Big Data im Internet der Bosch-Dinge und -Dienste

Lothar Baum[1], Wolfgang Blochinger[2], Michael Peters[3], Jörg Sommer[4]

[1] Corporate Research, [2-4] Corporate Sector Information Systems and Services
Robert Bosch GmbH
Postfach 30 02 20, 70442 Stuttgart
{Lothar.Baum | Wolfgang.Blochinger | Michael.Peters | Joerg.Sommer}@de.bosch.com

1 Motivation und Anwendungen bei Bosch

Bosch ist heute noch überwiegend ein Hersteller von *Dingen* der materiellen Welt. Was 1886 mit einfachen elektrotechnischen Apparaten begann, erstreckt sich heute über verschiedenste Anwendungsdomänen und Geschäftsbereiche. Kfz-Komponenten, Heizungsanlagen, Haushaltsgeräte und Industriesteuerungen gehören ebenso zum Produktportfolio wie Sicherheitssysteme, Beschallungsanlagen oder auch medizintechnische Produkte. Gut drei Viertel aller von Bosch hergestellten Produkte werden bereits heute bzw. in naher Zukunft durch embedded Software gesteuert. Dies erlaubt nicht nur eine effizientere und flexiblere Steuerung, sondern auch eine Vernetzung dieser Systeme. Auf diese Weise ist es erstmals möglich, genaue Informationen über Einsatz und *Leben* der Produkte im Feld zu sammeln und diese Informationen zur Produktoptimierung oder auch für neue Dienstleistungen zu nutzen. Bosch hat die Chancen des *Internets der Dinge und Dienste* früh erkannt und stellt sich in vielen Bereichen entsprechend strategisch auf.

2 Big Data und Data Mining

Ein Internet der Dinge und Dienste eröffnet den Zugang zu einer dramatisch wachsenden Zahl von Datenquellen – im Jahr 2020 werden 50 Milliarden [Ba13] *Connected Devices* weltweit aktiv sein, die potenziell Sensor- und Zustandsinformationen verfügbar machen. Auch wenn in konkreten Anwendungen nur ein Teil dieser Datenquellen zur Verfügung steht oder genutzt werden kann bzw. relevant ist, so sind die Datenmengen dennoch beachtlich. So generieren alleine die Logging-Informationen der ESP-Steuergeräte von ca. 400 000 BMW 5er der E60-Baureihe jährlich mehrere Gigabytes. Bei der Überwachung von Fertigungsparametern fallen in Bosch-Werken für einzelne Produktgruppen Daten in der Größenordnung von 50 Terabyte und mehr an. Der Wert dieser Daten ergibt sich jedoch erst durch eine systematische Analyse und den daraus ableitbaren Erkenntnissen – kurz: dem Data Mining. Während die mathematischen Grundlagen (Statistik, maschinelles Lernen) lange bekannt sind, besteht die besondere Herausforderung heute im Umgang mit extrem großen Datenmengen, in der teilweise hohen Geschwindigkeit, in der neue Daten anfallen sowie in der Heterogenität der Datenquellen.

3 Data Mining Plattform

Bosch stellt sich diesen Herausforderungen mit einer Reihe von Aktivitäten und Projekten, allen voran einem großen, geschäftsbereichsübergreifenden Data Mining Projekt. Eines der Hauptziele dieses Projekts ist der Aufbau und die Bereitstellung einer geeigneten Big Data Plattform. So entstand im vergangenen Jahr ein High-Performance Computing Cluster als Referenz-Umgebung speziell für Boschs Data Mining Fragestellungen. Diese Umgebung setzt auf bereits etablierte Standards und Open-Source-Lösungen wie dem Hadoop Ökosystem (HDFS, MapReduce, Hive, Hbase, und Mahout) [Ap14][Sa12][Wh12]. Im Rahmen des Data Mining Projekts wurde auch eine Vielzahl kommerzieller Lösungen analysiert und auf Eignung für Bosch-Anwendungen geprüft. Die sechs am weitesten verbreiteten Produkte Knime, RapidMiner, IBM SPSS, Revolution R, Statistica und SAS sowie drei weitere ausgewählte Tools (Alpine, HP Vertica und EMC Pivotal) wurden schließlich einem Benchmark auf einem einheitlichen Datensatz unterzogen. Es stellte sich heraus, dass dieser Satz an (nicht Bosch-spezifischen) echten Anwendungsdaten bereits ausnahmslos allen Tools Probleme bereitet. Während sich die meisten Tools dank guter Bedienoberflächen und breiter Datenquellenanbindung gut zum interaktiven Ausprobieren und Prototyping eignen, setzt Bosch aktuell für produktive Lösungen vorranging auf Eigenentwicklungen, etwa auf Basis von Mahout.

4 Herausforderungen aus industrieller Sicht

4.1 Nutzbarkeit der Daten hinsichtlich Qualität

Bevor Daten mit Analyse-Tools ausgewertet werden können, müssen sie in aller Regel zunächst geeignet aufbereitet werden. Datensätze sind im *echten Leben* nie perfekt: mal fehlen einzelne Datenpunkte, mal sind (Mess-)Fehler enthalten, mal wechseln Messverfahren oder Bezugswerte und machen Datenpunkte somit schwer vergleichbar. Je größer die Datenmengen werden, umso aufwendiger werden auch die notwendigen Bereinigungen. So erklärt sich, dass in typischen Data-Mining-Anwendungen über 80% des Aufwandes für das Sichten, Verstehen, Hinterfragen und Aufbereiten der Daten anfällt. Auch läßt sich im Vorhinein kaum genau sagen, ob bestimmte Data-Mining-Fragestellungen mit ausreichender Genauigkeit oder gar überhaupt mit den vorhandenen Daten beantwortet werden können. Oft stellt sich somit nicht die Technik (sprich: Analytik) als der Flaschenhals heraus, sondern vielmehr die Menge und Qualität der verfügbaren Daten.

4.2 Sicherheit

Große Datenmengen, insbesondere solche, die sich zur Analyse und zum Data Mining eignen, stellen einen enormen Wert dar. Damit besteht das Risiko, dass diese zum Ziel von externen Angriffen werden. Eine besondere Herausforderung beim Umgang mit Big Data ist daher der Schutz der Daten. Technische Ansätze wie Zugriffskontrolle,

Verschlüsselung oder auch Anonymisierung sind zwar bekannt, führen im Alltag allerdings auch zu Einschränkungen oder offenbaren Unzulänglichkeiten. Wie lassen sich beispielsweise Daten analysieren, ohne den gesamten Datenbestand offenlegen (entschlüsseln) zu müssen? Hier arbeitet Bosch z.B. an eigenen Verfahren zum Schutz der Privatsphäre, sogenannten *Privacy-enhancing Technologies*.

4.3 Zentrale vs. dezentrale Infrastruktur

Seit einigen Jahren werden bei Bosch zur Nutzung von Skaleneffekten und Erhöhung der Sicherheit IT-Ressourcen in wenigen Rechenzentren konsolidiert. Die damit einhergehende Automatisierung und Virtualisierung ermöglichen zusätzlich eine höhere Auslastung und eine höhere Flexibilität. Aus Sicht der IT-Infrastruktur ergibt sich nun die Frage, ob die in den (Produktions-) Werken anfallenden, großen Datenmengen von mehreren Terabytes wirtschaftlich über das globale Unternehmensnetz in den zentralen Rechenzentren gespeichert und verarbeitet werden können oder ob ein dezentraler Ansatz von Rechenleistung und Speicherkapazität für Big Data und Data Mining effizienter und effektiver ist. Dies hätte eine Umkehrung des Trends und eine De-Zentralisierung zur Folge.

4.4 Global Deployment

Eine weitere signfikante Herausforderung stellt die Etablierung einer unternehmensweiten, standardisierten und integrierten IT-Infrastruktur für die Untestützung des gesamten Data Mining Prozesses dar. Diese muss neben der Erstellung und Validierung von Modellen mittels leistungsfähiger *High Performance Computing* (HPC) Plattformen sowohl Aspekte der Datenaggregation und Aufbereitung als insbesondere auch Aspekte des Modell-Deployments in verschiedensten Produktionsumgebungen und Geschäftsprozessen möglichst nahtlos integrieren. Als wesentlichen Bestandteil für das Modell-Deployment setzt Bosch die *Predictive Model Markup Language* (PMML) zur einheitlichen Beschreibung von Modellen ein. Die Modell-Ausführung kann dann mittels einer in Produktionsumgebungen eingebetteten Modell Execution Server Infrastruktur, zum Beispiel auf Basis der Scoring Engine Adapa von Zementis, bewerkstelligt werden.

Literaturverzeichnis

[Ap14] Apache Software Foundation: Apache Hadoop Project Website, http://hadoop.apache.org. Zugriff am 07.03.2014.

[Ba13] Bassi, A. et al.: Enabling Things to Talk – Designing IoT solutions with the IoT Architectural Reference Model, Springer, Berlin, 2013.

[Sa12] Sammer, E.: Hadoop Operations. O'Reilly, 1st Edition, 2012.

[Wh12] White, T.: Hadoop – The Definitive Guide, O'Reilly, 3rd Edition, 2012.