Lfd.Nr.

Best Practice für Big Data Projekte

Leitfaden aus #Big Data in #Austria

Mario Meir-Huber, IDC
Martin Köhler, AIT

Bundesministerium für Verkehr, Innovation und Technologie

Danksagung

Der Leitfaden „Best Practice für Big Data Projekte“ wurde im Rahmen der Studie #Big Data in #Austria im Rahmenprogramm „IKT der Zukunft“ der Österreichischen Forschungsförderungsgesellschaft (FFG) vom Bundesministerium für Verkehr, Innovation und Technologie (BMVIT) beauftragt und gefördert. Wir bedanken uns bei den Fördergebern und bei folgenden Personen und Institutionen für deren wertvollen Beitrag welcher die erfolgreiche Durchführung der Studie ermöglicht hat: Diskussionsleitern des Startworkshops, Teilnehmern des Startworkshops, allen Teilnehmern der Umfragen, allen Interviewpartnern und bei den Teilnehmern der Disseminationsworkshops für wertvolle Kommentare, für wertvollen Input von Partnern aus der Wirtschaft und der Forschung in Bezug auf Informationen zu Anforderungen, Potenzialen und insbesondere zu konkreten Projektumsetzungen. Im Besonderen bedanken wir uns für die Unterstützung durch die Österreichische Computer Gesellschaft (OCG), den IT Cluster Wien, den IT Cluster Oberösterreich, Kollegen der IDC und des AIT, DI (FH) Markus Ray, Dr. Alexander Wöhrer, Prof. Peter Brezany, Dipl. Ing. Mag. Walter Hötzendorfer, Thomas Gregg, Gerhard Ebinger und Prof. Siegfried Benkner.

Projektpartner

IDC Central Europe GmbH
Niederlassung Österreich
Währinger Straße 61
1090 Wien

Austrian Institute of Technology GmbH
Mobility Department, Geschäftsfeld Dynamic Transportation Systems
Giefinggasse 2
1210 Wien

IDC Österreich bringt als international führendes Marktforschungsinstitut langjährige Erfahrung in der Marktanalyse, Aufbereitung und Sammlung von Wissen, Trendanalysen sowie deren Dissemination mit, welche für das Projekt essenziell sind. IDC hat bereits eine Vielzahl von Studien über Big Data erstellt, welche in diese Studie einfließen.

AIT Mobility bringt als innovatives und erfolgreiches Forschungsunternehmen einerseits langjährige Forschungs- und Lehrerfahrung im Bereich Big Data ein und liefert andererseits eine breite thematische Basis für domänenspezifische Analysen (z. B. im Verkehrsbereich). Darüber hinaus hat AIT langjährige Erfahrung in der Durchführung und Konzeption von wissenschaftlichen Studien.

Autoren

Mario Meir-Huber ist Lead Analyst Big Data der IDC in Zentral- und Osteuropa (CEE). In seiner Rolle beschäftigt er sich mit dem Big Data Markt in dieser Region und verantwortet die IDC Strategie zu diesem Thema. Hierfür befasst er sich nicht nur mit innovativen Konzepten und Technologien, sondern teilt sein Wissen mit Marktteilnehmern und Anwendern in diesem Feld. Vor seiner Zeit bei IDC war Mario Meir-Huber bei mehreren Unternehmen in führenden Rollen tätig und hat 2 Bücher über Cloud Computing verfasst. Für die Österreichische Computer Gesellschaft (OCG) ist er der Arbeitskreisleiter für Cloud Computing und Big Data.

Dr. Martin Köhler ist Scientist am Mobility Department der AIT Austrian Institute of Technology GmbH und beschäftigt sich mit Forschungsfragen bezüglich der effizienten Anwendung von Big Data und Cloud Computing Technologien im Mobilitätsbereich. Martin Köhler hat durch seine Mitarbeit an internationalen sowie nationalen Forschungsprojekten langjährige Forschungserfahrung in diesen Themenbereichen und ist Autor von zahlreichen wissenschaftlichen Publikationen. Neben dieser Tätigkeit ist er Lektor für Cloud Computing und Big Data an den Fachhochschulen Wiener Neustadt, St. Pölten und Technikum Wien sowie an der Universität Wien. Seit Herbst 2013 ist er einer der Leiter der OCG Arbeitsgruppe „Cloud Computing und Big Data“.

1 Einleitung

Der effiziente Einsatz von Big Data Technologien kann MarktteilnehmerInnen zusätzliches Wissen liefern, aber auch neue innovative Dienste generieren, die einen erheblichen Mehrwert mit sich bringen. Andererseits erfordert der effiziente Einsatz von Big Data Technologien fachspezifisches Wissen in unterschiedlichen Bereichen, beginnend bei den zu verwendenden Technologien, über Wissensextraktion und Aufbereitung, bis hin zu rechtlichen Aspekten. Diese breite thematische Aufstellung des Themas Big Data, von den rein technischen Aspekten für die Datenspeicherung und Verarbeitung, dem Wissensmanagement, den rechtlichen Aspekten für die Datenverwendung sowie den daraus resultierenden Potenzialen für Unternehmen (neue Dienste und Produkte sowie neues Wissen generieren) erfordert umfassende Kompetenzen für deren effiziente Umsetzung in spezifischen Projekten und für die Entwicklung neuer Geschäftsmodelle.

Dieser Leitfaden bietet Organisationen Richtlinien für die Umsetzung und Anwendung von Big Data Technologien um das Ziel der erfolgreichen Abwicklung und Implementierung von Big Data in Organisationen zu erreichen. Es werden Best Practices Modelle anhand von spezifischen Big Data Leitprojekte im Detail erörtert und anschließend wird auf dessen Basis ein spezifischer Leitfaden für die Umsetzung von Big Data Projekten in Organisationen präsentiert. In diesem Rahmen wird ein Big Data Reifegrad Modell, ein Vorgehensmodell, eine Kompetenzanalyse, unter anderem des Berufsbilds Data Scientist, sowie eine Referenzarchitektur für die effiziente Umsetzung von Big Data vorgestellt.

2 Identifikation und Analyse von Big Data Leitprojekten

Big Data umfasst unterschiedliche Technologieebenen und Charakteristiken. Des Weiteren werden Big Data Technologien in vielen verschiedenen Domänen eingesetzt und es ergeben sich jeweils differenzierte Anforderungen und Potenziale. In Kapitel **Fehler! Verweisquelle konnte icht gefunden werden.** wurde der Bereich Big Data definiert und es wurden Technologien sowie Marktteilnehmer in Österreich erhoben. In Kapitel **Fehler! Verweisquelle konnte nicht gefunden werden.** wurden Domänen identifiziert in welchen Big Data als relevante Technologie schon Einzug gehalten hat und zukünftige Potenziale gesehen werden. Diese domänenspezifischen Anforderungen und Potenziale sind gesondert herausgearbeitet. In vielen der oben analysierten Domänen wurden oder werden Big Data Projekte durchgeführt welche großen Einfluss auf die bereitgestellten Technologien, aber vor allem auf die jeweilige Domäne haben können. In diesem Kapitel werden mehrere Projekte aus unter anderem aus den Bereichen Mobilität und Gesundheitswesen vorgestellt und im Detail auf Big Data relevante Aspekte analysiert. Die vorgestellten Projekte bilden viele der essenziellen Technologien ab und haben großes Potenzial, den österreichischen Markt in Bezug auf Big Data weiterzuentwickeln. Des Weiteren wurden im Rahmen der Studie viele zusätzliche Projekte aus den unterschiedlichsten Bereichen im Rahmen von Interviews erhoben und diskutiert. Die Erkenntnisse aus diesen Erhebungen sind in der Umsetzung des Leitfadens sowie der Markt und Potenzialanalyse eingearbeitet.

Hier finden sie eine Kurzübersicht über die nachfolgend im Detail beschriebenen Projekte:

Projektname	Domäne	Projekttyp	Leitung (Österreich)	Abstract
Real-Time Data Analytics	Mobility	Internes Projekt	Austrian Institute of Technology	Bereitstellung einer flexiblen Infrastruktur für die Echtzeitanalyse von großen und dynamischen Datenströmen für mobilitätsbezogene Forschung
VPH-Share	Medizin	EU Projekt	Universität Wien	VPH-Share entwickelt die Infrastruktur und Services für (1) die Bereitstellung und gemeinsame Benutzung von Daten und Wissen, (2) die Entwicklung von neuen Modellen für die Komposition von Workflows, (3) die Förderung von Zusammenarbeit unterschiedlicher Stakeholder in dem Bereich Virtual Physiological Human (VPH).
Leitprojekt Handel	Handel	Nationales Projekt		Echtzeitdatenanalyse in Filialen
Katastrophenerkennung mit TRIDENC	Industrie	Nationales Projekt	Joanneum Research	Konzeption und Entwicklung einer offenen Plattform für interoperable Dienste, die ein intelligentes, ereignisgesteuertes Management von sehr großen Datenmengen und vielfältigen Informationsflüssen in Krisensituationen
Prepare4EODC	Weltraum	Nationales Projekt	TU Wien	Förderung der Nutzung von Erdbeobachtungsdaten für die Beobachtung globaler Wasserressourcen durch eine enge Zusammenarbeit mit Partnern aus der Wissenschaft, der öffentlichen Hand als auch der Privatwirtschaft.

Tabelle 1: Überblick der analysierten Projekte

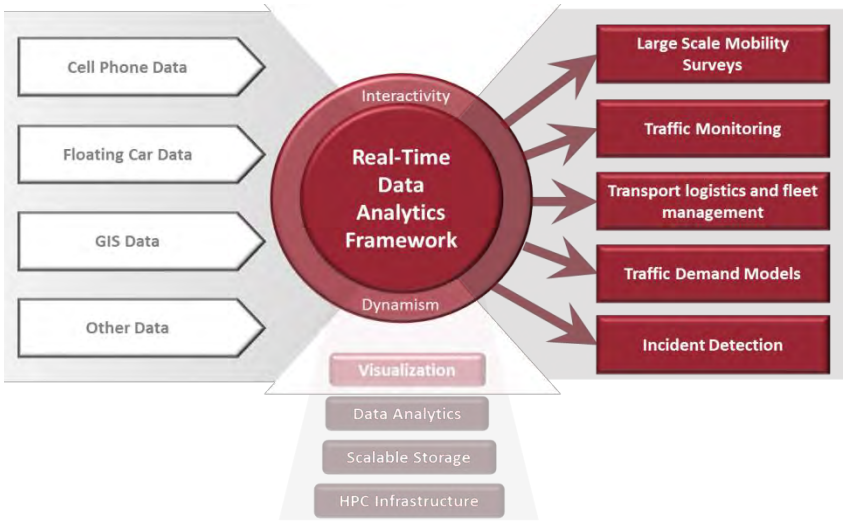
Nachfolgend werden diese Projekte näher beschrieben und der jeweilige Lösungsansatz wird detailliert präsentiert. Des Weiteren werden die Ergebnisse der Projektanalysen in Bezug auf die Big Data Charakteristiken (Volume, Veracity, Velocity, Value) sowie auf die verwendeten Technologien in Bezug auf den Big Data Stack (Utilization, Analytics, Platform, Management) vorgestellt. Für jedes Projekt wurden die jeweiligen projektspezifischen Anforderungen und Potenziale erhoben welche im Detail aufgeführt und analysiert sind.

2.1 Leitprojekt Verkehr: Real-time Data Analytics for the Mobility Domain

Real-time Data Analytics for the Mobility Domain	
Domäne	Transport
Projekttyp	Internes Projekt
Projektpartner	Austrian Institute of Technology (AIT) Mobility Department
Ansprechpartner	DI (FH) Markus Ray, AIT Tel: +43 50550 6658 Email: markus.ray@ait.ac.at
Projektbeschreibung	Dieses Projekt beschäftigt sich mit der Bereitstellung einer flexiblen Infrastruktur für die Echtzeitanalyse großer und dynamischer Datenströme für mobilitätsbezogene Forschung. Die Herausforderungen bestehen in der Verarbeitung und Verschneidung großer Datenmengen mit Mobilitätsbezug und deren Bereitstellung in Echtzeit. In den letzten Jahren ist die verfügbare Menge an für diese Domäne relevanten Daten enorm angestiegen (z.B. Verkehrssensoren wie Loop Detectors oder Floating Car Daten, Wetterdaten) und neue Datenquellen (z.B. Mobilfunkdaten oder Social Media Daten) stehen für detaillierte Analysen zur Verfügung. Dies schafft enorme Potenziale für innovative Lösungsansätze, die beispielsweise dynamische Aspekte (wie Echtzeitverkehrszustand) bei multi-modalen Tourenplanungen berücksichtigen (Bsp. Krankentransport), die semi-automatisierte Durchführung von vertieften Mobilitätserhebungen ermöglichen (z.B. für Verkehrsnachfrageerfassung), neuartige Erfassung multi-modaler Verkehrszustände etablieren oder die Bereitstellung profilbasierter individueller Mobilitätsinformationen ermöglichen. Solche Anwendungen erfordern technologisch innovative Konzepte, die diese Datenmengen in Echtzeit verarbeiten und analysieren können.
Lösungsansatz	Die technologische Lösung basiert auf der flexiblen Integration von High-Performance Computing (HPC) und Big Data Lösungen. Dieses integrierte Analyseframework nimmt folgende

Herausforderungen in Angriff:

- Effiziente verteilte Speicherung und Abfrage großer Datenmengen (inkl. anwendungsbezogener in-memory Lösungsansätze).
- Schnelle Laufzeit durch den Einsatz von HPC Konzepten
- Flexible modul-basierte Definition komplexer Analysearbeitsabläufe
- Unterstützung der Integration existierender Anwendungen
- Echtzeitverarbeitung und Integration heterogener Datenstreams



Eine integrierte modulare Data Pipeline mit dem MapReduce Programmierparadigma unterstützt die flexible Kombination unterschiedlicher applikationsspezifischer Module zu komplexen Arbeitsabläufen. Einerseits werden entsprechende Module für den Import und Export spezifischer Datenquellen angeboten, andererseits sind unterschiedliche Machine Learning Algorithmen und Funktionen implementiert. Diese können flexibel in der Map- sowie in der Reduce-Phase kombiniert werden, oder gesondert auf den HPC Ressourcen ausgeführt werden. Dies ermöglicht die flexible Ausnutzung unterschiedlicher Hardwareressourcen und Speichersysteme für applikationsspezifische Problemstellungen mit dem Ziel einer in Echtzeit durchgeführten Datenanalyse.

Anforderungen

Applikationsspezifische Anforderungen:

- Optimierung der multi-modalen Verkehrssystems: Bspw. Erhebung und Modellierung der Verkehrsnachfrage, Verkehrsüberwachung und Steuerung, Transportlogistik und Flottenmanagement, Empfehlen und Überprüfen von Verkehrsmaßnahmen.

Soziale Anforderungen:

- Richtlinien für den Umgang mit persönlichen Mobilitätsdaten

	Rechtliche Anforderungen: - Privacy und Security: Klare Regelung in Bezug auf personenbezogene Daten (rechtliche Rahmenbedingungen, Datenschutz, Sicherheit)	
Big Data Charakteristiken *(Welche Big Data Charakteristiken werden adressiert und in welchem Umfang?)*	**Volume**	Sehr große Datenmengen in den Bereichen Mobilfunkdaten, Floating Car Data und Social Media
	Variety	Komplexität in Bezug auf unterschiedlichste Datenstrukturen und -formate von Binärdaten, GIS-Daten bis zu Freitext
	Veracity	Komplexität in Bezug auf Echtzeitanbindung und Integration von unterschiedlichen Datenstreams Komplexität in Bezug auf Echtzeitausführung von komplexen Analysealgorithmen auf Basis von großen Datenmengen
	Value	Integration und Echtzeitanalyse von unterschiedlichen Datenquellen birgt enormes Potenzial in innovativen Methoden für unterschiedlichste Mobilitätsanwendungen
Big Data Stack *(Welche Big Data Ebenen und welche Technologien werden adressiert und verwendet?)*	**Utilization**	Die Analyse Ergebnisse können flexibel in Web-basierte Oberflächen integriert werden. Ziel ist die interaktive Analyse integrierter Datenquellen.
	Analytics	Das Analyseframework bietet Implementierungen diverser Machine Learning Algorithmen. Diese können auf Basis der darunterliegenden Plattform sowohl auf HPC Infrastruktur als auch Daten-lokal in Datenzentren ausgeführt werden.
	Platform	Das Projekt bedient sich des Hadoop Ökosystems und bindet mehrere Plattformen für die Ausführung von Daten-intensiven Applikation ein. Die Hadoop Ausführungsumgebung ist nahtlos in eigene Entwicklungen eines Pipelining Frameworks integriert und ermöglicht auf diese Weise, die skalierbare Analyse von (binären) Daten auf Basis applikationsspezifischer Algorithmen. Die enge Verzahnung des MapReduce Paradigmas mit Pipelining Konzepten ermöglicht die Daten-lokale Ausführung komplexer Analysealgorithmen.

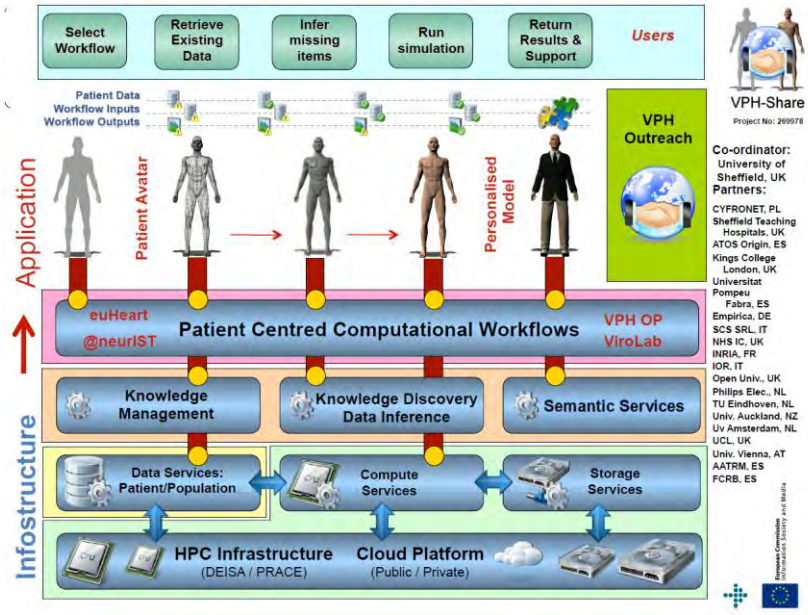
	Management	Innerhalb des Projekts werden unterschiedliche Datenspeicherungssysteme aus dem Hadoop Ökosystem verwendet. Die Storage und Rechenressourcen sind auf HPC Ressourcen und lokale Big Data Cluster aufgeteilt.
Potenziale	Bereitstellung eines Frameworks auf Basis von HPC und Big Data Technologien das die flexible Integration von Echtzeitdaten und zusätzlicher Datenquellen für innovative Forschungsmethoden im Mobilitätsbereich ermöglicht - Die gemeinschaftliche und integrierte Bereitstellung verschiedener Datenquellen fördert die Entwicklung neuer Forschungsfragen und innovativer Geschäftsfälle in den Bereichen HPC, Big Data sowie Mobilität. - Das erstmalige Ermöglichen eines gemeinsamen Zugriffs auf verschiedene verkehrsbezogene Datenquellen ermöglicht die Entdeckung neuer und die Verbesserung bestehender Verkehrsmodelle. - Die innovative Kombination von HPC und Big Data Technologien für komplexe Echtzeitdatenanalysen dient als Vorzeigebeispiel und liefert Anstöße für weitere Projekte.	

2.2 Leitprojekt Healthcare: VPH-Share

Projekt „VPH-Share: The Virtual Physiological Human, Sharing for Healthcare - A Research Environment“	
Domäne	Healthcare
Projekttyp	EU Projekt mit österreichischer Beteiligung (ICT for Health – Resource book of eHealth Projects - FP7)
Projektpartner	University of Sheffield, Coordinator Cyfronet Sheffield Teaching Hospitals Atos Kings College London Unversity of Pompeu Fabra, Spain Empirica SCS Supercomputing Solutions INRIA Instituto Ortopedico Rizzoli di Bologna

	Phillips The Open University Technische Universiteit Eindhoven The University of Auckland University of Amsterdam UCL **Universität Wien** Agència de Qualitat i Avaluació Sanitàries de Catalunya Fundació Clínic Barcelona
Ansprechpartner	Norman Powell BSc PhD VPH-Share Project Manager University of Sheffield Univ.-Prof. Dipl.-Ing. Dr. Siegfried Benkner Universität Wien Email: siegfried.benkner@univie.ac.at
Projektbeschreibung	Das Projekt VPH-Share ermöglicht die Integration von sorgfältig optimierten Services für die gemeinsame Benutzung und Integration von Daten, die Entwicklung von Modellen und die gemeinschaftliche Zusammenarbeit von unterschiedlichen Stakeholdern im Gesundheitsbereich auf Basis einer europäischen Cloud Infrastruktur. Innerhalb von VPH-Share werden meistens klinische Daten von individuellen Patienten (medizinische Bilder, biomedizinische Signale, Bevölkerungsdaten, ...) verarbeitet. Das Projekt umfasst unterschiedlichste Varianten von Operationen in Bezug auf diese Daten. Diese umfassen die sichere Speicherung, den sicheren Zugriff, die Annotierung, Inferenz und Assimilation, komplexe Bildverarbeitungsmechanismen, mathematische Modellierungen, die Reduktion der Daten, und die Repräsentation von aus den Daten generiertem Wissen. Das Projekt fokussiert auf die größte Herausforderung in diesem Bereich: der Schnittstelle zwischen dem Reichtum an Daten von medizinischen Forschungseinrichtungen und klinischen Prozessen. Das Ziel der Schaffung einer umfassenden Infrastruktur für die gemeinschaftliche Bereitstellung von Daten, Informationen und Wissen, deren Verarbeitung an Hand von Workflows und deren Visualisierung wird von einem europäischen Konsortium mit einem breiten Kompetenzmix in Angriff genommen und an Hand von vier Flagship Workflows (existierende FP6 und FP7 Projekte @neurIST, euHeart, VPHOP, Virolab) evaluiert.
Lösungsansatz	Das VPH-Share Projekt basiert auf einer Infrastruktur welche aus HPC und Cloud Ressourcen besteht. Auf Basis Web Service und RESTbasierte APIs für Compute Rund Storage Ressourcen angeboten. Diese Infrastrukturservices bieten die Grundlage für generische Datenservices und verteilten Services Wissensmanagement, Wissensentdeckung, Schlussfolgerungen, und semantischen Services. Diese Ebenen bilden die VPH-Share

Infostructure und werden für die verteilte und transparente Ausführung von unterschiedlichen Patientenzentrierten Workflows.



Anforderungen

Applikationsspezifische Anforderungen:

- Unterstützung der medizinischen Diagnose, der Behandlung, und der Prävention durch die Integration von diversen Datenquellen und der Ermöglichung von computergestützten Simulationen und Analysen

Soziale Anforderungen:

- Einbindung eines Ethikrats
- Richtlinien für klinische Daten
- Einfache und sichere Verfügbarmachung von verteilten Ressourcen
    - Für (Daten) Lieferanten
    - Für Forscher
    - Für Einrichtungen
    - Für Benutzer

Rechtliche Anforderungen:

- Unterschiedliche rechtliche Basis und Sicherheitsvorkehrungen für unterschiedliche Ressourcen
- Umsetzung auf Basis strenger Privacy und Security Richtlinien
- Komplexe verteilte Security Mechanismen

Big Data Charakteristiken *(Welche Big Data Charakteristiken werden adressiert und in welchem Umfang?)*	Volume	Mehr als 680 unterschiedliche Datensets aus verschiedenen Bereichen
	Variety	Komplexität in Bezug auf unterschiedlichste Datentypen und

		Formate (Bilddaten bis relationale Datenbanken)
	Veracity	Komplexität in Bezug auf integrierte Sicht auf heterogene und verteilt gespeicherte Datentypen Semantische Echtzeit-Integration der Daten auf Basis von Ontologien
	Value	Vereinheitlichte und integrierte Sicht (Ontologien) auf verteilte Daten verschiedener Stakeholder sowie die Verwendung von Computer-gestützten Simulationen bringt Mehrwert in der Diagnose, der Behandlung, und der Prävention innerhalb der medizinischen Forschung und für Ärzte Diese Technologien werden für verschiedene innovative Anwendungsfälle nutzbar gemacht: - @neurIST - VPHOP - euHeart - Virolab
Big Data Stack *(Welche Big Data Ebenen und welche Technologien werden adressiert und verwendet?)*	Utilization	Semantische Technologien (Linked Data und Ontologien) für die Wissensaufbereitung und die Integration von unterschiedlichen Datenquellen. Bereitstellung von Visualisierungstools für generiertes Wissen, integrierte Daten, sowie Ergebnisse komplexer datenintensiver Analysen und Simulationen.
	Analytics	Implementierung von medizinischen Simulationen und Analysen
	Platform	Multi-Cloud Plattform auf Basis von OpenSource Toolkits
	Management	Verteilte Cloud Lösung welche eine dezentrale Speicherung der Daten ermöglicht; Bereitstellung von verteilten und skalierbaren Cloud Services für die online Mediation von Daten.
Potenziale	Bereitstellung einer Referenzarchitektur für eine gemeinschaftliche Umgebung welche die Barrieren für die gemeinsame Benutzung von Ressourcen (Date, Rechen, Workflows, Applikationen) beseitigt. - Die gemeinschaftliche und integrierte Bereitstellung von	

verschiedenen Datenquellen kann die Entwicklung von neuen Forschungsfragen und von innovativen Geschäftsfällen fördern.

- Das erstmalige Ermöglichen eines gemeinsamen Zugriffs auf verschiedener klinischer Datenquellen kann die Entdeckung neuer beziehungsweise die Verbesserung bestehender Diagnose-, Behandlungs-, und Präventionsmethoden ermöglichen.
- Die innovative Kombination von verteilten Cloud Ressourcen zu einer gemeinschaftlichen europäischen Storage und Compute Cloud kann als Vorzeigebeispiel dienen und Anstöße für weitere Projekte liefern.

2.3 Leitprojekt Handel

Es wurde ein Handelsunternehmen untersucht, welches stark auf die Anwendung von Datenbezogenen Diensten setzt. Aus Policy-Gründen kann dieses Unternehmen jedoch nicht genannt werden.

Verbesserung der Wettbewerbsfähigkeit durch die Implementierung von Datenbezogenen Anwendungen	
Domäne	Handel
Projekttyp	Internes Projekt
Projektpartner	
Ansprechpartner	
Projektbeschreibung	Im Handel bestehen vielfältige Ansätze, komplexe Lösungen mit Daten zu verbinden. Hierbei erhofft man sich eine verbesserte Wettbewerbsfähigkeit und besseres Verständnis der Kunden. Im untersuchten Unternehmen wurde folgendes implementiert: • Echtzeitanalyse der Verkäufe einer Filiale. Dem Unternehmen stehen vielfältige Möglichkeiten zur Verfügung jede Filiale in Echtzeit zu beobachten. Hierbei werden verschiedene Dinge überprüft. Wesentlich ist die Beobachtung von Aktionen und die Analyse, wie sich diese pro Filiale unterscheiden. Ist am Wochenende das Produkt „XY“ verbilligt, so kann festgestellt werden wo es sich besser

		und wo schlechter verkauft und auf „Verkaufsspitzen“ besser reagiert werden. • Auffinden von Anomalien in einer Filiale. Das Unternehmen hat Erfahrungswerte und Muster über Verkäufe einer gewissen Produktkategorie in der Filiale. Ein Muster ist hier zum Beispiel „XY wird alle 15 Minuten verkauft“. Ist dem nicht so, so wird eine Benachrichtigung an FilialmitarbeiterInnen gesendet, mit der Aufforderung das jeweilige Produkt zu überprüfen. Es hat sich herausgestellt, das mit dem Produkt jeweils etwas nicht in Ordnung war. • Anpassen der Filialen an demographische Entwicklungen. In bestimmten Fällen ändert sich die Demographie des Einzugsgebietes stark. Dies kann durch die Stadtentwicklung oder aber auch durch Einfluss von Bevölkerungsgruppen der Fall sein. Dem Unternehmen stehen Verfahren zur Verfügung, welche es erlauben, diese zu erkennen und darauf zu reagieren. Eine Reaktion kann z.B. der Umbau der Filialen oder die Auswechselung des Produktsortiments sein.
Lösungsansatz		Die Anwendungen wurden mit vorhandenen Lösungen für den Handel realisiert. Im Hintergrund stand hierfür eine NoSQL-Datenbank welche vor allem auf dem Echtzeiteinsatz optimiert wurde.
Anforderungen		Applikationsspezifische Anforderungen: • Near-Realtime Analyse der Filialen bis max. 1 Minute Verzögerung • Speichern großer Datenmengen der Produkte • Mustererkennung der einzelnen Filialen und Produkten • Analyse des Verhaltens der Kunden und Erkennung von Abweichungen • Kombination von verschiedenen Datenquellen über demographische Daten Soziale Anforderungen: • Abspeicherung von Kundenverhalten Rechtliche Anforderungen: • Anonyme Verarbeitung von Kundendaten
Big Data Charakteristiken *(Welche Big Data Charakteristiken werden adressiert und in welchem Umfang?)*	**Volume**	Es werden viele Daten über Produkte, Muster und dergleichen abgespeichert. Dadurch entsteht ein großes Datenvolumen
	Variety	Daten kommen aus unterschiedlichen Quellen und müssen in ein Endformat

		für das Unternehmen übertragen werden
	Veracity	Daten über jeweilige Filialen müssen in Echtzeit überwacht werden können. Dadurch ergeben sich Anforderungen an die Geschwindigkeit der Daten
	Value	Die Daten liefern einen wesentlichen Mehrwert für das Unternehmen. Hierbei geht es primär darum, das Produktsegment zu optimieren. Dies steigert Umsatz und Gewinn des Unternehmens.
Big Data Stack *(Welche Big Data Ebenen und welche Technologien werden adressiert und verwendet?)*	**Utilization**	Die Analyse Ergebnisse können flexibel in Web-basierte Oberflächen integriert werden. Ziel ist die interaktive Analyse von integrierten Datenquellen.
	Analytics	Die Algorithmen basieren auf statistischen Modellen.
	Platform	Für die Analyse kommen für den Einzelhandel optimierte Systeme zum Einsatz.
	Management	Die Speichersysteme werden durch einen Partner/Outsourcingprovider zur Verfügung gestellt.
Potenziale	Es besteht noch viel Potenzial in Richtung Echtzeitanalyse. Nach aussagen des IT Verantwortlichen des Unternehmens steht dieses erst am Anfang der Möglichkeiten. Hierfür werden in den nächsten Monaten und Jahren viele Projekte stattfinden.	

2.4 Leitprojekt Industrie: Katastrophenerkennung mit TRIDEC

Real-Time fähige Anwendung zur Katastrophenerkennung	
Domäne	Industrie
Projekttyp	EU-Projekt
Projektpartner	Joanneum Research
Ansprechpartner	DI Herwig Zeiner Telefon: +43 316 876-1153 Fax: +43 316 876-1191 herwig.zeiner@joanneum.at

Projektbeschreibung	In TRIDEC (09/2010-10/2013) wurden neue real-time fähige Architekturen und Werkzeuge entwickelt, um Krisensituationen zu überwinden und Schäden oder negative Einflüsse nach Möglichkeit durch angepasste Entscheidungen abzuwehren. Zentrale Herausforderung war die Konzeption und Entwicklung einer offenen Plattform für interoperable Dienste, die ein intelligentes, ereignisgesteuertes Management von sehr großen Datenmengen und vielfältigen Informationsflüssen in Krisensituationen ermöglicht. Im Mittelpunkt stand die Arbeit an neuen Modellen und den dazugehörenden real-time fähigen Algorithmen für das Monitoring System. Das Paket beinhaltet auch einen entsprechenden Systemaufbau mit einer sicheren Kommunikations- und Analyseinfrastruktur.	
Lösungsansatz	Der technologische Ansatz wurde in zwei Anwendungsfeldern demonstriert, die sich beide durch das Auftreten extrem großer Datenmengen auszeichnen. Das erste Anwendungsszenario setzte den Fokus auf Krisensituationen, wie sie bei der Erschließung des Untergrundes durch Bohrungen auftreten können, einer für Geologen außerordentlich wichtigen, jedoch überaus teuren Aufschlussmethode. Bohrungen werden unter Verwendung von Sensornetzwerken permanent überwacht und Störungen im Bohrbetrieb frühzeitig ermittelt. In TRIDEC wurden vor allem an der Entwicklung neuer Analyseverfahren zur Erkennung und Vermeidung kritischer Situationen bei Bohrungen gearbeitet. Dazu zählt z.B. die Erkennung von Stuck Pipe Situationen genauso wir Vorschläge zur Durchführung krisenvermeidender Operationen (z.B. Ream & Wash).	
Anforderungen	Applikationsspezifische Anforderungen: • Real-Time Analyse von Katastrophenfällen • Real-Time Reaktion auf Katastrophenfälle	
Big Data Charakteristiken *(Welche Big Data Charakteristiken werden adressiert und in welchem Umfang?)*	**Volume**	-
	Variety	-
	Veracity	Industrieanlagen müssen in Echtzeit analysiert werden können. Dies geschieht in Abstimmung mit geografischen Daten.
	Value	Tritt ein Katastrophenfall wie z.B. ein Erdbeben oder ein Zunami ein, so kann frühzeitig darauf reagiert werden und der Schaden an Industrieanlagen minimiert werden.

Big Data Stack *(Welche Big Data Ebenen und welche Technologien werden adressiert und verwendet?)*	**Utilization**	Die Auswertungen werden durch eine graphische Oberfläche repräsentiert.
	Analytics	Die Algorithmen basieren auf statistischen Modellen.
	Platform	-
	Management	-
Potenziale		Hierbei handelt es sich primär um ein Forschungsprojekt. Österreichische Unternehmen, vor allem aus der Öl- und Gasindustrie sowie in produzierenden Unternehmen können dies Anwenden. Ferner besteht die Möglichkeit, ein Produkt daraus zu entwickeln.

2.5 Leitprojekt Weltraum: Prepare4EODC

Projekt Prepare4EODC	
Domäne	ASAP - Austrian Space Applications Programme
Projektpartner	Vienna University of Technology (TU Wien), Department of Geodesy and Geoinformation (GEO) EODC Earth Observation Data Centre for Water Resources Monitoring GmbH (EODC) GeoVille GmbH (GeoVille) Central Institute for Meteorology and Geodynamics (ZAMG) Catalysts GmbH (Catalysts) Angewandte Wissenschaft, Software und Technologie GmbH (AWST)
Ansprechpartner	Mag. Stefan Hasenauer Vienna University of Technology (TU Wien) Tel: +43 1 58801 12241 Email: stefan.hasenauer@geo.tuwien.ac.at DI Dr. Christian Briese EODC Earth Observation Data Centre for Water Resources Monitoring GmbH Tel: +43 1 58801 12211 Email: christian.briese@geo.tuwien.ac.at
Projektbeschreibung	Im Frühjahr 2014 wurde die EODC Earth Observation Data Centre for Water Resources Monitoring GmbH (EODC GmbH) im Rahmen eines „public-private Partnership“ gegründet. Das EODC soll die Nutzung von Erdbeobachtungsdaten für die Beobachtung globaler Wasserressourcen durch eine enge Zusammenarbeit mit Partnern

	aus der Wissenschaft, der öffentlichen Hand als auch der Privatwirtschaft fördern. Das Ziel dieser Kooperation ist der Aufbau einer kooperativen IT-Infrastruktur, die große Erdbeobachtungsdatenmengen (der neue Sentinel 1 Satellit der ESA liefert freie Daten von ca. 1,8 TB pro Tag) vollständig und automatisch prozessieren kann. Konkret wird das Projekt die folgenden Dienste vorbereiten: 1) Datenzugriff, Weiterverteilung und Archivierung von Sentinel-Daten, 2) eine Sentinel-1 Prozessierungskette, 3) ein Informationssystem für die Überwachung landwirtschaftlicher Produktion basierend auf Sentinel und 4) eine Cloud-Plattform für die wissenschaftliche Analyse von Satellitenbodenfeuchtigkeitsdaten.
Lösungsansatz	Entwicklung von Managementmethoden, grundlegender Software-Infrastruktur sowie ersten Datenverarbeitungsketten.
Anforderungen	Applikationsspezifische Anforderungen: near-realtime processing, advanced processing of large data volumes, long term preservation and reprocessing Soziale Anforderungen: public-private Partnership Rechtliche Anforderungen: Inspire Richtlinien
Big Data Charakteristiken ***(Welche Big Data Charakteristiken werden adressiert und in welchem Umfang?)***	1.8 TB pro Tag — Anbindung an den Vienna Scientific Cluster (VSC) Disk space: 3-5 PB
Potenziale	Das Ergebnis des Projektes wird positive Auswirkungen auf den Technologievorsprung Österreichs im Bereich der Erdbeobachtung haben. Darüber hinaus stellt die Erfassung globaler Wasserressourcen ein aktuelles Interesse der Gesellschaft dar. Neben nationalen Kooperationen sollen auch internationale Partner in die Aktivitäten des EODC eingebunden werden. Dadurch wird die internationale Rolle Österreichs im Bereich der Nutzung von Satellitendaten gestärkt.

3 Leitfaden für „Big Data“ Projekte

Für die effiziente und problemlose Umsetzung eines Big Data Projekts sind mannigfaltige Aspekte aus vielen unterschiedlichen Bereichen zu beachten welche über normales IT Projektmanagement und diesbezügliche Vorgehensmodelle hinausgehen. Durch die Neuartigkeit der Problemstellungen und der oftmaligen mangelnden unternehmensinternen Erfahrung entstehen zusätzliche Risiken, welche durch ein effizientes Big Data-spezifisches Projektmanagement verringert werden können.

In diesem Leitfaden für die Abwicklung eines Big Data Projekts wird auf die zu beachtenden Spezifika eingegangen und es werden entsprechende Rahmenbedingungen definiert. Der Leitfaden soll Organisationen eine Hilfestellung und Orientierung für die Umsetzung von Big Data Projekten bieten. Der Leitfaden gliedert sich in die Definition eines Big Data Reifegradmodells auf Basis dessen der aktuelle Status bezüglich Big Data in einer Organisation eingeschätzt werden kann und weitere Schritte gesetzt werden können. Des Weiteren wird ein Vorgehensmodell für die Umsetzung von Big Data Projekten vorgestellt welches sich in acht Phasen gliedert. Ein wichtiger Aspekt für die erfolgreiche Umsetzung eines Big Data Projekts ist die Entwicklung von entsprechender Kompetenz innerhalb der Organisation. Hierfür werden benötigte Kompetenzen und Berufsbilder definiert und deren Aufgabenbereiche erläutert. Bei der Speicherung und Verarbeitung von großen Datenmengen gilt es datenschutzrechtliche Aspekte und die Sicherheit der Daten zu beachten und zu gewährleisten. Hierfür sind grundlegende Informationen in Bezug auf Big Data dargestellt.

3.1 Big Data Reifegrad Modell

Der Einsatz von Big Data kann durch schnellere Entscheidungsprozesse als Basis für einen kompetitiven Vorsprung dienen. Dieser steigende Fokus auf Big Data Lösungen bietet einerseits große Gelegenheiten, andererseits ergeben sich dadurch auch große Herausforderungen. Das Ziel ist es nicht nur Informationen abgreifen zu können sondern diese zu analysieren und auf deren Basis zeitnah Entscheidungen treffen zu können.

In vielen Organisationen fehlt es derzeit an der Kompetenz und an dem nötigen Reifegrad um die gesamte Bandbreite von Technologien, Personalbesetzung bis zum Prozessmanagement in Bezug auf Big Data abzudecken. Diese Schritte sind aber für die effiziente und erfolgreiche Ausnutzung der vorhandenen Big Data Assets mit dem Ziel der durchgängigen Umsetzung von Big Data Analysen für die Optimierung von operationalen, taktischen und strategischen Entscheidungen notwendig. Die Fülle an technologischen und analytischen Möglichkeiten, der benötigten technischen und Management Fähigkeiten sowie der derzeit herrschende Hype in Bezug auf Big Data erschweren die Priorisierung von Big Data Projekten innerhalb von Organisationen.

In diesem Big Data Reifegrad Modell werden mehrere voneinander nicht unabhängige Anforderungen berücksichtigt welche die Einordnung einer Organisation in einen bestimmten Reifegrad ermöglichen:

- **Kompetenzentwicklung:** Ein ganzheitlicher Ansatz bezüglich Big Data innerhalb einer Organisation benötigt eine umfassende Strategie zur Entwicklung spezifischer

Kompetenzen im Umfeld von Big Data. Diese beinhalten technologische Kompetenzen in Bezug auf den gesamten Big Data Stack, Projektmanagement und Controlling Kompetenzen in Bezug auf Big Data, Kompetenzen in Business Development sowie Know-how bezüglich der Rechtslage und des Datenschutzes. Eine detailliertere Analyse der Kompetenzentwicklung wird in Kapitel 3.3 ausgearbeitet.

- **Infrastruktur:** Für die Verarbeitung und Analyse der Daten wird eine Anpassung der IT-Infrastruktur in Bezug auf die vorhandenen Datenquellen, der Anforderungen an die Analyse sowie deren Einbindung in die aktuell verfügbare Systemlandschaft benötigt. Des Weiteren ist für eine effiziente Umsetzung einer Big Data Strategie die Einbindung der unterschiedlichen Stakeholder innerhalb des Unternehmens erforderlich. Hierbei sollen alle Abteilungen welche entweder mit der Datenbeschaffung, Datenaufbereitung und Speicherung sowie alle Abteilungen die Daten für tiefgehende Analysen benötigen eingebunden werden um eine möglichst einheitliche an allen Anforderungen entsprechende Systemlandschaft zu ermöglichen.
- **Daten:** Die Menge an intern sowie extern verfügbaren Daten wächst kontinuierlich und diese sind in den unterschiedlichsten Formaten verfügbar. Neben der Anpassung der Infrastruktur an die gestiegenen Bedürfnisse auf Grund dieser Datenmenge müssen auch weitere Daten-spezifische Aspekte von Datenmanagement, Governance, Sicherheit bis zu Datenschutz berücksichtigt werden.
- **Prozessumsetzung:** Die Umsetzung von Big Data Projekten und der dadurch generierte Erkenntnisgewinn innerhalb einer Organisation wird die Aufdeckung von Ineffizienzen und die Identifizierung von neuen Interaktionsmöglichkeiten mit Kunden, Angestellten, Lieferanten, Partnern und regulatorischer Organisationen ermöglichen. Eine Gefahr bezüglich der vollständigen Ausschöpfung der Potenziale besteht durch die Möglichkeit einer Organisation die Geschäftsprozesse effizient umzugestalten und an die neuen Erkenntnisse anzupassen. Optimierte Geschäftsprozesse bieten eine höhere Chance in Richtung der schnellen Einführung von innovativen Big Data Anwendungsfällen und Betriebsmodellen.
- **Potenziale:** Aktuell stützen nur wenige Unternehmen ihre Geschäftsprozesse auf eine ganzheitlich definierte Big Data Strategie. Viele Big Data Projekte werden von der IT angestoßen und sind, unabhängig von der technologischen Umsetzung, oft nicht in die bestehende Geschäftsprozesse integriert. Eine enge Vernetzung der Geschäftsziele und Prozesse mit der Umsetzung in der IT kann einen innovativen geschäftsgetriebenen Investmentzyklus in einer Organisation schaffen, welcher die Ausschöpfung der Potenziale ermöglicht. Eine Gefahr besteht in zu hohen Erwartungen in erste Big Data getriebene Projekte durch den aktuell vorhandenen Hype bezüglich dieser Thematik. Diese Risiken können durch die Entwicklung und Umsetzung einer allgegenwärtigen Organisationsweiten Big Data Strategie minimiert werden.

Das hier vorgestellte Reifegradmodell kann als Unterstützung bei der Beurteilung der aktuell vorhandenen Fähigkeiten in Bezug auf die erfolgreiche Umsetzung von Big Data Projekten verwendet werden. Dieses Reifegrad Modell definiert sechs Ebenen welche unterschiedliche Reifegrade in der Umsetzung von Big Data Projekten definieren. Ausgehend von einem Szenario in welchem keinerlei Big Data spezifische Vorkenntnisse in einer Organisation vorhanden sind, erhöht sich der definierte Reifegrad bis zur erfolgreichen Umsetzung eines nachhaltigen Big Data Business innerhalb einer Organisation.

Folgend werden die einzelnen Ebenen im Detail vorgestellt sowie werden Ziele, Wirkungen und Maßnahmen beschrieben die von Organisationen in dieser Ebene gesetzt werden können.

Abbildung 1: Big Data Maturity Modell

3.1.1 Keine Big Data Projekte

In diesem Status hat die Organisation noch keine Aktivitäten im Big Data Bereich gestartet. Es wurden noch keine Erfahrungen gesammelt und es werden klassische Technologien für die Datenverwaltung und Analyse eingesetzt.

3.1.2 Big Data Kompetenzerwerb

In dieser Phase startet die Organisation erste Initiativen zum Erwerb von Big Data spezifischen Kompetenzen. Die Organisation definiert Early Adopters für den Big Data Bereich. Diese Early Adopters sind für die Sammlung und den Aufbau von Wissen innerhalb der Organisation zuständig. Des Weiteren werden mögliche Einsatzszenarien von Big Data Technologien in Bezug auf die IT Infrastruktur, einzelne Bereiche oder auch bereichsübergreifend entwickelt.

Diese Phase ist oft durch ein sehr hektisches und unkoordiniertes Vorgehen gekennzeichnet. Oft scheitern erste Versuchsprojekte an den verschiedensten Faktoren. Diese sind unzureichende technische Erfahrungen der MitarbeiterInnen, unzureichende Erfahrungen über die Potenziale von Big Data im Unternehmen, fehlendes Bewusstsein über die vorhandenen Daten und wie diese genutzt werden können sowie fehlende Management-Unterstützung. In vielen Fällen werden Projekte nicht zentral gesteuert sondern vielmehr ad-hoc erstellt, da ein gewisser Bedarf besteht. Das zentrale IT-Management ist nicht immer in Kenntnis der Projekte. In dieser Phase besteht eine erhöhte Gefahr, dass Projekte scheitern.

3.1.3 Evaluierung von Big Data Technologien

In dieser Phase wurden erste Big Data spezifische Kompetenzen erworben und mögliche Einsatzbereiche dieser Technologien wurden definiert. Die Organisation beginnt den strategischen Aufbau von Big Data Beauftragten. Die Organisation hat mit der Umsetzung von Big Data Projekten zur Optimierung von Geschäftsbereichen oder der IT Infrastruktur begonnen.

Die IT-Verantwortlichen innerhalb des Unternehmens beziehungsweise der Organisation beginnen, das Thema Big Data wesentlich strategischer zu sehen. Es werden erste Strategien und Projektmanagementtechniken evaluiert und umgesetzt. Vielfach handelt es sich jedoch noch um isolierte Versuchsballone, wo zum einem der technologische Kompetenzerwerb erleichtert werden sollte und zum anderen die jeweilige Implementierungsverfahren überprüft werden.

In Österreich findet man sich Großteils in dieser beziehungsweise der ersten Phase wieder. Die Verfasser der Studie konnten mit wichtigen IT-Entscheidungsträgern der heimischen Wirtschaft diesbezüglich sprechen.

3.1.4 Prozess für Big Data Projekte

Die initialen Big Data Projekte befinden sich vor dem Abschluss und erste Evaluierungsergebnisse sind vorhanden. Die Big Data Verantwortlichkeiten im Unternehmen sind klar definiert und strategische Maßnahmen für die Umsetzung von weiteren Big Data Projekten und deren Integration in Geschäftsprozesse sind gesetzt. Big Data hat sich in der Geschäftsstrategie der Organisation etabliert.

Hier findet sich auch der erste Ansatz von Top-Management-Unterstützung wieder. Es wurde erfolgreich bewiesen, dass Big Data Projekte einen gewissen Mehrwert im Unternehmen bieten und seitens der Geschäftsführung wird dies anerkannt. Die Big Data Verantwortlichen, welche international auch als „Data Scientists“ bekannt sind, sind nun im Unternehmen vorhanden und arbeiten Strategien und Prozesse für weitere Optimierungen aus. Ein Fokus liegt nun darin, herauszufinden welche weiteren Möglichkeiten es gibt und wie diese erfolgreich in die Tat umgesetzt werden können. Geschäftsprozesse dienen hierbei als wichtiges Mittel, diese Vorhaben zu leiten und zu koordinieren.

3.1.5 Gesteuerte Big Data Projekte

Big Data ist fixer Bestandteil der Organisationstrategie und es existiert ein klares Prozessmanagement für Big Data relevante Projekte. Erste Big Data Projekte wurden in Geschäftsprozesse integriert beziehungsweise wurden anhand dieser neue Geschäftsprozesse geschaffen.

Die in der Vorstufe begonnene Standardisierung wird konsequent umgesetzt und erweitert. Die Strategie wird nun ausgebaut und der Schritt zum nachhaltigen Big Data Business wird unternommen.

3.1.6 Nachhaltiges Big Data Business

Big Data ist zentraler Bestandteil der Strategie und wird zur Optimierung von unterschiedlichen Geschäftsprozessen eingesetzt. Die IT Infrastruktur setzt Big Data Technologien ein und bietet diese der Organisation für die Umsetzung neuer Projekte an. Big Data spezifische Aspekte sind in das Projektmanagement und die (IT) Strategie des Unternehmens integriert.

Die jeweiligen Big Data Verantwortlichen interagieren mit den unterschiedlichsten Abteilungen des Unternehmens und liefern ständig neue Innovationen. Diese Ebene ist jedoch nicht nur von der Big Data Strategie selbst abhängig sondern benötigt auch eine gewisse Stellung der IT im Unternehmen an sich. Hierfür ist es oftmals notwendig, dass jene Person, welche für die IT im Unternehmen verantwortlich ist, auch Teil der Geschäftsführung ist. Die letzte Stufe kann folglich nur erreicht werden, wenn auch andere Aspekte in diesem Umfang abgedeckt sind.

3.2 Vorgehensmodell für Big Data Projekte

Die Definition, Umsetzung und der Einsatz von standardisierten Vorgehensmodellen in IT Projekten ist sehr weit verbreitet. Neben dem Hinweis auf die Wichtigkeit der Umsetzung und Durchsetzung eines geeigneten Vorgehensmodells im Kontext von risikoreichen datengetriebenen Projekten werden diese hier nicht näher beleuchtet. In diesem Abschnitt werden vielmehr essenzielle Aspekte die in direktem Bezug zu Big Data stehen näher betrachtet. Auf deren Basis wird ein Big Data spezifisches Vorgehensmodell vorgestellt, welches während der Organisationsinternen, oder auch Organisationsübergreifenden, Umsetzung von Big Data Projekten unterstützend angewendet werden kann. Die einzelnen Schritte können unabhängig von dem aktuellen Reifegrad in einer Organisation gesehen werden. Das präsentierte Vorgehensmodell muss an die Anforderungen und Größe der Organisation angepasst werden und immer in Bezug auf die vorhandenen und einzubindenden Daten sowie der Unternehmensstrategie gesehen werden.

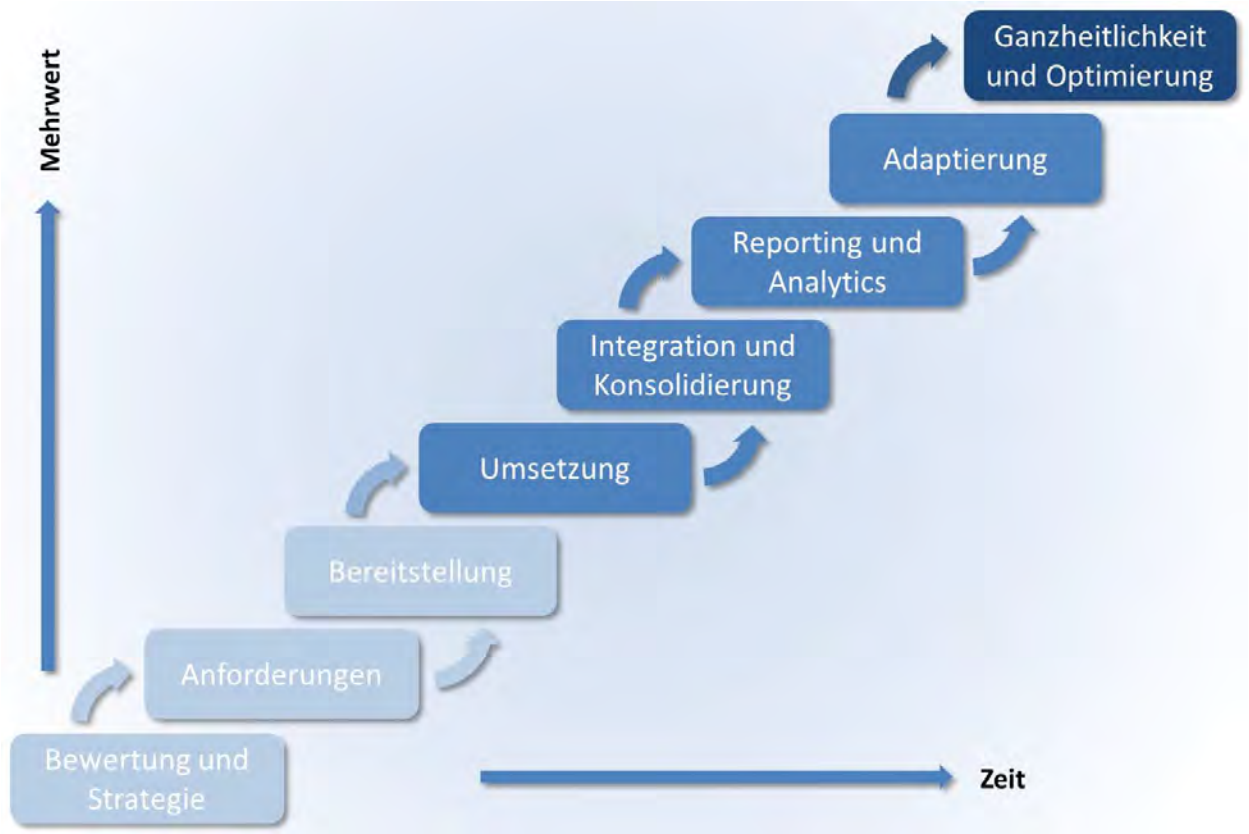


Abbildung 2: Vorgehensmodell für Big Data Projekte

Das Vorgehensmodell besteht aus acht Phasen welche zeitlich nacheinander dargestellt sind, deren Ergebnisse ineinander einfließen, und deren Anwendung von dem gewällten Projektmanagementmodell abhängig ist und dementsprechend angepasst werden müssen. Das Vorgehensmodell (siehe **Abbildung 2**) gliedert sich in die nachfolgend aufgelisteten Schritte:

- Bewertung und Strategie
- Anforderungen
- Vorbereitung
- Umsetzung
- Integration und Konsolidierung
- Reporting und Analytics
- Adaptierung
- Ganzheitlichkeit und Optimierung

Nachfolgend werden die einzelnen Phasen näher beleuchtet und ihre Spezifika dargestellt.

3.2.1 Bewertung und Strategie

Die Phase Bewertung und Strategie ist für die effiziente und erfolgreiche Umsetzung eines Big Data Projektes innerhalb einer Organisation sehr wichtig. Das Ziel dieser Phase ist die Ausarbeitung einer Strategie bezüglich der Umsetzung von Big Data spezifischen Thematiken innerhalb der Organisation. Diese Strategie umfasst Potenziale welche sich aus der Verwendung von Big Data ergeben, klar definierte Ziele in Bezug auf Geschäftsprozesse und Mehrwertgenerierung sowie Vorgehensmodelle, Herausforderungen, benötigte Kompetenzen und Technologiemanagement.

In einem ersten Schritt sollen innerhalb der Organisation die wichtigsten Ziele in Bezug auf Big Data sowie die daraus resultierenden möglichen Potenziale erhoben werden. Ziel hierbei ist es, innerhalb der Organisation Bewusstsein für die Möglichkeiten von Big Data zu schaffen und erste konkrete und machbare Ziele für die potentielle Umsetzung zu generieren.

Auf der Basis dieser Informationen wird in einem nächsten Schritt der aktuelle Status in Bezug auf Big Data erhoben. Dieser Status wird hier wie folgt definiert:

- Identifikation potentieller Geschäftsprozesse
- Identifikation von potentiellem Mehrwert
- Vorhandene interne Datenquellen
- Potentielle externe Datenquellen
- Status der Datenmanagement Infrastruktur
- Status der Datenanalyseinfrastruktur
- Status der Datenaufbereitung und Verwendung
- Vorhandene und benötigte Kompetenzen innerhalb der Organisation
- Analyse der Datenquellen und der Infrastruktur in Bezug auf Privacy und Sicherheit
- Bisherige Projekte im Big Data Bereich
- Big Data Strategie und Big Data in der Organisationsstrategie
- Evaluierung des aktuellen Reifegrades der Organisation in Bezug auf das Big Data Reifegrad Modell

Als Resultat dieser initialen Bewertung soll eine Einschätzung in Bezug auf das Big Data Reifegrad Modell entstehen. Die weiteren Schritte innerhalb der Organisation sollen in Einklang mit den definierten Ebenen des Big Data Reifegrad Modells getroffen werden. Ein weiteres Ergebnis dieser Phase ist eine Bestandsanalyse in Bezug auf Big Data welche bei weiterführenden Projekten nur adaptiert werden muss und eine Grundlage für die effiziente und erfolgreiche Abwicklung eines Big Data Projekts liefert.

Die Ergebnisse der Phase „Bewertung und Strategie“ werden wie folgt zusammengefasst:

- Strategie zu Big Data Themen
- Potenziale und Ziele in Bezug auf Big Data
- Adaptierung der organisationsinternen Strategie in Bezug auf Big Data relevante Themen
- Bewertung der organisationsinternen Umsetzung von Big Data
- Anwendung des Big Data Reifegrad Modells

Für die Durchführung der Phase „Bewertung und Strategie“ können unterschiedliche Methoden angewendet werden. Einige werden hier aufgelistet:

- Interne Workshops: Für die grundlegende Ausarbeitung einer Big Data Strategie und die Erhebung der Potenziale und Ziele wird die Abhaltung eines Big Data Strategie Workshops empfohlen. Ein weiteres Ziel dieses Workshops ist die Spezifikation von organisationsinternen Rollen. Es sollten die internen Big Data Verantwortlichkeiten abgeklärt werden und mindestens eine Person mit der internen Umsetzung und Verwaltung der Big Data Strategie verantwortet werden.
- Interne Umfragen: Anhand von organisationsinternen Umfragen können potentielle Ziele, der aktuelle Stand in Bezug auf Big Data, sowie strategische Optionen erhoben werden.

- Organisationsinterne Interviews: Durch spezifische Interviews zur Big Data Strategie, Potenziale und dem aktuellen Stand der Umsetzung können wichtige Informationen gesammelt werden. Durch persönlich vom Big Data Beauftragten durchgeführte Interviews kann mit Hilfe der Interviews eine Verankerung des Themas über mehrere Abteilungen hinweg erreicht werden und die Akzeptanz des Themas kann erhöht werden.
- Evaluierung der Marktsituation: Es wird erhoben, welche Big Data Techniken im Markt bereits erfolgreich umgesetzt wurden und was die eigene Organisation daraus lernen kann. Ziel ist es, jene Punkte die erfolgreich umgesetzt wurden zu analysieren und daraus Vorteile abzuleiten. Dies kann auf verschiedenste Art und Weise erfolgen:
    - Auffinden und Analysieren von Use Cases: mit klassischer Desktop-Recherche kann man eine Vielzahl an Big Data Use Cases für verschiedenste Domänen gewinnen. Oftmals sind diese jedoch in anderen geografischen Regionen entstanden, was Einfluss auf die tatsächliche Relevanz des Themas hinsichtlich Datensicherheit und –recht hat.
    - Aufsuchen von Konferenzen und Veranstaltungen: hierbei kann man anhand von erfolgreichen Use Cases lernen. Ferner besteht die Möglichkeit, sich mit Personen anderer Unternehmen auszutauschen.
    - Hinzuziehen von externen BeraternInnen: diese können entweder Branchenfremd oder Branchenaffin sein. Branchenaffine BeraterInnen haben bereits ein dediziertes Wissen über jeweilige Case-Studies, Möglichkeiten und relevanten Technologien. Branchenfremde BeraterInnen haben ein wesentlich breiteres Blickfeld, was unter Umständen neue Ideen bringt und nicht nur das gemacht wird, was andere Marktbegleiter ohnehin schon machen.

Generell wird empfohlen, die Bewertung und Strategie in drei Ebenen durchzuführen:

- Unternehmensinterne Bewertung und Strategie: hier wird vor allem evaluiert, was im Unternehmen vorhanden ist, welche Potenziale sich ergeben und welche Hemmfaktoren die Strategie gefährden könnten.
- Brancheninterne Bewertung und Strategie: In dieser Ebene wird evaluiert, was Marktbegleiter bereits implementiert haben, wo Potenziale gegenüber dem Mitbewerb bestehen und welche Nachteile das Unternehmen gegen eben diesen hat.
- Branchenunabhängige Bewertung und Strategie: In der letzten Ebene wird evaluiert, welche Themen außerhalb der jeweiligen Domäne von Relevanz sind. Dies soll einen wesentlich weiteren Blick auf Big Data Technologien schaffen und die Möglichkeit bringen, Innovation zu fördern.

3.2.2 Anforderungen

Die Phase „Anforderungen“ begründet sich auf der erfolgreichen Definition der Big Data Strategie und der umfassenden Bewertung der vorhandenen Big Data Technologien innerhalb einer Organisation. Die Phase Anforderungen ergibt sich aus etablierten IT Vorgehensmodellen mit dem Ziel die projektspezifischen Anforderungen im Detail zu erheben, diese zu analysieren, zu prüfen und mit dem Auftraggeber (intern sowie extern) abzustimmen.

In Bezug auf Big Data Projekte sollte eine erhöhte Aufmerksamkeit auf folgende Punkte gelegt werden:

- Hardware Anforderungen
    - In Bezug auf Big Data Charakteristiken und Big Data Stack
- Software Anforderungen
    - In Bezug auf Big Data Charakteristiken und Big Data Stack
- Funktionale Anforderungen
- Qualitätsanforderungen
- Datenschutz, Privacy, und Sicherheit

Bei der Umsetzung von Big Data Projekten können sich technische Schwierigkeiten in Bezug auf die Big Data Charakteristiken (Volume, Veracity, Velocity) ergeben. Diese stellen sowohl Software als auch Hardware vor neue Herausforderungen und sollten aus diesem Grund gesondert und detailliert behandelt werden. Hierbei müssen aktuell eingesetzte Technologien in Bezug auf deren Einsetzbarkeit mit Big Data evaluiert werden sowie falls notwendig die Anwendbarkeit von neuen Big Data Technologien im Organisationsumfeld bedacht werden.

Des Weiteren ergeben sich durch die Größe, Vielfalt und Geschwindigkeit der Daten additionale Anforderungen und Herausforderungen in Bezug auf die Funktionalität und Qualität. Diese Anforderungen und potentiell daraus resultierende Problematiken sollten ebenfalls vorab gesondert analysiert und evaluiert werden.

Ein wesentlicher Aspekt in Bezug auf die Anwendung von Big Data Technologien in Organisationen ist die Anwendung, Umsetzung und Einhaltung von Datenschutzrichtlinien sowie geeigneten Sicherheitsimplementierungen. Hierzu sollten von Projektbeginn an Datenschutz und Sicherheitsbeauftragte in die Projektabwicklung eingebunden werden. Gerade dem Umgang mit personenbezogenen Daten in Bezug auf Big Data Projekte muss gesonderte und detaillierte Beachtung geschenkt werden. Ebenso wichtig wie die Einhaltung von Datenschutzrichtlinien ist es, sich um gesellschaftliche/soziale Aspekte von Daten zu kümmern. Nur weil etwas nicht explizit durch eine Richtlinie, Verordnung oder geltendes Recht ausgeschlossen ist, heißt es nicht notwendigerweise, das dies unter sozialen Gesichtspunkten auch in Ordnung ist. Hierbei geht es vor allem um die Frage, wie die Organisation mit der Gesellschaft nachhaltig umgeht und wie der Balanceakt zwischen Geschäftsoptimierung mit Daten und der persönlichen Freiheit jedes einzelnen Individuums erfolgreich gemeistert wird.

Für die Analyse von Big Data spezifischen Anforderungen können organisationsinterne Standards und Spezifikationen angewendet werden. Wir schlagen vor folgende Anforderungen in Bezug auf Big Data zu erheben (IEEE Anforderungsspezifikation (IEEE, 1984)):

- Allgemeine Anforderungen
    - Produktperspektive (zu anderen Softwareprodukten)
    - Produktfunktionen (eine Zusammenfassung und Übersicht)
    - Benutzermerkmale (Informationen zu erwarteten Nutzern, z.B. Bildung, Erfahrung, Sachkenntnis)
    - Einschränkungen (für den Entwickler)
    - Annahmen und Abhängigkeiten (Faktoren, die die Entwicklung beeinflussen, aber nicht behindern z.B. Wahl des Betriebssystems)

    - Aufteilung der Anforderungen (nicht Realisierbares und auf spätere Versionen verschobene Eigenschaften)
- Spezifische Anforderungen
    - Funktionale Anforderungen
    - Nicht funktionale Anforderungen
    - Externe Schnittstellen
    - Design Constraints
    - Anforderungen an Performance
    - Qualitätsanforderungen
    - Sonstige Anforderungen

Ein weiterer wichtiger Aspekt in Bezug auf die Erhebung der Anforderungen ist die Erhebung der benötigten Kompetenzen für die erfolgreiche Abwicklung eines Big Data Projekts. Hierbei wird auf Kapitel 3.3 verwiesen in welchem die Kompetenzentwicklung in Bezug auf Big Data gesondert behandelt wird.

3.2.3 Vorbereitung

Das Ziel der Phase „Vorbereitung“ ist die Anpassung der aktuellen IT Infrastruktur an die Herausforderungen die sich aus der Umsetzung des Big Data Projekts ergeben. Auf Basis der Bewertung der aktuellen IT Infrastruktur in Bezug auf Software und Hardware ist eine umfangreiche Analyse der IT in der Organisation vorhanden. In der Phase „Anforderungen“ wurden detaillierte Anforderungen in Bezug auf die IT Infrastruktur (ebenfalls Hardware und Software) erhoben. In dieser Phase wird der aktuelle Bestand und dessen Leistungsfähigkeiten mit den erhobenen Anforderungen verglichen und eine Gap Analyse (Kreikebaum, Gilbert, & Behnam, 2011) durchgeführt. Diese hat das Ziel die Lücken zwischen IST und SOLL Zustand zu erheben. Des Weiteren bieten diese Ergebnisse die Grundlage für die Planung der weiteren Adaptierung der IT Infrastruktur in Bezug auf Big Data Projekte. Diese weitere Umsetzung sollte immer in Zusammenhang mit der Gesamtstrategie in Bezug auf Big Data gesehen werden und nicht projektspezifisch behandelt werden um mögliche Synergien zwischen Projekten, Anwendungsfeldern sowie Abteilungen optimal ausnutzen zu können.

Hierbei sollen folgenden Punkten zusätzliche Beachtung geschenkt werden:

- GAP Analyse in Bezug auf Big Data Infrastruktur: Hierbei wird analysiert, wie die aktuelle Kapazität aussieht und wie sich diese zukünftig entwickeln wird. Daraus soll sich ergeben, welche Anforderungen für die nächsten Jahre bestehen. Dies hat auch wesentlichen Einfluss auf die IT-Infrastrukturplanung, welche auf externe Speicher (Cloud) oder auch einem Rechenzentrum (intern/extern) ausgelegt werden kann. Neben der Speicherkapazität ist auch die Frage der Analysesysteme relevant. Diese haben andere technische Anforderungen als Speichersysteme. Ein wichtiger Faktor ist die Virtualisierung und Automatisierung der jeweiligen Systeme. Von gehobener Priorität ist die Frage danach, wie flexibel diese Systeme sind.
- GAP Analyse in Bezug auf Big Data Plattformen: Von höchster Priorität ist hier die Analyse der Skalierung und Flexibilität der aktuell und zukünftigen eingesetzten Plattformen. Des Weiteren sollen die jeweiligen Programmiermodelle überprüft werden, ob diese auch dem aktuellen Stand der Technik entsprechen.

- GAP Analyse in Bezug auf Big Data Analytics: Hierbei kommt die Fragestellung zum Tragen, welche Analysesoftware aktuell verwendet wird und wie diese der Strategie für zukünftige Anwendungen dient. Wichtig sind auch die aktuellen Algorithmen und deren Tauglichkeit.
- GAP Analyse in Bezug auf Big Data Utilization: Wichtig ist hierbei, wie sich die jeweiligen Anwendungen mit Big Data Technologien vereinbaren lassen.
- Evaluierung von neuen und vorhandenen Technologien: Damit soll festgestellt werden, wie sich die jeweiligen Technologien für die Umsetzung von Big Data Projekten eignen und wie die Eigenschaften in Bezug auf die Integration in die Systemlandschaft sind.
- Kompetenzen für Umsetzung von Infrastrukturmaßnahmen
- Sicherheit und Datenschutz in der Architektur

3.2.4 Umsetzung

Die Phase „Umsetzung“ beschäftigt sich mit der konkreten Implementierung und Integration der Big Data Lösung in die IT Systemlandschaft der Organisation. Bei der Umsetzung des Big Projekts sind neben der Anwendung von etablierten IT Projektmanagementstandards die Beachtung folgender Punkte essenziell:

- Möglichkeit der Integration in bestehende IT Systemlandschaft
- Möglichkeit der Skalierung der anvisierten Lösung in Bezug auf
    - Integration von neuen Datenquellen
    - Wachstum der integrierten Datenquellen
    - Geschwindigkeit der Datenproduktion
- Möglichkeit der Erweiterung der Lösung in Bezug auf
    - Analysealgorithmen für zukünftige Problemstellungen
    - Innovative Visualisierungsmöglichkeiten

Zusätzlich ist die Miteinbeziehung der definierten Big Data Strategie während der konkreten Implementierung des Big Data Projekts essenziell. Ein wichtiger Punkt hierbei ist die Erarbeitung einer organisationsweiten Big Data Infrastruktur welche für konkrete Projekte angepasst beziehungsweise erweitert werden kann. Ein Rahmen für die Umsetzung einer ganzheitlichen Big Data Architektur wird in Kapitel 3.5 näher erläutert.

Bereits während der Umsetzung ist ein laufender Soll/Ist Vergleich zu erstellen. Hierbei soll vor allem abgeklärt werden, ob die Ziele mit der umgesetzten Lösung übereinstimmen. Dies dient der Hebung des Projekterfolges. Eine genauere Lösung hierfür wird unter „Ganzheitlichkeit und Optimierung“ dargestellt.

3.2.5 Integration und Konsolidierung

Nachdem erfolgreichen Aufbau der (projektspezifischen) IT Infrastruktur und der Umsetzung des Big Data Projekts ist der nächste Schritt die effiziente Integration der (Projekt-) Infrastruktur und Software in die bestehende IT Systemlandschaft. Die vorher geschaffenen Schnittstellen werden genutzt um einen möglichst reibungsfreien Übergang zwischen den Elementen einer umfassenden Big Data Architektur zu schaffen. Des Weiteren werden die umgesetzten und eingesetzten Big Data Plattformen und Tools in die vorhandenen Toolchains eingebunden.

In dem Schritt „Integration und Konsolidierung“ werden nach der Integration der Hardware und Software in die IT Systemlandschaft auch die neuen Datenquellen in das System übergeführt. Ziele hierbei sind einerseits die Bereitstellung der Daten innerhalb der IT Infrastruktur sowie deren Integration in und mit aktuell vorhandenen Datenquellen. Die Art und Weise der Datenintegration hängt sehr stark von der gewählten Big Data Architektur und den Charakteristiken der Daten ab. Für Streaming Daten müssen gesonderte Vorkehrungen getroffen. Ein wichtiger Punkt für die Integration anderer Daten ist die flexible Kombination beziehungsweise die richtungsweisende Entscheidung zwischen klassischen Data Warehousing Zugängen und flexibleren Ansätzen aus dem Big Data Bereich welche die Art und Weise der Datenintegration stark beeinflussen.

Ziel der Integration der gesamten (projektspezifischen) Big Data Technologien in die IT Systemlandschaft ist die Konsolidierung der vorhandenen Systeme und die Schaffung einer ganzheitlichen Big Data Systemlandschaft die einfach für weitere Big Data Projekte genutzt werden kann und auf deren Basis Daten für die Verwendbarkeit in Geschäftsprozessen vorbereitet werden.

Hierbei ist eine elastische Plattform für die jeweiligen Benutzer des Unternehmens von Vorteil. Diese sollte mit wenig Aufwand für diese zur Verfügung stellen und sich am Cloud Computing Paradigma orientieren. Generell soll ein hohes Ausmaß an Self-Services erreicht werden. Hierbei sind mehrere Techniken möglich. Der Fokus soll jedoch auf diesen liegen:

- Infrastructure as a Service: Die Datenplattformen werden anhand von Infrastrukturdiensten zur Verfügung gestellt. Hierbei entfällt für die Benutzer das Management der darunter liegenden Hardware. Einzelne Benutzer müssen jedoch sehr detailliert planen, wie und welche Instanzen diese einsetzen wollen.
- Platform as a Service: Hierbei wird auch der Software-Stack automatisiert. Benutzer bekommen eine flexible Ausführungsplattform zur Verfügung. Damit kann z.B. Hadoop mit nur wenigen Mausklicks bereitgestellt werden und für jeweilige Lasten flexibel skaliert werden. Diese Form bietet die beste Unterstützung der Benutzer, bedeutet jedoch auch mehr Implementierungsaufwand.

3.2.6 Reporting und Analytics

Der Schritt „Reporting und Analytics“ befasst sich mit der Implementierung und Bereitstellung von Analysealgorithmen welche auf Basis der vorhandenen Infrastruktur und der vorhandenen Daten umgesetzt werden. Die (projektspezifische) Bereitstellung von Technologien und Methoden (siehe Kapitel **Fehler! Verweisquelle konnte nicht gefunden werden.**) beeinflusst die msetzung der Analyseverfahren maßgeblich. Die Auswahl der verwendeten Technologien beeinflusst hierbei die Möglichkeiten bei der Umsetzung und Entwicklung der Analyseverfahren und sollte aus diesem Grund mit großem Augenmerk, gerade in Bezug auf neue und innovative Szenarien, bedacht werden.

Weiters werden in diesem Schritt spezifische Analysemethoden evaluiert und auf Basis der vorhandenen Technologien umgesetzt und für die Integration in Geschäftsprozesse vorbereitet. Hierbei muss auf die Spezifika der einzelnen betroffenen Geschäftsprozesse sowie auf die Anwendungsdomäne (siehe Domänen in Kapitel **Fehler! Verweisquelle konnte nicht gefunden erden.**) eingegangen werden.

3.2.7 Adaptierung

Im Schritt „Adaptierung“ werden die umgesetzten Big Data Technologien auf allen Ebenen des Big Data Stacks in aktuelle und neue Geschäftsprozesse innerhalb des Big Data Projekts angewendet. Durch die Vielfältigkeit der integrierten Daten und die neuartigen Analyseverfahren hat dieser Schritt das Potenzial die etablierten Geschäftsprozesse signifikant und nachhaltig zu verändern. Ziel dieses Schritts ist die Adaptierung der Geschäftsprozesse auf Basis der vorhandenen Datenquellen um diese zu verbessern und Mehrwert zu generieren, also neuen Nutzen (neue Geschäftsfelder, effizientere Geschäftsprozesse, neue Geschäftsmodelle) für die Organisation zu ermöglichen.

3.2.8 Ganzheitlichkeit und Optimierung

Nach der erfolgreichen Abwicklung des Big Data Projekts und der entsprechenden Adaptierung der Geschäftsprozesse hin zu effizienteren und neuen Geschäftsmodellen auf Basis von vorhandenen Datenquellen ist es essenziell die gewonnenen Erkenntnisse nachhaltig innerhalb des Unternehmens zu verankern. Aus diesem Grund werden in dem Schritt „Ganzheitlichkeit und Optimierung“ wichtige Informationen aus der Abwicklung des Big Data Projekts extrahiert und aufbereitet. Ziel hierbei ist es, diese in die vorhandene Big Data Strategie einfließen zu lassen und diese auf Basis der neuen Erkenntnisse zu reflektieren sowie zu erweitern. Diese Erkenntnisse sollen nicht nur in die Big Data Strategie mit einfließen, sondern auch die Bewertung der aktuellen Situation bezüglich Big Data (Big Data Reifegrad Modell und Schritt 1) adaptieren um die Informationen in die Umsetzung von Nachfolgeprojekten einfließen lassen zu können.

Insbesondere werden in diesem Schritt Informationen zu folgenden Punkten gesammelt und in die Big Data Strategie eingearbeitet:

- Vorhandene Datenquellen
    - Typ, Big Data Charakteristiken, Architektur
- Vorhandene Technologien
    - Big Data Stack, Verwendung in Big Data Projekt
- Anforderungen und Herausforderungen
- Potenziale
    - Mehrwertgenerierung
    - Verbesserung von Prozessen
    - Neue Geschäftsprozesse

Durch die Abwicklung und Integration dieses Vorgehensmodells wird die vorhandene Big Data Strategie während der Umsetzung von Projekten verfeinert. Des Weiteren ist ein intaktes Wissensmanagement zu gesammelten Informationen und dem Big Data Reifegrad Modell notwendig und kann die erfolgreiche Umsetzung von Big Data Projekten in Organisationen unterstützen. In diesem Schritt muss auch kritisch reflektiert werden, ob die in den vorherigen Maßnahmen gesetzten Punkte auch zielführend umgesetzt wurden. Hierbei gilt es, Fehler zu analysieren und daraus zu lernen.

3.3 Kompetenzentwicklung

Im Zusammenhang mit der wertschöpfenden Umsetzung von Big Data Projekten in Unternehmen und Forschungseinrichtungen entsteht ein großer Bedarf an zusätzlicher Big Data spezifischer Kompetenz um die effiziente und nachhaltige Entwicklung von neuen Geschäftsprozessen voranzutreiben. Die benötigte Kompetenz umfasst den gesamten in **Fehler! Verweisquelle onnte nicht gefunden werden.** definierten Big Data Stack (Utilization, Analytics, Platform, Management) und es gewinnen neue Technologien, Kenntnisse und Fertigkeiten in diesen Bereichen immer größere Bedeutung. Um Kompetenzen in dem Bereich Big Data innerhalb einer Organisation aufzubauen, können unterschiedliche Maßnahmen gesetzt werden. Gezieltes externes Training von (zukünftigen) Big Data Spezialisten beziehungsweise die Anstellung von neuem in diesem Bereich spezialisiertem Personal können hierbei den ersten wichtigen Schritt setzen. Durch ein vielfältiges und spezialisiertes Angebot an internen Schulungsmaßnahmen kann das Wissen innerhalb einer Organisation verbreitet und vertieft werden.

Grundlage für eine wertschöpfende Umsetzung von Big Data ist die grundsätzliche Verfügbarkeit von hoch qualifiziertem Personal welche innerhalb der Organisation als „early adopters“ eingesetzt werden können. Hierfür bedarf es gerade für einen hochtechnologischen Bereich wie Big Data an grundlegend und wissenschaftlich fundiert ausgebildeten Fachkräften in allen Bereichen des Big Data Stacks. An dieser Stelle der Studie wird auf die Analyse der derzeit angebotenen tertiären Ausbildungen in Österreich (siehe Kapitel **Fehler! Verweisquelle konnte icht gefunden werden.**) verwiesen.

In den geführten Interviews und den abgehaltenen Workshops wurde von Seiten der Teilnehmer auf den grundsätzlichen Bedarf und ein derzeit zu geringes Angebot an hochqualifiziertem Personal in diesem Bereich hingewiesen. Auf Grund dieser Erfahrungen und dem international ersichtlichen Trend zu dem neuen Berufsbild „Data Scientist“ wird dieser Bereich nachfolgend näher beleuchtet.

3.3.1 Data Scientist

Das Berufsbild des Data Scientist hat in der letzten Zeit eine immer größere Bedeutung erlangt. In (Davenport & Patil, 2013) wird dieses Berufsbild als das attraktivste des 21ten Jahrhunderts beworben und in vielen Medien wird auf die Wichtigkeit gut ausgebildeter Fachkräfte für den Bereich Big Data hingewiesen (Bendiek, 2014), (Fraunhofer, 2014). Hierbei wird auch auf das Zitat von Neelie Kroes *„We will soon have a huge skills shortage for data-related jobs.“* (Speech – Big Data for Europe, European Commission – SPEECH/13/893 07/11/2013) hingewiesen. Ein weltweiter Anstieg an Ausschreibungen für dieses neue Berufsbild ist im letzten Jahr zu vermerken und das Berufsbild wird auf unterschiedliche Arten beschrieben. Die Beschreibungen des Berufsbilds Data Scientists reichen von Personen welche Wissen und Methodik aus Analytik, IT und dem jeweiligen Fachbereich vereinigen, über Personen welche Daten in Unternehmen durch die Analyse mit wissenschaftlichen Verfahren und der Entwicklung von prädiktiver Modellen schneller nutzbar machen.

Beispiele der detaillierten Diskussionen der benötigten Fachkenntnisse beschreiben Data Scientists als eine Rolle die aus der Weiterentwicklung von Business oder Data Analysts hervorgeht. Hierfür werden detaillierte Kenntnisse in den Bereichen IT, der Applikation, der

Modellierung, der Statistik, Analyse und der Mathematik benötigt (IBM, 2014). Ein weiterer wichtiger Punkt ist die Fähigkeit Erkenntnisse sowohl an leitende Stellen in Business und IT weitervermitteln zu können. Dementsprechend werden Data Scientists als eine Mischung aus Analysten und Künstlern beschrieben. In (KDNUggets, 2014) werden die benötigten Fähigkeiten von Data Scientists unter anderem folgendermaßen beschrieben: Es werden Kenntnisse in den Bereichen statistischer Analyse, High Performance Computing, Data Mining und Visualisierung benötigt. Im Speziellen werden Informatikkenntnisse über Mathematik, Mustererkennung, Data Mining, Visualisierung, Kenntnisse über Datenbanken, im speziellen Data engineering und data warehousing sowie entsprechendes Domänenwissen und unternehmerischer Scharfsinn gefordert.

In (Mason, 2014) wird der Bereich Data Scientist ebenfalls analysiert. Hier wird das von Organisationen geforderte Profil als Kombination aus Informatik, Hacking, Engineering, Mathematik und Statistik dargestellt und kritisch hinterfragt ob diese Kombination möglich ist *(„data scientists are awesome nerds“*). Eine weitere Diskussion des Berufsbilds wird auf (EMC2, 2014) dargestellt. Hier wird auf die innovative Kombination von quantitativen, technischen und kommunikativen Fähigkeiten mit Skepsis, Neugier und Kreativität verwiesen und somit der wichtige Aspekt der sozialen Fähigkeiten hervorgehoben.

Auf Basis breiter Diskussionen über das Berufsbild Data Scientist und Gesprächen im Rahmen der Studienworkshops, Interviews mit Stakeholdern, sowie vorhandenen Datengrundlagen auf Grund von IDC internen Studien und verfügbaren BITKOM Studien (BITKOM, 2013) wird das Berufsbild hier näher klassifiziert. Auf Grund der breiten Masse an geforderten Fähigkeiten ist die Vereinigung dieser in einer Person schwer zu erreichen.

In **Tabelle 2** wird ein Überblick über wichtige Kompetenzen für das Berufsbild Data Scientist gegeben:

Technische Kompetenzen	Soziale Kompetenzen	Wirtschaftliche Kompetenzen	Rechtliche Kompetenzen
Utilization - Visualisierung - Grafikdesign - Wissensmanagement	**Teamarbeit**	**Identifikation von innovativen Geschäftsmodellen**	**Datenschutz**
Analysis - Innovative Verknüpfung von Daten - Machine Learning - Statistik und Mathematik	**Kommunikation**	**Evaluierung und Umsetzung von innovativen Geschäftsmodellen**	**Ethik**
Platform - Skalierbare Programmierung - Datenmanagement	**Konfliktmanagement**	**Projektmanagement**	
Management - Infrastrukturentwicklung - Infrastrukturbereitstellung - Skalierbarer Speicher - Massive Infrastrukturen	**Führung**	**Projektcontrolling**	

Tabelle 2: Kompetenzen Big Data Scientist

Auf Basis der erhobenen Kenntnisse wird das Berufsbild näher kategorisiert. Hierbei werden fünf spezifische Kategorien analysiert und in **Abbildung 3** deren benötigte Fähigkeiten visualisiert. Unabhängig von der jeweiligen Kategorie wird auf die Wichtigkeit der Einbindung von domänenspezifischem Wissen hingewiesen. In dieser Studie unterscheiden wir zwischen folgenden Berufsbildern:

- Big Data Business Developer
- Big Data Technologist
- Big Data Analyst
- Big Data Developer
- Big Data Artist

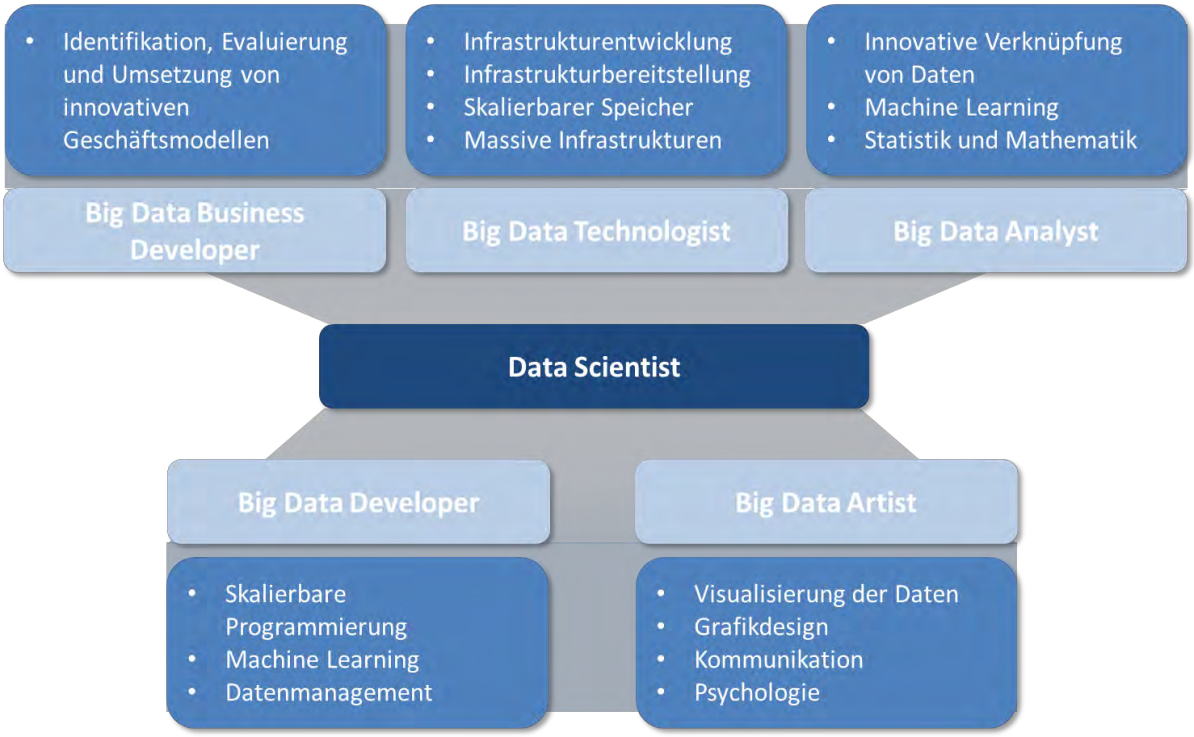


Abbildung 3: Das Berufsbild des Data Scientist

Big Data Business Developer

Das Berufsbild Big Data Business Developer setzt sich aus Kompetenzen in der Mehrwertgenerierung auf Basis von Daten und der Entwicklung aktueller und zukünftiger Geschäftsmodelle zusammen. Einerseits ist das Ziel eines Big Data Business Developers mithilfe von interner Kommunikation neue und innovative Geschäftsfelder und Möglichkeiten auf Basis der vorhandenen und zu integrierenden Daten aufzuspüren und diese gemeinsam mit Mitarbeitern zu entwickeln. Andererseits ist es dessen Aufgabe die Kundenbedürfnisse abzuschätzen, Kontakte zu Daten Providern, Partnern und Kunden zu knüpfen um maßgeschneiderte und innovative Daten-getriebene Produkte zu entwickeln. Demnach ist das Berufsbild als Schnittstelle zwischen der Ebene Utilization und dem Kunden zu sehen.

Big Data Technologist

Das Berufsbild des Big Data Technikers beschäftigt sich mit der Bereitstellung einer für Big Data Szenarien nutzbaren Infrastruktur für die Entwicklung neuer und innovativer Geschäftsmodelle. Hierfür sind breite Kenntnisse von Management und auch Platform Technologien erforderlich. Diese Kenntnisse umfassen die Entwicklung und Instandhaltung von großen Datenzentren, die Verwaltung von Big Data spezifischen Softwarelösungen für die skalierbare Speicherung von großen Datenmengen und auch die Bereitstellung von skalierbare Ausführungsumgebungen für Big Data Analysen.

Big Data Developer

Das Berufsbild des Big Data Entwicklers umfasst die skalierbare Implementierung von Big Data Analysen auf Basis von massiv parallelen Infrastrukturen und Plattformen. Hierfür werden herausragende Kenntnisse in der Parallelisierung von Programmen aus dem Bereich High Performance Computing sowie der skalierbaren Speicherung, effizienten Abfrage von Daten und

der Abfrageoptimierung aus dem Datenbankbereich benötigt. Dieser neue Entwicklertypus erfordert eine grundlegende Informatikausbildung in diesen Bereichen und eine hohe Flexibilität gegenüber neuen Technologien.

Big Data Analyst

Das Berufsbild des Big Data Analysten beschäftigt sich mit dem Auffinden von neuen Verknüpfungen und Mustern in Daten. Hierfür werden fundierte Kenntnisse in den Bereichen Machine Learning, Mathematik und Statistik benötigt. Big Data Analysten entwickeln mathematische oder statistische Modelle, setzen diese mit Hilfe von Big Data Entwicklern um, und wenden diese auf große Datenmengen an um neue Zusammenhänge und Informationen zu generieren.

Big Data Artist

Das Berufsbild des Big Data Artist ist für die visuelle Darstellung und Kommunikation des Mehrwerts für den Endbenutzer und den Kunden zuständig. Hierfür werden einerseits fundierte Kenntnisse in den Bereichen skalierbare Visualisierung, Grafikdesign und Human Computer Interaction für die computergestützte Darstellung der Informationen benötigt. Des Weiteren sollten Big Data Artists eine gute Ausbildung in den Bereichen Kommunikation und Psychologie besitzen um den Effekt der Darstellungsform auf das Gegenüber abschätzen zu können und in das Design einfließen lassen zu können.

3.4 Datenschutz und Sicherheit

Datenschutz und Sicherheit sind in Bezug auf Big Data ein sehr wichtiges Thema. In der Umsetzung einer Big Data Strategie, insbesondere in der Umsetzung von Big Data Projekten, muss diesen Themen eine große Bedeutung beigemessen werden. Datenschutz und Sicherheit umfassen mehrere Dimensionen und diese müssen vor und während der Umsetzung sowie während des Betriebs betrachtet werden. Diese Dimensionen reichen von rechtlichen Fragestellungen aus datenschutzrechtlicher Sicht, Sicherheitsstandards welche für die Umsetzung eines Projektes eingehalten werden müssen, bis zu technischen Umsetzungen des Datenschutzes und der Sicherheit der Infrastruktur. Gerade in Bezug auf Big Data Projekte erlangt die Durchsetzung und Beachtung von Datenschutz enorme Bedeutung. Aktuelle Technologien ermöglichen die Speicherung, Verknüpfung, und Verarbeitung von Daten in noch nie dagewesenen Ausmaßen. In diesem Bereich muss auf die Balance zwischen technologischen Möglichkeiten, rechtlichen Rahmenbedingungen, und persönlichen Rechten auf Datenschutz hingewiesen werden und diese muss in weiterer Folge aufrechterhalten werden. Technologische Umsetzung von Datenschutz und Sicherheit sind grundsätzlich vorhanden doch es wird auf die große Bedeutung der Weiterentwicklung von diesen technologischen Möglichkeiten in Bezug auf die weitere Vernetzung von Daten sowie der rechtlichen Rahmenbedingungen hingewiesen.

In Bezug auf Datenschutz und Sicherheit wurden mehrere Studien in Österreich durchgeführt. In mehreren Leitfäden wird die Situation in Bezug auf Datenschutz und Sicherheit in Österreich von diesbezüglichen Experten im Detail ausgearbeitet und der Öffentlichkeit zur Verfügung gestellt. In diesem Kapitel wird aus diesem Grund auf vorhandene Analysen dieses Bereichs verwiesen und auf die Wichtigkeit von Datenschutz und Sicherheit und diesbezüglichen definierten rechtlichen

Rahmenbedingungen im Bereich Big Data verwiesen. Für eine detailliertere Analyse auf wird auf folgende frei zugängliche Leitfäden in Bezug auf Österreich der EuroCloud[1] und des IT-Cluster Wien[2] sowie auf den BITKOM Leitfaden für Deutschland verwiesen:

- EuroCloud, Leitfaden: Cloud Computing: Recht, Datenschutz & Compliance, 2011 (EuroCloud, 2011)
- IT Cluster Wien, Software as a Service – Verträge richtig abschließen 2., erweiterte Auflage, 2012 (IT Cluster Wien, 2012)
- BITKOM, Leitfaden: Management von Big-Data-Projekten, 2013 (BITKOM, 2013)

3.5 Referenzarchitektur

Für die effiziente Umsetzung eines Big Data Projekts innerhalb einer Organisation ist die Wahl einer geeigneten Architektur essenziell. Neben den zahlreichen verfügbaren Technologien und Methoden sowie der großen Anzahl an verfügbaren Marktteilnehmer (siehe Kapitel **Fehler! erweisquelle konnte nicht gefunden werden.**) sind die Spezifika der verfolgten Business Cases und deren Anforderungen sowie Potenziale in der jeweiligen Domäne für die Implementierung einer geeigneten Architektur, der entsprechenden Frameworks sowie deren Anwendung in der Organisation notwendig. Um diese wichtigen Entscheidungen innerhalb einer Organisation zu unterstützen werden hier zuerst unterschiedliche Fragestellungen für die Wahl der Architektur diskutiert und im Anschluss wird eine Referenzarchitektur für Big Data Systeme für alle genannten Szenarien definiert.

Um die richtigen technologischen Lösungen für Big Data Projekte zu wählen sollten davor einige Fragestellungen geklärt werden. Derzeit ist hierbei die größte Herausforderung die Wahl der technologischen Plattform unter Rücksichtnahme der Verfügbarkeit von Kompetenzen innerhalb der eigenen Organisation. Wie in (Big Data Public Private Forum, 2013) dargestellt gibt es derzeit eine Vielzahl an Diskussionen bezüglich Anforderungen an Technologien welche zu einem großen Teil aus den verfügbaren Daten resultieren. Darüber hinaus werden aber die benötigten Kompetenzen (siehe Kapitel 3.1.2) und das Verständnis für die Herausforderungen und Technologien zu wenig beleuchtet und sind nicht ausreichend vorhanden. Aus diesem Grund hat Yen Wenming in (Yen, 2014) folgende drei Schlüsselfragen definiert und eine grundlegende Kategorisierung von Problemstellungen und Architekturen dargestellt.

- Daten sollen für die Lenkung von Entscheidungen verwendet werden und nicht für deren grundsätzliche Verfügbarkeit gespeichert und analysiert werden.
- Die Analysemethoden sollen regelmäßig an die aktuellen Bedürfnisse angepasst werden.
- Automatisierung der Prozesse ist eine essenzielle Herausforderung um mehr Experimente und Analysen ausführen zu können. Dies ermöglicht die effiziente Betrachtung neuer Fragestellungen und hebt dadurch das Innovationspotenzial.

Des Weiteren wird in **Abbildung 4** eine Referenzarchitektur für Big Data Systeme auf Basis des Big Data Stacks dargestellt. Die vorgestellte Architektur umfasst den gesamten Daten Lebenszyklus innerhalb einer Organisation und bezieht auch bestehende Systeme (z.B. ERP Systeme, Relationale Datenbanken, Data Warehouse) mit ein.

[1] http://www.eurocloud.at/

[2] http://www.clusterwien.at/overview/de/

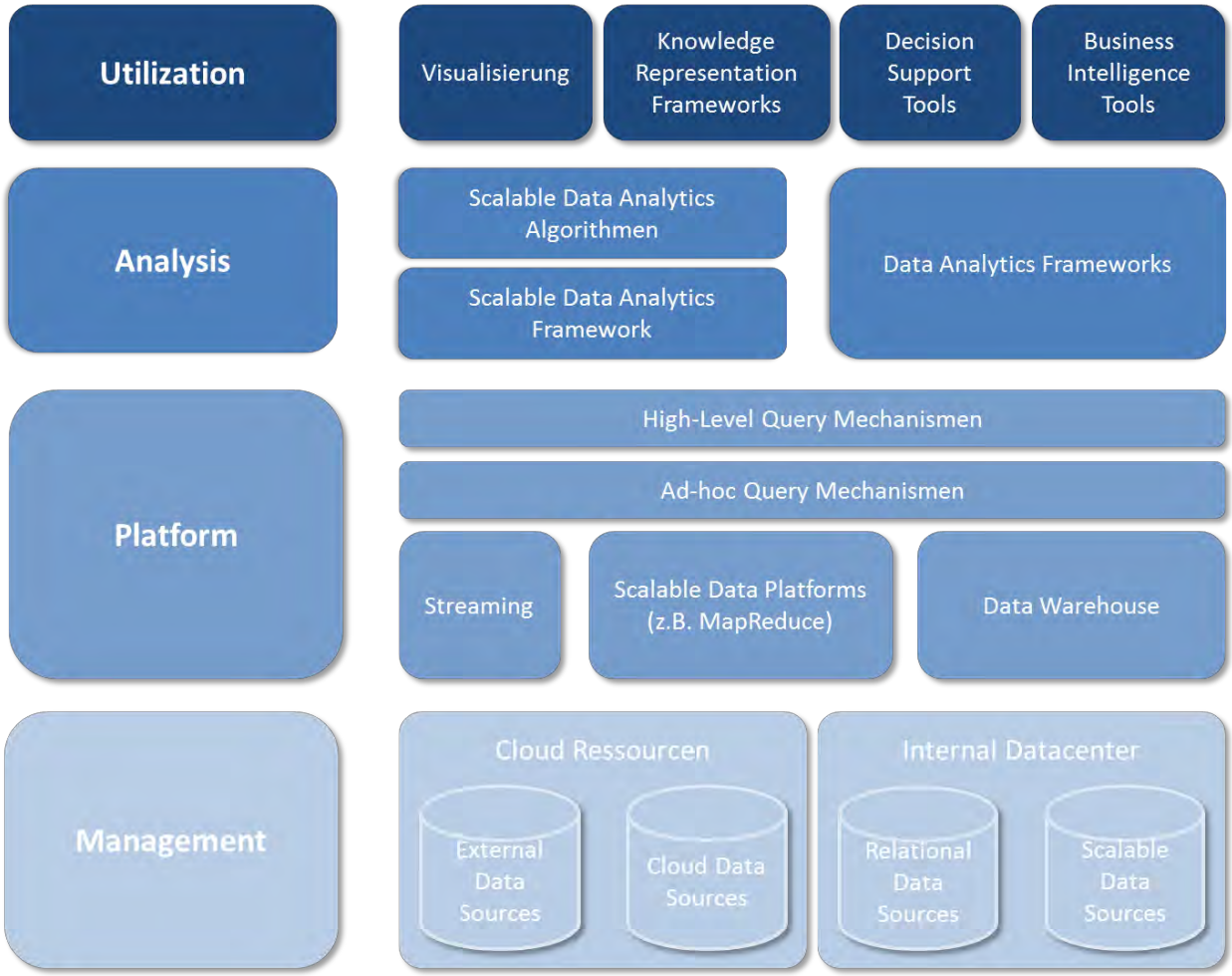


Abbildung 4: Referenzarchitektur für Big Data Systeme

Management Ebene

Die Management Ebene umfasst die Bereitstellung der Infrastruktur sowie die Datenquellen an sich. Die bereitgestellte Infrastruktur wird hierbei in Cloud-basierte Ressourcen und interne Datenzentren unterteilt. Unter Cloud-basierten Systemen werden hierbei einerseits klassische Cloud Ressourcen wie elastische Rechen- und Speicherressourcen verstanden welche von einer Organisation explizit für die Speicherung oder die Berechnung angemietet werden. Andererseits werden in diese Kategorie auch externe Datenquellen welche nach dem Everything as a Service Prinzip (Banerjee & et. al., 2011) angeboten werden eingeordnet. Dies beinhaltet, unter anderem, externe Sensor Systeme (z.B. GPS, Mobilfunk, Wettersensoren) sowie soziale Medien (z.B. soziale Netzwerke und Nachrichtendienste) welche als Service im Internet verfügbar sind. Unter internen Datenzentren werden alle unternehmensinternen Serverlandschaften unabhängig von deren Größe zusammengefasst. Hierbei kann es sich um explizite Big Data Datenzentren großer Firmen wie auch um einzelne Datenbankserver handeln welche in den Big Data Lebenszyklus eingebunden werden. Auf Basis dieser Rechen- und Speicherressourcen werden Daten in unterschiedlichen Formaten und Größen abgespeichert und hierfür werden diverse Systeme verwendet. Für die effiziente Umsetzung eines Big Data Systems werden die Installation oder Anmietung/Einmietung eines Datenzentrums auf Basis von Commodity Hardware (dessen Größe abhängig von der Organisation und der Datenmenge ist) und die Verwendung eines einfachen skalierbaren Speichersystems empfohlen. Hierfür werden von unterschiedlichen Anbietern auch

spezielle vorgefertigte Lösungen angeboten (siehe Kapitel **Fehler! Verweisquelle konnte nicht efunden werden.**). Gemeinhin wird eine Erweiterung der bestehenden IT Landschaft mit Big Data Infrastruktur und die flexible Integration der bestehenden Datenquellen empfohlen.

Ein weiterer essenzieller Punkt auf der Management Ebene ist die effiziente und verteilte Speicherung von großen Datenmengen auf Basis der zur Verfügung stehenden Ressourcen. Eine große Bedeutung haben hierbei verteilte Dateisystemen wie zum Beispiel das Hadoop Distributed File System (HDFS) erlangt. Diese ermöglichen die transparente Daten-lokale Ausführung von Plattform Lösungen. Neben verteilten Dateisystemen können aufkommende NoSQL Systeme für die verteilte und skalierbare Speicherung von großen Datenmengen verwendet werden. Diese bieten eine höhere Abstraktion der Daten bei hoher Skalierbarkeit. Neben diesen Big Data Systemen wird das Daten Ökosystem durch relationale Datenbanken ergänzt.

Plattform Ebene

Die Plattform-Ebene beschäftigt sich mit der effizienten Ausführung von Datenanalyseverfahren auf großen Datenmengen wofür massiv parallele, skalierende und effiziente Plattformen benötigt werden. Die Wahl der richtigen Architektur innerhalb eines Projekts hängt stark von den konkreten Fragestellungen und der Charakteristiken der typischen Datenanalysen ab. Diese werden in (Yen, 2014) im Zuge von Big Data häufig in „Batch Systeme“, „Interaktive Systeme“ und „Stream Verarbeitung“ unterteilt. **Tabelle 3** gibt einen Überblick über die Charakteristiken dieser Architekturen auf Basis dieser Einteilung. Die Systeme werden anhand der Dauer von Abfragen, des unterstützten Datenvolumens, des Programmiermodells und der Benutzer verglichen. Während Systeme für die Batch Verarbeitung für längere Analysen auf sehr großen Datenmengen welche Ergebnisse nicht sofort zur Verfügung stellen müssen eingesetzt werden können, werden interaktive Systeme meistens auf geringeren Datenmengen, dafür aber mit kurzen Zugriffszeiten, eingesetzt. Systeme für die Echtzeit-Verarbeitung von Datenstreams werden gesondert angeführt. Durch die neuesten Entwicklungen im Bereich der Big Data Systeme (z.B. Apache YARN und Apache Spark) verschwimmen die Grenzen zwischen Batch Verarbeitung und Interaktiven Systemen immer mehr. Ziel für die Zukunft ist es hier gemeinsame Systeme für unterschiedliche Anwendungsfälle auf Basis derselben Daten zu entwickeln. Diese Systeme unterstützen die Ausführung von unterschiedlichen Programmiermodellen welche für differenzierte Szenarien gedacht sind.

	Batch Verarbeitung	Interaktive Systeme	Stream Verarbeitung
Abfragedauer	Minuten bis Stunden	Millisekunden bis Minuten	Laufend
Datenvolumen	TBs bis PBs	GBs bis PBs	Durchgängiger Stream
Programmiermodell	MapReduce, BSP	Abfragen	Direkte Azyklische Graphen
Benutzer	Entwickler	Entwickler und Analysten	Entwickler

Tabelle 3: Big Data Architekturen (Kidman, 2014)

In der beschriebenen Referenzarchitektur sind diese Systeme auf Grund dieser Entwicklungen anders eingeteilt. Streaming Lösungen werden für die Echtzeitanbindung von (externen) Datenquellen sowie Sensorsystemen benötigt. Diese Lösungen können Daten in Echtzeit analysieren beziehungsweise die vorverarbeiteten Daten in das Big Data System einspeisen. Skalierbare Datenplattformen verfolgen massiv parallele Programmierparadigmen und werden auf Basis von internen Datenzentren (oder von Cloud Ressourcen) bereitgestellt. Drittens umfasst dieser Bereich auch klassische Data Warehouses welche in Organisationen verfügbar sind. Im Unterschied zu der vorhergehenden Einteilung (siehe **Tabelle 3**) sind Ad-Hoc und High-Level Abfragesprachen über den skalierbaren Plattformen und Date Warehouses angeordnet. Diese können als zusätzliche Abstraktionsschicht für die einfachere Benutzung und den interaktiven Zugriff auf dieselben Daten verwendet werden.

Analytics Ebene

Der Bereich Analytics beschäftigt sich mit der Informationsgewinnung aus großen Datenmengen auf Basis von mathematischen Modellen, spezifischen Algorithmen oder auch kognitiven Ansätzen. Das Ziel hierbei ist es unter anderem, neue Modelle aus Datenmengen zu erkennen, vorhandene Muster wiederzufinden oder neue Muster zu entdecken. Auf dieser Ebene werden Methoden aus dem Machine Learning, der Mathematik oder der Statistik angewendet. Hierfür stehen Organisationen unterschiedlichste Technologien von kommerziellen Betreibern als auch aus dem OpenSource Bereich bereit. Die größte sich hier stellende Herausforderung ist die Anpassung der Algorithmen an massiv parallele Programmiermodelle der Plattform Ebene um diese auf riesigen Datenmengen skalierbar ausführen zu können.

In der Referenzarchitektur werden drei Subbereiche dargestellt: Data Analytics Frameworks, Scalable Data Analytics Frameworks und Scalable Data Analytics Algorithms. In Organisationen werden häufig klassische Data Analytics Frameworks produktiv in den unterschiedlichsten Bereichen eingesetzt. Derzeitiges Ziel der Hersteller ist es diese Frameworks an die Herausforderungen im Bereich Big Data anzupassen. Dies geschieht meistens durch die nahtlose Unterstützung und Integration von unterschiedlichen massiv parallelen Programmiermodellen. Da es sich hierbei um einen wichtigen aber derzeit noch nicht vollständig abgeschlossenen Prozess handelt werden diese Frameworks gesondert dargestellt. Neben klassischen Data Analytics Frameworks entstehen immer mehr neue Frameworks welche speziell

für die massiv parallele Ausführung von Datenanalysen entwickelt werden. Hierbei entstehen abstrahierte Programmiermodelle welche die Parallelisierung (teilweise) vor den Anwendern verstecken können und gleichzeitig hochperformante Bibliotheken für Machine Learning, Mathematik und Statistik in das Framework einbauen. Die Entwicklung dieser Frameworks ist derzeit ein aufstrebender Bereich und gerade diese werden den Bereich Big Data in den nächsten Jahren vorantreiben.

Algorithmen für spezifische Problemstellungen werden auf Basis der zur Verfügung stehenden Frameworks für Datenanalyse entwickelt und bereitgestellt. Hierbei ist auf die Anwendbarkeit der Algorithmen auf große Datenmengen und der möglichst transparenten massiven Parallelisierung dieser zu achten. Mittlerweile existieren einige skalierbare Machine Learning Frameworks welche ausgewählte Algorithmen in skalierbarer Form bereitstellen. Generell ist für die Analyseebene ein breites Spektrum an Kompetenz (richtige Algorithmen, massiv parallele und skalierbare Implementierungen, Verwendung von Big Data Frameworks) von Nöten und dies ist auch in der Auswahl der Tools zu beachten.

Utilization Ebene

Der Einsatz von Big Data-Technologien und -Verfahren wird von Unternehmen und Forschungseinrichtungen unabhängig von dem Geschäftsfeld mit einem bestimmten Ziel angedacht: Big Data-Technologien sollen Mehrwert generieren und dadurch die aktuelle Marktsituation der Organisation stärken. Utilization Technologien bilden diese Schnittstelle zwischen dem Benutzer und den darunter liegenden Technologien. Hierbei wird im Rahmen der Referenzarchitektur einerseits auf die notwendige Integration in die vorhandene Toolchain in Bezug auf Wissensmanagement, Business Intelligence und entscheidungsunterstützende Systeme verwiesen. Des Weiteren kann die Integration von Wissensrepräsentationssystemen, skalierbaren und interaktiven Visualisierungssystemen sowie von Visual Analytics als einfache und interaktive Schnittstelle für Benutzer das Potenzial von Big Data Technologien innerhalb der Organisation enorm heben.

Literaturverzeichnis

AIT Mobility Department. (2013). *Strategy 2014-2017.* Vienna.

al., D. V. (2014). *IDC's Worldwide Storage and Big Data Taxonomy, 2014.* Framingham: IDC.

al., D. V. (2012). *Perspective: Big Data, Big Opportunities in 2013.* Framingham: IDC.

al., D. V. (2013). *Worldwide Big Data Technology and Services 2013–2017 Forecast.* Framingham: IDC.

Alpaydin, E. (2010). *Introduction to Machine Learning, second edition.* London, England: The MIT Press.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks, 54(15)* , S. 2787-2805.

Banerjee, P., & et. al. (2011). Everything as a service: Powering the new information economy. *Computer 44.3* , S. 36-43.

Battre, D., Ewen, S., Hueske, F., Kao, O., Markl, V., & Warneke, D. (2010). Nephele/PACTs: A Programming Model and Execution Framework for Web-Scale Analytical Processing. *Proceedings of the ACM Symposium on Cloud Computing (SoCC)* .

Bendiek, S. (2014). *itdaily*. Abgerufen am 08. April 2014 von http://www.it-daily.net/analysen/8808-die-zukunft-gehoert-den-data-scientists

Berners-Lee, T. (2001). The semantic web. *Scientific American* , S. 28-37.

Bierig, R., Piroi, F., Lupu, M., Hanburry, A., Berger, H., Dittenbach, M., et al. (2013). Conquering Data: The State of Play in Intelligent Data Analytics.

Big Data Public Private Forum. (2013). *Consolidated Technical Whitepapers.*

Big Data Public Private Forum. (2013). *First Draft of Sector's Requisites.*

BITKOM. (2012). *Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte.*

BITKOM. (2013). *Leitfaden: Management von Big-Data-Projekten.*

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. In *International Journal on Semantic Web and Information Systems (IJSWIS) 5.3* (S. 1-22).

BMVIT. (2014). *Bundesministerium für Verkehr, Innovation und Technologie* .

Booch, G., Rumbaugh, J., & Jacobson, I. (1998). The Unified Modeling Language (UML). *http://www. rational. com/uml/(UML Resource Center)* .

Brewer, E. (02 2012). Pushing the CAP: Strategies for Consistency and Availability. *IEEE Computer, vol. 45, nr. 2* .

Brewer, E. (2000). Towards Robust Distributed Systems. *Proceedings of 9th Ann. ACM Symp. on Principles of Distributed Computing (PODC 00)* .

Brown, B., Chui, M., & Manyika, J. (2011). *Are you ready for the era of 'big data'?* http://unm2020.unm.edu/knowledgebase/technology-2020/14-are-you-ready-for-the-era-of-big-data-mckinsey-quarterly-11-10.pdf: McKinseyQuarterly.

Bundeskanzleramt. (2014). *Digitales Österreich*. Abgerufen am April 2014 von http://www.digitales.oesterreich.gv.at/site/5218/default.aspx

Buyya, R., Broberg, J., & Goscinski, A. (2011). *Cloud Computing: Principles and Paradigms.* Wiley.

Cheng T. Chu, Sang K. Kim, Yi A. Lin, Yuanyuan Yu,, & Gary R. Bradski, Andrew Y. Ng, Kunle Olukotun. (2006). Map-Reduce for Machine Learning on Multicore. *NIPS* .

Cisco. (2014). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018.*

Cloudera. (kein Datum). *Cloudera Impala, Open Source, Interactive SQL for Hadoop*. Abgerufen am 2 2014 von http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html

Colling, D., Britton, D., Gordon, J., Lloyd, S., Doyle, A., Gronbech, P., et al. (01 2013). Processing LHC data in the UK. *Philisophical transactions of the royal society* .

Cooperation OGD Österreich: Arbeitsgruppe Metadaten. (2013). *OGD Metadaten - 2.2.* Abgerufen am April 2014 von https://www.ref.gv.at/uploads/media/OGD-Metadaten_2_2_2013_12_01.pdf

Critchlow, T., & Kleese Van Dam, K. (2013). *Data-Intensive Science.* CRC Press.

Davenport, T., & Patil, D. (2013). *Harvard Business Review*. Von http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ abgerufen

Dax, P. (2011). Transparenz und Innovation durch offene Daten. *futurezone.at* .

Dax, P., & Ledinger, R. (November 2012). *Open Data kann Vertrauen in Politik stärken*. Von http://futurezone.at/netzpolitik/11813-open-data-kann-vertrauen-in-politik-staerken.php abgerufen

Dean, J., & Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM, vol. 51* , S. 107-113.

(2013). *Demystifying Big Data - A Practical Guide To Transforming The Business of Government.* TechAmericaFoundation.

Dowd, K., Severance, C., & Loukides, M. (1998). *High performance computing. Vol.2.* O'Reilly.

DuBois, L. (2013). *Worldwide Storage in Big Data 2013–2017 Forecast.* Framingham: IDC.

Edlich, S., Friedland, A., Hampe, J., Brauer, B., & Brückner, M. (2011). *NoSQL Einstieg in die Welt nichtrelationaler Datenbanken.* Hanser.

Eibl, G., Höchtl, J., Lutz, B., Parycek, P., Pawel, S., & Pirker, H. (2012). *Open Government Data – 1.1.0.*

Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., & Bae, S.-H. (2010). Twister: A Runtime for Iterative MapReduce. *19th ACM International Symposium on High Performance Distributed Computing* .

Ekanayake, J., Pallickara, S., & Fox, G. (2008). MapReduce for Data Intensive Scientific Applications. *IEEE International Conference of eScience* .

EMC2. (2014). Abgerufen am 08. April 2014 von https://infocus.emc.com/david_dietrich/a-data-scientist-view-of-the-world-or-the-world-is-your-petri-dish/

EU Project e-CODEX. (2013). Von http://www.e-codex.eu/ abgerufen

EU Project Envision. (2013). Von http://www.envision-project.eu/. abgerufen

EU Project PURSUIT. (2013). Von http://www.fp7-pursuit.eu/PursuitWeb/. abgerufen

EuroCloud. (2011). *Leitfaden: Cloud Computing: Recht, Datenschutz & Compliance.*

Europäische Kommission. (25. 1 2012). *Vorschlag für VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES*. Von http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:DE:PDF abgerufen

European Commission. (2010). *The European eGovernment Action Plan 2011-2015.*

Fayyad, U., Piatestsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Maganzine* .

Federal Chancellery of Austria. (2013). *What is eGovernment?* Von http://oesterreich.gv.at/site/6878/default.aspx abgerufen

Forbes. (2013). *A Very Short History Of Big Data*. Abgerufen am 16. 08 2013 von http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Fraunhofer. (2014). Abgerufen am 08. April 2014 von http://www.iais.fraunhofer.de/data-scientist.html

George, L. (2011). *HBase: The Definitive Guide.* O'Reilly.

Gorton, I., & Gracio, D. (2012). *Data-Intensive Computing: A Challenge for the 21st Century.* Cambridge.

Grad, B., & Bergin, T. J. (10 2009). History of Database Management Systems. *IEEE Annals of the History of Computing Volume 31, Number 4* , S. 3-5.

Grothe, M., & Schäffer, U. (2012). *Business Intelligence.* John Wiley & Sons.

Gu, Y., & Grossman, R. (2009). Sector and Sphere: the design and implementation of a high-performance data cloud. *Philosophical Transactions of the royal society* .

Gu, Y., & Grossman, R. (2007). UDT: UDP-based data transfer for high-speed wide area networks. *Computer Networks* .

Habala, O., Seleng, M., Tran, V., Hluchy, L., Kremler, M., & Gera, M. (2010). Distributed Data Integration and Mining Using ADMIRE Technology. *Grid and Cloud Computing and its Applications, vol. 11 no. 2* .

Haerder, T., & Reuter, A. (1983). Principles of transaction-oriented database recovery. *ACM Computing Surveys, Vol. 15, Nr. 4* , S. 287-317.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, 2nd Edition.* The Morgan Kaufmann Series in Data Management Systems.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space (1st edition), Synthesis Lectures on the Semantic Web: Theory and Technology.* Morgan & Claypool.

Heise, A., Rheinländer, A., Leich, M., Leser, U., & Naumann, F. (2012). Meteor/Sopremo: An Extensible Query Language and Operator Model. *BigData Workshop (2012)* .

Hewitt, E. (2010). *Cassandra: The Definitive Guide.* O'Reilly.

Hey, T., & Trefethen, E. (2005). Cyberinfrastructures for e-Science. Science. *Science Magazine* .

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. *Microsoft Research* .

HL7. (2013). *HL/ Electronic Health Record*. Von http://www.hl7.org/ehr/. abgerufen

Hüber, B., Kurnikowski, A., Müller, S., & Pozar, S. (2013). *Die wirtschaftliche und politische Dimension von Open Government Data in Österreich.*

IBM. (2011). *IBM Big Data Success Stories.*

IBM. (2014). *What is a data scientist?* Von http://www-01.ibm.com/software/data/infosphere/data-scientist/ abgerufen

IEEE. (1984). *Guide to Software Requirements Specification. ANSI/IEEE Std 830-1984 .* Piscataway/New Jersey: IEEE Press.

IHT SDO. (2013). *SNOMED CT*. Von http://www.ihtsdo.org/snomed-ct/ abgerufen

IT Cluster Wien. (2012). *Software as a Service – Verträge richtig abschließen 2., erweiterte Auflage.*

Jackson, M., Antonioletti, M., Dobrzelecki, B., & Chue Hong, N. (2011). Distributed data management with OGSA-DAI. *Grid and Cloud Database Management* .

Johnston, W. (1998). High-Speed, Wide Area, Data Intensive Computing: A Ten Year . *Seventh IEEE International Symposium on High Performance Distributed Computing* .

KDNUggets. (2014). *KDNUggets*. Abgerufen am 08. April 2014 von http://www.kdnuggets.com

Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. In *Visual Data Mining* (S. 76-90). Berlin: Springer.

Kell, D. (2009). In T. Hey, S. Tansley, & K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery.*

Khoshafian, S., Copeland, G., Jagodits, T., Boral, H., & Valduriez, P. (1987). *A Query Processing Strategy for the Decomposed Storage Model.* ICDE.

Kidman, A. (2014). Abgerufen am 08. April 2014 von http://bigdataanalyticsnews.com/choose-best-tool-big-data-project/

Koehler, M. (2012). A service-oriented framework for scientific Cloud Computing.

Koehler, M. e. (10 2012). The VPH-Share Data Management Platform: Enabling Collaborative Data Management for the Virtual Physiological Human Community. *The 8th International Conference on Semantics, Knowledge & Grids, Beijing, China* .

Koehler, M., & Benkner, S. (2009). A Service Oriented Approach for Distributed Data Mediation on the Grid. *Eighth International Conference on Grid and Cooperative Computing* .

Koehler, M., Kaniovskyi, Y., Benkner, S., Egelhofer, V., & Weckwerth, W. (2011). A cloud framework for high troughput biological data processing. *International Symposium on Grids and Clouds, PoS(ISGC 2011 & OGF 31)069* .

Kompa, M. (2010). Wird Island die Schweiz der Daten? *Telepolis* .

Kreikebaum, H., Gilbert, D. U., & Behnam, M. (2011). *Strategisches Management.* Stuttgart: Kohlhammer.

Leavitt, N. (14. 02 2010). Will NoSQL Databases Live Up to Their Promise? *IEEE Computer,vol.43, no.2* .

Malewicz, G., Austern, M., Bik, A., Dehnert, J., Horn, I., Leiser, N., et al. (2010). Pregel: a system for large-scale graph processing. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* .

Markl, V. (28. November 2012). *Big Data Analytics – Technologien, Anwendungen, Chancen und Herausforderungen.* Wien.

Markl, V., Löser, A., Hoeren, T., Krcmar, H., Hemsen, H., Schermann, M., et al. (2013). *Innovationspotentialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen (Big Data Management).* BMWi.

Mason, H. (2014). Abgerufen am 08. April 2014 von http://www.forbes.com/sites/danwoods/2012/03/08/hilary-mason-what-is-a-data-scientist

McGuiness, D., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation* .

McKinsey&Company. (04 2013). *How big data can revolutionize pharmaceutical R&D*. Von http://www.mckinsey.com/insights/health_systems_and_services/how_big_data_can_revolutionize_pharmaceutical_r_and_d abgerufen

Melnik, S., Gubarev, A., Long, J., Romer, G., Shivakumar, S., Tolton, M., et al. (2010). Dremel: Interactive Analysis of Web-Scale Datasets. *Proceedings of the VLDB Endowment* .

Message Passing Interface Forum. (21. September 2012). MPI: A Message-Passing Interface Standard.

Moniruzzaman, A. M., & Akhter Hossain, S. (2013). NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. *ArXiv e-prints* .

NESSI. (2012). *Big Data - A New World of Opportunities.*

Open Government Data. (2013). Abgerufen am 2013 von http://gov.opendata.at

OpenMP Application Program Interface, Version 4.0. (July 2013). OpenMP Architecture Review Board.

Organisation of American States. (2013). *e-Government*. Von http://portal.oas.org/Portal/Sector/SAP/DptodeModernizaci%C3%B3ndelEstadoyGobernabilidad/NPA/SobreProgramadeeGobierno/tabid/811/language/en-US/default.aspx abgerufen

Poole, D., & Mackworth, A. (2010). *Artificial Intelligence, Foundations of Computational Agents.* Cambridge University Press.

Popper, K. (2010). *Die beiden Grundprobleme der Erkenntnistheorie: aufgrund von Manuskripten aus den Jahren 1930-1933.* Vol. 2. Mohr Siebeck.

Proceedings of ESWC. (2013).

Sabol, V. (2013). Visual Analytics. TU Graz.

Sendler, U. (2013). *Industrie 4.0–Beherrschung der industriellen Komplexität mit SysLM (Systems Lifecycle Management).* Springer Berlin Heidelberg.

Shim, J., Warketin, M., Courtney, J., Power, D., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, Volume 33, Issue 2* .

Tachmazidis, I. et al. (2012). Scalable Nonmonotonic Reasoning over RDF data using. MapReduce. *Joint Workshop on Scalable and High-Performance Semantic Web Systems (SSWS+ HPCSW 2012).*

Tiwari, S. (2011). *Professional NoSQL.* wrox.

Top500 Supercomputer Sites. (08 2013). Von http://www.top500.org/. abgerufen

Trillitzsch, U. (2004). Die Einführung von Wissensmanagement. *Dissertation* . St. Gallen, Schweiz.

Ussama, F., & et. al. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. American Association for Artificial Intelligence .

Valiant, L. (1990). A bridging model for parallel computation. *Communications of the ACM, vol.33, no. 8* , S. 103-111.

W3C. (2013). *OWL Web Ontology Language*. Von http://www.w3.org/TR/webont-req/#onto-def. abgerufen

Wadenstein, M. (2008). *LHC Data Flow.*

Wahlster, W. (1977). *Die repräsentation von vagem wissen in natürlichsprachlichen systemen der künstichen intelligenz.* Universität Hamburg, Institut für Informatik.

Warneke, D., & Kao, O. (2009). Nephele: Efficient Parallel Data Processing in the Cloud. *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers* .

Wirtschaftskammer Österreich. (2013). *Die österreichische Verkehrswirtschaft, Daten und Fakten - Ausgabe 2013.*

Wrobel, S. (2014). Big Data – Vorsprung durch Wissen. *BITKOM Big Data Summit.*

Yen, W. (2014). *Using Big Data to Advance Your Cloud Service and Device Solutions*. Abgerufen am 08. April 2014 von http://channel9.msdn.com/Events/Build/2014/2-645

Zhang, Y., Meng, L., Li, H., Woehrer, A., & Brezany, P. (2011). WS-DAI-DM: An Interface Specification for Data Mining in Grid Environments. *Journal of Software, Vol 6, No 6* .