Automatisierte Verarbeitung von Fernerkundungsdaten mit Cloud-Technologien

STEPHAN PLABST[1]

Zusammenfassung: Dieser Beitrag stellt kurz das Projekt IQmulus aus dem 7. Forschungsrahmenprogramm der Europäischen Union vor. Es wird dabei ein Schwerpunkt auf die automatisierte Verarbeitung von nutzerdefinierten Workflows gelegt. Dabei kommen Technologien aus dem Umfeld des Cloud-Computing und Big Data zum Einsatz. Dies wird beispielhaft am Anwendungsfall Änderungserkennung (Change Detection) von Gebäuden im Gebäudebestand aufgezeigt. Abschließend wird ein Ausblick auf die weiteren Projektziele gegeben

1 Das Projekt IQmulus

1.1 Projektziele

Neu entwickelte Datenerfassungstechniken stellen ein Mittel für die schnelle und effiziente, mehrdimensionale und räumliche Datenerfassung dar. All diese Systeme liefern Punktwolken, oft mit anderen Sensordaten angereichert, wodurch große Mengen an Rohdaten produziert werden.

Das Projekt IQmulus – „A High-volume Fusion and Analysis Platform for Geospatial Point Clouds, Coverages and Volumetric Data Sets“, das aus Mitteln des 7. Forschungsrahmenprogramms (FP7) der Europäischen Union gefördert wird, hat zum Ziel eine Plattform zu entwickeln, die die benötigten Funktionen bereitstellt, um aktuelle Forschungsergebnisse aus der Datenverarbeitung und Visualisierung zu integrieren. Mit dieser Plattform sollen Fragestellungen aus dem realen Leben bearbeitet und beantwortet werden. Beispiele hierfür sind das Echtzeitmonitoring von Flutereignissen, die Änderungserkennung in flächendeckenden Datenbeständen oder die Generierung aktueller Seebodenmodelle.

Angesichts der breiten Auswahl von verschiedenen, verfügbaren Sensoren und der großen Menge an so gewonnenen Daten, kombiniert mit der Absicht die aus den Daten gewonnen Erkenntnisse in einem möglichst kurzen Zeitraum bereitzustellen, muss die angestrebte Plattform sowohl bei der Verarbeitung als auch bei der Datenspeicherung hochgradig skalierbar sein. Dabei kommt besonders den vier Aspekten Vielfalt, Menge, Geschwindigkeit und Analytik, die häufig mit dem Begriff Big Data verbunden sind, eine sehr große Bedeutung zu (Quelle: http://www.iqmulus.eu).

1.2 Architektur

Die Entwicklung für die Plattform ist auf mehrere Teams aufgeteilt, die sich mit verschiedenen Aspekten und Komponenten beschäftigen. Die Integration der verschiedenen Komponenten erfolgt dabei mit Methoden der Kontinuierlichen Integration (Continous Integration), bei der die

[1] M.O.S.S. Computer Grafik Systeme GmbH, Hohenbrunner Weg 13, 82024 Taufkirchen;
E-Mail: splabst@moss.de

Ergebnisse der Entwickler direkt in der Plattform zum Einsatz kommen. Dabei kommt dem Test der entwickelten Software eine sehr große Bedeutung zu (PLABST et al. 2014).

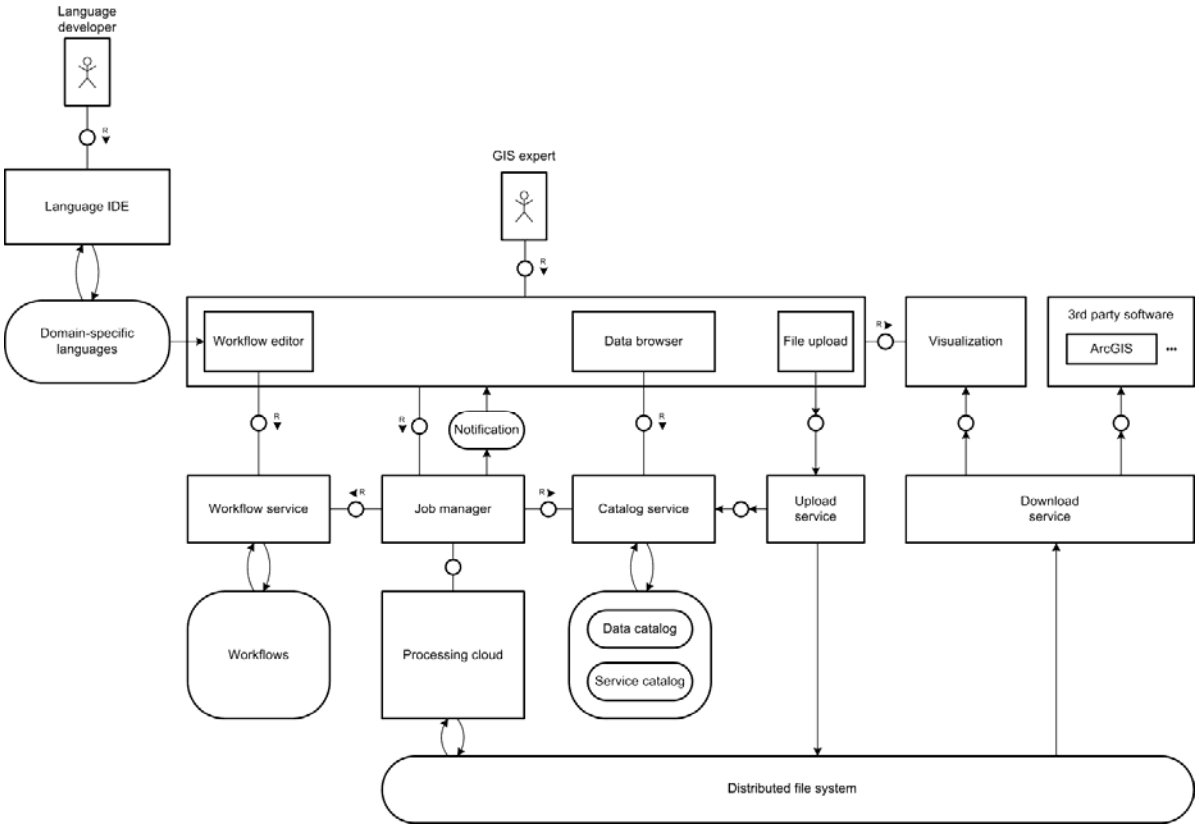

Abb. 1: Architektur der IQmulus-Plattform (schematisch), Quelle: ZULKOWSKI et al. 2014

Um die schnelle Skalierbarkeit der einzelnen Prozesse gewährleisten zu können, kommen in der praktischen Umsetzung der Architektur Technologien aus dem Umfeld des Cloud-Computing zum Einsatz. Dazu gehören unter anderem die serverseitige Verarbeitung von Prozessen, die Virtualisierung der eingesetzten Server auf der vorhandenen Hardware und die Kapselung der Prozesse in Containern mittels Docker (http://www.docker.com). Zusätzlich wird Software aus dem Hadoop-Projekt (http://hadoop.apache.org) eingesetzt, um einerseits einen dynamisch erweiterbaren Speicherplatz für die großen Datenmengen zur Verfügung zu haben und um andererseits eine spezielle Ausführungsumgebung für auf Big Data optimierte Analysen bereit zu stellen (DEAN & GHEMAWAT 2008). In Abb. 1 sind diese beiden Komponenten mit „Distributed file system“ und „Processing cloud“ beschrieben.
Die Datenbereitstellung erfolgt dabei über einen Uploaddienst oder den direkten Zugriff auf das verteilte Dateisystem. Die Daten werden anschließend in einem Katalogdienst registriert und stehen für die weitere Verarbeitung in den mittels Workflow definierten Prozessen zur Verfügung.
Die Ergebnisse der Workflows werden wieder im verteilten Dateisystem zur Verfügung gestellt und können über die Plattform visualisiert werden.

1.3 Workflow-Definition in einer natürlichen Sprache

Ziel des Projekts ist es, die Verarbeitung der sehr großen Datenmengen möglichst weitgehend zu automatisieren und durch Einsatz der Cloud- und Big Data-Technologien gegenüber den klassischen Ansätzen der Geoinformatik zu beschleunigen. Dies ermöglicht es, hochdynamische Datensätze zu analysieren und möglichst in Echtzeit Hinweise für Entscheider zu geben. Gleichzeitig soll die Plattform aber auch die Langzeitanalyse von sehr großen Daten ermöglichen, die bisherige Systeme teilweise überfordern.

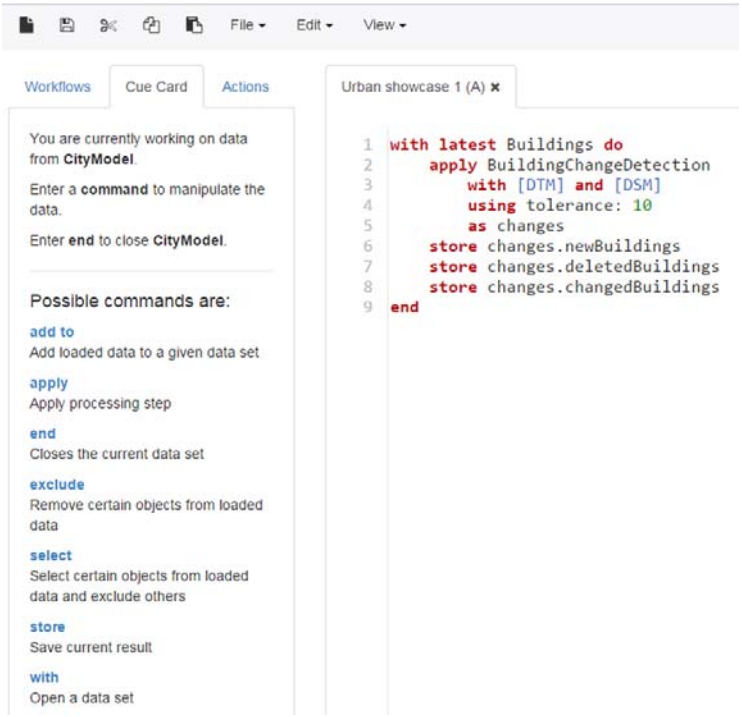

Abb. 2: Workfloweditor, Quelle: [KRÄMER/SENNER, 2014]

Um die verschiedenen Anforderungen aus der Nutzergemeinschaft abdecken zu können, ist es notwendig, die Definition der Workflows für die Ausführung der Analyseprozesse so flexibel aber auch so einfach als nötig zu gestalten. Dabei soll es auch Benutzern, die keine speziellen GIS- oder Programmierkenntnisse besitzen, möglich sein, einen Workflow entsprechend den eigenen Bedürfnissen zu definieren.

Die Plattform stellt dazu bestimmte Analyseroutinen – gekapselt als sogenannte Prozessdienste (STUMPF et al. 2014) – zur Verfügung. Die eigentliche Definition des Workflows erfolgt dabei in einer der natürlichen Sprache ähnlichen Festlegung. Diese Sprache nennt sich Domain Specific Language (DSL), die speziell an die Bedürfnisse für Verarbeitung und Analyse räumlicher Daten angepasst ist (KRÄMER & SENNER 2014). In dieser Sprache stehen die Verben für bestimmte Prozessschritte und die Objekte für entweder ausgewählte oder automatisch festgelegte Datensätze.

Die Sprache ist erweiterbar gestaltet, um auf weitere Anforderungen reagieren zu können. Zusätzlich ist es möglich die Sprache zu übersetzen, um Sie an Nutzerkreise aus anderen Sprachräumen anzupassen. Zum Zeitpunkt der Erstellung dieses Beitrags liegt die DSL nur in Englisch vor.

Für die Festlegung eines Workflows wird dabei ein visueller Editor zur Verfügung gestellt, der den Benutzer beim Editieren unterstützt. Zusätzlich wird dem Benutzer eine Auswahl an vordefinierten Workflows angeboten, die von Spezialisten vorab erstellt wurden.

Der in Abb. 2 dargestellte Workflow ist ein Beispiel für eine Änderungserkennung und wird in Kapitel 2 genauer besprochen.

1.4 Workflow-Steuerung zur hochgradigen Automatisierung von Abläufen

Die in der DSL und damit in einer annähernd natürlichen Sprache definierten Workflows werden anschließend in ein technisches Format übersetzt, das eine automatisierte, serverseitige Ausführung ermöglicht. Als Sprache wurde hier ein XML-Dialekt gewählt, der die Abläufe auf Basis gefärbter Petri-Netze definiert (ZULKOWSKI et al. 2014). Dabei wird der Workflow in eine Abfolge von einzelnen Prozessen mit jeweils verbundenen Ein- und Ausgängen überführt. Die einzelnen Prozesse entsprechen dabei ausführbaren Programmen, die Datenkonversionen oder -analysen durchführen.

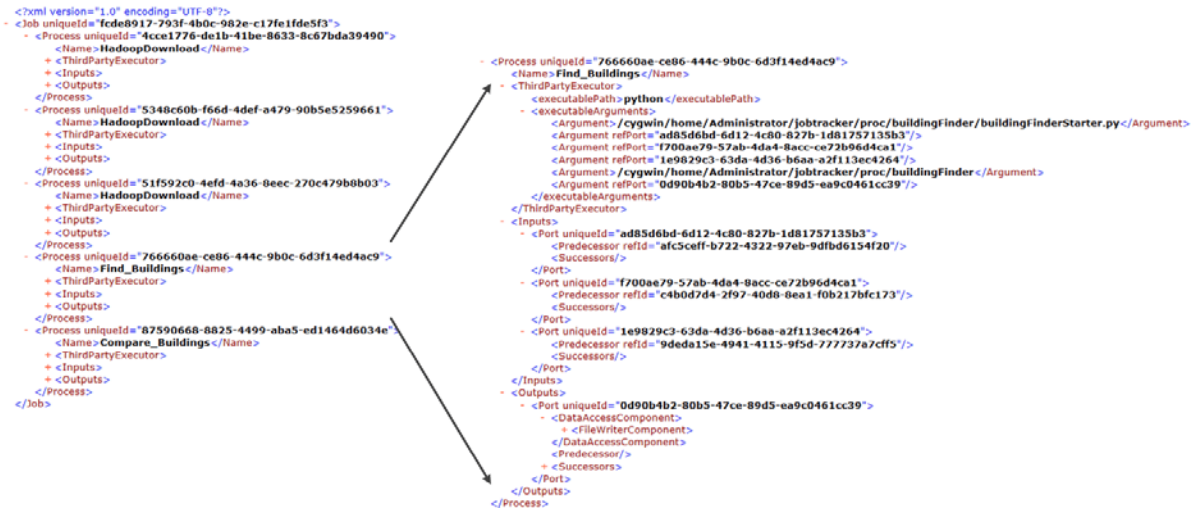

Abb. 3: Workflowbeschreibung in XML

Dabei ist es auch möglich innerhalb der Prozessketten Zweige zu definieren, die parallel oder in Form einer Schleife nacheinander abgearbeitet werden, Dies wird in den Workflows als Multiplexer bezeichnet. Ein solcher Workflow ist in Abb. 4 dargestellt. In dem dort dargestellten Workflow wird der Multiplexer verwendet, um das Bearbeitungsgebiet in mehrere Teilgebiete aufzuteilen und die eigentliche Analysefunktion massiv parallelisieren zu können.

Zur Abarbeitung der so beschriebenen Prozessketten kommt der novaFACTORY Process Chain Manager (novaFACTORY PCM) zum Einsatz. Dieser kann auch unabhängig von der Definition der Workflows im Rahmen einer DSL eingesetzt werden und ermöglicht es komplexe Abläufe ohne notwendige Nutzerinteraktion abzuarbeiten und zu automatisieren. Aufgrund der eindeutigen Beschreibung in Form des XML-Dokuments kann ein so definierter Workflow jederzeit wiederholt werden.

Da die Definition der durchzuführenden Prozesse ebenfalls innerhalb der XML-Datei erfolgt, kann flexibel auf notwendige Anforderungen der Anwender reagiert werden. Ebenso einfach ist es möglich neue Entwicklungen in die Prozessketten aufzunehmen.

2 Anwendungsbeispiel Änderungserkennung von Gebäuden

2.1 Change Detection

Aufbauend auf den bisherigen Ergebnissen des Projekts wird ein Workflow für die Änderungserkennung (Change Detection) von Gebäuden im Gebäudebestand vorgestellt. Dabei kommt die Software novaFACTORY der Firma M.O.S.S. Computer Grafik Systeme GmbH zum Einsatz.

Ziel dieses Workflows ist es automatisiert und großflächig Veränderungen gegenüber dem aktuell gespeicherten Gebäudebestand zu erkennen. Dies kann vor allem dabei helfen Änderungen am Bestand schneller als bisher in die Datenbasis zu übernehmen, aber auch um Abweichungen zwischen den vorhandenen Daten und den realen Gebäuden zu erkennen. Das System soll hier zur Verfahrensunterstützung dienen und kann die eigentlichen Messarbeiten nur unterstützen und nicht ersetzen.

Die flächendeckende Erkennung der Gebäude erfolgt dabei nicht auf Basis terrestrischer Messungen, sondern auf Basis der vorhandenen Fernerkundungsdaten. Die Erkennung der Gebäude kann dabei entweder in 2D auf Basis von Rasterdaten, z.B. aus Digitalen Orthophotos, oder in 3D auf Basis eines Digitalen Oberflächenmodells, z.B. aus LiDAR-Daten oder Luftbildmatching, erfolgen. Im Folgenden wird der Workflow für die Erkennung aus einem Oberflächenmodell in Form einer Punktwolke beschrieben.

2.2 Workflow

Ausgangsdaten für diese Erkennung sind das aktuelle und flächendeckende Digitale Oberflächenmodell aus LiDAR-Daten in Form einer Punktwolke, das aktuelle Digitale

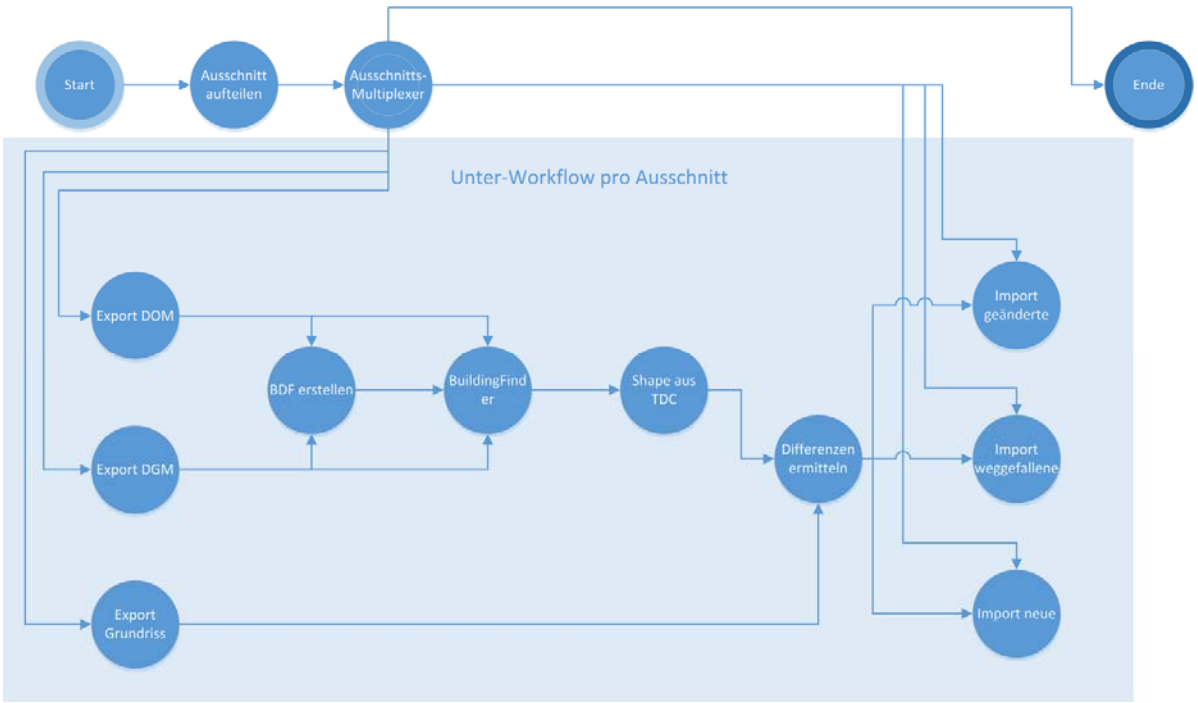

Abb. 4: Workflow für Change Detection (schematisch), Quelle: PLABST et al. 2014

Geländemodell als GRID und der aktuelle Gebäudebestand in Form von Vektordaten. Der Workflow führt zuerst eine Vorverarbeitung der Punktwolke durch, um die Daten möglichst optimiert der anschließenden Datenprozessierung zuführen zu können. Diese Vorverarbeitung beinhaltet z.B. die Filterung anhand verschiedener Merkmale der Punktwolke (speziell der Klassifizierung der Punkte) und die Kachelung der Daten für die nächsten Prozessschritte.
Auf Basis der so vorbereiteten Punktwolke wird anschließend eine Erkennung von Gebäuden – speziell der Gebäudedächer aus dem Oberflächenmodell – durchgeführt. Dazu kommt die Software tridicon® BuildingFinder der Firma 3DCon GmbH zum Einsatz. Die Software ermittelt dabei aus der Punktwolke ein dreidimensionales Gebäudemodell. Für die weiteren Prozessschritte sind nur noch die Grundrisse der Gebäude notwendig. Diese entsprechen dabei – als Besonderheit bei der Erkennung von Gebäuden aus Fernerkundungsdaten – nicht dem eigentlichen, amtlichen Gebäudegrundriss sondern der Form der Dachflächen. Dies ist für den Zweck der beschriebenen Anwendung ausreichend.
Die so erzeugten Gebäudegrundrisse werden mit den vorhandenen Gebäudegrundrissen aus dem Datenbestand verglichen, um die Änderungen zwischen dem Prozessierungsergebnis und den vorhandenen Daten zu finden. Dabei wird eine Toleranz beim Vergleich angesetzt, um Differenzen zwischen den Datenbeständen, die aufgrund der unterschiedlichen Entstehungsweisen und nicht aufgrund tatsächlicher Änderungen am Gebäudebestand entstanden sind, Rechnung zu tragen. Zum anderen wird eine Aggregation der Einzelgebäude durchgeführt, da bei der automatisierten Erkennung von Grundrissen aus Fernerkundungsdaten eine Unterscheidung zwischen Einzelgebäuden mit einer gemeinsamen Kante, z.B. Reihenhäuser, schwieriger ist.

Abb. 5: Erkennungsergebnis

Die Ergebnisse dieses Vergleichs sind Gebäudegrundrisse, die neu hinzugekommen sind (z.B. Neubauten), die geändert wurden (z.B. Umbauten oder Abbruch und Neubau) oder nicht mehr existieren (z.B. Abbruch). Dies gilt jeweils im Vergleich zwischen dem Datenbestand, der aus den Fernerkundungsdaten erkannt wurde und dem vorhandenen Datenbestand. Dieses Vergleichsergebnis wird in einem eigenen Datenbestand abgelegt und dem Bearbeiter zur Prüfung vorgelegt.
Auf Basis der Ergebnisse ist es möglich, weitere Maßnahmen zu veranlassen. Dies kann eine Kontrolle der Ergebnisse mittels eines weiteren Datenbestands (z.B. durch einen Vergleich mit den Orthophotos) oder die Beauftragung eines Messtrupps zur punktuellen Aktualisierung der Bestandsdaten sein. Wenn die Genauigkeit der vorliegenden Ergebnisdaten ausreicht, können diese auch direkt in den aktuellen Bestand übernommen werden.

Falls notwendig ist es möglich, den automatischen Ablauf des Workflows in Form eines Stopppunktes zu unterbrechen, um einen manuellen Eingriff zu ermöglich. Beispielsweise kann dies die Kontrolle der Prozessierungsergebnisse sein oder die Einbindung eines Programmes, das Benutzereingaben erfordert.

2.3 Beschleunigung der Rechenleistung

Die Verarbeitung sehr großer, flächendeckender Datenbestände dauert in der Regel sehr lange, da die Verarbeitungsdauer mit der Datenmenge korreliert. Bei aktuellen, landesweiten LiDAR-Kampagnen können hier Datengrößen von mehreren Terrabyte und einer Punktzahl von mehreren 100 Milliarden Punkten pro Datensatz entstehen. Die begrenzenden Faktoren für die Weiterverarbeitung dieser Daten sind hier vor allem die begrenzte Prozessorleistung und der performante Zugriff auf die Datenbestände (DUMBILL 2012).
Der Dateizugriff kann durch die Verwendung verteilter Dateisysteme und eine intelligente Speicherung der Daten in diesen Dateisystemen erfolgen (BOEHM 2014). Im Projekt wird hier auf das Dateisystem HDFS des Hadoop-Projekts gesetzt.
Die Rechenleistung kann vor allem durch Parallelisierung der Algorithmen verbessert werden. Hierbei spielt entweder die Parallelisierung durch Nutzung mehrere lokaler CPU-Kerne eine Rolle oder die Parallelisierung durch Nutzung von Cloud- und Big Data-Technologien. Wenn Algorithmen für die Anwendungsfälle neu geschrieben werden, so können hier spezielle Prozessierungsumgebungen, wie Hadoop, genutzt werden (ELDAWY & MOKBEL 2013). Im vorliegenden Fall wurde eine kommerzielle Software verwendet, die auf dem verwendeten Server die vorhandenen Ressourcen so gut als möglich ausnutzte. Sie zeigte eine Auslastung auf den vorhandenen Prozessorkernen. Eine Umstellung auf ein Framework wie Hadoop war aber nicht möglich.
Hierfür bietet die in novaFACTORY integrierte Workflowsteuerung novaFACTORY PCM die Möglichkeit unter Verwendung von Cloud-Technologien bei Bedarf dynamisch Serverknoten für die Berechnung hinzuzufügen. Durch die erfolgte Kachelung der Daten können die beiden Prozessschritte der Gebäudeerkennung und des Vergleichs der Datenbestände auf mehrere Server verteilt und dort gestartet werden. Durch diese Besonderheit ist es möglich die Vorteile einer Cloud-Umgebung zu nutzen, ohne dass die verwendete Software speziell dafür angepasst werden muss (PLABST et al. 2014). Ebenso kann die Software ohne Aufteilung auf mehrere Serverknoten in einer klassischen GIS-Umgebung verwendet werden.

3 Fazit und Ausblick

Durch die Definition der Domain Specific Language auf Basis der natürlichen Sprache und die direkte Übersetzung in maschinenlesbaren Code vereinfachen die Ergebnisse des Projekts die Definition komplexer Workflows für die Anwender. Dabei kann auf einen Bestand bereits vorhandener Definitionen, die von Experten erstellt wurden, zurückgegriffen werden. Ebenfalls ist es möglich, eigene Workflows auf Basis dieser Vorlagen neu zu erstellen.
Die Verwendung von Technologien aus dem Umfeld von Big Data und Cloud Computing ermöglicht es Algorithmen aus den Bereichen der Geoinformatik und der Fernerkundung zu

beschleunigen. Damit wird es möglich andere Nutzerkreise und Anwendungen für Geodaten zu definieren und zu erschließen.
Das Projekt IQmulus hat noch eine Laufzeit über knapp zwei weitere Jahre. In diesen steht die weitere Entwicklung und Verfeinerung von Algorithmen für Big Data im Vordergrund. Zusätzlich sollen die für die Continous Integration notwendigen Testverfahren verbessert. Darüber hinaus soll die bereits existierende IQmulus Plattform und Zusammenarbeit mit der Nutzergemeinschaft verbessert.
Mit dem Produkt novaFACTORY ist es möglich, bereits heute auf Erkenntnisse aus dem Projekt IQmulus direkt zuzugreifen.

4 Danksagung

Teile der hier vorgestellten Arbeit werden durch die EU-Zuwendung FP7-ICT-2011-318787 (IQmulus) unterstützt.

5 Literaturverzeichnis

BOEHM, J., 2014: File-centric Organization of large LiDAR Point Clouds in a Big Data Context. Workshop on Processing Large Geospatial Data Cardiff, UK, 08.07.2014. http://rs.tudelft.nl/~rlindenbergh/workshop/BoehmIQmulus.pdf

DEAN, J. & GHEMAWAT, S., 2008: MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, S. 1-13.

DUMBILL, E., 2012: What is big data? O'Reilly Radar vom 11.01.2012. O'Reilly Media, Inc. http://radar.oreilly.com/2012/01/what-is-big-data.html

ELDAWY, A. & MOKBEL, M.F., 2013: A Demonstration of SpatialHadoop: an efficient MapReduce Framework for Spatial Data. Proceedings VLDB Endowment 6, Seite 1230-1233. http://db.disi.unitn.eu/pages/VLDBProgram/pdf/demo/p744-mokbel.pdf

GAŠEVIC, D., DJURIC, D. & DEVEDŽIC, V., 2009: Model Driven Engineering and Ontology Development. 2nd ed., Springer.

KRAUT, V., KIEßLICH, N., RUDOLF, H., SCHÄFER, M. & KRÄMER, M., 2014: DSL Tool Chain – Prototype Release. Deliverable D3.3 des IQmulus Projektes.

KRÄMER, M. & SENNER, I., 2014: Processing DSL Specification - Baseline Version. Deliverable D2.4.1 des IQmulus Projektes.

KRÄMER, M. & STEIN, A., 2014: Automated urban management processes: integrating a graphical editor for modular domain-specific languages into a 3D GIS. Proceedings of the 19th international conference on urban planning and regional development in the information society GeoMultimedia.

PLABST, S., KRÄMER, M., RUDOLF, H., KRAUT, V. & STUMPF, A., 2014: Integrated Prototype – Version 1. Deliverable D6.1 und D6.2 des IQmulus Projektes.

STUMPF, A., CANCOUËT, R., PIETE, H., DELACOURT, C., SPAGNUOLO, M., CERRI, A., SIRMACEK, B. & LINDENBERGH, R., 2014: Change Detection and Dynamics Toolkit. Deliverable D4.5.1 des IQmulus Projektes.

ZULKOWSKI, M., PLABST, S., KRÄMER, M., KIEßLICH, N., 2014: Control Components – Vertical Prototype Release. Deliverable D3.2 des IQmulus Projektes.