Universität Augsburg

Fakultät für Angewandte Informatik

Masterarbeit

im Studiengang Informatik

zur Erlangung des akademischen Grades
Master of Science

Thema:	<Thema der Arbeit>
Autor:	Gerald Siegert MatNr. 1450117
Version vom:	2. Januar 2018
1. Betreuerin:	Prof. Dr. X
2. Betreuer:	Prof. Dr. Y

Eidesstattliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum: Unterschrift:

Zusammenfassung

Abstract

Inhaltsverzeichnis

Abbildungsverzeichnis

Listingverzeichnis

Kapitel 1

Einleitung

1.1 Testen von Software und Systemen

Softwaretests sind in der heutigen Zeit eine wichtige Grundlage im Bereich der Qualitätssicherung bei Softwareprojekten. Abhängig von Komplexität und Sicherheitsanforderungen werden meist zwischen 30 und 60 Prozent der Kosten einer Software für die Qualitätssicherung und somit das Testen der Software verwendet (11). Ohne Softwaretests hätte heutige Software zahlreiche Fehler. Um die Wichtigkeit von Sofwaretests zu unterstreichen sieht z. B. das V-Modell (vgl. Abbildung 1.1) für jeden Entwicklungsschritt am Ende eine entsprechende Testphase vor, woher auch der Name V-Modell stammt.

Nun ist es natürlich sehr aufwändig und daher mit einem vertretbaren Aufwand kaum machbar, jeden Test manuell durchzuführen. Daher wird versucht, möglichst viel zu automatisieren. Vor allem bei Unit-Tests ist dies mithilfe des *XUnit*-Framework sehr einfach möglich. Dabei werden zunächst einzelne Testfälle erstellt und können im Anschluss aufgrund der aktuellen Codebasis jederzeit ausgeführt werden. Automatisierte Tests können auch dazu genutzt werden, um einen einzelnen Test mit verschiedenen Eingaben durchzuführen. Dadurch können verschiedene Eingabeklassen (wie negative oder positive Ganzzahlen) mit sehr geringem Aufwand in einem Test genutzt werden und somit verschiedene Testfälle direkt ausgeführt werden. Durch diese Testautomatisierung können somit zahlreiche Kosten eingespart werden.

Weitere Testframeworks gibt es auch zum Testen von sicherheitskritischen Systemen, welche nicht immer reine Software sein müssen. Dies ist einer der wichtigsten Anwendungsfälle für sogenannte *Model Checker* (MC). *Model Checking* (MC) wird dazu verwendet, um ein Modell auf seine Spezifikation zu testen. Dazu werden abhängig von den aktivierten Komponentenfehlern verschiedene Zustände des Modells vollautomatisch ausgeführt und anhand der dadurch im System auftretenden Fehler entschieden, ob die Spezifikation erfüllt wird (vgl. Abschnitt 2.1) (8, 9).

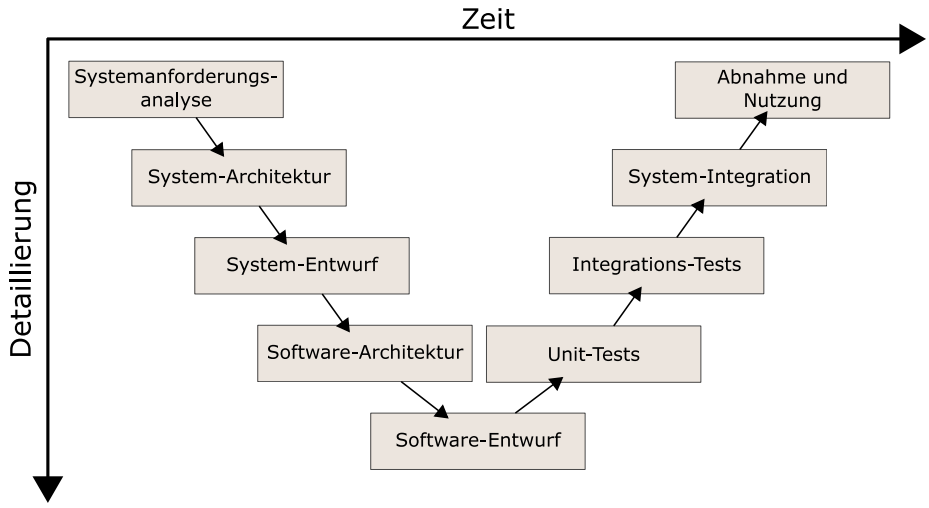

Abbildung 1.1: Aufbau des V-Modells (12)

1.2 Problemstellung

Nun gibt es zahlreiche MC, wie z. B. *LTSmin*[1]. Auch am Institut für Software & Systems Engineering der Universität Augsburg wurde in den letzten Jahren mit S# (sprich *Safety Sharp*) ein entsprechendes Framework zum testen von sicherheitskritischen und selbst-adaptiven Systemen basierend auf dem MC-Ansatz entwickelt (9, 10). Nun ist das aktuelle Vorhaben, mithilfe von S# ein reales Serversystem zu testen, welches bereits als theoretisches Modell basierend auf der ZNN.vom-Fallstudie, bekannt aus der Dissertation von Shang-Wen Cheng (5), implementiert wurde. Als reales System soll nun *Apache Hadoop*[2] getestet werden, welches in der Industrie im Bereich Datenverarbeitung eingesetzt wird. Mit Hadoop ist es möglich, ein Servercluster zu erstellen, auf dem anschließend dafür entwickelte Anwendungen ausgeführt werden und somit große Datenmengen zu verarbeiten. Es soll daher nun getestet und analysiert werden, wie sich Hadoop unter verschiedenen Lastprofilen verhält und dabei bestimmte Fehler auftreten, wenn z. B. einer der Hadoop-Nodes ausfällt.

Bei der ZNN.com-Fallstudie als reines Modell gab es bereits eine ähnliche Aufgabenstellung, welche im Positionspapier (6) genauer erläutert wurde. Das Hauptziel dieses Projektes ist daher nun, anstatt eines reinen Modells ein reales System zu testen. Dafür wird ein modellbasierter Ansatz als Testkonzept genutzt und Hadoop zunächst als Modell nachgebildet. Dieses Modell wird dann dazu genutzt, um ein reales Hadoop-Cluster entsprechend zu steuern und mithilfe des MC-Ansatzes zu testen, wie sich das reale System unter bestimmten Bedingungen verhält und dabei zu ermitteln, wann es nicht mehr funktionsfähig ist.

[1] `http://fmt.cs.utwente.nl/tools/ltsmin/`
[2] `https://hadoop.apache.org/`

1.2.1 Aufbau des Modells

Zunächst muss natürlich erst einmal der Versuchsaufbau selbst in S# modelliert werden. Ein Modell beinhaltet in S# zunächst einmal die Komponenten des Systems und deren Zusammenhänge, also wie die Komponenten miteinander agieren. Wichtig sind in einem S# -Modell aber auch mögliche Komponentenfehler, welche bekannt sind und jederzeit auftreten können. Komponentenfehler werden bereits in der Designphase eines Modells eingearbeitet und können bei der späteren Ausführung flexibel aktiviert und deaktiviert werden, um die Probleme des zu testenden Systems zu ermitteln (vgl. Abschnitt 2.2).

Um nun Hadoop in S# zu modellieren wird zunächst ein Konzept erstellt, in dem ausgearbeitet wird, welche Komponenten und Komponentenfehler relevant sind. Anschließend müssen deren Zusammenhänge und wesentlichen Eigenschaften ausgearbeitet werden und in das Konzept übernommen werden. Sobald das Konzept fertig ausgearbeitet ist, kann das Modell in S# implementiert werden.

1.2.2 Mapping zum realen System

Nachdem die Basis des Modells steht, kann die Funktionalität entwickelt werden. Dazu werden in S# nur Basisfunktionen eingebaut, um mit dem realen System kommunizieren zu können. Dies geschieht mit einer Art Treiber, welcher mithilfe von mehreren SSH-Verbindungen mit dem realen Hadoop-Cluster kommuniziert und so das Mapping zwischen Modell und realem System übernimmt. Jede der SSH-Verbindungen ist dabei nur für einen Einsatzzweck gedacht, sodass es Verbindungen für u. A. folgende Einsatzzwecke gibt:

- Starten von Benchmark-Anwendungen
- Monitoring des realen Cluster
- Injektion von Komponentenfehler

Der Vorteil von mehreren Verbindungen liegt darin, dass jede Verbindung unabhängig ist und nicht auf die Antwort des zuvor gestarteten Programms warten muss. So ist es möglich, mithilfe mehrere Verbindungen mehrere Anwendungen parallel zu starten und jede Rückgabe auszuwerten und währenddessen verschiedene Komponentenfehler zu aktivieren.

1.2.3 Erstellung der Lastprofile

Sobald das Grundmodell steht, können die Testfälle selbst entwickelt werden. Als Testfälle dienen dazu unterschiedliche Lastprofile, um verschiedene Auslastungsgrade und Nutzungsszenarien zu simulieren. Dazu sollen die Lastprofile verschiedene Benchmarks

beinhalten, deren einzelne Anwendungen kombiniert oder alleine auf dem realen System ausgeführt werden:

- Hadoop Mapreduce Examples
- Intel HiBench
- SWIM (Statistical Workload Injector for Mapreduce)

Eine Besonderheit bildet der SWIM-Benchmark, welcher sehr Ressourcenintensiv ist und daher auf einem *Single Node Cluster*, also einem kompletten Hadoop-Cluster auf nur einem Computer, sehr zeitintensiv sein kann. Der Intel HiBench basiert auf einzelnen Bestandteilen der Mapreduce Examples. Die Examples wiederum sind zahlreiche voneinander unabhängige Beispielanwendungen für Hadoop. Dadurch besteht die Möglichkeit, abhängig davon, welche Anwendungen einzeln oder parallel gestartet werden, unterschiedliche Profile zu simulieren. Daher muss zunächst auch geprüft werden, welcher Benchmark welche Möglichkeiten bietet, um die benötigten Testfälle bzw. Lastprofile zu erstellen und so den dynamischen Teil des zu testenden Modells zu erstellen.

1.2.4 Erstellen und Ausführen der Tests

Sobald Modell und Testfälle stehen, kann mit der Erstellung der Tests fortgefahren werden. Die Tests müssen nun so erstellt werden, dass sie sich einerseits auf veränderte Bedingungen des realen Clusters anpassen, aber auch automatisiert die einzelnen Anwendungen der Lastprofile aktivieren und ausführen. Dies schließt auch unterschiedliche Profile für die Aktivierung der Komponentenfehler ein. Zum einen kann nur eine Simulation ohne Fehler gestartet werden, zum anderen aber auch unterschiedliche Komponentenfehler aktiviert werden. Der MC von S# besitzt dazu auch Möglichkeiten, um Komponentenfehler kombiniert auszuführen. Dazu werden basierend auf zuvor definierte *Constraints* Komponentenfehler aktiviert, um so typische Probleme des realen Systems zu simulieren. Basierend darauf wird dann ermittelt, welche Fehler nun im realen System auftauchen.

1.2.5 Evaluierung der Ergebnisse

Je nachdem welche *Constraints* bei der Ausführung genutzt werden, sind nun unterschiedliche Fehler und Daten im realen System ermittelt worden, welche zum Abschluss evaluiert werden müssen. Einige Erwartungen sind da natürlich bereits im Voraus klar: Sollte es zu einem Netzwerk- oder Serverausfall eines Hadoop-Nodes kommen, muss das System dies selbstständig erkennen und die Anwendung an einen anderen Node abgeben. Dabei sollte das System auch erkennen, welche anderen Nodes bereits beschäftigt sind und entsprechend auf dem von Hadoop genutzten *Load Balancer* einen

Node auswählen. Neben einer Fehleranalyse können aber auch die Laufzeiten unter bestimmten Bedingungen analysiert werden.

Kapitel 2

Relevante Technik

2.1 Model Checking

Model Checking (MC) ist eine Möglichkeit, um Systeme zu testen und zu verifizieren. Dazu werden vom *Model Checker* (MC) alle möglichen Systemzustände in einem *Brute-Force*-Ähnlichem Vorgehen getestet und somit alle möglichen Szenarien getestet. Die Anzahl der Zustände kann sehr schnell 10^{120} oder mehr betragen (8, 4).

Ein MC nutzt, wie der Name schon sagt, ein Modell des Systems, um das System zu testen. Wie bei jeder anderen modellbasierten Technik ist daher die Qualität des MC nur so gut wie das darauf zugrunde liegende Modell.

2.2 S#

Wie in Abschnitt 1.2 erwähnt, wird am Lehrstuhl das Framework S# entwickelt. Da es in C# entwickelt wurde und C# auch zum Entwickeln von Modellen und dazugehörigen Testszenarien genutzt wird, können zahlreiche Features des .NET-Frameworks bzw. der Sprache C# im Speziellen genutzt werden. S# vereint dabei die Simulation, die Visualisierung, modellbasierte Tests sowie das MC der Modelle (9, 10). Dadurch können alle Schritte einer vollständigen Analyse inkl. Modellierung direkt im Visual Studio ausgeführt werden und somit auch alle Features der IDE und von .NET, wie z. B. die Debugging-Werkzeuge, genutzt werden. Um das MC durchzuführen, hat S# jedoch einige Einschränkungen, u. A. sind Schleifen und Rekursionen nur eingeschränkt bzw. nicht möglich. Die größte Einschränkung ist allerdings, dass während der Laufzeit keine neuen Objektinstanzen erzeugt werden können, sodass alle benötigten Instanzen bereits während der Initialisierung des Modells erzeugt werden müssen (9).

Um nun ein System testen zu können, muss dieses zunächst mithilfe von C# -Klassen und -Instanzen modelliert werden. Die dafür verwendeten Modelle sind meist stark vereinfacht und bilden nur die wesentlichen Aspekte der realen Systeme ab. Für

```
public class YarnNode : Component
{
  // fault definition, also possible: new PermanentFault()
  public readonly Fault NodeConnectionError = new TransientFault();

  // interaction logic (Members, Properties, Methods...)

  // fault effect
  [FaultEffect(Fault = nameof(NodeConnectionError))]
  internal class NodeConnectionErrorEffect : YarnNode
  {
    // fault effect logic
  }
}
```

Listing 2.1: Grundlegender Aufbau einer S# -Komponente

einen korrekten Test ist es jedoch wichtig, dass das Modell des Systems vergleichbar mit dem echten System ist.

Listing 2.1 zeigt den typischen, grundlegenden Aufbau einer S# -Komponente. Jede Komponente des Modells muss von `Component` erben, um als S# -Komponente definiert zu sein. Jede Komponente kann nun temporäre (`TransientFault`) oder dauerhafte (`PermanentFault`) Komponentenfehler enthalten, welche zunächst innerhalb der Komponente definiert werden. Der Effekt eines Komponentenfehlers wird anschließend in der entsprechenden Unterklasse definiert, welche von der Hauptklasse (hier `YarnNode`) erbt und mithilfe des Attributs `FaultEffect` dem dazugehörigen Komponentenfehler zugeordnet wird (10).

Um die Modelle zu testen, kommt in S# die *Deductive Cause-Consequence Analysis* (DCCA) zum Einsatz. Die DCCA ermöglicht eine vollautomatisch und MC-basierte Sicherheitsanalyse, wodurch selbstständig die Menge der aktivierten Komponentenfehler ermittelt wird, mit denen sich das Gesamtsystem nicht mehr rekonfigurieren kann und somit ausfällt. Je nach Konfiguration können dazu auch Heuristiken genutzt werden, welche die Analyse beschleunigen und genauer machen können (7). Dabei werden die verschiedenen aktivierten Komponentenfehler während der Analyse in tolerierbare und nicht-tolerierbare Fehler unterschieden. Tolerierbare Komponentenfehler werden dazu genutzt, die Grenzen der Selbstkonfiguration des Systems zu ermitteln. Dabei wird für jeden Systemzustand nach einer Rekonfiguration durch die DCCA eine neue Fehlermenge ermittelt, mit der das System gerade noch so lauffähig ist. Das Auftreten eines tolerierbaren Komponentenfehler ist also gleichbedeutend mit einem einfachen Fehler im System, welcher die gesamte Funktionsweise des Systems nicht massiv einschränkt und es sich noch selbst rekonfigurieren kann. Sobald jedoch ein Fehler auftritt, durch den es dem System nicht mehr möglich ist, sich selbst zu rekonfigurieren, wurde ein nicht-tolerierbarer Fehler gefunden, durch den das System nicht mehr funktionsfähig ist (9).

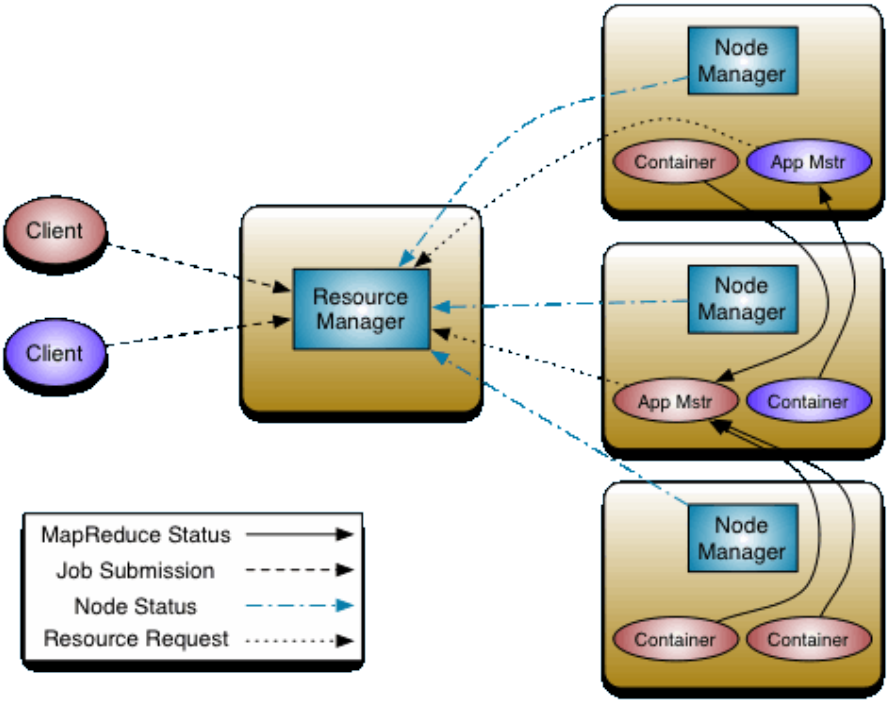

Abbildung 2.1: Architektur von YARN (1)

2.3 Apache Hadoop

Apache Hadoop ist ein Open-Source-Software-Projekt, in dem Software für verteilte Systeme entwickelt wird. Hadoop wird von der *Apache Foundation* entwickelt und bietet verschiedene Komponenten an, welche vollständig skalierbar sind, von einer einfachen Installation auf einem PC bis hin zu einer Installation über mehrere Server in einem Serverzentrum. Hadoop besteht hauptsächlich aus folgenden Kernmodulen (3):

Hadoop Common Gemeinsam genutzte Kernkomponenten

Hadoop Distributed File System Kurz HDFS, Verteiltes Dateisystem

Hadoop YARN Framework zur Verteilung und Ausführung von Tasks und das dazugehörige Ressourcen-Management

Hadoop MapReduce YARN-Basiertes System zum Verarbeiten von großen Datenmengen

Hadoop ermöglicht es dadurch, sehr einfach mit Tasks umzugehen, welche große Datenmengen verarbeiten. Da es für Hadoop nicht relevant ist, auf wie vielen Servern es läuft, kann es beliebig skaliert werden, wodurch entsprechend viele Ressourcen zur Bearbeitung und Speicherung von großen Datenmengen zur Verfügung stehen können.

Die Kernidee der Architektur von YARN ist die Trennung vom Ressourcenmanagement und Scheduling. Dazu besitzt der Master den *ResourceManager*, welcher für das gesamte System zuständig ist und die Anwendungen im System überwacht. Er besteht

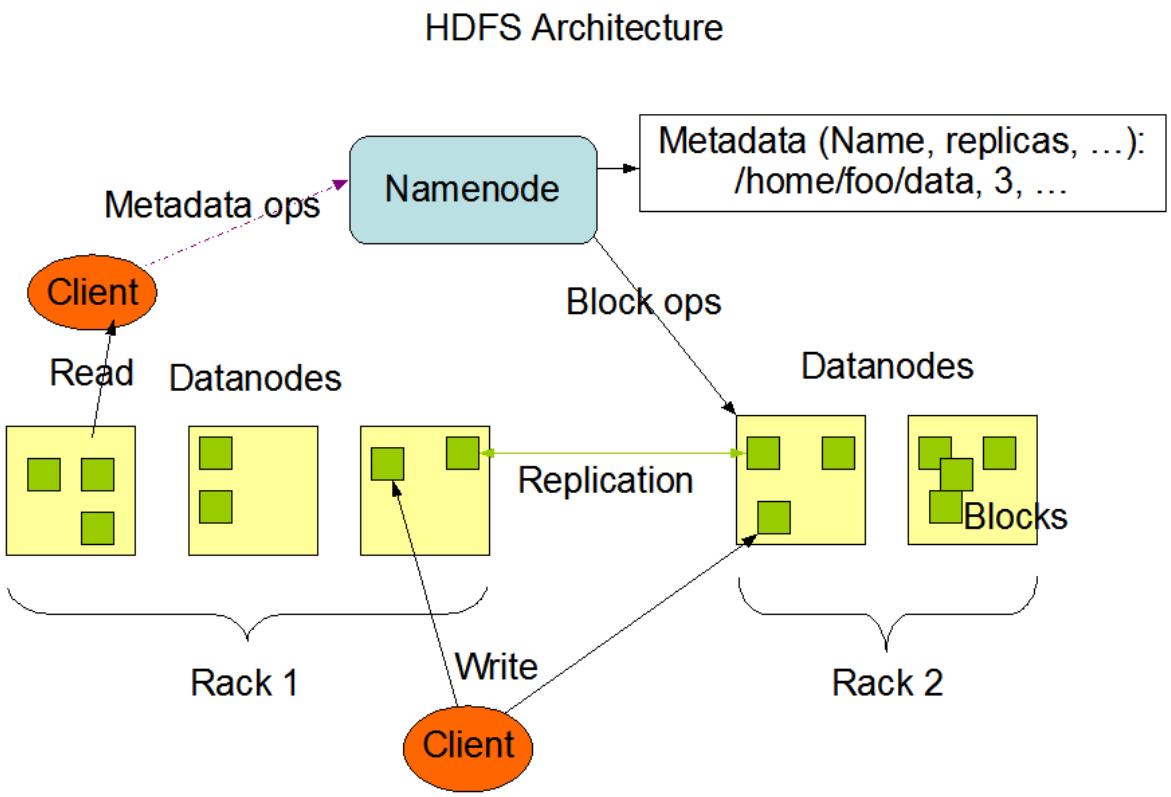

Abbildung 2.2: Architektur des HDFS (2)

aus zwei Kernkomponenten, dem *ApplicationsManager* und dem *Scheduler*. Der ApplicatonsManager ist für die Annahme und Ausführung von einzelnen Anwendungen zuständig, denen der Scheduler die dafür notwendigen Ressourcen zuteilt und überwacht. Für jeden Slave-*Node* im Hadoop-System gibt es dazu einen *NodeManager*, welcher die lokalen Ressourcen des Nodes überwacht und dem ResourceManager mitteilt. Jede Anwendung besitzt jeweils einen eigenen *ApplicationMaster*, welcher für das Monitoring und die Kommunikation mit dem ResourceManager und NodeManager zuständig ist und die dazu notwendigen Informationen bereit stellt. Jede YARN-Anwendung bzw. Job oder Task besteht zudem aus einem oder mehreren *Containern*, in welchen die Tasks ausgeführt werden. Jeder Bestandteil eines Tasks kann auf jedem beliebigen Node ausgeführt werden (1).

Das HDFS basiert auf der gleichen Architektur wie YARN und besitzt ebenfalls einen Master und mehrere Slaves, welches in der Regel die gleichen Nodes sind wie bei YARN sind. Der *NameNode* ist als Master für die Verwaltung des Dateisystems zuständig und reguliert den Zugriff auf die darauf gespeicherten Daten. Die Daten selbst werden in mehrere Blöcke aufgeteilt auf den *DataNodes* gespeichert. Um den Zugriff auf die Daten im Falle eines Node-Ausfalls zu gewährleisten, wird jeder Block auf anderen Nodes repliziert. Dateioperationen (wie Öffnen oder Schließen) werden direkt auf den DataNodes ausgeführt, sie sind darüber hinaus auch dafür verantwortlich, dass Clients die Daten lesen oder beschreiben können (2).

Literaturverzeichnis

[1] Apache Software Foundation: *Apache Hadoop YARN*. `http://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/YARN.html`. Version: 1 2016

[2] Apache Software Foundation: *HDFS Architecture*. `http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html`. Version: 1 2016

[3] Apache Software Foundation: *Welcome to Apache™ Hadoop®!* `http://hadoop.apache.org/`. Version: 12 2017

[4] Baier, J.-P. Christel; K. Christel; Katoen: *Principles of model checking*. MIT Press, 2008. – ISBN 978–0–262–02649

[5] Cheng, S.-W. : *Rainbow: Cost-effective Software Architecture-based Self-adaptation*. Pittsburgh, PA, USA, Carnegie Mellon University, Diss., 2008. – AAI3305807

[6] Eberhardinger, B. ; Habermaier, A. ; Reif, W. : Toward Adaptive, Self-Aware Test Automation. In: *2017 IEEE/ACM 12th International Workshop on Automation of Software Testing (AST)*, 2017, S. 34–37

[7] Eberhardinger, B. ; Habermaier, A. ; Seebach, H. ; Reif, W. : Back-to-Back Testing of Self-organization Mechanisms. In: Wotawa, F. (Hrsg.) ; Nica, M. (Hrsg.) ; Kushik, N. (Hrsg.): *Testing Software and Systems: 28th IFIP WG 6.1 International Conference, ICTSS 2016, Graz, Austria, October 17-19, 2016, Proceedings*. Springer International Publishing. – ISBN 978–3–319–47443–4, 18–35

[8] Grumberg, O. ; Clarke, E. ; Peled, D. : Model checking. (1999)

[9] Habermaier, A. ; Eberhardinger, B. ; Seebach, H. ; Leupolz, J. ; Reif, W. : Runtime Model-Based Safety Analysis of Self-Organizing Systems with S#. In: *2015 IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, 2015, S. 128–133

[10] Habermaier, A. ; Leupolz, J. ; Reif, W. : Unified Simulation, Visualization, and Formal Analysis of Safety-Critical Systems with S#. In: Beek, M. H. (Hrsg.) ; Gnesi, S. (Hrsg.) ; Knapp, A. (Hrsg.): *Critical Systems: Formal Methods and Automated Verification: Joint 21st International Workshop on Formal Methods for Industrial Critical Systems and 16th International Workshop on Automated Verification of Critical Systems, FMICS-AVoCS 2016, Pisa, Italy, September 26-28, 2016, Proceedings*. Springer International Publishing. – ISBN 978–3–319–45943–1, 150–167

[11] Polo, M. ; Reales, P. ; Piattini, M. ; Ebert, C. : Test Automation. In: *IEEE Software* 30 (2013), Jan, Nr. 1, S. 84–89. http://dx.doi.org/10.1109/MS.2013.15. – DOI 10.1109/MS.2013.15. – ISSN 0740–7459

[12] Pätzold, M. ; Seyfert, S. : *V-Modell*. https://commons.wikimedia.org/wiki/File:V-Modell.svg. Version: Januar 2010. – Lizenz: CC-BY-SA 3.0