Blablabla

9.1 Zusammenfassung

Blablba

9.2 Ausblick

Kapitel 10

Verzeichnisse

Literaturverzeichnis

- [Agn14] AGNEESWARAN, VIJAY: *Big Data Analytics Beyond Hadoop*. Pearson, 2014.
- [AST07] ANDREW S. TANENBAUM, MAARTEN VAN STEEN: *Distributed Systems - Principles and Paradigms*. Pearson, Prentice Hall, Second Edition Auflage, 2007.
- [Bis06] BISHOP, C.M.: *Pattern Recognition and Machine Learning*. Springer, 2006.
- [BJA06] BENJAMIN J. ANDERSON, DEBORAH S. GROSS, ET AL.: *Adapting K-Medians to Generate Normalized Cluster Centers*. Carleton College, Northfield Minnesota, 2006.
- [Bro98] BROSIUS, FELIX: *SPSS 8: Professionelle Statistik*. International Thompson Publishing, 1998.
- [CBD08] CHUONG B DO, SERAFIM BATZOGLOU: *What is the expectation maximization algorithm?* Nature Publishing Group, 2008.
- [Cou13] COUNCIL, NATIONAL RESEARCH: *Frontiers in Massive Data Analysis*. National Academic Press, 2013.
- [CV14] CLAUDIUS VIELHAUER, TOBIAS SCHEIDAT: *Fusion von biometrischen Verfahren zur Benutzerauthentifikation*. Otto-von-Guericke Universität Magdeburg Advanced Multimedia and Security Lab (AMSL), 2014.
- [GJ13] GARETH JAMES, DANIELA WITTEN, ET AL.: *An Introduction to Statistical Learning with Applications in R*. Springer, 4 Auflage, 2013.
- [HS83] H.A. SIMON, P. LANGLEY, G.L. BRADSHAW: *Rediscovering chemistry with the BACON system - Machine learning, an artificial intelligence approach*. Tiogra Publishing Co., 1983.
- [JA03] JOHANNES ASSFALG, CHRISTIAN BÖHM, ET. AL: *Knowledge Discovery in Databases*. Ludwig Maximilians Universität München, Institut für Informatik, 2003.
- [JC14] JÜRGEN CLEVE, UWE LÄMMEL: *Data Mining*. De Gruyter/Oldenbourg Wissenschaftsverlag, 1 Auflage, 2014.

- [JWS90] JUDE W. SHAVLIK, THOMAS G. DIETTERICH: *Reading in Machine Learning*. Morgan-Kaufman Publishers, Inc., 1 Auflage, 1990.
- [Mit15] MITCHELL, TOM M.: *Machine Learning*. McGraw Hill, Draft of January 31, 2015 Auflage, 2015.
- [Ng02] NG, RAYMOND T.: *CLARANS: a method for clustering objects for spatial data mining*. Nummer DOI: 10.1109/TKDE.2002.1033770. Dept. of Comput. Sci., British Columbia Univ., Vancouver, BC; IEEE Transactions on Knowledge and Data Engineering (Impact Factor: 1.82), 10 2002.
- [NG12] NORBERT GRONAU, CORINNA FOHRHOL: *Wettbewerbsfaktor Analytics – Reife-grade ermitteln, Wirtschaftlichkeitspotentiale entdecken*. Addison-Wesley, 2012.
- [NRP05] NIKHIL R. PAL, Kuhu PAL, ET. AL: *A Possibilistic Fuzzy c-Means Clustering Algorithm*, Band 13. IEEE TRANSACTIONS ON FUZZY SYSTEMS, 2005.
- [Rus11] RUSSOM, PHILIP: *TDWI Best Practices Report: Big Data Analytics*. TDWI Research, 2011.
- [Sat12] SATHI, DR. ARVIND: *Big Data Analytics: Disruptive Technologies for Changing the Game*. MC Press, First Edition Auflage, 2012.
- [SW97] S.M. WEISS, N. INDURKHIA: *Predictive data mining: A practical guide*. Morgan-Kaufman, 1997.
- [TR10] THOMAS RAUBER, GUDULA RÜNGER: *Parallel Programming: For Multicore and Cluster Systems*. Springer, 2010.
- [UF96] U.M. FAYYAD, G. PIATETSKY-SHAPIO, ET AL.: *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- [Whi13] WHITE, TOM: *Hadoop: The Definitive Guide*, Band 2nd Edition. O'Reilly, 2013.
- [WKH12] WOLFGANG KARL HÄRDLE, LÉOPOLD SIMAR: *Applied Multivariate Statistical Analysis*. Springer, 3 Auflage, 2012.

Internetquellen

- [BG12] BALTES-GÖTZ, B.: *Logistische Regressionsanalyse mit SPSS, Universität Trier, ZIMK* URL: <http://www.uni-trier.de/fileadmin/urt/doku/logist/logist.pdf>, 06 2012.
- [BIT14] BITKOM: *Big-Data-Technologien – Wissen für Entscheider* URL: http://www.bitkom.org/files/documents/BITKOM_Leitfaden_Big-Data-Technologien-Wissen_fuer_Entscheider_Febr_2014.pdf, 2 2014.
- [DGC14] DR. GOUTAM CHAKRABORTY, MURALI KRISHNA PAGOLU: *Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining* URL: <http://support.sas.com/resources/papers/proceedings14/1288-2014.pdf>, 08 2014.
- [DJF06] DR. J. FÜRNKRANZ, DR. G. GRIESER: *Maschinelles Lernen: Symbolische Ansätze* URL: <http://www.ke.tu-darmstadt.de/lehre/archiv/ws0607/mldm/>, 2006.
- [DML13] DAVID M. LANE, DAVID SCOTT, ET AL.: *Introduction to Statistics, Rice University, University of Houston* URL: <http://onlinestatbook.com/2/index.html>, 11 2013.
- [Dum11] DUMBHILL, EDD: *What is big data?* URL: <http://radar.oreilly.com/2012/01/what-is-big-data.htm>, 11 2011.
- [Enz14] ENZMANN, DR. DIRK: *Lineare Regression, Universität Hamburg, Juristische Fakultät* URL: <http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Lehre/StatIKrim/Regression.pdf>, 11 2014.
- [Gha04] GHARAMANI, ZOUBIN: *Unsupervised Learning* URL: <http://mlg.eng.cam.ac.uk/zoubin/papers/ul.pdf>, 09 2004.
- [Hor14] HORTONWORKS: *Hadoop Distributed File System* URL: <http://hortonworks.com/hadoop/hdfs/>, 04 2014.
- [Lou10] LOUKIDES, MIKE: *What is data science?* URL: <http://radar.oreilly.com/2010/06/what-is-data-science.html>, 10 2010.

- [Met13] METZ, CADE: *Spark: Open Source Superstar Rewrites Future of Big Data* URL: <http://www.wired.com/2013/06/yahoo-amazon-amplab-spark/all/>. Wired.com, 06 2013.
- [Mir11] MIRKES, E. M.: *University of Leicester: K-means and K-medoids* URL: http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html. University of Leicester, 01 2011.
- [Tur10] TURNER, JAMES: *Hadoop: What it is, how it works and what can it do* URL: <http://radar.oreilly.com/2011/01/what-is-hadoop.html>, 3 2010.
- [Xia13] XIA, C.: *Work Structure of MapReduce* URL: http://xiaochongzhang.me/blog/wp-content/uploads/2013/05/MapReduce_Work_Structure.png, 05 2013.
- [Xin13] XIN, GONZALES ET AL: *A Resilient Distributed Graph System on Spark* URL: <https://amplab.cs.berkeley.edu/publication/graphx-grades>. Amplab UC Berkeley, 08 2013.

Abbildungsverzeichnis

2.1	Der Machine-Learning-Prozess: Phase 1 - Lernphase	10
2.2	Der Machine-Learning-Prozess: Phase 2 - Prediction-Phase (Vorhersage) . . .	10
2.3	Ein konkretes Word-Count-Beispiel für MapReduce [Xia13]	20
3.1	Übersicht des BDAS mit den vom AMPLab empfohlenen Bibliotheken.	22
3.2	Der BDAS. Abbildung aus „Big Data Analytics Beyond Hadoop“, S 15 [Agn14]	23
3.3	Der Datamanagement-Layer im BDAS mit HDFS und Tachyon	24
3.4	Die Bestandteile der MLbase [Lou10]	27
5.1	Clusteraufbau mit Spark [SP14]	32
5.2	Schematische Darstellung der Funktionsweise von Spark va14	34

Tabellenverzeichnis

Listings

Anhang A

Zusätze

A.1 Quelltext