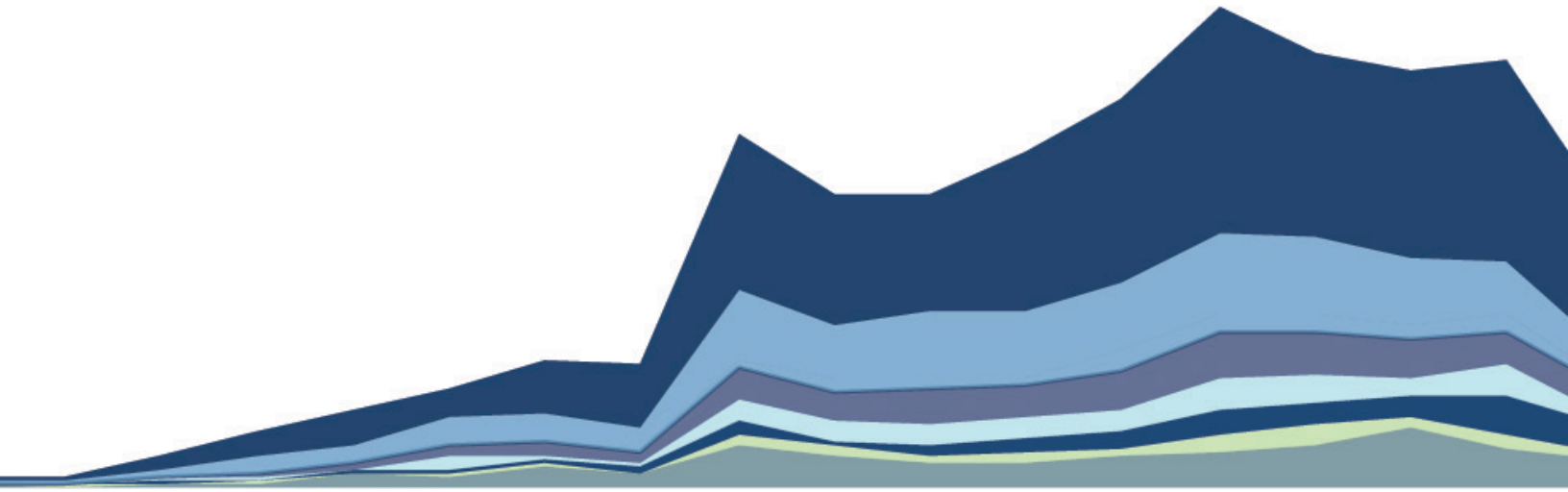


Jeff Feng, Produktmanager bei Tableau

# Die Big Data-Vision von Tableau

Wir leben in einem neuen Zeitalter. Daten sind heute ein wichtiger „Rohstoff“ für den geschäftlichen Erfolg der Unternehmen und läuten die nächste industrielle Revolution ein. Im 18. und 19. Jahrhundert haben neue Herstellungsverfahren die Fertigungstechnik vollständig verändert, und auf ähnliche Weise öffnet das Zeitalter von Big Data ein gänzlich neues Kapitel, wie Daten produziert, analysiert und genutzt werden.

Es liegt eine gewisse Ironie darin, dass Big Data sowohl Chancen als auch Gefahren birgt. Datenbestände werden zum wesentlichen Unterscheidungsmerkmal zwischen extrem rentablen und schwächelnden Unternehmen. Angesichts des enormen Volumens, rasanten Wachstums und der Vielfalt der Daten stoßen relationale Datenbankverwaltungssysteme an Kapazitäts- und Kostengrenzen.

Unternehmen setzen daher zunehmend auf Big Data-Technologien wie Hadoop-, Spark- und NoSQL-Datenbanken, um ihre gesteigerten Anforderungen an große Datenmengen zu erfüllen. Zur Bereitstellung der Technologien verwenden Unternehmen interne sowie Cloud-basierte Modelle. Darüber hinaus werden Konzepte von Hadoop zur Erweiterung des Angebots in schnelle analytische Datenbanken integriert oder direkt über Konnektoren verbunden. Zudem umfassen schnelle analytische Datenbanke und Data Warehouses entweder Hadoop-Konzepte zur Erweiterung des Angebots oder direkte Konnektorenzu Hadoop. Bei aller Weiterentwicklung und Konsolidierung der Big Data-Landschaft erweist sich ein Thema als beständig: Unternehmen müssen in der Lage sein, ein gängiges Analysetool zu nutzen, um auf ihre Daten zuzugreifen, ganz gleich, ob es große oder kleine Volumen sind und unabhängig vom Speicherort.

# Inhaltsverzeichnis

## Die Tableau-Strategie für (Big) Data

Tableau hat sich das Ziel gesetzt, Benutzer dabei zu unterstützen, ihre Daten sichtbar und (be)greifbarzu machen. Dieser Weg, so unsere Grundüberzeugung, ist nur gangbar, wenn Daten demokratisiert werden, also wenn „die Personen, die die Daten kennen, berechtigt sind, Fragen zu den Daten zu stellen“. Wissensarbeiter sollten einfach und selbstverständlich auf ihre Daten zugreifen können, ganz gleich, wo diese gespeichert sind. Dieselben Wissensarbeiter sollten zudem ihre Daten analysieren und daraus Erkenntnisse ziehen können, ohne die Hilfe einer kleinen Elite aus Informatikspezialisten und Entwicklern in Anspruch nehmen zu müssen.

Unabhängig vom Datenvolumen ist die Visualisierung von Daten wichtig, da sie Informationen in Erkenntnisse und Maßnahmen verwandeln kann. Besondere Bedeutung hat die Visualisierungstrategie für Big Data, da die mit dem Speichern, Vorbereiten und Abfragen von Daten verbundenen Kosten viel höher sind. Unternehmen müssen daher gut strukturierte Datenquellen nutzen und Best Practices konsequent anwenden, um ihren Wissensarbeitern direkte Abfragen in Big Data zu ermöglichen. In jüngster Zeit hat der Bereich Big Data erhebliche Innovationen durchlaufen. Somit steht eine Vielzahl von Optionen zur Verfügung, von denen jede ihre eigenen Stärken besitzt. Tableau verfolgt die Vision, jede beliebige Big Data-Plattform zu unterstützen, die für unsere Benutzer relevant wird, und sie dabei zu unterstützen, mit ihren Daten in Echtzeit zu interagieren.

Zur Umsetzung dieser Big Data-Vision konzentriert sich Tableau auf sechs Säulen:

1. **Breiter Zugang zu Big Data-Plattformen**: Ein Bestandteil unserer Vision ist die speicherortunabhängige Analyse von Big Data. Tableau unterstützt gegenwärtig mehr als 40 verschiedene Datenquellen sowie zahlreiche weitere Quellen, die über unsere Erweiterungsoptionen zugänglich gemacht werden können. Unsere Anwender profitieren auch von neuen Datenquellen auf dem Markt, sodass wir diese fortlaufend in unser Produkt integrieren werden, um die Hürde für den Zugriff auf Daten niedrig zu halten.

   Derzeit werden die folgenden Konnektoren im Rahmen des Big Data-Ökosystems unterstützt:

   - **Hadoop**: Cloudera Impala & Hive, Hortonworks Hive, MapR Hive, Amazon EMR mit Impala & Hive, Pivotal HAWQ, IBM BigInsights
   - **NoSQL**: MarkLogic, Datastax
   - **Spark**: Apache Spark SQL
   - **Cloud**: Amazon Redshift, Google BigQuery
   - **Betriebsdaten**: Splunk
   - **Schnelle analytische Datenbanken**: Actian Vectorwise & ParAccel, Teradata Aster, HP Vertica, SAP Hana, SAP Sybase, Pivotal Greenplum, EXASOL EXASolution

2. **Self-Service-Visualisierung von Big Data für Geschäftsanwender**: Geschäftliche Nutzer können ihre Daten mit Drag&Drop-Vorgängen visualisieren, ohne komplexe Anweisungen in SQL, Java-Code oder MapReduce schreiben zu müssen. Tableau vereinfacht die Aufgaben bei der Datenanalyse: Anwender erhalten schneller als je zuvor visuelle Erkenntnisse zu ihren Daten.

3. **Hybride Datenarchitektur zur Optimierung der Abfrageperformance**: Tableau unterstützt Direktverbindungen mit Datenquellen oder importiert Daten in den Arbeitsspeicher. Die direkte Konnektivität eignet sich optimal für Verbindungen mit schnellen, interaktiven Abfrage-Engines und großen Datensätzen. Langsamere Datenquellen lassen sich aufwerten und beschleunigen, indem ein Datenextrakt erstellt und in die speicherresidente Daten-Engine importiert wird.

4. **Datenverschmelzung für Analysen mehrerer Datenquellen**: Verteilte Daten sind häufig eine noch größere Herausforderung als Big Data. Die Daten der Analysten befinden sich nur selten in einem kompakten Datenpaket an einer Stelle. In der Regel verteilen sich die Daten auf mehrere Standorte sowie auf unterschiedliche Technologien und Plattformen. Mit Tableau können Anwender die verschiedensten Datenquellen erschließen, indem sie Big Data mit anderen Datenquellen verschmelzen von Big Data mit anderen Datenquellen (z. B. Salesforce-, MySQL-, Excel-Dateien), sodass die Datenbestände nicht verschoben werden müssen.

5. **Insgesamt verbessserte Plattform-Abfrageperformance**: Angesichts der wachsenden Datenvolumen investiert Tableau fortlaufend in grundlegende Verbesserungen der Abfrageperformance, mit denen eine Interaktion mit Daten in Echtzeit unterstützt wird. Ganz aktuell wurden die Funktionen für parallele Abfragen, Abfrageverschmelzung und Zwischenspeicherung externer Abfragen integriert. Tableau nutzt nun auch die Vektorisierung bei Prozessoren, die dafür ausgelegt sind.

6. **Leistungsfähige und einheitliche visuelle Schnittstellen mit Daten**: Tableau bietet benutzerfreundliche Analysetools für das Filtern von Daten, die Erstellung von Prognosen sowie Trendlinienanalysen. Tableau kann darüber hinaus die Benutzeraktionen deuten und die beste Möglichkeit für die Darstellung der Daten anhand visueller Best Practices auswählen. Tableau stellt zudem eine einheitliche visuelle Schnittstelle mit Daten bereit, die für alle Datenquellen gleich ist, nachdem eine Datenverbindung hergestellt wurde.

Unsere Vision steht im Einklang mit der Gesamtentwicklung der Datenlandschaft von heute. Es ist zur neuen Normalität geworden, dass viele Kunden mit den unterschiedlichsten Big Data-Technologien arbeiten. Neben den Data Warehouses sind inzwischen Technologien wie Hadoop und Spark aufgrund ihrer Speicher- und Verarbeitungsmöglichkeiten Bestandteil der Datenarchitektur. Parallel dazu zeigt sich, dass bei Data Warehouses entsprechend ihrer Hadoop-Bereitstellungen eine Restrukturierung erfolgt. Häufig werden NoSQL-Datenbanken aufgrund ihrer flexiblen Datenmodelle, der geringen Latenzzeit und dem anwendungsspezifischen Design als Backend für Anwendungen gegenüber relationalen Datenbanken bevorzugt. Allgegenwärtig sind mittlerweile auch die Cloud-Datenquellen, da Geschäftsprozesse vorzugsweise in Cloud-fähigen CRM- und ERP-Systemen verwaltet werden. Das nutzungsabhängige Abrechnungsmodell wird für Datenspeicherung und -verarbeitung in der Cloud immer beliebter. Angesichts der vielfältigen und flexiblen Backends benötigen Benutzer ein Frontend-Tool wie Tableau, das sich flexibel mit Big Data-Plattformen, Cloud-Datenquellen und relationalen Datenbanken verbinden lässt und Methoden zur raschen und flexiblen Datenanalyse bereitstellt.

## Wie arbeitet Tableau mit Big Data?

Die Hauptkomponenten von Tableau sind VizQL und die Daten-Engine. VizQL ist eine proprietäre Technologie, mit denen Anwender ihre Daten unmittelbar darstellen und ein sofortiges visuelles Feedback erhalten können. Mit VizQL steht Benutzern ein einziges Visualisierungstool zur Erstellung einer breiten Palette an grafischen Übersichten für jede Aktion zur Verfügung, beispielsweise Balkendiagramme, Liniendiagramme und Karten. Die Daten-Engine ermöglicht hingegen eine komprimierte, speicherresidente und spaltenbasierte Darstellung der Daten. Sie ist in die Direktverbindungstechnologie von Tableau integriert. Diese Technologie führt in hohem Maße optimierte, plattformspezifische SQL-Abfragen der Datenbank aus. Tableau kann so riesige Datenmengen in Echtzeit visualisieren, ohne Daten verschieben zu müssen.

In den folgenden Abschnitten werden Zugriffsverfahren und Sicherheitsvorkehrungen für Big Data mit Tableau erläutert und weitere spezielle Funktionen für Hive vorgestellt.

**Datenzugriff**

Ein ausgereiftes Verbindungsmodell ist der Schlüssel für die Arbeit mit Big Data. Unsere benannten Konnektoren für Big Data nutzen das ODBC-Protokoll sowie datenbankspezifische Funktionen durch Optimierung der gesendeten SQL-Abfragen:

**SQL-basierte Verbindungen**

Durch Nutzung von SQL verfügt Tableau über Schnittstellen zu Hadoop, NoSQL-Datenbanken und Spark. Der von Tableau generierten SQL-Anweisungen entsprechen der Norm ANSI SQL-92. Der Einsatz von SQL ist sehr lohnend, denn die Sprache ist äußerst kompakt (ein Ausdruck), quelloffen, standardisiert und es gibt keine Bibliotheksabhängigkeiten. Zudem ist SQL sehr vielfältig und ausdrucksstark. Mittels SQL lassen sich u. a. Join-Vorgänge, Funktionen, Kriterien, Zusammenfassungen, Gruppierungen und verschachtelte Vorgänge darstellen.

**ODBC**

Die von Tableau verwendeten Treiber nutzen den Programmierstandard Open Database Connectivity (ODBC) als Übersetzungsschicht zwischen SQL- und SQL-ähnlichen Datenschnittstellen, die von diesen Big Data-Plattformen bereitgestellt werden. Für Hadoop sind das Schnittstellen wie Hive Query Language (HiveQL), Impala SQL, BigSQL und Spark SQL. Zur Erzielung der bestmöglichen Leistung erfolgt eine Optimierung der generierten SQL. Aggregationen, Filterfunktionen und andere SQL-Vorgänge werden auf die Big Data-Plattformen verlagert.

**NoSQL-Schnittstellen**

Der Name weist bereits darauf hin, dass NoSQL-Datenbanken („Not only SQL“) neben relationalen Datenformaten auch Daten speichern, die nicht in relationalen Formaten modelliert wurden. Ebenso können NoSQL-Datenbanken auch SQL-ähnliche Schnittstellen unterstützen. Derzeit unterstützt Tableau MarkLogic und DataStax Enterprise als benannte Konnektoren mit SQL-ähnlichen Schnittstellen. Mit MarkLogic sind Verbindungen zu Volltextsuchen oder komplexe Suchvorgänge in unstrukturierten Daten oder alten relationalen Datensätzen möglich. Über DataStax Enterprise und Cassandra unterstützen wir eine Hive ODBC-Schnittstelle, die mit HiveQL auf den partitionierten Zeilenspeicher von Cassandra zugreift.

**Datensicherheit**

Die Implementierung unabhängiger visueller Self-Service-Analysen ist auf Unternehmensebenenur möglich, wenn Authentifizierungs- und Datenzugriffsfragen geklärt sind. Wir befinden uns zurzeit auf einem mehrstufigen Weg, um den sicheren Datenzugriff auf Big Data-bezogene Datenquellen zu ermöglichen.

Aktuell bieten wir LDAP- oder Kerberos-Authentifizierung für Tableau Desktop-Benutzer, die eine sichere Verbindung mit Hive Server 2-Clustern über Cloudera Hadoop, Hortonworks Hadoop oder MapR Hadoop benötigen. Für Verbindungen mit Cloudera Impala bieten wir mit der zusätzlichen Unterstützung für die einmalige Anmeldung und den delegierten Zugriff mit Kerberos eine Erweiterung der vorhandenen Unterstützung mit nativem Active Directory, SAML und dem integrierten Authentifizierungssystem von Tableau. Anwender profitieren von einer nahtlosen Benutzererfahrung, da Benutzer aufgrund der Anmeldung auf ihrem lokalen Rechner keine erneute Anmeldung bei Tableau Server oder anderen Impala-Live-Datenquellen benötigen. Für IT-Administratoren stellt die Kompatibilität von Tableau mit Apache den Schutz vertraulicher Daten sicher, denn Benutzer sehen nur die Daten, zu deren Nutzung sie berechtigt sind. Dank der Zusammenarbeit von Tableau und Cloudera bei der Delegierung von Benutzerrechte für Impala können wir sicherstellen, dass Benutzer über eine stabile, automatisierte Backend-Authentifizierung auf Impala als Live-Datenquelle zugreifen können. Zukünftig ist die Unterstützung für die einmalige Anmeldung und den delegierten Zugriff über Kerberos für viele weitere Datenquellen geplant.

**Spezielle Funktionen in Hadoop Hive**

Hadoop ist nahezu zum Synonym für Big Data-Technologie geworden. Hadoop erweitert zudem die Datenverarbeitungsmöglichkeiten, die im Vergleich zu herkömmlichen Datenbanken auf dem Speicher-Layer durchführbar sind, erheblich. Tableau bietet somit eine Reihe einzigartiger Funktionen für Verbindungen mit Hadoop Hive. Dazu gehören:

- **XML-Verarbeitung**: Tableau stellt eine Reihe benutzerdefinierter Funktionen für die Verarbeitung von XML-Daten mit XPath bereit. Mit diesen Funktionen können Benutzer Inhalte extrahieren, einfache Analysen durchführen und XML-Daten filtern.
- **Web- und Textverarbeitung**: Zusätzlich zu den XPath-Operatoren bietet die Hive-Abfragesprache viele Möglichkeiten für die Arbeit mit gängigen Webelementen und Textdaten:
    - **JSON-Objekte**: Zum Abrufen von Datenelementen aus Zeichenfolgen, die JSON-Objekte enthalten.
    - **URLs**: Zum Extrahieren von Komponenten einer URL (z. B. Protokolltyp oder Hostname) oder zum Abrufen des Werts, der mit einem bestimmten Abfrageschlüssel in einer Liste mit Schlüssel- und Wertparametern zugeordnet ist.
    - **Textdaten**: Zum Suchen und Ersetzen von Text in Hive von Tableau.
- **On-the-Fly ETL**: Dank benutzerdefinierter SQL können Benutzer ihre Datenverbindungen mit komplexen Join-Bedingungen, Vorfilterfunktionen und Vorab-Aggregationen definieren.

- **SQL-Anfangsdaten**: Benutzer können mehrere SQL-Anweisungen angeben, die unmittelbar nach Einrichtung einer Datenverbindung zusammen ausgeführt werden sollen, um vorwiegend Performancemerkmale zu optimieren oder eine benutzerdefinierte Datenverarbeitungslogik zu erstellen.
- **Benutzerdefinierte Analyse mit UDFs und MapReduce**: Tableau ermöglicht Benutzern die Implementierung von UDFs, benutzerdefinierten Aggregatsfunktionen und willkürlichen SQL-Ausdrücken von Hive mithilfe von „Durchlauffunktionen“. Diese Funktionen werden in der Regel als Java Archive (JAR-Dateien) ausgestaltet, die für das gesamte Hadoop-Cluster kopiert werden können. Mit benutzerdefinierter SQL verfügen Benutzer außerdem über explizite Steuerungsmöglichkeiten für MapReduce-Vorgänge.

## Nutzungsszenarien: Tableau und Big Data

Big Data-Einsteiger werden entdecken, dass es zwei wesentliche Szenarien für den Einsatz von Tableau für ihre Datenbestände gibt: Datenuntersuchung und Datenvisualisierung.

**Datenuntersuchung**

Unternehmen erfassen und speichern alle Arten von Daten, ohne im Vorfeld den Analysezweck zu bestimmen, da sie annehmen, dass die Daten künftig hilfreiche Erkenntnisse ermöglichen werden. Daten in Webprotokollen, Serverprotokollen, Clickstreams, sozialen Medien und Sensordaten werden nicht mehr verworfen, sondern auf Datenplattformen wie Hadoop erfasst. Diese Vorgehensweise ist flexibel und bietet einen Spielraum für experimentelle Analysen. Mit Tableau lassen sich allgemeine Datentrends ausloten und visualisieren, bevor erhebliche Ressourcen in Produktionszwecke investiert werden.

Bei EMC wird Tableau eingesetzt, um die vom Leistungssensor erfassten Daten auf Hadoop zu analysieren. Tom Hudgins, Solution Engineer bei EMC, erläutert: „Meine Datenbank enthält ca. 70 Billionen Zeilen, die mit Tableau analysiert werden. Wir analysieren mit Tableau intelligente Zählerdaten – Informationen zum Energieverbrauch, die von Haushalten und Unternehmen zurückkommt. Unternehmen, die sich durch diesen Informationsfluss arbeiten und die wertvollen Informationen herausfiltern können, die zu neuen Einsichten und Erkenntnisse führen, über die vorher vielleicht noch nie nachgedacht wurde, machen den Unterschied zwischen Erfolg und Misserfolg aus.“

**Datenvisualisierung**

Nach der Festlegung, welche Analysen ermöglicht oder durchgeführt werden sollen, sollte sich die Datenvisualisierungsstrategie auf die Performanceoptimierung konzentrieren. Reaktionsschnelle Dashboards und die Live-Kommunikation mit Daten sind nur möglich, wenn die richtige Menge von Daten und Detailgenauigkeit aus den Big Data herausgefiltert werden kann. IT-Administratoren und Tableau-Benutzer haben Tools zur Verfügung, die Daten an der richtigen Detailtreue ausrichten. Ebenso wichtig ist eine Datenverarbeitungsplattform, die interaktive Analysen unterstützt. Best Practices für die Performancemaximierung:

- Schnelle interaktive Abfrage-Engine
- Benutzerdefinierte Anpassung der Verbindungsperformance für Live-Abfragen
- Optimierung der Extrakte durch Übersichts-, Filter- und Sampling-Funktionen
- Nutzung der Best Practices für Datenbanken, wie etwa Partitionierung

Das Team von Rosenblatt Securities hat seinen Ansatz für Big Data wiederholt getestet und optimiert. „Mit Tableau haben wir in einem Team von 5 Mitarbeitern Aufgaben erledigt, für die bereits 50 Mitarbeiter sehr viel Zeit gebraucht hätten“, sagt Scott Burrill, Partner und Managing Director. „Wir erstellen für 800 Wertpapiere Prognosen, um in Echtzeit zu ermitteln, ob wir uns an einem Einstiegs- oder Ausstiegszeitpunkt befinden. Wir können extrem schnell abgeleitete Analysen für Hunderttausende verschiedene Felder durchführen und diese einbringen, visualisieren, auswerten und zu Geschichten verarbeiten. Tableau hat uns ein vollständig neues Feld eröffnet, denn wir können auf Ergebnisse reagieren, ohne dass wir wie früher ein Sampling durchführen müssen. Wir haben Einblick auf ganze Populationen von Daten.“

## Zusammenfassung

Das Zeitalter der Big Data ist angebrochen. Die Menge der Daten nimmt ständig und immer schneller zu. Um die neue Normalität im Datenbereich bewältigen zu können, verlagern viele Unternehmen ihre Dateninfrastruktur auf Hadoop, Spark, NoSQL und schnelle analytische Datenbanken. Tableau kann Anwendern wie den Mitarbeitern bei EMC und Rosenblatt Securities für die täglichen Aufgaben visuelle Einblicke in Big Data ermöglichen.

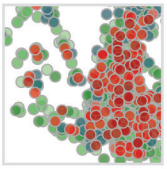
## Über den Autor

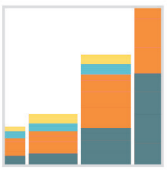
**Jeff Feng – Product Manager, Tableau Software (@jtfeng)**

Jeff Feng ist Produktmanager bei Tableau Software und zuständig für Roadmap, Strategie und die Entwicklung neuer Funktionen im Bereich Big Data-Produkte. Seine Produkte revolutionieren die Arbeit mit Daten. Vor seiner Anstellung bei Tableau war Jeff Feng als Managementberater bei McKinsey & Co tätig, wo er Fortune 500-High-Tech-Unternehmen zu Unternehmens-, Technologie- und Produktstrategien beriet, sowie Programmmanager bei Apple bei der Markteinführung des iPhone 4. Jeff Feng besitzt einen MBA-Abschluss der MIT Sloan School of Management sowie einen Master- und Bachelor-Abschluss der University of Illinois.

## Über Tableau

Tableau unterstützt die Anwender dabei, ihre Daten anschaulich und verständlich aufzubereiten. Mit Tableau analysieren und visualisieren die Nutzer vorhandene Informationen blitzschnell und teilen die Ergebnisse mit anderen. Mehr als 26.000 Unternehmen weltweit nutzen Tableau im Büro und unterwegs für schnelle Analysen. Zehntausende Nutzer verwenden Tableau Public, um anderen Personen Daten in Blogs und auf Websites zur Verfügung zu stellen. Laden Sie die kostenlose Testversion herunter und erleben Sie, wie Tableau Sie unterstützen kann: www.tableau.com/de-de/trial.





## Ähnliche Whitepapers

Fünf Best Practices für Tableau und Hadoop

Sieben Tipps für den Erfolg mit Big Data

Fostering a Data-Driven Culture:
A Special Report from the Economist Intelligence Unit and Tableau

Big Data: The Next Industrial Revolution

Tableau Software und Big Data

Aberdeen Group: Maximizing the Value of Analytics and Big Data

Alle Whitepapers anzeigen

## Weitere Ressourcen

· Kostenlose Testversionen herunterladen

· Produkt-Demo

· Schulungen und Lernprogramme

· Community und Support

· Kundenberichte

· Lösungen