Datenmanagement in der Cloud für den Bereich Simulationen und Wissenschaftliches Rechnen

Peter Reimann, Tim Waizenegger, Matthias Wieland und Holger Schwarz

Institut für Parallele und Verteilte Systeme, Universität Stuttgart
Vorname.Nachname@ipvs.uni-stuttgart.de

Abstract: Für Organisationen, die Simulationen nicht als ihr Kerngeschäft verstehen und sie daher eher sporadisch durchführen, lohnt sich der Betrieb einer eigenen Recheninfrastruktur nur selten. Dies betrifft z. B. kleine und mittlere Unternehmen sowie einige wissenschaftliche Institutionen. Besserung können öffentliche Cloud-Infrastrukturen als Plattform für die Ausführung von Simulationen verschaffen. Das Datenmanagement in der Cloud ist aber speziell für den Bereich Simulationen noch weitgehend unerforscht. In diesem Beitrag identifizieren wir daher noch offene Fragestellungen bzgl. des Datenmanagements von Simulationen in der Cloud. Dies betrifft vor allem die Datenbereitstellung und inwieweit nutzer- und simulationsspezifische Anforderungen an das Datenmanagement in der Cloud eingehalten werden können. Wir untersuchen Technologien, welche sich diesen Fragestellungen widmen, und diskutieren, ob und wie sie in der Cloud sowie für Simulationen einsetzbar sind. Weiterhin skizzieren wir wichtige zukünftige Forschungsthemen.

1 Einleitung

Einige Organisationen, z. B. kleine und mittlere Unternehmen oder manche wissenschaftliche Institutionen, führen Simulationen eher sporadisch durch, sodass sich der Betrieb einer eigenen Recheninfrastruktur nur selten lohnt. Eine Alternative sind öffentliche Cloud-Infrastrukturen als Plattform für die Ausführung von Simulationen. Ein Kernaspekt von Cloud-Infrastrukturen ist, dass Nutzer der Cloud die für ihre Anwendungen benötigten Ressourcen nicht explizit vorhalten müssen, sondern sie dynamisch und schnell vom Cloud-Anbieter bereitgestellt bekommen. Ebenso können die Ressourcen bedarfsgerecht und elastisch an die individuell benötigte Leistungsfähigkeit angepasst werden, von sehr kleinen Berechnungen bis hin zu komplexen Anwendungen mit hoher Parallelität. Gerade wenn Simulationsberechnungen nur sporadisch durchgeführt werden, bieten diese bedarfsgerechten, dynamischen und kostengünstigen Provisionierungsmöglichkeiten einen erheblichen Vorteil gegenüber einer eigenen Recheninfrastruktur.

Cloud-Anbieter stellen ihre Dienste in verschiedenen Abstraktionsstufen bereit, die in Infrastructure-, Platform- und Software-as-a-Service eingeteilt werden (IaaS, PaaS, SaaS) [MG11]. Auf der untersten Ebene bieten sie mit IaaS-Diensten virtualisierte Hardware wie Rechenleistung, Speicher und Netzwerk an, die universell zum Betrieb beliebiger Anwendungen dienen. In PaaS-Angeboten werden Plattform- oder Middleware-Dienste zur

Verfügung gestellt, z. B. Datenbankdienste oder Plattformen für Web-Anwendungen. Die Kategorie SaaS umfasst vollständige (Anwender-)Software, deren Betrieb und Wartung der Cloud-Anbieter übernimmt. Zur Provisionierung und Verwaltung solcher Dienste gibt es bereits eine Reihe von Lösungen. Ein repräsentatives Beispiel ist der OASIS-Standard *Topology and Orchestration Specification for Cloud Applications* (TOSCA) [OA13], der auch schon im Bereich von Simulationen erfolgreich Einzug gehalten hat [VHKL13]. Die spezifischen Fragestellungen, die sich Cloud-Nutzer bzgl. des Datenmanagements ihrer Simulationsanwendungen in der Cloud stellen, blieben bisher aber weitgehend unbeachtet:

- Wie können Nutzer Eingabedaten für eine Simulation in der Cloud bereitstellen, und wie werden ihnen die Ergebnisdaten wieder zur Verfügung gestellt?
- Wie können Nutzer sicherstellen, dass alle ihre individuellen nicht-funktionalen Anforderungen an das Datenmanagement in der Cloud eingehalten werden?

Bleiben diese Fragen unbeantwortet, kann sich dies negativ auf die Nutzerakzeptanz von Cloud-Diensten für Simulationen auswirken. Dieser Beitrag liefert dazu folgende Mehrwerte und ist wie folgt aufgebaut:

Abschnitt 2 leitet – anhand einer Beispielsimulation von Strukturänderungen in Knochen [Kr13] – die wichtigsten Anforderungen her, die aus Nutzersicht für das Datenmanagement in Simulationen beachtet werden müssen. In Abschnitt 3 wird der Einsatz des OASIS-Standards TOSCA für die Bereitstellung von Simulationssoftware beschrieben. Abschnitt 4 diskutiert die erste offene Fragestellung bzgl. der Bereitstellung von Eingabedaten und der Rückgabe der Ergebnisdaten von Simulationsabläufen. Zudem wird ein Vorschlag für die Umsetzung dieser Datenbereitstellung und Datenrückgabe vorgestellt, der auf dem SIMPL-Rahmenwerk (*SimTech, Information Managemenent, Processes, and Languages*) basiert [RS13, Re11]. Abschnitt 5 widmet sich Technologien zur Einhaltung der in Abschnitt 2 identifizierten Nutzeranforderungen. Dabei diskutieren wir, inwieweit diese Technologien in der Cloud einsetzbar sind, und skizzieren wichtige zukünftige Forschungsthemen. Abschnitt 6 fasst abschließend die wichtigsten Aspekte zusammen.

2 Datenmanagement in Simulationen

Die Untersuchung komplexer Probleme erfordert häufig die Kopplung von Simulationsmodellen verschiedener wissenschaftlicher Anwendungsgebiete. Als Beispiel betrachten wir eine Simulation von zeitabhängigen Strukturänderungen in Knochen, die z. B. bei Heilungsprozessen nach Knochenbrüchen relevant ist [Kr13]. Abb. 1 zeigt, wie diese Simulation Modelle aus den Anwendungsgebieten Biomechanik und Systembiologie koppelt. Das biomechanische Simulationsmodell beschreibt das Verhalten von Knochen auf einer makroskopischen Gewebeebene. Das auf der Finite-Elemente-Methode (FEM) basierende Pandas-Rahmenwerk[1] bietet hierzu eine numerische Implementierung. Es wandelt auf Knochen wirkende externe Belastungen in eine charakteristische Lösung für die interne

[1] http://www.mechbau.uni-stuttgart.de/pandas/index.php

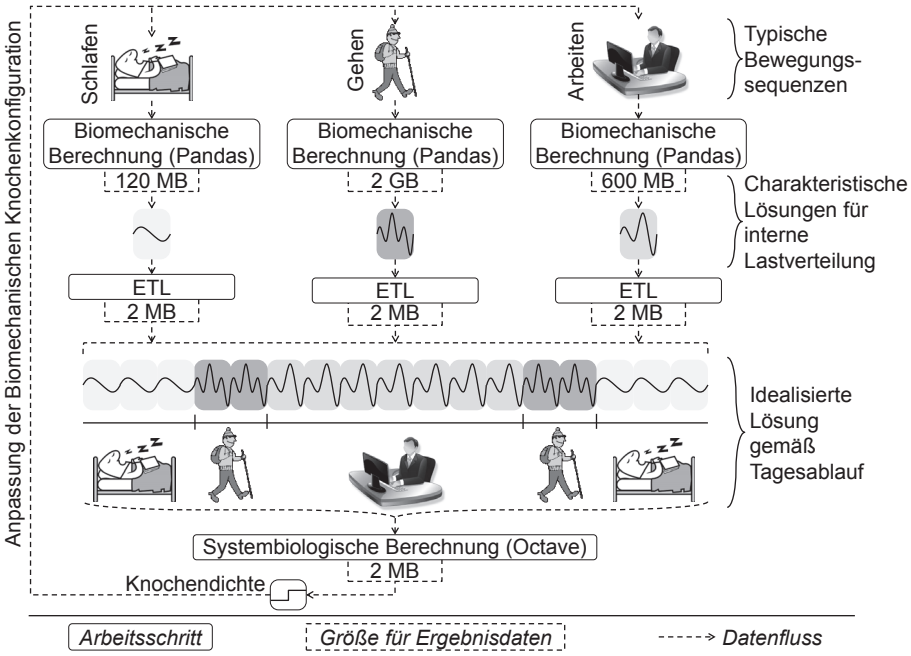


Abbildung 1: Prozess für die gekoppelte Simulation von Strukturänderungen in Knochen

Lastverteilung innerhalb des Knochengewebes um. Dies erfolgt jeweils unabhängig für typische Bewegungssequenzen einer Person. Als Ergebnis speichert Pandas pro Punkt des FE-Gitters und pro Zeitschritt bis zu 20 Variablen in einer Datenbank. Abhängig von der Anzahl der Gitterpunkte und der Anzahl der Zeitschritte ergibt dies ein Datenvolumen zwischen 100 MB und mehreren GB für jede Bewegungssequenz.

Die von Pandas berechneten Lösungen berücksichtigen allerdings keine zellularen Reaktionen im Knochengewebe. Genau hier kommt die systembiologische Simulation ins Spiel, die die Bildung oder den Abbau des Gewebes auf Basis der Interaktion von Zellen bestimmt. Dies kann mithilfe der Rechenumgebung GNU Octave[2] umgesetzt werden. Vorbereitende ETL-Prozesse (Extraktion, Transformation, Laden) filtern passende Daten aus der Datenbank von Pandas und aggregieren sie anschließend über alle Zeitschritte. Die einzelnen charakteristischen Lösungen von Pandas werden zudem zu einer idealisierten Lösung zusammengefasst, die dem approximierten Tagesablauf der Person entspricht. Diese Lösung wird schließlich in CSV-basierten Dateien (*comma-separated values*) gespeichert, sodass Octave sie einlesen kann. Die systembiologische Simulation rechnet die genaue Knochendichte nach Ablauf eines Tages aus. Diese wird genutzt, um die Konfiguration des biomechanischen Modells anzupassen. Der gesamte Prozess wiederholt sich, bis die vom Anwender vorgegebene Anzahl an Tagen berechnet wurde.

[2] http://www.gnu.org/software/octave/

2.1 Anforderungen an das Datenmanagement

Für das Datenmanagement in Simulationen ergeben sich eine Reihe spezifischer Anforderungen, welche für die an den Simulationsergebnissen interessierten Nutzer wichtig sind. Um diese Anforderungen zu identifizieren, haben wir mehrere Szenarien aus der Literatur (siehe z. B. [SR09, TDG07]), verschiedene Simulationssoftware wie Pandas oder Octave sowie reale Simulationen analysiert. Neben der in Abschnitt 2 vorgestellten Simulation gehören hierzu die von Fehr et al. und Rommel et al. beschriebenen Beispiele [FE11, RK11]. Nachfolgend diskutieren wir die aus Nutzersicht wichtigsten Anforderungen.

Datensicherheit: Im Unternehmenskontext beschreiben Daten häufig schutzwürdiges geistiges Eigentum. Weiterhin müssen zahlreiche gesetzliche Bestimmungen eingehalten werden, welche die Offenlegung von Informationen regulieren. Wissenschaftliche Institutionen möchten z. B. Forschungsergebnisse nicht vor einer Publikation der Arbeit offenlegen oder im Fall von Studien die persönlichen Daten der Teilnehmer schützen. Dies alles macht adäquate *Vorkehrungen zur Datensicherheit* notwendig.

Datenqualität: Mathematische Simulationsmodelle sind lediglich Abbildungen der realen Probleme, die untersucht werden sollen. Die numerische Implementierung dieser Modelle bringt häufig weitere Approximationen und damit Ungenauigkeiten mit ein. Dies impliziert hohe Anforderungen an die *Qualität von Simulationsergebnissen* [Re14].

Effizienz und Optimierung: Die Gesamtgröße der in einzelnen Simulationsläufen involvierten Daten kann zwischen wenigen 100 KB und einigen TB liegen. Die Größe und Komplexität der Daten sind entscheidende Faktoren für die Ausführungszeit von Simulationen. Dies führt zwangsläufig zu Anforderungen bzgl. der *Effizienz* des Datenmanagements und bzgl. der Unterstützung entsprechender *Optimierungsmöglichkeiten* [Vr07].

Reproduzierbarkeit und Nachvollziehbarkeit: Ein weiterer wichtiger Aspekt ist die Sicherstellung der *Reproduzierbarkeit und Nachvollziehbarkeit einer Simulation und ihrer Ergebnisse* [Fr08]. In Bezug auf das Datenmanagement muss hierbei z. B. dafür Sorge getragen werden, dass sämtliche Ergebnisdaten einer Simulation nie verloren gehen und sie dem Nutzer nach den Berechnungen wieder zur Verfügung gestellt werden. Gleiches gilt auch für Log- oder Provenance-Informationen [CVu12, Fr08], die während der Simulation gesammelt werden. Provenance-Informationen beschreiben den detaillierten Simulationsablauf, z. B. die Herkunft der Eingabedaten und wie diese weiterverarbeitet werden.

3 Softwarebereitstellung in der Cloud

Die manuelle Installation und Konfiguration einer Simulationssoftware in einer Cloud ist sehr zeitaufwändig. Um diese Softwarebereitstellung zu automatisieren, kann der OASIS-Standard TOSCA eingesetzt werden [OA13]. TOSCA erlaubt es, beliebige Anwendungen mit geringem Aufwand interoperabel in verschiedenen Cloud-Umgebungen bereitzustellen. Den Kern von TOSCA bildet eine Modellierungssprache, welche die Definition einer Anwendung und ihres Aufbaus in einer sog. Diensttopologie ermöglicht. Diese Topologie besteht aus Anwendungskomponenten und deren Beziehungen. Des weiteren können

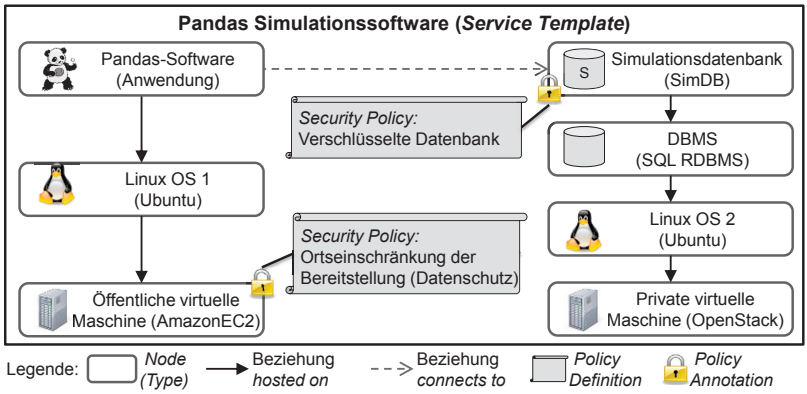


Abbildung 2: Beispiel einer TOSCA Diensttopologie für die Simulationssoftware Pandas

Deployment- und Management-Pläne als Workflows erstellt werden, um die Installation und die Verwaltung einer Anwendung zu automatisieren. Nachfolgend illustrieren wir beispielhaft, wie TOSCA zur Bereitstellung der in Abschnitt 2 betrachteten Simulationssoftware Pandas verwendet werden kann.

3.1 Bereitstellung der Simulationssoftware Pandas

Abb. 2 stellt die mit TOSCA definierte Topologie der Pandas-Software dar. Die Topologie besteht aus zwei Server-Stacks: einer für die Simulationssoftware (links) und einer für die Datenhaltung (rechts). Der Simulationssoftware-Stack wird auf einer virtuellen Maschine in einer öffentlichen Cloud-Umgebung betrieben. Dies bietet einen Kostenvorteil, da die Nutzer keine eigene Infrastruktur einrichten und verwalten müssen, sondern diese nur bedarfsgerecht anmieten, wenn sie tatsächlich eine Simulation durchführen wollen. Weiterhin kann die Infrastruktur jederzeit skaliert werden, je nachdem wie rechenintensiv die Simulation ist. Auf dieser öffentlichen virtuellen Maschine ist ein Linux-Betriebssystem installiert (dargestellt durch die Beziehung *hostet-on*), auf welchem wiederum die Pandas-Software läuft. Die Pandas-Software ist mit dem Datenbank-Stack über eine Beziehung *connects-to* verbunden, d. h. die beteiligten Komponenten kommunizieren über ein Netzwerk. Im Gegensatz zum Software-Stack nutzt der Datenbank-Stack eine private virtuelle Maschine, welche z. B. auf einem OpenStack-System des Nutzers betrieben wird. Dadurch kann die Datensicherheit auf hohem Niveau gehalten werden, da die Datenspeicherung nicht öffentlich erfolgt. Auf dieser virtuellen Maschine ist in einem Linux-Betriebssystem ein SQL-Datenbanksystem installiert, das die Simulationsdatenbank verwaltet.

Prinzipiell bietet TOSCA noch vielfältigere Modellierungsmöglichkeiten. So können beispielsweise an jeder Anwendungskomponente und an der gesamten Topologie über Policies frei definierbare Anforderungen annotiert werden, welche bei der Installation, Konfiguration und während des Betriebs der Anwendung eingehalten werden müssen. Im betrachteten

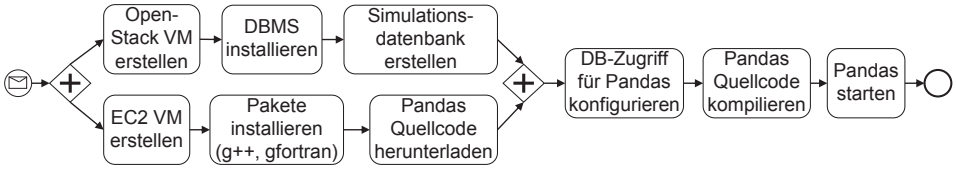


Abbildung 3: TOSCA Deployment-Plan für Pandas

Anwendungsfall wird als Sicherheitsanforderung z. B. der Ort eingeschränkt, an dem der Server für den linken Software-Stack läuft. Weiterhin definiert die Topologie, dass die Simulationsdatenbank ihre Daten verschlüsseln muss.

Um eine mit TOSCA definierte Anwendung in der Cloud aufzusetzen, können verschiedene Laufzeitumgebungen für TOSCA eingesetzt werden. Ein Beispiel ist OpenTOSCA, das als Eingabe ein *Cloud Service Archive* (CSAR) benötigt, welches zusätzlich zur Diensttopologie alle für die Anwendung nötigen Artefakte und Softwarepakete enthält [Bi13].

3.2 Plan-basierte Bereitstellung und Verwaltung

Ein weiterer zentraler Teil von TOSCA sind Pläne: *Deployment-Pläne* nehmen die Installation von Anwendungen vor, während *Management-Pläne* Anwendungen zu deren Laufzeit verwalten und überwachen. Abb. 3 illustriert einen Deployment-Plan für die Simulationssoftware Pandas. Dieser besteht aus zwei parallel ablaufenden Pfaden, welche die beiden in Abb. 2 gezeigten Stacks aufsetzen. Im oberen Pfad erstellt der Plan zunächst eine virtuelle Maschine auf dem privaten OpenStack-System, installiert darauf ein Datenbanksystem und erzeugt die Simulationsdatenbank. Unten erstellt er analog eine virtuelle Maschine auf Amazon EC2, installiert die für Pandas notwendigen Linux-Pakete und lädt den Pandas-Quellcode von einem Repository herunter. Danach wird die parallele Verarbeitung synchronisiert, damit die Verknüpfung der beiden Stacks über die Konfiguration des DB-Zugriffs erfolgen kann. Weiterhin muss der Pandas-Quellcode kompiliert und Pandas anschließend gestartet werden, sodass es bereit ist, die nachfolgenden Simulationsberechnungen durchzuführen. Dies kann durch Management-Pläne erfolgen, was im nächsten Abschnitt anhand der Datenbereitstellung und Datenrückgabe näher erläutert wird.

4 Datenbereitstellung in der Cloud

Eine wesentliche Herausforderung an die Datenbereitstellung und Datenrückgabe in der Cloud ergibt sich aus dem Wunsch, Simulationen sowie die heterogenen Datenlandschaften aus verschiedenen Anwendungsgebieten zu koppeln (siehe Abschnitt 2). Um eine nahtlose Kopplung zu ermöglichen, muss eine entsprechende Lösung hinreichend *generisch* sein. Dies bedeutet, sie muss insbesondere alle von den verschiedenen Anwendungsgebieten

und ihrer Nutzer benötigten Datenformate und Datenmanagementoperationen unterstützen. Ein weiterer wichtiger Aspekt ist, dass die an den Simulationsergebnissen interessierten Wissenschaftler ihre Simulationen häufig selbst implementieren und dabei auch einen Großteil des Datenmanagements umsetzen müssen. Um Wissenschaftler dabei nicht mit komplexen Implementierungsdetails dieses Datenmanagements zu überfordern, sollte eine Lösung für die Datenbereitstellung eine geeignete *Abstraktionsunterstützung* bieten [RS13].

Da TOSCA eine hohe Modellierungsmächtigkeit für die Softwarebereitstellung bietet, ist zur Erreichung einer generischen Lösung eine Integration der Datenbereitstellung in TOSCA-Plänen ratsam. Bisher kann die Datenbereitstellung in TOSCA-Plänen nur durch proprietäre Services umgesetzt werden, welche wiederum keine generische Lösung darstellen [OA13]. In diesem Abschnitt diskutieren wir zunächst mögliche Alternativen für die Umsetzung der Datenbereitstellung. Als Ergebnis dieser Diskussion stellt sich schließlich heraus, dass das SIMPL-Rahmenwerk (*SimTech, Information Managemenent, Processes, and Languages*) den oben beschriebenen Herausforderungen an eine generische Lösung und an eine geeignete Abstraktionsunterstützung am besten begegnet [RS13, Re11]. Folglich erläutern wir abschließend, wie SIMPL mit TOSCA integriert werden kann.

4.1 Diskussion von Alternativen für die Datenbereitstellung

Standard-Tools für die Definitionen und Ausführung von ETL-Prozessen, wie z. B. die von IBM[3] , Pentaho[4] und Talend[5] angebotenen Lösungen, bieten zwar vielfältige und daher in Simulationen generisch einsetzbare Möglichkeiten zur Umsetzung von Datenmanagementoperationen. Vertrauen Wissenschaftler aber ausschließlich diesen ETL-Tools, müssen sie umfangreiche Implementierungsdetails spezifizieren. Zum Beispiel betrifft dies die in Abschnitt 2 angesprochenen ETL-Prozesse zum Filtern, Aggregieren und Transformieren der Ergebnisdaten von Pandas. Hierbei müssen Wissenschaftler komplexe SQL-Anfragen definieren oder sogar Skript- bzw. Programmiersprachen zur Implementierung des Datenmanagements einsetzen [RS13]. Wissenschaftler besitzen zwar eine hohe Expertise in ihrem Anwendungsbereich der Simulation, weisen aber i. d. R. eingeschränkte Fähigkeiten bzgl. dieser Anfrage- oder Programmiersprachen auf. Daher bieten solche ETL-Tools keine für Wissenschaftler adäquate Abstraktionsunterstützung.

Eine Alternative zur Umsetzung der Datenbereitstellung in TOSCA-Plänen bieten die Workflow-Produkte von IBM, Microsoft und Oracle mit sog. SQL-Aktivitäten [Vr08]. Eine solche Aktivität beinhaltet eine SQL-Anweisung und sendet diese Anweisung bei der Ausführung der Aktivität an ein externes Datenbanksystem. Die Einschränkung auf SQL sorgt aber dafür, dass diese Lösung in vielen Simulationen nicht einsetzbar ist, insbesondere wenn wie bei der in Abschnitt 2 beschriebenen Simulation auch Zugriffe auf Dateien nötig werden. Das wissenschaftliche Workflow-System Kepler bietet ähnliche Aktivitäten für den Zugriff auf SQL-Datenbanksysteme, Dateisysteme und Sensornetze [Lu06]. Allerdings gibt es für jeden Typ von Datenressource und sogar für verschiedene Typen von Datenmana-

[3] http://www.ibm.com/software/data/integration/info_server/
[4] http://www.pentaho.de/
[5] http://de.talend.com/

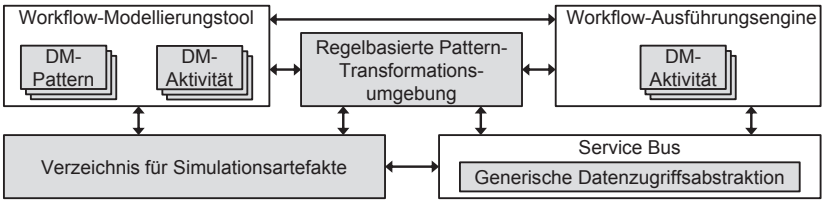


Abbildung 4: Zentrale Komponenten des SIMPL-Rahmenwerks, vgl. [RS13, Re11]

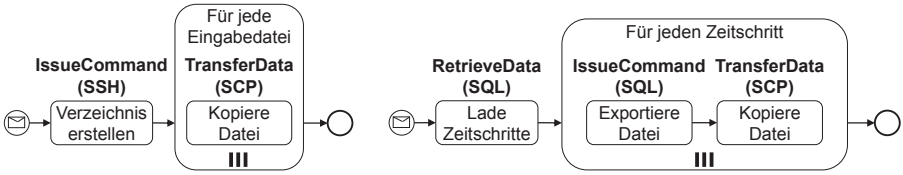
gementoperationen ganz unterschiedliche proprietäre Aktivitätstypen. Die Entwickler von Kepler haben selbst eingeräumt, dass dies keine generische Lösung darstellt und dass die große Anzahl an Aktivitätstypen die Workflowmodellierer sogar überfordern kann [Ba10].

Im Gegensatz zu den bisher diskutierten Ansätzen bietet das SIMPL-Rahmenwerk eine generische sowie einfach zu benutzende Lösung für die Datenbereitstellung [RS13, Re11]. Abb. 4 zeigt, wie SIMPL die Architektur eines Workflowsystems erweitert. Der Service Bus stellt Workflowmodellierern über eine *generische Datenzugriffsabstraktion* einen einheitlichen Zugriff auf beliebige externe Datenressourcen zur Verfügung [Re11]. Das Workflow-Modellierungstool und die Workflow-Ausführungsengine bieten zudem eine Unterstützung für ebenso generische Datenmanagementaktivitäten (*DM-Aktivitäten*). Die Aktivitäten können beliebigen externen Datenressourcen zugeordnet werden und beliebige Befehle in deren Befehlssprachen beinhalten – also nicht nur SQL-Anweisungen, sondern z. B. auch Shell-Kommandos für den Zugriff auf Dateien. Als weiterführende Abstraktionsunterstützung und als Alleinstellungsmerkmal gegenüber anderen Ansätzen ermöglicht eine Erweiterung des Workflow-Modellierungstools die Nutzung von Datenmanagementpatterns (*DM-Patterns*) als Workflow-Bausteine [RS13]. Jedes Pattern fasst mehrere feingranulare Workflow-Schritte zusammen, was bereits die Anzahl der für Workflowmodellierer sichtbaren Schritte reduziert. Weiterhin müssen diese nur wenige abstrakte Parameterwerte für die Patterns festlegen, anstatt komplexe Implementierungsdetails zu spezifizieren. Die *regelbasierte Patterntransformationsumgebung* bildet solche parametrisierten Patterns mithilfe von Metadaten aus dem *Verzeichnis für Simulationsartefakte* auf ausführbare Workflowfragmente oder Services ab [RS13].

4.2 Integration von SIMPL mit TOSCA

Um die Datenbereitstellung und Datenrückgabe für Simulationen in der Cloud umzusetzen, können die SIMPL DM-Aktivitäten und die Patterns in TOSCA-Pläne integriert werden. Dazu muss lediglich die Workflowausführungsengine in OpenTOSCA [Bi13] um die DM-Aktivitäten erweitert werden. Alle anderen Komponenten des SIMPL-Rahmenwerks können über ihre Service-Schnittstellen nahtlos an OpenTOSCA angebunden werden.

Abb. 5 zeigt die mit SIMPL DM-Aktivitäten angereicherten Management-Pläne für die Datenbereitstellung bzw. Datenrückgabe der Simulationssoftware Pandas. Der Plan für



(a) Datenbereitstellung (b) Datenrückgabe

Abbildung 5: TOSCA-Pläne mit SIMPL-Aktivitäten zur Datenbereitstellung und -rückgabe für Pandas

die Datenbereitstellung wird nach dem in Abb. 3 gezeigten Deployment-Plan ausgeführt. Er erzeugt zunächst über eine IssueCommand-Aktivität ein Verzeichnis auf der virtuellen Maschine von Pandas, in das später die Eingabedateien kopiert werden können. Dazu setzt die Aktivität über eine SSH-Schnittstelle ein entsprechendes Shell-Kommando an die virtuelle Maschine von Pandas ab. Anschließend kopiert der Plan über eine parallele Schleife und über eine darin eingebettete TransferData-Aktivität jede relevante Eingabedatei mithilfe eines Secure Copy (SCP) Befehls in das zuvor erzeugte Verzeichnis. Der Plan für die Datenrückgabe wird nach den Simulationsberechnungen ausgeführt. Eine gängige Praxis ist, für jeden Zeitschritt der biomechanischen Simulation eine Datei aus der Pandas-Datenbank zu exportieren und diese Datei auf einen anderen Rechner zu kopieren. Dort kann z. B. eine entsprechende Zeitreihe visualisiert oder analysiert werden. Der Plan lädt zunächst über eine RetrieveData-Aktivität mit eingebetteter SQL-Anweisung eine Liste aller Zeitschritte aus der Datenbank. Die nachfolgende Schleife iteriert über diese Liste und exportiert für jeden Zeitschritt die passende Datei mithilfe einer IssueCommand-Aktivität und einer darin eingebetteten SQL Export-Anweisung. Schließlich kopiert eine TransferData-Aktivität diese Datei auf den Zielrechner, womit die Datenrückgabe abgeschlossen ist.

5 Einhaltung nicht-funktionaler Anforderungen

Die formale Beschreibung von Anforderungen wird in TOSCA über Policies abgedeckt (siehe Abschnitt 3). Verwenden Cloud-Nutzer TOSCA, können sie mit dem Policy4TOSCA-Rahmenwerk beliebige Anforderungen definieren, insbesondere auch die in Abschnitt 2.1 diskutierten [Wa13]. Die Policies werden dabei entweder global an einer gesamten Dienst-topologie oder lokal an einzelne ihrer Komponenten angehängt. Anschließend müssen die über Policies definierten Anforderungen mit geeigneten technischen Mitteln umgesetzt werden, was für jede der in Abschnitt 2.1 identifizierten Anforderungen sehr unterschiedlich erfolgen kann. Daher betrachten wir im Folgenden vor allem diese technische Umsetzung der Policies. Für die einzelnen in Abschnitt 2.1 identifizierten Anforderungen diskutieren wir hierzu jeweils den aktuellen Forschungsstand rund um Simulationen und/oder Cloud und skizzieren einige offene Forschungsthemen.

Datensicherheit: Die Einhaltung der von den Policies vorgegebenen Anforderungen an Datensicherheit wird durch die genutzte TOSCA-Laufzeitumgebung umgesetzt. Dies kann

von verschiedenen Umgebungen auf unterschiedliche Art erfolgen. Für Policies, die lokal an einzelne Dienstkomponenten angehängt sind, wählt OpenTOSCA entsprechende Implementierungen und Konfigurationen der Komponenten aus, für welche die Policies gelten [Bi13]. Für global an eine Diensttopologie annotierte Policies ändert OpenTOSCA die Deployment-Pläne, sodass die Policies zur Deployment-Zeit beachtet werden.

Datenqualität: Reiter et al. beschreiben ein Rahmenwerk zur Definition und Evaluierung von Datenqualitätsanforderungen in Simulationsworkflows [Re14]. Workflowmodellierer können ihre Anforderungen an die Datenqualität über Policies direkt in Workflows definieren, z. B. an Workflowaktivitäten oder bei Kontrollflussentscheidungen. Die Evaluation der Datenqualität kann entweder im Workflowsystem erfolgen oder auf Ebene der Dienste, welche die Daten liefern bzw. verarbeiten. Um die Datenqualität zu verbessern, können z. B. diese Dienste ausgetauscht oder die Parametrisierung der Dienstaufrufe geändert werden [Re14]. Die von Reiter et al. betrachteten Workflows entsprechen in TOSCA den Plänen (siehe Abschnitt 3.2). Im Policy4TOSCA-Rahmenwerk werden Policies nicht wie bei Reiter et al. in Workflows bzw. Plänen definiert, sondern in den Diensttopologien. Hier muss in Zukunft untersucht werden, welche der beiden Varianten für die Datenqualität von Simulationen in der Cloud am besten geeignet ist. Die Evaluation der Datenqualität sowie die Maßnahmen zu deren Verbesserung erfolgen hingegen analog zu OpenTOSCA – auf Ebene der Workflows bzw. Pläne oder auf Ebene der Dienstkomponenten.

Effizienz und Optimierung: Um die *Effizienz* von Cloud-Anwendungen zu steigern, können insbesondere Optimierungen der TOSCA Diensttopologien und der Pläne erfolgen. Sollten z. B. die Effizienzanforderungen an die Datenhaltung der Software Pandas sehr hoch sein, so kann die in Abb. 2 gezeigte Diensttopologie geändert werden. Der Datenbank-Stack könnte in einer öffentlichen Cloud und als effizienter, skalierbarer und kostengünstiger Datenbankdienst umgesetzt werden. Allerdings muss dabei mit negativen Auswirkungen auf die Datensicherheit gerechnet werden. Hierbei müssen ganzheitliche Methoden entwickelt werden, welche flexibel auf solche sich konkurrierenden Anforderungen reagieren können. Vrhovnik et al. schlagen zudem einen Ansatz zur *Optimierung* von Workflows mit eingebetteten SQL-Anweisungen vor [Vr07]. Solch ein Ansatz kann prinzipiell auch auf die in Abschnitt 4.2 beschriebenen, mit SIMPL-Aktivitäten angereicherten TOSCA-Pläne für die Datenbereitstellung und Datenrückgabe angewandt werden. Als ein zukünftiges Forschungsthema muss die Einschränkung auf SQL allerdings überwunden werden, insbesondere da bei vielen Simulationen auch Zugriffe auf Dateien nötig sind.

Reproduzierbarkeit und Nachvollziehbarkeit: Da das Datenmanagement in der Cloud im Fokus dieses Beitrags steht, gehen wir für die *Reproduzierbarkeit* davon aus, dass Simulationsberechnungen selbst bei gleicher Eingabe immer das gleiche Ergebnis liefern. Darüber hinaus muss dafür gesorgt werden, dass die Softwarebereitstellung, die Datenbereitstellung und die Berechnungsabläufe erneut durchführbar sind. In der Laufzeitumgebung OpenTOSCA enthält ein CSAR-Archiv eine vollständige Beschreibung der Software sowie alle Pläne, die zum Deployment der Software, zur Datenbereitstellung und für die Berechnungsabläufe benötigt werden [Bi13]. Die auf diese Weise beschriebenen Prozessschritte einer Simulation können daher jederzeit wiederholt werden. Die Eingabedaten selbst werden allerdings nicht in einem CSAR-Archiv gehalten. Sie müssen für eine erneute Durchführung einer Simulation vom Nutzer vorgehalten oder bei der vorherigen Durchführung in der Cloud

abgelegt werden. Ein offenes Thema in diesem Kontext ist die Langzeitarchivierung der großen Eingabe- und Ergebnisdaten einer Simulation. Die *Nachvollziehbarkeit* einer Simulation kann über Log- oder Provenance-Informationen sichergestellt werden [CVu12, Fr08]. Cuevas-Vicenttín et al. zeigen hierbei die Notwendigkeit auf, unterschiedliche Provenance-Informationen aus verschiedenen Quellen in der Cloud zu integrieren, um sie so Wissenschaftlern besser zugänglich zu machen [CVu12]. Als Basis können bereits verfügbare Systeme verwendet und an die Bedürfnisse von Simulationen angepasst werden. Beispiele sind Systeme für die Kontrollpflicht der Auftragsdatenverarbeitung in der Cloud [Se13].

6 Zusammenfassung und Ausblick

Speziell für Organisationen, die Simulationen eher sporadisch durchführen, geht der Trend für die Ausführung dieser Simulationen zu Cloud-Umgebungen. Die Hauptgründe sind die bedarfsgerechten, dynamischen und automatisierbaren Provisionierungsmöglichkeiten für benötigte Ressourcen. Die Bereitstellung von Software in der Cloud wird bereits von verschiedenen Systemen und Standards, wie z.B. TOSCA unterstützt. Die Fragestellungen bzgl. des Datenmanagements von Simulationen in der Cloud, blieben bisher aber weitgehend unbeachtet. In diesem Beitrag untersuchten wir zum Einen verschiedene Möglichkeiten für die Datenbereitstellung und Datenrückgabe in der Cloud. Wir schlugen hierfür das generische und einfach zu benutzende SIMPL-Rahmenwerk vor [RS13, Re11] und zeigten, wie es mit TOSCA integriert werden kann. Zum Anderen untersuchten wir verschiedene Technologien für die Definition und Einhaltung simulationsspezifischer Nutzeranforderungen. Das Policy4TOSCA-Rahmenwerk kann hierbei als gute Basis dienen, um nicht-funktionale Anforderungen zu definieren [Wa13]. Bezüglich der Einhaltung dieser Anforderungen diskutierten wir zudem verschiedene offene Forschungsthemen. In diesem Rahmen wichtige Themen sind insbesondere die in Abschnitt 5 diskutierten Optimierungsmöglichkeiten sowie die dort skizzierte Verwaltung und Integration von Provenance-Informationen.

Danksagung: Die Autoren danken der Deutschen Forschungsgemeinschaft (DFG) sowie dem Bundesministerium für Wirtschaft (BMWi) für die Förderung des Projekts im Rahmen des Exzellenzclusters Simulation Technology bzw. im Rahmen von CloudCycle (01MD11023). Außerdem danken wir Dimka Karastoyanova und Frank Leymann sowie den Mitgliedern ihrer Forschungsgruppe für ihre hilfreiche Kooperation.

Literatur

[Ba10] D. Barseghian et al. Workflows and Extensions to the Kepler Scientific Workflow System to Support Environmental Sensor Data Access and Analysis. *Ecological Informatics*, 5(1), 2010.

[Bi13] T. Binz et al. OpenTOSCA – A Runtime for TOSCA-based Cloud Applications. In *Tagungsband der 11. International Conference on Service Oriented Computing (ICSOC)*, Berlin, Deutschland, 2013.

[CVu12] V. Cuevas-Vicenttín et al. Scientific Workflows and Provenance: Introduction and Research Opportunities. *Datenbank-Spektrum*, 12(3), 2012.

[FE11] J. Fehr und P. Eberhard. Simulation Process of Flexible Multibody Systems with Non-modal Model Order Reduction Techniques. *Multibody System Dynamics*, 25(3), 2011.

[Fr08] J. Freire et al. Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3), 2008.

[Kr13] R. Krause et al. Scientific Workflows for Bone Remodelling Simulations. *Applied Mathematics and Mechanics*, 13(1), 2013.

[Lu06] B. Ludäscher et al. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice and Experience*, 18(10), 2006.

[MG11] P. Mell und T. Grance. The NIST Definition of Cloud Computing, 2011.

[OA13] OASIS. *Topology and Orchestration Specification for Cloud Applications Version 1.0*, Mai 2013.

[Re11] P. Reimann et al. SIMPL - A Framework for Accessing External Data in Simulation Workflows. In GI, Hrsg., *Datenbanksysteme für Business, Technologie und Web (BTW)*, Kaiserslautern, Deutschland, 2011.

[Re14] M. Reiter et al. Quality of Data Driven Simulation Workflows. *Journal of Systems Integration*, 5(1), 2014.

[RK11] J. B. Rommel und J. Kästner. The Fragmentation-Recombination Mechanism of the Enzyme Glutamate Mutase Studied by QM/MM Simulations. *Journal of the American Chemical Society*, 26(133), 2011.

[RS13] P. Reimann und H. Schwarz. Datenmanagementpatterns in Simulationsworkflows. In GI, Hrsg., *Datenbanksysteme für Business, Technologie und Web (BTW)*, Magdeburg, Deutschland, 2013.

[Se13] A. Selzer. Die Kontrollpflicht nach § 11 Abs. 2 Satz 4 BDSG im Zeitalter des Cloud Computing. *Datenschutz und Datensicherheit - DuD*, 37(4), 2013.

[SR09] A. Shoshani und D. Rotem. *Scientific Data Management: Challenges, Technology, and Deployment*. Computational Science Series. Chapman & Hall, 2009.

[TDG07] I. Taylor, E. Deelman und D. Gannon. *Workflows for e-Science - Scientific Workflows for Grids*. Springer, London, UK, 2007.

[VHKL13] K. Vukojevic-Haupt, D. Karastoyanova und F. Leymann. On-demand Provisioning of Infrastructure, Middleware and Services for Simulation Workflows. In *Tagungsband der 6. IEEE International Conference on Service Oriented Computing and Applications*, Kauai, HI, USA, 2013.

[Vr07] M. Vrhovnik et al. An Approach to Optimize Data Processing in Business Processes. In *Tagungsband der 33. International Conference on Very Large Data Bases (VLDB)*, Wien, Österreich, 2007.

[Vr08] M. Vrhovnik et al. An Overview of SQL Support in Workflow Products. In *Tagungsband der 24. International Conference on Data Engineering*, Cancùn, México, 2008.

[Wa13] T. Waizenegger et al. Policy4TOSCA: A Policy-Aware Cloud Service Provisioning Approach to Enable Secure Cloud Computing. In *Tagungsband der 3. International Conference on Secure Virtual Infrastructures (DOA-Trusted Cloud'13)*, Graz, Österreich, 2013.