redis

Latenz ist der neue Systemausfall

WARUM GESCHWINDIGKEIT DER NEUE SCHLÜSSELFAKTOR IST

Zusammenfassung

Da Unternehmen zunehmend Anwendungen nutzen, die auf verschiedenen Infrastrukturen, in verschiedenen Regionen und bei verschiedenen Anbietern gehostet werden, gerät die Geschwindigkeit, mit der Endbenutzer auf diese Anwendungen zugreifen können, unter Druck.

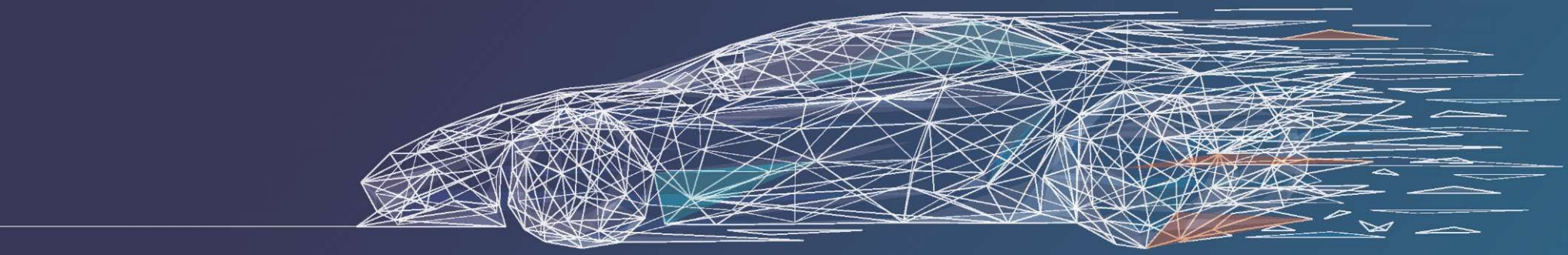
Hinzu kommt die Tatsache, dass Anwendungen heutzutage aus einer beträchtlichen Anzahl verschiedener Komponenten bestehen, was zu einer Verschlechterung der Benutzerfreundlichkeit führt.

Diese beiden Eigenschaften—die Modularität der Anwendungen und die Komplexität der Infrastruktur—können direkt zu einer schlechteren Anwendungsleistung führen.

Aus diesem Grund ist die Geschwindigkeit der Datenschicht, der gemeinsamen horizontalen Schicht der Anwendung, von entscheidender Bedeutung.

Die Verwendung einer geografisch replizierten Datenschicht bei gleichzeitiger Vermeidung von Problemen mit Dateninkonsistenzen ist eine Herausforderung, die alle IT-Verantwortlichen lösen müssen.

Durch den Einsatz einer Datenschicht, die Ihre Daten über Clouds und die ganze Welt hinweg vereinheitlicht, können Organisationen einige der inhärenten Einschränkungen überwinden, mit denen Technologie-Teams seit Jahrzehnten zu kämpfen haben, und so ihren Endnutzern bessere Erfahrungen bieten.



Einleitung

Digitale Teams haben die letzten zehn Jahre damit verbracht, die permanente Verfügbarkeit ihrer digitalen Ressourcen sicherzustellen, und dies ist ihnen auch weitestgehend gelungen. Hohe Verfügbarkeit ist jetzt die Norm.

Unternehmen haben dieses hohe Niveau an Digitalisierung und Verfügbarkeit zum Teil dadurch erreicht, dass sie die Vorteile der Cloud für sich nutzen: einfache Skalierbarkeit, modulare Dienste und verfeinerte Architekturmuster. Alle diese Eigenschaften ermöglichen positive Ergebnisse, wenn auch mit der Kehrseite einer gesteigerten Komplexität. Diese Komplexität wirkte sich zunächst am stärksten auf die Verfügbarkeit aus und führte zu dem, was wir die ‚Epoche der Verfügbarkeit' nennen. Nachdem Organisationen jedoch immer besser verstehen, wie sie eine hohe Verfügbarkeit liefern können, stellen sie fest, dass es noch andere Probleme zu lösen gibt.

Mittlerweile geht die Epoche der Verfügbarkeit ihrem Ende zu, da immer mehr Unternehmen die Verringerung der Latenzzeit als nächsten Schlüssel zur Erschließung der von ihnen angestrebten Ergebnisse ins Auge fassen.

Sie verstehen zunehmend, dass langsame Produkte und Dienste genauso gut überhaupt nicht verfügbar sein könnten—dass Latenz der neue Systemausfall ist.

Leider ist das Lösen von Latenzproblemen oft schwieriger als das Gewährleisten einer hohen Verfügbarkeit. Während diese durch geeignete Technik, ein höheres Maß an Redundanz sowie eine bessere Überwachung und Sichtbarkeit verbessert werden kann, unterliegt das Phänomen der Latenz den Gesetzen der Physik.

Um die Latenzzeit so weit wie möglich zu verringern, müssen Organisationen das Wesen und die Einflussfaktoren der Latenz verstehen und darüber hinaus über klare, eindeutige Leitlinien verfügen, um die Latenzzeit für die Benutzer ihrer Anwendungen und Webseiten so weit wie möglich zu verringern.

Wenn zu viel Latenz heute einem Systemausfall gleichkommt, finden Sie hier das nötige Wissen, um die niedrigste Latenz zu erzielen, die physikalisch erreicht werden kann.

Organisationen erkennen zunehmend, dass langsame Produkte und Dienste genauso gut überhaupt nicht verfügbar sein könnten—dass Latenz den neuen Systemausfall darstellt.

Nicht die Großen fressen die Kleinen, sondern die Schnellen fressen die Langsamen

Schnelles Handeln ist die neue Norm. Während die Devise früher Analyse und Sorgfalt lautete, sieht die betriebliche Realität heute so aus, dass Unternehmen schneller als je zuvor innovative Lösungen finden müssen, um der Konkurrenz voraus zu sein. Die operative Landschaft für jede Organisation verändert sich rasant und Erfolg wird denjenigen beschieden sein, die am besten auf diese Dynamik reagieren können.

WARUM AGILITÄT SO WICHTIG IST: ZEIT IST VON ENTSCHEIDENDER BEDEUTUNG

Unsere heutige Welt unterscheidet sich deutlich von der Situation vor nur wenigen Jahren, und das Tempo der Veränderungen wird immer rasanter. Vor diesem Hintergrund ist es wichtiger denn je, Organisationen die Fähigkeit zu geben, schnell zu reagieren. Es ist wichtig, die Veränderungen in der Gesellschaft zu verstehen, um besser begreifen zu können, wie entscheidend schnelles Handeln tatsächlich ist.

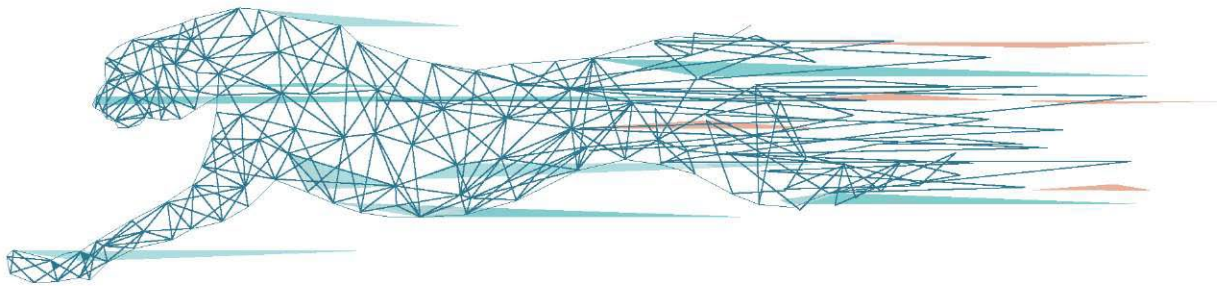
Bereits 2011 schrieb Marc Andreessen, ein bekannter Unternehmer, Investor, Vorstands-mitglied und Erfinder des Netscape-Webbrowsers, einen inzwischen berühmten Meinungsbeitrag für ***The Wall Street Journal*** und erläuterte, warum Software die Welt verschlingt.

In seinem Essay stellte Andreessen seine Theorie über die Größe, den Umfang und die Geschwindigkeit dieses Wandels vor und vertrat die Ansicht, dass „wir uns mitten in einem dramatischen und weitreichenden technologischen und wirtschaftlichen Wandel befinden, in dem Softwareunternehmen im Begriff sind, große Teile der Wirtschaft zu übernehmen."

Während dieser Wandel jedoch tatsächlich von großer Bedeutung ist, kommt der Geschwindigkeit dieses Wandels die größte Relevanz zu. Grundlegend für die Fähigkeit, schnell agieren zu können, ist die Fähigkeit, Änderungen vorzunehmen, eine Vielzahl von Tools zu nutzen und das beste Endnutzererlebnis zu bieten. Wenn das alles sehr nach der Arbeitsweise von Unternehmen im Silicon Valley klingt, ist das verständlich. Tatsache ist, dass viele der Unternehmen, die mit Erfolg traditionelle Branchen umkrempeln, immer mehr wie unternehmerische Technologieunternehmen aussehen.

Es ist fast zehn Jahre her, dass Andreessen diesen Aufsatz geschrieben hat, und viele seiner Vorhersagen haben sich bestätigt. Es ist zwar zu einem Klischee geworden, Tesla, Uber, Lyft, Netflix und Airbnb als Beispiele für die digitale Disruption anzuführen, doch es steht außer Frage, dass Führungskräfte sowohl im Taxi- als auch in Gastgewerbe von einer Flutwelle ungeahnten Ausmaßes überrollt wurden. Jenseits des Klischees ist es jedoch wichtig, festzustellen, wie viel Aufwand diese Unternehmen betreiben, um auf ihren Anwendungen das schnellstmögliche Kundenerlebnis zu bieten: Geschwindigkeit ist in der Tat wichtig.

Es ist eine Binsenweisheit, dass schnelles Handeln in einem organisatorischen Umfeld von der Bereitstellung digitaler Erfahrungen abhängt, die ihrerseits diese Geschwindigkeitsmerkmale aufweisen. Latenz ist der neue Hemmschuh für organisatorischen Wandel.



Die Welt verändern auf dem Weg der Digitalisierung

Unzählige Unternehmen mit traditioneller Organisationsstruktur knüpfen ihre Hoffnungen auf zukünftigen Erfolg an den Prozess der Digitalisierung. Es lohnt sich, ein paar Beispiele näher zu betrachten, um ein Gefühl für die Größenordnung zu bekommen

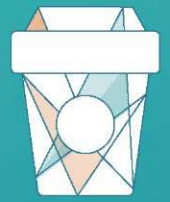
DIGITAL JOE: STARBUCKS SETZT AUF DIGITAL-FIRST

Starbucks CEO Kevin Johnson war einst Führungskraft bei Microsoft. Seine Erfahrungen bei dem Technologie-Giganten, der ebenfalls in Seattle ansässig ist, halfen ihm, digitales Denken in seiner neuen Funktionen bei einer ganz anderen Art von Unternehmen anzuwenden.

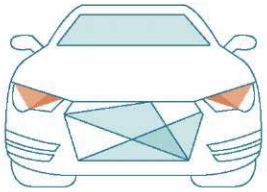
Johnson spricht in Bezug auf den digitalen Weg von Starbucks Klartext: „Wo andere versuchen, eine mobile App zu entwickeln, hat Starbucks eine durchgängige Kundenplattform geschaffen, die auf Loyalität basiert."

Die wichtigste digitale Innovation des Unternehmens dreht sich um seine mobile App zum Bestellen und Bezahlen. Die Fokussierung auf die App ist im Grunde eine kundenorientierte Strategie, da sie die grundlegenden Bedürfnisse der Konsumenten anspricht: Komfort, Vermeidung von Warteschlangen, Schnelligkeit bei der Abwicklung und so weiter. In Verbindung mit dem umfangreichen Treueprogramm bietet die App Starbucks den perfekten Rahmen für Zusatzverkäufe und die Vermarktung an die Verbraucher/innen. Ebenso wichtig ist, dass die App riesige Mengen an Nutzerdaten an das Unternehmen zurückleitet, wodurch es die Gewohnheiten und Wünsche seiner Kunden besser verstehen kann.

Starbucks hat massiv in die Schaffung von digitalen Berührungspunkten für seine Kunden investiert und mit seiner enormen globalen Präsenz war die Verfügbarkeit der Anwendungen—sowohl in Bezug auf die reine Betriebszeit als auch auf die Latenzzeit—von entscheidender Bedeutung.



Die Fokussierung auf die App ist im Grunde eine kundenorientierte Strategie, da sie auf die grundlegenden Bedürfnisse der Verbraucher/innen eingeht.

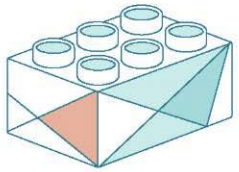


AUDI: AUTOMOBILHERSTELLER ODER DIGITALES UNTERNEHMEN?

Die ohnehin hart umkämpfte Automobilbranche steht kurz- bis mittelfristig vor massiven Umwälzungen. Neue Vertriebsmodelle, der Aufstieg der Elektrofahrzeuge und das autonome Fahren verändern die Spielregeln für die Automobilhersteller. Angesichts dieser Herausforderungen hat Audi die Absatzmethoden für seine seine Fahrzeuge verändert.

Die 2012 eingeführte Audi City bietet ein tiefgreifendes Markenerlebnis, das den Besuchern das virtuelle Erkunden der gesamten Audi-Produktpalette ermöglicht, sogar in innerstädtischen Geschäften, die nicht über genügend Platz für Ausstellungsräume verfügen.

Audi ist eine Luxusmarke, und der Schritt des Unternehmens, den eigenen Vertriebskanal zu sprengen, wurde nicht auf die leichte Schulter genommen. Audi investierte viel in den Aufbau eines virtuellen Einzelhandelserlebnisses, das so authentisch wie ein physisches sein sollte. Teil dieses Prozesses war die Nutzung verschiedener unterschiedlicher Berührungspunkte, Anwendungsformfaktoren und Anzeigekonzepte. All dies mit den Erwartungen der Nutzer an ein geschmeidiges Erlebnis in Einklang zu bringen, war eine technologische Herausforderung, die ein völlig neues Denken erforderte.



LEGO: VON PLASTIKBAUSTEINEN ZU DIGITALEN BLÖCKEN

Die LEGO Gruppe ist der berühmte dänische Hersteller des gleichnamigen Kinderspielzeugs. Doch nach einer langen Expansionsphase von 1970 bis 1991 erlebte LEGO von 1992 bis 2004 einen stetigen Rückgang des Geschäfts. Im Jahr 2004 stand das Unternehmen kurz vor dem Konkurs.

Als LEGO auf der Kippe stand, war das Unternehmen gezwungen, eine umfassende Umstrukturierung einzuleiten. Seine digitaleTransformation konzentrierte sich auf die Erschließung neuer Einnahmequellen über Filme, Spiele für Mobilgeräte und mobile Anwendungen.

Als LEGO diesen Prozess in Angriff nahm, war, eines der zentralen Probleme, mit denen sich das Unternehmen konfrontiert sah, die Belastung seiner Performance durch Zehntausende von Kindern, die gleichzeitig ihre verschiedenen LEGO-Apps und Spiele nutzten. Das Management entschied, dass Schnelligkeit—sowohl bei der Innovation als auch bei der Lieferung seiner digitalen Produkte—eine nicht-verhandelbare Bedingung darstellte.

Das Management bei LEGO schrieb vor, dass Schnelligkeit—sowohl bei der Innovation als auch bei der Bereitstellung seiner digitalen Produkte—eine nicht verhandelbare Bedingung darstellte.

Die zwei Epochen digitaler Bereitstellung

Organisationen haben in Bezug auf die digitale Bereitstellung zwei Epochen erlebt. Zuerst mussten sie sich mit der Epoche der Verfügbarkeit auseinandersetzen. Heute, da die Frage der Verfügbarkeit weitgehend gelöst ist, beginnt die Epoche der Schnelligkeit.

DIE EPOCHE DER VERFÜGBARKEIT: BETRIEBSZEIT IST ENTSCHEIDEND

Mit dem Aufkommen des Internets und der Gründung von Unternehmen wie Amazon, eBay und Netflix begannen Konzerne damit, das Potenzial dieser neuen Technologien und Geschäftsmodelle zu erkunden. In den Anfangstagen der digitalen Transformation verfolgten IT-Teams vor allem eine einzige Kennzahl: die Betriebszeit. Organisationen auf dem Weg in die digitale Welt hatten einen Fokus: sicherzustellen, dass ihre Webseiten und Anwendungen überall und zu jeder Zeit verfügbar waren. Diese Phase, die wir als „Epoche der Verfügbarkeit" bezeichnen, war durch Tools und Konzepte zur Gewährleistung der Zuverlässigkeit von Webseiten gekennzeichnet.

Die Verfügbarkeitsepoche förderte eine enorme Innovation und alles in dem Bestreben, die Anzahl der 9er-Ziffern in der Metrik für die prozentuale Betriebszeit zu erhöhen. Die Zusammenführung von Entwicklungsabteilung und operativem Betrieb in die kombinierte DevOps-Rolle sollte die Entwicklung von Anwendungen beschleunigen und die Zuverlässigkeit erhöhen. Leistungsstarke Werkzeuge und Plattformen zur Überwachung von Anwendungen und Infrastrukturen wurden geschaffen, um diesen heiligen Gral zu erreichen: immer höhere Prozentsätze für die Betriebszeit in einer sich immer schneller entwickelnden Umgebung.

Auch wenn das Ziel „fünf mal neun" leicht über die Lippen kommet, ist es wichtig zu verstehen, was 99.999 % Betriebszeit tatsächlich bedeuten: nicht mehr als 26 Sekunden Ausfallzeit pro Monat. Immer mehr Unternehmen nähern sich solchen Betriebszeitstatistiken an oder erreichen sie sogar durch hochwertige technologische Lösungen und ein eingehendes Verständnis dessen, was zur Vorbeugung gegen Betriebsausfälle erforderlich ist. Daher können sich die ClOs auf die Optimierung anderer Bereiche konzentrieren. Infolgedessen gewinnen Bereiche, die einst ignoriert wurden, nun an entscheidender Bedeutung.

Da das Problem der Verfügbarkeit weitgehend gelöst ist, beginnt für die Unternehmen nun die ‚Epoche der Schnelligkeit'.

DIE STATISTIK ALS HINWEIS AUF DAS ENDE DER EPOCHE DER VERFÜGBARKEIT

Unternehmen agierten während der letzten zehn oder sogar zwanzig Jahre unter der Prämisse, dass für die Zunahme von digitalen Berührungspunkten mit Kunden die grundlegende Verfügbarkeit dieser Berührungspunkte ausschlaggebend sei. Eine ganze Generation von IT-Fachkräften war besessen von den Metriken der Verfügbarkeit und den Tools für ihre Optimierung.

Es gibt jedoch einige fundamentale Aspekte, die für diese Praktiker eine neue Wendung der Verhältnisse mit sich bringen. Abgesehen von der gestiegenen Komplexität, die sie durch ihre eigenen Bemühungen um die Betriebszeit herbeigeführt haben, existieren außerdem externe Faktoren, die kritische Voraussetzungen für eine möglichst geringe Latenzzeit schaffen.

Mit der massenhaften Nutzung mobiler Touchpoints ändert sich auch die Art und Weise, wie die Verbraucher Daten konsumieren sowie ihre Anforderungen an die Unmittelbarkeit. Verbraucherinnen und Verbraucher nutzen ihre mobilen Geräte, um sich besser über die Produkte und Dienstleistungen zu informieren, die für sie wichtig sind. 80 % der Verbraucher suchen nach Produktinformationen, Bewertungen und Preisen auf ihrem Smartphone, während sie in einem Geschäft einkaufen.

Und dieser Trend, Informationen zu konsumieren, ist nur der Anfang, denn Verbraucherinnen und Verbraucher tätigen auch Transaktionen auf neue Weise. Ein Drittel aller Einkäufe während der Weihnachtseinkaufssaison 2018 wurde über Smartphones abgewickelt.

Leider neigen Unternehmen dazu, ihre eigene Fähigkeit zur Bereitstellung guter Erfahrungen zu überschätzen. Untersuchungen von Qualtrics ergaben, dass zwar 60 % der Unternehmen davon ausgehen, ein gutes mobiles Erlebnis zu bieten, aber nur 22 % der Verbraucher/innen sehen dies ebenso.

All dies deutet auf den Bedarf nach Schnelligkeit hin. Mobiles Surfen findet in verschiedenen Kontexten statt – beim stationären Surfen, beim Gehen, in Geschäften und in kurzen Pausen—all diese Situationen verlangen mehr denn je nach Schnelligkeit.

EINE MAHNENDE GESCHICHTE: DISNEYS VERUNGLÜCKTE MARKTEINFÜHRUNG

Im vergangenen Jahr setzte Disney einen Großteil seines zukünftigen Erfolgs auf die Einführung von Disney+, dem hochkarätigen Videostreaming-Dienst des Unternehmens. Wie viele andere Unternehmen, die in einem Bereich außerhalb ihrer üblichen Liefermethoden von sich reden machen wollen, hat auch Disney+ die Markteinführung groß angekündigt und die Kunden auf das bevorstehende Erlebnis eingestimmt.

Unglücklicherweise begannen die Kunden unmittelbar nach dem Start von Disney+ damit, sich über die schlechte Performance des Dienstes zu beschweren: Verlängerte Pufferzeiten, Ausfälle und generelle Verzögerungen verhinderten, was ein fröhlicher Eröffnungstag hätte werden können. Die Kritik war eindeutig: Ein Service, der mit schlechter Geschwindigkeit liefert, ist genauso schlecht wie einer, der gar nicht erst verfügbar ist.

DIE EPOCHE DER SCHNELLIGKEIT: WER RASTET, HAT SCHON VERLOREN

In den letzten fünf Jahren haben die meisten Unternehmen ein gutes Verständnis für Betriebszeiten erworben. In der Zwischenzeit haben ihre Dienstanbieter viel unternommen, um mehrere Redundanzen in ihre Plattformen einzubauen und sicherzustellen, dass der Weg zu einer nahezu perfekten Verfügbarkeit einfach zu beschreiten ist. Werkzeuge zur Überwachung, Verfahren zur Entwicklung der Standortzuverlässigkeit und die Bemühungen um Ausfallsicherheit im Falle von unvermeidlichen Störungen haben geholfen, das zu gewährleisten, was Endnutzer heutzutage erwarten: Webseiten und Anwendungen, die immer verfügbar sind, wenn sie gebraucht werden.

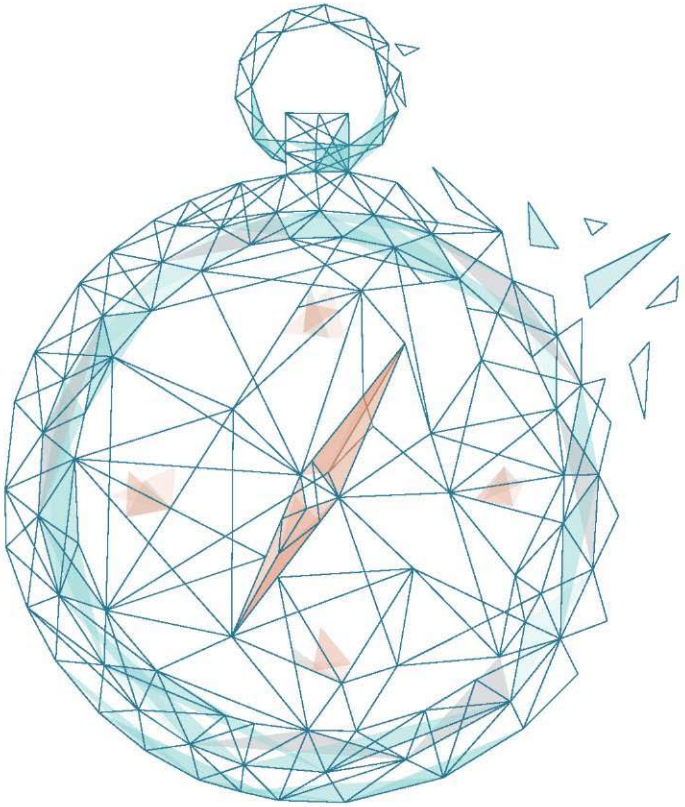
Allerdings haben alle diese zusätzlichen technischen Entwicklungen und der Ausbau immer komplexerer Architekturen in dem Bestreben, die widerstandsfähigsten Anwendungen bereitzustellen, neue Herausforderungen mit sich gebracht, die ebenso gravierend sind wie die Betriebszeit.

Wir befinden uns eindeutig auf dem Weg in eine zweite Epoche, eine Phase, in der Zuverlässigkeit zu einem Grundpfeiler geworden ist, während Schnelligkeit jetzt das Differenzierungsmerkmal im Wettbewerb darstellt. Kundenentscheidungen, die früher mit Zeit und Analyse getroffen wurden, erfolgen zunehmend in Sekundenschnelle. Und wenn Ihre Webseite mehr als diesen Sekundenbruchteil zum Hochladen benötigt oder Ihr Streaming-Dienst unter Ladehemmungen und Pufferzeiten leidet, haben Sie schon verloren.

Falls Sie annehmen, dass die Unzufriedenheit der Kunden sich nicht auf ihre Konsumgewohnheiten auswirken wird, sollten Sie dies noch einmal überdenken. Wie in einem Forbes-Artikel aus dem Jahr 2019 berichtet wurde (Wie schnell ist schnell genug? Mobile Ladezeiten beeinflussen das Kundenerlebnis und den Umsatz): „Eine langsam ladende Seite auf einem mobilen Gerät stellt nicht nur die Geduld der Verbraucher auf die Probe. Sie kann das „Versagen" bei der Kundenerfahrung sein, das Sie einen Verkauf kostet. Dies ist die wichtigste Erkenntnis aus dem Bericht zur Schnelligkeit von Webseiten. Diese Studie, die sich mit den Meinungsäußerungen von 1150 Verbrauchern und Unternehmen befasst, zeigte, dass die Geschwindigkeit einer Website einen entscheidenden Faktor für das Kaufverhalten darstellt."

Zudem sind die Auswirkungen einer unzureichenden Seitengeschwindigkeit sind nicht unbedeutend: „Fast 70 % der Konsumenten geben an, dass die Seitengeschwindigkeit ihre Kaufbereitschaft beeinflusst. Darüber hinaus verringert eine langsame Ladezeit auch die Wahrscheinlichkeit, dass sie in Zukunft wiederkommen. Eine Aufschlüsselung der Daten zeigt, dass 22 % der Käuferinnen und Käufer erklärten, sie würden die Registerkarte schließen, 15 % sagten, sie würden die Webseite eines Konkurrenten besuchen und 12 % würden einer Freundin/einem Freund von ihrer negativen Erfahrung berichten.

Wenn die neue Epoche durch die Notwendigkeit definiert wird, eine möglichst geringe Latenzzeit zu gewährleisten, über welche Dinge müssen Unternehmen dann nachdenken, um dieses Ziel zu erreichen?



Bereitstellung von Schnelligkeit in einer komplexen Welt

In seinem bahnbrechenden Beitrag aus dem Jahr 2013 über The Composable Enterprise™, beschrieb Jonathan Murray, ehemaliger CTO der Warner Music Group, die Zukunft der Technologie im Kontext der Forderungen von Unternehmen nach Schnelligkeit und Agilität. Basierend auf seiner lebenslangen Erfahrung mit der Umsetzung der digitalen Strategien von Großunternehmen, beschrieb Murray das ‚Composable Enterprise' (Unternehmen als Komposition) folgendermaßen: „Geschäftsfunktionen, Prozesse, Organisationen, Lieferantenbeziehungen und Technologie müssen als Bausteine betrachtet werden, die je nach Bedarf neu konfiguriert werden können, um der sich ändernden Wettbewerbslandschaft gerecht zu werden."

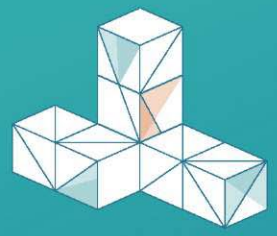
Dieses neue Komponenten-Betriebsmodell (COM, Component Operating Model) erfordert einen, Lego-Baustein'-Ansatz für die Entwicklung und Implementierung von Prozessen und der Organisationen, die diese unterstützen. Die Umsetzung eines COM-basierten Ansatzes wird tiefgreifende Auswirkungen auf die Struktur von Organisationen und das Wesen der Arbeit haben.

Geschäftskonzepte auf der Grundlage von COM werden die traditionellen IT-Infrastrukturen und Organisationen erheblich belasten. Unsere derzeitigen IT-Dienstleistungen wurden für ein statisches—und funktional oft isoliertes—Betriebsmodell entwickelt. IT muss dynamisch viel anpassungsfähiger werden, um mit dem heutigen Geschäftstempo Schritt halten zu können.

„Ein neuer Komponenten-Architekturmodell-Ansatz (CAM) für IT-Infrastruktur, Anwendungen und Dienstleistungen wird erforderlich sein, um sicherzustellen, dass IT in der Lage ist, die geschäftlichen Anforderungen zu erfüllen. Die Zeit zwischen der Ermittlung eines Geschäftsbedarfs und der Bereitstellung der erforderlichen IT-Lösung muss Stunden und Tage betragen, statt Monate und Jahre."

Dieser vor einigen Jahren geschriebene, vorausschauende Beitrag von Murray beschreibt die neue Normalität in Organisationen. In den letzten Jahren hat sich die Art und Weise, wie Infrastrukturen genutzt und Anwendungen erstellt werden, grundlegend verändert. Mit dem Aufkommen von Containern, Microservices Architekturen, diskreten modularen Anwendungstools und dergleichen bedeutet die Aufrechterhaltung der Funktionsfähigkeit einer Anwendung ein Jonglieren mit Dutzenden von Diensten, Regionen, geografischen Gebieten, Dienstanbietern und mehr.

Während diese Kompositionsfähigkeit die Produktivität der Entwickler und die organisatorische Agilität fördert, hat sie ihren Preis. Es scheint, dass die Bereitstellung einer niedrigen Latenzzeit unter diesen Bedingungen ein Hirngespinst ist.



In den letzten paar Jahren hat sich die Art und Weise, wie Infrastrukturen genutzt und Anwendungen entwickelt werden, grundlegend verändert.

Schnelle Daten in einer dezentralen Datenverarbeitungsumgebung

Wie wir aus Murrays bahnbrechender Arbeit über die Kompositionsfähigkeit in modernen Anwendungen und Infrastrukturen wissen, gibt es keinen einfachen monolithischen Datenstapel mehr, auf dem Anwendungen aufgebaut werden. In dem Bestreben, Entwicklern und ihren Unternehmen die größtmögliche Flexibilität und Geschwindigkeit zu bieten, nutzen wir stattdessen eine große Anzahl modularer Entwicklerdienste, unterschiedliche Infrastrukturmuster, verschiedene Hosting-Ansätze und eine massive geografische Verteilung von Anwendungen. Gleichzeitig versuchen wir, diese Anwendungen so schnell wie möglich für Benutzer auf der ganzen Welt bereitzustellen.

In dieser Zeit massiver Komplexität könnte man leicht annehmen, es gäbe kein gemeinsames Gefüge, auf das sich Unternehmen stützen können – so als sei ihre Welt ständig im Fluss und kontinuierlichen Veränderungen unterworfen.

Es gibt jedoch einen roten Faden, der sich durch all die verschiedenen Angelegenheiten von Unternehmen zieht, und das sind ihre Daten. Indem wir eine Datenschicht als konsistenten und einheitlichen Faden betrachten, der von allen anderen Teilen des Stapels genutzt wird, ermöglichen wir es Unternehmen, Sinn in das Chaos zu bringen.

Und durch die Wahl einer Datenschicht, die für dezentrale Umgebungen entwickelt wurde, schnellste Verarbeitungszeiten aufweist und branchenführende Belastbarkeit bietet, können wir genau das liefern, was ein Unternehmen verlangt.

Ein Schlüsselfaktor, mit dem Unternehmen sicherstellen können, dass ihre Anwendungen sowohl belastbar als auch schnell sind, ist das Arbeiten innerhalb eines konsistenten Datenschichtkonstrukts. Und eine konsistente Datenbasis beginnt mit einer Datenbank, die scheinbar unmögliche Ziele erreichen kann: Verteilte Architekturen, Konsistenz, Flexibilität und Schnelligkeit.

Moderne Ansätze zur Verringerung der Latenzzeit

Wie wir gesehen haben, werden Anwendungen zunehmend unter Verwendung von Microservices aufgebaut: Sie nutzen eine Vielzahl verschiedener Komponenten mit unterschiedlichen Ansätzen für die Infrastruktur, werden an verschiedenen Orten gehostet, von Menschen überall genutzt und auf viele verschiedene Plattformen verteilt.

Da sich Daten an so vielen Orten befinden und über so viele verschiedene Netzwerke übertragen werden, sind auch die vielen Gelegenheiten für das Enstehen von Datenkonflikten kaum verwundernlich. Zum Umgang mit diesen Konflikten wurden konfliktfreie replizierte Datentypen (CRDT) entwickelt, mit denen Daten über mehrere Standorte hinweg repliziert werden können.

Mit CRDTs können einzelne Replikate unabhängig und gleichzeitig aktualisiert werden, ohne dass sie durch Abstimmung koordiniert werden müssten.

Ohne CRDTs können gleichzeitige Aktualisierungen mehrerer Replikate der gleichen Daten zu Inkonsistenzen führen, da keine Koordinierung zwischen den Computern erfolgt, auf denen die Replikate gehostet werden.

Mit CRDTs können jedoch alle Ungereimtheiten, die sich aus diesem verteilten Ansatz ergeben, gelöst werden. CRDTs wurden zunächst in Situationen eingesetzt, in denen Massenverteilung die Norm darstellt (wie z. B. Online-Chatsysteme, Internet-Glücksspiele, Audio- und Videostreaming). Sie werden aber zunehmend auch in allgemeineren Anwendungen verwendet.

Es steckt viel Technologie dahinter, die gewährleistet, dass ein CRDT funktioniert, aber am einfachsten kann man es sich so vorstellen, dass ein CRDT eine Datenschicht bereitstellt, auf der Replikate autonom agieren und dennoch Konsistenz geboten wird.

Da sich Daten an so vielen Orten befinden und über so viele verschiedene Netzwerke übertragen werden, sind auch die vielen Gelegenheiten der Entstehung von Datenkonflikten kaum verwunderlich.

Weiterführung des Cache

Bei traditionellen Datenbankmodellen ist der Datenbankstandort vom Cache getrennt. Stellen Sie sich die Datenbank als die städtische Zentralbibliothek und den Cache als die lokale Bibliotheksfiliale vor, in der die beliebtesten Bücher aufbewahrt werden, um die häufigsten Ausleihwünsche zu erfüllen. Wenn die Liste der beliebtesten Bücher unverändert bleibt, mag das gut funktionieren, doch wenn sich die Lesegewohnheiten ändern und neue Bücher in der Gunst der Leserinnen und Leser auftauchen und wieder verschwinden, wird es schwieriger.

Und diese Idee der schnellen Überprüfung verschiedener Informationseinheiten gegenüber einer konstanten Basis ist genau die Metapher für moderne Anwendungen—die gesamte Zusammensetzbarkeit, von der Murray gesprochen hat, führt dazu, dass von vielen unterschiedlichen Diensten und Orten aus zu verschiedensten Zeitpunkten auf Daten in der Datenbank zugegriffen werden muss.

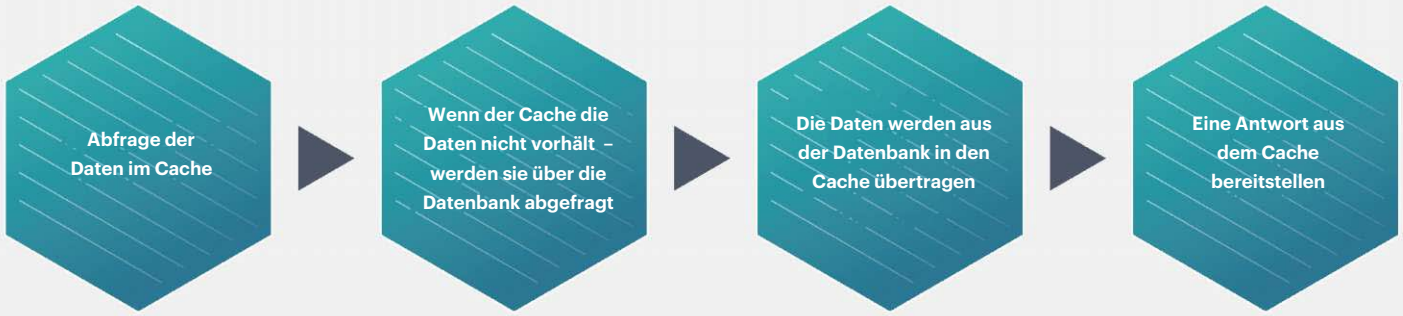
In einer Welt mit immer mehr diskreten Diensten und daher immer mehr Stellen, an denen etwas schief gehen kann, ist das traditionelle Modell nicht ideal (siehe das nachstehende Diagramm). In Anwendungen, wo es bei den Datenmodellen eher um die häppchenweise Übertragung vieler Informationsbrocken geht, ist das Cachemodell unter Umständen nicht der schnellste Weg, um Daten dorthin zu bringen, wo sie benötigt werden.

Hier kommt das Konzept einer einzelnen Datenschicht ins Spiel—durch die Nutzung einer einzelnen Schicht als Ersatz für die Datenbank/Cache-Kombination wird die Komplexität der Datenschicht reduziert. Im Gegenzug wird die verteilte und modulare Anwendung, die heute die Norm darstellt, erheblich beschleunigt.

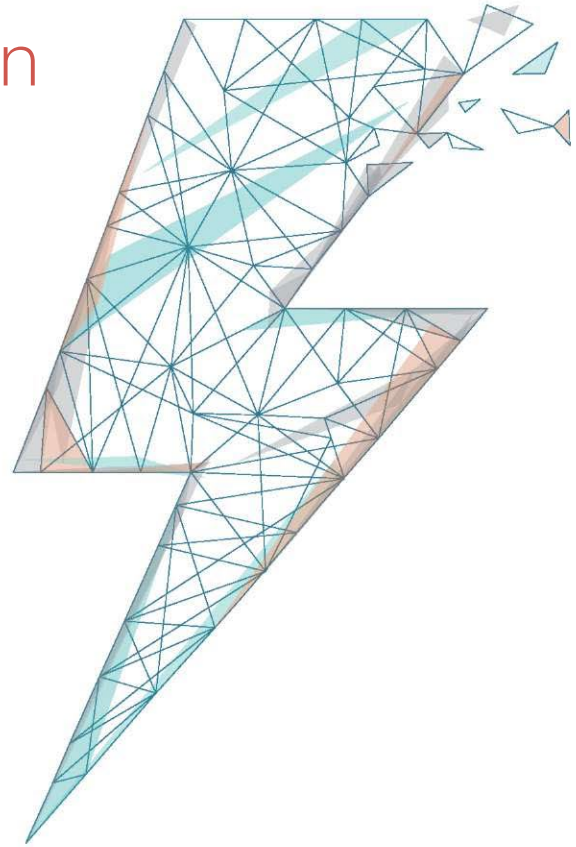
Der zusätzliche Vorteil besteht darin, dass die Verringerung der Anzahl der Komponenten auf der Datenschicht auch die Latenzzeit verringert. Während individuelle Komponenten einer komplexen Datenschicht schnell sein können, reduziert ein einziger Datenspeicher die Anzahl der Netzwerk-Hops, was die Dinge unweigerlich verlangsamt.

Anstelle eines Cache nutzen daher viele moderne Datenbanken In-Memory-Techniken, bei denen Datenspeicher anstelle von externen Festplatten für die Speicherung verwendet werden. Dies ist von entscheidender Bedeutung, da die Geschwindigkeit nicht durch mehrere Speicherebenen eingeschränkt wird, wenn alles im Datenspeicher gesichert ist. Bei einem Cache-basierten Modell wird der im Cache gespeicherte Datenbestand zum Flaschenhals, der die Gesamtgeschwindigkeit drosselt.

In einem traditionellen Modell bedeutet ein Datenzugriff für die Anwendung den folgenden Handlungsablauf:



Schnelligkeit: Ein Byte nach dem anderen



Traditionelle Datenbanken sind, wie wir oben gesehen haben, auf externe Speicherung für ihren Cache angewiesen. Bis vor kurzem und in den letzten 50 Jahren erfolgte die Speicherung auf physischen, rotierenden Festplatten und daher waren die meisten traditionellen Datenbankkonzepte für diesen Zweck optimiert.

Da Festplatten jedoch physische Geräte sind, unterliegen sie den einschränkenden Bedingungen der physischen Welt. Um diese physischen Einschränkungen zu umgehen, wurde eine Reihe von Betriebsbeschränkungen geschaffen. Es ist zwar ein technischer Umweg, aber die Details der physikalischen Festplattentechnologie wirken sich stark auf die Datenbankgeschwindigkeit aus.

Der springende Punkt ist jedoch, dass die Grenze zwischen moderner Speicherung und Arbeitsspeicher immer unschärfer wird. Das Vorrücken von Solid-State-Laufwerken (SSDs) und anderen neuen Speicherkonzepten bedeutet, dass diese technischen Behelfslösungen, entwickelt für eine Welt, die den physikalischen Geschwindigkeitsbeschränkungen mechanischer Geräte unterliegt, nicht länger erforderlich sind. Es bedeutet auch, dass die Speicherung in mehrere Ebenen aufgeteilt werden kann, so dass alle Daten in einem schnellen Speicher aufbewahrt werden können und kein separater Cache mehr benötigt wird.

Das Nettoergebnis für diejenigen, die versuchen, Schnelligkeit zu erreichen, ist eine schnellere Datenschicht, auf der unsere Anwendungen aufsetzen können.

Das Aufkommen neuer Speicherkonzepte bedeutet, dass technische Behelfslösungen, entwickelt für eine Welt, die den physischen Geschwindigkeitsbeschränkungen mechanischer Geräte unterliegt, nicht länger erforderlich sind.

Skalierung ohne Geschwindigkeitseinbußen

Es ist schön und gut, eine Anwendung zu entwickeln, die bei begrenzter Nutzung schnell läuft, aber was passiert, wenn Ihr Durchsatz enorm ansteigt? Dieses Problem erhofft sich jeder Anwendungsentwickler, der auf Popularität und Verbreitung setzt.

Skalierung erfolgt jedoch auf zweierlei Weise—***vertikal,*** bezogen darauf, wie viele Daten über die Datenschicht hinweg übertragen werden, aber auch ***horizontal,*** bezogen auf die schiere Menge der vorhandenen Informationen.

Organisationen müssen eine Datenschicht aufbauen, die diese Formen der Skalierung schrittweise und nahtlos zulässt. Hierzu müssen eine Reihe von unterschiedlichen Faktoren berücksichtigt werden: die Fähigkeit, die Datenschicht an mehreren Orten auszuführen und verschiedene Arten von Arbeitsspeicher und Speicher zu verwenden, die Fähigkeit, Daten je nach Häufigkeit ihrer Nutzung in Datenebenen anzuordnen und schließlich die Möglichkeit zur globalen Skalierung.

Wenden wir uns nun diesem letzten Bereich zu. Diese gesamte Kapazität, im Speicher zu arbeiten und zu sichern, ist vorteilhaft, doch wenn sich Ihre Anwendung weltweit verbreiten soll, bleibt Ihnen dann immer noch das gleiche niedrige Latenzniveau erhalten?

WIR LEBEN IN EINER MEHRKERNWELT

Moderne Datenverarbeitung erfolgt zunehmend über einen Mehrkernkontakt. Der Begriff „Mehrkern“ (Multi-Core) bezieht sich einfach auf Rechenvorgänge, bei denen zwei oder mehr individuelle Recheneinheiten innerhalb einer CPU vorhanden sind. Die an die CPU gesendeten Anweisungen können auf separaten Kernen gleichzeitig verarbeitet werden, was die Gesamtgeschwindigkeit erhöht.

Die Nutzbarmachung von Multi-Core-Architekturen kann eine Herausforderung darstellen. Organisationen, die eine höchst skalierbare Datenschicht nutzen möchten, müssen sich darüber Gedanken machen. Ist Ihre Datenschicht in der Lage, auf einem einzigen Cluster horizontal zu skalieren, um die beste Skalierung mit der geringsten Latenz bereitzustellen?

VERTIKALE SKALIERUNG

HORIZONTALE SKALIERUNG

Organisationen müssen eine Datenschicht einrichten, die diese Skalierung schrittweise und nahtlos ermöglicht.

EIN BESUCH BEIM CAP-THEOREM

Da dieses Papier unweigerlich von denjenigen verwendet werden wird, die global verteilte Anwendungen mit einer Leistung erstellen wollen, die derjenigen lokal angesiedelter Anwendungen ähnelt, lohnt sich ein Blick auf einige Einschränkungen im Zusammenhang mit dezentralen Datenschichten.

Vor etwa 20 Jahren entwickelte der Informatiker Eric Brewer das CAP-Theorem, das sich auf verteilte Anwendungen und speziell auf die Daten bezieht, die von diesen Anwendungen erzeugt und verwendet werden.

Das CAP-Theorem besagt in einfachster Form, dass jedes vernetzte System mit gemeinsam genutzten Daten nur zwei von drei wünschenswerten Eigenschaften haben kann; Konsistenz (C) äquivalent zu einer einzigen aktuellen Kopie der Daten; hohe Verfügbarkeit (A) dieser Daten (für Aktualisierungen); und Partitionstoleranz der Netzwerke (P).

So gesehen bedeutete das CAP-Theorem in diesen frühen Tagen, in denen es nur um Schnelligkeit ging, dass diejenigen Ansätze, die am ehesten die größte Schnelligkeit und die höchste Anwendungsverfügbarkeit (Netzwerkpartitionen und Hochverfügbarkeit) bieten würden, auch Dateninkonsistenz zur Folge haben mussten.

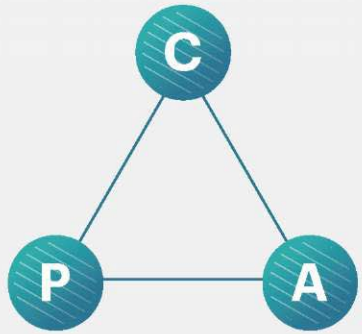
In den Jahrzehnten seit der Einführung des CAP-Theorems wurden jedoch neue Ansätze für den Umgang mit verteilten Systemen entwickelt, die dieses theoretisch unmögliche Kunststück vollbringen können: Datenkonsistenz, Verfügbarkeit und Partitionstoleranz. Das Aufkommen neuer Datenansätze bedeutet, dass wir niedrige Latenzzeiten haben können, ohne auf Datenkonsistenz verzichten zu müssen.

Dies ist zwar nicht der Ort für eine definitive technische Abhandlung, aber es ist wichtig, dass diejenigen, die Verantwortung für die Anwendungen ihres Unternehmens tragen, die Grundzüge der Funktionsweise moderner Anwendungen verstehen.

Wie bereits erwähnt, wird es in einer Welt, in der Anwendungen zwangsläufig verteilt sind, mehrere Knoten geben, die in vielen einzelnen Anwendungen enthalten sind. In dieser Mehrknoten-Situation gibt es zwei allgemeine Optionen: **Aktiv-Passiv-Daten** oder **Aktiv-Aktiv-Daten.**

DAS CAP-THEOREM

KONSISTENZ
Gleichbedeutend mit dem Vorhandensein einer einzigen aktuellen Kopie der Daten



PARTITIONEN
Die Toleranz gegenüber Netzwerk-Partitionen

VERFÜGBARKEIT
Die hohe Verfügbarkeit jener Daten

Das Aufkommen neuer Datenansätze bedeutet, dass wir niedrige Latenzzeiten haben können, ohne auf Datenkonsistenz verzichten zu müssen.

VEREINHEITLICHTE DATENSCHICHT

Datenebenen, der Teil der Software, der die Datenanfragen verarbeitet, können entweder Aktiv Aktiv oder Aktiv-Passiv sein.

Aktiv-Aktiv (manchmal auch dual-aktiv genannt) ist ein Ansatz, bei dem jeder Knoten Zugang zu einer replizierten Datenbank hat, wodurch jeder Knoten Zugang zu einer einzelnen Anwendung hat und diese nutzen kann. Diese Technologie ermöglicht die Konsistenz der Daten für Ihre Anwendungen über verschiedene Umgebungen hinweg (Server, Hybrid, Multi-Cloud) und sogar für Anwendungen, die über die ganze Welt verteilt sind. In einem Aktiv-Aktiv-System wird die Belastung aus allen Anfragen auf die gesamte verfügbare Verarbeitungskapazität symmetrisch aufgeteilt. Bei einem Störfall auf einem Knoten nimmt ein anderer Knoten im Netzwerk dessen Platz ein.

Ein Aktiv-Aktiv-Cluster besteht in der Regel aus mindestens zwei Knoten, die beide gleichzeitig die gleiche Art von Dienst aktiv ausführen. Da mehr Knoten für den Dienst zur Verfügung stehen, wird auch eine deutliche Verbesserung des Durchsatzes und der Antwortzeiten im Vergleich zu einem Aktiv-Passiv-Ansatz erzielt.

AKTIV-PASSIV

Ein Aktiv-Passiv-Cluster besteht ebenfalls aus mindestens zwei Knoten. Wie die Bezeichnung ‚Aktiv-Passiv' jedoch andeutet, sind nicht alle Knoten aktiv.

In einem Cluster mit zwei Knoten muss zum Beispiel, wenn der erste Knoten bereits aktiv ist, der zweite Knoten passiv oder im Bereitschaftsmodus sein. Der passive (auch als Failover bezeichnete) Knoten dient als Reserve und ist bereit, den Betrieb zu übernehmen, sobald der aktive (auch als primär bezeichnete) Server nicht mehr arbeiten kann oder die Verbindung unterbrochen wurde.

Wenn Klienten eine Verbindung zu einem Zwei-Knoten-Cluster in einer Aktiv-Passiv-Konfiguration herstellen, verbinden sie sich mit nur einem Server. Mit anderen Worten: Alle Klienten verbinden sich mit demselben Server. Wie bei der Aktiv- Aktiv-Konfiguration ist es wichtig, dass die beiden Server genau dieselben Einstellungen haben. Dies wird Redundanz genannt und stellt sicher, dass sich Daten nahtlos zwischen den Knoten vervielfältigen können.

Wenn Änderungen an den Einstellungen des Primärservers vorgenommen werden, müssen diese Änderungen auf den Failover-Server übertragen werden. Wenn also der Failover-Server übernimmt, können die Klienten keinen Unterschied feststellen.

Wenn die Latenz der neue Störfall ist, dann sind die Latenzwerte eindeutig umso niedriger, je näher sich ein Knoten am Anwendungsnutzer befindet. Wir müssen daher einen Weg finden, um Anwendungen global zu verteilen (da die Verlagerung von Knoten in die Nähe von Anwendungsnutzern die Latenz verringert) und gleichzeitig Konsistenz gewährleisten. Glücklicherweise haben wir in dieser Hinsicht einige Unterstützung.

AUF SCHNELLIGKEIT AUSGELEGT

Konfliktfreie Replikation ist ein Konzept, das die Existenz mehrerer Kopien (Replikate) von Daten an mehreren Orten auf konsistente Weise ermöglicht. Es ist eine sehr wichtige Methode, um niedrige Latenzzeiten für verteilte Anwendungen zu gewährleisten, aber es gilt noch andere Aspekte zu berücksichtigen. Wie bereits erwähnt, speichern moderne Datenbanken, die für die niedrigste Latenzzeit moderner Anwendungen konzipiert sind, Daten im Arbeitsspeicher. Durch den Wegfall eines externen Cache können wir den Umfang des erforderlichen Datenverkehrs reduzieren.

Während herkömmliche Datenbanken für Anwendungsfälle konzipiert wurden, bei denen 10 oder 100 Millisekunden für die Verarbeitungszeit akzeptabel waren, ist in der heutigen Welt, in der für Anwendungen sofortige Reaktionszeiten gefordert werden, eine Leistung im Bereich von weniger als einer Millisekunde eine Notwendigkeit.

VERSAGEN IST OKAY, WENN ES SCHNELL VONSTATTEN GEHT

Failover ist, wie der Name schon sagt, ein automatisiertes System, bei dem für den Fall, dass ein Knoten aus irgendeinem Grund ausfällt, ein anderer, replizierter Knoten die Lücke schließt. Failover lässt sich zwar leicht einrichten, aber die Geschwindigkeit dieses Failovers bestimmt die Auswirkungen des Ausfalls für den Endnutzer.

Um die niedrigste Latenz in einer Welt zu gewährleisten, in der Knotenausfälle unvermeidlich sein können, ist es wichtig, dass die Mehrknoten-Datenschicht die Ausfallsicherung so schnell wie möglich bereitstellen kann.

Zusammenfassung

In der modernen Welt müssen Organisationen, die auf die Bereitstellung von digitalen Erlebnissen angewiesen sind, sicherstellen, dass ihre Stakeholder die Anwendungen nutzen können, wann und wo immer sie wollen. Die Benutzer von heute verlangen jedoch nicht nur kontinuierlichen Zugriff, sondern auch praktisch sofortige Leistung. Latenzzeiten können in einer Zeit, in der sich die Welt von der Epoche der Verfügbarkeit zur Epoche der Schnelligkeit bewegt, genauso schädlich sein wie die Nichtverfügbarkeit von Anwendungen.

Glücklicherweise haben wir heute Möglichkeiten, die vor einem Jahrzehnt noch nicht zur Verfügung standen. Viele Hindernisse für die Bereitstellung schneller Anwendungen, darunter das CAP-Theorem, wurden überwunden. Und jetzt haben Organisationen die Möglichkeit, eine konfliktfreie Datenschicht zu nutzen, unabhängig davon, wie viele Replikate verwendet werden.

Durch den Einsatz von Datenbanken, die vollständig im Arbeitsspeicher arbeiten und im Aktiv-Aktiv-Modus ausgeführt werden, können wir schnellere Datenbanken als bisher bereitstellen und die geringe Latenzzeit bieten, die Benutzer der Anwendungen heute verlangen.

Dies sollte für jedes Unternehmen als dringende Aufgabe betrachtet werden - Ihre Konkurrenten und Mitbewerber bieten die schnelle Anwendungen an, die Ihre Kunden fordern - den Luxus „Zeit“ haben Sie nicht.

Über den Autor - Ben Kepes

Ben Kepes ist ein Technologieanalyst, Kommentator und Berater, der sich in den letzten anderthalb Jahrzehnten als weltweit anerkannter Experte in den Bereichen Cloud Computing, Unternehmenstechnologie und digitale Transformation einen Namen gemacht hat.

Bens Kommentare wurden vielfach veröffentlicht, unter anderem in Forbes, Wired und The Guardian, und er wurde als Redner zu zahlreichen Konferenzen in den Bereichen Technologie, Wirtschaft und zu allgemeinen Themen eingeladen.

@benkepes

Über Redis

Moderne Unternehmen sind auf die Leistungsfähigkeit von Echtzeitdaten angewiesen. Mit Redis können Unternehmen sofortige Erfahrungen in einer äußerst zuverlässigen und skalierbaren Weise bereitstellen.

Redis ist die weltweit populärste In-Memory-Datenbank und kommerzieller Anbieter von Redis Enterprise, das überragende Leistung, konkurrenzlose Zuverlässigkeit und unübertroffene Flexibilität für Personalisierung, maschinelles Lernen, IoT, Suchabfragen, E-Commerce, Social Media und Metering-Lösungen weltweit bereitstellt.

Redis, das in führenden Analystenberichten über NoSQL, In-Memory-Datenbanken, operative Datenbanken und Database-as-a-Service (DBaaS) immer wieder als führend eingestuft wird, genießt das Vertrauen von mehr als 7400 Unternehmenskunden, darunter fünf Fortune-10-Unternehmen, drei der vier Kreditkartenaussteller, drei der fünf größten Kommunikationsunternehmen, drei der fünf größten Gesundheitsunternehmen, sechs der acht größten Technologieunternehmen und vier der sieben größten Einzelhändler.

Redis Enterprise ist als Service in öffentlichen und privaten Clouds, als herunterladbare Software, in Containern und für hybride Cloud-/On-Premises-Implementierungen verfügbar und unterstützt beliebte Redis-Anwendungsfälle wie Hochgeschwindigkeitstransaktionen, Job- und Warteschlangenmanagement, Benutzersitzungsspeicher, Echtzeit-Dateneingabe, Benachrichtigungen, Content-Caching und Zeitreihendaten.

Hauptsitz
700 E El Camino Real Suite 250
Mountain View, CA 94040

Tel: +1 (415) 930-9666
redis.com

Folgen Sie uns