Smart Data

Smart-Data-Technologien

des BMWi-Technologieprogramms
„Smart Data – Innovationen aus Daten“

Impressum

Herausgeber
Smart-Data-Begleitforschung
FZI Forschungszentrum Informatik
Außenstelle Berlin
Friedrichstr. 60, 10117 Berlin
www.smart-data-programm.de

Konzeption und Gestaltung
LoeschHundLiepold Kommunikation GmbH, Berlin

Stand
November 2015

Druck
WIRmachenDRUCK GmbH

Bildnachweis
Victoria – Fotolia.com (Titel), FZI Forschungszentrum Informatik (S. 4, 22), Michael Beigl (S. 22), Christoph Heinz (S. 22), Christoph Boden (S. 22), Volker Markl (S. 23), Henrik Oppermann (S. 23), Nico Schlitter (S. 23), Till Riedel (S. 23), Bernhard Neumair (S. 23)

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Inhalt

Editorial 4

1 Von Technologien zu Standards 5
1.1 Bestehende Plattformen 6
1.2 Standardisierungsgrad bestehender Plattformen 8
1.3 Standards: de jure oder de facto? 8

2 Big Data in Motion 10
2.1 Auf dem Weg von Rohdaten zu Dashboards 10
2.2 Wohin geht die Reise? 11

3 Smart Data Technology Stack 12
3.1 Anwendungsbeispiele aus der Druckindustrie 12
3.2 Anforderungen an Datenhaltung, -zugriff und -verarbeitung 13
3.3 Die Gesamtarchitektur und Fazit 15

4 Einführung in Apache Flink 16
4.1 Apache Flink: eine Streaming Dataflow Engine 16
4.2 Robuste Ausführungsumgebung mit automatischer Optimierung 16
4.3 Prägnante APIs für typsicheren und wartbaren Code 17
4.4 Vielfältige Integrationsmöglichkeiten und wachsende Community 18
4.5 Basistechnologie im Berlin Big Data Center 18

5 Das SDIL Smart Data Testbed 19
5.1 Die technische Basis des SDIL und deren Nutzung 19
5.2 Der Data Innovation Cycle 20

Über die Autoren 22

Fußnoten 24

Abbildungsverzeichnis 25

Editorial

Prof. Dr. Dr. h. c. Stefan Jähnichen, FZI Forschungszentrum Informatik / Smart-Data-Begleitforschung

Sehr geehrte Leserinnen und Leser,

Big Data ist in aller Munde und die Auswertung von großen Datenmengen ist eine Herausforderung, die uns in der Industrie, im Dienstleistungssektor, in der Mobilität, der Gesundheitswirtschaft, im Energiebereich und in anderen Branchen noch viele Jahre beschäftigen wird. Immer öfter fällt dabei auch die Bezeichnung „Smart Data“, in Anlehnung an die vernetzenden Konzepte Smart Home, Smart Grid, Smart Car oder gar Smart Senior.

Doch was bedeutet eigentlich „Smart“ im Zusammenhang mit der Nutzung großer Datenmengen? „Big Data“ bezeichnet zunächst Datenmengen, deren Größe, Komplexität oder Heterogenität zu Ressourcenengpässen führt und die deshalb mit bisherigen Methoden nicht mehr analysiert oder verarbeitet werden können. Der Ansatz von Smart Data hingegen geht über diesen Bezug nur auf die Größe der Datenmenge hinaus: Smart Data sind aus den heterogenen und sehr komplexen Datensätzen abgeleitete nutzbringende, hochwertige und abgesicherte Informationen, bei denen man neben der technischen Beherrschung der Datenmenge auch die Datenqualität, die Datensicherheit, den Datenschutz und den Nutzen im Blick hat. Grundlage für diese Aufwertung und Nutzung von Datenbeständen sind Technologien, die eine Verarbeitung erst möglich machen.

Auf dem Markt gibt es eine Vielzahl von Tools, die teils auf Cloud-Technologien basieren und es ermöglichen, große Datenmengen skalierbar zu verarbeiten. Besonders die Open-Source-Tools „Spark“ und „Flink“ der Apache Foundation werden in diesem Zusammenhang immer wieder genannt, denn sie erlauben jenseits von proprietären Lösungen wie Matlab die Verarbeitung großer Datenmengen für wenig Geld. Dies ist besonders für die Forschung von großem Nutzen, aber auch die Wirtschaft kann davon profitieren, insbesondere dann, wenn sich Firmen wie USU dieser Technologien annehmen und sie für den professionellen Einsatz bereithalten. Abseits der Open-Source-Tools bieten namhafte Software-Hersteller wie die Software AG, IBM und SAP ebenfalls eine Reihe von Werkzeugen an. Und auch die Big-Data-Kompetenzzentren BBDC und SDIL haben sich Big-Data-Technologien und deren Einsatz verschrieben. Speziell wenn es um eben diesen Einsatz und die mögliche Kombination verschiedener Technologien geht, sind Standards von besonderer Bedeutung und bedürfen daher besonderer Aufmerksamkeit.

In den verschiedenen Projekten des Technologieprogramms „Smart Data – Innovationen aus Daten“ des Bundesministeriums für Wirtschaft und Energie kommen diese Technologien zum Einsatz. Die Begleitforschung des Technologieprogramms hat daher im Juni 2015 einen Technologie-Workshop mit über 50 Experten durchgeführt, bei dem erfahrene Anwender aus Wissenschaft und Praxis verschiedene Big-Data-Tools vorstellen und deren jeweilige Anwendungskontexte und Nutzungsmöglichkeiten diskutieren konnten. Die Ergebnisse und Inhalte dieses Workshops sind in der vorliegenden Broschüre zusammengestellt, die einen kurzen Überblick über den aktuellen Stand der Technik gibt.

Herzlich bedanken möchte ich mich bei allen Autorinnen und Autoren für ihre Mitarbeit, durch die sie wesentlich zum Gelingen dieser Technologiebroschüre beigetragen haben. Allen Leserinnen und Lesern wünsche ich viel Freude bei der Lektüre.

Ihr Prof. Dr. Dr. h. c. Stefan Jähnichen,
FZI Forschungszentrum Informatik und Leiter der Smart-Data-Begleitforschung

1 Von Technologien zu Standards

Dr. Alexander Lenk, FZI Forschungszentrum Informatik / Smart-Data-Begleitforschung[1]

Die Bezeichnung von Daten als „Rohstoff des 21. Jahrhunderts“ ist inzwischen weltweit zum geflügelten Wort geworden. Grund dafür ist das enorme Potential, das Daten zur Steigerung des globalen Wirtschaftswachsums und der Wettbewerbsfähigkeit Deutschlands zugeschrieben wird. Besonders vielversprechend ist dabei die wirtschaftlich sinnvolle Nutzung großer, bereits erfasster Datenbestände, auch „Big Data“ genannt. Dabei werden enorme Datenmengen aus vielfältigen Quellen gesammelt und ausgewertet und anschließend wirtschaftlich nutzbar gemacht. Um nur einige Zahlen zu nennen: Studien prognostizieren einen rasanten Anstieg des weltweiten Umsatzvolumens durch Big Data auf über 50 Mrd. Euro bis zum Jahr 2017, sechs Mrd. Euro davon entfallen allein auf Deutschland.[2]

Doch Zuwachs und Komplexität der Daten stellen die Unternehmen auch vor vielfältige Probleme. Laut Studien sind nach wie vor 80 Prozent der weltweit anfallenden Daten unstrukturiert und 95 Prozent mit bislang gängigen Verfahren nicht automatisiert auswertbar.[3] Der Grund hierfür ist, dass die Daten oft in reiner Textform statt in Datenbanken vorliegen. Das erschwert das Sammeln und Verarbeiten dieser Datenmengen. Zudem mangelt es an technischen Lösungen, durch die derartige Daten analysiert, verarbeitet und mit Zusatzinformationen angereichert werden können. Für eine wirtschaftliche und gesellschaftliche Nutzung ist eine entsprechende Datenaufbereitung jedoch unerlässlich. Der Weg hin zu verwertbaren Informationen aus heterogenen und unstrukturierten Massendaten führt über semantische Technologien. Diese weisen den rohen Daten eine Bedeutung zu und bilden somit die Grundlage für deren intelligente Verarbeitung, Strukturierung und Nutzung. Durch diesen „Veredelungsprozess“ wird „Big“ Data erst zu „Smart“ Data.[4]

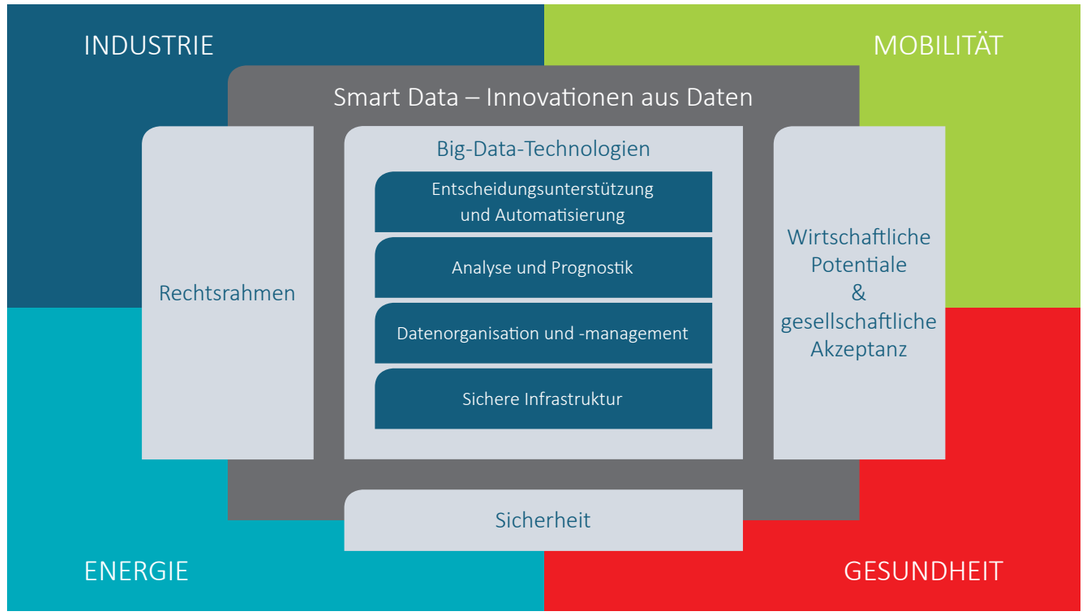


Abbildung 1: das Technologieprogramm „Smart Data – Innovationen aus Daten“[5] des BMWi

Der Begriff „Smart Data“ beschreibt also den Entwicklungsschritt von den ursprünglich unstrukturierten Massendaten aus Big Data hin zu generiertem Wissen. Dieses Wissen bietet die Grundlage für technologische Innovationen in den unterschiedlichsten Branchen und Arbeitsgebieten, die im Rahmen des Technologieprogramms „Smart Data – Innovationen aus Daten“ vom Bundesministerium für Wirtschaft und Energie (BMWi) gefördert werden. Dabei wurden dreizehn Leuchtturmprojekte ausgewählt, die sich in vier Anwendungsfelder gliedern: Industrie, Mobilität, Gesundheit und Energie (siehe Abbildung 1).

In jedem der Anwendungsfelder sind interessante Innovationen zu erwarten, vor allem durch die Kombination der aktuell auf dem Markt verfügbaren Smart-Data-Basistechnologien. Für diese Technologien lässt sich wiederum eine Unterteilung in vier Themen vornehmen: sichere Infrastruktur, Datenorganisation und -management, Analyse und Prognostik sowie Entscheidungsunterstützung und Automatisierung.

- Im Bereich Sichere Infrastruktur steht die Verarbeitung der Daten im Vordergrund. Qualitätseigenschaften wie Performance, Skalierbarkeit, Zuverlässigkeit, Verfügbarkeit und Sicherheit sind hier von besonderer Relevanz. Vor allem Infrastrukturen wie Cloud Computing haben sich dabei bisher als geeignet erwiesen.
- Im Bereich Datenorganisation und -management werden die Herausforderungen erfasst, die sich aus der Komplexität der unstrukturierten vielfältigen Daten ergeben. Zudem wird hier der Einsatz semantischer Technologien betrachtet, die die Massendaten in einen Kontext setzen und so ihre Bedeutung und ihren Inhalt maschinell interpretierbar machen.
- Die Herausforderungen im Umgang mit Massendaten werden in der Analyse und Prognostik mit Hilfe analytischer Verfahren aus den Bereichen Statistik und Semantik bearbeitet, z. B. mit Verfahren des maschinellen Lernens oder der künstlichen Intelligenz.
- Auf Basis der hieraus abgeleiteten Prognosen können in der Entscheidungsunterstützung und Automatisierung die Geschäfts- und Entscheidungsprozesse in Unternehmen unterstützt werden.[6]

Neben den beschriebenen Anwendungsfeldern und Basistechnologien werden im Rahmen des Technologieprogramms die drei Querschnittsthemenfelder „Sicherheit“, „Rechtsrahmen“ sowie „Wirtschaftliche Potentiale & gesellschaftliche Akzeptanz“ behandelt, die fach- und projektübergreifende Themen und Fragestellungen adressieren:

- Das Thema Sicherheit bildet die technische Basis für die anderen Querschnittsthemen. Im Fokus stehen dabei Verfahren zur Abwehr ungewollter Zugriffe sowie Sicherheitslösungen für die Datenverarbeitung, -speicherung und -anonymisierung.[7]
- Das Querschnittsthema Rechtsrahmen umfasst Fragen der rechtlichen Rahmenbedingungen bei der Entwicklung und Nutzung neuer Technologien im Kontext von Smart Data und formuliert Empfehlungen zur Weiterentwicklung des gesetzlichen Rahmens.
- Im Themengebiet „Wirtschaftliche Potentiale & gesellschaftliche Akzeptanz“ werden neue Smart-Data-Geschäftsmodelle und die speziellen Herausforderungen im Bereich der Informations- und Netzwerkökonomie betrachtet. Zentrale Aufmerksamkeit liegt hierbei auch auf Fragen der gesellschaftlichen Akzeptanz derartiger disruptiver Technologieinnovationen.[8]

1.1 Bestehende Plattformen

Heute kommen bereits unterschiedliche Smart-Data-Plattformen zur Auswertung großer Datenmengen in Forschung und Industrie zum Einsatz. Betrachtet man den inneren Aufbau dieser Plattformen, so zeigt sich eine sechsgliedrige Taxonomie[9] (siehe Abbildung 2).

- Zur Datenhaltung stehen verschiedene Technologien zur Verfügung, zu denen insbesondere verteilte Dateisysteme (DFS, z. B. HDFS) sowie relationale, nichtrelationale und In-Memory-Datenbanken wie z. B. MySQL, Cassandra oder HANA zählen.
- Beim Datenzugriff besteht die Herausforderung darin, dass das die exponentiell anwachsende Menge an Daten so bereitgestellt wird, dass sie verarbeitet werden kann. Hier werden spezielle Methoden wie Streaming-Verfahren oder Abfragesprachen zusammengestellt. In diesem Zusammenhang sind Paradigmen bzw. Modelle wie MapReduce, CEP oder Lambda zu nennen, die durch Tools wie Hadoop oder Esper unterstützt werden.
- Im Bereich der analytischen Verarbeitung sind unterschiedliche Verfahren wie Predictive Analytics, Data Mining, maschinelles Lernen oder künstliche Intelligenz zu nennen. Bei diesen Verfahren kommen wiederum verschiedene Tools und Programmiersprachen wie SPSS, R, Matlab oder MLlib zum Einsatz.
- Ziel der Visualisierung ist es, Informationen und Wissen für Entscheidungsträger oder Analysten bereitzustellen und aufzubereiten.
- Bei der Datenintegration wird sichergestellt, dass verschiedene Datenquellen integriert und in einer Smart-Data-Umgebung zur Verfügung gestellt werden. Daten aus verschiedenen Quellen und mit unterschiedlichen Strukturen werden in einer gemeinsamen Struktur zusammengeführt, sodass die Verarbeitung der Informationen mit gleichen Formaten erfolgen kann. Dabei wird jedoch nicht der Inhalt der Daten, sondern lediglich ihre Struktur bzw. ihre Metadaten verändert.

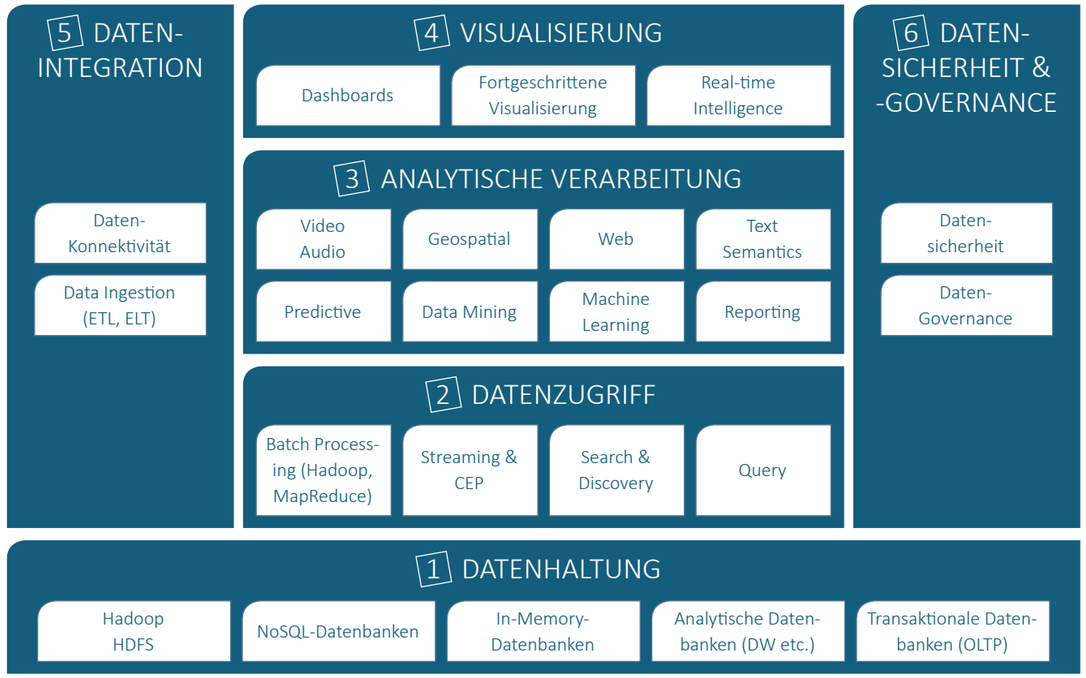


Abbildung 2: Taxonomie einer Smart-Data-Plattform und Beispiele für Technologien[10]

- Unter Datensicherheit und -Governance versteht man insbesondere die Einhaltung und Umsetzung von Regeln und Bestimmungen zum Datenschutz und zur IT-Sicherheit, die von besonderem Interesse für Smart-Data-Projekte sind.[11]

1.2 Standardisierungsgrad bestehender Plattformen

Wie auch bei anderen technologischen Entwicklungen sind für Smart Data Standardisierungs- und Vereinheitlichungsprozesse von besonderer Bedeutung, denn nur über festgelegte Standards lässt sich die Interoperabilität unterschiedlicher Lösungen und Akteure gewährleisten. Darüber hinaus garantieren Standards auch die Unabhängigkeit der Anbieter. Die oben beschriebenen Themenbereichen von Smart Data unterscheiden sich in Bezug auf die Standardisierungsgrade der Technologien derzeit jedoch noch stark.

So zählen die Datenhaltung und die analytische Verarbeitung zu den Bereichen, in denen sich mittlerweile eine vergleichsweise kleine Zahl von maßgeblichen Technologien und konkreten Anwendungen etabliert hat. Die reibungslose Zusammenarbeit zwischen den konkreten Lösungen funktioniert dabei allerdings bisher nur eingeschränkt. In den Bereichen des Datenzugriffs und der Datenintegration etablieren sich zwar bereits erste Verfahren, endgültig durchgesetzt hat sich aber noch keines. Beim Datenzugriff mangelt es derzeit außerdem noch an neuen, ausgereifteren Zugriffsmethoden für große Datenmengen. Da bei der Datenintegration grundsätzlich verschiedene Ebenen der Datenveredelung wie auch unterschiedlichste Verfahren und Lösungen zur Datenhaltung und zum Datenzugriff einbezogen werden müssen, ist auch hier die Herausbildung von Standards noch nicht weit fortgeschritten. In den Bereichen Visualisierung, Datensicherheit und Governance schließlich steht die Herausbildung entsprechender standardisierter Vorgehensweisen derzeit noch ganz am Anfang. Für alle Themenbereiche gilt jedoch, dass in entsprechenden Standardisierungsaktivitäten großes Potential liegt, auch wenn die Etablierung der Standards voraussichtlich noch einige Zeit in Anspruch nehmen wird.[12]

1.3 Standards: de jure oder de facto?

Wenn man also von Standardisierungsprozessen auf dem Gebiet „Smart Data“ spricht, wird schnell deutlich, dass sich dieses in zwei Teile gliedert: Im kleinen Rahmen sind bereits Standards in Form von vorherrschenden oder maßgeblichen Vorgehensweisen und konkreten Lösungen vorhanden und sollten auch zukünftig weiter vorangetrieben werden. Speziell im Bereich „Smart Data“ ist es für Deutschland wichtig, eine Vorreiterrolle bei der Standardisierung zu spielen und auch Methoden, die in Projekten entwickelt werden, einheitlich zur Verfügung zu stellen.

Im größeren Kontext der Smart-Data-Plattformen dagegen sind die Standardisierung- und die Interoperabilität der derzeitigen Smart-Data-Technologien noch unzureichend. Dies reduziert zum einen die Austauschbarkeit konkreter Lösungen – und damit die Handlungsspielräume einzelner Akteure – und schränkt zum anderen das Potential für eine nutzen- und gewinnbringende Integration unterschiedlicher Lösungen und Datenbestände über Akteursgrenzen hinweg bedeutend ein. Ein wichtiges Ziel im Zusammenhang mit der Standardisierung von Smart Data ist die Herstellung echter Austauschbarkeit und Interoperabilität. Hierzu muss sowohl die Unabhängigkeit der Anwender von einzelnen Anbietern gefördert als auch derzeit bestehende Kooperationshemmnisse abgebaut werden.

Die technische Komplexität und Vielschichtigkeit von Smart-Data-Plattformen könnte allerdings verhindern, dass Standards von einem Gremium definiert und (quasi-)regulatorisch etabliert werden. Wahrscheinlicher ist es daher, dass Standardisierungsprozesse auf dem Gebiet der Smart-Data-Plattformen durch die Schaffung von Fakten, also über die Technologien selbst, erfolgen.[13] Zusammen mit Standardisierungsinitiativen für einzelne Methoden, Architekturen, Vorgehensmodelle etc. ergibt sich hier jedoch eine gute Chance, das gesamte Feld von Smart Data über eine Kombination aus De-facto-Standards und klassischen Standards von Gremien ganzheitlich zu erfassen.

2 Big Data in Motion

Dr. Christoph Heinz, Software AG / Smart-Data-Projekte iTESA und sd-kama

Das Thema Big Data hat in den letzten Jahren für viel Aufmerksamkeit gesorgt. Mit Big Data ist die Hoffnung verbunden, aus der ungeheuren Masse von strukturierten und unstrukturierten Daten, die heutzutage in Unternehmen anfallen, die tatsächlich relevanten Informationen herausdestillieren zu können. Eine häufig anzutreffende Charakterisierung von Big Data basiert dabei auf den „vier Vs“: Volume, Variety, Velocity und Value. Dabei stellt die Velocity, also die hohe Geschwindigkeit, mit der kontinuierlich Daten eintreffen, eine besondere Herausforderung für entsprechende Big-Data-Plattformen dar. Eine zentrale Problemstellung vor diesem Hintergrund ist es, aus Datenquellen, aus denen permanent Massen neuer Daten „sprudeln“, sofort die relevanten Erkenntnisse zu gewinnen und zu verwerten. Sei es, dass eine günstige Kaufgelegenheit im Finanzmarkt genutzt, ein Betrugsversuch mit Kreditkarten erkannt oder der drohende Ausfall einer Produktionsmaschine verhindert werden soll.

Klassische Big-Data-Technologien speichern zumeist erst größere Datenmengen, um diese dann auf mehreren Rechnerknoten verteilt parallel zu analysieren, womit jedoch eine zeitliche Verzögerung einhergeht. Im Gegensatz dazu versucht man bei Streaming Analytics, Daten, wenn möglich, bereits bei ihrem Eintreffen zu analysieren und direkt zu reagieren. Dieser neue Ansatz wird sogar noch effektiver, wenn man ihn mit den klassischen Technologien kombiniert. Die Möglichkeit, ausgefeilte Analysen auf historischen Daten auszuführen, stellt quasi das Referenzwissen zur Verfügung, mit dem aktuelle Echtzeit-Analysen auf strömenden Daten mit historischen Erkenntnissen in Bezug gesetzt werden können. So kann man etwa im Strom aktuell laufender Kreditkarten-Transaktionen nach in der Vergangenheit beobachteten Betrugsmustern suchen.

2.1 Auf dem Weg von Rohdaten zu Dashboards

Wenn man nun die beiden Ansätze einer Echtzeit-Analyse und einer nachgelagerten Analyse von Big Data betrachtet, ergeben sich grundlegende Unterschiede in der internen Mechanik. Bei einer zeitlich versetzten Analyse wird oftmals im Batchbetrieb eine Vielzahl von Prozessen auf einem Cluster von Rechnern ausgeführt, um so die Analyse der Gesamtheit der Daten auf möglichst viele Schultern verteilen zu können. Bei einer Echtzeit-Analyse werden hingegen die neu eintreffenden Daten kontinuierlich analysiert, wobei jedoch nur ein jeweils aktueller Ausschnitt der Daten und nicht die Gesamtheit der bereits empfangenen Daten betrachtet wird. Das Marktforschungsunternehmen Forrester veröffentlichte 2014 eine Marktanalyse zu diesem unter dem Namen „Big Data Streaming Analytics“[14] bekannten Thema. Die Software AG, die in dieser Marktstudie als Leader genannt wird, kombiniert zur Lösung dieses Problems mehrere Technologien im Rahmen ihrer Streaming-Analytics-Plattform[15] (siehe Abbildung 3).

Auf der untersten Ebene stehen die unterschiedlichen Datenquellen zur Verfügung, zu denen insbesondere Datenströme verschiedenster Natur zählen. Um Letztere effizient zu übertragen, ist eine performante Messaging-Infrastruktur erforderlich. Diese muss eine Vielzahl von Technologien unterstützen, seien es Protokolle wie HTML5 Websockets, MQTT, Plattformen wie mobile Clients auf iOS/Android/Windows-Phone-Basis oder Programmiersprachen wie Python, C# oder etwa Java.

Ergänzt wird diese Messaging-Infrastruktur durch eine Speichertechnologie, welche die Daten hochverfügbar im Hauptspeicher vorhält, sodass ein schneller Zugriff darauf möglich ist. Im fordernden Smart-Data-Umfeld muss dieser Datenspeicher auch in der Lage sein, mit dem Volumen der Daten zu skalieren, indem neue Rechnerknoten mit zusätzlicher Hauptspeicher-Kapazität integriert werden können. Neben dem aktuellen Ausschnitt aus den Daten können in diesem Datenspeicher zusätzlich Daten gehalten werden, die für die Anreicherung der Live-Daten mit Kontextinformationen erforderlich sind, wie etwa Kundenstammdaten.

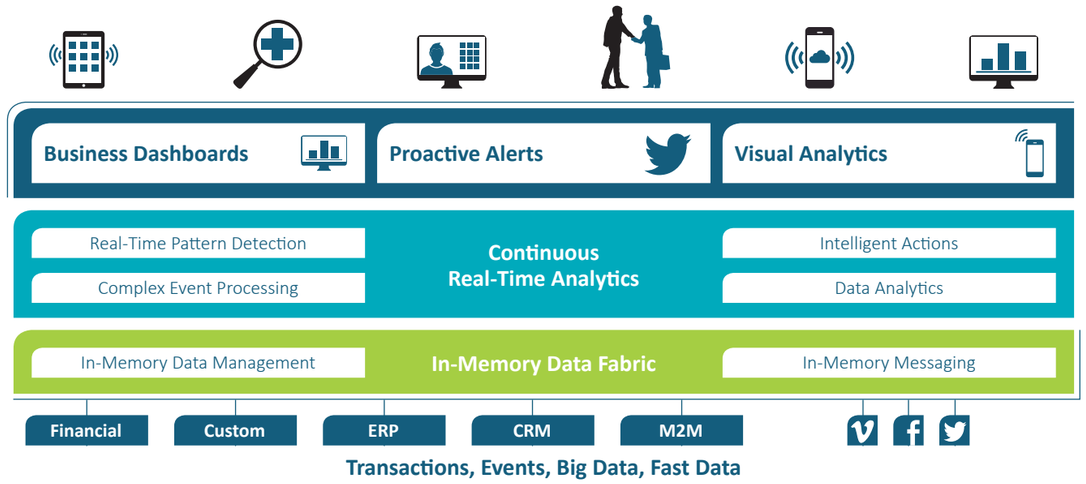


Abbildung 3: die Streaming-Analytics-Plattform der Software AG

Um die angebundenen Datenströme in Echtzeit zu analysieren, empfiehlt sich der Einsatz von Complex Event Processing (CEP). Ein CEP-Server empfängt über die Messaging-Komponente kontinuierlich neue Daten und analysiert diese direkt bei deren Eintreffen. Die Analyselogik wird dabei vorab definiert und entsprechend für neue Daten ausgeführt. Sie beinhaltet üblicherweise Operationen wie Aggregation, Filtern und Join und ist bezüglich eines Zeitfensters definiert, wie beispielsweise „Berechne die Summe aller Transaktionen einer Kreditkarte in den letzten 10 Minuten“. Auch hier wird aus Gründen der Performanz der Hauptspeicher für die Berechnungen herangezogen. Mittels einer solchen CEP-Komponente sollen die wesentlichen Informationen aus der Flut eintreffender Daten extrahiert und somit das Datenvolumen massiv reduziert werden.

Die finale Komponente innerhalb eines solchen Streaming Analytics Stacks widmet sich der visuellen Aufbereitung der Ergebnisse. Eine solche Visualisierungskomponente ermöglicht es, kontinuierlich einströmende Analyse-Ergebnisse in ansprechenden, übersichtlichen Dashboards darzustellen. Dabei ist die Darstellung nicht nur auf diese Ergebnisse beschränkt: Auch Daten aus anderen Quellen können zusätzlich integriert und miteinander kombiniert werden, bevor sie visualisiert werden. Damit hilft die Visualisierungskomponente auch bei der Integration der zu visualisierenden Ergebnisse.

2.2 Wohin geht die Reise?

Mit Streaming Analytics ergibt sich eine interessante Weiterentwicklung im Big-Data-Kontext. Indem Daten direkt und ohne Verzögerung analysiert werden, können Unternehmen sowohl aussichtsreiche als auch kritische Situationen frühzeitig erkennen und entsprechend zeitnah darauf reagieren. Dieser Ansatz bedeutet jedoch keinesfalls das Ende des klassischen Big-Data-Ansatzes. Vielmehr birgt gerade die Kombination beider Ansätze, also „Big Data in motion“ plus „Big Data at rest“, eine Menge Potential. Kurzfrist-Analysen auf Datenströmen dienen dazu, aktuelle Situationen bezüglich eines überschaubaren Zeithorizonts zu analysieren, während Langfrist-Analysen helfen, belastbare Muster in der Gesamtheit der Daten zu erkennen und als Kontextwissen bereitzustellen. Da die zu erwartende Datenmenge in den nächsten Jahren noch ansteigen wird, ist es besonders interessant, die Evolution der vorgestellten Technologien sowie der zu erwartenden neuen Technologien zu beobachten. Zu Letzteren sind etwa der vermehrte Einsatz von SSD-Festplatten und neuerer Speichertechnologien oder die Entwicklung skalierbarer Data-Mining-Verfahren zu zählen.

3 Smart Data Technology Stack

Henrik Oppermann, USU Software AG / Smart-Data-Projekte SAKE und Smart Regio

Zunehmend mehr Sensoren in Maschinen und Anlagen sowie die weite Verbreitung von Remote-Plattformen verändern die Basis, auf der Unternehmen ihre Services anbieten. Im Zuge von Industrie 4.0 und des Internets of Things (IoT) entdecken immer mehr Unternehmen das Potential, auf Basis dieser Daten neue Service-Produkte anbieten zu können. Viele Unternehmen generieren aber mehr Daten, als sie speichern, korrelieren und analysieren können. Da die Daten schnell und in großen Mengen erzeugt werden, können Unternehmen häufig nicht schnell genug reagieren, um die Informationen in Smart Services zu verwandeln. Der hier vorgestellte Technologieansatz auf einer flexibel skalierbaren Plattform kann riesige Datenmengen (Volume) sehr schnell (Velocity) in einer einheitlichen Architektur verarbeiten.

3.1 Anwendungsbeispiele aus der Druckindustrie

Durch die andauernde Konsolidierung in der Druckindustrie verlagert sich der Kundenstamm der Heidelberger Druckmaschinen AG (HDM) beispielsweise mehr und mehr hin zu Druckereien, die im Dreischichtbetrieb (24/7) mit engen Lieferfristen hochautomatisiert produzieren. Das führt dazu, dass HDM auf einen veränderten Servicebedarf reagieren muss. Hierbei geht es im Wesentlichen um das frühzeitige Erkennen von Unregelmäßigkeiten anhand von online verfügbaren Maschinendaten. Ziel ist es, durch geplante Servicemaßnahmen aufkommende Störungen zu erkennen und zu eliminieren, bevor sie den Produktionsprozess stören. Damit wird die höchste Verfügbarkeit der Maschine dank intelligent geplanter, proaktiver Serviceeinsätze erreicht. Der Qualität der Vorhersagen kommt dabei eine besondere Bedeutung zu, da Fehlalarme mit unnötigem Maschinenstillstand und Produktionsausfall extrem teuer werden können.

Zur Auslegung des Systems und damit zur Auswahl der Komponenten ist das Projekt SAKE von folgenden Anforderungen ausgegangen:

- Auswertung von Logdateien für über 2.000 Druckmaschinen
- jeden Tag 4 Mio. Datensätze pro Maschine (ca. 300 MB): Neue Logdateien sollen sofort prozessiert werden, zur schnellen Weiterleitung von Warnmeldungen sowie zur Aufbereitung von Daten für Dashboards oder Drittsysteme
- Daten sollen bis zu drei Jahre für weitere Auswertungen vorgehalten werden (ca. 600 TB)
- Symptomscanner muss online auch Auswertungen auf historischen Daten ausführen können
- Machine-Learning- und Data-Mining-Ansätze werden auf sehr großen Datenmengen angewendet (ca. 600 Mrd. Datensätze) und dienen zum Aufdecken von Zusammenhängen bei unbekannten Fehlern
- Ingenieure untersuchen, ob ein Kausalzusammenhang besteht; wenn ja, wird dies als Regel formuliert und mittels CEP kontinuierlich überprüft: Qualitätsgesicherte Auswertung unbekannter Zusammenhänge auf Echtzeitdaten

Damit stellen sich neben vielen anderen funktionalen Anforderungen für die Plattform vor allem die beiden Big-Data-Dimensionen Volume und Velocity als maßgebliche Designparameter heraus. In der Literatur wird als Referenz gerne die Lambda-Architektur[16] genannt, die für die beiden Dimensionen separate Datenverarbeitungsschichten vorsieht: den Batch Layer für große historische Datenmengen und den Speed Layer für die Verarbeitung von Realtime-Daten. Dieser Ansatz der separaten Verarbeitung widerspricht aber der Anforderung, historische Daten schon in der Verarbeitung von Logdatenströmen (Realtime) zu ermöglichen. Als wesentliche nichtfunktionale Anforderungen waren für die Auswahl der Komponenten die Langlebigkeit und die Unterstützung durch eine Community wichtig, um die Investitionssicherheit zu erhöhen und die Entwicklungskosten zu senken.

Entlang der oben dargestellten Taxonomie will das Smart-Data-Projekt SAKE die Auswahl der wesentlichen Komponenten begründen und die Gesamtarchitektur zusammenfassen.

Datenbank	Vorteile	Nachteile
Apache Cassandra	- Linear skalierbar - Hohe Schreibrate - Große Community	- Datenmodellierung muss Abfragewege berücksichtigen - Eingeschränkte Abfragemöglichkeiten
Apache HBase	- In vielen Punkten vergleichbar mit Cassandra	- Setzt auf HDFS auf - Master/Slave-Architektur - Keine Query Language
Apache Hive	- SQL-ähnliche Abfragemöglichkeiten	- Komplexität bei Datenaktualisierungen - Batchorientiert - Nur eingeschränkt geeignet für interaktive Abfragen
MongoDB	- Flexible Abfragemöglichkeiten - Große Community	- Geringe Schreibrate - Aufwändiges Clustersetup/Anpassung - Dokumentorientiert, schemafrei
MySQL/Galera Cluster	- SQL	- Geringe Schreibrate - Eingeschränkte Skalierbarkeit

Abbildung 4: die Auswahl der Datenbank

3.2 Anforderungen an Datenhaltung, -zugriff und -verarbeitung

Als wesentliche Kriterien in Bezug auf die Datenhaltung ergaben sich a) eine lineare Skalierbarkeit in den zweistelligen Terabyte-Bereich (und damit Abbildung eines DB-Clusters), b) eine hohe Schreibrate, c) Unterstützung von Replikation für die Ausfallsicherheit sowie d) Produktionsreife und Investitionssicherheit.

Die Tabelle (Abbildung 4) fasst unsere Betrachtung von Systemen und die Auswahl zusammen.

Derzeit ist der größte bekannte Apache-Cassandra-Cluster bei Apple mit mehr als 75.000 Knoten und über 10 PB Datenvolumen installiert und demonstriert deutlich das Thema Skalierung. Weitere bekannte Nutzer sind Google und Netflix. Letzteres Unternehmen verfügt über mehr als 2.500 Knoten und verzeichnet über 1 Billion Zugriffe am Tag.[17]

Ergänzend wird für die Archivierung der historischen Daten auf das verteilte Dateisystem HDFS zurückgegriffen und Logdateien zusätzlich über das Parquet-Dateiformat abgelegt. Datenspeicherung in Dateien bietet sich an für Daten, die in sehr großer Menge anfallen, nach der Vereinnahmung nicht mehr umstrukturiert oder aktualisiert werden müssen und nur sequentiell gelesen werden. Das Parquet-Format zeichnet sich darüber hinaus durch hohe Kompressionsraten, parallele Verarbeitbarkeit und eine sehr hohe Schreib- und Lesegeschwindigkeit aus.

Wesentliche Kriterien zur Technologieauswahl bei Datenzugriff und -verarbeitung waren a) die Unterstützung von Datenströmen und historischen Daten, b) die interaktive Datenanalyse für große Datenmengen (historische Daten für Monate), c) hoch performante Unterstützung für iterative Algorithmen (Machine Learning), d) eine große Community (Ökosystem) und Produktionsreife.

Folgende Tabelle fasst die Betrachtung und Auswahl zusammen:

Datenbank	Vorteile	Nachteile
Apache Spark	- Hohe Performanz durch In-Memory-Computing - Große Community - Gute Skalierbarkeit - Einbettung in Java - Großes Ökosystem	- Datenströme nicht für harte Echtzeitanwendungen geeignet
Apache Hadoop + Apache Storm	- Sehr großes Ökosystem - Große Community	- Map-Reduce-Jobs nicht für interaktive Zwecke geeignet - Für kleinere Cluster hohe Anzahl von Verwaltungsdiensten - Lambda-Architektur: höhere Komplexität
Apache Flink (früher: Stratosphere)	- Hohe Performanz durch In-Memory-Computing - Gute Skalierbarkeit - Richtige Datenströme - Iterative Transformationen auf Collections	- Community/Ökosystem noch im Aufbau

Abbildung 5: die Auswahl der Verarbeitungsplattform

Dabei garantiert das umfangreiche Ökosystem von Spark[18] und seine weitreichende Unterstützung[19] (u. a. durch IBM[20] und Yahoo[21]) die Sicherheit der Investitionen und ermöglicht es gleichzeitig, die Entwicklungskosten zu senken. So kann durch die in Spark integrierte Machine-Learning-Bibliothek MLlib[22] auf eine Vielzahl von Algorithmen zugegriffen werden (siehe auch den nächsten Abschnitt).

Bei der analytischen Verarbeitung kommen neben der in Spark integrierbaren MLlib auch Eigenentwicklungen im Bereich der Algorithmen und Complex Event Processing (CEP) zum Einsatz. Denn hier sollen die Auswertungen bei Bedarf durch Anpassung der Konfiguration leicht horizontal und vertikal verteilt werden können und mit dem gleichen Werkzeug für Datenströme und Stapelverarbeitung (historische Daten) formulierbar sein. Zudem müssen komplexere Auswertungsschritte in Java-Klassen formuliert und ausgelagert werden können (höhere Ausdrucksmächtigkeit als z. B. SQL) und eine deklarative Beschreibung des Event Processing Networks möglich sein. Für diese Anforderungen sind aber gängige CEP-Engines nicht ausgelegt: Gerade der Einbezug historischer Daten ist bei diesen meist nicht vorgesehen. Wir haben daher eine eigene CEP-Engine und -Entwicklungsumgebung auf Basis von Eclipse[23] realisiert. Ebenso existiert eine Management-Konsole für die Verwaltung der verschiedenen Auswertungen und das Deployment derselbigen. Eine eigens entwickelte Sprache (Domain Specific Language, DSL) ermöglicht hierbei eine leichtere Formulierung von Auswertungsregeln.

Für die interaktive Diagnose wurde ein Notebook-Konzept auf Basis von Apache Zeppelin[24] realisiert, mit dem Maschinen-Analysten in verschiedenen Skriptsprachen Hypothesen und Auswertungen formulieren, auf Echtdaten ausprobieren und deren Auswirkungen überprüfen können. Es stellt ein wichtiges Instrument zur Formulierung von Überwachungsregeln dar.

3.3 Die Gesamtarchitektur und Fazit

In der folgenden Abbildung ist der Big Data Stack der USU Software AG dargestellt. Als Besonderheit haben wir das im Hadoop-Umfeld gängige Tool zur Ressourcenverwaltung Apache YARN durch Apache Mesos ersetzt. Grund war das genauere Ressourcen-Management für unterschiedliche Frameworks, z. B. das Zusammenspiel von Spark, Cassandra etc. Als Nachteil anzusehen ist, dass das Cluster-Management wenig intuitiv ist. Das System ist derzeit On-Premise und als Private Cloud betreibbar. Im Rahmen der Smart-Data-Projekte SAKE und SmartRegio wird u. a. eine Public-Cloud-Version entwickelt.

Ein wesentlicher Nutzen neben der präzisen Vorhersage von Fehlern in Anlagen und Maschinen (Predictive Maintenance) ist der Erkenntnisgewinn durch die tiefe Analyse und Vergleiche auf Basis von Massendaten und Clusterbildung. Durch Peer-Group-Vergleiche lassen sich so etwa Optimierungspotentiale in Maschinen gleichen Typs identifizieren. Zudem können die gewonnenen Erkenntnisse für eine Verbesserung der Kundenservices genutzt werden, etwa durch die Optimierung des Maschinenbetriebs oder den Einbau besserer oder günstigerer Ersatz- und Bauteile.

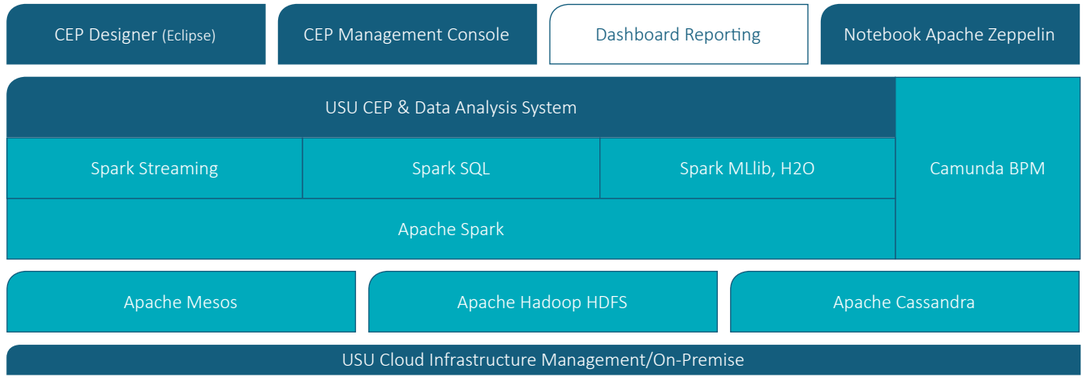


Abbildung 6: Systemarchitektur der USU-Industrial-Big-Data-Lösung

4 Einführung in Apache Flink

Prof. Dr. Volker Markl, Christoph Boden, TU Berlin / Berlin Big Data Center (BBDC)

Apache Flink[25] ist eine skalierbare und deklarative Open-Source-Big-Data-Analytics-Plattform für Batch- und Streamdatenverarbeitung, die u. a. Konzepte massiv-paralleler Datenverarbeitungssysteme wie etwa Apache Hadoop und paralleler Datenbanksysteme vereint. Als vergleichsweise neues Big-Data-Analytics-Framework in der Apache Software Foundation ermöglicht Apache Flink Data Scientists, sich auf ihre ursprünglichen Analyseprobleme zu konzentrieren, anstatt sich mit den Problemen der Programmierung von skalierbaren, verteilten Systemen auseinanderzusetzen. Während Big-Data-Systeme der „ersten Generation“, wie das auf dem MapReduce-Paradigma basierende System Apache Hadoop, Aggregationen und einfache Berechnungen auf sehr große Datenmengen skalieren können, bietet Apache Flink die Möglichkeit, komplexe Analysen sowohl für „Data at Rest“- als auch „Data in Motion“-Anwendungsfälle zu realisieren und somit Smart-Data-Anwendungen zu schaffen.

4.1 Apache Flink: eine Streaming Dataflow Engine

Im Kern ist Apache Flink eine Streaming-Engine, die sowohl Batch- als auch kontinuierliche Streamverarbeitung ermöglicht. Dies grenzt das System von anderen Big-Data-Analytics-Lösungen wie etwa Apache Spark ab, welches Streaming lediglich durch die Verarbeitung von Microbatches auf die darunter liegende Batch-Engine abbildet. Auf der Anwenderseite bietet Flink eine durch Datenflüsse inspirierte DataSet-API in Java und Scala sowohl für Batch- als auch für Realtime-Streaminganwendungsfälle. Darüber hinaus bietet Flink ein knotenzentrisches Programmiermodell für die Analyse von Graphen (Gelly), eine Python-API sowie eine Table-API mit einer SQL-ähnlichen Anfragesprache für die einfache Transformation von Daten. Zusätzlich bietet FlinkML, eine Bibliothek für maschinelles Lernen von Apache Flink, bereits Implementierungen von populären maschinellen Lernverfahren wie etwa Support Vector Machines (SVMs) und Alternating Least Squares (ALS) Recommender sowie ein Optimization Framework für optimierungsbasierte Lernverfahren.

4.2 Prägnante APIs für typsicheren und wartbaren Code

Musste man selbst für einfachste Analysen in Apache Hadoop noch überproportional viel „Glue code“ schreiben, so sind selbst komplexere Algorithmen in

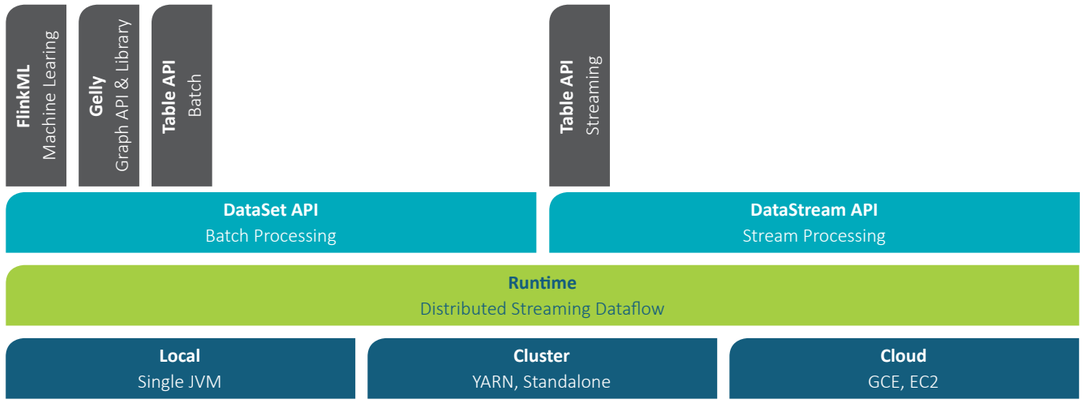


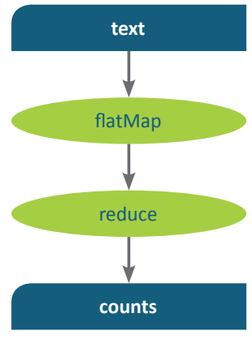
Abbildung 7: Apache Flink – Eine Open-Source-Plattform für skalierbares Batch und Stream Processing


```
ExecutionEnvironment env =
    ExecutionEnvironment.getExecutionEnvironment();

DataSet<String> text = readTextFile (input);

DataSet<Tuple2<String, Integer>> counts= text
    .map (1 -> 1.split („\\W+“))
    .flatMap ((String[] tokens,
                Collector<Tuple2<String, Integer>> out) -> {
      Arrays.stream(tokens)
      .filter(t -> t.length() > 0)
      .forEach(t -> out.collect(new Tuple2<>(t, 1)));
    })
    .groupBy(0)
    .sum(1);

env.execute(„Word Count Example“);
```




```
val env = ExecutionEnvironment
    .getExecutionEnvironment

val input = env.readTextFile(textInput)

val counts = text
    .flatMap { line => line.split(„\\W+“) }
    .filter { term => term.nonEmpty }
    .map { term => (term, 1) }
    .groupBy(0)
    . sum(1)

env.execute()
```

Abbildung 8: das WordCount-Beispiel, implementiert in der Java- (links) und Scala-DataSet-API (rechts) von Apache Flink

Apache Flinks DataSet-API sehr prägnant implementiert. Abbildung 8 illustriert dies anhand des wohlbekannten WordCount-Beispiels mit Quellcode in Flinks Java- und Scala-DataSet-API für Batch-Anwendungen. Darüber hinaus können auch Kollektionen von benutzerdefinierten Java-Objekten („POJOS – Plain Old Java Objects“) verarbeitet werden.

Die Streaming-API ist analog gestaltet, operiert jedoch auf einem DataStream-Objekt. Streamingoperatoren können Zustände in Form von Java- bzw. Scala-Objekten speichern. Darüber hinaus bietet Flink mit „Windowed Streams“ die Möglichkeit, einen endlichen Stream aus einem unendlichen Stream mittels (Sliding) Windows über den Stream zu erzeugen. Hierbei handelt es sich um die spezielle Form eines Zustands, welcher durch das System gemanaged wird. Apache Flink stellt einen Fault-Tolerance-Mechanismus bereit, welcher diesen Zustand im Fehlerfall wiederherstellen kann.

4.3 Robuste Ausführungsumgebung mit automatischer Optimierung

Neben den Programmierschnittstellen bietet Flink eine sehr robuste Ausführungsumgebung mit nativer Unterstützung von iterativen Algorithmen. Eine eigene Speicherverwaltung innerhalb der JVM ermöglicht es, Anwendungen jenseits des Hauptspeichers zu skalieren und den Garbage-Collection-Overhead von Java-Programmen zu reduzieren, sowie einen kostenbasierten Optimierer, welcher Batch-Programme automatisch optimiert. Hierbei gilt es z. B., teure Operationen (wie Shuffle und Sortieren) nach Möglichkeit zu vermeiden sowie Zwischenergebnisse zu cachen. Im Gegensatz zu Apache Spark, einem weiteren Open-Source-System für Big Data Analytics, wird diese Optimierung in allen APIs ausgeführt. Abbildung 3 fasst die wesentlichen Unterschiede zwischen Apache Hadoop, Apache Spark und Apache Flink nochmals zusammen.

	hadoop MapReduce	Spark	Flink
API	MapReduce auf Key/ Value-Paaren	Transformationen auf Key/ Value Collections	Iterative Transformationen auf Collections
Paradigma	MapReduce	RDD	Zyklische Datenflüsse
Optimierung	Keine	Optimierung von SQL-Queries	Optimierung in allen Data-Set-APIs
Ausführung	Batchsortierung	Batch mit Memory Pinning	Stream mit Out-of-Core-Algorithmen

Abbildung 9: Vergleich der verschiedenen Big-Data-Analytics-Frameworks Apache Hadoop, Apache Spark und Apache Flink

4.4 Vielfältige Integrationsmöglichkeiten und wachsende Community

Apache Flink integriert sich in eine Fülle von weiteren Open-Source-Tools. Flink selbst kann Daten aus Quellen wie HDFS (unterstützt alle Hadoop-Inputformate einschließlich Apache Avro, HCatalog/Hive), HBase, Tachyon und anderen konsumieren. Flink bietet Kompatibilitäts-APIs für Hadoop MapReduce, Apache Storm und weitere Projekte wie Apache Zeppelin, Apache SAMOA, Apache MRQL. Des Weiteren führen die Projekte Apache Mahout, Google Cloud Dataflow und Cascading Unterstützung für Flink als Backend ein.

Apache Flink ist ursprünglich unter dem Namen Stratosphere[26] aus der DFG-Forschungsgruppe „Stratosphere: Information Management in the Cloud“, einer wissenschaftlichen Kollaboration der Technischen Universität Berlin, der Humboldt-Universität zu Berlin und dem Hasso-Plattner-Institut in Potsdam, hervorgegangen. Die weitere Entwicklung wurde u. a. gefördert durch die EIT ICT Labs, die DFG, IBM, Oracle, Hewlett-Packard, Deutsche Telekom, BMBF, BMWi sowie durch die Europäische Kommission (im Rahmen des FP7-Förderprogramms). Heute hat Apache Flink eine schnell wachsende Entwickler-Community mit über 80 Mitgliedern aus Ungarn, Schweden, Frankreich und Deutschland. Zudem weckte es auch reges Interesse in der Industrie: Auf der ersten zweitägigen Flink-Entwicklerkonferenz „Flink Forward“ im Oktober 2015 in Berlin haben diverse Unternehmen wie etwa ResearchGate, Bouygues Telecom, Amadeus, Telefonica, Capital One, Ericsson oder die Otto Group von ihren erfolgreichen Projekterfahrungen mit Apache Flink berichtet.

4.5 Basistechnologie im Berlin Big Data Center

Apache Flink stellt auch die Basistechnologie des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Kompetenzzentrums „Berlin Big Data Center“ (BBDC)[27] dar. Das BBDC verfolgt hierbei das Ziel, die bisher getrennt nebeneinander existierenden Forschungsgebiete „Maschinelles Lernen“ und „Datenmanagement“ zu verschmelzen. Es gilt, ein deklaratives, hochskalierbares Open-Source-System zu entwickeln, welches in der Lage ist, die Spezifikation, automatische Optimierung, Parallelisierung und Hardware-Adaption sowie die fehlertolerante, effiziente Ausführung von Datenanalyseproblemen mittels verschiedener Methoden des maschinellen Lernens, der linearen Algebra, der Statistik, der Wahrscheinlichkeitstheorie, der Computerlinguistik sowie der Signalverarbeitung durchzuführen. Die Technologie wird exemplarisch anhand der Anwendungsbereiche „Materialforschung“, „Medizin“ sowie „Informationsmarktplätze“ validiert. Des Weiteren wird Apache Flink als Querschnittstechnologie unter anderem in den durch das BMWi geförderten Smart-Data-Projekten SmartDataWeb und Smart4Mobility sowie im durch das BMBF geförderte Big-Data-Projekt GeoMultiSense eingesetzt.

Weitere Informationen zum Projekt inklusive ausführlicher Dokumentation aller Programmierschnittstellen sowie Beispielen sind unter https://flink.apache.org/ zu finden.

5 SDIL Smart Data Testbed

Prof. Dr. Michael Beigl, Prof. Dr. Bernhard Neumair, Dr. Till Riedel, Nico Schlitter, Karlsruher Institut für Technologie (KIT) / Smart Data Innovation Lab (SDIL)

Das Smart Data Innovation Lab (SDIL) ist ein seit 2014 bestehendes bundesweites Labor für Smart-Data-Forschung auf Basis von In-Memory-Big-Data-Analysetechnologien zur Analyse komplexer Datenbestände mit intelligenten Verfahren. Das SDIL ist eine Austausch- und Betriebsplattform, um die Kooperation zwischen Wirtschaft, öffentlicher Hand und Forschung im Bereich Big-Data- und Smart-Data-Technologien zu beschleunigen. Unternehmen und Forschungseinrichtungen haben sich dazu als Partner in SDIL-Communities (Industrie 4.0, Energie, Smart Cities und Medizin) organisiert und einen Rahmen zwischen den beteiligten Partnern vereinbart, um eine intensive Kooperation zwischen Wirtschaft und Wissenschaft im Bereich Smart Data zu erreichen.

Aus den SDIL-Communities entstehen Projektideen, die nach einer typischerweise nur zwei Wochen dauernden Antragsphase über den SDIL-Lenkungsausschuss bewilligt werden. Bewilligten Forschungsprojekten im Bereich der Big-Data-Analyse stellt das SDIL dann eine erstklassige Infrastruktur zur In-Memory-Big-Data-Analyse bereit. Am Karlsruher Institut für Technologie (KIT), dem Träger des SDIL, wird eine leistungsstarke IT-Infrastruktur von der Hardware- und Software-Industrie bereitgestellt, mit der die Wissenschaft gemeinsam mit der Wirtschaft hoch performante Arbeit mit Big Data durchführen kann.

5.1 Die technische Basis des SDIL und deren Nutzung

Technisch besteht das SDIL aus mehreren von SAP und IBM bereitgestellten NUMA-Knoten mit insgesamt über 200 Prozessorkernen, welche mit derzeit zwölf Terabyte Hauptspeicherkapazität ausgestattet sind und so schnelle Analysezyklen und interaktives Arbeiten mit Daten ermöglichen. Hierzu werden neben projektspezifischen Teilen drei Kernsoftwareplattformen angeboten: Terracotta BigMemory Max (Software AG), SAP HANA (SAP), Watson Foundations/BigInsights (IBM). Diese sind auf die Plattform optimiert und es werden Lizenzen für mit der Plattform kompatible Analysewerkzeuge bereitgestellt.

Um hierauf mit geringen Einstiegshürden eine realitätsnahe Forschung zu ermöglichen, liefern Industrie-

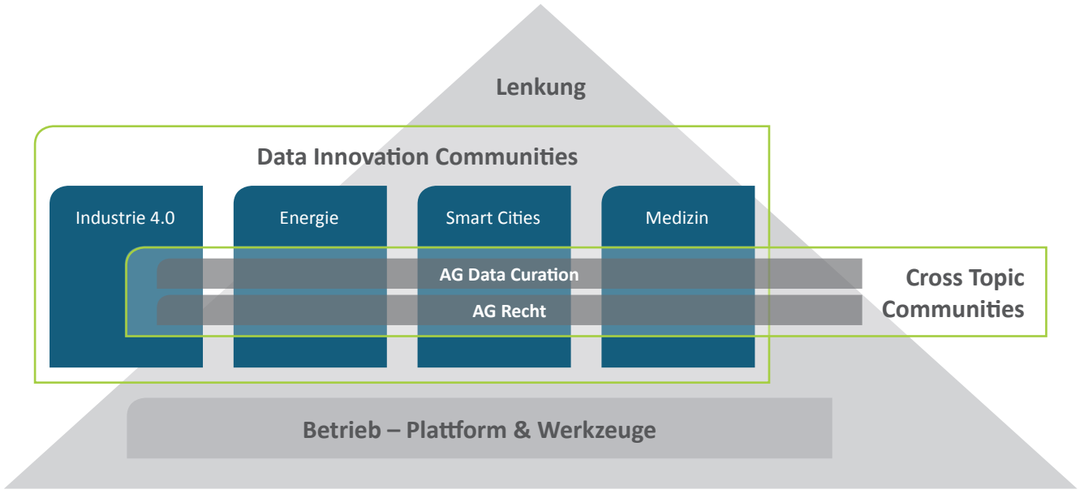


Abbildung 10: die Struktur des Smart Data Innovation Labs

partner Datenquellen aus der Praxis. Die Forschung konzentriert sich dabei auf die strategisch wichtigen Felder Industrie 4.0, Energie, Smart Cities und Medizin. Diese Datenquellen werden ergänzt durch Daten der öffentlichen Hand sowie im Internet frei verfügbare Datenquellen. Neue Datenquellen und Forschungsfragen werden hierzu in themenspezifischen „Data Innovation Communities“ (DICs) diskutiert, die jeweils von einem industriellen und einem akademischen Partner geleitet werden. Die regelmäßigen Workshops dienen außerdem dazu, neue Partner zu gewinnen und fachspezifische Diskussionen anzuregen.

Aus den DICs wurden bereits mehre Pilotprojekte innerhalb des SDIL angestoßen und durch das Strategieboard genehmigt. Dies schließt beispielsweise das vom BMWi im Rahmen des Technologieprogramms „Smart Data – Innovationen aus Daten“ geförderte Projekt sd-kama ein. sd-kama verfolgt das Ziel, eine Informationsplattform aufzubauen, die ein effizientes und effektives Katastrophenmanagement ermöglicht. Das Herzstück von sd-kama bildet eine Smart-Data-Plattform, deren sicheren technischen Betrieb das Smart Data Innovation Lab innerhalb des Projektes übernimmt. Durch eine geteilte Infrastruktur entstehen automatisch Synergien zwischen geförderten und industriellen Projekten.

Im Smart-City-Bereich sind dies beispielsweise Projekte zum intelligenten Ressourcenmanagement (Siemens). Die Bandbreite von laufenden, abgeschlossenen und geplanten Projekten ist bereits jetzt enorm. So stehen im Energiebereich Projekte zur Modellierung dezentraler Märkte (SAP, KIT) und zu prädiktiven Analytics für das Energiemanagement (IBM, EnBW) im Fokus. Im Health-Bereich stehen Smart-Data-Projekte zu Brain Analytics (FZ Jülich, IBM), Makulardegeneration (LMU, IBM) und Wirbelsäulenverletzungen (Uniklinik Düsseldorf) im Vordergrund. Bezüglich der Datenquellen sind die laufenden und geplanten Projekte aus der Industrie-4.0-Community am breitesten angelegt. Die Projekte umfassen neben der prädiktiven Wartung (TRUMPF, SAP, KIT), Log-Analysen (ABB, KIT, IBM) nutzerprofiladaptive HMIs, optimales Nesting und smartes Schweißen (alle TRUMPF, KIT, IBM), sowie das Maschinelle Lernen in der Lackieranlage (Daimler, KIT, IBM). Parallel dazu bietet das vom baden-württembergischen Forschungsministerium (MWK) geförderte SDIL-Projekt „Smart Data Solution Center Baden-Württemberg“ auch kleinen und mittleren Unternehmen im Land die Möglichkeit, ihre gesammelten Daten kostenfrei auf dem SDIL voranalysieren zu lassen und so Smart-Data-Potentiale zu erkennen.

Ein Katalog von Datenquellen, die entweder frei oder nur innerhalb von SDIL-Projekten zur Verfügung stehen, ist zurzeit in Arbeit. Der Datenlieferant hat hierbei die Möglichkeit, zwischen individuell zugeschnittenen Modellen zu wählen. Auch von der Analyse geschlossener Datenquellen sollen nach einem Fair-Share-Prinzip industrielle und akademische Partner profitieren. So soll das Smart Data Innovation Lab in der nächsten Ausbaustufe ein Open-Source-Repositorium bieten, in dem Code-Artefakte abgelegt werden, die im Rahmen der Forschung entstehen und dann von anderen Projekten wiederverwendet werden können.

5.2 Der Data Innovation Cycle

Das SDIL wird dadurch vollständig von den beteiligten Partnern getragen. Daneben fördert das BMBF im Rahmen des neuen Projekts SDI-X die Weiterentwicklung der Werkzeuge, Betriebskonzepte und Prozesse rund um Smart Data innerhalb des SDIL. Das SDIL basiert auf einem Forschungs- und Innovationszyklus für Big- und Smart-Data-Analysen (genannt „Data Innovation Cycle“), dargestellt in Abbildung 11. Dieser bietet die Möglichkeit, die Geschwindigkeit von Analyseaktivitäten zu bewerten. Auf diese Weise sollen die Effekte neuer Prozesse, Werkzeuge und Betriebskonzepte messbar gemacht werden. Dieser Zyklus soll durch die in SDI-X entworfenen Konzepte und Mechanismen langfristig gefördert und aufrechterhalten werden.

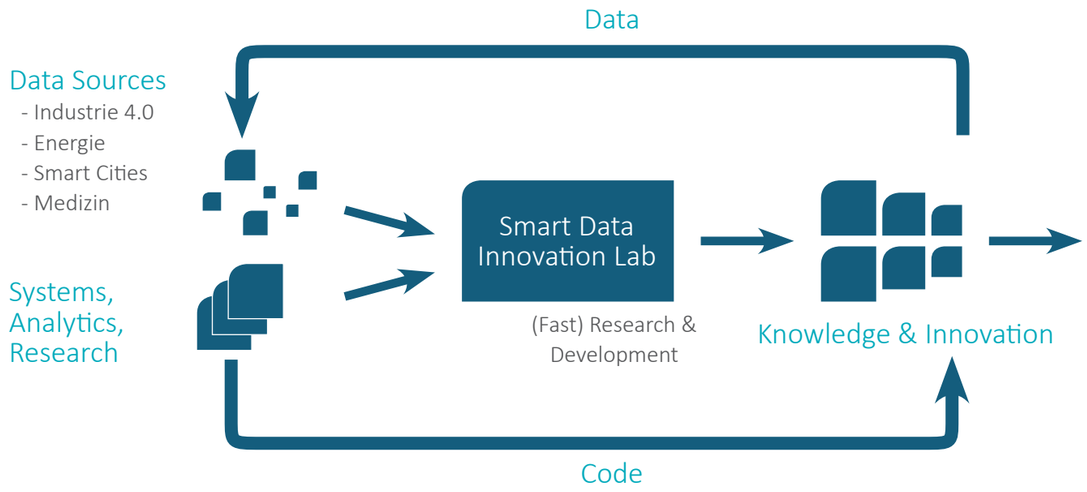


Abbildung 11: Forschungs- und Innovationszyklus im Smart Data Innovation Lab

- Best-Practice-Prozesse für Smart Data: Um konzentriert und zeitnah an komplexen Fragestellungen von Smart Data arbeiten zu können, ist eine Untermauerung von Analysetätigkeiten und -projekten durch Best-Practice-Prozesse notwendig.
- Best-Practice-Werkzeuge für Smart Data: Grundlage für Innovationen auf Basis von Big Data ist der flexible Einsatz geeigneter Verfahren und Algorithmen zur Datenanalyse auf verschiedenen Analytics-Plattformen (spezialisierte Hard- und Softwaresysteme). Zu diesem Zweck werden Projektpartner in der Anwendung der Werkzeuge geschult und bei der Spezialisierung und Erweiterung der Werkzeuge, besonders im Bereich „Maschinelles Lernen“ beraten.
- Best-Practice-Betriebskonzepte für Smart Data: Wichtige Aspekte dieses Vorhabens sind die Betriebskonzepte zur nachhaltigen Bereitstellung einer innovativen technischen Infrastruktur. Diese sind Grundlage für zeitnahe Big-Data-Innovationen, insbesondere in mittleren bis großen Kooperationsforschungsprojekten. Weiterhin wird SDI-X skalierbare und nachhaltige Lösungen entwerfen, um die Privatheit von Daten über den gesamten Lebenszyklus eines Analyseprojektes innerhalb des SDIL zu gewährleisten. Entsprechende Anonymisierungskonzepte sollen die Erfüllung der Datenschutzbestimmungen garantieren.
- Beratung für Smart-Data-Projekte: Projekte werden gezielt im Einsatz einer neuen Komponente beraten, beispielsweise durch Workshops und Tutorials mit Teilnehmern mehrerer Projekte.

Mit dem Ziel, datengetriebene Innovationsprozesse in den Bereichen Industrie 4.0, Energie, Smart Cities und Medizin durch geeignete Prozesse, Werkzeuge und Betriebskonzepte der Analyseinfrastruktur zu beschleunigen, soll das SDIL Gruppen aus Industrie und Forschung helfen, Projekte zeitnah durchzuführen. Dazu tragen neben einer leistungsfähigen und sicheren Infrastruktur auch rechtliche und organisatorische Rahmenwerke bei. Um Prozesse rund um das Thema Smart Data kontinuierlich zu verbessern, ist das SDIL jedoch weiterhin auf die rege Beteiligung angewiesen, welche durch die Data Innovation Communities angetrieben wird. Durch die breite Unterstützung seitens der Industrie ist bereits jetzt eine technische und organisatorische Grundlage geschaffen worden, die allen interessierten Anwendern und Entwicklern von Smart Data offen steht und von der Forschung und Innovation in Deutschland profitieren können.

Über die Autoren

PROF. DR. MICHAEL BEIGL

... ist Inhaber des Lehrstuhles für Pervasive Computing Systems und der Forschungsgruppe Technology for Pervasive Computing (TECO) am Karlsruher Institut für Technologie (KIT). Zuvor war Beigl Professor für verteilte und ubiquitäre Systeme und leitete die DUS-Forschungsgruppe an der TU Braunschweig. Er ist einer der Initiatoren und der Sprecher des Smart Data Innovation Labs (SDIL).

CHRISTOPH BODEN

... ist stellvertretender Projektkoordinator des Berlin Big Data Centers (BBDC) und wissenschaftlicher Mitarbeiter am Fachgebiet Datenbanksysteme und Informationsmanagement (DIMA) der Technischen Universität Berlin. Er hat Wirtschaftsingenieurwesen in der Fachrichtung Informations- und Kommunikationssysteme an der Technischen Universität Dresden, der Technischen Universität Berlin sowie der University of California, Berkeley, studiert.

DR. CHRISTOPH HEINZ

... ist Software Engineer Specialist bei der Software AG. Vor seiner Tätigkeit bei der Software AG war Herr Heinz Head of Marketing and Sales bei einem Start-up im Bereich des Complex Event Processing. Der Diplom-Wirtschaftsmathematiker hat an der Universität Marburg im Bereich Informatik promoviert.

DR. ALEXANDER LENK

... leitet die Außenstelle des FZI Forschungszentrums Informatik in Berlin und die Begleitforschung des vom Bundeswirtschaftsministerium geförderten Technologieprogramms „Smart Data – Innovationen aus Daten“. Er ist Forschungsfeldleiter „Cloud Computing und Cloud Services“ am FZI und Gastwissenschaftler am Karlsruher Institut für Technologie (KIT) sowie an der TU Berlin. Seine Forschungsschwerpunkte liegen auf den Themen Cloud Computing, Big Data und Informationswirtschaft.

PROF. DR. VOLKER MARKL

... leitet das Fachgebiet Datenbanksysteme und Informationsmanagement (DIMA) an der Technischen Universität Berlin sowie die Forschungsgruppe „Intelligente Analyse von Massendaten – Smart Data“ am Deutschen Forschungszentrum für künstliche Intelligenz (DFKI). Prof. Dr. Markl ist Direktor des vom BMBF eingerichteten Kompetenzzentrums „Berlin Big Data Center“ (BBDC) und leitet das vom BMWi geförderte „Forum Smart Data“.

PROF. DR. BERNHARD NEUMAIR

... ist Inhaber des Lehrstuhls „Management komplexer IT-Systeme“ am Institut für Telematik der Fakultät für Informatik am Karlsruher Institut für Technologie (KIT) sowie Geschäftsführender Direktor des Steinbuch Centres for Computing (SCC).

HENRIK OPPERMANN

... ist Business Unit Manager bei der USU Software AG und leitet den Forschungsbereich der USU Gruppe. Dort verantwortet er den Aufbau von Industrial Big Data Lösungen und koordiniert die Arbeiten der Smart Data Projekte. Herr Oppermann hat ein Studium des Maschinenbaus mit Fachrichtung Verfahrenstechnik an der TU München abgeschlossen und im Anschluss am Karlsruher Institut für Technologie (KIT) ein Aufbaustudium in Wirtschaftsingenieurwesen absolviert.

DR. TILL RIEDEL

... ist akademischer Rat am Lehrstuhl für Pervasive Computing Systems, Laborleiter der Forschungsgruppe Technology for Pervasive Computing (TECO) und operativer Leiter Anwendung des Smart Data Innovation Lab (SDIL) am Karlsruher Institut für Technologie (KIT). Riedel hat am KIT Informatik studiert und promoviert.

NICO SCHLITTER

... ist wissenschaftlicher Mitarbeiter am Steinbuch Centre for Computing (SCC) des Karlsruher Instituts für Technologie (KIT) und operativer Leiter Infrastruktur des Smart Data Innovation Labs (SDIL). Schlitter hat Informatik an der TU Chemnitz studiert.

Fußnoten

1 Teile dieses Textes wurden schon in englischer Sprache veröffentlicht in Lenk et al., Towards a Taxonomy of Standards in Smart Data, Proceedings of the 2015 IEEE Conference on Big Data, IEEE, 2015.

2 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Smart-Data-Auftaktkonferenz-Begleitforschung.pdf?__blob=publicationFile&v=1, Zugriff am 13.07.15.

3 Vgl. http://smart-data.fzi.de/memorandum/, Zugriff am 13.07.15.

4 Vgl. ebd.

5 Vgl. http://www.bmwi.de/DE/Mediathek/publikationen,did=601310.html

6 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/smart-data-innovationen-aus-daten.html, Zugriff am 15.07.15.

7 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151534, Zugriff am 13.07.15.

8 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151528, Zugriff am 13.07.15.

9 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151528, Zugriff am 13.07.15.

10 Vgl. Nationaler IT-Gipfel, Arbeitsgruppe 2 – Projektgruppe Smart Data. „Smart Data – Potenziale und Herausforderungen. Strategiepapier. Hamburg 2014.

11 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151528, Zugriff am 13.07.15.

12 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151528, Zugriff am 13.07.15.

13 Vgl. http://www.digitale-technologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_doc Id=151528, Zugriff am 13.07.15.

14 Vgl. “The Forrester Wave™: Big Data Streaming Analytics Platforms, Q3 2014“, http://www.forrester.com/pimages/rws/reprints/document/113442/oid/1-V33X9V, Zugriff am 01.09.2015.

15 Vgl. http://www.softwareag.com/corporate/products/default.asp, Zugriff am 01.09.2015.

16 Nathan Marz, James Warren: Big Data – Principles and best practices of scalable realtime data systems, April 2015. ISBN 9781617290343.

17 http://cassandra.apache.org/ http://spark.apache.org/

18 http://spark.apache.org/

19 https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark

20 https://www-03.ibm.com/press/us/en/pressrelease/47107.wss

21 https://spark-summit.org/2013/talk/feng-hadoop-and-spark-join-forces-at-yahoo/

22 http://spark.apache.org/mllib/

23 https://eclipse.org/

24 https://zeppelin.incubator.apache.org/

25 https://flink.apache.org/

26 Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, Felix Naumann, Mathias Peters, Astrid Rheinländer, Matthias J. Sax, Sebastian Schelter, Mareike Höger, Kostas Tzoumas, and Daniel Warneke. 2014. The Stratosphere platform for big data analytics. The VLDB Journal 23, 6 (December 2014).

27 http://www.bbdc.berlin/

Abbildungsverzeichnis

Abbildung 1
Das Technologieprogramm „Smart Data – Innovationen aus Daten“ des BMWi

Abbildung 2
Taxonomie einer Smart-Data-Plattform und Beispiele für Technologien

Abbildung 3
Die Streaming-Analytics-Plattform der Software AG

Abbildung 4
Die Auswahl der Datenbank

Abbildung 5
Die Auswahl der Verarbeitungsplattform

Abbildung 6
Systemarchitektur der U-Industrial-Big-Data-Lösung

Abbildung 7
Apache Flink – Eine Open-Source-Plattform für skalierbares Batch und Stream Processing

Abbildung 8
Das WordCount-Beispiel, implementiert in der Java- (links) und Scala-DataSet-API (rechts) von Apache Flink

Abbildung 9
Vergleich der verschiedenen Big-Data-Analytics-Frameworks Apache Hadoop, Apache Spark und Apache Flink

Abbildung 10
Die Struktur des Smart Data Innovation Labs

Abbildung 11
Forschungs- und Innovationszyklus im Smart Data Innovation Lab