

TOP TEN
Big Data
TRENDS

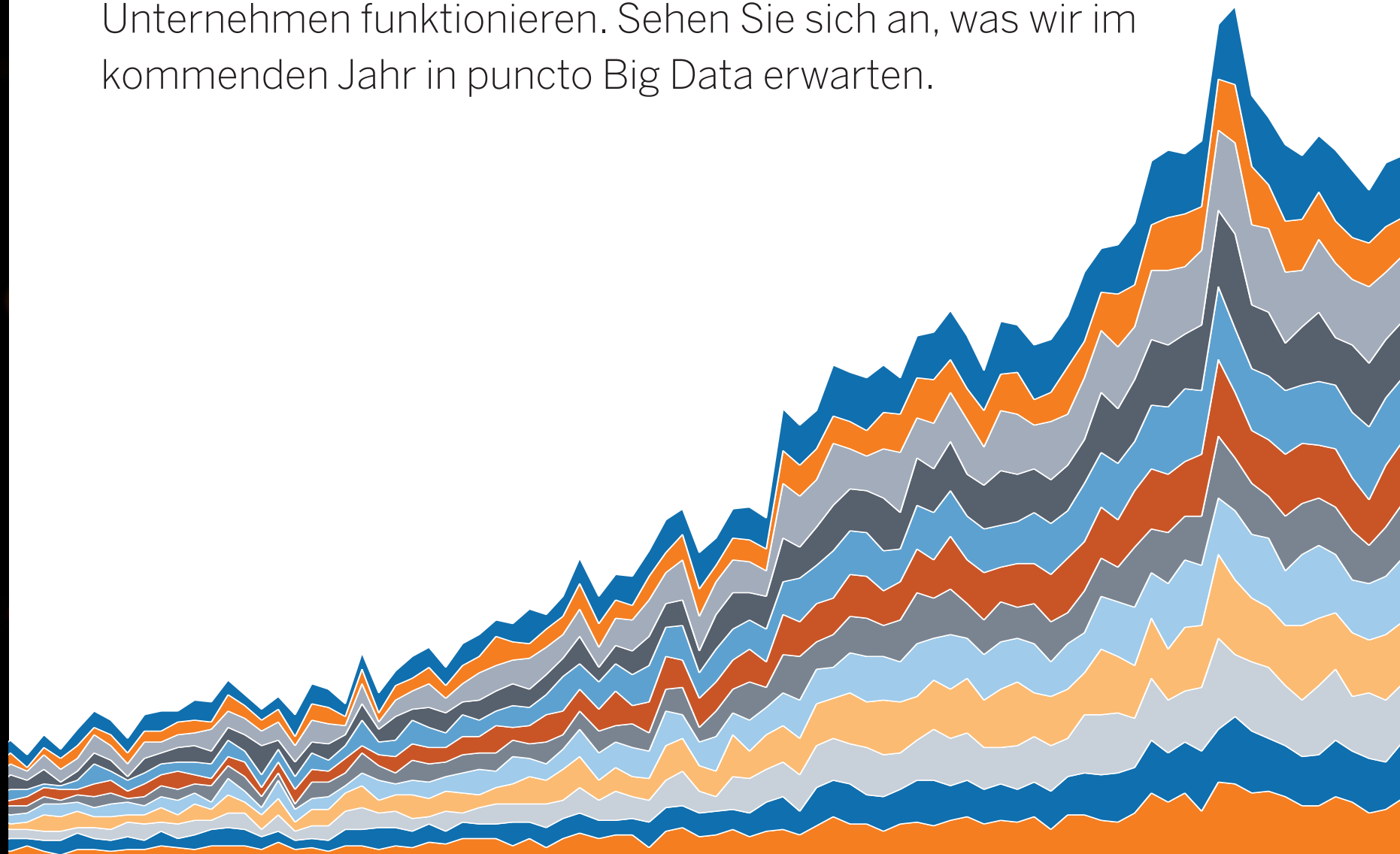




Die 10 wichtigsten Big Data-Trends

Immer mehr Organisationen speichern, verarbeiten und generieren Wert aus Daten in allen erdenklichen Formen und Mengen. Es werden immer mehr Systeme genutzt, die große Volumen von strukturierten und unstrukturierten Daten unterstützen. Der Markt wird Plattformen fordern, die den Hütern der Daten helfen, Big Data zu regulieren und zu sichern, und gleichzeitig den Endbenutzer diese Daten analysieren lassen. Diese Systeme werden reifen, sodass sie gut innerhalb der IT-Systeme und -Standards der Unternehmen funktionieren. Sehen Sie sich an, was wir im kommenden Jahr in puncto Big Data erwarten.

Jedes Jahr stoßen wir bei Tableau einen Dialog über die Entwicklungen in der Branche an, um eine Vorstellung davon zu bekommen, welche Big Data-Trends das folgende Jahr bestimmen werden. Unsere Vorhersagen fürs kommende Jahr:



BIG DATA

1

Big Data wird schnell und leichter zugänglich: Mehr Optionen, um Hadoop zu beschleunigen

Sicher, auf Hadoop ist es möglich, das maschinelle Lernen zu nutzen und Sentiment-Analysen durchzuführen, aber als Erstes wird häufig die Frage gestellt: Wie schnell ist die interaktive SQL? SQL ist letztendlich die Datenleitung der Geschäftsanwender, die Hadoop-Daten für schnellere, leichter wiederholbare KPI-Dashboards und Untersuchungsanalysen nutzen wollen.

Dieser Wunsch nach Schnelligkeit treibt die Einführung beschleunigter Datenbanken voran. Beispiele hierfür sind [Exasol](#) und [MemSQL](#), Hadoop-basierte Speicher wie [Kudu](#) sowie Technologien, die raschere Abfragen ermöglichen. Diese Abfragebeschleuniger verwenden SQL-on-Hadoop-Engines ([Apache Impala](#), [Hive LLAP](#), [Presto](#), [Phoenix](#) und [Drill](#)) sowie OLAP-on-Hadoop-Technologien ([AtScale](#), [Jethro Data](#) und [Kyvos Insights](#)) und verwischen die Trennlinie zwischen herkömmlichen Data Warehouses und der Big Data-Welt.

Big Data ist nicht mehr nur Hadoop: Für Hadoop maßgeschneiderte Tools werden ausgemustert

In früheren Jahren konnten wir mit der Big Data-Welle den Aufstieg mehrerer Technologien beobachten, die dem Analysebedarf auf Hadoop entsprachen. Aber Unternehmen mit komplexen, heterogenen Umgebungen wollen keinen isolierten BI-Zugriffspunkt für eine einzige Datenquelle (Hadoop) mehr einführen. Antworten auf ihre Fragen sind in einer Vielzahl von Quellen verborgen: von Aufzeichnungssystemen über Cloud-Warehouses bis hin zu strukturierten und unstrukturierten Daten, sowohl von Hadoop- als auch von Nicht-Hadoop-Quellen. (Übrigens bereiten sich auch relationale Datenbanken auf Big Data vor. SQL Server hat beispielsweise kürzlich JSON-Support hinzugefügt.)

Die Kunden werden die Analyse aller Daten fordern. Daten- und quellenunabhängige Plattformen werden florieren, während für Hadoop maßgeschneiderte Plattformen, die sich nicht für verschiedene Anwendungsfälle bereitstellen lassen, außen vor bleiben werden. Der [Verkauf von Platfora](#) dient als Frühindikator für diesen Trend.



Unternehmen nutzen von Anfang an sog. Data Lakes, um Werte zu schaffen

Ein Data Lake ist wie ein künstlicher Stausee. Zunächst wird ein Damm gebaut (ein Cluster), der dann mit Wasser gefüllt wird (Daten). Nachdem der See angelegt wurde, wird das Wasser (die Daten) für verschiedene Zwecke genutzt, etwa um Elektrizität zu erzeugen, es zu trinken und zur Erholung (prädiktive Analysen, ML, Netzsicherheit usw.).

Bis jetzt war das Füllen des Sees Selbstzweck. Dies wird sich ändern, weil die wirtschaftliche Rechtfertigung für Hadoop schwieriger wird. Unternehmen fordern eine wiederholbare und agile Nutzung des Sees für schnellere Antworten. Sie werden die Wirtschaftlichkeit sorgfältig prüfen, bevor sie in Personal, Daten und Infrastruktur investieren. Dies stärkt die Partnerschaft zwischen **Geschäftsabteilung und IT**. Und Selfservice-Plattformen gewinnen Anerkennung als Tools, um von Big Data zu profitieren.

4

Architekturen reifen und drängen Universal-Frameworks zurück

Hadoop ist nicht mehr nur eine Plattform mit Stapelverarbeitung für Datenwissenschaftler. Es ist zu einer Mehrzweck-Engine für Ad-hoc-Analyse geworden. Es wird sogar zur Betriebsberichterstattung für tägliche Arbeitslasten eingesetzt, was herkömmlicherweise von Data Warehouses erledigt wurde.

Organisationen decken diesen Hybridbedarf, indem sie auf fallspezifische Architekturkonzepte setzen. Sie suchen nach den verschiedensten Faktoren, etwa Personas, Fragen, Volumen, Zugriffshäufigkeit, Datengeschwindigkeit und Aggregationsniveau, bevor sie sich auf eine Datenstrategie festlegen. Diese modernen Architekturen werden bedarfsgesteuert sein. Sie kombinieren die besten Selfservice-Tools zur Datenaufbereitung, Hadoop Core und Analyseplattformen für Endbenutzer auf eine Weise, die sich neu konfigurieren lässt, wenn sich der Bedarf weiterentwickelt. Die Flexibilität dieser Architekturen wird letztendlich entscheidend sein für die Wahl der Technologie.



5

Vielfalt, nicht Volumen oder Geschwindigkeit, ist entscheidend für Big Data-Investitionen

Gartner definiert Big Data mit drei Vs: High-Volume, High-Velocity, High-Variety (Datenbestände mit hohem Volumen, hoher Geschwindigkeit und großer Vielfalt). Während alle drei Vs wachsen, wird die Vielfalt zur größten Triebkraft von Big Data-Investitionen, wie aus einer von New Vantage Partners **kürzlich durchgeführten Umfrage** hervorging. Dieser Trend wird sich fortsetzen, da die Unternehmen mehr Quellen integrieren und sich auf die „**lange Schleppe**“ **der Big Data konzentrieren wollen**. Vom schemafreien JSON über verschachtelte Typen in anderen Datenbanken (relational und NoSQL) bis zu strukturierten Daten (Avro, Parquet, XML) vervielfältigen sich die Datenformate und die Konnektoren werden entscheidend. Unternehmen werden Analyseplattformen auch weiterhin danach beurteilen, ob sie Direktverbindungen zu diesen verschiedenen Quellen herstellen können.

Spark und Machine Learning lassen Big Data erstrahlen

Apache Spark, ursprünglich eine Komponente des Hadoop-Systemumfelds, wird jetzt zur Big Data-Plattform erster Wahl für Unternehmen. Bei einer **Umfrage** unter Datenarchitekten, IT-Managern und BI-Analysten bevorzugten fast 70 % der Befragten Spark gegenüber dem etablierten MapReduce, das stapelorientiert und nicht für interaktive Anwendungen oder die Verarbeitung von Echtzeit-Streams geeignet ist.

Diese Fähigkeit zu großen Rechenoperationen mit Big Data haben Plattformen hervorgebracht, die rechenintensives Machine Learning, KI und grafische Algorithmen bieten. Microsoft Azure ML war aufgrund seiner Anfängerfreundlichkeit und der einfachen Integration mit vorhandenen Microsoft-Plattformen besonders erfolgreich. Die Bereitstellung von ML für die Massen wird mehr Modelle und Anwendungen zur Folge haben, wodurch Petabytes von Daten generiert werden. Wenn Maschinen lernen und Systeme intelligent werden, wird sich das Augenmerk auf Anbieter von Selfservice-Software richten, um zu sehen, wie sie den Zugriff der Endbenutzer auf diese Daten ermöglichen.

Die Konvergenz von IoT, Cloud und Big Data schafft neue Gelegenheiten für Selfservice-Analysen

Bald wird alles mit einem Sensor ausgestattet sein, der Daten zurück in die Zentrale schickt. IoT generiert Riesenmengen an strukturierten und unstrukturierten Daten, und ein wachsender Anteil dieser **Daten wird von Cloudservices bereitgestellt**. Die Daten sind häufig heterogen und befinden sich auf zahlreichen relationalen und nicht relationalen Systemen von Hadoop-Clustern bis zu NoSQL-Datenbanken. Während Innovationen bei Speicherung und Managed Services den Erfassungsprozess beschleunigt haben, sind der Zugriff auf die Daten und das Verstehen der Daten selbst immer noch die größte Herausforderung auf der letzten Meile. Das Resultat ist, dass die Nachfrage nach Analysetools ansteigt, die nahtlos eine Verbindung zu einer Vielzahl von in der Cloud gehosteten Datenquellen herstellen und diese kombinieren können. Mit solchen Tools können Unternehmen sämtliche Datentypen untersuchen und visualisieren, ganz gleich, wo sie gespeichert sind. Auf diese Weise lassen sich versteckte Geschäftsgelegenheiten in der IoT-Investition aufdecken.

Selfservice-Datenaufbereitung wird zum Mainstream, wenn die Endbenutzer beginnen, Big Data zu gestalten

Geschäftsanwendern den Zugriff auf Hadoop-Daten zu gestatten, ist eine der größten Herausforderungen unserer Zeit. Der Boom bei den Selfservice-Analyseplattformen hat diese Umstellung erleichtert. Aber die Geschäftsanwender wollen Zeitaufwand und Komplexität der Datenaufbereitung für die Analyse noch weiter reduzieren. Dies ist besonders wichtig, wenn es um eine Vielzahl von Datentypen und -formaten geht.

Agile Selfservice-Tools zur Datenaufbereitung ermöglichen es nicht nur, dass Hadoop-Daten direkt an der Quelle aufbereitet werden können, sondern stellen die Daten auch als Momentaufnahmen bereit, damit sie schneller und einfacher untersucht werden können. Wir haben in diesem Bereich eine Vielzahl von Innovationen seitens Unternehmen beobachtet, die sich auf die Big Data-Datenaufbereitung durch Endbenutzer konzentrieren, wie etwa [Alteryx](#), [Trifacta](#) und [Paxata](#). Diese Tools senken die Einstiegsbarrieren für [Hadoop-Nachzügler](#) und [-Späteinsteiger](#). Sie werden auch weiter an Zugkraft gewinnen.

Big Data wird erwachsen: Hadoop als Beitrag zu Unternehmensstandards

Wir beobachten einen wachsenden Trend, dass Hadoop zu einem Kernstück des IT-Umfelds im Unternehmen wird. Die Investitionen in die Sicherheits- und Governance-Komponenten im Umfeld der Unternehmenssysteme werden zunehmen. Apache Sentry bietet ein System zur Durchsetzung detaillierter, rollenbasierter Autorisierung für Daten und Metadaten, die in einem Hadoop-Cluster gespeichert sind. [Apache Atlas](#), das als Teil der Data Governance-Initiative entwickelt wurde, ermöglicht den Organisationen die Anwendung einer konsistenten Datenklassifizierung im gesamten Datensystemumfeld. [Apache Ranger](#) bietet eine zentralisierte Sicherheitsverwaltung für Hadoop.

Mittlerweile erwarten die Kunden derartige Funktionen von ihren RDBMS-Plattformen auf Unternehmensebene. Diese Funktionen stehen im Vordergrund neu aufkommender Big Data-Technologien und beseitigen eine weitere Hürde für die Einführung im Unternehmen.

Das Aufkommen der Metadatenkataloge hilft den Mitarbeitern, Big Data zu finden, deren Analyse sich lohnt

Über einen langen Zeitraum vernichteten Unternehmen Daten, weil zu viel verarbeitet werden musste. Mit Hadoop können Sie viele Daten verarbeiten, aber die Daten sind im Allgemeinen nicht so organisiert, dass sie gefunden werden können.

Metadaten können den Benutzern helfen, relevante Daten zu entdecken und zu verstehen, deren Analyse mit Selfservice-Tools sich lohnt. Diese Lücke beim Kundenbedarf wird von Unternehmen wie [Alation](#) und [Waterline](#) ausgefüllt, die maschinelles Lernen einsetzen, um das Auffinden von Daten in Hadoop zu automatisieren. Die Katalogdateien verwenden Tags, entdecken Beziehungen zwischen Datenbeständen und liefern sogar Abfragevorschläge über durchsuchbare Benutzeroberflächen. Dies hilft sowohl den Datenkonsumenten als auch den Dateneigentümern dabei, den Zeitaufwand zu reduzieren, um Daten vertrauenswürdig zu machen, sie zu finden und präzise abzufragen. Im kommenden Jahr wird die Selfservice-Untersuchung als natürliche Erweiterung der Selfservice-Analyse an Bedeutung gewinnen und sich einer größeren Nachfrage erfreuen.



Über Tableau

Die Integration der Datenvisualisierung in Ihre Einzelhandelssysteme und -prozesse ist einfacher als Sie denken.

Tableau Software unterstützt Ihre Mitarbeiter dabei, Daten zu sehen und zu verstehen – unabhängig von ihrem Umfang oder davon, in wie vielen unterschiedlichen Systemen sie gespeichert sind. Das nahtlose Zusammenspiel von PC und iPad ermöglicht dabei das schnelle Verknüpfen, Zusammenführen, Visualisieren und Teilen von Datendashboards. Erstellen und veröffentlichen Sie Dashboards mit automatisierten Datenupdates und teilen Sie sie mit Kollegen, Partnern oder Kunden – ganz ohne Programmierkenntnisse. Testen Sie Tableau jetzt kostenlos.

[TABLEAU.COM/DE-DE/TRIAL](https://tableau.com/de-de/trial)