OFFIS Gesundheit
INSTITUT FÜR INFORMATIK

Analyse von strukturierten und unstrukturierten Daten in Analytischen Informationssystemen

SE 2014 – Doctoral Symposium

Mirco Josefiok

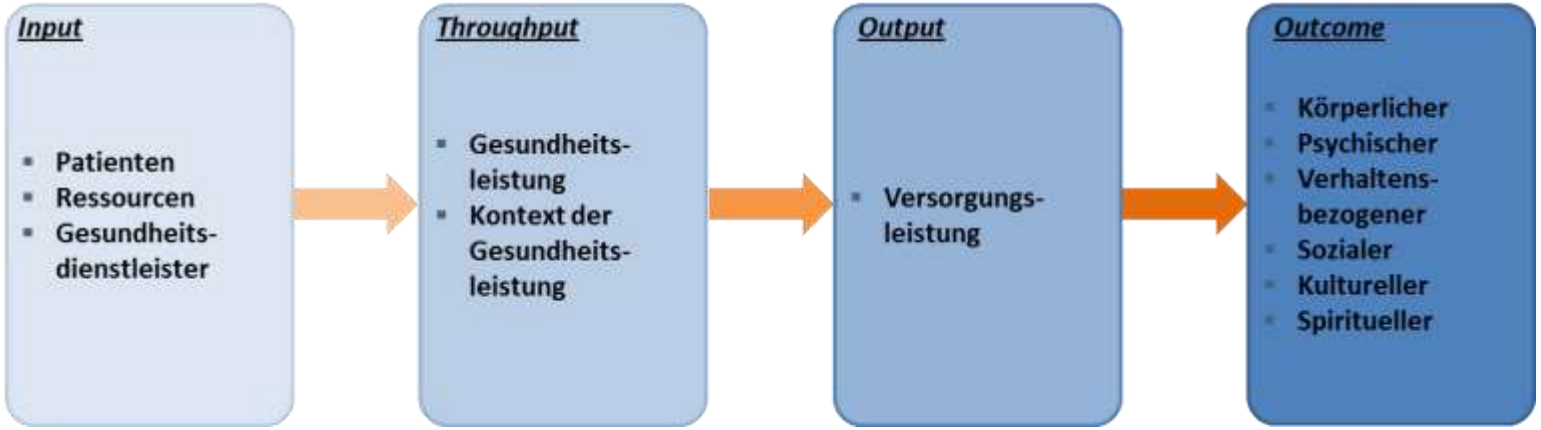
26.02.2014

2 Anwendungsfall Versorgungsforschung

Versorgungsforschung ist ein fachübergreifendes Forschungsgebiet, das grundlagenorientiert

- den Versorgungsbedarf (Input),
- die Versorgungsstrukturen bzw. Prozesse und Ergebnisse (Throughput),
- die erbrachten Versorgungsleistungen (Output) und
- den Zugewinn an Gesundheits- bzw. Lebensqualität (Outcome)

beschreibt und die Bedingungszusammenhänge kausal erklärt,



3 Datenquellen für die Versorgungsforschung

Klassische Datenquellen

- Krankenhäuser
- Niedergelassene Ärzte
- Krankenkassen
- Apotheken
- Reports zum Gesundheitswesen
- Personal Health Records
- Krankenhausdiagnosestatistik

Neue Datenquellen

- Mobile Geräte
- Apps
- Social Media
- Sensoren

Daten, die für die Versorgungsforschung relevant sind, aber nicht aus dem Gesundheitswesen kommen, werden in den kommenden Jahren zunehmen.

[Safran2012, pfaff2011lehrbuch]

4 Datenquellen für die Versorgungsforschung

Klassische Datenquellen

- **Krankenhäuser**
- Niedergelassene Ärzte
- Krankenkassen
- Apotheken
- Reports zum Gesundheitswesen
- **Personal Health Records**
- Krankenhausdiagnosestatistik

Neue Datenquellen

- Mobile Geräte
- Apps
- Social Media
- **Sensoren**

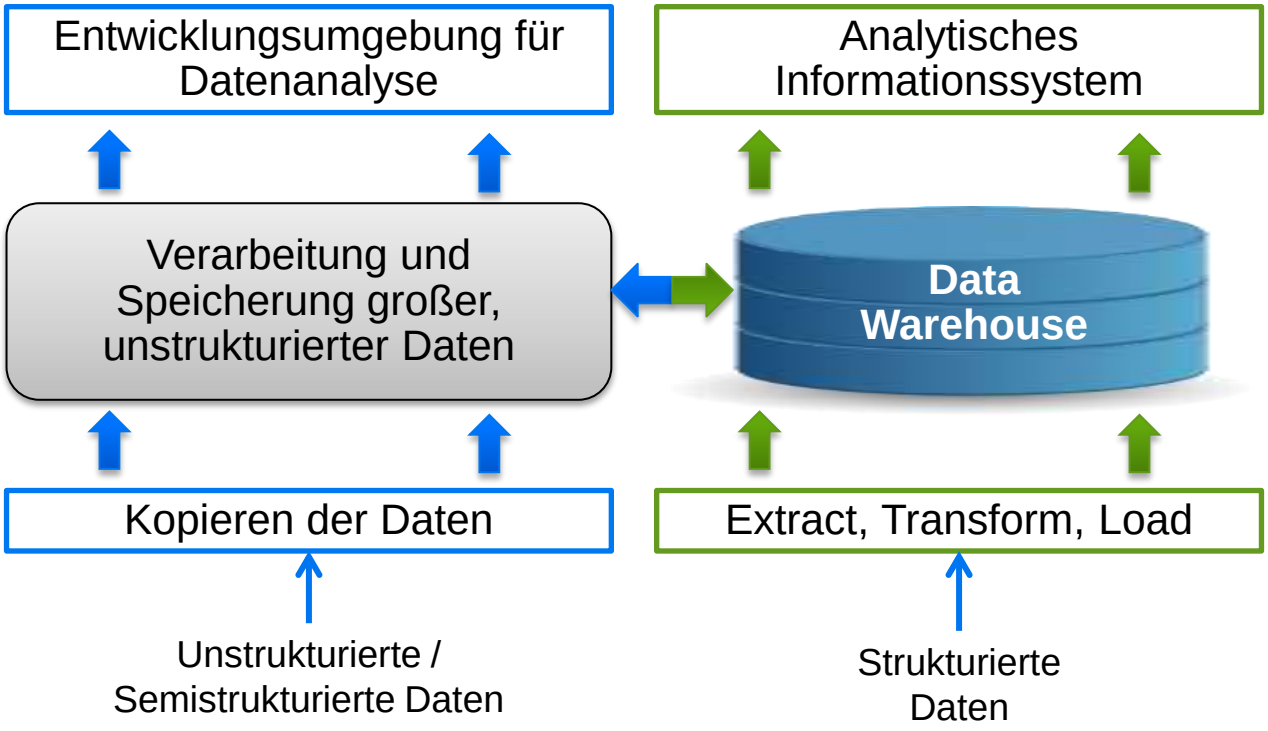
[Safran2012, pfaff2011lehrbuch]

5 Herausforderungen bei der Datenanalyse für die Versorgungsforschung

- Für die Beantwortung von Fragestellungen im Rahmen der Versorgungsforschung werden verschiedene Datenquellen herangezogen:
  - Neben Primärdaten (Erhebungen, Befragungen etc.) auch immer mehr Sekundärdaten
  - Sekundärdaten sind oft schnell und sehr umfangreich verfügbar, aber zumeist nicht für die wissenschaftliche Verwendung aufbereitet
- Für genauere Analysen und Vorhersagen werden sukzessive weitere Datenquellen erschlossen:
  - Setzt einen komplexen und langwierigen Integrationsprozess voraus
  - Ggf. nicht möglich alle Daten aus unterschiedlichen Quellen in eine einheitliche Struktur zu bringen
- Erkenntnisgewinn kann durch die notwendige Aufbereitung der Daten gemindert werden
- Nicht jedes Unternehmen / Einrichtung verfügt über die notwendigen Mittel für den Aufbau einer geeigneten Infrastruktur

[pfaff2011lehrbuch, Schaeffer2013]

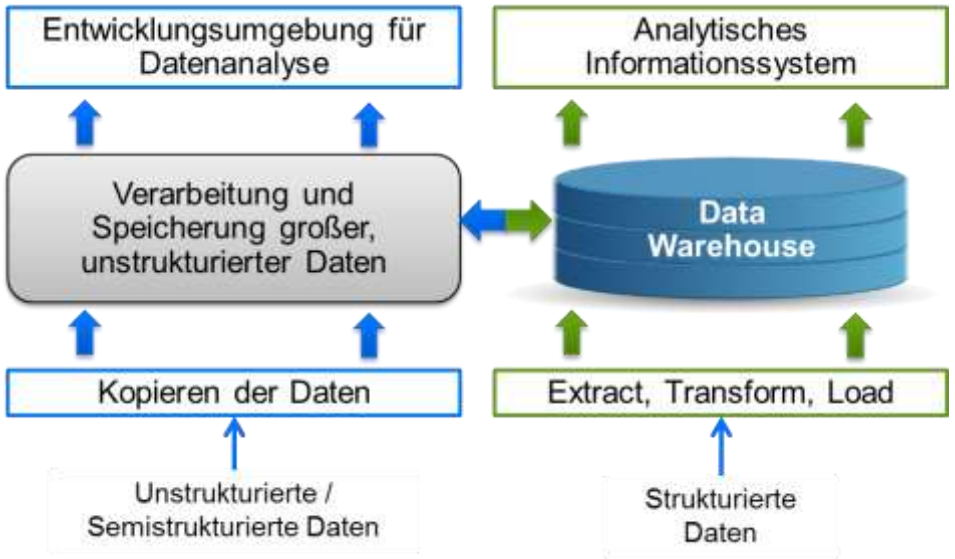
6 Problemstellung



- AIS sind bislang **nicht** in der Lage, Analysen von **unstrukturierten** und **strukturierten** Daten durchzuführen
- Für Unternehmen ist es i.d.R. **nicht wünschenswer**t mehrere Analyseplattformen zu betreiben

[Awadallah2013]

Es besteht die Notwendigkeit für ein Analytisches Informationssystem, welches sowohl strukturierte, als auch unstrukturierte Daten innerhalb einer Analyse verarbeiten kann.



8 Forschungsfrage

„Wie kann ein AIS so erweitert werden, dass eine Verbesserung der Analysequalität mittels einer gemeinsame Analyse von strukturierten und unstrukturierten Daten erreicht wird?“

9 Eigener Ansatz

Artefakte

- Konzept
  - fachliches Konzept beschreibt Fragestellungen, Nutzungsszenarien und Auswertungsmöglichkeiten für kombinierte Analysen
  - technisches Konzept beschreibt Integration von AIS mit Big Data-Technologien
- Prototyp
  - Entwicklung eines integrierten AIS auf Basis des Konzepts
  - Bereitstellung des Prototypen als Service
  - Technische Machbarkeit zeigen
  - Vorarbeiten aus Projekten nutzen
- Softwarearchitektur
  - Erfahrungen und Ergebnisse in eine Softwarearchitektur nach IEEE 42010 überführen
  - Generalisierung / Wiederverwertbarkeit der Erkenntnisse sicherstellen

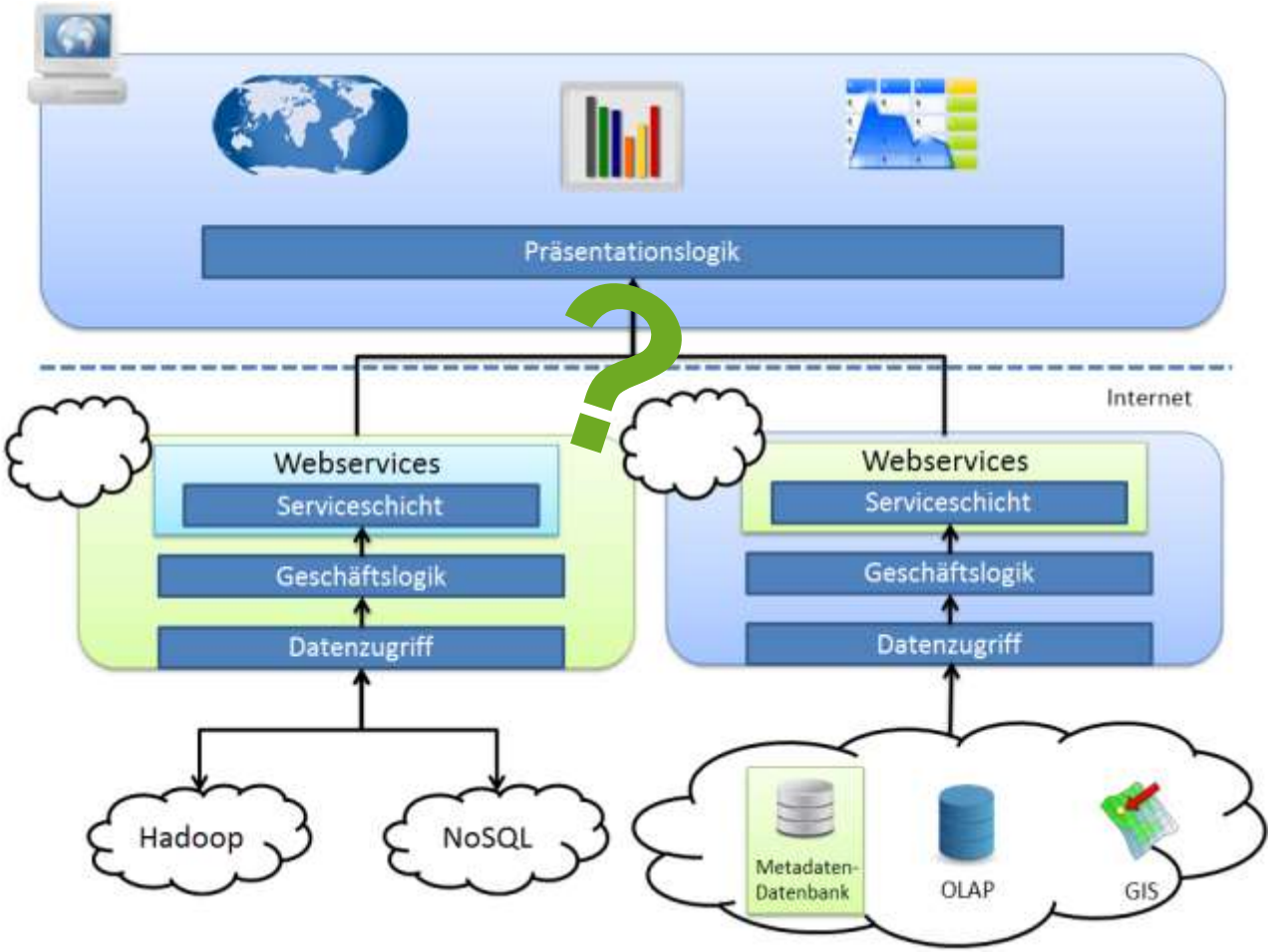
10 Eigener Ansatz

Konzept

- Fachliches Konzept
  - Herausstellen von geeigneten Fragestellungen
  - Erarbeiten von möglichen Antwortszenarien
  - Ermitteln von gewünschten Analyseverfahren
  - Erheben von entsprechenden Auswertungsmechanismen
- Technisches Konzept
  - Ermitteln der technischen und funktionalen Anforderungen
  - Integration von herkömmlichen AIS mit Big Data-Technologien
  - Entwerfen einer Analyseplattform für die Analyse von strukturierten und unstrukturierten Daten

11 Eigener Ansatz

Prototyp



12 Eigener Ansatz

Softwarearchitektur

- Beschreibung der Softwarearchitektur nach IEEE 42010
  - dies umfasst die eigentliche Beschreibung der Architektur
  - die Sicht auf die Architektur entlang eines definierten Standpunktes welche durch das Anliegen der Stakeholder festgelegt wurde
- Inhalte der Softwarearchitektur können sich nach dem Rational Unified Process richten

User Interface	• Standards und Werkzeuge
Business	• Best Practices • Standards und Werkzeuge für die Umsetzung
Middleware	• Transaktionsmanagement • API Design
System Software	• System Management • Auswahlprozess

13 Verwandte Arbeiten

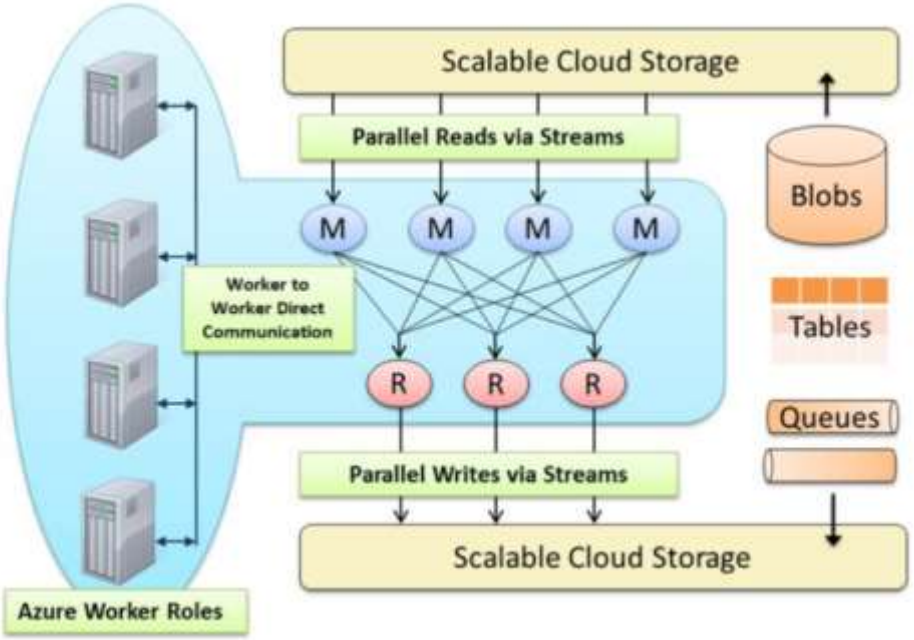
Data Warehouseing and OLAP over Big Data

- Konzeptionelle Arbeiten [Cuzzocrea2011, Cuzzocrea2013, Cuzzocrea2013a]
- Identifikation von relevanten Technologien
  - z.B. Hadoop, Hive, Cloud etc.
- Erarbeiten von offene Fragestellungen und möglichen Forschungsrichtungen
  - Erstellen von OLAP Cubes
  - Aggregation von Big Data-Daten verbessern
  - Effizienteres Verarbeiten von Daten
  - Erarbeiten von passenden Architekturen
- Kein Fokus auf die Erweiterung von AIS

14 Verwandte Arbeiten

Project Daytona

- Bereitstellung von Analysediensten als Cloud-Service
- Basierend auf Diensten von Windows Azure
- Einfache Erstellung eigener (einfacher) Algorithmen für die Analyse
- Upload der eigenen Daten zu Windows Azure
- Ausführen der definierten Algorithmen
- Beispielclient für das Überwachen der Aufgaben und Ansicht der Ergebnisse
- Keine (direkte) Einbeziehung strukturierter Daten

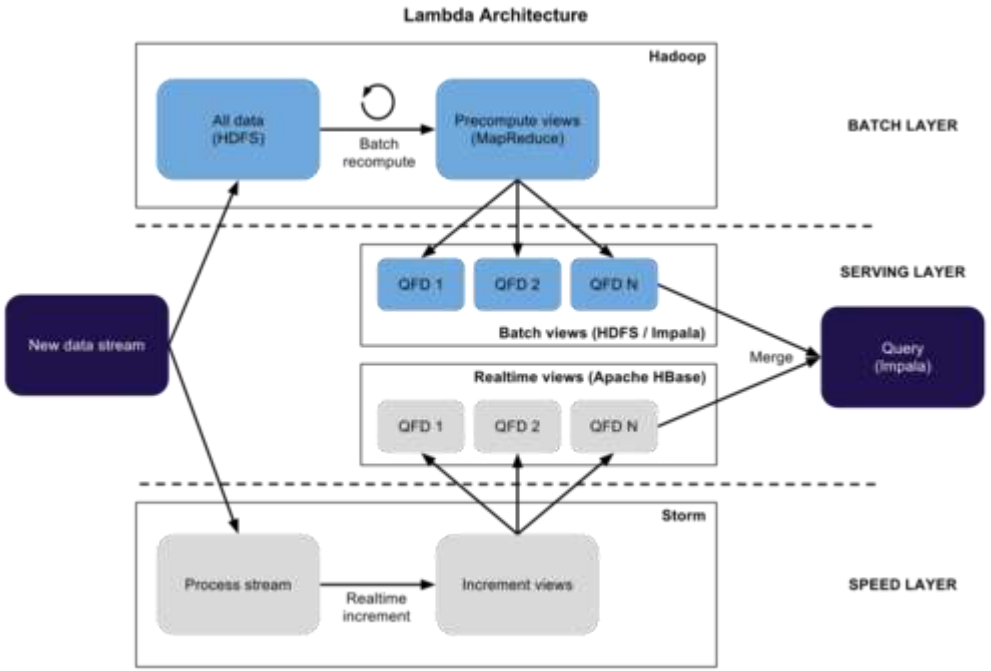


[Barga2012]

15 Verwandte Arbeiten

Lambda Architecture

- Architektur für die verteilte Verarbeitung großer Datenmengen
- Entstanden bei Twitter
- Ausschließlich technischer Ansatz
- Keine Einbeziehung strukturierter Daten



[Marz2014]

16 Verwandte Arbeiten

Übersicht

	strukt. Daten	unstruk. Daten	Integration mit AIS	Datenübernahme in DWH
Cuzzocrea et al.	++	++	-	+
Project Daytona	o	++	o	o
Lambda	-	++	-	+
Ansatz MJo	++	+	++	o

++ vorhanden; + tlw. Vorhanden; - nicht vorhanden; o nicht im Fokus der Betrachtung

17 Evaluation

- Prototypische Umsetzung und Überführung in Softwarearchitektur
  - Realisierbarkeit der Konzepte zeigen
  - Mögliche Generalisierung durch Beschreibung in einer Softwarearchitektur herbeiführen
- Fallstudie in der Versorgungsforschung
  - Voraussetzung:
    - Strukturierte und unstrukturierte Daten
    - Fachexperten
    - Eigene Expertise
  - Gemeinsam mit Fachexperten soll ein Konzept für die Anwendung erarbeitet werden

18 Zusammenfassung

- Ziel dieser Arbeit ist die Beantwortung der Forschungsfrage
  - Wie kann ein AIS so erweitert werden, dass eine Verbesserung der Analysequalität mittels einer gemeinsame Analyse von strukturierten und unstrukturierten Daten erreicht wird?
  - Domäne der Anwendung ist die Versorgungsforschung
- Dazu werden
  - Ein Konzept für eine integrierte Analyselösung zur gemeinsamen Analyse von strukturierten und unstrukturierten Daten entwickelt
  - Ein Prototyp entwickelt, mit welchem die technische Machbarkeit gezeigt wird
  - Die Ergebnisse und Erfahrungen in eine Softwarearchitektur überführt
- Verwandte Arbeiten werden berücksichtigt
  - Bisher keine Konzepte für eine Zusammenführung von strukturierten und unstrukturierten Daten bekannt

19 Literatur

[Awadallah2013] Awadallah, Amr ; Graham, Dan: Hadoop and the Data Warehouse: When to Use Which, 2013

[Barga2012] Barga, Roger S. ; Ekanayake, Jaliya ; Lu, Wei: Project Daytona: Data Analytics as a Cloud Service. In: 2012 IEEE 28th International Conference on Data Engineering : IEEE, 2012 — ISBN 978-0-7695-4747-3, pp. 1317–1320

[Bartel2012] Bartel, Jörg ; Böken, Arnd ; Decker, Björn ; Falkenberg, Guido ; Guzek, Robert ; Janata, Steve ; Keil, Thomas ; Kisker, Holger ; Konrad, Ralf ; et al.: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Berlin, 2012

[baron2013big] Baron, P: Big Data für IT-Entscheider: Riesige Datenmengen und moderne Technologien gewinnbringend nutzen : Hanser Fachbuchverlag, 2013 — ISBN 9783446433397

[BauerGünzel200812] Bauer, A. ; Günzel, H. (eds.): Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung. 3., überar. ed. : dpunkt Verlag, 2008 — ISBN 9783898645409

[Chamoni2010] Chamoni, P. ; Gluchowski, P.: Analytische Informationssysteme - Einordnung und Überblick. In: Analytische Informationssysteme : Business Intelligence-Technologien und -Anwendungen, 2010 — ISBN 9783540292869, pp. 3–16

20 Literatur

[Cuzzocrea2013] CUZZOCREA, ALFREDO ; SONG, IY ; DAVIS, KC: Analytics over large-scale multidimensional data: the big data revolution! In: 14th international workshop on Data (2011), pp. 101–103 — ISBN 9781450309639

[Cuzzocrea2011] CUZZOCREA, ALFREDO ; BELLATRECHE, L ; SONG, IY: Data warehousing and OLAP over big data: current challenges and future research directions. In: … on Data warehousing and OLAP (2013), pp. 67–70 — ISBN 9781450324120

[Cuzzocrea2013a] CUZZOCREA, ALFREDO: Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures. In: 2013 IEEE 37th Annual Computer Software and Applications Conference, Ieee (2013), pp. 481–483 — ISBN 978-0-7695-4986-6

[Marz2014] MARZ, NATHAN ; WARREN, JAMES: Big Data Principles and best practices of scalable realtime data systems : Manning Publications Company, 2014

[Safran2012] SAFRAN, CHARLES: The Imperative of Big Data for Public Health Transformation, 2012

[Schaefer2013] SCHAEFFER, DM ; OLSON, PC: Big Data Options For Small And Medium Enterprises (2013), Nr. 2010, pp. 209–213