WHITE PAPER

Lösungen für Storage-Aufgaben für **Big Data und Analytik** im Petabyte-Maßstab

SCALE TO WIN

INFINIDAT

Abstract

Big Data- und Analytik-Workloads bringen für Unternehmen neue Herausforderungen mit sich. Die erfassten Daten stammen aus Quellen, die vor zehn Jahren noch gar nicht existierten. Es werden Daten von Mobiltelefonen, maschinengenerierte Daten und Daten aus Webseiten-Interaktionen erfasst und analysiert. In Zeiten knapper IT-Budgets wird die Lage zusätzlich dadurch verschärft, dass die Big Data-Volumen immer größer werden und zu enormen Speicherproblemen führen.

Das vorliegende White Paper informiert über die Probleme, die Big Data-Anwendungen für Storage-Systeme mit sich bringen, sowie darüber, wie die Auswahl der richtigen Storage-Infrastruktur Big Data- und Analytik-Anwendungen optimieren kann, ohne das Budget zu sprengen.

Einführung: Big Data bedeutet Stress für die Storage-Infrastruktur

Big Data-Anwendungen haben zu einer explosionsartigen Zunahme der Datenmenge geführt, die Unternehmen online bereitstellen und analysieren müssen. Als Folge davon sind auch die Kosten der Datenspeicherung als Prozentsatz des IT-Gesamtbudgets explodiert. Eine neuere Befragung von Unternehmen, deren Tätigkeit auf Big Data und Analytik basieren, ergab die folgenden fünf wichtigsten Anwendungsfälle als Ursachen für diese Datenexplosion:

1. Verbesserter Service und Support für Kunden

2. Digitale Sicherheit, Erkennung von Angriffen sowie Erkennung und Verhinderung betrügerischer Aktivitäten

3. Betriebliche Analysen

4. Auswertung von Big Data

5. Verbesserte Data Warehouse-Nutzung

Ein verbesserter Kundenservice ist ein ständiges Anliegen für Unternehmen jeder Größenordnung. „Wie kann ich meine Kundenbeziehungen verbessern? Indem ich die Kundenbeziehungen pflege und verbessere, wird mein Kunde sicher weiterhin und noch öfter bei mir kaufen." Bei diesen Daten handelt es sich meist um Informationen zu Präferenzen, die aus verschiedenen Quellen, durch Online-Navigation und über Suchkriterien gewonnen werden. Diese während des gesamten Kunden-Lifecycles gesammelten Daten werden ausgewertet und zusammen mit den Kunden-Datensätzen gespeichert.

Digitale Sicherheit und Erkennung von Angriffen sind für die Anwender sehr wichtig. Diese Daten werden in Echtzeit erfasst und analysiert und sind normalerweise maschinengeneriert. Die Analyseergebnisse müssen sofort zurückgemeldet werden, um diese Aktivität relevant zu machen. Dies erfordert eine schnelle Speicherung mit viel Kapazität, da maschinengenerierte und von Sensoren gelieferte Daten viel Speicherplatz erfordern.

Für betriebliche Analysen werden Daten erfasst (oft Sensordaten von anderen Maschinen) und genutzt, um Bereiche für betriebliche Verbesserungen zu erkennen sowie um Störungen zu isolieren und Break-Fix-Analysen durchzuführen. Fertigungsunternehmen erfassen sekundengenaue Daten zu Roboter-Aktivitäten in der Produktion und wollen damit nicht nur Informationen über den aktuellen Status gewinnen, sondern auch über Möglichkeiten zur Verbesserung der Abläufe. Wie bei der Erkennung von Angriffen werden diese Daten in Echtzeit generiert und analysiert und die Ergebnisse müssen für eine sinnvolle Nutzung gespeichert und zurückgemeldet werden. Anders als bei der Erkennung von Angriffen sind sämtliche Daten wichtig, da sie Maschinen- und Prozesstrends aufzeigen, die nachfolgend nützliche Informationen liefern.

Nutzung von Big Data: Dass es sich um Big Data handelt, wird erst erkennbar, wenn diese Daten erfasst werden und wenn sich herausstellt, welche Daten noch fehlen. Hierzu ist es erforderlich, eine möglichst große Datenmenge zu sammeln.

Verbesserte Data Warehouse-Nutzung: „Wie kann ich meine vorhandenen Analytikdaten, die normalerweise in Warehouse- oder Mart-Form zur Verfügung stehen, mit zusätzlichen Daten aus externen Quellen verbessern, um die Genauigkeit zu erhöhen, die Ausführungszeiten zu verkürzen und die benötigten Antworten zu erhalten, ohne hierfür das Rad neu erfinden zu müssen?" Data Warehouses werden zunehmend auch von kleineren Unternehmen eingesetzt, da die Analyse von Transaktionsdaten für jedes Unternehmen und auf jeder Ebene erforderlich wird.

Alle diese Anwendungsfälle erfordern immer mehr Speicherplatz und Rechenleistung. Big Data werden heute als Produktivdaten angesehen, so dass die Verfügbarkeit, Wiederherstellbarkeit und Performance ebenso wichtig ist, wie bei den Transaktionssystemen im Unternehmen. Die IT-Budgets werden, wie bereits erwähnt, eher kleiner als größer. Diese diametral entgegengesetzten Kräfte führen zu einer Veränderung in der Speicherbranche. Wie kann man mit weniger Mitteln mehr erreichen, ohne dabei Kompromisse bei der Zuverlässigkeit, Effizienz oder Performance von Systemen und Anwendungen eingehen zu müssen?

Die Anforderungen von Big Data- und Analytik-Workloads an die Datenspeicherung für Unternehmensanwendungen lassen sich wie folgt zusammenfassen:

- Hervorragende Performance

- Extrem hohe Speicherdichte
- Hervorragende Verfügbarkeit und hohe Zuverlässigkeit
- Einfache Handhabung, Verwaltung und Bereitstellung
- Attraktive niedrige Cost of Ownership

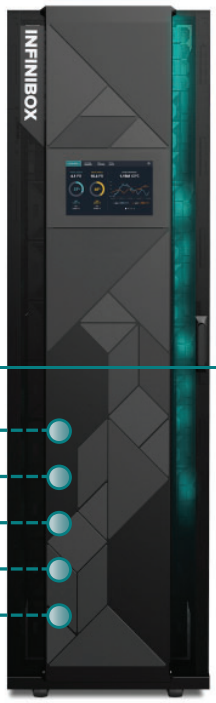
Abhilfe schafft hier eine neuartige Speicherarchitektur wie die InfiniBox® von Infinidat.

Hohe Performance

Große Datenbestände, umfangreiche Analytik-Anwendungen und zeitkritische Nachfragen nach Resultaten machen eine hohe Leistung der verwendeten Speichersysteme zu einem wichtigen Kriterium. Bei der InfiniBox wird eine maximale Performance ohne Tuning oder Optimierung erreicht. Die InfiniBox verwendet handelsübliche Standard-Hardwarekomponenten (CPU/Hauptspeicher/HDDs/SSDs) in einer ausgeklügelten Storage-Architektur, um die maximale Performance aus den eingesetzten 480 Near-Line SAS-Plattenspeichern herauszuholen. Ein zentrales Element des Kernsystem-Codes ist die Fähigkeit zur Analyse realer Anwendungsprofile und zur gezielten Festlegung von Cache-Prefetch- und Destage-Algorithmen. Das System ist speziell für reale Profile ausgelegt und bietet optimale Performance unter diesen Bedingungen. Diese Fähigkeit ist eine Kernkompetenz der InfiniBox-Architektur.

GROSSE DATENBESTÄNDE

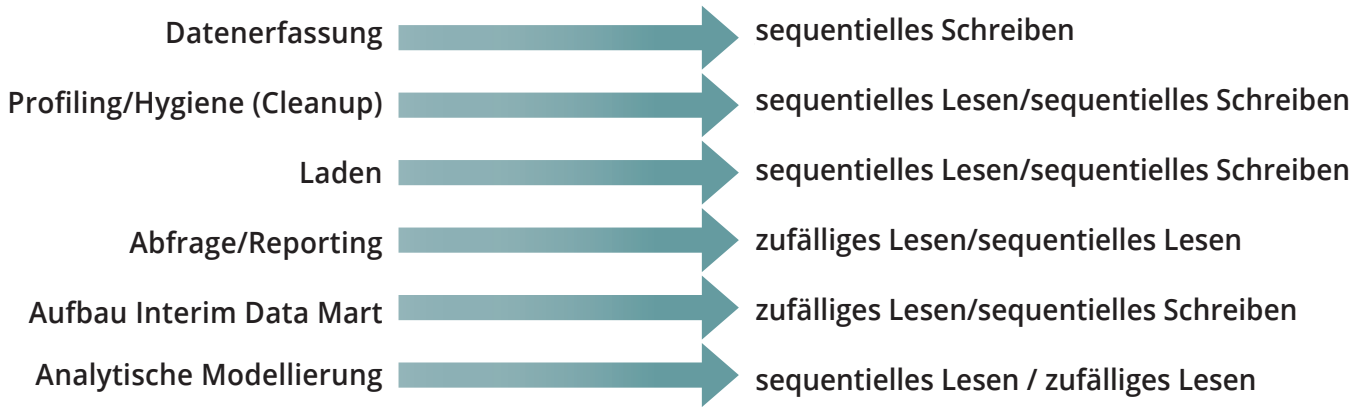
Große Datenbestände stellen hohe und einzigartige Anforderungen an Storage-Arrays in Unternehmen, da ihr E/A-Profil unvorhersehbar ist und oft die Speicherstruktur überfordert. Dies führt zu hohen Latenzen und damit zu längeren Laufzeiten der Analytikprozesse. Da jedoch manche Analytikaktivitäten sehr latenzabhängig sind, kann dies für die Endanwender der Anwendung zu Verzögerungen führen. Speicherplattformen mit begrenzter Cache-Größe können von solchen Workloads überfordert werden – jedoch nicht die InfiniBox. Die InfiniBox verwendet einen fortschrittlichen Cache-Management-Algorithmus (Neutal Cache) mit DRAM und SSDs, um die Cache-Treffer zu erhöhen und die Latenzzeiten zu verringern.



Anwendungsfälle für Big Data und Analytik	Speicheranforderungen für Big Data und Analytik
Exploration	Gleichbleibend hohe Performance
Kundenbetreuung	Hohe Kapazität
Sicherheit und Schutz vor Betrug	Höchste Verfügbarkeit und Zuverlässigkeit
Betriebliche Analysen	Einfach zu managen und bereitzustellen
DWH-Simulation	Attraktive Cost of Ownership

BEI BIG DATA VERWENDETE DIVERSE E/A-PROFILE UND MUSTER

In vielen Analytik-Umgebungen weisen die E/A-Profile alle folgenden Eigenschaften auf:



Viele dieser Eigenschaften treten erfahrungsgemäß gleichzeitig auf, andere werden durch bestimmte Aktivitäten wie Datensicherungen oder Datenladen/ETL veranlasst. InfiniBox kann eine Vielzahl von E/A-Arten gleichzeitig unterstützen. Die Datenarchitektur virtualisiert die Speicherung für jedes Volume, indem sie alle 480 Spindeln der InfiniBox-Einheit parallel belegt, wobei ein ausgeklügeltes Layout mit verteilter Datenparität zur Anwendung kommt.

Zusätzlich bietet die InfiniBox hochentwickelte Fähigkeiten zur Verbesserung von Schreibzugriffen. Mit einem speziellen patentierten multimodalen Protokoll-Schreibverfahren erzielt Infinidat eine erhebliche Verbesserung der Effizienz von aus dem Cache heraus gespeicherten Schreibzugriffen. Dies spielt für die Datenerfassungs- und ETL-Phasen eine sehr wichtige Rolle.

BLOCK-MANAGEMENT – AUF DIE GRÖSSE KOMMT ES AN

Viele Analytik-Workloads können das E/A-Profil spontan ändern. Aber im Allgemeinen verwenden weitaus die meisten Big Data- und Analytikanwendungen E/A-Operationen mit großen Blöcken beim Einlesen von Daten aus Speichern und beim Reduzieren, Sortieren, Vergleichen und Schreiben der aggregierten Daten. Große Blöcke können herkömmliche Storage-Plattformen vor Probleme stellen, da die meisten Storage-Umgebungen nicht für die Unterstützung großer Blöcke ausgelegt sind.

Hohe Dichte

Infinidat kann ein System mit einer effektive Kapazität von mehreren Petabytes in einem einzigen 19-Zoll-Rack konfigurieren. Das InfiniBox-Speichersystem ist ein modernes, vollkommen rastersymmetrisches, rein aktives Controller (Knoten)-System mit einer fortschrittlichen mehrschichtigen Caching-Architektur. Die Datenarchitektur umfasst ein Datenverteilungsmodell mit doppelter Parität (Wide Stripe). Dieses Modell verwendet eine einzigartige Kombination von zufälliger Datenverteilung und Paritätsschutz. Dies gewährleistet maximale Datenverfügbarkeit und minimalen Daten-Footprint. Jedes einzelne Volume, das in einer einzigen InfiniBox-Einheit generiert wird, speichert kleine Teile der Daten auf jedem der 480 Plattenlaufwerke der Einheit. Die mit InfiniBox mögliche nutzbare Speicherkapazität pro Einheit ist die höchste in der Speicherbranche.

Hohe Verfügbarkeit und Zuverlässigkeit

Die Analytik verfügbar zu halten, spielt eine wichtige Rolle für jedes Speichersystem. Die InfiniBox-Architektur bietet eine robuste, hochverfügbare Speicherumgebung mit „sieben Neunen“-Verfügbarkeit. Dies entspricht einer Ausfallzeit von weniger als drei Sekunden pro Jahr! Unsere Kunden berichten von keinen Datenverlusten auch bei

Ausfall mehrerer Plattenspeicher. InfiniBox bietet Features für durchgängige Business-Kontinuität, einschließlich asynchronem Remote-Mirroring und Snapshots. Mit Hilfe der Snapshots erfordert die Wiederherstellung einer Datenbank nur den Zeitaufwand für die Zuordnung der Volumes zu Hosts – Minuten statt Stunden, wie bei einer herkömlichen Datensicherungs- und -wiederherstellung.

Bedienerfreundliches automatisiertes Provisioning und Management

Die InfiniBox-Architektur sowie die elegante und unkomplizierte Web-basierte GUI und die integrierte Befehlszeilen-Schnittstelle ermöglichen ein schnelles Aufsetzen und Verwalten des Speichersystems. Gegenüber herkömmlichen Konzepten werden bei der Speicherverwaltung enorme Zeiteinsparungen erzielt. Durch die offene Architektur der InfiniBox und ihrer umfassenden Unterstützung für RESTful API ermöglichen Plattformen wie OpenStack und Docker die Ausführung der Speicheradministration auf der Anwendungsebene, ohne dass diese Aktivitäten über die InfiniBox GUI veranlasst werden müssen.

Die InfiniBox bietet ein Managementsystem, das Storage-Pools und Volumes nach Benutzern getrennt behandeln kann. Durch die Unterstützung von Multi-Tenancy-Features können Nutzer von Anwendungen, wie sie beispielsweise in einer privaten Cloud-Umgebung eingesetzt werden, den der betreffenden Nutzergruppe zugeordneten Speicher sehen und managen.

Sehr niedrige Total Cost of Ownership

Hohe Performance, extreme Verfügbarkeit, höchste Datendichte und Bedienerfreundlichkeit tragen alle zu beispiellos niedrigen TCO bei. Dies ist wichtig für Umgebungen, in denen aufgabenkritische Datenbanken zu immer kompakteren physikalischen Einheiten konsolidiert werden müssen.

Splunk und InfiniBox

Einer der zahlreichen Big Data-Anwendungsfälle ist die Unterstützung von Splunk-Clustern. Das aktuelle Standard-Einsatzmodell für Splunk-Cluster ist der Aufbau eines Clusters aus zahlreichen kostengünstigen Knoten. Jeder Knoten besitzt gleich viel Rechenleistung, Hauptspeicher und dedizierte Speicherkapazität. Die meisten Anwender, die erstmals eine Splunk-Umgebung aufbauen, müssen in der ersten Planungsphase die richtige Kombination von Ressourcen für dieses Design auf Knotenebene bestimmen. Anschließend verbinden sie mehrere dieser Knoten miteinander zu einer großen IT-Umgebung für ihre bevorzugte Analytikanwendung. Hier tritt dann meist das Problem auf, dass der Speicherplatz schon lange vor der Rechenkapazität im Cluster knapp wird. Anwendern mit dedizierten Speichern bleibt dann nur die Möglichkeit, den Cluster um weitere Knoten zu vergrößern. Dieses Modell mit dediziertem Speicher hat nur den Haken, dass dann auch wieder mehr Rechenleistung hinzugefügt wird (die möglicherweise nicht benötigt wird).

Durch Einführung des Block-basierten SAN-Storage-Systems mit InfiniBox anstelle von dedizierten Festplatten pro Knoten besteht keine Beschränkung mehr auf die einem Knoten zugeordneten Speicherkapazität. Es können dynamisch mehr LUNs hinzugefügt werden, oder es kann mehr Speicherplatz pro LUN für jeden Cluster-Knoten zugeteilt werden. Es muss erst Rechenleistung hinzugefügt werden, wenn diese benötigt wird. Der Grad der Splunk-Datenredundanz kann reduziert werden, da die Daten mit „sieben Neunen"-Verfügbarkeit vollständig geschützt werden..

Schlussbemerkung

Die InfiniBox schließt die Lücke zwischen hoher Performance und hoher Kapazität für Big Data-Anwendungen. Sie ermöglicht einem Unternehmen, das Big Data- und Analytik-Projekte realisiert, seine Business-Ziele wrklich zu erreichen: Kostensenkung, kontinuierliche und tiefgehende Skalierung der Kapazität und unkompliziertes und effektives Management – ohne jeden Kompromiss bei Performance oder Zuverlässigkeit. All dies leistet eine effektive und effiziente Unterstützung für Big Data-Anwendungen zu einem attraktiven Preis.