Beitrag G: Andreas Abecker, Torsten Brauer, Johannes Kutterer, Karsten Schnitter, Jens Nimis, Patrick Wiener

Überlegungen zu einer Spatial Big Data Architektur im BigGIS Projekt

Andreas Abecker[1], Torsten Brauer[1], Johannes Kutterer[1], Karsten Schnitter[1], Jens Nimis[2], Patrick Wiener[2]

[1] disy Informationssysteme Gmbh,vname.nname@disy.net
[2] Hochschule Karlsruhe - Technik und Wirtschaft, vname.nname@hs-karlsruhe.de

Abstract

In the recent years, a number of data-management and data-analytics applications and technologies under the label “big data” has found much interest among academics and practitioners. From our point of view, the respective researchers and commercial providers, up to now, neglected to a large extent both the spatial dimension of potential big-data applications and their usage potential in the area of environment protection and environment management. Hence, the RTD project BigGIS investigates methods and tools for “Spatial Big Data” in three environment-related application scenarios. In this paper, some basic definitions and considerations are explained, the three application scenarios of BigGIS are presented and some initial insights regarding software architectures for Spatial Big Data solutions in the area of geo data and environmental applications are presented.

Zusammenfassung

Unter der Überschrift “Big Data” machen in den vergangenen Jahren eine Reihe neuartiger Anwendungen und Technologien für Datenmanagement und -analyse viel von sich reden, in Wissenschaft und Praxis. In der Wahrnehmung der Autoren wurden dabei bisher sowohl die räumliche Komponente möglicher Big Data Anwendungen als auch Anwendungsideen im Kontext Umweltschutz und -verwaltung eher stiefmütterlich

behandelt. Daher werden im BMBF-Projekt BigGIS anhand dreier Anwendungsszenarien im Umweltbereich Methoden und Techniken für „Spatial Big Data“ untersucht. Der vorliegende Beitrag erläutert einige grundlegende Begriffe und Überlegungen, stellt die Anwendungsszenarien in BigGIS vor und skizziert einige erste Einsichten zur Umsetzung von Spatial Big Data Lösungen im Geo-/Umweltdatenbereich.

1 Motivation und Überblick

1.1 Motivation

Geoinformationssysteme (GIS) – bzw. allgemeiner: räumliche Visualisierungen und Datenanalysen innerhalb von Geodateninfrastrukturen (GDI) – werden seit Langem für die Verarbeitung räumlicher Daten (also Daten mit Orts- bzw. Raumbezug) in vielfältigen Anwendungsgebieten in der öffentlichen Verwaltung, wie Stadtplanung, Flurneuordnung, Verkehrsplanung, Raum- und Umweltplanung oder auch im Feuer- und Katastrophenschutz sowie im Umwelt- und Naturschutz verwendet. Hinzu kommt in jüngerer Zeit eine wachsende Anzahl von Geodatennutzungen in der Privatwirtschaft, mit zunehmender kommerzieller Bedeutung in Anwendungsgebieten wie Logistik, Landwirtschaft, Geomarketing, Standort- und Gebietsplanung, lokationsbasierten Informations- und Werbediensten etc. Zu den Kernfunktionen der verwendeten IKT-Systeme gehören neben der Erfassung der Daten auch räumliche Abfragen sowie die Visualisierung und die Interpolation von Attributwerten in der Fläche aus wenigen Messwerten. Mithilfe von GIS können beispielsweise Umweltschutzbehörden herausfinden, welche Biotope im Fall eines Chemieunfalls, bei einem Hochwasser oder einem Waldbrand tangiert sind. Im Feuer- und Katastrophenschutz werden GIS in Lageinformations- und Einsatzunterstützungssystemen eingesetzt, um betroffene Areale und potenzielle Maßnahmen aus den verfügbaren Informationen abzuleiten oder um Evakuierungen zu koordinieren.

In jüngerer Zeit wird von solchen GIS zunehmend erwartet, dass sie rapide anwachsende Datenmengen im Tera- bis Petabytebereich an Archivdaten und aktuellen Fernerkundungsdaten zusammen mit vielfältigen Geodaten, historischen und aktuellen Sensordaten (Zeitreihendaten) sowie Meldungen von Bürgern integriert und zeitnah nutzbar machen. So wurde im Katastrophenmanagement beispielsweise die *„International Charter Space and Major Disasters“* etabliert, die im Katastrophenfall

weltweit von den Hilfsorganisationen aktiviert werden kann. Wird die Charta aktiviert, werden die Sensoren geeigneter Satelliten so programmiert, dass sie beim folgenden Überflug über das Krisengebiet entsprechende Aufnahmen liefern.

Weiterhin sollen zukünftige GIS nicht mehr nur Abfragen auf vergangenen Informationen ermöglichen, sondern ebenfalls die Prognose *zukünftiger* Entwicklungen, z.B. von Sensor¬werten sowie deren Abhängigkeiten, um Entwicklungen und zukünftige Probleme zu erkennen und vorausschauendes Handeln zu ermöglichen. Bedingt wird diese Entwicklung durch die zunehmende Verbreitung kostengünstiger vernetzter Sensoren (mobile Sensorik bis hin zu einfachen Mobiltelefonen), mit denen auch die Öffentlichkeit direkt und unkompliziert Daten beitragen kann, sowie durch die zunehmende Bedeutung noch unstrukturierter Daten in Form von Fernerkundungsdaten (von Satelliten, u.a. aus dem im Aufbau befindlichen Programm COPERNICUS, Flugzeugen und UAS *(Unmanned Aerial Systems)*). Zudem gewinnen bodengestützte Fernerkundungssensoren wie kommunale Sensorsysteme in Form von Überwachungskameras und Befahrungsdaten mit Laserscannern, sowie bildgebende Systeme an Bedeutung. Darüber hinaus existieren immer mehr textuelle Informationen sowie auch annotierte Fotografien mit geo-temporalen Bezügen, insbesondere im Social Web. Diese zusätzlichen Informationsquellen erlauben prinzipiell die Modellierung kausaler Zusammenhänge und robustere, multivariate Verfahren zur Approximation, Prognose, und Maßnahmenableitung. GIS stehen durch diese Entwicklungen vor Herausforderungen in allen vier Dimensionen, die definitorisch für das neue Technologiefeld Big Data sind: *Volume, Variety, Velocity, Veracity.*

Solchen Fragestellungen widmet sich das FuE-Verbundprojekt BigGIS, das vom Bundesministerium für Bildung und Forschung im Rahmen des Förderschwerpunkts „Big Data“ unterstützt wird.

1.2 Überblick zum BigGIS-Projekt

Ziel des im April 2015 gestarteten BMBF-Forschungsvorhabens **BigGIS** („Prädiktive und präskriptive Geoinformationssysteme basierend auf hochdimensionalen geo-temporalen Datenstrukturen“) ist die Erforschung, prototypische Umsetzung und Evaluierung von Techniken, Modellen und Methoden, die in vielfältigen Anwendungsfällen Entscheidungen auf Basis von großen Mengen an zeitlich-strukturierten Geo-

daten aus unterschiedlichen Quellen (wie insbesondere Fernerkundung, Crowdsourcing und dem Social Web, aber auch aus Legacy-Systemen zur Geodatenverarbeitung) unterstützen.

Mit dem zu schaffenden System soll die integrierte Verwaltung, Analyse und Visualisierung von zeitlichen und räumlichen, strukturierten und unstrukturierten Daten verbessert bzw. überhaupt erst möglich gemacht werden. Auf Basis dieser Daten sollen **deskriptive, prädiktive und präskriptive Analysen** unterstützt werden. Die Verarbeitung der Daten in dem zu entwickelnden System soll dabei schnell genug erfolgen, um z.B. auch die Gesundheits- und Umweltsicherung im Katastrophenfall zu unterstützen. Die Forschungs- und Entwicklungsarbeiten werden getrieben von den praktischen Erfordernissen dreier konkreter Anwendungsszenarien:

Katastrophenschutz: Entscheidungsunterstützung bei komplexen Schadenslagen am Beispiel von Schadgas-Situationen.

Umweltmonitoring: Umweltmanagement am Beispiel des Managements invasiver Tier- und Pflanzenarten mit Auswirkungen auf die menschliche Gesundheit (wie z.B. Eichen-Prozessionsspinner, Asiatische Tigermücke und Beifußblättrige Ambrosie).

Smart City und Gesundheit: Die Förderung der Gesundheit von Menschen in Städten am Beispiel der Umwelteinflüsse Feinstaub und Temperatur.

Im BigGIS-**Projektkonsortium** arbeiten folgende Organisationen zusammen:

- FZI Forschungszentrum Informatik am Karlsruher Institut für Technologie
    - Shared Research Group „Corporate Services and Systems“ (Prof. Thomas Setzer): Projektkoordination, deskriptive und präskriptive Analysen
    - Abteilung „Logistics and Supply Chain Optimization“ (Prof. Stefan Nickel): Präskriptive Analysen, Anwendungsfokus auf dem Szenario Smart City und Gesundheit
    - Abteilung „Wissensmanagement“ (Prof. Rudi Studer): Metadaten, semantische Technologien, Crowdsourcing (nutzergenerierte Inhalte)
- Universität Konstanz, Lehrstuhl „Datenanalyse und Visualisierung“ (Prof. Daniel Keim):
    - Visual Analytics für räumliche und zeitliche Big Data

- Hochschule Karlsruhe Technik und Wirtschaft, Fachgebiet Datenbanksysteme und Webtechnologien (Prof. Jens Nimis):
    - Gesamtarchitektur für das Projekt, Infrastrukturmanagement
- Disy Informationssysteme GmbH, Karlsruhe
    - Experten für GIS und räumliche Datenanalyse, GDI, Software-Integration, Experten für Anwendungen der Umweltverwaltung; Anwendungsfokus auf dem Szenario Umweltmanagement
- Exasol AG, Nürnberg:
    - Datenbankexperten, In-Memory-Datenbanken, Cluster Computing, Kompressionsalgorithmen
- EFTAS Fernerkundung Technologietransfer GmbH, Münster:
    - Erzeugung und Analyse von Fernerkundungsdaten von Satelliten und UAS; Anwendungsfokus auf dem Szenario Katastrophenschutz
- Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW), Karlsruhe:
    - Datenbereitstellung für alle Szenarien, Anforderungsgeber und Evaluationspartner aus Sicht der öffentlichen Verwaltung, Anwendungsfokus auf dem Szenario Umweltmanagement

Weitere assoziierte Partner für die Szenarien Katastrophenschutz und Smart City sind das THW Karlsruhe und die Stadtverwaltung Karlsruhe.

1.3 Aufbau dieses Beitrags

Im Folgenden wird zunächst (Kapitel 2) der Begriff Spatial Big Data etwas näher beleuchtet. Nach einer Arbeitsdefinition für Big Data folgen deren Erweiterung zum Begriff Spatial Big Data, einige Argumente, wieso wir eine wachsende Bedeutung für diese Thematik erwarten, sowie eine Auflistung einiger interessanter Werkzeuge für Spatial Big Data. Danach (Kapitel 3) werden die drei Anwendungsszenarien im BigGIS-Projekt mit ihren Datenquellen und Zielsetzungen erläutert. In Kapitel 4 werden anhand einer noch unvollständigen und allgemeinen Architekturskizze zu BigGIS

einige erste Erkenntnisse und Überlegungen zur Software-Architektur für Spatial Big Data erläutert. Mit Kapitel 5 folgt eine kurze Zusammenfassung.

2 Zum Begriff Spatial Big Data

2.1 Arbeitsdefinition Spatial Big Data

Wie es schon bei früheren IT-Trends der Fall war, ist der Begriff **Big Data** nicht klar wissenschaftlich-technisch definiert, sondern entsteht aus der Zusammenschau weithin akzeptierter Zielsetzungen und Charakterisierungen aus dem Marketing großer IT- und Beratungsfirmen mit einem gewissen Vorrat an modernen Technologien und Lösungsansätzen.[6] Darauf basierend, schlagen wir folgende **Arbeitsdefinition** vor:

Man redet von einer Big Data Anwendung, wenn in mindestens einer der Dimensionen

- **Volume** (Datenmenge, z.B. im Bereich von Terabyte aufwärts)
- **Velocity** (Geschwindigkeit der Datenentstehung, z.B. Tausende von Sensor-Messwerten pro Sekunde)
- **Variety** (Heterogenität von Daten, entstehend z.B. aus der Kombination unstrukturierter und (semi-) strukturierter Daten)
- **Veracity** (Zuverlässigkeit der Daten, z.B. bei Nutzung von Social Media Inhalten)

Ausprägungen vorliegen, die eine effiziente und effektive Verarbeitung und Nutzung der Daten mit konventionellen Methoden des Datenmanagements und der Datenanalyse auf konventioneller Hardware unmöglich machen. Dann kommen neue Methoden der Datenverarbeitung zum Einsatz, wie insbesondere In-Memory Datenbanken, NoSQL und Graph-Datenbanken, Parallelisierung in Clustern durch den MapReduce Ansatz und moderne Verfahren der Datenstromanalyse.

Auch wenn hier einige Begriffe „fließend“ oder auch kontextabhängig sind („konventionelle Hardware“) bzw. sich auf den Entstehungszeitpunkt der Big Data Begrifflichkeit ab etwa 2005 bzw. im engeren Sinne ab 2011 beziehen („neue Methoden“), lassen sich mit dieser Arbeitsdefinition Big Data Fragestellungen gut identifizieren.

[6] Vgl. [BITKOM, 2014], siehe auch https://de.wikipedia.org/wiki/Big_Data oder https://www.gi.de/service/informatiklexikon/detailansicht/article/big-data.html

Die Zielsetzungen, die man in Big Data Projekten verfolgt, entsprechen überwiegend denen, die man bereits in den Bereichen Data Mining, Text Mining und Business Intelligence verfolgt hat, nur unter Verwendung der oben erwähnten neuen Technologien zum Umgang mit den genannten „4 V“ – und zusätzlich häufig unter Hinzukommen eines starken (Nah-)Echtzeitaspekts.

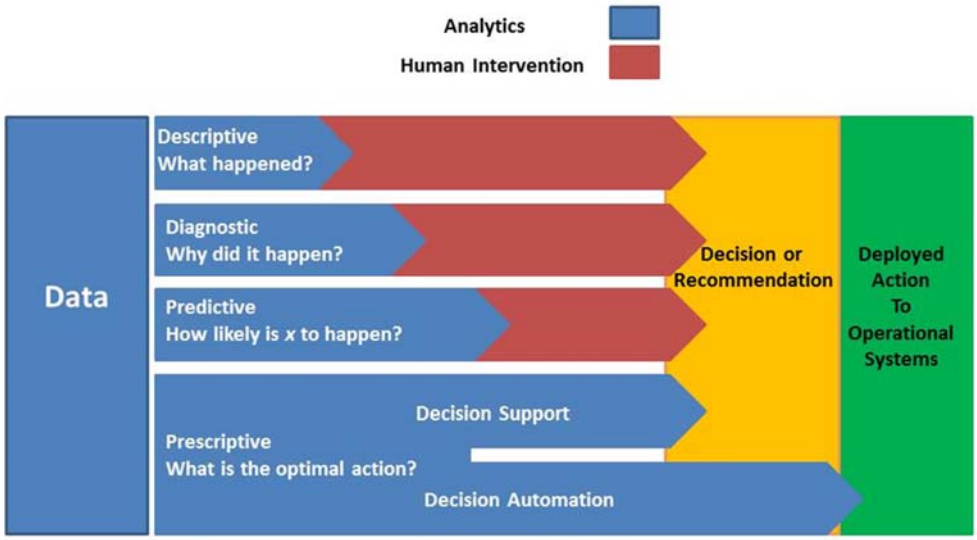

Abbildung G-1: Analytics von der Deskription bis zur Präskription, angelehnt an Gartner

(Quelle: http://www.sv-europe.com/blog/10-reasons-organisation-ready-prescriptive-analytics/)

Die grundlegenden Analysefragestellungen sind in Abbildung G-1 gegenübergestellt:

- **Deskriptive** Analysen stellen interessante Zusammenhänge und Abläufe dar und modellieren sie mit geeigneten Methoden.
- **Diagnostische** Analysen leiten die unterliegenden Kausalzusammenhänge ab.
- **Prädiktive** Verfahren erzeugen Prognosen und Vorhersagen aus Modellen und Kausalzusammenhängen.
- **Präskriptive** Verfahren erzeugen aus Beobachtungen, Modellen und Vorhersagen entscheidungsunterstützende oder sogar -automatisierende Algorithmen.

Im weiteren Verlauf fassen wir für die Darstellung von BigGIS die deskriptiven und die diagnostischen Fragestellungen in einer Kategorie zusammen.

Wichtig ist in der Grafik der durch die Aufteilung von roten und blauen Balken dargestellte Zusammenhang zwischen Aufgaben, die i.d.R. menschliche Interaktion erfordern und solchen, die sich automatisch ausführen lassen. In der Literatur zu Big Data werden häufig Systeme vorgestellt, die bspw. unter Zuhilfenahme von Methoden der Erkennung komplexer Ereignisse (Complex Event Processing, CEP) Überwachungs- und Steuerungsprobleme *vollautomatisch* lösen. Dabei sollte man nicht übersehen, dass normalerweise *vor* der Installation eines solchen Systems (in der Grafik weiter unten angesiedelt) erst die sehr viel stärker manuellen Stufen (in der Grafik weiter oben) des Systemverstehens, der Modellbildung und dann der Definition von Ereignismustern stehen.

Wir erweitern nun den Begriff Big Data zum Begriff **Spatial Big Data**, wenn in einem Big Data Problem eine signifikante örtliche oder räumliche Komponente – in den Daten und/oder bei ihren Auswertungen – vorliegt.

Dies umfasst mindestens zwei interessante Sachverhalte:

- Geodaten selber (insbesondere komplexere Geodaten als reine Sachdaten mit assoziierten Punktgeometrien) sind wesentlicher Gegenstand der Betrachtung (z.B. bei der Analyse von Fahrzeugtrajektorien für die Verkehrsplanung und -steuerung oder bei der Analyse von Bewegungsprofilen von Personen für die Panikforschung).
- Sachdaten mit Ortsbezug werden in nichttrivialer Weise räumlich ausgewertet (z.B. langsame Verlagerung bestimmter Wohnmilieus innerhalb einer Stadt, Korrelationsanalysen räumlich auftretender Phänomene, u.ä.).

Die Erfahrungen mit bekannten Anwendungen der „Spatial Intelligence/ Location Intelligence“ zeigen, dass die algorithmische Komplexität räumlicher Analysen häufig eine Größenordnung höher liegt als bei Datenanalysen ohne räumliche Dimension. Daher kann das obige Kriterium für Big Data Anwendungen „effiziente und effektive Verarbeitung … mit konventionellen Methoden … auf konventioneller Hardware unmöglich“ hier u.E. auch schon bei niedrigeren Ausprägungen der obigen vier V gegeben sein. Außerdem sei darauf hingewiesen, dass Geodatenanalysen in der Praxis häufig mit hoher Datenheterogenität *(Variety)* und niedriger Datenqualität zu kämpfen haben.

2.2 Bedeutung von Spatial Big Data

Wir gehen davon aus, dass mittelfristig die Menge von Spatial Big Data Anwendungen stark ansteigen wird – weil auch die Menge von dafür nutzbaren Daten stark anwachsen wird. Dies wird durch verschiedene Entwicklungen befördert:

- Immer billigerer und einfacherer Zugang zu (immer detaillierteren) Satellitendaten (vgl. COPERNICUS-Programm der EU oder jüngere Entwicklungen im Bereich Kleinsatelliten bis zu Nanosatelliten).
- Zunehmende Anzahl von Forschungsprojekten, kommerziellen Anbietern und Anwendungsideen im Bereich Fernerkundung durch unbemannte Flugobjekte (UAS).
- Immer preisgünstigere und leistungsfähigere In-situ-Sensorik mit (Nah-)Echtzeit-Datenfernübertragung für verschiedenste umweltrelevante Themen, z.B. im Hochwasserbereich.
- Ebenso zunehmend preisgünstige und leistungsfähigere mobile Sensorik mit (Nah-)Echtzeit-DFÜ, die auf Fahrzeugen montiert (z.B. in der *Precision Agriculture*) oder an Smartphones gekoppelt werden kann.
- Wachsende technische Möglichkeiten, Nutzungsszenarien und Benutzerakzeptanz für die Verfolgung beweglicher Objekte (Fußgänger-Tracking, Fahrzeug-Tracking, Warenverfolgung im Internet-of-Things und mit Industry 4.0 Anwendungen, usw.).
- Nutzergenerierte Geodaten (*Volunteered Geographic Information* aus Ansätzen zur Bürgerbeteiligung (Participatory Sensing, Citizen Observatories), wie z.B. bei privaten Wetterstationen, Open Street Map, Mängelmeldern, Artenfinder etc.
- Georeferenzierte (oder georeferenzierbare) Social-Media Inhalte werden mehr und mehr. Gerade in Business-Anwendungen finden diese große Beachtung, aber auch bspw. in Katastrophenszenarien.

All diese Trends lassen uns vermuten, dass verschiedenartigste Anwendungen von Spatial Big Data in der nahen Zukunft entstehen werden / können, mit Bezug auf Geo- und Umweltdaten z.B. in Themenfeldern wie Smart City Überwachung und Steuerung, Präzisionslandwirtschaft, Verkehrsmanagement, Katastrophenschutz und -rettung, erneuerbare Energien und Smart Grid oder Klimaanpassung.

2.3 Technologien für Spatial Big Data

Zum Zeitpunkt der Antragstellung von BigGIS waren erst sehr wenige Werkzeuge und Initiativen zum Umgang mit Spatial Big Data weit verbreitet. Inzwischen gibt es einige interessante Entwicklungen mehr. Zum Selbststudium für interessierte Leser listen wir einige wichtige Werkzeuge auf, auch wenn sie in unserem Projekt (bisher) nicht genutzt werden:

- **Geomesa** – ermöglicht durch verteilte Indexierung die Speicherung, Abfrage und Verarbeitung räumlich-zeitlicher Daten auf verteilten Cloud-Datenbanken wie Accumulo, HBase, Cassandra und Kafka.
- **GeoJinni / Spatial Hadoop** – realisiert MapReduce für räumliche Daten, so dass große Mengen räumlicher Daten mit Hadoop cluster-basiert gespeichert und verarbeitet werden können.
- **Geotrellis** – unterstützt Hochleistungsverarbeitung für Rasterdaten (inkl. Vektor-zu-Raster Transformationen und umgekehrt) via REST-Dienste oder im Batch-Betrieb.
- **Rasdaman** – ist eine hochleistungsfähige Array-Datenbank mit Anwendungen im Rasterdatenmanagement für GIS, aber auch in anderen Anwendungen des Sensordatenmanagements und des Scientific Computing; die Entwickler sind eng mit der OGC Big Data Standardisierung vernetzt.
- **GeoSpark** – realisiert eine In-Memory Cluster-Datenbank für große räumliche Datenmengen (Vektordaten) sowie wichtige geometrische Operationen und räumliche Anfragen.

Darüber hinaus bewerben die großen internationalen Software „Komplettsortimentanbieter“ innerhalb ihrer Big Data Stacks auch Geodaten-Operationen, so wie z.B. IBM [Klein et al, 2015], Oracle [Oracle, 2016] und SAP (Spatial-Funktionen von SAP HANA). Ebenso veröffentlichen die großen GIS-Hersteller in jüngster Zeit auch zunehmend ihre Pläne, Big Data Werkzeuge als Geodatenbanken zu verwenden[7] Dies zeigt zunächst einmal, dass die GIS-Welt und der Big Data Hype sich nicht mehr gegenseitig

[7] Zum Beispiel Esri mit ArchGIS auf SAP HANA (http://geospatial-solutions.com/esri-arcgis-to-support-sap-hana-as-enterprise-geodatabase/) oder Hexagon mit GeoMedia auf SAP HANA (http://www.hexagongeospatial.com/about-us/news/hexagon-geospatial-releases-software-that-integrates-natively-with-sap-hana).

ignorieren können / wollen. Inwieweit diese Ansätze alle schon technisch und betriebswirtschaftlich durchdacht und „frei von Kinderkrankheiten“ sind, muss sich noch zeigen.

3 Die Anwendungsszenarien im BigGIS Projekt

3.1 Anwendungsszenario 1: Katastrophenschutz

Untersucht wird die Entscheidungsunterstützung bei komplexen Schadenslagen am Beispiel von Schadgas-Situationen. Es wird beispielhaft die Ausbreitung einer Schadgaswolke betrachtet, wie sie z.B. bei einem Brand in einer chemischen Fabrik entstehen könnte. Ähnliche Fragestellungen tauchen aber natürlich in vielerlei Katastrophenszenarien auf, z.B. bei Hochwasser, Waldbränden, Chemie-Unfällen oder Terroranschlägen. Ziel ist es, dass Einsatzkräfte innerhalb von 15 Minuten nach Eintreffen vor Ort ein möglichst genaues und umfassendes Schadensbild haben, das auch kontinuierlich aktualisiert wird.

Dazu wird unter Federführung des Projektpartners EFTAS der Ansatz des *Micro Rapid Mapping* entwickelt: Dafür wird ein Mikro-Flugroboter (AiD-MC8 Octocopter) mit einer RGB-Kamera, einer Thermalkamera, einer Hyperspektral-Kamera und einem RTK GPS ausgestattet. Die durch Aufklärungsflüge erhaltenen Fernerkundungsdaten sollen dann mit offiziellen topographischen Karten und Katasterdaten abgeglichen und ergänzt werden, insbesondere z.B. um Daten über kritische Infrastrukturen, gefährdete Bevölkerung, Naturschutzgebiete, Wasserschutzgebiete usw. Später könnte diese Information zur Komplettierung und Aktualisierung des Lagebilds ergänzt werden durch In-situ-Sensorik (z.B. der Feuerwehr) und durch nutzererzeugte Inhalte bzw. Social Media Inhalte, sofern sich betroffene Bürger mit Internetverbindung im gefährdeten Gebiet aufhalten.

Die dominante Big Data Dimension ist hier eindeutig ***Volume***, wenn man bedenkt, dass eine 10-minütige Befliegung durch den RGB Videostream mit Full-HD-50 Daten schon etwa 175 GB Rohdaten und durch die Hyperspektral-Kamera mit 5 Cubes/sec und 125 Spektralkanälen etwa 520 GB Rohdaten liefert. Da diese großen Datenvolumina aber möglichst schnell für die Analyse ausgewählt und ausgewertet werden müssen, kann man das Problem auch als spezifische Ausprägung der Dimension ***Velocity*** verstehen. Weiterhin ergeben sich durch Integration zusätzlicher

Daten (Bestandsdaten, In-situ-Sensorik, Social Media Daten) wachsende Herausforderungen mit ***Variety*** und auch mit ***Veracity***.

Als Analyse-Fragestellungen können/sollen untersucht werden:

- **Deskriptiv**: Zusammenführung von Beobachtungsdaten und Bestandsdaten; Kombination der verfügbaren Daten und Informationen in ein konsistentes Lagebild; Analyse und Visualisierung der aktuellen Gefährdungssituation.
- **Prädiktiv**: Vorhersage möglicher zukünftiger Szenarien zur Ausbreitung des Schadstoffs in Abhängigkeit von geographischen Gegebenheiten und Wetterbedingungen.

Präskriptiv: Planung des effizienten Einsatzes von Sensorik, um einen größtmöglichen Informationsgewinn für die deskriptiven und prädiktiven Verfahren zu erzielen (Befliegungen, In-situ-Sensorik oder gezielte Anfragen an Helfer).

3.2 Anwendungsszenario 2: Umweltmonitoring

Im von der LUBW koordinierten Anwendungsszenario Umweltmonitoring werden **invasive Spezies** (Tier- oder Pflanzenarten) betrachtet, die zum Teil erhebliche Gesundheitsprobleme für die Bevölkerung (wie Allergien), ökologische (Verdrängen einheimischer Arten) oder ökonomische Schäden (Schäden an Früchten und Nutzpflanzen) verursachen (wie z.B. Eichen-Prozessionsspinner, Asiatische Tigermücke und Beifußblättrige Ambrosie). Im Projekt wird beispielhaft die **Kirschessigfliege** betrachtet *(drosophila suzukii)*, welche vor etwa 10 Jahren aus Asien über Amerika nach Europa eingeschleppt wurde und in Deutschland seit 2011 eine ernsthafte Gefährdung für den Wein- und den Obstbau darstellt. Sie befällt alle weichschaligen Früchte wie z.B. Süßkirschen, Erdbeeren, Brombeeren, Himbeeren, Johannisbeeren, Heidelbeeren, Pfirsiche sowie Kelter- und Tafeltrauben und hat in einigen Gebieten bereits große Ernteausfälle bewirkt. Die Kirschessigfliege hat sehr kurze Fortpflanzungszyklen: sie benötigt nur 12 bis 14 Tage bis zur nächsten Generation und kann bis zu 13 Generationen pro Jahr aufweisen. Die Kirschessigfliege befällt unterschiedliche Früchte in verschiedenen Phasen des Jahres, sie nutzt Rückzugs- und Überwinterungsräume zwischen den Befallphasen der Obstkulturen.

Ziel im Projekt ist es, die Verteilungsmuster in Abhängigkeit von Vegetation, Wetter und ggf. anderer relevanter Parameter (Landnutzung, Topologie, Gegenmaßnahmen)

zu beschreiben, verstehen und vorherzusagen. Als Datenbasis zur Verbreitung dienen zunächst offizielle Daten der Verwaltung (Fallenfunde) – aktuell und historisch. Hinzu kommen Geobasisdaten, (möglichst hochaufgelöste) Landnutzungs- und Landbeckungsdaten (idealerweise sollten z.B. für Wälder und Obstplantagen die vorhandenen Baumarten bekannt sein) sowie (ebenfalls möglichst hoch aufgelöste) Wetterbeobachtungen und -vorhersagen. In Zukunft sind auch nutzergenerierte Daten vorstellbar (wenn bspw. Landwirte auch unabhängig von offiziellen Fallenfunden lokale Sichtungen melden oder wenn Bürger ehrenamtlich Sichtungen melden).

Die dominante Big Data Dimension ist in diesem Szenario ***Variety***. Mit Ausnahme von Wetterdaten, die auch große Datenmengen repräsentieren können, ist die Datenlage sogar im ersten Ansatz sehr viel *dünner* als man sich das wünschen würde.

Als Analyse-Fragestellungen können/sollen untersucht werden:

- **Deskriptiv**: Interpolation und Visualisierung der historischen Entwicklungen des Auftretens der invasiven Art (räumlich-zeitliche Muster), idealerweise korreliert mit möglichen erklärenden Faktoren.
- **Prädiktiv**: Vorhersage des Auftretens (Ausbreitungsprognosen, räumlich und zeitlich, sowie Kausalanalyse zu den Treibern der Ausbreitung).
- **Präskriptiv**: Planung des Einsatzes von semi-stationärer und/oder mobiler Sensorik zur Detektion von Schädlingen; effiziente Einsatzplanung von Teams/ Geräten zur Schädlingsbekämpfung (je nach betrachteter Art, z.B. manuelles Zerstören der Laichmöglichkeiten, Ausreißen der Pflanzen, Köderfallen, Insektizide etc.).

3.3 Anwendungsszenario 3: Smart City und Gesundheit

Ein immer wichtiger werdender Anwendungsbereich für GIS ist das Verstehen und Verbessern der Strukturen, Abläufe, und Zusammenhänge in Städten, mit dem Ziel, das Leben und die Gesundheit der Bürger zu verbessern. Konkret wird das urbane Mikroklima und seine Wirkung auf den Menschen beeinflusst von Faktoren wie Klima und Wetter (Temperatur, Wind, Luftfeuchtigkeit, direkte Sonneneinstrahlung, Regen), Luft (Feinstaub, Ozonwerte, Stickoxide, …), Lärm, Pollen und anderen allergenen Stoffen, der Landnutzung und -bedeckung, der Verteilung von städtischem Grün und Bäumen und der Architektur (Belüftungskorridore in der Bebauung, Beschattungseffekte, Speicherfunktion für Wärme, …).

Im Projekt wird ein Fokus auf die Temperatur gelegt und sog. *Urban Heat Islands* betrachtet. Gemäß Gesundheitsberichterstattung des Bundes sterben jedes Jahr zahlreiche Menschen an den Folgen von Hitze. Karlsruhe, als eine der wärmsten Städte in Deutschland, bildet hierbei zudem mehrere noch stärker ausgeprägte Hitzeinseln aus, bei denen die Temperatur um mehrere Grade höher sein kann als nur wenige hundert Meter entfernt. Aufgrund der geringen Dichte an Messstationen existieren jedoch bisher keine Modelle, die eine Analyse der Hitze- und Temperaturbelastungen für einzelne Wohnviertel oder gar für Straßenzüge oder bestimmte Plätze ermöglichen. Aufgrund einer erwarteten weiteren Erwärmung wäre ein besseres Verständnis für die Abhängigkeiten von der Geo-Position aber zwingend notwendig.

In BigGIS sollen historische Zeitreihen von Temperaturmessungen in Karlsruhe kombiniert werden mit Infrarot-Satellitenaufnahmen, welche die (Oberflächen-) Temperaturen allerdings lediglich in einer Granularität von mehreren hundert Metern in Frequenzen von einigen Wochen ermitteln können. Mit Hilfe dieser Daten, sowie (unzuverlässigen) von Bürgern gemeldeten Daten soll ein feingranulareres Temperaturmodell erstellt werden, welches von den Temperaturen an den Messstationen auf die restlichen Positionen im Stadtgebiet schließen lässt. Danach sollen Modelle entwickelt werden, die auf Basis von Wetter und beispielsweise Windprognosen für die Stadt Karlsruhe feingranulare innerstädtische Temperaturvorhersagen ermöglichen. Über Messfahrten zur Ermittlung von tatsächlichen Temperaturen sollen die Ergebnisse der Approximationen und Prognosen dann verifiziert werden, um Aussagen zur Modellgüte ableiten zu können und die Modelle zu kalibrieren. Da mobile Messungen und Messfahrten Kosten verursachen, sind die Messfahrten so zu planen, dass ein maximaler Informationsgewinn für die Modellierung erreicht werden kann.

Datenquellen außer den oben genannten können 3D-Stadmodelle sein, Daten privater Wetterstationen sowie Befliegungsdaten mit einem Thermalscan. Die dominante Big Data Dimension ist hier zunächst *Variety*, bei zunehmender Nutzung privat bereitgestellter Daten auch *Veracity*.

Als Analyse-Fragestellungen können/sollen untersucht werden:

- **Deskriptiv**: Interpolation bzw. Approximation und Visualisierung der Hitzebelastung aus grobgranularen, unvollständigen und teilweise unzuverlässigen Daten.
- **Prädiktiv**: Feinkörnige Vorhersage innerstädtischer Hitzebelastung.

- **Präskriptiv**: Planung von Messstationen und Messfahrten mit maximalem Informationsgewinn.

4 Software-Architektur für Spatial Big Data

Im bisherigen Projektverlauf lag der Schwerpunkt der Arbeiten auf einer genauen Analyse der Anwendungsszenarien und der entsprechenden Datenbeschaffung, auf Vor- und Zuarbeiten, der Schaffung der Basisinfrastruktur und auf punktuellen algorithmischen Überlegungen zu Visualisierungs- und Analyseverfahren. Eine einheitliche Projektarchitektur wurde noch nicht entwickelt, zumal die drei Anwendungsszenarien aus Big Data Sicht auch so unterschiedlich gelagert sind, dass noch nicht klar ist, ob eine einheitliche, umfassende Software-Lösung für alle drei Anwendungen sinnvoll ist. Trotzdem sind im Verlauf der Projektarbeiten einige Beobachtungen und Einsichten entstanden, die wir anhand der in Abbildung G-2 gezeigten, noch unvollständigen Illustration der Verarbeitungsideen in BigGIS erläutern.

(1) Als Basis-Framework zur Installation und Verteilung der Projektentwicklungen und zur betriebssystemunabhängigen, skalierbaren Ressourcenbereitstellung hat sich **Docker**[8] als sehr interessante Lösung herausgestellt.

(2) Grundsätzlich sollte man angesichts vieler schöner Marketing-Demos von Standalone Big Data Applikationen nicht übersehen, dass man sich im Bereich der Geo- und Umweltdaten sehr häufig innerhalb oder im Kontext von existierenden Softwarelandschaften (Geodateninfrastrukturen, GDI) der öffentlichen Verwaltung bewegt, die man nicht einfach durch neue Ansätze wird *ersetzen* können, sondern durch Spatial Big Data Lösungen *ergänzt* werden sollen, zu diesen also interoperabel sein müssen. Hier spielen bewährte (und ggf. weiter zu entwickelnde) **Standards** (insbesondere des **OGC**) eine große Rolle.[9] Für eine Spatial Big Data Software-Lösung sind sowohl auf Seite der Datenquellen entsprechende Schnittstellen zu schaffen (z.B. auf der Basis des OGC Sensor-Observation-Service für Sensordatenströme[10] oder mithilfe der OGC SensorThings API[11] für die

[8] https://de.wikipedia.org/wiki/Docker_(Software) und https://www.docker.com/
[9] Vgl. http://www.opengeospatial.org/projects/groups/bigdatadwg und http://external.opengeospatial.org/twiki_public/BigDataDwg/
[10] http://docs.opengeospatial.org/wp/07-165r1/
[11] https://en.wikipedia.org/wiki/SensorThings_API

Einbindung von IoT-Datenströmen) als auch auf der Seite der Bereitstellung von Berechnungsergebnissen durch entsprechende standardkonforme Schnittstellen (hier gewährleistet durch die Einbindung von Disy Cadenza[12], das die verschiedenen OGC Schnittstellen unterstützt).

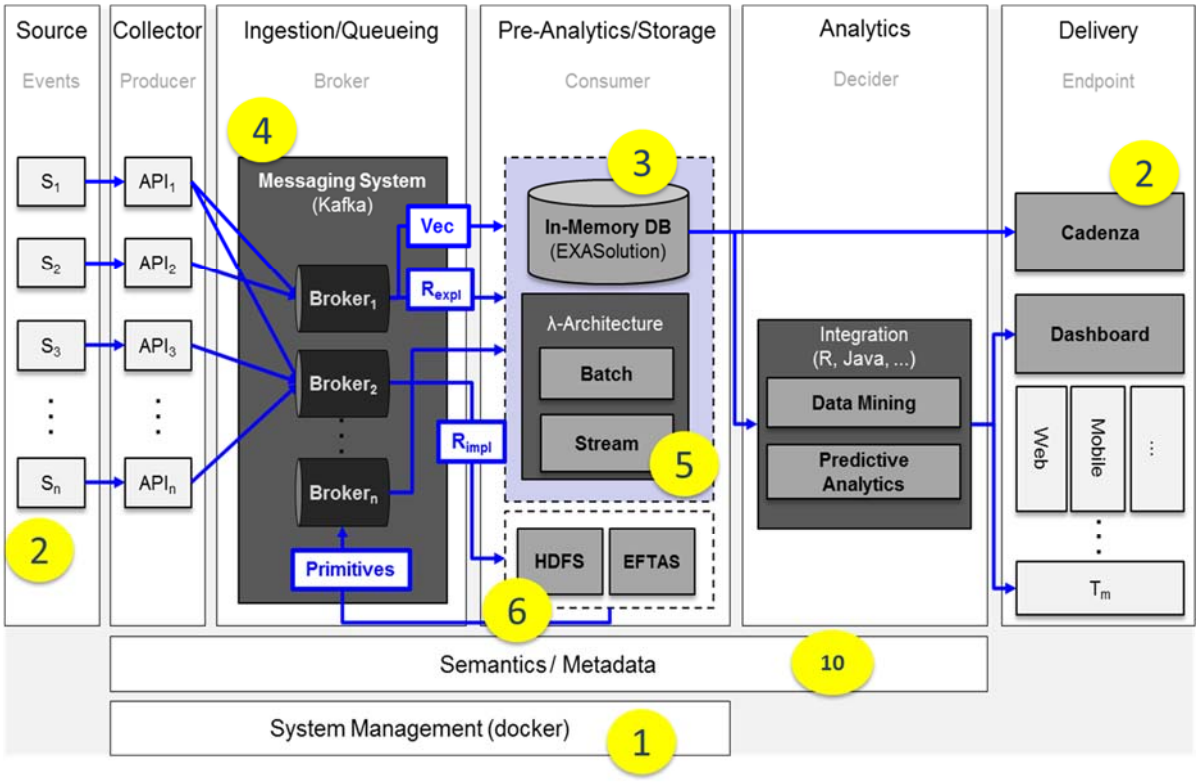

Abbildung G-2: Initiale Stufen-Architektur für (große) Geodaten(ströme)

(3) Eine weitere „einfache“ Erkenntnis ist die, dass gerade für kompliziertere Vektordatenverarbeitungen die Performanz moderner **In-Memory-Datenbanken** definitiv ein Gewinn sein kann. Disy hat hier bisher überwiegend mit kommerziellen Produkten gearbeitet (EXAsolution, SAP HANA, Oracle Spatial and Graph). Alle bieten bereits ähnliche „Standardfunktionalitäten“ für die Geodatenverarbeitung an, teilweise noch mit „Kinderkrankheiten“, aber in ständiger Verbesserung und mit sehr guten Performanzwerten. Diese Produkte können ohne Zweifel auch problemlos in existierende GDIen integriert werden. Eine interessante Option ist hier auch, dass meist die Möglichkeit gegeben ist, **R-Skripte** eng mit der Datenbanklösung zu verzahnen.

[12] http://www.disy.net/produkte/cadenza.html

(4) Zum Integrieren und Verteilen von Eingangsdaten bietet sich das Message Brokering System Apache **Kafka**[13] als de-facto Standard an, das sowohl mit sehr großen Datenmengen als auch sehr hoher Datenfrequenz umgehen kann.

(5) Die Datenstromverarbeitung und verschiedene Lernalgorithmen werden voraussichtlich auf Apache **Spark**[14] basieren, das eine Cluster-basierte Parallelisierung von Algorithmen zur Big Data Analyse darstellt.

(6) Für die Massendatenspeicherung und Batch-Verarbeitungen wird voraussichtlich auf das Apache Hadoop Distributed File System (**HDFS**) sowie proprietäre Lösungen von EFTAS für die Rohdaten aus der Fernerkundung zurückgegriffen.

(7) Auch über (6) hinausgehend (und in der Abbildung noch nicht dargestellt), spielen ohne Zweifel **Rohdaten aus der Fernerkundung** eine besondere Rolle. Nicht nur, dass sie in unseren bisherigen Szenario-Analysen die einzige Quelle für tatsächlich sehr signifikante *Datenvolumina* darstellen. Es ist auch so, dass Rohdaten zwar eventuell für spätere Neu-/Nachuntersuchungen gespeichert werden sollten, aber normalerweise nicht *per se* von Interesse sind, sondern nur nach mehr oder weniger komplexen Interpretationsprozessen, bis hin zur semantischen Objekterkennung (Bildanalyse aus Raster- zu interpretierten Vektordaten). Dies wird im Bild dadurch angedeutet, dass einerseits Rasterdaten (**R**) und andererseits Vektordaten (**Vec**) fließen, wobei die Rasterdaten noch weitgehend uninterpretiert sein können (die gefragte Information also noch implizit enthalten ist - $\mathbf{R_{impl}}$) oder aber bereits das Ergebnis einer Klassifikation / Analyse sein können (so dass die gefragte Information schon explizit im Rasterbild dargestellt ist - $\mathbf{R_{expl}}$). Die selben Rohdaten können auch mit verschiedenen Algorithmen nach verschiedenen Merkmalen untersucht werden, so dass die späteren Verarbeitungsschritte den früheren Interpretationsschritten mitteilen müssen, mittels welcher **Primitive** die gefragte Information repräsentiert werden soll. Es ist außerdem zu überlegen, ob nicht für Fernerkundungsrohdaten neben den üblichen Datenpfaden eine spezielle „Abkürzung“ (an den üblichen Daten-Einlagerungsprozessen vorbei) erforderlich ist und ob man *überhaupt* alle Rohdaten in die Big Data Plattform überführen will und kann. Wenn bspw. in einem Notfallszenario eine gewisse Nah-Echtzeit

[13] http://kafka.apache.org/ und https://en.wikipedia.org/wiki/Apache_Kafka
[14] http://spark.apache.org/ und https://de.wikipedia.org/wiki/Apache_Spark

Funktionalität unerlässlich ist und daher aus einem UAS durch einen Downlink die Daten *während des Fluges* direkt per Funk übertragen werden sollen, kann es durchaus vorkommen, dass die Datenübertragung langsamer ist als die Datenerzeugung. In diesem Fall muss untersucht werden, ob nicht schon im UAS eine Vorselektion oder sogar schon eine Vorinterpretation durchgeführt wird, so dass nur noch vermutlich relevante oder bereits voranalysierte Daten ins Back-End System übertragen werden.

(8) Was ebenfalls in der Grafik noch nicht dargestellt wird: in realen komplexeren Anwendungsszenarios wird man auch **interaktivere Aufgaben** bearbeiten müssen, bei denen der Anwender manuell mit Daten und Algorithmen arbeitet (vgl. Abbildung G-1), sich bestimmte Datenselektionen oder -sichten für die weitere Verarbeitung definiert, ggf. ganze Workflows für die Datenverarbeitung konfiguriert usw. – wie das aus dem Data Mining schon lange bekannt ist. Dieses Element des benutzerfreundlichen Umgangs mit Big Data Algorithmen und Datenpools wird leicht übersehen, weil die sichtbareren Big Data Anwendungen häufig im Bereich der Vollautomatisierung angesiedelt sind und durch reine unidirektionale Pipeline-Architekturen realisiert werden können. Dennoch ist hier noch einiges an Entwicklungsarbeit für praxistaugliche Werkzeuge zu leisten. In der Grafik wird dies dadurch angedeutet, dass auf der Delivery-Seite (beim Endanwender) das Cadenza-Tool dargestellt ist, das dem Data Analyst / Data Scientist Werkzeuge und Benutzungsoberflächen für die Datenanalyse zur Verfügung stellen soll. Wenn man die Software-Architektur dahingehend komplettiert, ergeben sich daraus insbesondere **Feedback-Schleifen**, weil u.U. Ergebnisse oder Zwischenergebnisse von Berechnungen (ggf. mit Metadaten zu ihrer Erzeugung) wieder in die Datenpools zurückgespeist werden müssen.

(9) Eine weitere allgemeine Bemerkung, die in der Grafik schwer an einem Punkt zu verorten ist: In marketing-orientierten Publikationen zu Big Data wird leicht der Eindruck vermittelt, die neuen Technologien würden allein aufgrund der schieren Datenmenge in Verbindung mit intelligenten Algorithmen die (aufwändigen) ETL-Prozesse des klassischen Data Warehousing unnötig machen. Tatsächlich ist es häufig aber nur so, dass man zur schnellen Verarbeitung der großen Datenmengen zunächst einmal alle verfügbaren Daten erfasst und dann erst später im Batch-Betrieb syntaktische, semantische oder qualitätssichernde Transformationen

durchführt. Aus ETL wird also **ELT**. Gerade im Geodatenbereich, wo semantische Heterogenitäten oft noch stärker hervortreten als in anderen Domänen und wo schon eine unterschiedliche räumliche oder zeitliche Auflösung von Datenreihen diese unvergleichbar werden lassen kann, gehen wir davon aus, dass sich Transformationsprozesse in aller Regel nicht komplett vermeiden lassen. Da außerdem in unseren vorliegenden Szenarioanalysen die deutlich dominante Problemdimension *Variety* ist (und nicht *Volume*), ist auch vorstellbar, dass einige „T-Prozesse“ des ETL/ELT schon früh in der Pipeline und eventuell auch als Teil der Datenstromverarbeitung möglich oder sogar notwendig sind.

(10) Geht man davon aus, dass wie bereits mehrfach gesagt, *Variety* die wichtigste Big Data Dimension in unseren Szenarien darstellt und daher komplexe ELT-Prozesse nicht zu vermeiden sind, stellt sich die Frage, ob nicht zumindest die Konfiguration dieser ELT-Prozesse teil- oder vollautomatisiert werden kann. Hierfür sollen im Projekt wissensbasierte Methoden auf der Basis **semantischer Metadaten** genutzt werden. Diese können beschreiben, welche Algorithmen für welche Datencharakteristika geeignet sind, wie gewisse Datenpools entstanden sind, wie gewisse Daten hinsichtlich ihrer Zuverlässigkeit einzuschätzen sind usw.

5 Zusammenfassung

Wir haben die Motivation, Struktur, wesentliche Grundlagen und die Zielsetzungen des BigGIS-Projekts dargestellt. Dabei wurden die Datengrundlagen und die deskriptiven, prädiktiven und präskriptiven Analysefragestellungen der drei Anwendungsszenarien Katastrophenschutz, Umweltmonitoring und Smart City und Gesundheit vorgestellt. Dominante Big Data Dimension ist immer *Variety*; *Volumen* und *Velocity* können durch die Einbindung von Methoden der Fernerkundung hinzukommen, *Veracity* durch die Verwendung nutzergenerierter Daten. Erste Überlegungen zu einer Software-Architektur für das Projekt bauen auf die Frameworks Docker, Kafka und Spark, haben im Kern die In-Memory Datenbank von Exasol, erweitert um R-Skripte, sowie HDFS für die Massendatenspeicherung. Spezifische Fragestellungen, denen man sich im Projekt widmet, sind unter anderem die effiziente Behandlung von Fernerkundungsrohdaten und -interpretationsprozessen sowie die Verwendung semantischer Metadaten für komplexe Transformationsprozesse und effektive Benutzerinteraktion. Gerade auch

angesichts der *Variety* und *Veracity* Fragestellungen in unseren Anwendungsfällen sehen wir eine große Herausforderung im Überbrücken der Lücke zwischen komplexer Datenlage, vielfältigen Algorithmen und problemspezifischen, endanwendertauglichen Benutzungsoberflächen. Außerdem liegt ein Schwerpunkt der konkreten Projektarbeiten auf algorithmischen Fragestellungen bei den deskriptiven, prädiktiven und präskriptiven Aufgaben in den jeweiligen Szenarien.

Danksagung. *Teile der Texte, Bilder und Darstellungen gehen auf verschiedene Partner im BigGIS-Projektkonsortium zurück. BigGIS (http://biggis-project.eu/) wird finanziell unterstützt vom Bundesministerium für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01IS14012A bis 01IS14012G. Weitere Unterstützung findet das Projekt bei seinen assoziierten Partnern, insbesondere der Stadt Karlsruhe sowie beim Staatlichen Weinbauinstitut Freiburg, das für die Forschungen zur Kirschessigfliege seine Daten beisteuert.*

6 Literaturverzeichnis

[BITKOM, 2014]

BITKOM-Arbeitskreis Big Data: Big-Data-Technologien – Wissen für Entscheider. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V., Berlin, 2014. URL: https://www.bitkom.org/Bitkom/Publikationen/Big-Data-Technologien-Wissen-fuer-Entscheider.html . (Letzter Zugriff: 03.08.2016)

[Klein et al, 2014]

L.J. Klein, F.J. Marianno, C.M. Albrecht, M. Freitag, S. Lu, N. Hinds, X. Shao, S. Bermudez Rodriguez, H.F. Hamann: PAIRS: A scalable geo-spatial data analytics platform. In: BIG DATA '15 - Proceedings of the 2015 IEEE International Conference on Big Data, p. 1290-1298, IEEE Computer Society, Washington DC, USA, 2015.

[Oracle, 2016]

Oracle Inc.: Oracle Big Data Spatial and Graph. Oracle Data Sheet. URL: http://download.oracle.com/otndocs/products/bigdata-spatialandgraph/bdsg-data-sheet.pdf . Siehe auch: http://docs.oracle.com/bigdata/bda45/index.htm . Oracle Inc., 2016. (Letzter Zugriff: 03.08.2016)