Gesprengte Ketten

Smart Data, deklarative Datenanalyse, Apache Flink

Volker Markl
Technische Universität Berlin / DFKI GmbH
Sekr. E-N 7, Einsteinufer 17 / DFKI Projektbüro Berlin, Alt-Moabit 91c
10587 Berlin, Germany / 10559 Berlin, Germany
+49 30 314 23555
volker.markl@tu-berlin.de / volker.markl@dfki.de

Informatik Spektrum 01/2015 Preprint

Abstract

Das letzte Jahrzehnt war durch die Digitalisierung von praktisch allen Lebensbereichen gekennzeichnet. Wirtschaft, Wissenschaft und Gesellschaft stehen nun riesige Mengen von stetig wachsenden, oftmals heterogenen Daten zur Verfügung. Allerdings sind diese Daten weder Informationen noch Wissen. Sie sind erst wertvoll, wenn sie verfeinert und analysiert werden, so dass aus Rohdaten „Smart Data“ werden. Nur dann können die ökonomischen und sozialen Potenziale vollständig entfaltet werden, beispielsweise im Hinblick auf Prozessoptimierung, Massenindividualisierung oder andere Formen von Erkenntnis- oder Effizienzgewinnen. Leider stehen dem breiten Einsatz von „Big Data“-Anwendungen derzeit noch hohe Einstiegshürden entgegen. Das häufig genannte Berufsbild des „Data Scientist“ ist vom Anforderungsprofil so komplex, dass es nur wenige Experten gibt, die dieses erfüllen. In diesem Artikel identifizieren wir, dass Ausbildung alleine die Knappheit an „Data Scientists“ nicht beheben kann. Vielmehr sind unterstützende Technologien erforderlich, um das volle Potential von „Big Data“ in der Breite zu entfalten. Wir skizzieren die Technologie „Apache Flink“ als einen ersten Schritt, sowie die Vision des Berlin Big Data Center (BBDC) zur intelligenten Analyse von Massendaten.

1. Einführung

Heutzutage hört man oft “Daten sind das neue Öl!”. Genauso wie Öl, sind Daten eine komplexes Produkt, welches durch zahlreiche Verarbeitungs- und Verfeinerungsschritte entsteht. Ebenso kann man eine Analogie zu “Big Data” herstellen. Datenbohrtürme sind dann, z.B., Informationsextraktions- und Informationsintegrationsmethoden, welche Information aus den grundlegenden Rohdaten extrahieren und semantisch anreichern. Die Raffinerien entsprechen Datenanalyse- und Dataminingalgorithmen, Systemen und Werkzeugen, welche die Daten aggregieren, gruppieren und die Daten somit neu ordnen, um sie in Erkenntnisse und verwertbare Informationen umzuwandeln. Wir sehen bereits einen ganzen Wirtschaftszweig an Vertriebsnetzen rund um Big Data entstehen, wie z.B. Informationsmarktplätze, die semantisch angereicherte und erweiterte Daten verkaufen. Im Transport und Logistikwesen werden zunehmend Big Data Lösungen zur Fahrzeugortung und zur Optimierung des Flottenmanagement eingesetzt. Industrie 4.0 verwendet Big Data Analysen, um zukünftig intelligente Fertigungsprozesse zu ermöglichen. Auch im Gesundheitswesen wird zunehmend an Big Data Anwendungen gearbeitet. Big Data wird die wissenschaftlichen Prozesse nicht nur beschleunigen, sondern sogar viele wissenschaftliche Prozesse ändern und tiefgreifende Auswirkungen auf Wirtschaft, Wissenschaft und die Gesellschaft im Ganzen haben.

In diesem Zusammenhang entsteht das neue Berufsbild des „Data Scientist“, laut dem Harvard Business Review „The Sexiest Job of the 21st Century“. Ein Data Scientist soll „Big Data“ beherrschen und durch Einsatz von fortgeschrittenen Analysemethoden aus den Datenmassen Vorhersagen und Handlungsempfehlungen ableiten. Allerdings sind „Data Scientists“ rar. Diese Knappheit von „Data Scientists“ hat mehrere Gründe. Zum einen sind zur Analyse von heterogenen Daten tiefe Kenntnisse aus verschiedenen Gebieten der Datenanalyse erforderlich, beispielsweise aus den Bereichen Graphen- und Netzwerkanalyse, Maschinelles Lernen, Mathematik, Statistik, Signal- und Sprachverarbeitung. Zum anderen waren in der Vergangenheit die Disziplinen der Datenanalyse und skalierbare Datenverarbeitung nicht eng verzahnt, was jedoch für einen souveränen Umgang mit großen Datenmengen mit geringer Latenz erforderlich ist. Ferner findet die Datenanalyse jedoch nicht in einem luftleeren Raum statt. Somit sind neben den Kenntnissen in der Datenanalyse und der skalierbaren Datenverarbeitung auch noch Domänenwissen in einem Anwendungsgebiet, beispielsweise Industrie 4.0, Logistik, Gesundheit,

Energie, Marktforschung sowie ggf. weiterführende Kenntnisse bezüglicher wirtschaftlicher, rechtlicher oder gesellschaftlicher Fragen von Nöten. Die Nachfrage nach derartigen Experten wächst stetig. Dementsprechend sind auch die Kosten für die Analyse von Big Data sehr hoch. Schlimmer noch ist die Tatsache, dass trotz vieler neuer Ausbildungsprogramme (an Universitäten weltweit) und motivierten Studenten es trotzdem unmöglich sein wird, diese „eierlegendenden Wollmilchsäue“ zu schaffen, da die benötigten Fähigkeiten sehr komplex sind und sich stark voneinander unterscheiden (wie in Abb. 1 dargestellt). Das das Wall Street Journal vom 10.4.2012 spricht in diesem Zusammenhang von „[…] Little Talent” als Big Data's Big Problem“.

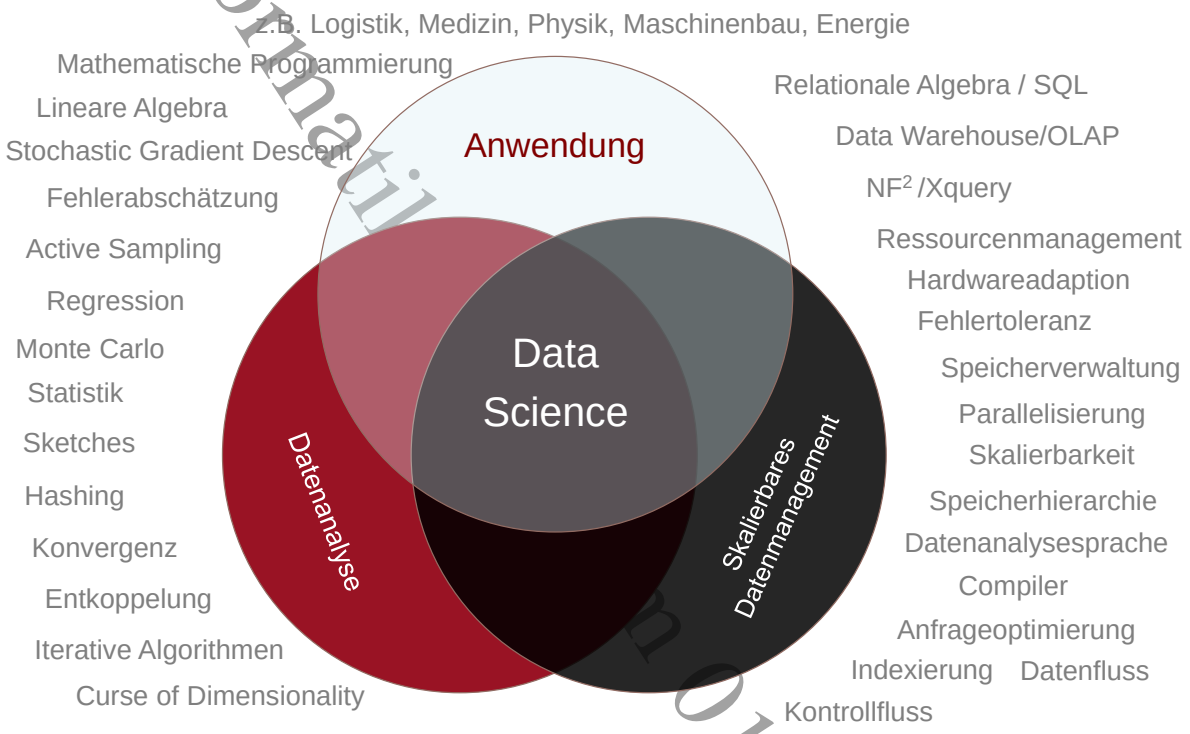

Abb. 1 Ein „Data Scientist“ ist eine *eierlegende Wollmilchsau*

Vor dem Trend zu „Big Data“ waren die wenigen Programmierer oder Analysten, die eine tiefergehende Analyse von sehr großen Datenmengen realisieren konnten, im Kontext des wissenschaftlichen Rechnens und Supercomputing beschäftigt. Diese Experten mussten tief in Trickkiste der Systemprogrammierung greifen, um parallele Algorithmen zur Datenanalyse in MPI unter Reduktion von Kommunikation sowie unter optimaler Ausnutzung von Speicher- und Plattenressourcen zu realisieren. Dank High-Level Programmierung, Compilern und Datenbanksystemen sowie insbesondere der Anfragesprache SQL mussten sich für viele Jahrzehnte Software Entwickler und normale Anwender keine Gedanken über die Skalierbarkeit bezüglich ihrer Datenbankauswertungen machen. Durch „Big Data“ haben die derzeitigen Technologien jedoch ihre Grenzen erreicht. SQL kann keine iterativen Algorithmen effizient verarbeiten. Jedoch sind derartige Algorithmen Bestandteil jeder nicht-trivialen Datenanalyse, sei es bei Algorithmen des Maschinellen Lernens, z.B. Regression, Support-Vektor-Maschinen, Entropie Maximierung, oder Algorithmen der Signalverarbeitung, der Graphanalyse, oder der Optimierung. Die Anforderungen dieser Verfahren bezüglich skalierbarer Datenverarbeitung sind weitaus höher sind als ehemals bei der relationalen Algebra. So muss die skalierbare Datenverarbeitung neben den klassischen relationalen Operationen der Selektion, Projektion, Kreuzprodukt, Vereinigung und Differenz auch mit benutzerdefinierten Funktionen und Iterationen sowie mathematischen Operationen, insbesondere Operationen auf Matrizen, umgehen, und diese auf parallelen und verteilten Rechnerarchitekturen zur Ausführung bringen. Dies zusammen mit der Tatsache, dass wir durch „Moore’s Gesetz“ Leistungssteigerungen bei Rechnerarchitekturen nur noch durch Parallelisierung, Verteilung und Spezialisierung erreichen werden, hat die Datenverarbeitung erheblich verkompliziert und dazu geführt, dass die Analyse von „Big Data“ wenigen Universalexperten vorbehalten ist.

Im Zeitalter von Mehrkernprozessoren und heterogenen CPUs sowie Cloud Computing und NoSQL müssen wir dafür sorgen, dass die gut etablierten Konzepte deklarativer Sprachen (also der Inbegriff relationaler Datenbanksysteme) ihren Weg in „Big Data“-Systeme finden. Um dies auch Wirklichkeit werden zu lassen, muss die Forschung einige Herausforderungen lösen. Zum Beispiel (i) eine (deklarative) Datenanalysesprache entwickeln, die keine Systemprogrammierungsfähigkeiten benötigt, (ii) in dieser Sprache geschriebene Programme auf die gewählte Ausführungsplattform übersetzen und automatisch an Rechnerarchitektur, Datenverteilung und System-

Informatik Spektrum 2012 Preprint

last anpassen und (iii) dies alles auf eine skalierbare Weise. Im Detail bedeutet dies, Ausführungsstrategien zu erarbeiten, die sowohl verteilt und parallelisiert funktionieren und gleichzeitig sowohl in-memory als auch Sekundärspeicherverarbeitung von fortgeschrittenen Algorithmen auf riesigen Datenmengen erlauben. Um diese Herausforderungen erfolgreich zu lösen, müssen unter anderem Forschungsgruppen in *Übersetzerbau, Datenanalyse, Datenbanktechnologie, Verteilte Systeme und Maschinelles Lernen* zusammenarbeiten und ihre Lösungsansätze mit Anwendern aus verschiedensten Bereichen validieren. Wir müssen neue skalierbare Algorithmen und Systeme entwickeln, die in der Lage sind, riesige, heterogene Datenmengen –und ströme zu organisieren und daraus Informationen zu extrahieren und zu integrieren. Dafür müssen deklarative Anfragesprachen um die Konzepte von Iterationen und benutzerdefinierten Funktionen sowie dem Management von Zustand erweitert werden, um die Spezifikationen und Verarbeitung von fortgeschrittenen Analysemethoden aus den Bereichen des Maschinellen Lernens, der Signalverarbeitung, der mathematischen Optimierung, der Sprach-, Audio- oder Videoverarbeitung oder anderen Bereichen der Datenanalyse zu ermöglichen.

Abb. 2: Tiefe Analysen von Big Data ist das Ziel!

Derzeit befindet sich „Big Data“ hinsichtlich des Reifegrads der Technologien noch in der Steinzeit. Diese Steinzeit stellt sich dar in Algorithmen, die entweder in primitiven Sprachen ausgedrückt werden und handoptimiert werden müssen (z.B. in MPI, MapReduce), oder in relational-orientierten Sprachen (z.B. SQL, XQuery, Pig, Hive und JAQL) ausgedrückt werden und dann mit nicht-optimierbaren externen Treiberprogrammen für iterative Algorithmen kombiniert werden, oder in „analysefreundlicheren“ Sprachen und Systemen (z.B. R oder Python), die nicht für „Big Data“ skalieren. Erforderlich ist nun der Übergang von der Steinzeit in eine neue Epoche von „Big Data“. Diese muss die Stärke deklarativer Sprachen, nämlich die automatische Optimierung, Parallelisierung und Anpassung des gleichen Programmes an die Rechnerarchitektur (abhängig von der Datenverteilung, Datengröße, Datenrate und Systemauslastung), erhalten und damit die Produktivität der „Data Scientists“ steigern, was wiederum zu reduzierten Analysekosten, schnellerer Analysezeit und insgesamt zu einem breiteren Zugang und einer breiteren Anwendung von Datenanalysen in Wirtschaft, Wissenschaft und Gesellschaft führen wird.

2. Ein erster Schritt: Stratosphere und Apache Flink

Mit finanzieller Unterstützung der DFG[1], der EIT[2] ICT Labs, von BMBF[3] und BMWi[4] und der Europäischen Kommission (im Rahmen des FP7 Programmes) sowie von mehreren Unternehmen hat ein großes Team von

[1] Die Deutsche Forschungsgemeinschaft (DFG) fördert Stratosphere als Forschergruppe FOR 1306. An dieser Forschergruppe sind TU Berlin, HU Berlin und das Hasso-Plattner-Institut der Universität Potsdam beteiligt.

Informatik Spektrum 01/2015 Pre-Print

Forschern und Entwicklern einen ersten Schritt zur Erforschung und Entwicklung einer neuen Infrastruktur zur Analyse von „Big Data“ unternommen, zunächst publiziert unter dem Namen Stratosphere [1,2], seit 2014 als Open-Source System unter dem Namen Apache Flink [9]. In Apache Flink kann man massiv parallele Datenanalysen mithilfe eines funktionalen Programmiermodells durchführen. Durch das Konzept von Bulk- und Delta-Iterationen kann Apache Flink iterative Verfahren der Datenanalyse effizient realisieren und damit Operatoren zur Informationsextraktion und Integration zusammen mit Operatoren für tiefgreifende Datenanalysen in einem System verarbeiten. Dadurch fasst Flink es viele spezialisierte Systeme zur Datenverarbeitung oder Maschinellem Lernen in einer Umgebung zusammen [3]. Das System, veröffentlicht unter der Apache 2.0 Software Lizenz, kann sowohl auf einzelnen Rechnern, in Rechnerclustern, oder - ohne spezielle Installation - in Hadoop Clustern via Yarn installiert werden. Technologisch gesehen ist Flink ist ein Hybridsystem, das sowohl Datenbanktechnologien als auch Map/Reduce-Technologien in sich vereint. Damit besitzt Flink die Fähigkeiten zur Verarbeitung von komplexen benutzerdefinierten Funktionen, zur Ableitung des Datenbankschemas zum Analysezeitpunkt („schema on read“) und zum Umgang mit komplexe Datentypen sowie die Skalierbarkeit von Map/Reduce Systemen und kombiniert diese mit den Konzepten der Spezifikation und Optimierung aus Datenbanksystemen. Flink biete *Java* und *Scala* als Sprachschnittstellen, sowie eine Schnittstelle zur Graphenanalyse namens *Spargel*, angelehnt an das spezialisierte System Pregel. Die Architektur von Flink ist in Abbildung 3 dargestellt.

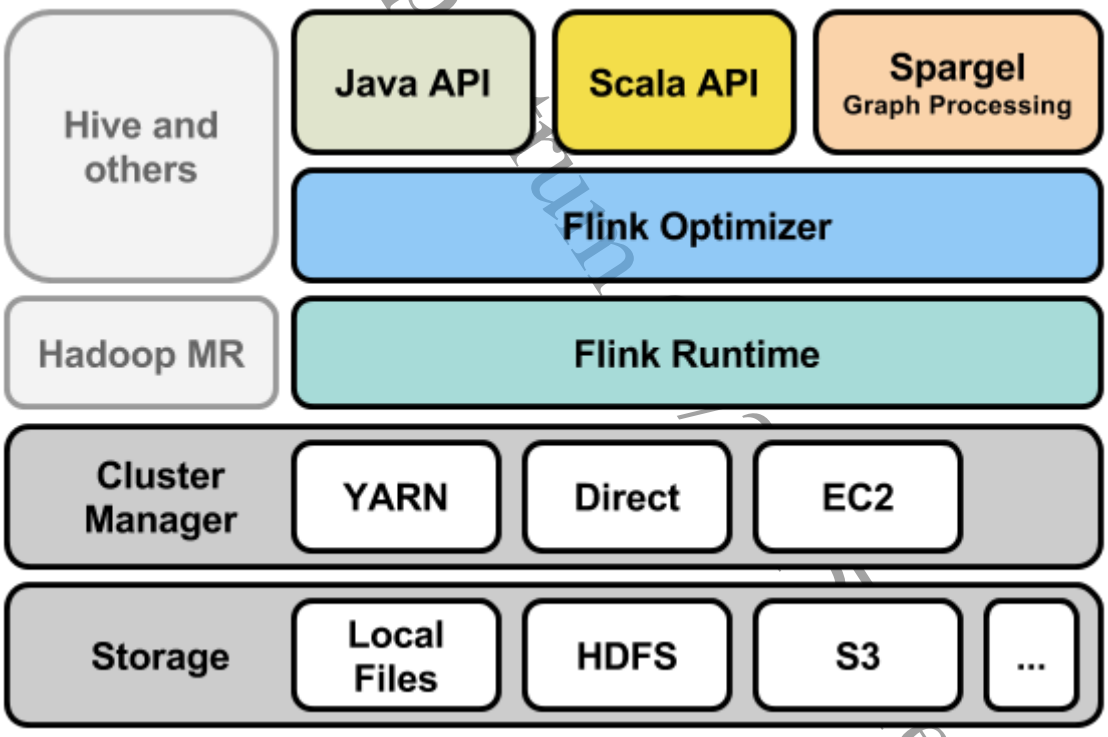

Abb. 3 Die Architektur von Stratosphere/Flink

Informatik Spektrum – Pre-Print

[2] Die ICT Labs des European Institute of Innovation and Technology (EIT) unterstützt Apache Flink im Rahmen des EUROPA Projektes. An diesem sind neben der TU Berlin als Partner SICS, INRIA, ELTE/SZTAKI sowie mehrere Firmen beteiligt.

[3] Das Bundesministerium für Bildung und Forschung fördert die Weiterentwicklung von Flink im Rahmen des Kompetenzzentrums „Berlin Big Data Center“ (BBDC). Am BBDC sind neben der TU Berlin das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI), das Zuse-Institut Berlin, die Beuth Hochschule, sowie das Fritz-Haber-Institut der Max-Planck-Gesellschaft beteiligt.

[4] Das Bundesministerium für Wirtschaft und Energie (BMWi) fördert die Weiterentwicklung von Flink im Rahmen der Projekte SmartDataWeb und SD4M.

3. Wie geht es weiter?

Um auf großen, sich verändernden Datenmengen iterative und zustandsbehaftete Analyseprogramme mit nur geringer Latenz zu verarbeiten und dabei auch noch deklarative Spezifikationsmöglichkeiten zu ermöglichen, sind neue Methoden und Techniken im Hinblick auf Datenanalyseverfahren wie auch im Hinblick auf Systementwicklung erforderlich. Die bisherigen Arbeiten an Technologien jenseits von Hadoop wie Stratosphere/Flink [1,2,10], Spark [4], ePIC [5], Asterix [6] haben eine Grundlage für die skalierbare Verarbeitung von komplexen Datenanalyse geschaffen. Diese Systeme liefern eine inspirierende Basis für weitere Forschungen, müssen jedoch angepasst und erweitert werden.
Neben Technologien umfasst „Big Data“ (oft auch „Smart Data“ genannt) einen viel weiteren Bereich und bietet Möglichkeiten und Herausforderungen in 5 Dimensionen: Technologie, Anwendung, Wirtschaft, Recht und Gesellschaft (siehe Abbildung 4).

Abb. 4 Die 5 Dimensionen von „Big Data“

Im Folgenden skizzieren wir ausgewählte Aspekte entlang dieser Dimensionen und verweisen für Details auf eine Studie [7].

Technologie: Wie oben beschrieben, benötigen wir skalierbare Systeme und Plattformen für Datenanalysen und insbesondere Technologien, die dabei helfen, die Fachkräftelücke zu schließen. Daneben bestehen Herausforderungen auch im Bereich der Datenanalysealgorithmen, insbesondere für die Analyse von Datenströmen, im Umgang mit heterogenen Rechnerarchitekturen, sowie bei der in-situ Analyse von Daten am Erzeugungsort.

Anwendung: Viele neue Anwendungen entstehen in der Informationswirtschaft, wie z.B. Informationsmarktplätze, welche Daten anreichern und verkaufen. Informationsmarktplätze verbessern wirksam die Situation der Informationswirtschaft, insbesondere im Hinblick auf mittelständische Unternehmen und Startups. Andere Beispiele umfassen die personalisierte Medizin, Industrie 4.0 und die digitale Geisteswissenschaften. Die Informatik muss Ihre Algorithmen, Systeme und Technologien anhand derartiger Anwendungen validieren.

Wirtschaft: Die Chancen und Risiken in der wirtschaftlichen Dimension liegen in neuen Geschäftsmodellen und im Paradigmenwechsel bei der Inhaltsverteilung (z.B. Preisfestsetzung für Informationen, die Rolle von Open-Source-Software). Insbesondere für die Wirtschaftsinformatik ergeben sich hier spannende Herausforderungen und Forschungsfragestellungen.

Recht: Aus juristischer Sicht entstehen durch Big Data, zusätzlich zu den gängigen Aspekten wie Datenschutz und Datensicherheit, neue Herausforderungen in Bezug auf Eigentum, Haftung und Insolvenz.

Soziale Dimension: Die datengetriebene Innovation wird tiefgreifende Auswirkungen auf die Gesellschaft im Ganzen haben, u.a. in Bezug auf das Zusammenleben, Nachrichten und demokratische Prozesse.

Informatik Spektrum 1/2015 Preprint

Das vom BMBF geförderte Kompetenzzentrum „Berlin Big Data Center“ (BBDC) [8] hat sich für die Jahre 2014 bis 2018 das Ziel gesetzt, einige Herausforderungen in der technologischen Dimension anzugehen und Lösungen in ausgewählten Anwendungen zu demonstrieren Gleichzeitig plant das Zentrum, durch die Technologien und Anwendungen auch Inspiration für die Forschung in den anderen Dimensionen von Big Data zu geben und mit Partnern Forschungsfragen auch in diesen Bereichen zu beleuchten. Aufbauend auf Apache Flink wird das BBDC die deklarative Spezifikation von komplexen Datenanalysen ermöglichen und auf diese Weise die Welt des Maschinellen Lernens mit der Welt der Datenanalyse zu verschmelzen. Um dies zu erreichen, verfolgt das Berliner Big Data Center die folgenden sieben Ziele im Kontext von Forschung, Innovation, sowie Aus- und Weiterbildung:

- **Kompetenzen bündeln** in skalierbarem Datenmanagement, Datenanalyse und Big Data Anwendungen.
- **Durchführung von Grundlagenforschung**, um neuartige und automatische skalierbare Technologien zu entwickeln, die in der Lage sind tiefgreifende Analysen von „Big Data“ durchzuführen.
- **Entwicklung eines integrierten, deklarativen, hoch-skalierbaren Open-Source-Systems**, welches in der Lage ist die Spezifikation, automatische Optimierung, Parallelisierung und Hardwareadaption und die fehlertolerante, effiziente Ausführung von Datenanalyseproblemen mittels verschiedener Methoden (z.B. Methoden des maschinellen Lernens, linearen Algebra, Statistik/Wahrscheinlichkeitstheorie, Computerlinguistik oder Signalverarbeitung) durchzuführen.
- **Wissens- und Technologietransfer**, um Innovation in Firmen und Startups zu unterstützen.
- **Ausbildung von Data Scientists** mit Bezug auf die fünf Big Data Dimensionen (d.h., Anwendung, Wirtschaft, Recht, soziale und technologische Dimension).
- **Menschen in die Lage zu versetzen „Smart Data" zu nutzen**, um neue Information aus massiven Datenmengen zu ermitteln.
- **Die Allgemeinheit in die Lage zu versetzen, korrekte datengetriebene Entscheidungsfindung durchzuführen.**

4. Schlusswort

Hinter Stratosphere und Flink steht ein sehr großes Forschungs- und Entwicklungsteam, welches nicht vollständig an dieser Stelle genannt werden kann. Ich danke insbesondere allen Mitgliedern meines Fachgebiets, sowie ehemaligen Doktoranden, Post-Docs und Studenten, die Stratosphere und Apache Flink geschaffen haben und weiterentwickeln. Ferner danke ich meinem Kollegen Prof. Dr. Odej Kao und seinen Mitarbeitern an der Technischen Universität Berlin sowie unseren vielen Partnern an der Humboldt Universität Berlin, dem Hasso Plattner Institut in Potsdam, den EIT ICT Labs und der gesamten Apache Flink Community dafür, dass wir zusammen ein System bauen konnten und weiterentwickeln dürfen, welches unsere Vision realisiert. Außerdem danke ich den Kollegen und Mitarbeitern von TU Berlin, TU München und Universität Münster, mit denen ich die Innovationspotentialanalyse zu „Big Data“ durchführen durfte, sowie Herrn Prof. Dr. Klaus-Robert Müller und meinen weiteren Kollegen im Berlin Big Data Center (BBDC) für die gemeinsame Gestaltung der Vision des BBDC sowie deren Umsetzung.

Literaturverzeichnis

1. A. Alexandrov, R. Bergmann, S. Ewen, et al.: “The Stratosphere Platform for Big Data Analytics,” VLDB Journal 05/2014.
2. Stratosphere, http://www.stratosphere.eu, zuletzt besucht am 17.11. 2014
3. S. Ewen, K. Tzoumas, M. Kaufmann, et al.: “Spinning Fast Iterative Data Flows,” PVLDB 5(11): 1268-1279 (2012).
4. M. Zaharia, M. Chowdhury, M. J. Franklin, et al: “Spark: cluster computing with working sets,” HotCloud (2010).
5. D. Jiang, G. Chen, B. C. Ooi, K.-L. Tan, S. Wu: “epiC: an Extensible and Scalable System for Processing Big Data,” PVLDB 7(7): 541-552 (2014).
6. S. Alsubaiee, Y. Altowim, H. Altwaijry, et al: “ASTERIX: An Open Source System for Big Data Management and Analysis.” PVLDB 5(12): 1898-1901 (2012).
7. Markl, V.; Krcmar, H; Hoeren, T.: „Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen“, http://www.dima.tu-berlin.de/menue/research/big_data_management_report/ , zuletzt besucht am 17.11.2014
8. Berlin Big Data Center, http://bbdc.berlin, zuletzt besucht am 18.11.2014
9. Apache Flink Incubator Project, http://flink.incubator.apache.org/, zuletzt besucht am 17.11., 2014

Informatik Spektrum 01/2015 Pre-Print