Technologie Workshop „Big Data”

22. Juni 2015

FZI Forschungszentrum Informatik
am Karlsruher Institut für Technologie

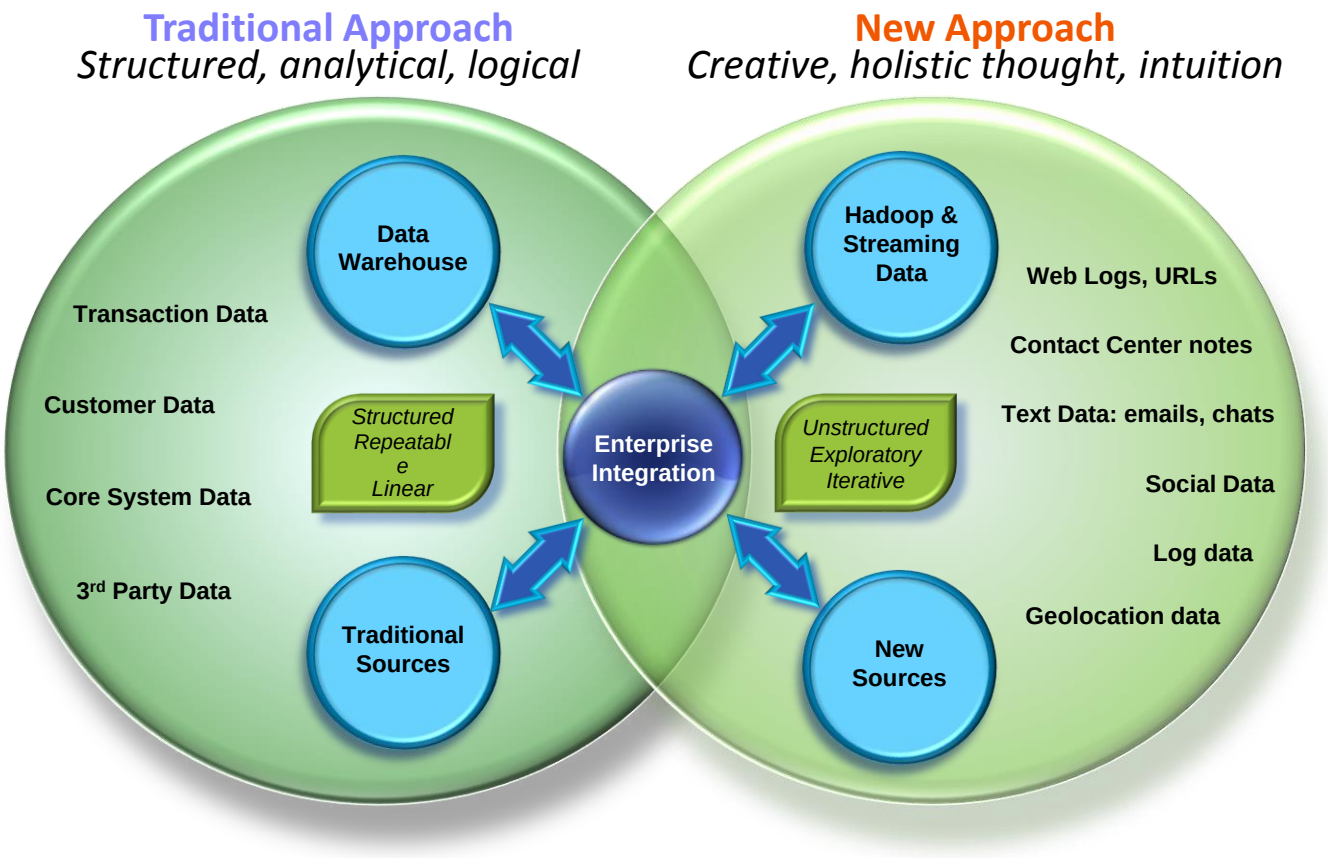
Axel J. Schwarz
Mobil: +49 171 5619419
E-Mail: axel.j.schwarz@de.ibm.com

IBM Smarter Analytics für „Big Data”

v1.3

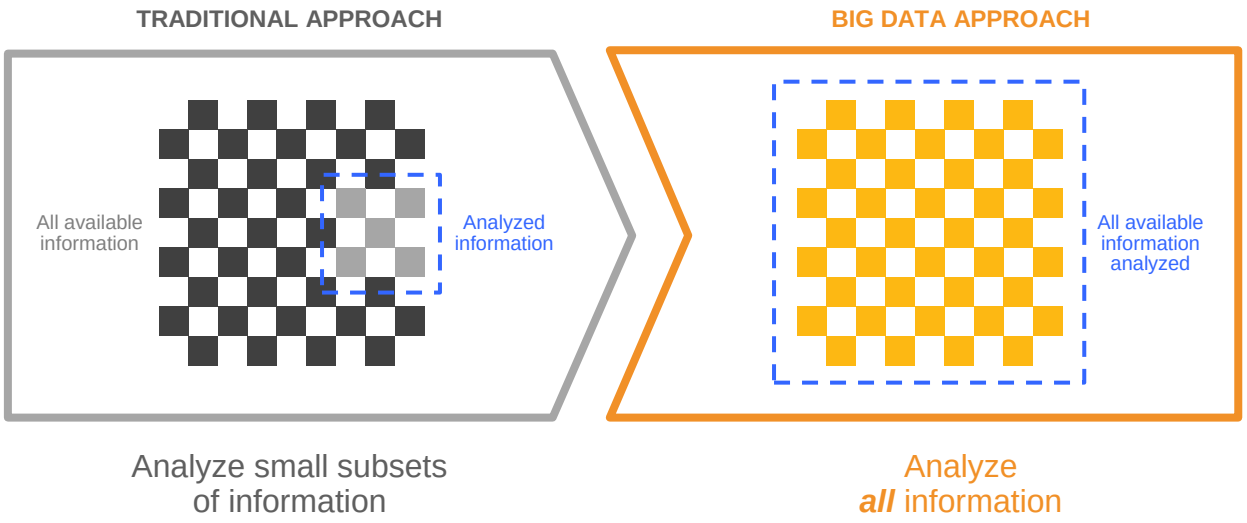
Paradigm shift enabled by „Big Data"

Analytics is expanding from enterprise data to big data, creating new opportunities for competitive advantage



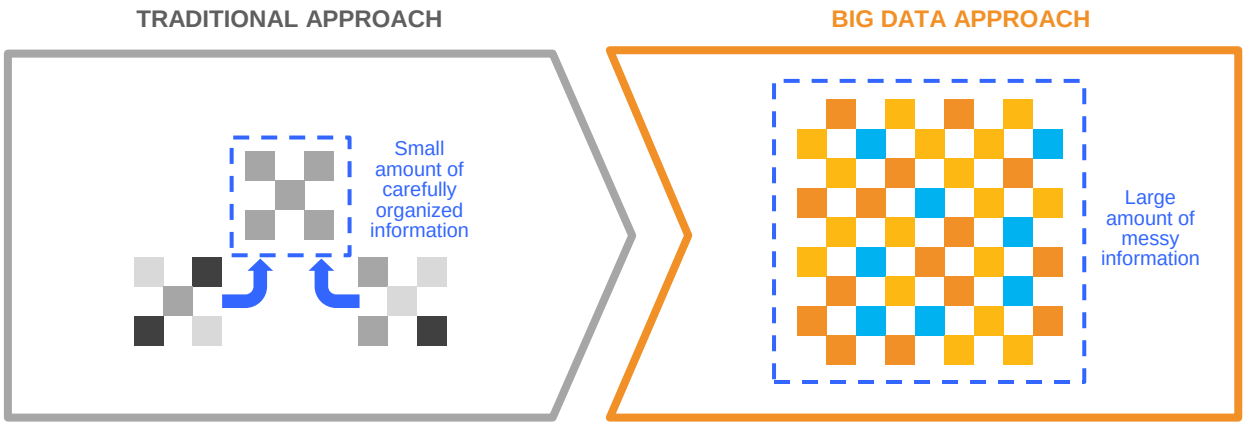
Paradigm shift enabled by „Big Data“

Leverage more of the data being captured



Paradigm shift enabled by „Big Data“

Reduce effort required to leverage data

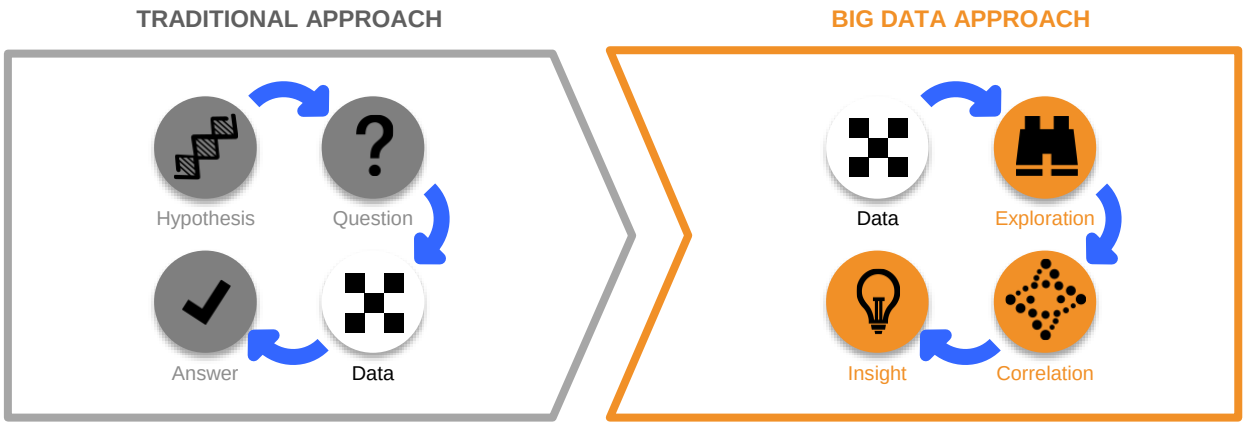


Carefully cleanse information ***before*** any analysis

Analyze information as is, cleanse as needed

Paradigm shift enabled by „Big Data"

Data leads the way—and sometimes correlations are good enough

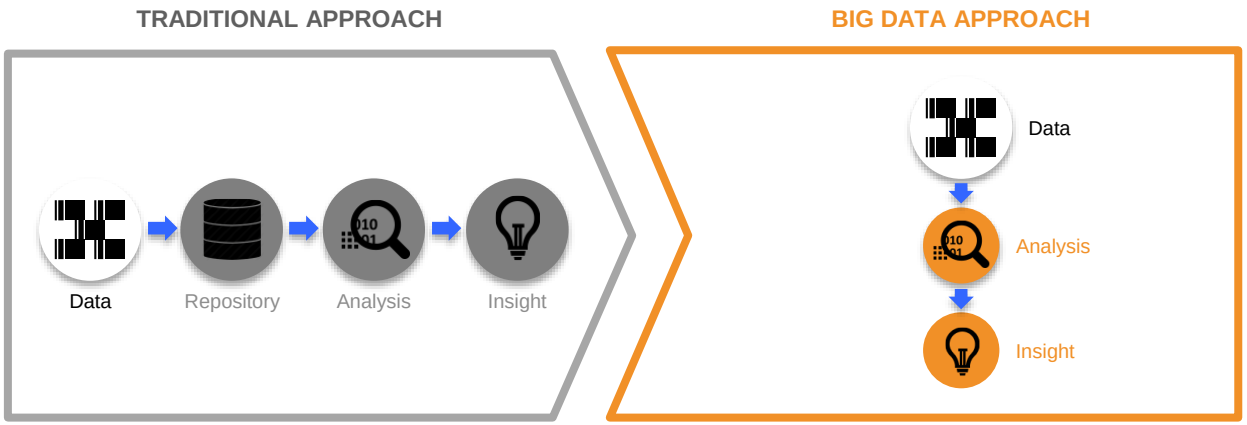


Start with hypothesis and test against selected data

Explore ***all*** data and identify correlations

Paradigm shift enabled by „Big Data"

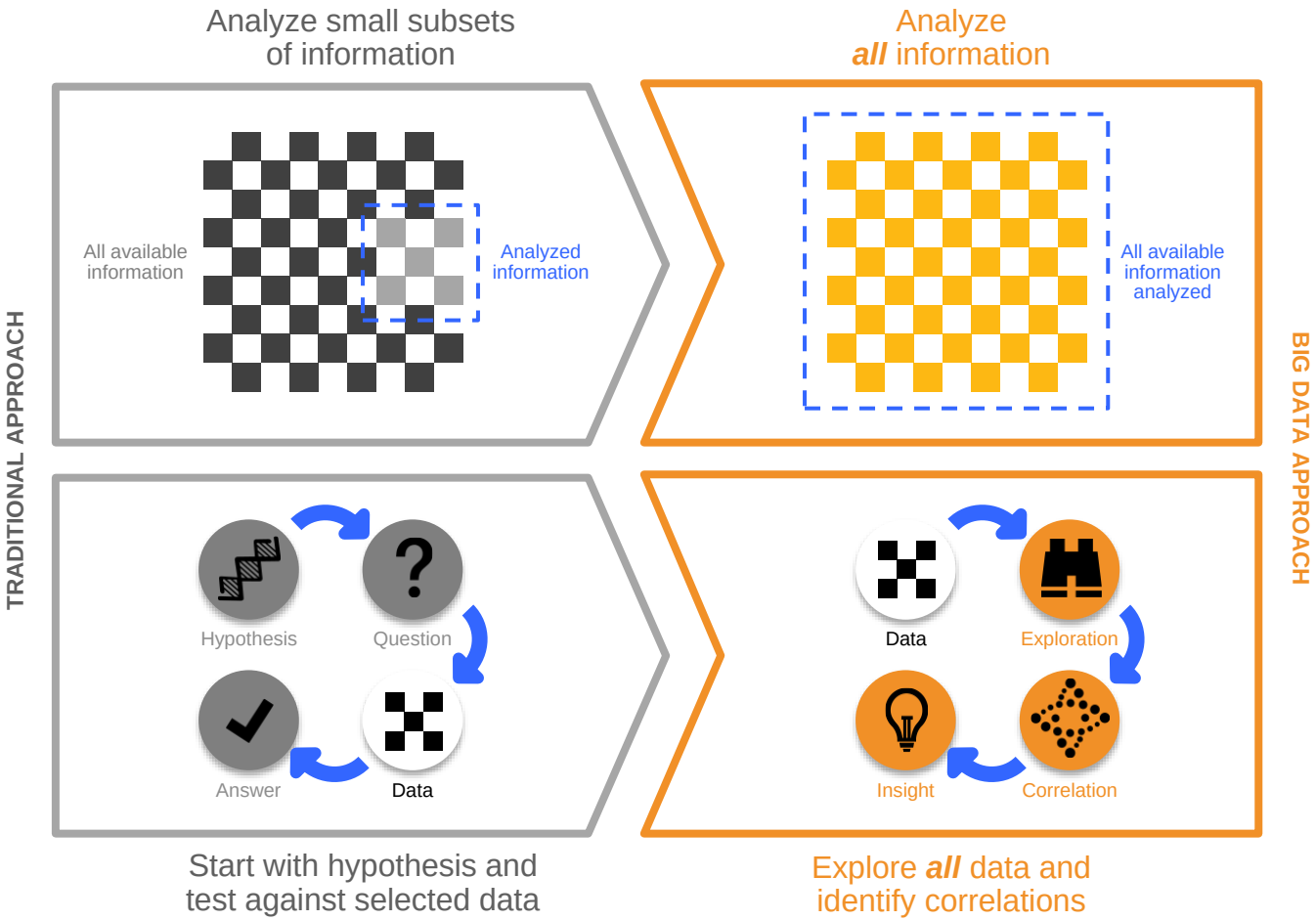
Leverage data as it is captured



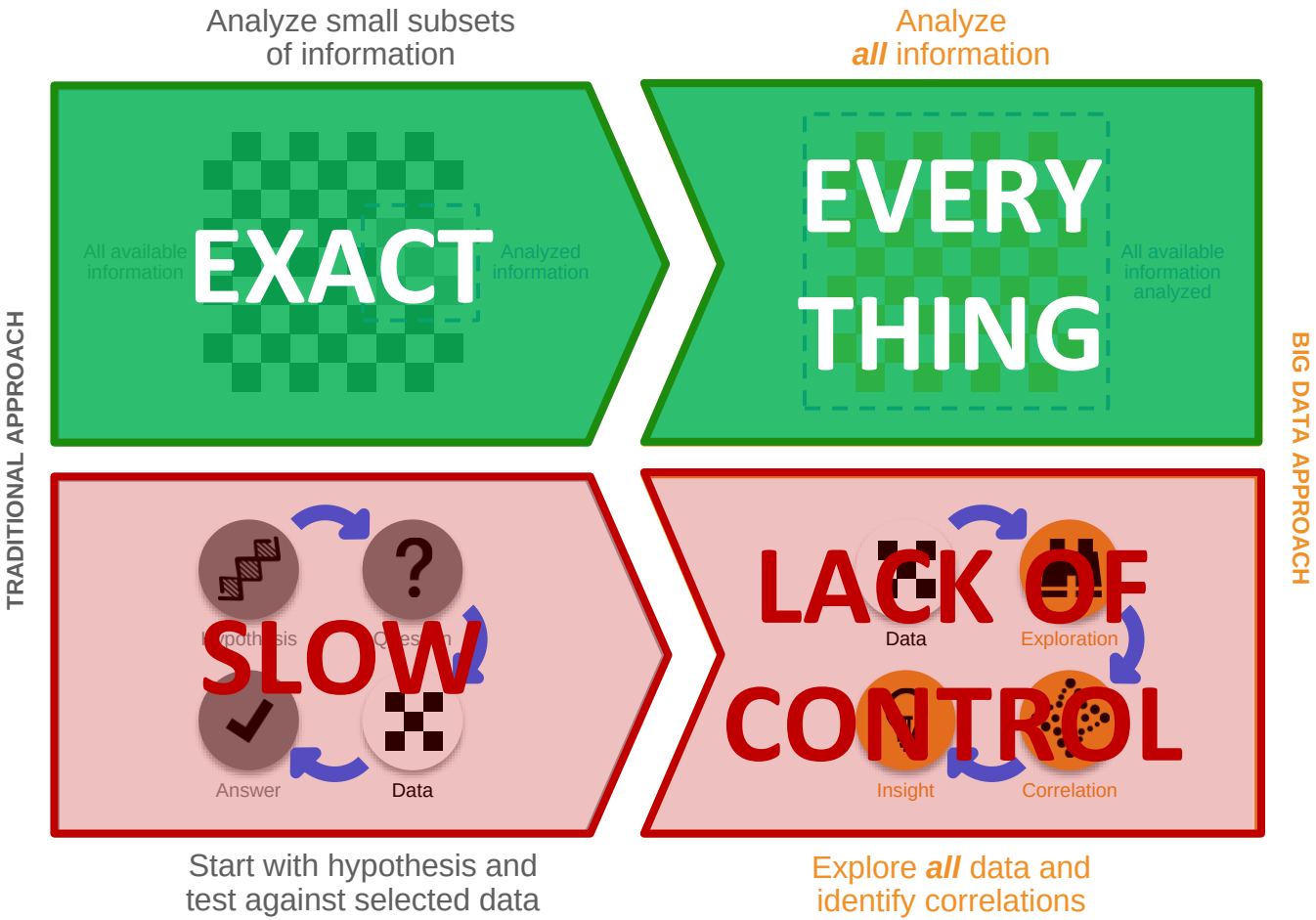
Analyze data ***after*** it's been processed and landed in a warehouse or mart

Analyze data ***in motion*** as it's generated, in real-time

Paradigm shift enabled by „Big Data" - Summary

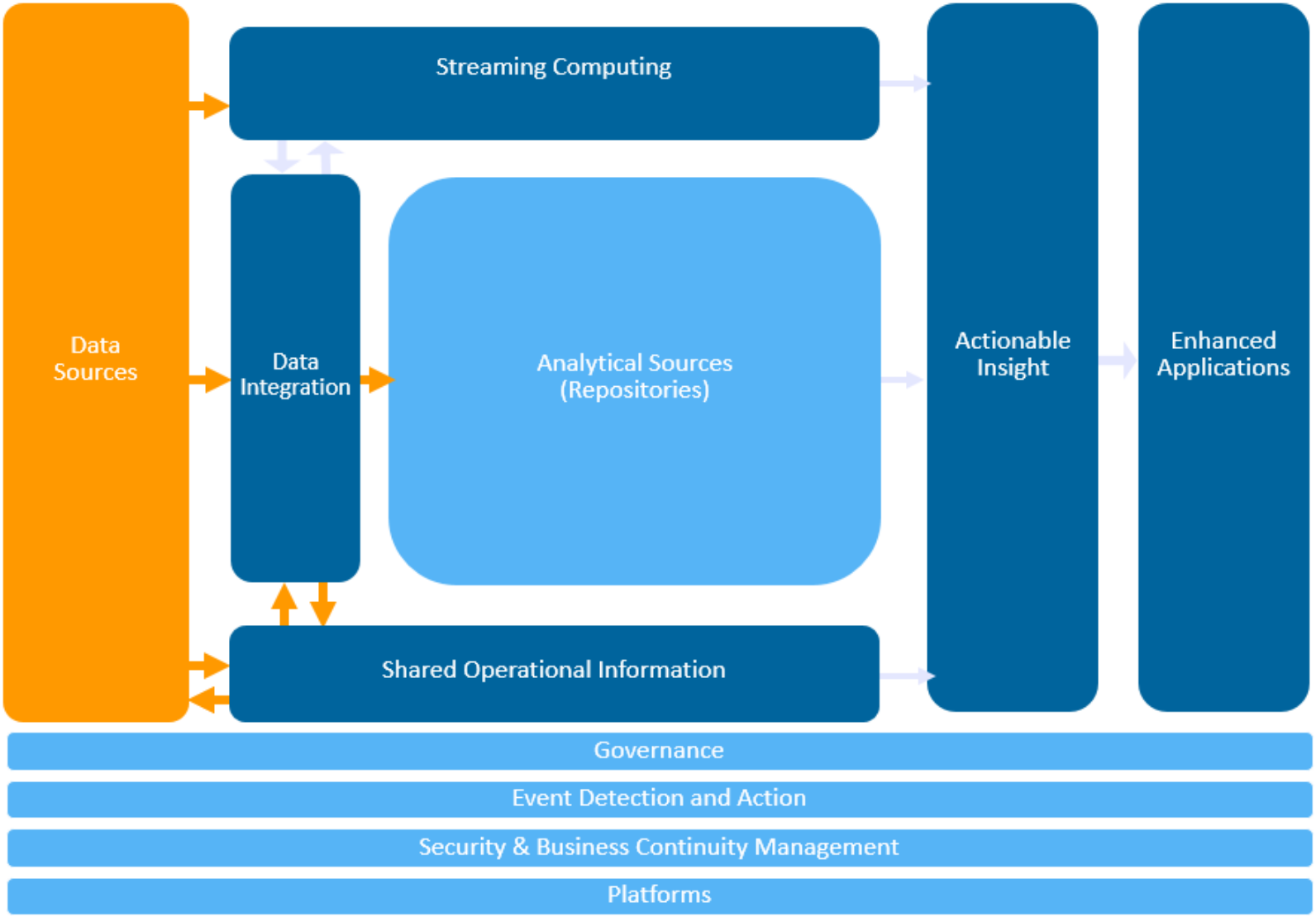


Paradigm shift enabled by „Big Data" - Summary



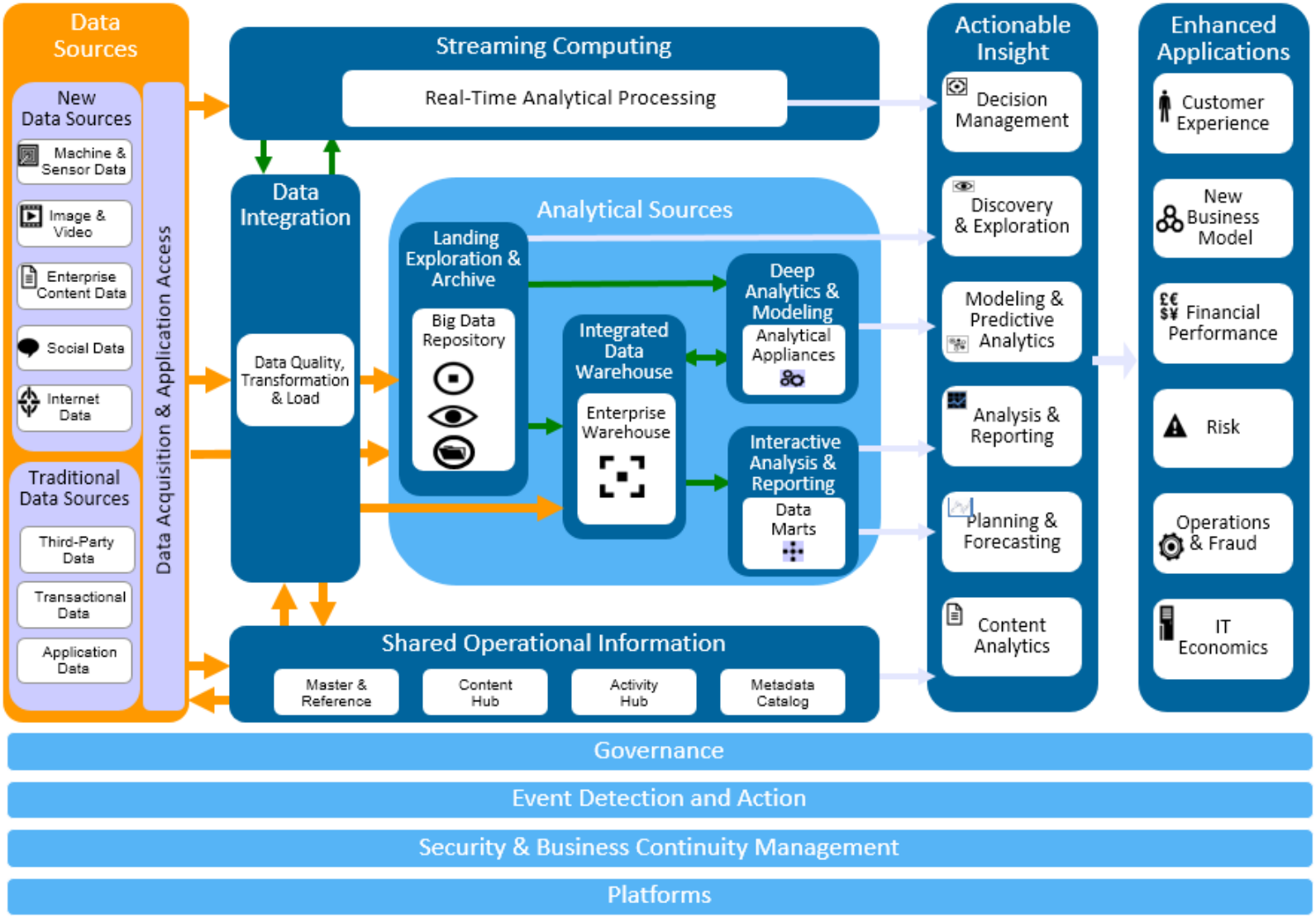
IBM Big Data & Analytics Referenz-Architektur

Komponenten - Überblick



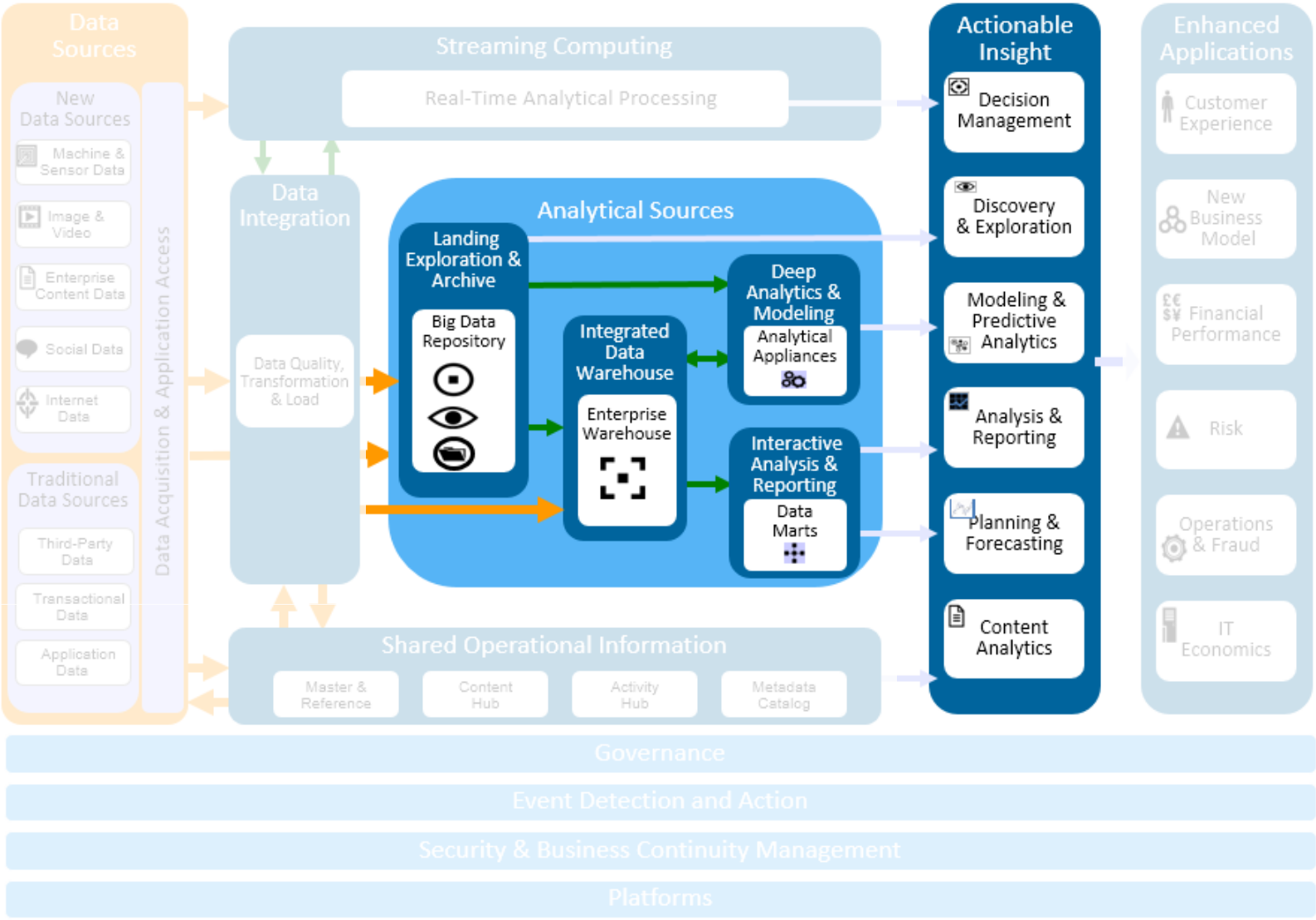
IBM Big Data & Analytics Referenz-Architektur

Komponenten - Detaillierung



IBM Big Data & Analytics Referenz-Architektur

Komponenten – Fokus



In-Memory „Big Data“ Processing - Apache Spark
IBM Investing in Four Catalysts for Big Data Adoption

Open Source Innovation

Technical Standards

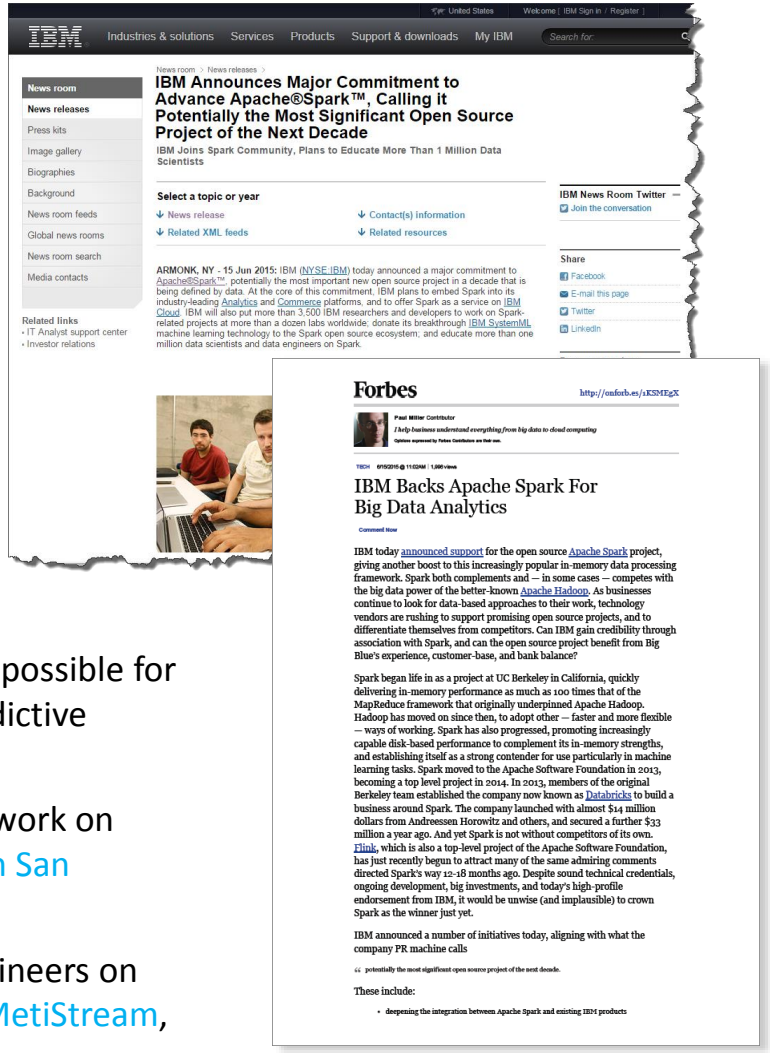
Familiar Interfaces & Integration with Established Tools

New Analytics Capabilities

In-Memory „Big Data" Processing - Apache Spark
…potentially the most significant open source project of the next decade.

To further accelerate open source innovation for the Spark ecosystem, IBM is taking the following actions:

- IBM will build Spark into the core of the company's analytics and commerce platforms.
- IBM's Watson Health Cloud will leverage Spark as a key under-pinning for its insight platform, helping to deliver faster time to value for medical providers and researchers as they access new analytics around population health data.
- IBM will open source its breakthrough IBM SystemML machine learning technology and collaborate with Databricks to advance Spark's machine learning capabilities.
- IBM will offer Spark as a Cloud service on IBM Bluemix to make it possible for app developers to quickly load data, model it, and derive the predictive artifact to use in their app.
- IBM will commit more than 3,500 researchers and developers to work on Spark-related projects […], and open a Spark Technology Center in San Francisco […].
- IBM will educate more than 1 million data scientists and data engineers on Spark through extensive partnerships with AMPLab, DataCamp, MetiStream, Galvanize and BigData University MOOC.



IBM Industries & solutions Services Products Support & downloads My IBM

News room
News releases
Press kits
Image gallery
Biographies
Background
News room feeds
Global news rooms
News room search
Media contacts

Related links
- IT Analyst support center
- Investor relations

News room > News releases >

IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open Source Project of the Next Decade

IBM Joins Spark Community, Plans to Educate More Than 1 Million Data Scientists

Select a topic or year
News release
Contact(s) information
Related XML feeds
Related resources

ARMONK, NY - 15 Jun 2015: IBM (NYSE:IBM) today announced a major commitment to Apache®Spark™, potentially the most important new open source project in a decade that is being defined by data. At the core of this commitment, IBM plans to embed Spark into its industry-leading Analytics and Commerce platforms; and to offer Spark as a service on IBM Cloud. IBM will also put more than 3,500 IBM researchers and developers to work on Spark-related projects at more than a dozen labs worldwide; donate its breakthrough IBM SystemML machine learning technology to the Spark open source ecosystem; and educate more than one million data scientists and data engineers on Spark.

IBM News Room Twitter
Join the conversation

Share
Facebook
E-mail this page
Twitter
LinkedIn

Forbes http://onforb.es/1KSMEgX

Paul Miller Contributor
I help business understand everything from big data to cloud computing
Opinions expressed by Forbes Contributors are their own.

TECH 6/15/2015 @ 11:02AM | 1,095 views

IBM Backs Apache Spark For Big Data Analytics

Comment Now

IBM today announced support for the open source Apache Spark project, giving another boost to this increasingly popular in-memory data processing framework. Spark both complements and — in some cases — competes with the big data power of the better-known Apache Hadoop. As businesses continue to look for data-based approaches to their work, technology vendors are rushing to support promising open source projects, and to differentiate themselves from competitors. Can IBM gain credibility through association with Spark, and can the open source project benefit from Big Blue's experience, customer-base, and bank balance?

Spark began life in as a project at UC Berkeley in California, quickly delivering in-memory performance as much as 100 times that of the MapReduce framework that originally underpinned Apache Hadoop. Hadoop has moved on since then, to adopt other — faster and more flexible — ways of working. Spark has also progressed, promoting increasingly capable disk-based performance to complement its in-memory strengths, and establishing itself as a strong contender for use particularly in machine learning tasks. Spark moved to the Apache Software Foundation in 2013, becoming a top level project in 2014. In 2013, members of the original Berkeley team established the company now known as Databricks to build a business around Spark. The company launched with almost $14 million dollars from Andreessen Horowitz and others, and secured a further $33 million a year ago. And yet Spark is not without competitors of its own. Flink, which is also a top-level project of the Apache Software Foundation, has just recently begun to attract many of the same admiring comments directed Spark's way 12-18 months ago. Despite sound technical credentials, ongoing development, big investments, and today's high-profile endorsement from IBM, it would be unwise (and implausible) to crown Spark as the winner just yet.

IBM announced a number of initiatives today, aligning with what the company PR machine calls

"potentially the most significant open source project of the next decade.

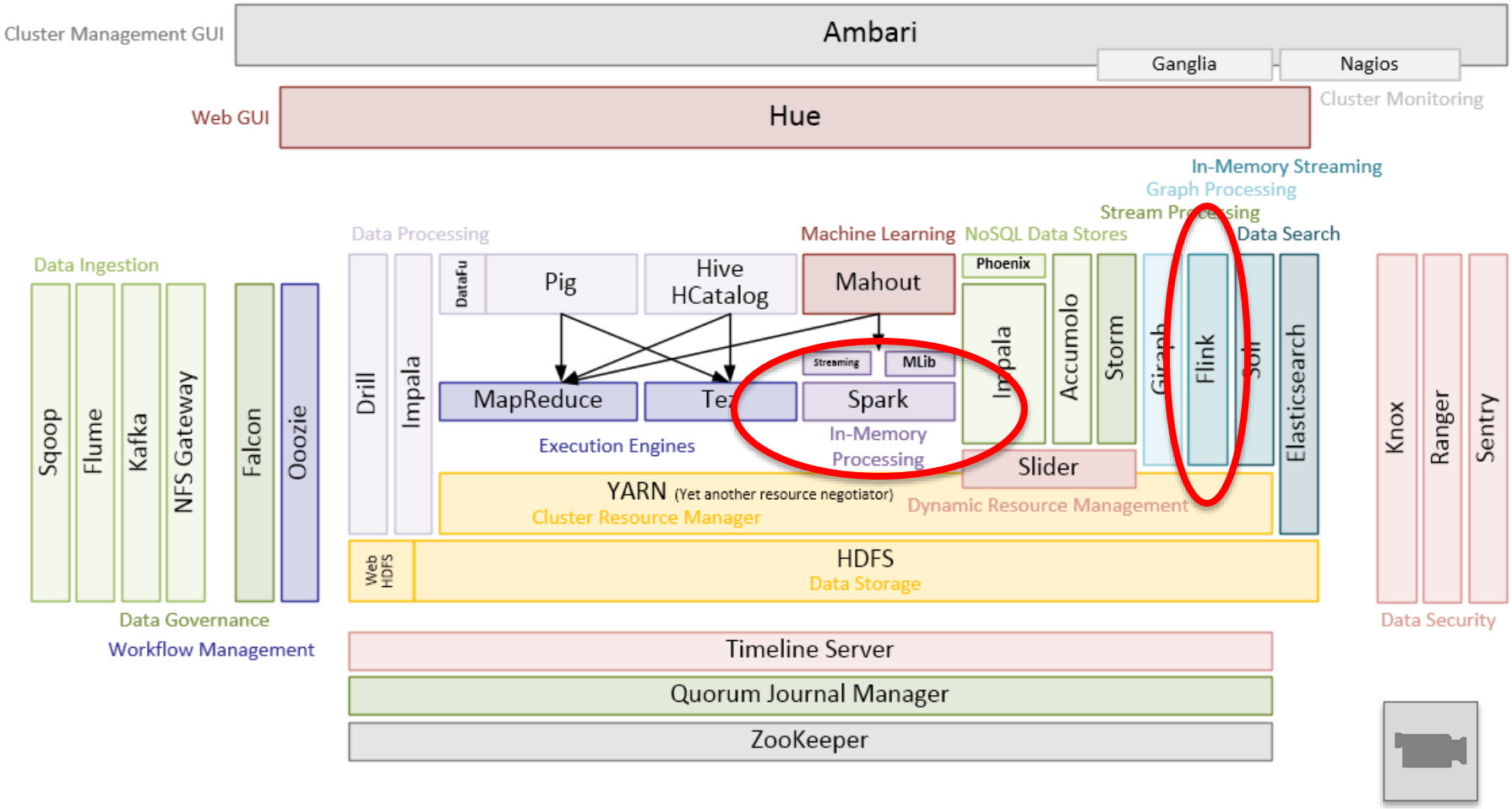
These include:

- deepening the integration between Apache Spark and existing IBM products

In-Memory „Big Data" Processing - Apache Spark
The Combination: The Flexibility of Spark on a Stable Hadoop Platform

Unlimited Scale	Ease of Development
In-Memory Performance	Enterprise Platform
Wide Range of Data Formats	Combine Workflows

In-Memory „Big Data“ Processing - Apache Spark
Die wichtigsten Bewohner des „Hadoop-Zoos”

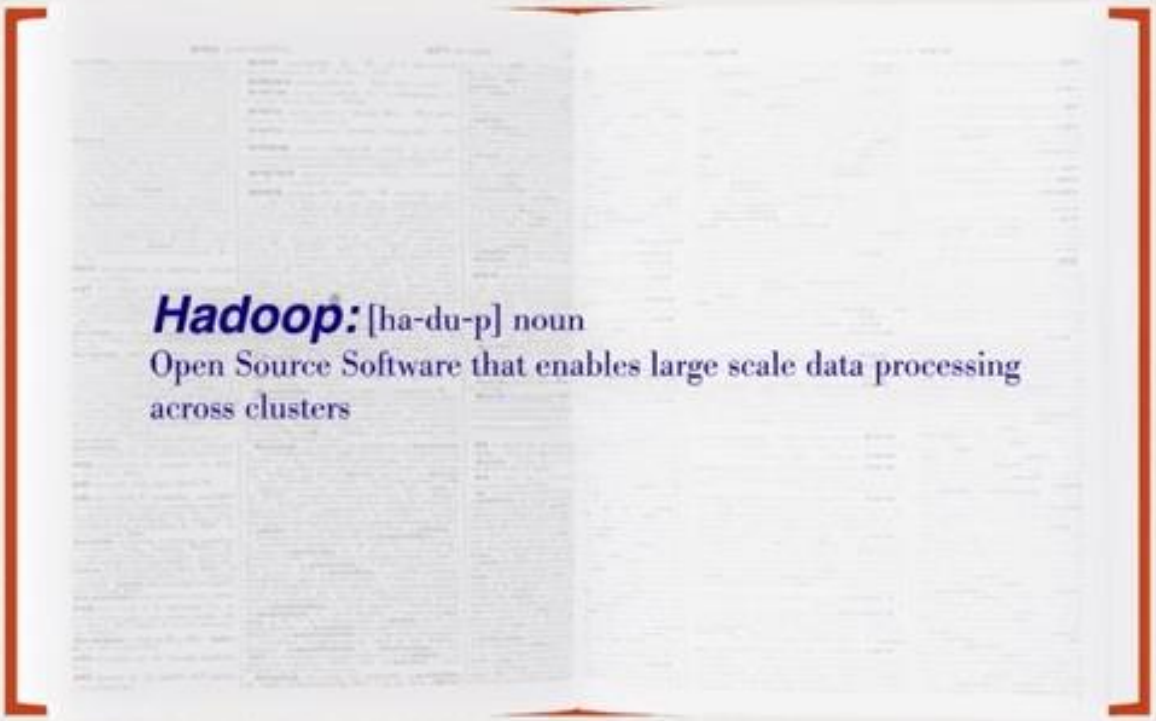


Quelle: Uwe Seiler: Zoo voller Gehege, in: iX Developer Big Data, 2/2015, S.37

In-Memory „Big Data“ Processing - Apache Spark

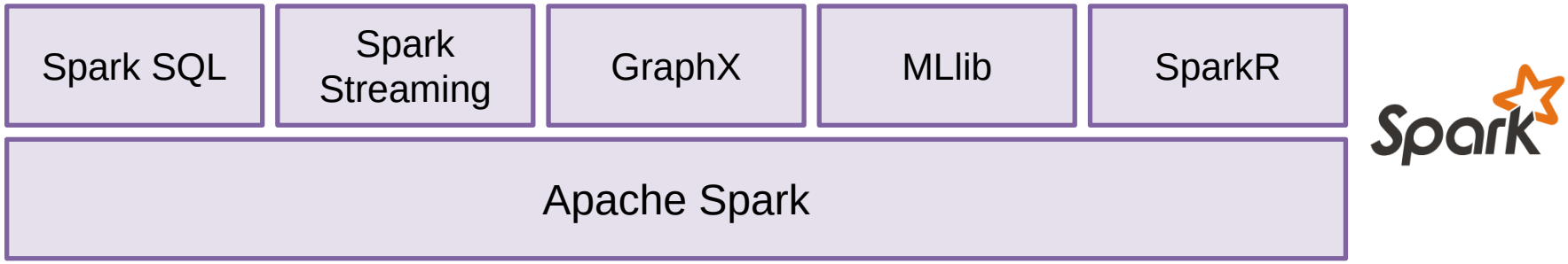
What’s the scoop with Hadoop?

Der Link zum Video findet sich im Anhang dieser Präsentation.

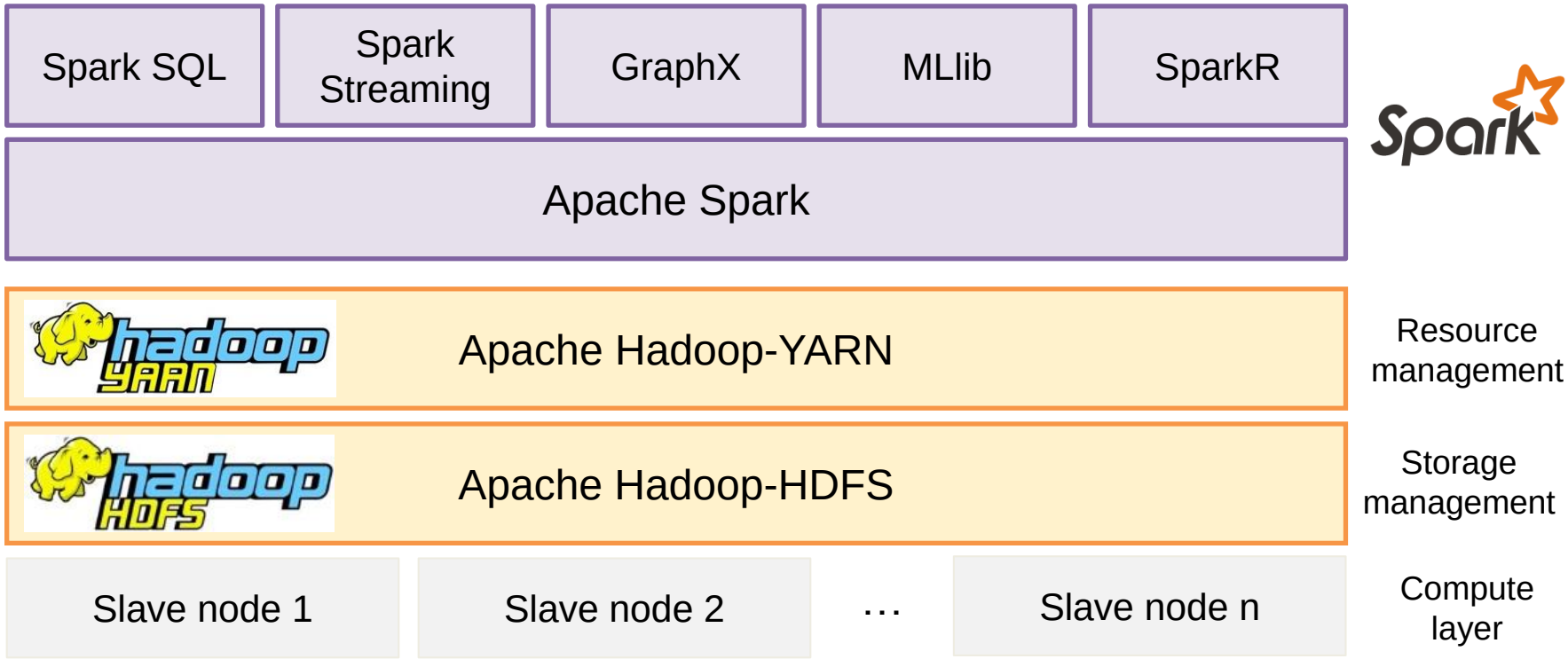


In-Memory „Big Data“ Processing - Apache Spark

Spark Libraries



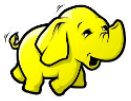
In-Memory „Big Data“ Processing - Apache Spark
Spark on Hadoop

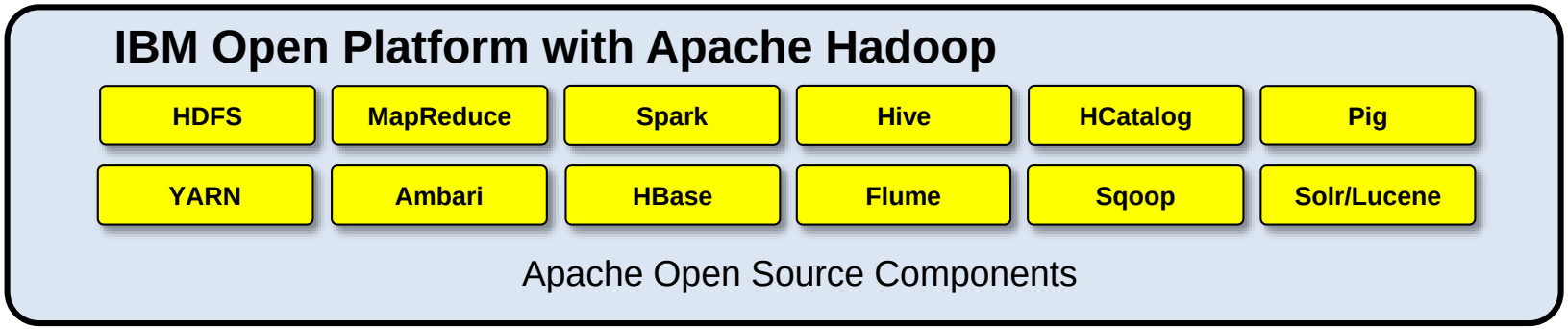


In-Memory „Big Data" Processing - Apache Spark
IBM Open Platform with Apache Hadoop

- 100% open source code
  - Commitment to currency: "days, not months"
  - Includes Spark

- Free for production use
  - Decoupled Apache Hadoop from IBM analytics and data science technologies
  - Production support offering available

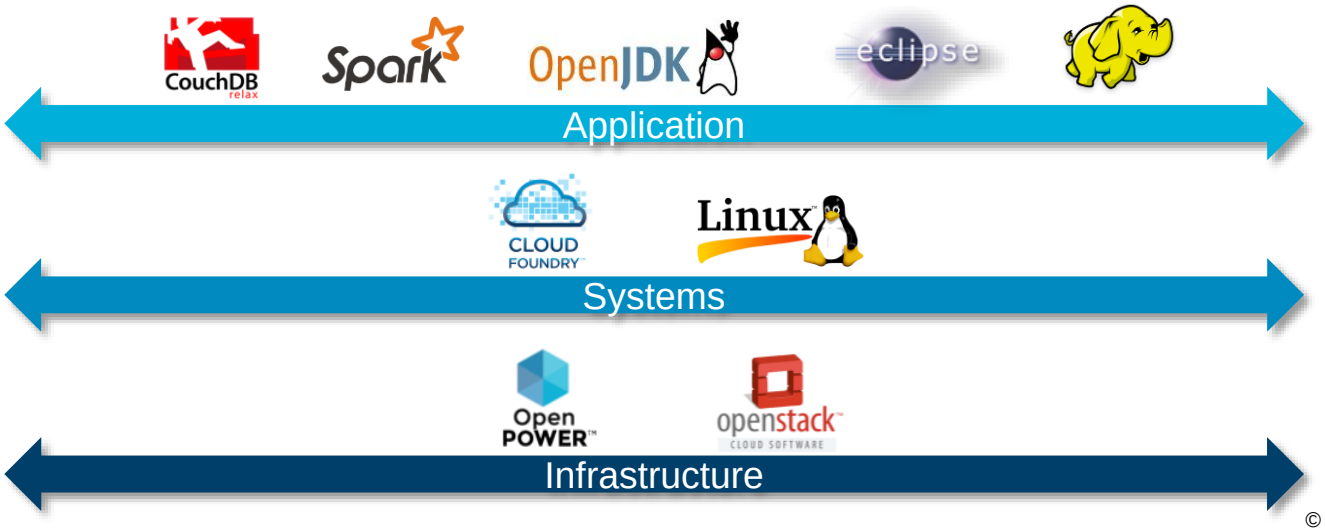




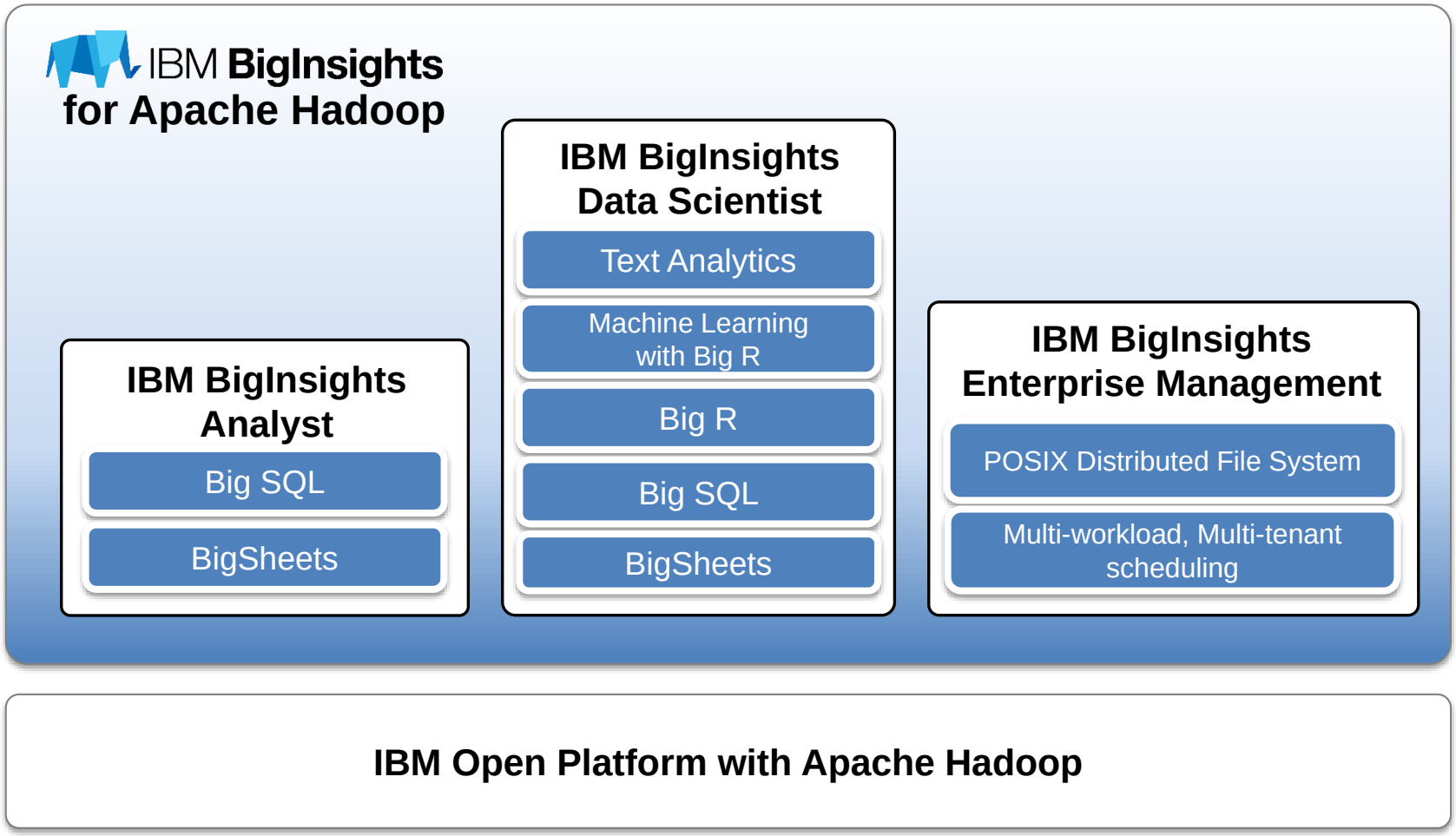
In-Memory „Big Data" Processing - Apache Spark
IBM is Committed to Open Source

- Open source technologies are the base for IBM software and solutions

- IBM's long history of deep open source commitment
  - **Apache Software Foundation**: Founding member in 1999
  - **Cloud Foundry**: #1 contributor; Basis for Bluemix
  - **OpenStack:** #4 contributor; Basis for IBM's IaaS
  - **Linux**: #3 contributor; IBM first enterprise backer of Linux
  - **Hadoop/Spark:** Extensive investment in open source contribution; Integration with Analytics software



In-Memory „Big Data" Processing - Apache Spark
IBM BigInsights for Apache Hadoop



SIDAP

Skalierbares Integrationskonzept zur Datenaggregation, -analyse, -aufbereitung von großen Datenmengen in der Prozessindustrie

Bundesministerium für Wirtschaft und Energie

Smart Data

Projekt Partner Teilprojekte Zusatzinfos

02 Partner

Das Projektkonsortium besteht aus international führenden Unternehmen der deutschen Prozessindustrie, Unternehmen der Informationstechnik, Sensor- und Aktortechnik sowie einem Forschungsinstitut. Unterstützt wird das Projekt desweiteren von großen Industrieverbänden. Innerhalb des Konsortiums wird unterschieden zwischen geförderten und assoziierten Partnern.

Geförderte Partner

Geförderte Partner erhalten Fördergelder vom Projektträger, typischerweise etwa 35% der Fördersumme. Im Gegenzug sind die Beteiligten dazu verpflichtet sich aktiv in das Projekt einzubringen und vereinbarte Leistungen zeitgerecht und eigenverantwortlich zu liefern.

Bayer Technology Services (BTS), Bayer AG

Die Bayer AG ist ein weltweit tätiges Unternehmen mit Kernkompetenzen auf den Gebieten Gesundheit, Agrarwirtschaft und hochwertigen Materialien. Die Bayer-Tochter Bayer Technology Services (BTS) ist ein führendes Technologieunternehmen und arbeitet an ganzheitlichen Technologien und Lösungen entlang des gesamten Lebenszyklus von Produkten, Prozessen und Anlagen der chemisch-pharmazeutischen Industrie. Die Arbeit des Konzerns umfasst die Entwicklung von Technologieplattformen sowie Innovationen, den bestmöglichen und kostengünstigsten Anlagenbau und die Unterstützung für einen sicheren und effizienten Anlagenbetrieb mit Serviceleistungen, Know-how und maßgeschneiderten Werkzeugen.

Science For A Better Life

IBM Deutschland GmbH

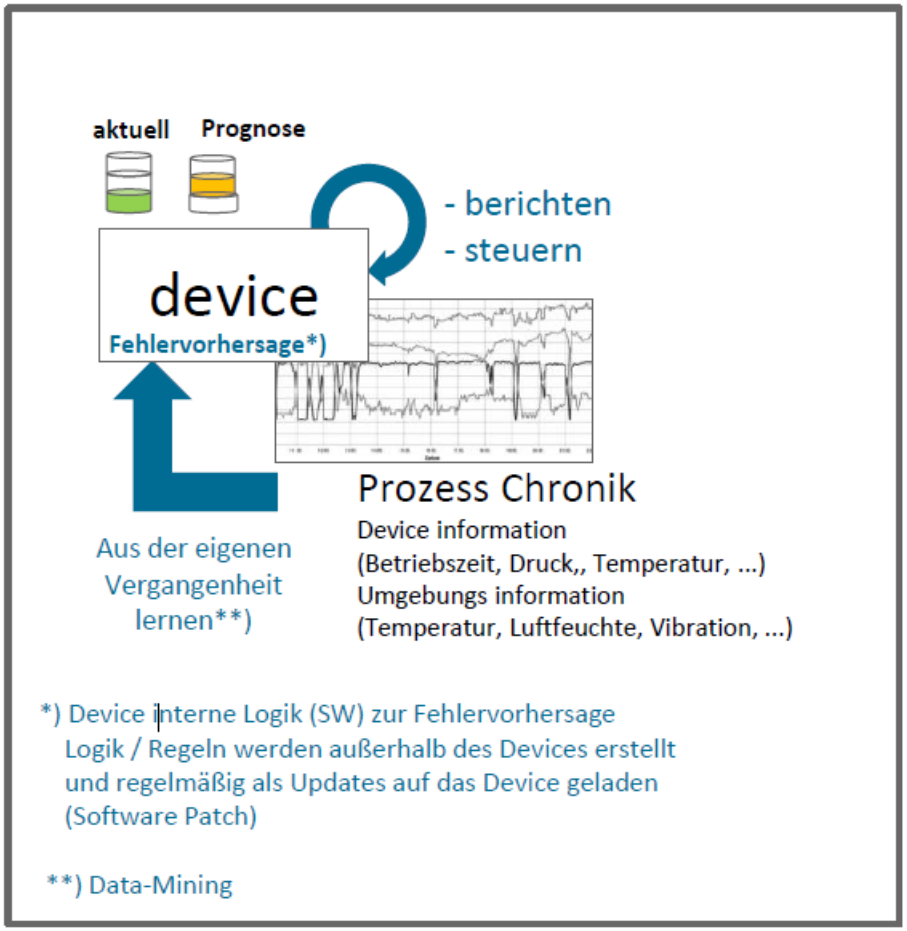
IBM gehört zu den weltweit größten Anbietern im Bereich Informationstechnologie (Hardware, Software und Services) und B2B-Lösungen. Das Lösungsportfolio der IBM reicht vom Supercomputer über Software und Dienstleistungen, inklusive Beratungsleistungen, bis zur Finanzierung. Mit der Entwicklung der unternehmensorientierten Big-Data-Plattform kann das gesamte Spektrum der geschäftlichen Herausforderungen im Bereich Big Data in Angriff genommen werden. Die Plattform vereint konventionelle Technologien, die sich gut für strukturierte Routineaufgaben mit neuen Technologien eignen, die auf hohe Geschwindigkeit und Flexibilität ausgerichtet sind und sich für die Ad-hoc-Datenuntersuchung, -erkennung und die unstrukturierte Analyse anbieten. IBM arbeitet im Rahmen von Industrie 4.0 an der Integrationsplattform für Engineeringwerkzeuge und stellt diese SIDAP zur Verfügung.

AIS, TU München

Der Lehrstuhl für Automatisierung und Informationstechnik (AIS) beschäftigt sich unter der Leitung von Frau Professor Vogel-Heuser seit über zehn Jahren mit dem gesamten Engineering Lebenszyklus im

Link: http://www.sidap.de

Exkurs: SIDAP
Echtzeit-Betriebsanalyse – Betrieb auf Basis globaler Erfahrung



Device: Gerät, Apparat, Einrichtung, Verfahren, Einheit

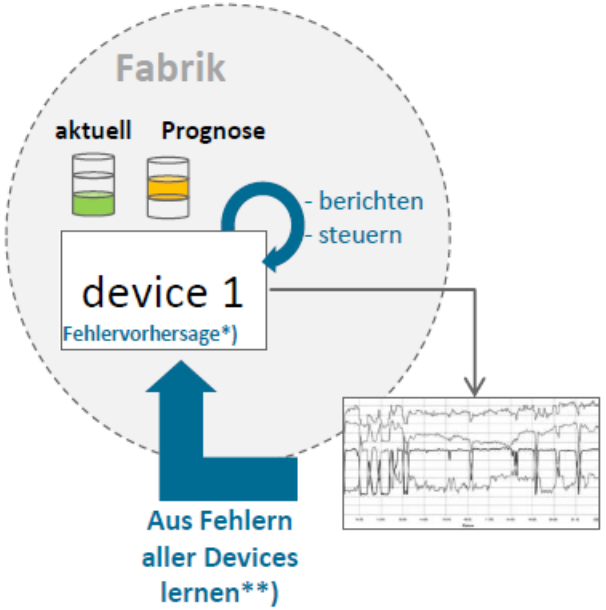
Otto von Bismarck, Winston Churchill:

„Der Weise lernt von den Fehlern der Anderen, der Dumme aus den Eigenen”

Quelle: Dr. T. Pötter, Dr. M. Steffen (beide Bayer Technology Services)

Exkurs: SIDAP
Echtzeit-Betriebsanalyse – Betrieb auf Basis globaler Erfahrung

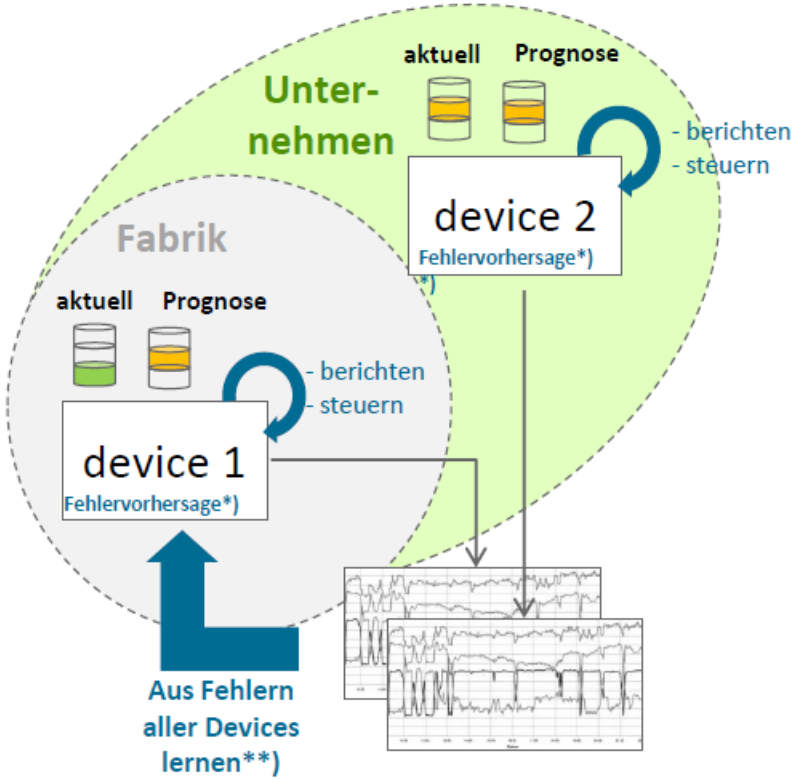
“Dumm”

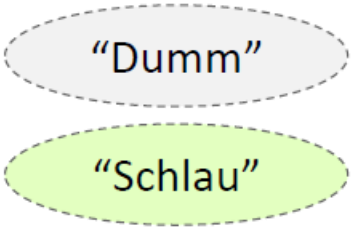


*) Device interne Logik (SW) zur Fehlervorhersage .. Logik / Regeln werden außerhalb des Devices erstellt und regelmäßig als Updates auf das Device geladen . (Software Patch)
**) Data-Mining und Big-Data Techniken

Quelle: Dr. T. Pötter, Dr. M. Steffen (beide Bayer Technology Services)

Exkurs: SIDAP
Echtzeit-Betriebsanalyse – Betrieb auf Basis globaler Erfahrung



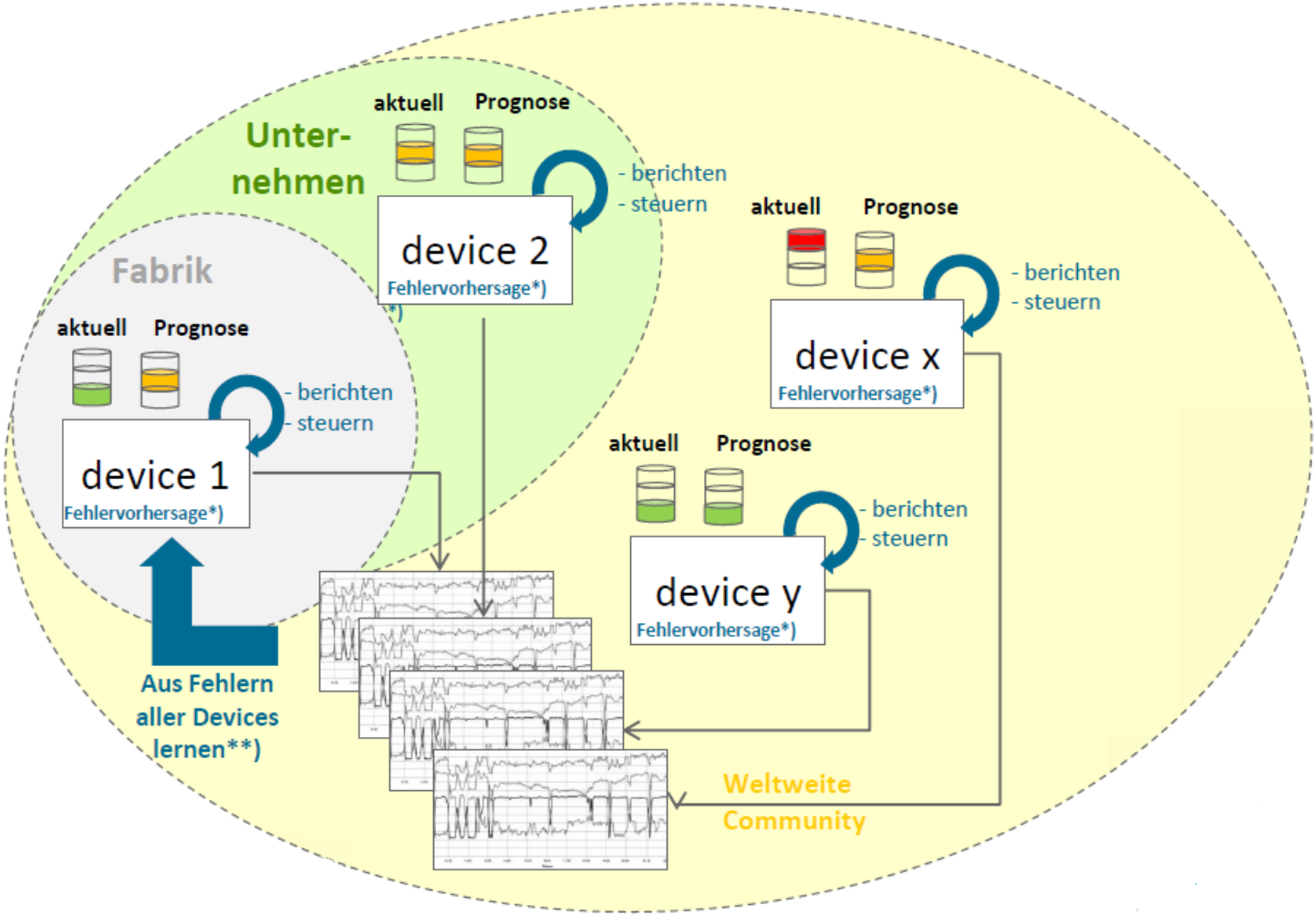


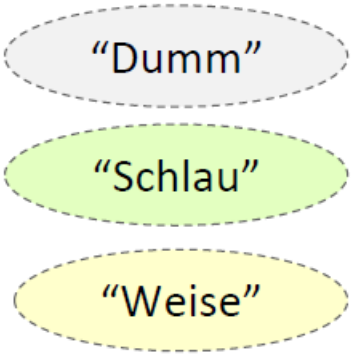
*) Device interne Logik (SW) zur Fehlervorhersage .. Logik / Regeln werden außerhalb des Devices erstellt und regelmäßig als Updates auf das Device geladen . (Software Patch)

**) Data-Mining und Big-Data Techniken

Quelle: Dr. T. Pötter, Dr. M. Steffen (beide Bayer Technology Services)

Exkurs: SIDAP
Echtzeit-Betriebsanalyse – Betrieb auf Basis globaler Erfahrung





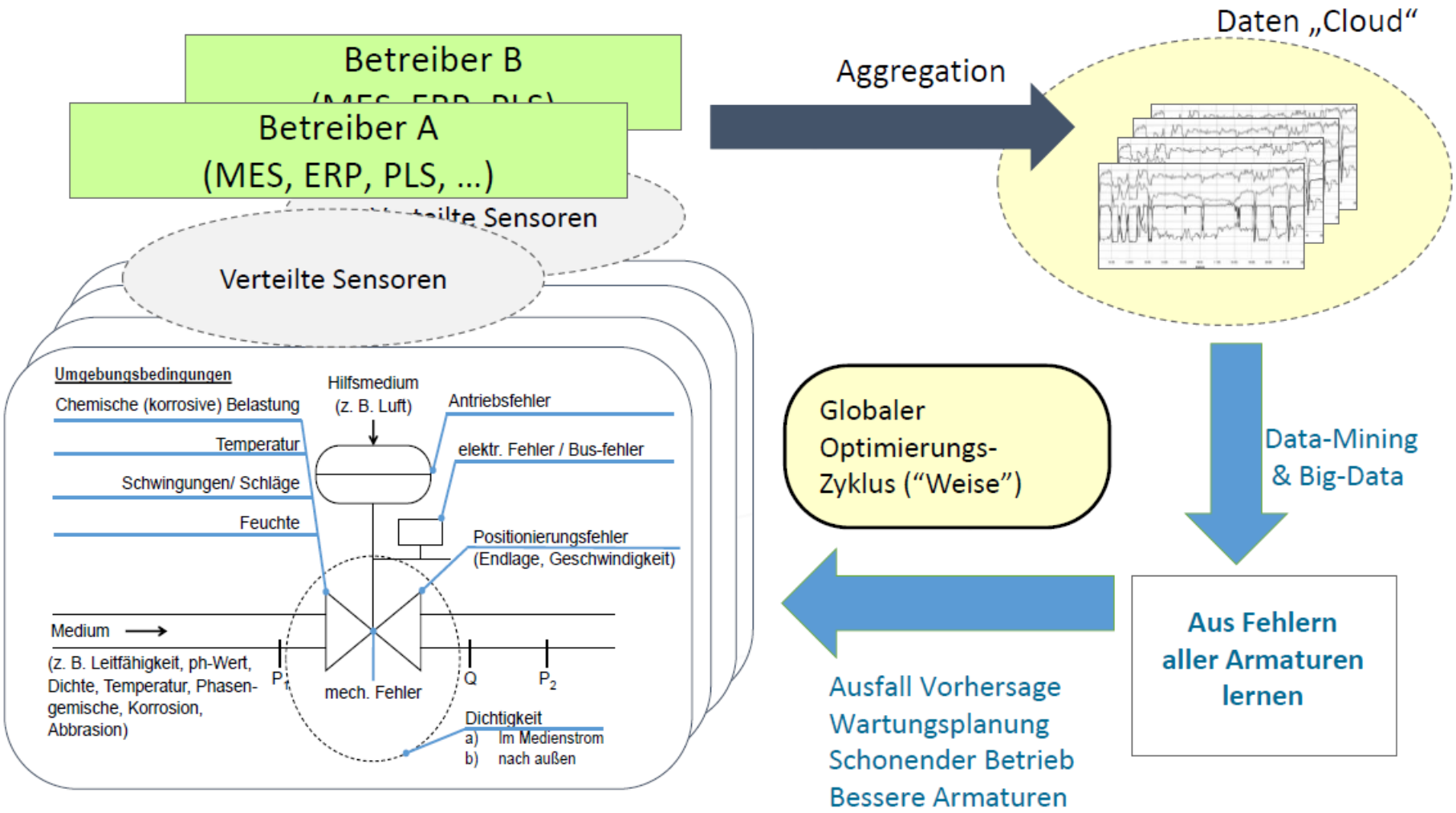
Die Qualität der Vorhersagen von Fehlern wächst mit der Anzahl der Einträge in die Datenbasis.

*) Device interne Logik (SW) zur Fehlervorhersage .. Logik / Regeln werden außerhalb des Devices erstellt und regelmäßig als Updates auf das Device geladen . (Software Patch)

**) Data-Mining und Big-Data Techniken

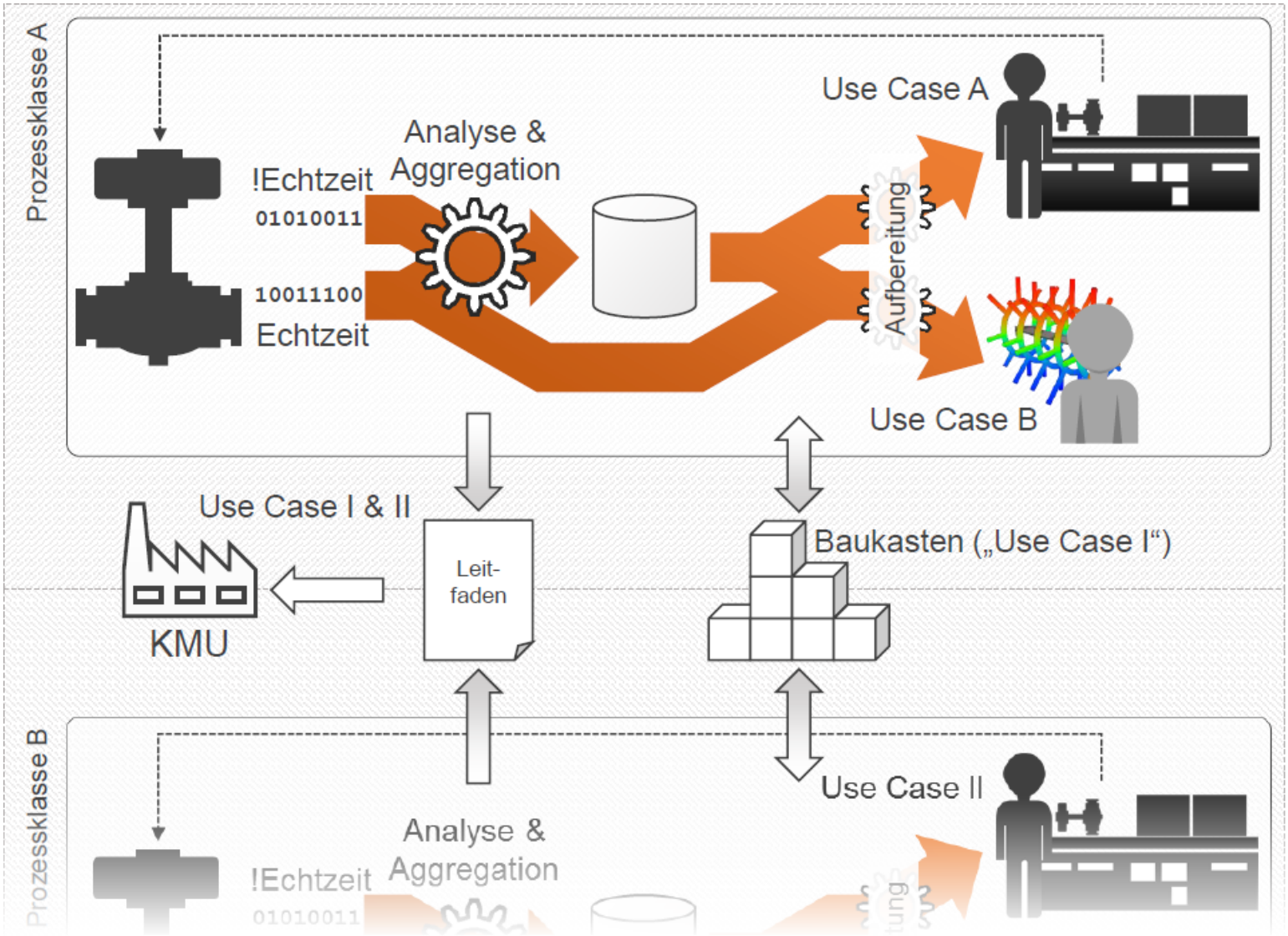
Quelle: Dr. T. Pötter, Dr. M. Steffen (beide Bayer Technology Services)

Exkurs: SIDAP
Datenaggregation im Rahmen von SIDAP

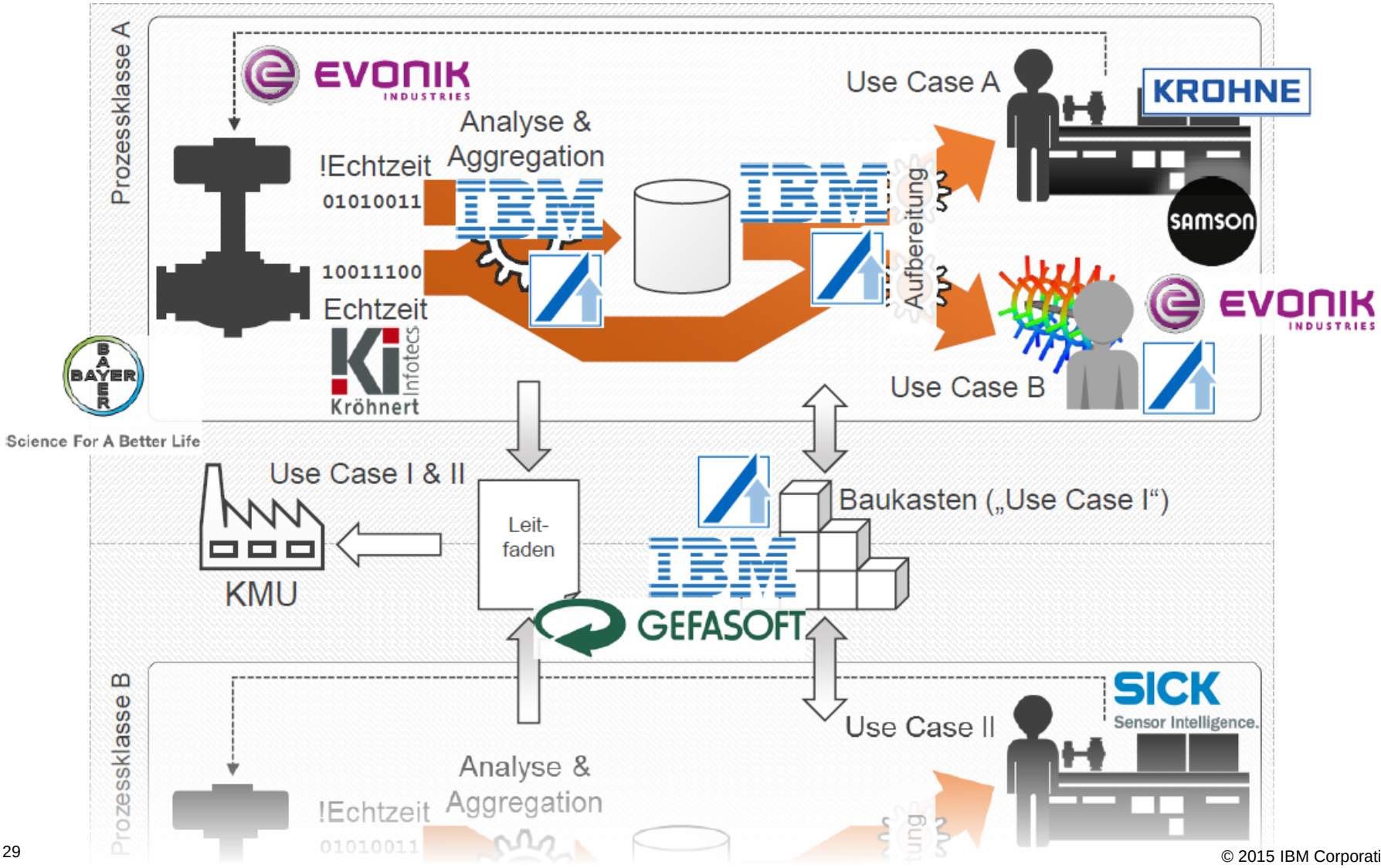


Quelle: Dr. T. Pötter, Dr. M. Steffen (beide Bayer Technology Services)

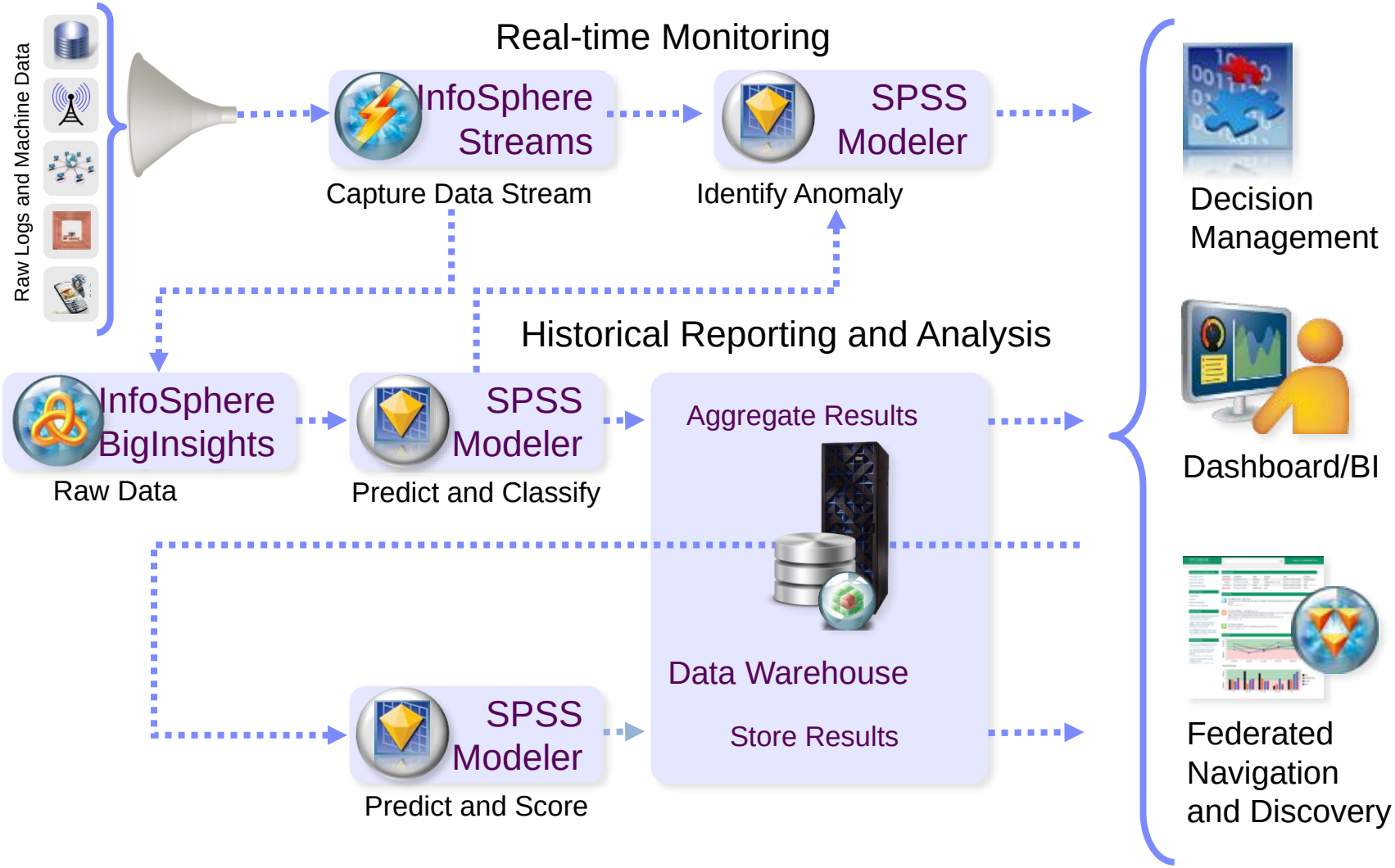
Exkurs: SIDAP
Use-Cases



Exkurs: SIDAP
Use-Cases und Aufgabenbereiche der Partner



Exkurs: SIDAP
Echtzeit-Betriebsanalyse (Operations Analysis)



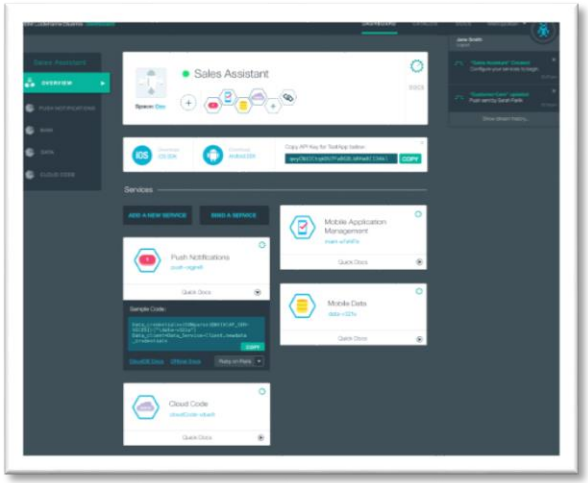
Exkurs: Internet of Things
IBM’s Internet of Things Foundation

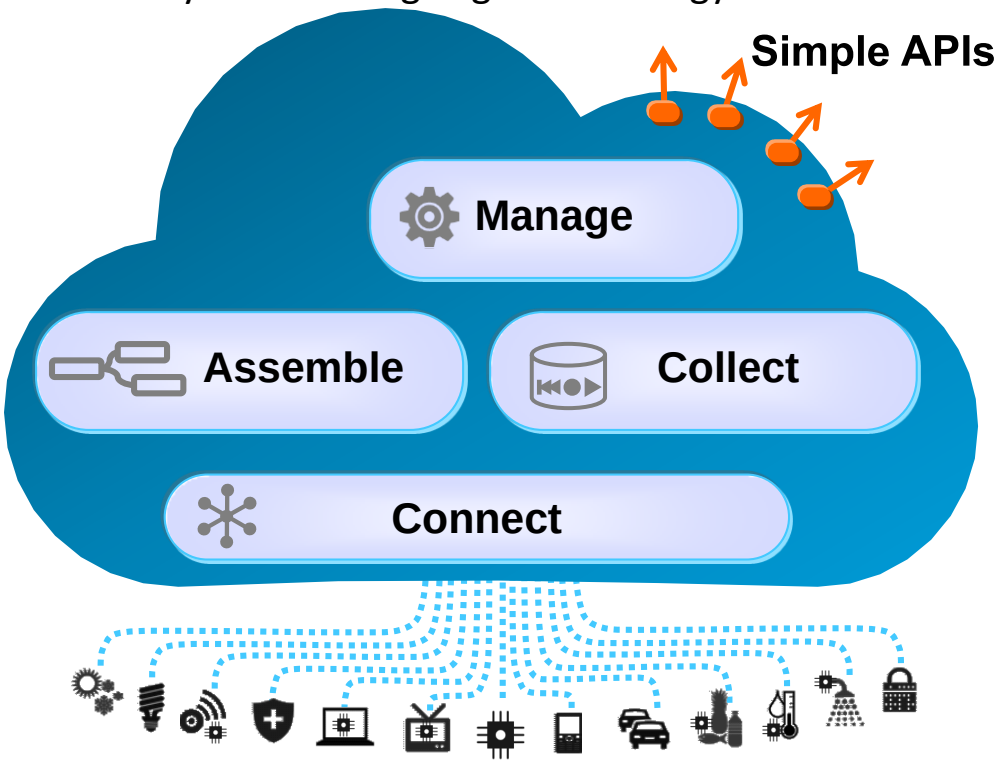
- Secure Device Registration
- Scalable Device Connectivity
- Historian
- Visual wiring
- PAYG SaaS pricing
- Powered by IBM MessageSight technology

IBM Bluemix™
www.bluemix.net

Simple APIs

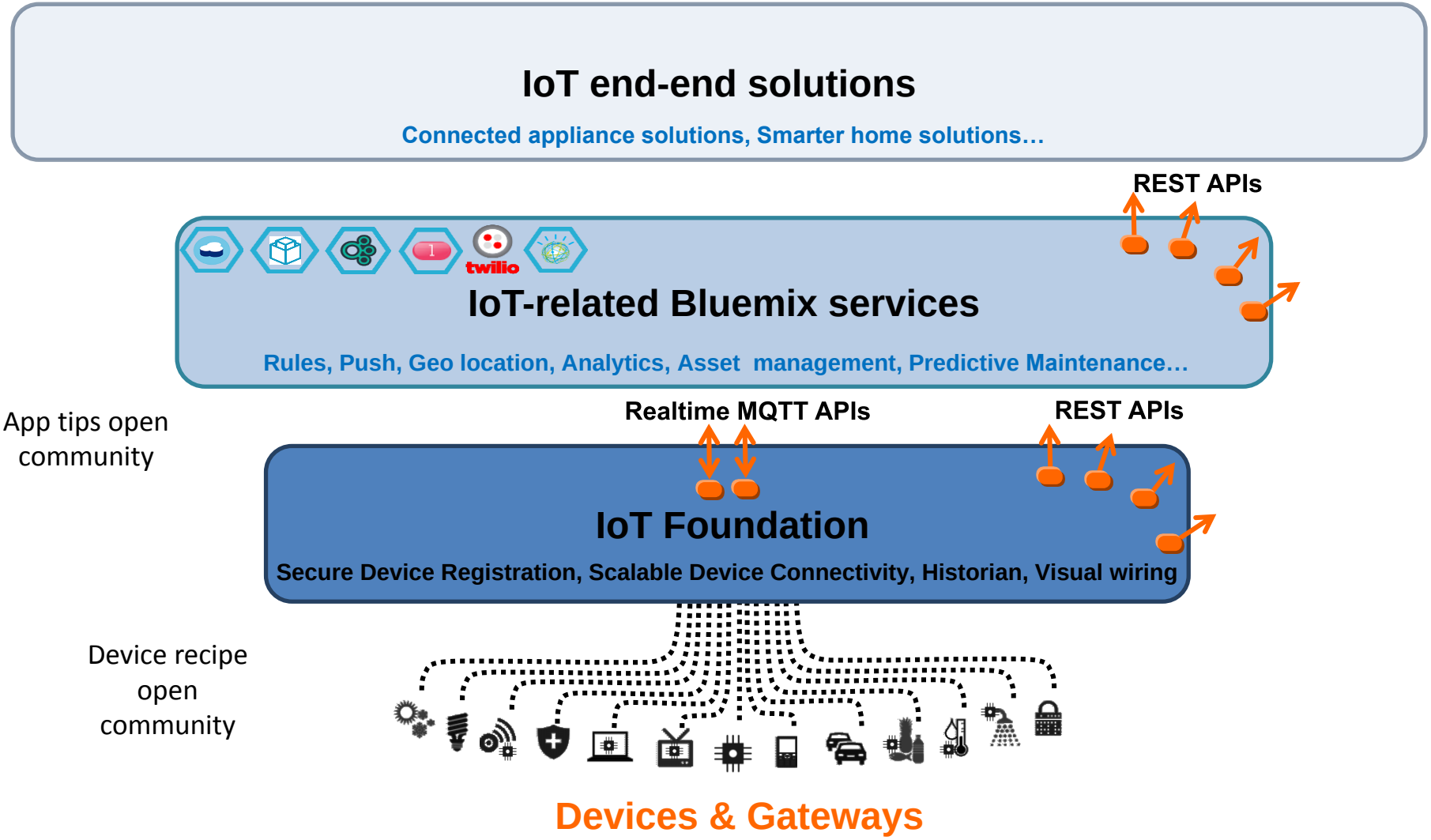
Compose





Mobile & Internet of Things

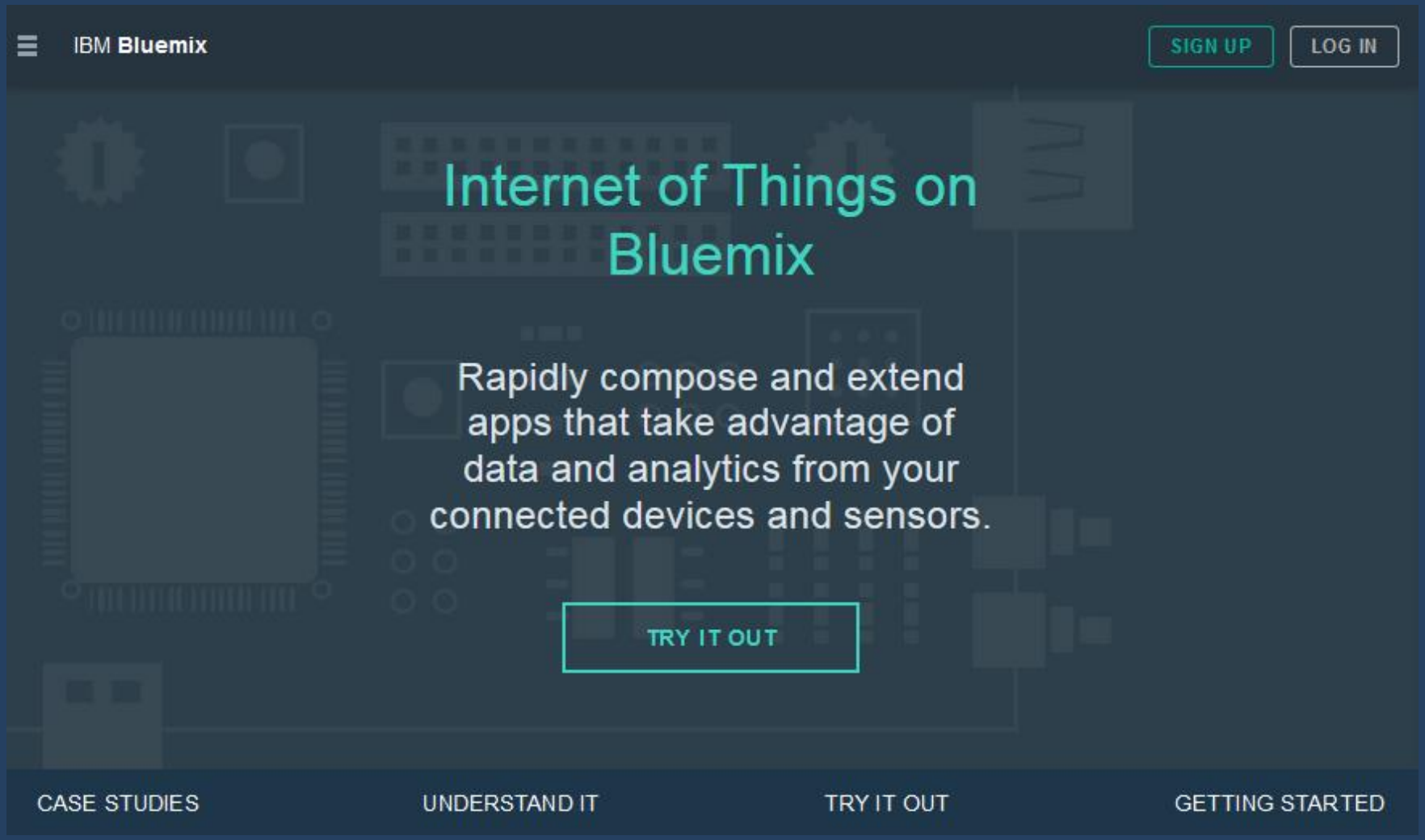
Exkurs: Internet of Things
IoT becomes a Composable Business



Exkurs: Internet of Things
IBM Bluemix – How to build a smarter App

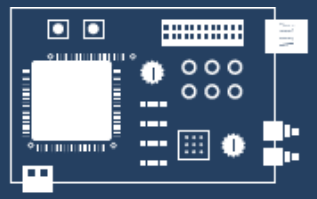
Der Link zum Video findet sich im Anhang dieser Präsentation.

Exkurs: Internet of Things
IBM Internet of Things Foundation



Exkurs: Internet of Things

IBM Internet of Things Foundation – Status Quo

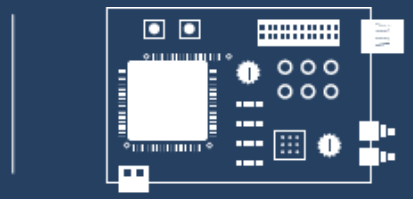


Your device or gateway

We start with your device, be it a sensor, a gateway or something else.
To find out how to get it connected, search our recipes.

Exkurs: Internet of Things

IBM Internet of Things Foundation – Status Quo



Your device or gateway

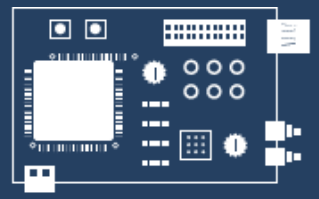
We start with your device, be it a sensor, a gateway or something else.
To find out how to get it connected, search our recipes.

MQTT

Your device data is sent securely up to the cloud using the open, lightweight MQTT messaging protocol.

Exkurs: Internet of Things

IBM Internet of Things Foundation – Status Quo



Your device or gateway

We start with your device, be it a sensor, a gateway or something else. To find out how to get it connected, search our recipes.

MQTT

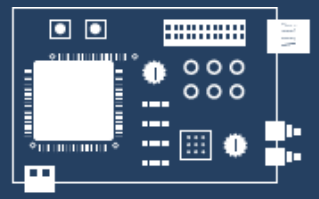
Your device data is sent securely up to the cloud using the open, lightweight MQTT messaging protocol.

IBM Internet of Things Foundation

This is the hub of all things IBM IoT. This is where you can setup and manage your connected devices so that your apps can access their live and historical data.

Exkurs: Internet of Things

IBM Internet of Things Foundation – Status Quo



Your device or gateway

We start with your device, be it a sensor, a gateway or something else. To find out how to get it connected, search our recipes.

MQTT

Your device data is sent securely up to the cloud using the open, lightweight MQTT messaging protocol.

IBM Internet of Things Foundation

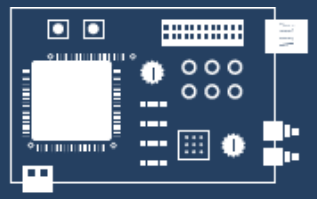
This is the hub of all things IBM IoT. This is where you can setup and manage your connected devices so that your apps can access their live and historical data.

REST & Real-time APIs

Use our secure APIs to connect your apps with the data coming from your devices.

Exkurs: Internet of Things

IBM Internet of Things Foundation – Status Quo



Your device or gateway

We start with your device, be it a sensor, a gateway or something else.
To find out how to get it connected, search our recipes.

MQTT

Your device data is sent securely up to the cloud using the open, lightweight MQTT messaging protocol.

IBM Internet of Things Foundation

This is the hub of all things IBM IoT. This is where you can setup and manage your connected devices so that your apps can access their live and historical data.

REST & Real-time APIs

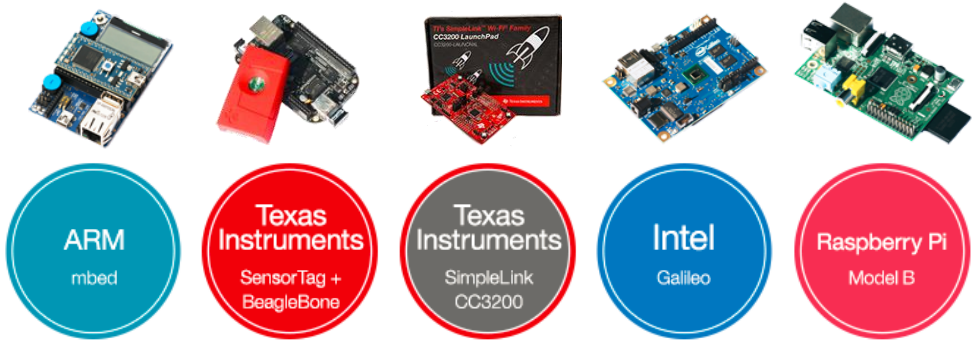
Use our secure APIs to connect your apps with the data coming from your devices.

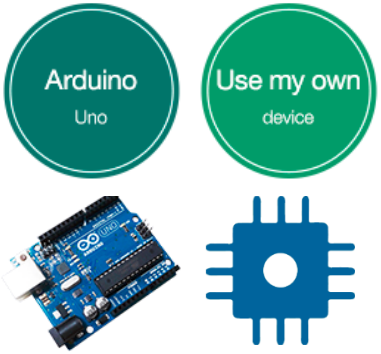
Your application and analytics

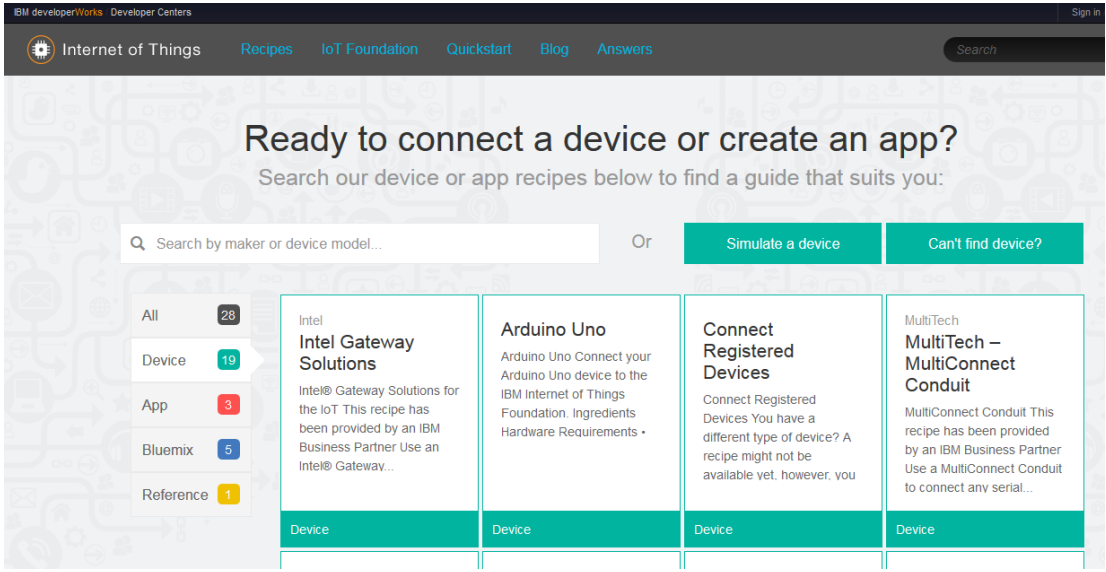
Create applications within IBM Bluemix, another cloud, or your own servers to interpret the data you now have access to!

Exkurs: Internet of Things
Device Recipes make it faster

Connect

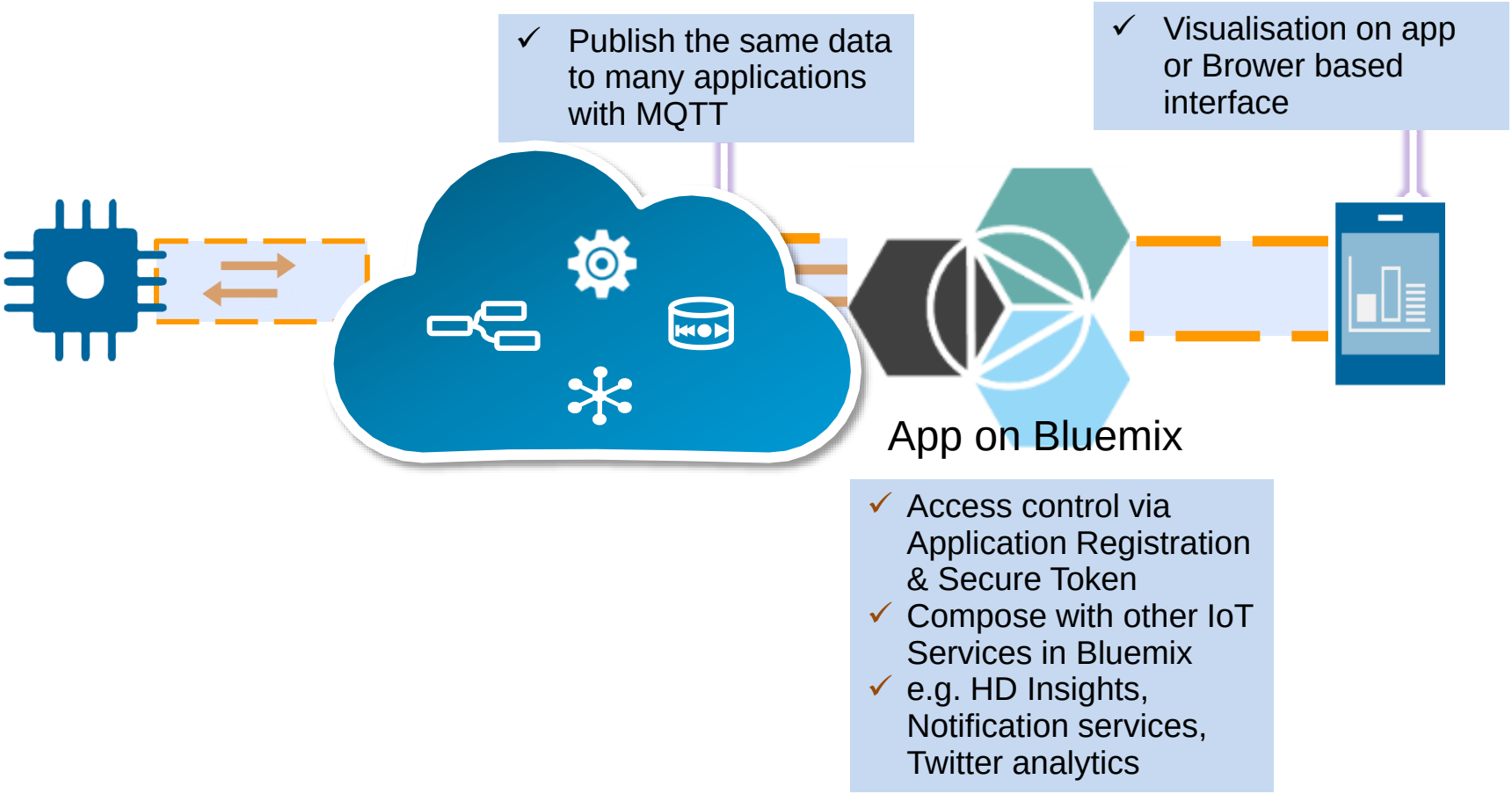




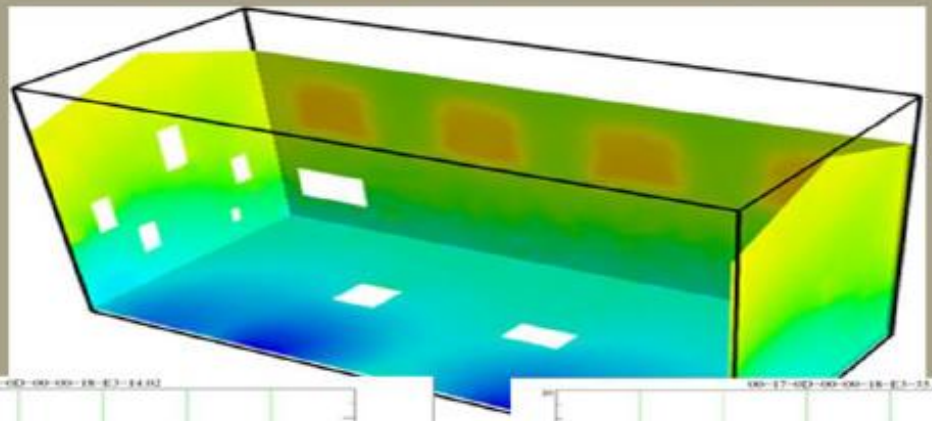


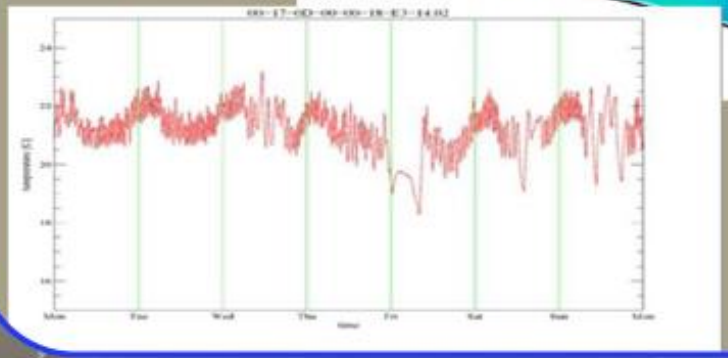
Exkurs: Internet of Things
Application Connection

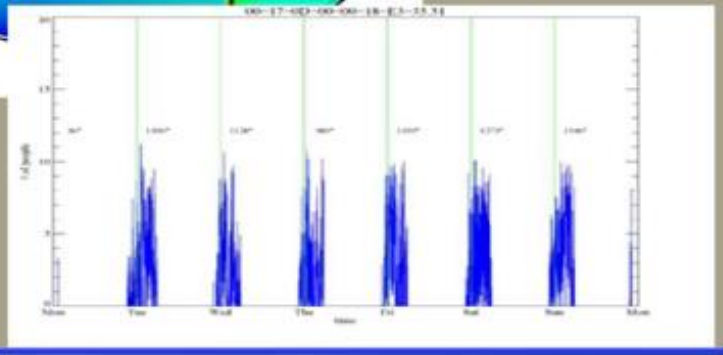
Connect



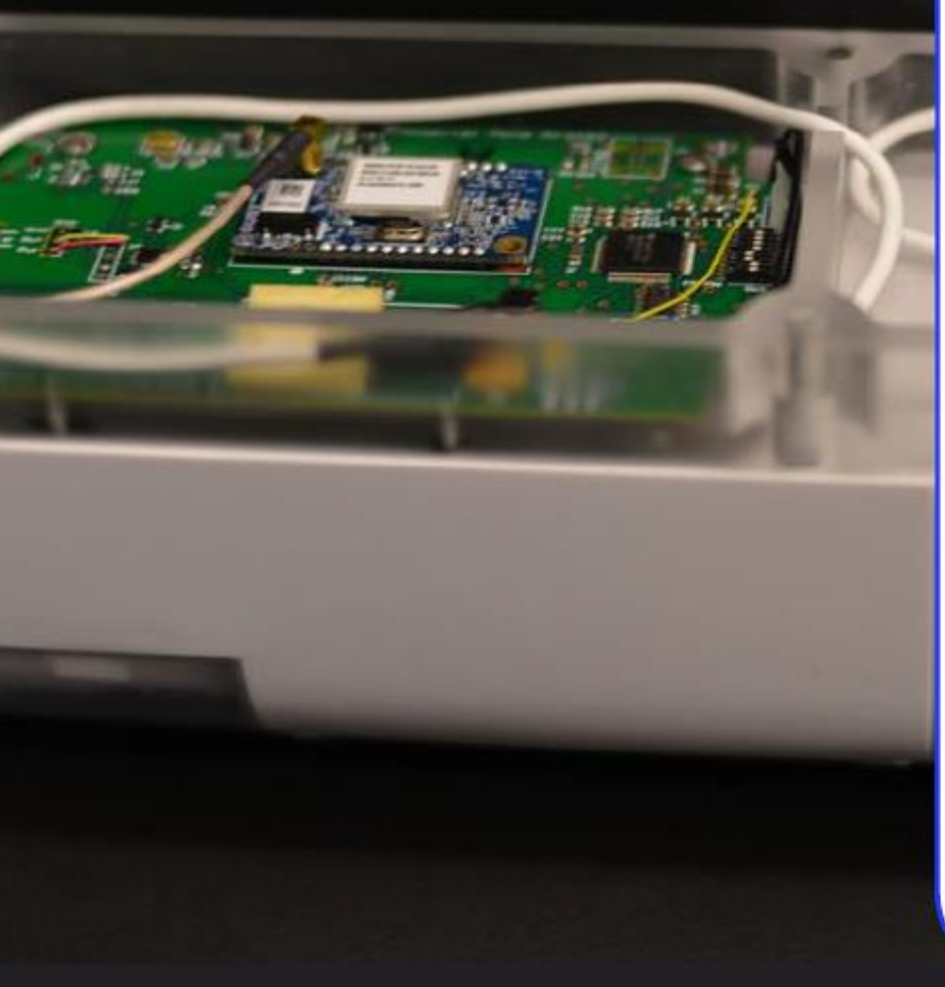
Exkurs: Internet of Things Beispiel

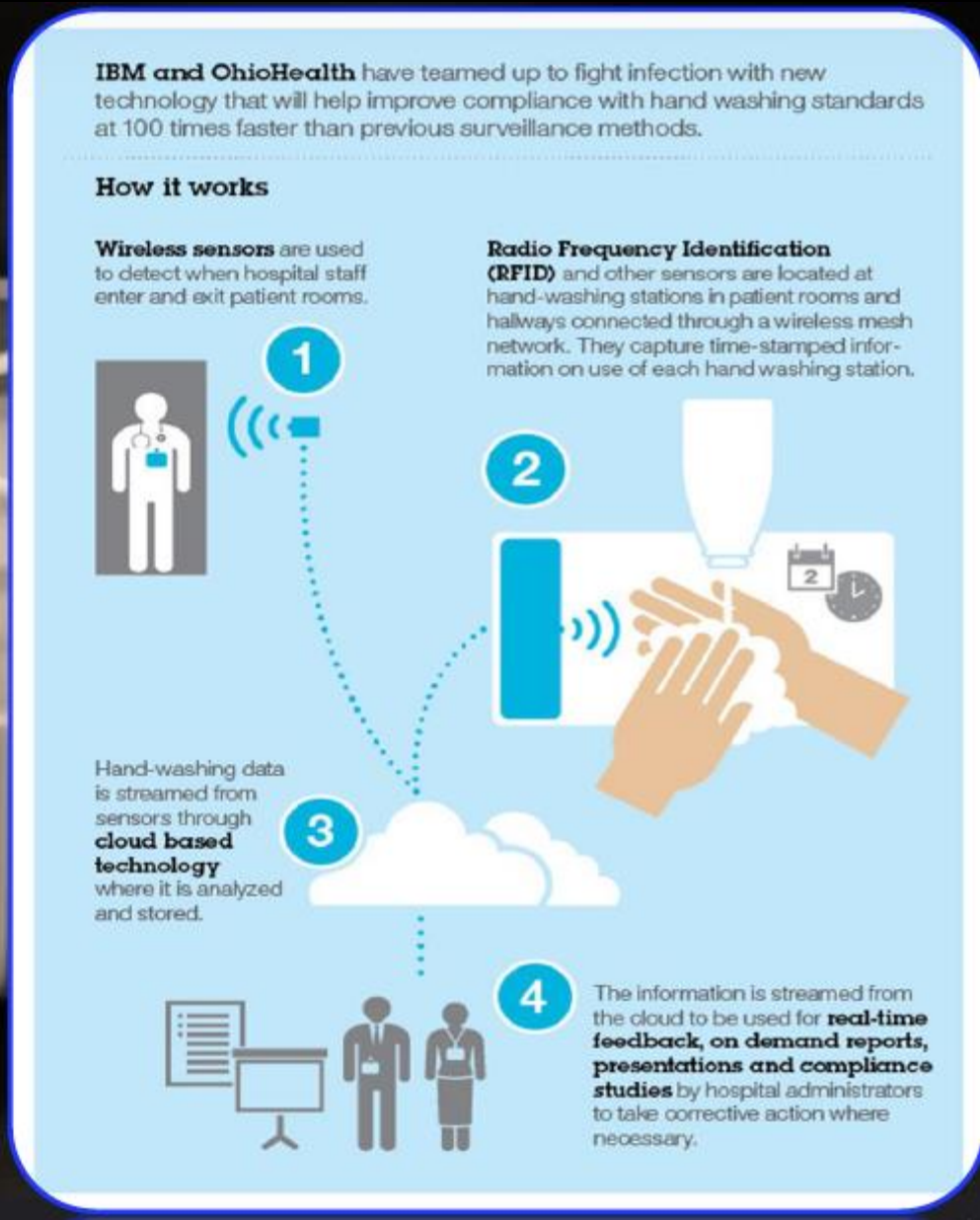




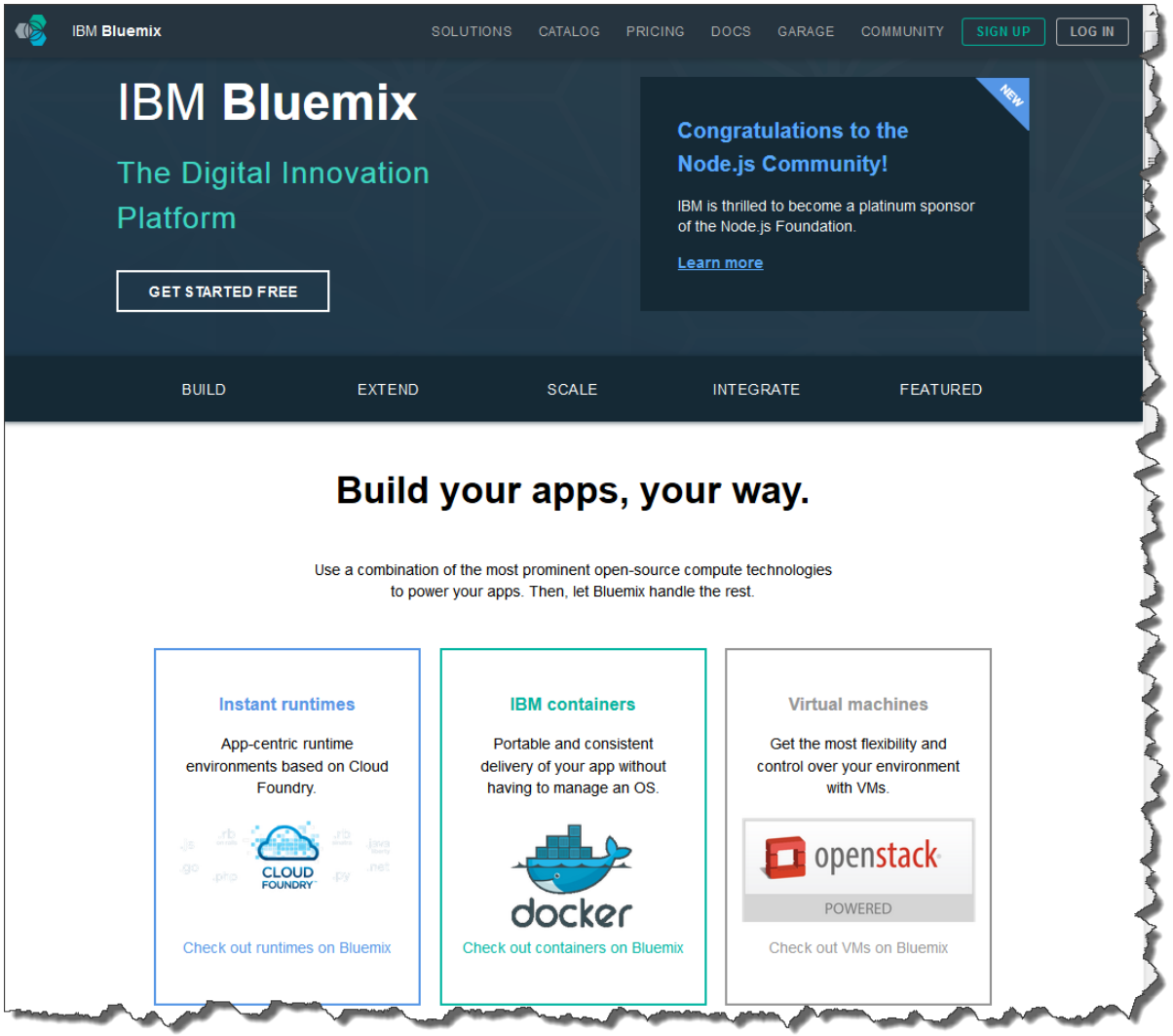


Exkurs: Internet of Things
Beispiel



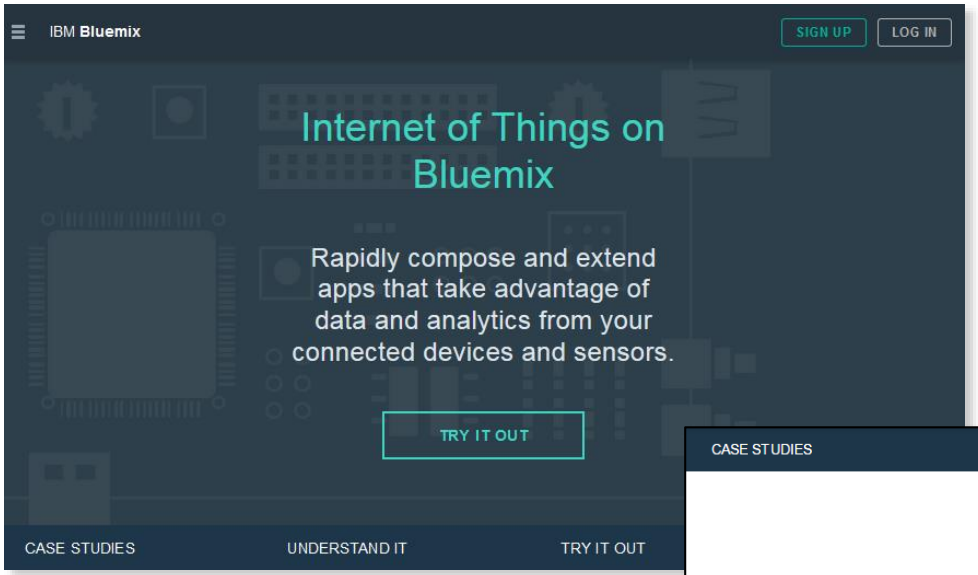


Empfehlung Ia
IBM Bluemix

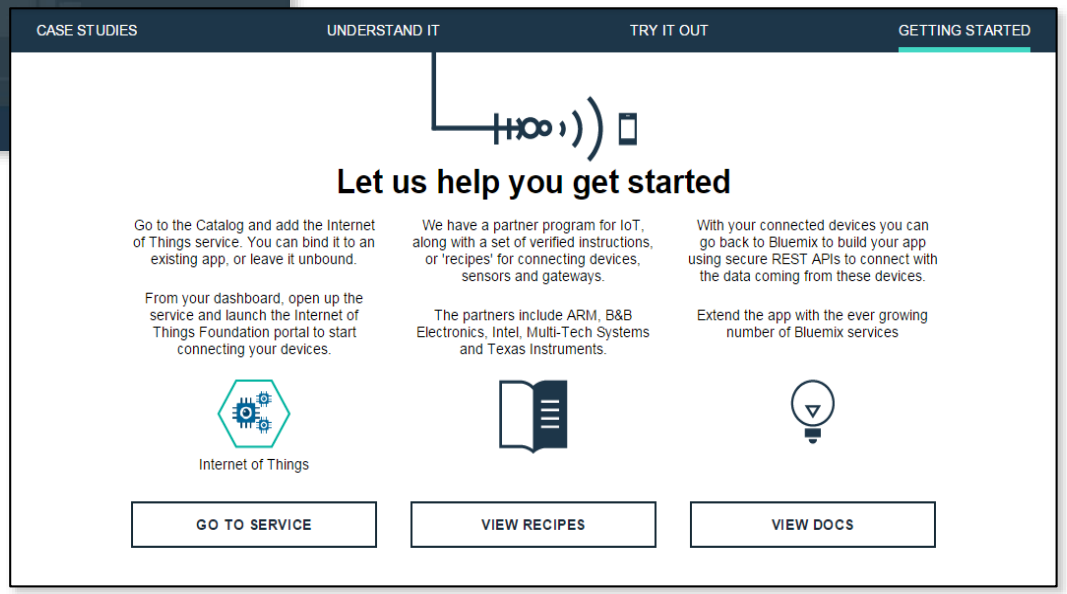


Link:
http://www.ibm.com/bluemix

Empfehlung Ib
IBM Internet of Things Zone auf Bluemix



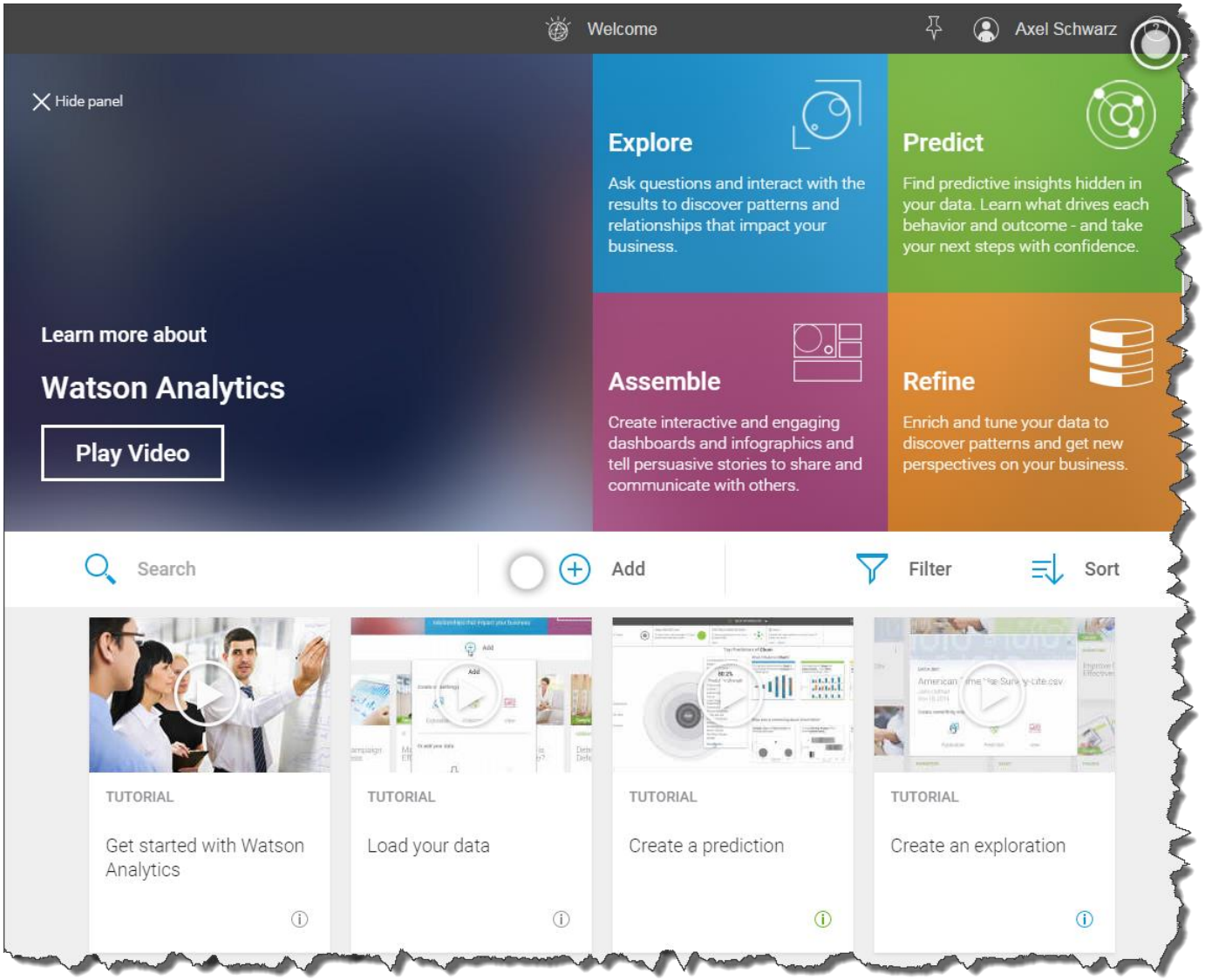
IBM **Internet of Things Foundation**



IoT Zone in Bluemix
[ibm.biz/try_iot](ibm.biz/try_iot)

oder
https://bluemix.net/solutions/iot

Empfehlung II
IBM Watson Analytics



Link:
http://www.ibm.com/analytics/watson-analytics/

Empfehlung III
„Big Data" University

Link:
http://bigdatauniversity.com/

Dipl. Wirtsch.-Ing.
Axel J. Schwarz
Mobik: +49 171 5619419
E-Mail: axel.j.schwarz@de.ibm.com

Software Client Architect

Vielen Dank!

Quellen und Referenzen

Literatur/Blog-Beiträge

- Ryan Baxter: Bluemix and the Internet of Things
  http://ryanjbaxter.com/2014/07/16/bluemix-and-the-internet-of-things/
- Bernard Marr: Spark Or Hadoop -- Which Is The Best Big Data Framework?
  http://www.forbes.com/sites/bernardmarr/2015/06/22/spark-or-hadoop-which-is-the-best-big-data-framework/2/
- Paul Miller - IBM Backs Apache Spark For Big Data Analytics
  http://www.forbes.com/sites/paulmiller/2015/06/15/ibm-backs-apache-spark-for-big-data-analytics/

Videos

- IBM IBM Big Data: How it works - https://www.youtube.com/watch?v=u5jWC89xBzI
- What’s the Scoop with Hadoop? - https://www.youtube.com/watch?v=BcGqiJXcDo8
- How to build a smarter app - https://www.youtube.com/watch?v=E5TKNbWl2Co
- Learn how to create a mobile app quickly using IBM BlueMix! - https://www.youtube.com/watch?v=DbXN2mz90b0

Links

- IBM Emerging Technology - Node-RED: http://nodered.org
- IBM developerWorks - IoT: https://www.ibmdw.net/iot/
- IBM developerWorks - Build a cloud-ready temperature sensor with the Arduino Uno and the IBM IoT Foundation:
  http://www.ibm.com/developerworks/cloud/library/cl-bluemix-arduino-iot1/index.html Bluemi

Bildquellen

- IBM sowie lizenzfreie digitale Bilder der Medien-Bibliotheken: 123RF (http://www.123rf.com) und Stock.XCHNG (http:///www.sxc.hu).