Dauerlauferprobungen

Big Data in weltweit verteilten Tests agiler analysieren

Durch immer mehr Sensordaten stoßen Testfahrzeuge an die Bandbreitengrenze der verwendeten Ethernet-Infrastruktur. Es ist wenig realistisch, die enormen Datenmengen zuerst zu einem zentralen Cluster zu bewegen, um dann mit der Analyse zu beginnen. NorCom zeigt, wie mithilfe der sogenannten Distributed Query Engine und eines Daten-Lifecycle-Managements – auf Basis der hauseigenen DaSense-Plattform – performante Analysen auf global verteilten Datensätzen unmittelbar ausgeführt und zugänglich gemacht werden können.

AUTOREN

Dr. Andreas Pawlik
ist Teamleiter Data Science bei NorCom in München.

Thomas Bonfer
ist Lead Data Scientist, spezialisiert auf Autonomes Fahren bei NorCom in München.

Ludwig Oser
ist Lead Data Scientist, spezialisiert auf Antriebssysteme, bei NorCom in München.

DIE ZEITKRITISCHE ANALYSE GROSSER DATENMENGEN

In der Automobilindustrie steht die Auswertung großer Datenmengen in allen Phasen der Produktentwicklung immer stärker im Fokus. Während der ständig laufenden Testfahrten fallen über einen längeren Zeitraum große Mengen an Messdaten an verschiedenen global verteilten Standorten an. Diese Daten müssen zeitnah ausgewertet werden, um die darin enthaltenen Informationen in den Entwicklungsprozess einfließen zu lassen.

Ein zeitaufwendiger Transport von großen Datensätzen in eine zentralisierte Umgebung ist in diesem engen Zeitrahmen nicht möglich. Nur durch die sofortige Auswertung können Testläufe agil geplant und Entwicklungskosten reduziert werden.

Die in diesem Artikel vorgestellte Software trägt zur Lösung dieser gewaltigen Herausforderungen bei. Basierend auf DaSense [1], einer Big-Data- und Advanced-Analytics-/Deep-Learning-Plattform, ist sie speziell auf die Anforderungen der Automobilbranche zugeschnitten. Die hier herausgestellte Funktion von DaSense, das Arbeiten mit global verteilten Daten, wurde von NorCom zum Patent angemeldet, und zusammen mit einem großen Automobilhersteller wurde die Lösung an mehreren Standorten erfolgreich erprobt.

BIG-DATA-ENTWICKLUNGS-PLATTFORM DASENSE

DaSense ist eine auf der bekannten Big-Data-Technologie Hadoop basierende Plattform [2], die Analysen in einem Rechen-Cluster mit Apache Spark [3] für Daten bis zum Petabyte-Bereich skaliert. Hierbei wird in einem ersten Schritt ausgenutzt, dass für die Auswertung einer Testfahrt auch nur die Daten dieser einen Testfahrt betrachtet werden müssen. Dieser Teil der Analyse kann somit massiv parallel auf mehreren Clustern ablaufen, und die Anzahl an Testfahrten, die gleichzeitig ausgewertet werden können, ist lediglich durch die Größe der Rechen-Clusters beschränkt. Im nachfolgenden Schritt werden die einzelnen Teil-Ergebnisse dann zu einer Gesamtanalyse zusammengefasst und weiteren Schritten zugeführt, zum Beispiel einer Visualisierung.

Eine in der Praxis oft benötigte Analyse, die bereits ohne jegliche eigene Programmierung von DaSense abgedeckt wird, ist die Ereignissuche. Hierbei werden für verschiedene Signal- und Metadaten Grenzwerte festgelegt, um Ereignisse in den Daten zu identifizieren (zum Beispiel Schaltvorgänge, Temperaturbereiche et cetra). Zunächst werden diese Ereignisse in allen vorliegenden Messdaten identifiziert und die Ergebnisse pro Messfahrt zurückgeliefert. Dieses Ergebnis ist im Vergleich zum gesamten Datensatz sehr klein, weshalb die Grundidee von DaSense wie folgt lautet: „Move the algorithm, not the data." Anstatt also die Daten zu bewegen, wird die Auswertung zu den Daten transferiert. Im anschließenden Schritt werden diese Einzelergebnisse dann zu einer Gesamtstatistik zusammengeführt.

Ein einzelnes Ereignis kann durch lediglich zwei Werte – die Anfangs- und die Endzeit – repräsentiert werden. Mehrere Ereignisse in der gleichen Messfahrt werden so zu einer Zeitintervall-Liste zusammengefasst. Diese ist in DaSense als Austauschformat definiert, um mehrere Applikationen (Apps) zu einem Big-Data-fähigen Workflow zusammenzuschließen. So kann die Search-App die gefundenen Ereignisse an weitere Analyse-Apps weiterreichen (zum Beispiel für eine Klassierung), an Visualisierungs-Apps (zum Beispiel an eine Histogramm-App) oder an die Export-App, die die gefundenen Ereignisse in verschiedenen Dateiformaten zur Verfügung stellen kann, um sie mit anderen Tools weiter zu bearbeiten, **BILD 1**. Der Datenfluss bleibt dabei größtenteils virtuell, ein entscheidendes Design-Merkmal für die Skalierung auf große Daten.

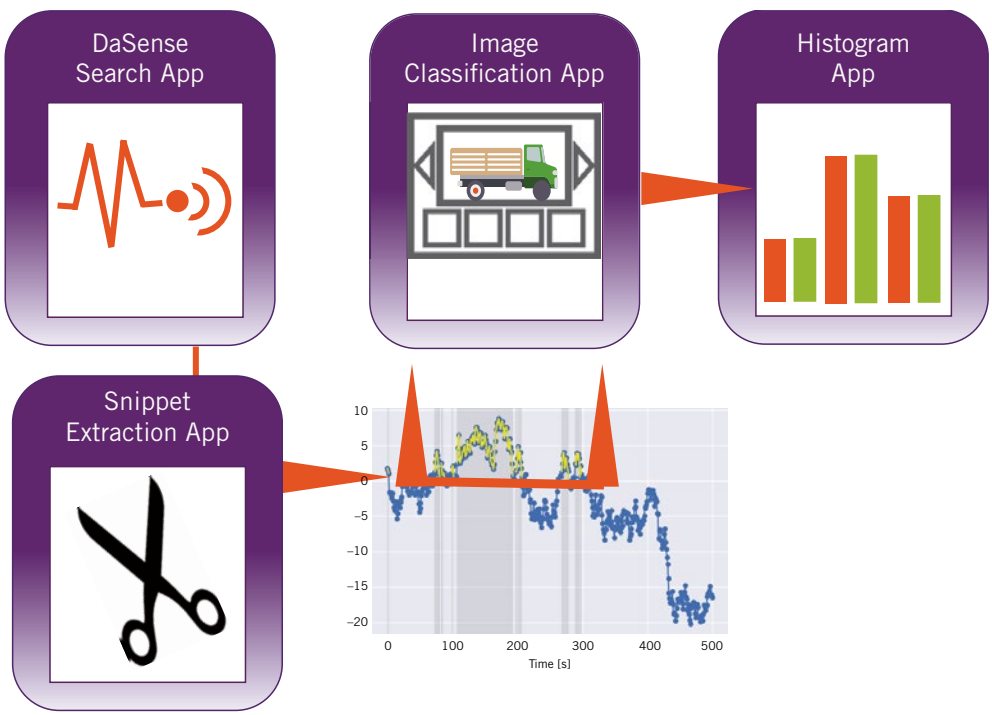

BILD 1 In DaSense werden nur die relevanten Ereignisse, nicht die gesamten Daten zur Analyse weitergereicht (© NorCom)

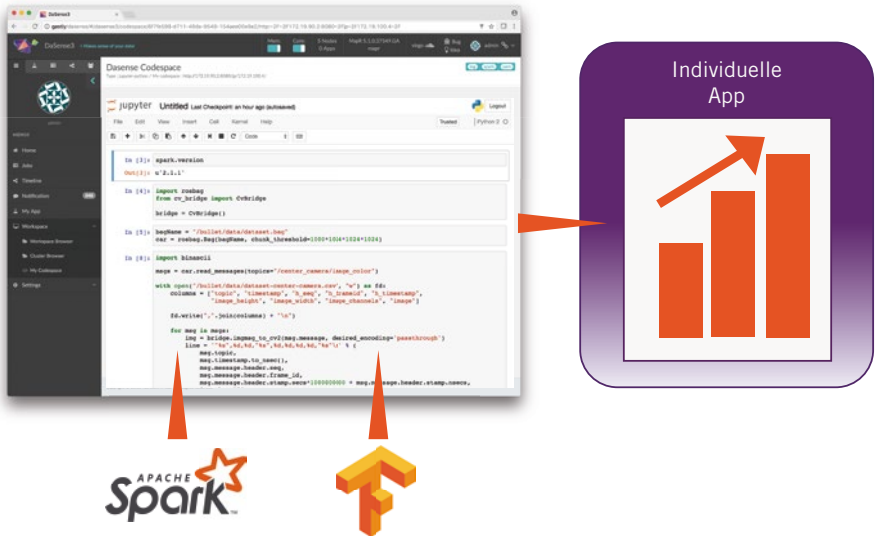

BILD 2 In DaSense können Abfragen programmiert werden und – bei Bedarf – Apps zur Mehrfachverwendung extrahiert werden (© NorCom)

Auswertungen, die nicht über die Standard-Apps von DaSense abgedeckt werden, können jederzeit über eine komfortable Programmierschnittstelle realisiert werden. Hierfür stellt die Big-Data-Plattform eine auf Python basierte domänenspezifische Sprache (DSL) zur Verfügung, die für die Auswertung von verteilt abgelegten Zeitreihen optimiert ist. In der interaktiven Entwicklungsumgebung können damit komplexe Analysen umgesetzt werden, bei denen sich DaSense automatisch um die Parallelisierung kümmert. Fertige Analysen können für die wiederholte Ausführung einfach und in Eigenregie in produktive Apps überführt werden, **BILD 2**.

VERTEILTE ANALYSEN MIT DER DISTRIBUTED QUERY ENGINE

Eine Einschränkung im klassischen Ansatz von Big Data mit Hadoop ist, dass die auszuwertenden Daten an einem Cluster-Standort zur Verfügung stehen müssen, eine Verteilung der Daten über mehrere Standorte hinweg ist nicht vorgesehen. Die Distributed Query Engine (DQE) erweitert diesen Ansatz um den nächsten logischen Schritt: das Ausführen von Analysen auf mehreren Clustern zur gleichen Zeit, **BILD 3**. Dabei spielt es keine Rolle, ob es sich bei den angebundenen Clustern um On-premise- oder Cloud-Lösungen handelt. Auf jedem einzelnen Cluster ist eine Instanz von DaSense installiert, die zu einem großen Netzwerk zusammengeschaltet werden. Analysen werden damit über mehrere Rechenzentren hinweg ausführbar.

Für die DQE wurde eine eigene Job Engine implementiert, mit der auch komplexe Workflows von Jobs abgebildet werden können. Die DQE kümmert sich um die Planung der einzelnen Jobs und das Verschicken von Algorithmen. Sollte sich ein Austausch von Rohdaten einmal nicht verhindern lassen – zum Beispiel, wenn Signale aus verschiedenen Testfahrten direkt miteinander verglichen werden sollen –, so bietet die DSL die Funktionalität, um die Rohdaten automatisch auf das nötigste Maß zu beschneiden und zu komprimieren.

Dieser Ansatz besitzt einen weiteren entscheidenden Vorteil: In vielen Ländern gibt es gesetzliche Einschränkungen, die den Transfer der Rohdaten über die Landesgrenzen hinaus erschweren oder verhindern. Mit der DQE können auch in diesem Fall die Messfahrten ausgewertet werden, ohne dass die Daten dabei den Ort ihrer Erzeugung verlassen.

DATA-LIFECYCLE-MANAGEMENT

Mithilfe des „Data Lifecycle Management“ ist es dem DaSense-System möglich, zu jedem Zeitpunkt den Aufenthaltsort der Daten zu verfolgen und gegebenenfalls den physischen Transport der Daten zu planen. Denn auch wenn die global verteilten Daten sofort für Analysezwecke zur Verfügung stehen, kann ein späterer Transport der Daten durchaus nützlich sein. Vor allem um die Ausfallsicherheit der teils mobilen lokalen Cluster zu gewährleisten, wird man die anfallenden Daten in der Regel weiterhin in ein zentrales Cluster überführen.

Das Data-Lifecycle-Management verfolgt dabei, wo sich die Daten zu einem gegebenen Zeitpunkt befinden, um die verteilte Analyse zu jedem Zeitpunkt zu gewährleisten, **BILD 4**. Die Daten werden erst von einem Standort entfernt, wenn sie komplett zu einem anderen Standort übertragen worden sind und aktuell keine Analyse auf dem Datensatz durchgeführt wird. Hierdurch können die Daten auch auf „langsamen“ Wegen (zum Beispiel per Post) bewegt werden und stehen dabei jederzeit für Auswertungen zur Verfügung.

ZUSAMMENFÜHRUNG DER KOMPONENTEN

Die Distributed Query Engine und das Data-Lifecycle-Management stellen zwei unabhängige Erweiterungen der DaSense-Plattform dar, allerdings entfalten sie ihr gesamtes Potenzial erst im Zusammenspiel. Im ersten Schritt wird eine Analyse mithilfe des DaSense Query Optimizers untersucht und gegebenenfalls durch die folgenden Module optimiert:

- Algebraic Library – algebraische Regeln zur Modifizierung einer Analyse
- Cost-Prediction – eine Reihe von Funktionen, welche die Kosten zur Berechnung einer Analyse vorhersagen

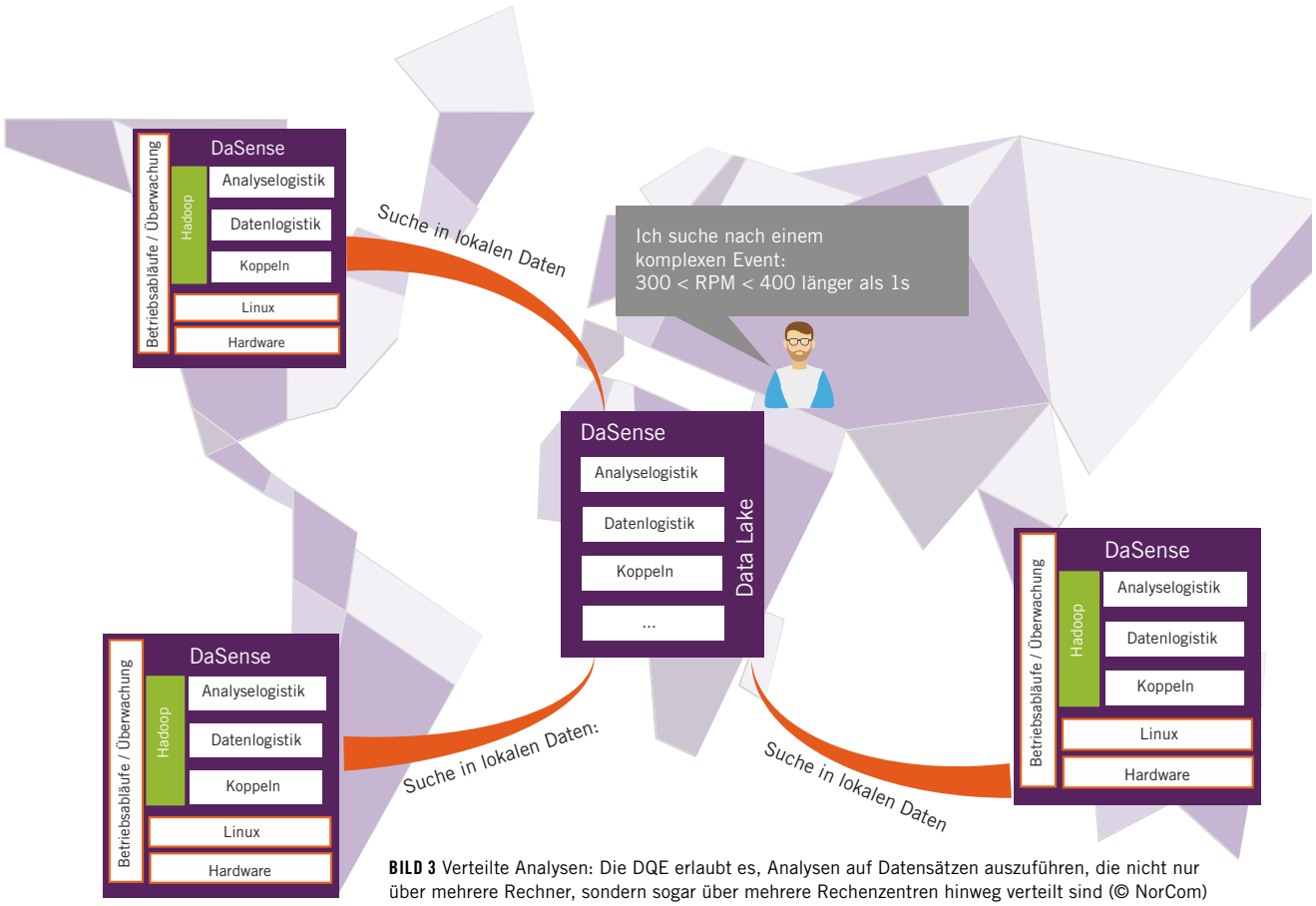

BILD 3 Verteilte Analysen: Die DQE erlaubt es, Analysen auf Datensätzen auszuführen, die nicht nur über mehrere Rechner, sondern sogar über mehrere Rechenzentren hinweg verteilt sind (© NorCom)

BILD 4 Data-Lifecycle-Management: Zu jedem Zeitpunkt liegen die Daten auf mindestens einer DaSense-Instanz zur Analyse bereit; nach dem Transport werden redundante Daten gegebenenfalls gelöscht (© NorCom)

- Execution Engine Capabilities – für jeden Standort können Informationen wie die Anzahl freier Ressourcen oder technische Spezifikationen der verfügbaren Hardware abgerufen werden
- Bandwidth-Graph – die verfügbare Bandbreite für In- und Output an den einzelnen Standorten
- Metadata Store – eine Datenbank zur Verwaltung von Metadaten von analysierbaren Daten – dazu zählt zum Beispiel die Datengröße.

Der Query Optimizer erstellt über eine vordefinierte Kostenfunktion eine Vorhersage der anfallenden Kosten für eine Analyse, basierend auf Faktoren wie zum Beispiel die erwarteten Laufzeiten, den Bedarf an Ressourcen, der Netzwerkbandbreite et cetera. Ziel des Query Optimizers ist es, eine gegebene Analyse mithilfe der algebraischen Regeln so zu modifizieren, dass minimale Kosten entstehen.

Nach der Optimierung entsteht ein Distributed Execution Graph (DEG), dessen Komponenten auf unterschiedlichen Standorten ausgeführt werden sollen. Diese Aufgabe wird vom Asynchronous Distributed Execution Coordinator übernommen. Dieser Koordinator sendet dann Teile des DEG zu den einzelnen Standorten, aggregiert die Ergebnisse und gibt diese letztlich an den Anwender zurück.

Ein ausschlaggebender Faktor bei der Kostenoptimierung ist die Datenlokalität. Ähnlich wie Hadoop, folgt auch DaSense bei der verteilten Suche dem Paradigma, unnötige Datentransfers zu vermeiden. An dieser Stelle kommt das Data-Lifecycle-Management ins Spiel. Zum einen verwaltet es Informationen über den Standort und die Größe der Daten. Zum anderen können über den Data Movement Planner Datenbewegungen angestoßen werden. Dieser Planer kennt eine Reihe von Regeln – entweder manuell festgelegt oder mittels Machine-Learning-Methoden aus Datenbewegungsmustern abgeleitet –, die durch den Planer interpretiert werden können. Der Data Mover führt schließlich Kopiervorgänge durch; ein Clean-up Agent ist für das Entfernen von Daten verantwortlich. Durch die intelligente Planung von Datenbewegungen können die Kosten von häufig ausgeführten Analysen deutlich reduziert werden, **BILD 5**.

ANWENDUNG IM DAUERLAUF

Im Fahrzeugdauerlauf werden in immer kürzerer Zeit immer mehr Messdaten aufgezeichnet. Da die Fahrzeuge weltweit im Einsatz sind, wird es zunehmend schwieriger, diese Messdaten für die Analyse in die Zentrale zu übertragen. In einem Projekt erprobte NorCom mit einem OEM eine verteilte Messtechnik, um die neuen Herausforderungen im Dauerlauf zu lösen, **BILD 6**. Dabei kommen Messdatenboxen zum Einsatz, an denen die in den Dauerlauffahrzeugen verbauten Logger beim Schichtwechsel direkt vor Ort entladen werden. Die Messdatenboxen dienen nicht nur als Zwischenspeicher, sondern fungieren gleichzeitig als flexible, transportable und weltweit integrierbare Analyseplattformen. Grundlage ist das oben beschriebene Analyseverfahren mit DaSense, das die Anforderungen an die Übertragungsbandbreite erheblich verringert und eine Auswertung auch großer Datenmengen fast in Echtzeit ermöglicht. Die Implementierung des Projekts erfolgte in zwei Phasen. In der Phase 1 wurde DaSense im zentralen Datenzenter sowie auf einer Messdatenbox installiert. In das Datenzenter und auf der Box abgeladene Daten werden von DaSense automatisch in ein Big-Data-fähiges Format transformiert und den Anwendern für die schnelle Suche, Reporting, sowie Root-Cause-Tiefenanalysen [4] zur Verfügung gestellt. In Phase 2 wurden mit DaSense über Datacenter und Messdaten-Box geografisch übergreifende Analysen realisiert.

AUSBLICK

Mit DaSense hat NorCom eine Messdatenlösung entwickelt, die schnelle und

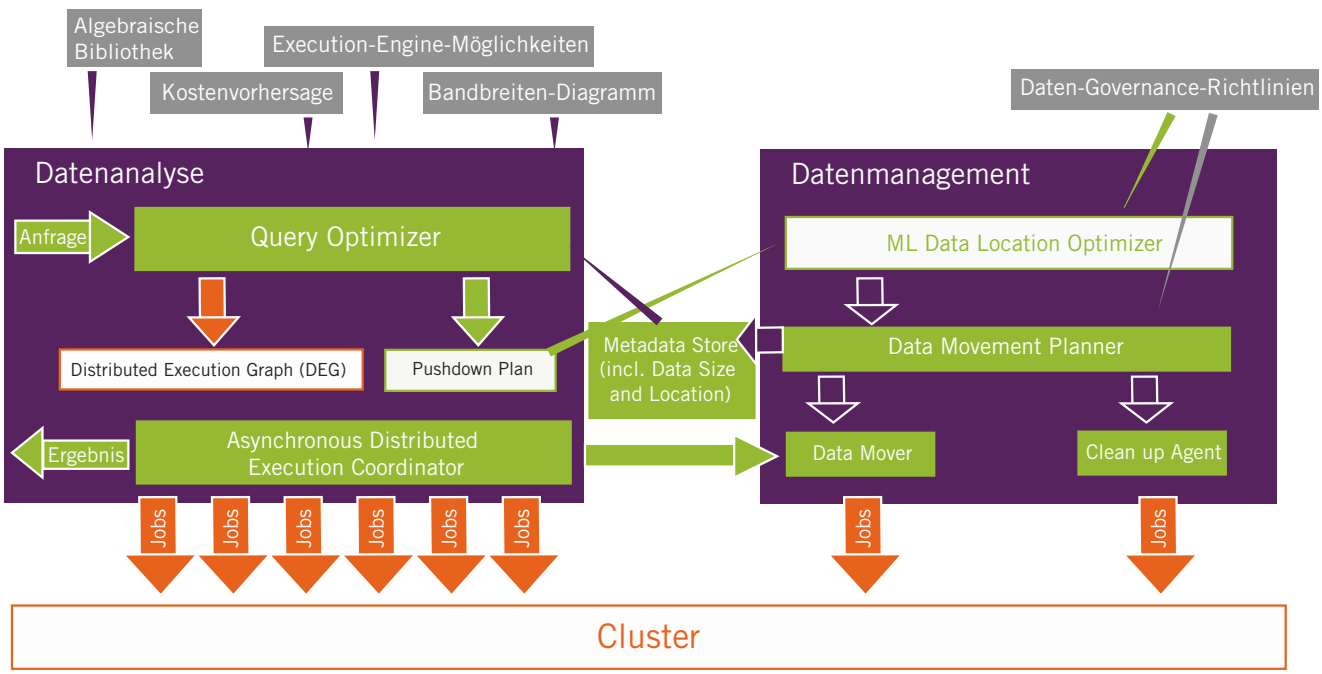

BILD 5 Durch den Einsatz verschiedener Komponenten kann die Laufzeit der Analysen optimiert und Daten entsprechend bewegt werden (© NorCom)

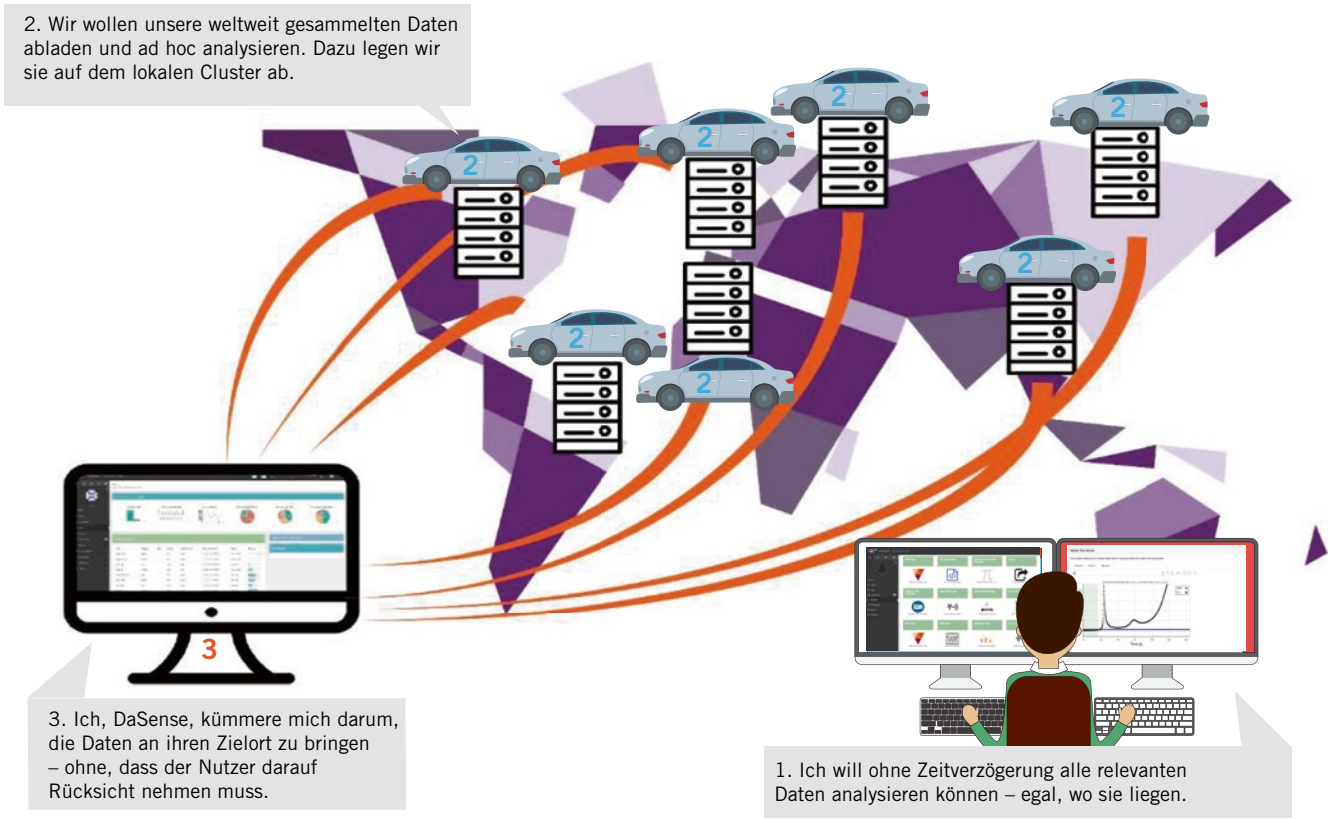

BILD 6 Im Dauerlauf werden Messdaten in großen Mengen global verteilt generiert; die Analyse soll trotzdem auf allen Daten ohne Verzögerung möglich sein, bevor die Daten zu einem zentralen Cluster transferiert werden; falls nötig können die Daten trotzdem, ohne Downtime, zum zentralen Cluster überführt werden (© NorCom)

skalierende Analysen aus der Ferne ermöglicht, selbst wenn die Daten geografisch über mehrere Standorte verteilt sind.

Aktuell erfolgt die Erweiterung der Funktionalität zur Einbindung von GPUs zur Beschleunigung von Analysen, insbesondere der Anwendung tiefer neuronaler Netze für das maschinelle Lernen, was derzeit bereits auf einzelnen Standorten erfolgt. Damit eignet sich die Big-Data- und die Advanced-Analytics-/Deep-Learning-Plattform für den Einsatz auch über den Dauerlauf hinaus, zum Beispiel für die weltweite Entwicklung von Algorithmen für das autonome Fahren [5].

LITERATURHINWEISE

[1] Abthoff, T.: Big Data Technologien in der Fahrzeugentwicklung. In: ATZelektronik 11 (2016), Nr. 5, S. 48-51
[2] http://hadoop.apache.org/, aufgerufen am 14.08.2018
[3] https://spark.apache.org/, aufgerufen am 14.08.2018
[4] Abthoff, T.: Interaktive Big Data Analytik in der Motorenentwicklung: MTZ 77 (2016), Nr. 12, S. 62-66
[5] Pawlik, A. et.al.: Big data for assisted and
[6] autonomous driving. Proceedings, 18th International Stuttgart Symposium, 2018