Por qué el almacenamiento es el héroe desconocido de la AI

En la carrera hacia la inteligencia artificial general (AGI), la tecnología de almacenamiento está marcando el ritmo. Mientras que los algoritmos y la computación son el centro de atención, el almacenamiento potencia los avances en AI. Durante la revolución de flash, 15K00 discos se estancaron a medida que el rendimiento computacional se duplicaba cada dos años, pero la virtualización habilitada para flash y, hoy en día, las cargas de trabajo impulsadas por GPU impulsan una mayor innovación de almacenamiento junto con las demandas de eficiencia, sustentabilidad y confiabilidad.

Los primeros esfuerzos de AI se vieron limitados por la complejidad algorítmica y la escasez de datos, pero a medida que los algoritmos avanzaban, surgieron cuellos de botella en la memoria y el almacenamiento. El almacenamiento de alto

rendimiento desbloqueó innovaciones como ImageNet, que impulsó los modelos de visión, y GPT-3, que requería petabytes de almacenamiento. Con 400 millones de terabytes de datos generados diariamente, el almacenamiento debe administrar cargas de trabajo a escala de exabytes con latencia de submilisegundos para potenciar el aprendizaje automático cuántico y AGI. A medida que la AI progresaba, cada ola de innovación impuso nuevas demandas de almacenamiento, impulsando avances en capacidad, velocidad y escalabilidad para adaptarse a modelos cada vez más complejos y conjuntos de datos más grandes.

- **Aprendizaje automático clásico (1980s-2015):** El reconocimiento de voz y los modelos de aprendizaje supervisado impulsaron el crecimiento del conjunto de datos de megabytes a gigabytes, lo que hizo que la recuperación de datos y la organización fueran cada vez más críticas.
- **Revolución del aprendizaje profundo (2012-2017):** Modelos como AlexNet y ResNet superaron las demandas de almacenamiento, mientras que Word2Vec y GloVe avanzaron en el procesamiento de lenguaje natural, cambiando al almacenamiento NVMe de alta velocidad para conjuntos de datos a escala de terabytes.
- **Modelos básicos (2018 a la actualidad):** BERT presentó conjuntos de datos a escala de petabytes, con GPT-3 y Llama 3 que requieren sistemas Meta escalables y de baja latencia como Meta’s Tectonic para manejar billones de tokens y mantener un rendimiento de 7TB/s.
- **Leyes de escalamiento de Chinchilla (2022):** Chinchilla enfatizó el crecimiento de conjuntos de datos sobre el tamaño del modelo LLM, lo que requirió almacenamiento de acceso paralelo para optimizar el rendimiento.

El almacenamiento no solo respalda la AI, sino que lidera el camino y da forma al futuro de la innovación mediante la administración eficiente y a escala de los datos en constante crecimiento del mundo. Por ejemplo, las aplicaciones de AI en la conducción autónoma dependen de plataformas de almacenamiento capaces de procesar petabytes de datos de sensores en tiempo real, mientras que la

investigación genómica requiere un acceso rápido a conjuntos de datos masivos para acelerar los descubrimientos. A medida que la AI continúa superando los límites de la administración de datos, los sistemas de almacenamiento tradicionales enfrentan desafíos crecientes para mantenerse al día con estas demandas cambiantes, lo que destaca la necesidad de soluciones diseñadas específicamente.

Cómo las cargas de trabajo de AI tensan los sistemas de almacenamiento tradicionales

Consolidación de datos y administración del volumen

Las aplicaciones de AI administran conjuntos de datos que van desde terabytes hasta cientos de petabytes, lo que supera ampliamente las capacidades de los sistemas de almacenamiento tradicionales como NAS, SAN y el almacenamiento heredado con conexión directa. Estos sistemas, diseñados para cargas de trabajo precisas y transaccionales, como generar informes o recuperar registros específicos, tienen dificultades con las demandas de agregación pesadas de la ciencia de datos y los patrones de acceso amplios y de alta velocidad de las cargas de trabajo de AI/ML. La capacitación modelo, que requiere recuperación masiva de datos por lotes en todo el conjunto de datos, destaca esta desalineación. Las arquitecturas rígidas, las limitaciones de capacidad y el rendimiento insuficiente de la infraestructura tradicional la hacen inadecuada para la escalabilidad y la velocidad de la AI, lo que destaca la necesidad de plataformas de almacenamiento diseñadas específicamente.

Cuellos de botella de rendimiento para el acceso a datos de alta velocidad

El análisis en tiempo real y la toma de decisiones son esenciales para las cargas de trabajo de AI, pero las arquitecturas de almacenamiento tradicionales suelen crear cuellos de botella con IOPS insuficientes, ya que se diseñaron para tareas transaccionales moderadas en lugar de las demandas de lectura/escritura

paralela intensivas de AI. Además, la alta latencia de los discos giratorios o los mecanismos de almacenamiento en caché obsoletos retrasan el acceso a los datos, lo que aumenta el tiempo de comprensión y reduce la eficiencia de los procesos de AI.

Manejo de diversos tipos de datos y cargas de trabajo

Los sistemas de AI manejan datos estructurados y no estructurados, incluidos texto, imágenes, audio y video, pero las soluciones de almacenamiento tradicionales luchan con esta diversidad. A menudo se optimizan para datos estructurados, lo que resulta en una recuperación lenta y un procesamiento ineficiente de formatos no estructurados. Además, la indexación deficiente y la administración de metadatos dificultan la organización y búsqueda de diversos conjuntos de datos de manera eficaz. Los sistemas tradicionales también enfrentan problemas de rendimiento con archivos pequeños, comunes en los modelos de lenguaje de capacitación, ya que la sobrecarga de metadatos altos provoca retrasos y tiempos de procesamiento más largos.

Limitaciones de la arquitectura heredada

El efecto acumulativo de estos desafíos es que las arquitecturas de almacenamiento tradicionales no pueden seguir el ritmo de las demandas de las cargas de trabajo de AI modernas. Carecen de la agilidad, el rendimiento y la escalabilidad necesarios para respaldar los diversos requisitos de datos de alto volumen de la AI. Estas limitaciones destacan la necesidad de soluciones de almacenamiento avanzadas que están diseñadas para manejar los desafíos únicos de las aplicaciones de AI, como la escalabilidad rápida, el alto rendimiento, la baja latencia y la gestión de datos diversos.

Desafíos clave del almacenamiento en AI

Las cargas de trabajo de AI imponen demandas únicas a los sistemas de almacenamiento, y abordar estos desafíos requiere capacidades avanzadas en las siguientes áreas:

- **Consolidación unificada de datos:** Los silos de datos fragmentan información valiosa, lo que requiere consolidación en una plataforma unificada que admite diversas cargas de trabajo de AI para un procesamiento y capacitación sin problemas.
- **Capacidad y rendimiento escalables:** Una plataforma de almacenamiento robusta debe administrar diversos perfiles I/O y escalar de terabytes a exabytes, lo que garantiza un acceso de baja latencia y alta tasa de transferencia. Al permitir la escalabilidad sin interrupciones, la plataforma permite que las cargas de trabajo de AI se expandan sin problemas a medida que crecen las demandas de datos, manteniendo operaciones fluidas e ininterrumpidas.
- **Flexibilidad de escalabilidad horizontal y horizontal:** El manejo del acceso transaccional de latencia baja para bases de datos de vectores y cargas de trabajo de alta simultaneidad para capacitación e inferencia requiere una plataforma que ofrezca ambas capacidades.
- **Confiabilidad y tiempo de actividad continuo:** A medida que la AI se vuelve fundamental para las empresas, el tiempo de actividad del 99,9999 % es esencial. Una plataforma de almacenamiento debe admitir actualizaciones y actualizaciones de hardware sin interrupciones, lo que garantiza operaciones continuas sin tiempo de inactividad visible para los usuarios finales.

Optimización del almacenamiento en todo el proceso de AI

Las soluciones de almacenamiento efectivas son esenciales en cada etapa del proceso de AI, desde el curado de datos hasta el entrenamiento y la inferencia, ya que permiten que las cargas de trabajo de AI funcionen de manera eficiente y a escala. Los procesos de AI requieren un almacenamiento que pueda manejar sin problemas las tareas sensibles a la latencia, escalar para satisfacer las demandas de alta simultaneidad, admitir diversos tipos de datos y mantener el rendimiento

en entornos distribuidos.

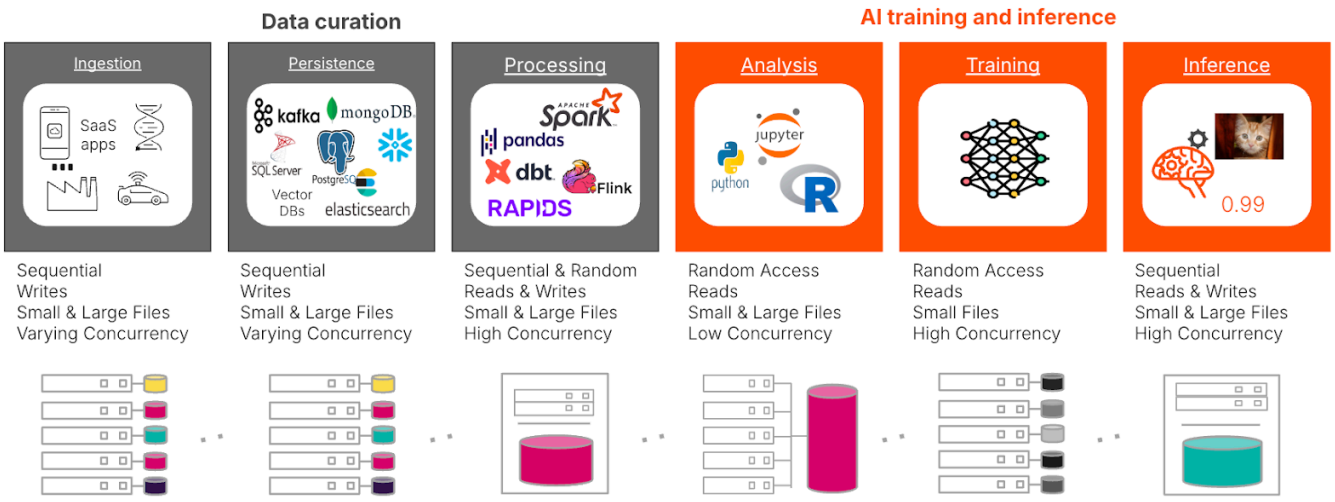

Figura 1: Los patrones de almacenamiento para la AI son variados y requieren una plataforma diseñada para el rendimiento multidimensional.

En la etapa de selección de datos, la administración de conjuntos de datos a escala de petabyte a exabyte comienza con la ingesta, donde el almacenamiento debe escalar sin problemas para manejar volúmenes de datos masivos y, al mismo tiempo, garantizar un alto rendimiento. Las aplicaciones en tiempo real, como la conducción autónoma, requieren un almacenamiento de latencia baja capaz de procesar los datos entrantes al instante. Los módulos DirectFlash® (DFM) se destacan en estos escenarios al eludir las arquitecturas tradicionales de SSD para acceder directamente a flash NAND, lo que ofrece un rendimiento más rápido y uniforme con una latencia significativamente reducida. En comparación con los SSD y SCM heredados, los DFM también ofrecen una mayor eficiencia energética, lo que permite a las organizaciones cumplir con las demandas de las cargas de trabajo de AI a gran escala mientras optimizan el consumo de energía y mantienen un rendimiento predecible bajo una alta simultaneidad.

Durante la persistencia, las soluciones de almacenamiento de datos deben admitir la retención a largo plazo y la accesibilidad rápida para los datos de acceso frecuente. El paso de procesamiento es clave para preparar datos para la capacitación, donde el almacenamiento debe administrar una variedad de tipos y

tamaños de datos de manera eficiente, manejando datos estructurados y no estructurados en formatos como NFS, SMB y objetos.

En la fase de entrenamiento e inferencia de AI, el entrenamiento de modelos genera demandas de lectura/escritura intensivas, lo que requiere arquitecturas de escalabilidad horizontal para garantizar el rendimiento en varios nodos. Los sistemas de control de versiones y control de puntos de control eficientes son fundamentales en esta etapa para evitar la pérdida de datos. Además del control, las arquitecturas emergentes como la generación aumentada por recuperación (RAG) presentan desafíos únicos para los sistemas de almacenamiento. RAG se basa en la recuperación eficiente de bases de conocimiento externas durante la inferencia, lo que exige un almacenamiento de baja latencia y alta tasa de transferencia capaz de manejar consultas simultáneas y paralelas. Esto ejerce presión adicional sobre la administración de metadatos y la indexación escalable, lo que requiere arquitecturas de almacenamiento avanzadas para optimizar el rendimiento sin cuellos de botella.

Al alinear las soluciones de almacenamiento con las necesidades específicas de cada etapa del proceso, las organizaciones pueden optimizar el rendimiento de la AI y mantener la flexibilidad necesaria para respaldar las demandas de AI en evolución.

Conclusiones

El almacenamiento es la columna vertebral de la AI, ya que la creciente complejidad del modelo y la intensidad de los datos impulsan demandas exponenciales en la infraestructura. Las arquitecturas de almacenamiento tradicionales no pueden satisfacer estas necesidades, lo que hace que la adopción de soluciones de almacenamiento ágiles y de alto rendimiento sea esencial.

La relación simbiótica entre la AI y las plataformas de almacenamiento significa avances en el almacenamiento, no solo soporte, sino que también acelera el progreso de la AI. Para las empresas que recién comienzan a explorar la AI, la flexibilidad es crucial: Necesitan un almacenamiento que pueda escalar a medida

que crecen sus necesidades de datos y procesamiento, admitir varios formatos (p. ej., archivo, objeto) e integrarse fácilmente con las herramientas existentes.

Las organizaciones que invierten en plataformas de almacenamiento modernas se posicionan a la vanguardia de la innovación. Esto requiere:

- **Evaluación de la infraestructura:** Identifique las limitaciones actuales y las áreas de mejora inmediata.
- **Adoptar soluciones escalables:** Implemente plataformas que ofrezcan flexibilidad, alto rendimiento y crecimiento sin interrupciones.
- **Planificación para necesidades futuras:** Manténgase a la vanguardia de las tendencias emergentes para garantizar que la plataforma evolucione con los desarrollos de AI.

Al priorizar las plataformas de almacenamiento como un componente central de la estrategia de AI, las organizaciones pueden desbloquear nuevas oportunidades, impulsar la innovación continua y mantener una ventaja competitiva en el futuro basado en datos.

¿Necesita más información?

Visite la página de soluciones de AI

Mire la repetición del seminario web: "Consideraciones para una infraestructura de AI empresarial estratégica acelerada"

Descargue el informe técnico: "La plataforma de Pure Storage para AI"