



© [M] Subaru;
Syda Productions | Fotolia

Interaktive Big-Data-Analytik in der Motorenentwicklung

AUTOR



Dr. Tobias Abthoff
ist Mitglied des Vorstands und
verantwortet die Bereiche
Technology & Business Development
der NorCom IT AG in München.

Die Auswertung von Messdaten spielt eine wichtige Rolle in der modernen Motorenentwicklung. Hierbei entstehen im Flottenbetrieb inzwischen Datenmengen, die aufgrund ihrer Größe mit herkömmlichen Analysewerkzeugen nur noch unvollständig erschlossen werden können. Basierend auf Big-Data-Technologien bietet NorCom mit DaSense eine Lösung an, die die interaktive Verarbeitung von riesigen Datenmengen erlaubt.

FEHLERANALYSE BEI GROSSEN DATENMENGEN

Mit ständig steigenden Anforderungen in der Fahrzeugentwicklung wird die Verarbeitung großer Datenmengen ein immer bedeutenderer Faktor. So wird es immer wichtiger, Fehlerquellen zu identifizieren, die aufgrund ihrer geringen Häufigkeit im normalen Entwicklungsbetrieb kaum zu erkennen sind, sondern erst in der Großserie als Summenbeanstandung – etwa als Kundenfeedback, dass „das

Fahrzeug ruckelt“ – auffallen.

Um solche Beanstandungen zu beheben, ist es wichtig, sehr genaue Informationen über Zusammenhänge zwischen Fehlern und ihren möglichen Ursachen zu bekommen. Solche Zusammenhänge finden sich mit hoher Wahrscheinlichkeit in den vielen Fehlerspeichern der Fahrzeuge. Weil die Fehler nur sehr selten auftreten, werden unter Umständen enorme Datenmengen benötigt, um eine signifikante Aussage treffen zu können. Die Datenmengen können hierbei durch-

aus im Petabyte-Bereich oder darüber liegen, was die Möglichkeiten von traditionellen Datenverarbeitungsmethoden bei Weitem übersteigt.

Um mit solchen Datenmengen arbeiten zu können, sind, vor allem im Umfeld der Internetindustrie, die sogenannten „Big-Data“-Technologien entstanden. Dieser Beitrag geht auf die Möglichkeiten in der Motorenentwicklung ein, die Big-Data-Technologien eröffnen.

Wie in [1] beschrieben, hat die NorCom IT AG das Thema Big Data in der Fahr-

zeugentwicklung eingeführt und dabei DaSense vorgestellt, eine Entwicklungsplattform der nächsten Generation für die Big-Data-Algorithmenentwicklung in der Automobilindustrie. Hierbei liegt die Stärke von DaSense darin, dass Big-Data-Technologien speziell für die Abläufe der Automobilentwicklung angepasst werden. Über interaktive Schnittstellen wird es Entwicklern ermöglicht, ihr spezifisches Know-how mit Big-Data-Methoden zu kombinieren, **BILD 1**. Zusammen mit kurzen Analysezeiten wird die Auswertung großer Datenmengen mit dieser Software zu einem effektiven Bestandteil des Entwicklungsprozesses.

Der folgende Abschnitt gibt zunächst einen Überblick über die relevanten Konzepte von DaSense, Details finden sich in [1].

WAS IST BIG DATA?

Der Begriff „Big Data“ ist eine Sammelbezeichnung für eine ganze Reihe von Technologien, die im Umfeld der großen Internetunternehmen entstanden sind. Eine der bekanntesten und am weitesten verbreiteten Big-Data-Technologien ist Hadoop [2] mit seinem zugehörigen Softwareökosystem, das auch als technologische Grundlage von DaSense

dient. Auf Hadoop aufbauende Big-Data-Lösungen basieren auf einer Reihe von Kernkonzepten, die es ermöglichen, große Datenmengen hochparallelisiert zu verarbeiten.

CLUSTER

Ein grundlegendes Konzept der Hadoop-Architektur ist der Computercluster. Ein solcher Cluster besteht in der Regel aus sehr vielen einzelnen Computern, auf denen Analysen parallelisiert abgearbeitet werden. Dazu werden die einzelnen Rechner („Knoten“) so miteinander vernetzt, dass sie für den Benutzer über eine einheitliche Schnittstelle zur Verfügung stehen. Eine wichtige Eigenschaft eines Hadoop-Clusters ist, dass die einzelnen Knoten aus Standardkomponenten bestehen, die sich leicht und kostengünstig beschaffen lassen. Desweiteren lässt sich ein Cluster nahezu nach Belieben mit wachsenden Datenmengen erweitern. Dabei sorgt die Hadoop-Infrastruktur dafür, dass Analysen skalieren und Analysezeiten pro Dateneinheit nahezu gleichbleiben.

DER „DATA LAKE“

Der Begriff des Data Lake umschreibt

die Datenspeicherung in einem Cluster. In einem Cluster sind die einzelnen Knoten sowohl für Berechnungen als auch für die Datenspeicherung zuständig. Um die effiziente parallele Abarbeitung von Analyseaufgaben auf dem Cluster zu ermöglichen, besteht die wichtigste Grundfunktion des Data Lake darin, Analysen und die zugehörigen Daten effizient zusammenzubringen. Weitere charakteristische Merkmale des Data Lake sind Fehlertoleranz und Datenvielfalt.

Zur Fehlertoleranz gehört insbesondere, dass Ausfälle von Knoten miteingeplant sind und selbstständig kompensiert werden. Auch können defekte Knoten im laufenden Betrieb ausgetauscht werden.

Die Unterstützung von vielfältigen Formaten ist besonders wichtig für die hochkomplexen Daten, die in automobilen Entwicklungsprozessen anfallen. Das Spektrum reicht hier von einfachen zeitbasierten Messreihen, wie sie etwa aus elektronischen Steuergeräten ausgelesen werden, über komplexe Metadaten, bis hin zu Sensordaten, die von Kameras oder Radargeräten stammen.

MAPREDUCE



BILD 1 DaSense ermöglicht dem Fahrzeugentwickler den interaktiven Umgang mit riesigen Datenmengen und bietet eine Vielzahl von Möglichkeiten, Entwickler-Know-how mit Big Data zu kombinieren (© NorCom)

Die Beschaffenheit eines Data Lake erfordert eine neue Sichtweise auf die Art und Weise, wie Daten analysiert und verarbeitet werden. Das am weitesten verbreitete und sehr bekannte Verfahren ist MapReduce, das 2004 von Google vorgestellt wurde [3]. Das grundlegende Prinzip ist hier, die einfache Datenanalyse in zwei Schritte einzuteilen: das „Mapping“ führt eine beliebige Anzahl von Analysen parallel auf dem Cluster aus, deren Ergebnisse dann im „Reduce“-Schritt zusammengeführt werden und das Analyseresultat liefern. Bei Bedarf können auch mehrere Mapping-Schritte hintereinander ausgeführt werden. Darauf aufbauend gibt es inzwischen auch allgemeinere und erheblich flexiblere Verfahren zur Datenanalyse auf dem Cluster wie Spark [4], das ebenfalls in DaSense zum Einsatz kommt.

DASENSE: BIG DATA FÜR DIE AUTOMOBILINDUSTRIE

DaSense bringt die Vorteile und Stärken der Big-Data-Technologien in die Automobilentwicklung und übernimmt alle notwendigen Anpassungen. Darüber hinaus bietet die Software ein einheitliches Big-Data-Gesamtkonzept, von der Teststrecke bis hin zum Entwickler-schreibtisch. Im Folgenden soll die Anwendung der oben angedeuteten

Konzepte anhand der Umsetzung eines Projektes in der Motorenentwicklung mit DaSense demonstriert werden.

DETEKTION UND BEWERTUNG VON SCHALTRUCKLERN

Schwingungen, die im Motor auftreten, etwa durch Lastwechsel im Getriebe, können sich für den Passagier als „Ruckeln“ bemerkbar machen. Ein wichtiger Aspekt, etwa bei der Entwicklung von Fahrzeugen mit Automatikgetrieben, ist, solche Ruckler möglichst zu dämpfen und somit den Fahrkomfort zu erhöhen. In dem hier vorgestellten Szenario werden Messdaten verarbeitet, die über eine Testperiode von zwei Tagen eingefahren wurden und in der Form von circa 1600 MDF4-Dateien mit einer Gesamtgröße von über 60 GB vorliegen. Mithilfe von DaSense können in kurzer Zeit aus diesen Daten fast 250.000 Schaltvorgänge extrahiert und nach ihrem Ruckeln bewertet werden. Basierend auf dieser Bewertung unterstützt DaSense den Entwickler bei der Suche nach statistisch signifikanten Zusammenhängen im gesamten Datenkorpus.

VON MESSDATEN ZU BIG DATA

Nach einer Testfahrt werden die Fahr-

zeugdaten aus elektronischen Steuergeräten ausgelesen und liegen im MDF4-Format [5] zunächst auf herkömmlichen Datenträgern vor. Zum Ende einer Testserie werden diese Daten dann auf den Data Lake von DaSense übertragen und stehen somit zur weiteren Verarbeitung auf dem Cluster bereit. Die wichtigste Aufgabe bei der Übertragung ist die Aufschlüsselung der MDF4-Dateien und die Übertragung der enthaltenen Messreihen in ein Cluster-freundliches Datenbankformat. DaSense bietet die vollautomatisierte Konvertierung von MDF4-Dateien in das Parquet-Format [6], das besonders zur Filterung und Suche in Datensätzen in Hadoop optimiert ist. Eine MDF4-Datei enthält in der Regel die Daten aus mehreren hundert Messkanälen, und jeder dieser Kanäle wird extrahiert und in der Parquet-Datenbank abgelegt.

SCHALTVOGÄNGE IDENTIFIZIEREN UND EXTRAHIEREN

DaSense bietet eine Vielzahl von Möglichkeiten, mit deren Hilfe ein Entwickler Ereigniskombinationen in Messkanälen identifizieren und extrahieren kann. In diesem Beispiel werden zur Identifikation eines Schaltvorganges zwei Kanäle betrachtet: „Gang“ und „ZielGang“. Hierbei enthält „Gang“ den gerade tatsächlich eingelegten Gang und „ZielGang“ den zu schaltenden Gang. Findet kein Schaltvorgang statt, dann enthalten beide Kanäle den gleichen numerischen Wert. Entschieden die Schaltelektronik zu einem bestimmten Zeitpunkt, dass eine Schaltung stattfinden soll, dann setzt sie zunächst den Wert im Kanal „ZielGang“ auf den zu schaltenden Gang und nach Abschluss des Schaltvorganges wird dann „Gang“ wieder auf den gleichen Wert wie „ZielGang“ gesetzt. Die Zeitintervalle, in denen „Gang“ ungleich „ZielGang“ gilt, entsprechen genau den während einer Fahrt durchgeführten Schaltvorgängen. Zur komfortablen Formulierung solcher Bedingungen bietet die Software eine sogenannte domänenspezifische Sprache für die Analyse zeitbasierter Messreihen an. Tatsächlich genügt es, dass ein Benutzer nicht viel mehr als den Ausdruck „Gang ! = ZielGang“ angeben muss, um in einem Mapper-Schritt eine Extraktion der zu Schaltvorgängen gehörenden Zeitabschnitte auf allen verfügbaren Daten auf dem Cluster anzustoßen

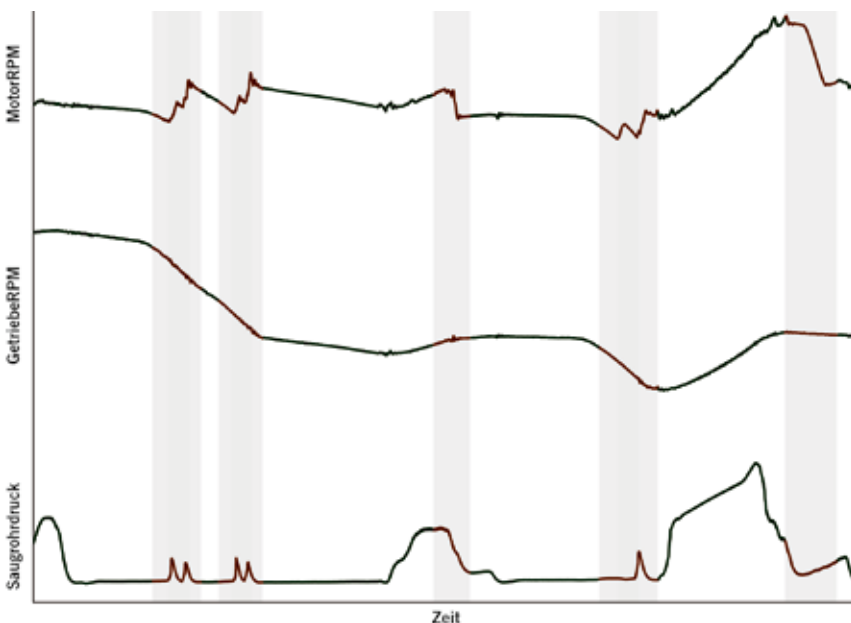


BILD 2 Extraktion von zu Schaltvorgängen gehörenden Zeitabschnitten aus allen Messkanälen; die einzelnen (durch graue Bänder angedeuteten) Zeitabschnitte werden als eigene Messreihen wieder in einer Parquet-Datenbank abgespeichert (schematisch) (© NorCom)

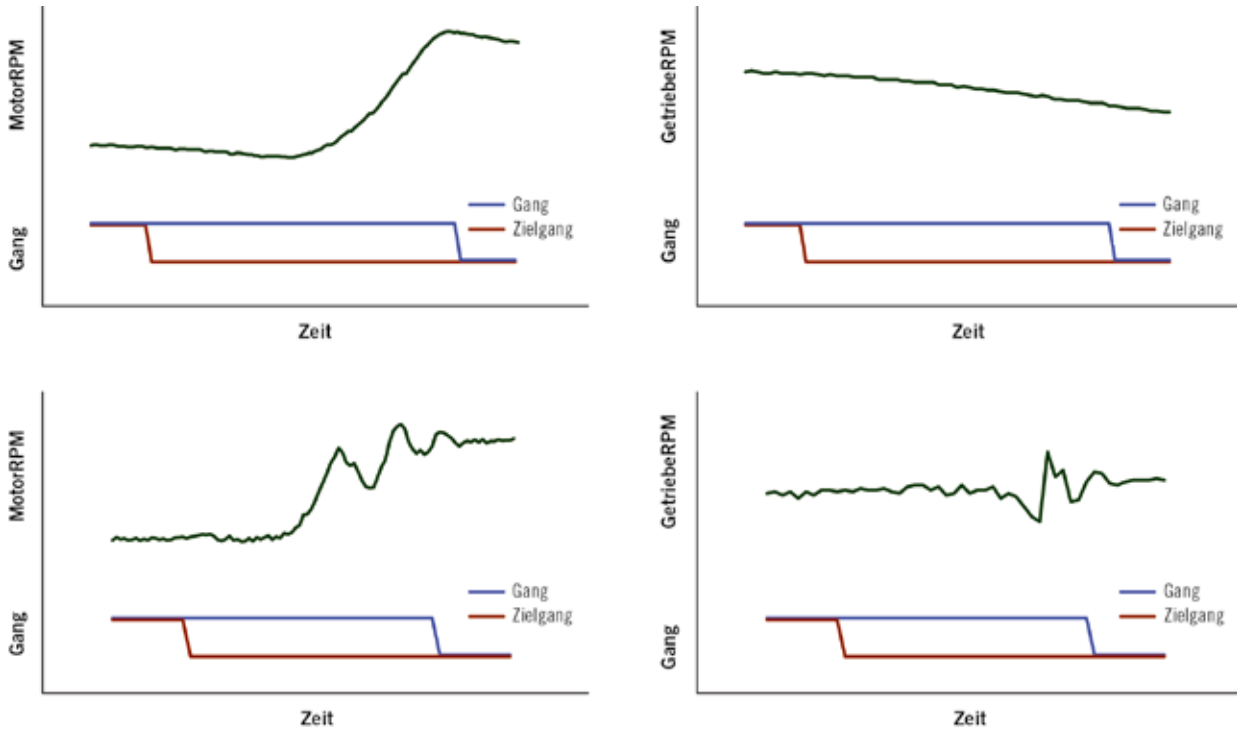


BILD 3 Zwei Beispiele für den Verlauf der Drehzahl für ruhige (oben) und stark ruckelnde (unten) Schaltungsverläufe; der eigentliche Schaltverlauf findet statt, während die Kanäle „Gang“ und „ZielGang“ unterschiedliche Werte annehmen; die Schwingungen erstrecken sich oft über das Schaltende hinaus (schematisch) (© NorCom)

(siehe auch **BILD 7** in [1]).

In einem zweiten Schritt verwendet der Entwickler die gefundenen Zeitabschnitte, um die zugehörigen Zeitausschnitte aus allen anderen verfügbaren Kanäle zu erhalten. Das Resultat wird in einer neuen Parquet-Datenbank gespeichert, in der aus jedem Kanal jeder zu einem Schaltvorgang gehörende Zeitabschnitt als eigener Eintrag abgelegt ist, **BILD 2**.

BEWERTUNG VON SCHALTRUCKLERN

Wie aus der Fahrzeugmechanik bekannt ist, lassen sich motorverursachte Schwingungen im Fahrzeug am genauesten an der Drehzahl able-

sen. Typischerweise können Drehzahlen an mehreren Stellen im Fahrzeug abgelesen werden. Zur Illustration betrachten wir hier die beiden Kanäle „MotorRPM“ und „GetriebeRPM“. **BILD 3** zeigt jeweils einen guten und einen schlechten Verlauf für einen Runterschaltvorgang in diesen Kanälen. Für den Kanal „MotorRPM“ sieht man, dass die Drehzahl während des Schaltvorganges stark ansteigt, während der Kanal „GetriebeRPM“ einen linearen Verlauf hat. Im Falle eines Schaltrucklers werden beide Verläufe von Schwingungen überlagert.

Um eine statistische Auswertung zu ermöglichen, bietet DaSense die Möglichkeit, für solche Verläufe beliebige Bewertungen zu erstellen und diese in einem

Mapper-Schritt auf den gesamten Datenbestand anzuwenden. Um etwa für den Kanal „MotorRPM“ eine Bewertung für Schwingungsereignisse zu erhalten, bietet sich beispielsweise eine Fourier-Analyse an, mit deren Hilfe ein Maß für Schwingungen etwa im relevanten Bereich von 3 bis 10 Hz definiert werden kann. Mit DaSense lässt sich diese Größe als Analysefunktion leicht implementieren und auf dem Cluster per Knopfdruck auf alle Schaltvorgänge anwenden. Der Anwender erhält dann innerhalb weniger Sekunden eine statistische Verteilung dieser Größe nach Gängen sortiert, wie in der linken Hälfte von **BILD 4** exemplarisch für einige Gänge gezeigt. Wir stellen fest, dass einige Verteilungen deutlich breiter sind als andere, was darauf hindeutet, dass

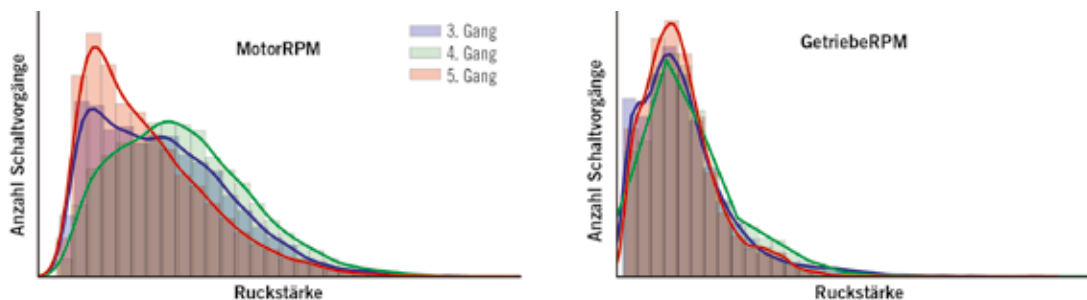


BILD 4 Verteilung von Ruckstärken in den Kanälen „MotorRPM“ und „GetriebeRPM“; die Verteilungen für „MotorRPM“ weisen in beiden Fällen für verschiedene Gänge charakteristische Unterschiede auf (schematisch) (© NorCom)

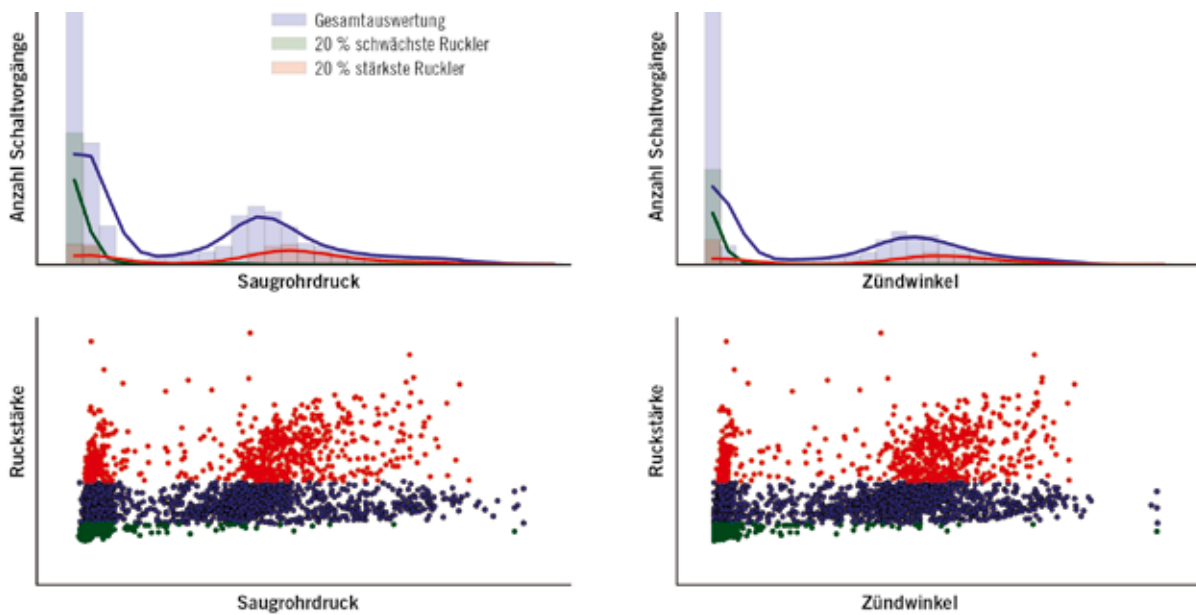


BILD 5 Unten: Ruckelstärke als Funktion von Saugrohrdruck (links) und Zündwinkel (rechts); oben: es ist zu erkennen, dass ein großer Teil der Ruckler bei hohem Saugrohrdruck und frühem Zündzeitpunkt auftritt (schematisch) (© NorCom)

Schaltvorgänge in diesen Gängen in der Regel unruhiger verlaufen.

Um den Effekt für bereits vorgenommene Maßnahmen zur Ruckeldämpfung zu überprüfen, ist es naheliegend, nun die gleiche Analyse auf dem Kanal „GetriebeRPM“ durchzuführen. Allerdings sind hier die Schwingungen bereits in den etwas höherfrequenten Bereich verlagert. Hinzu kommt, dass dieser Kanal nur mit 50 Hz gesampelt wird, im Gegensatz zu 100 Hz im Kanal „MotorRPM“. Dies führt dazu, dass die Fourier-Analyse kein zuverlässiges Maß mehr liefert. Stattdessen können wir uns hier den einfachen Grundverlauf zunutze machen und wir stellen fest, dass die Standardabweichung zu einem linearen Fit ein gutes Maß für ein Ruckelereignis ist. Auch die neue Analysefunktion ist in DaSense leicht implementiert und kann auf dem Cluster hochparallelisiert ausgeführt werden. Die rechte Hälfte von **BILD 4** zeigt die Verteilung dieser Größe für einige Gänge. Man sieht, dass die Verteilungen schmalere und Gang für Gang regulärer verlaufen und somit die bereits vorgenommenen Maßnahmen zur Ruckeldämpfung gewirkt haben. Jedoch weisen die Flanken der Verteilungen immer noch eine signifikante Anzahl von starken Ruckelereignissen auf und es sind immer noch Unterschiede in den Gängen erkennbar.

KORRELATIONSANALYSE

Um weitere Aufschlüsse zu bekommen, interessiert man sich nun für Vorgänge in anderen Kanälen oder Kanalkombinationen während eines Schaltvorgangs. Ähnlich wie für Ruckelereignisse in Drehzahlkanälen stellt DaSense eine Vielzahl von Möglichkeiten zur Verfügung, solche Vorgänge zu identifizieren, zu filtern – etwa nach Gängen – und in Kombination mit spezifischem Entwickler-Know-how zu bewerten (siehe auch [1] für einen Überblick über die Möglichkeiten). Darauf aufbauend können nun mit DaSense die verschiedensten Methoden des Machine Learning zum Einsatz kommen. Zur Illustration wurde auf den Messdaten eine einfache Korrelationsanalyse durchgeführt. Dazu wurden in DaSense ähnliche Bewertungen wie die zuvor beschriebenen auf andere Kanäle angewendet und die empirische Korrelation der gewonnenen Verteilungen mit denen der Schaltruckler im Kanal „GetriebeRPM“ berechnet. **BILD 5** zeigt für einen spezifischen Gang die Verteilungen zweier Kanäle, die eine sehr hohe Korrelation zu Schaltrucklern aufweisen. Die oberen Diagramme zeigen die Verteilungen von Saugrohrdruck und Zündwinkel während der Schaltvorgänge. Dabei ist die Gesamtstatistik blau eingezeichnet und die Teilstatistiken für die jeweils 20 % schwächsten und stärksten

Ruckler in grün beziehungsweise rot. Die unteren Diagramme zeigen jeden Schaltvorgang als einzelnen Messpunkt, wobei die Ruckstärke über den zugehörigen Saugrohrdruck beziehungsweise Zündwinkel eingetragen ist. Es lässt sich ablesen, dass die meisten starken Ruckelereignisse bei einem hohen Saugrohrdruck und frühem Zündzeitpunkt auftreten.

Die hier gezeigten Ergebnisse wurden auf einem kleinen Cluster mit sechs Knoten berechnet. Pro Schaltereignis wurden Ausschnitte aus circa 700 Messkanälen extrahiert und statistisch bewertet. Die Laufzeiten für statistische Bewertungen und Korrelationsanalysen lagen im Bereich von einigen Sekunden bis zu einigen Minuten.

LITERATURHINWEISE

- [1] Abthoff, T.: Big-Data-Technologien in der Fahrzeugentwicklung In: ATZelextronik 11 (2016), Nr. 5, S. 42-47
- [2] Welcome to Apache Hadoop! Online: <http://hadoop.apache.org>, aufgerufen am 05.05.2016
- [3] Dean, J.; Ghemawat, S.: MapReduce: Simplified Data Processing on large Clusters. OSDI'04: Sixth Symposium on Operating System Design, San Francisco, 2004
- [4] Zaharia, M.; Chowdhury, M.; Franklin, M. J.; Shenker, S.; Stoica, I.: Spark: Cluster Computing with Working Sets. HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, Boston, 2010



READ THE ENGLISH E-MAGAZINE

Test now for 30 days free of charge:
www.mtz-worldwide.com