PURESTORAGE®

Tech Webinar

Pure Orange Shot

2024 Pure Storage Webinar Series

#2 두번째 시리즈

데이터 파이프라인을 통한 AI 가속화 전략 with NVIDIA
(RAG를 활용한 LLM 향상과 Pure Storage 및 NVIDIA를 통한 기업 AI 가속화)

2024년 6월 26일(수) 오후 2시

PURESTORAGE®

데이터 파이프라인을 통한 AI 가속화 전략 with NVIDIA

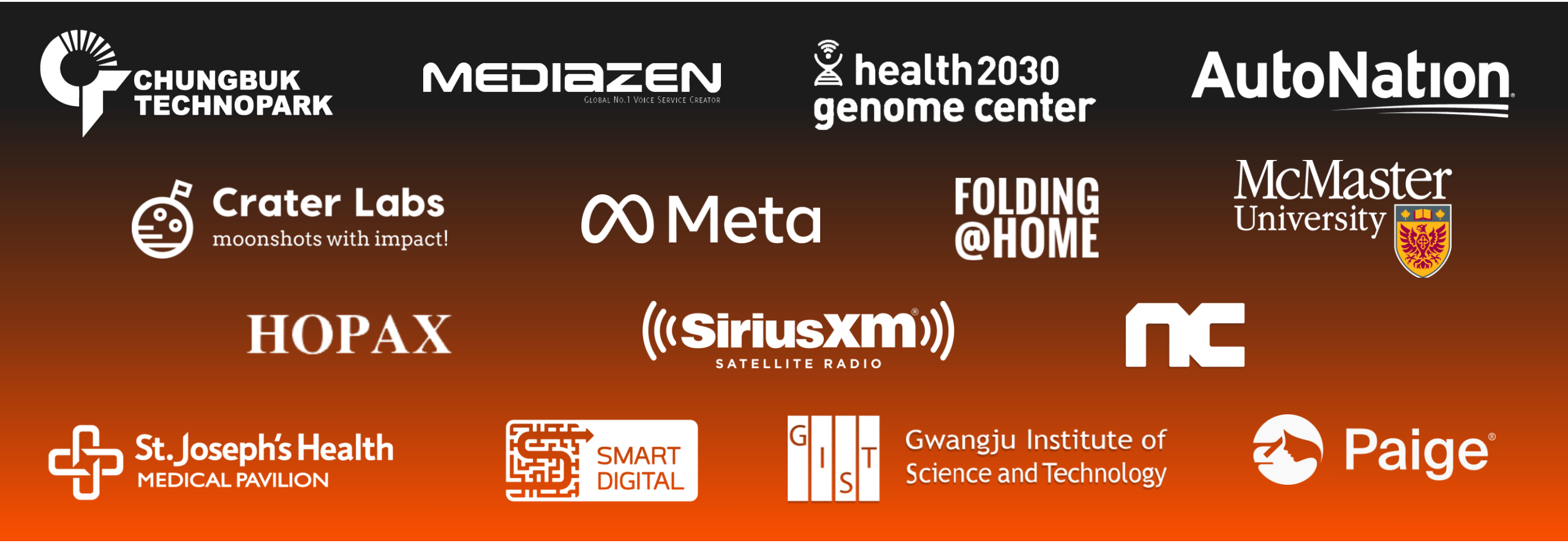
(RAG를 활용한 LLM 향상과 Pure Storage 및 NVIDIA를 통한 기업 AI 가속화)

김기배 상무
Sr. Channel Manager
Pure Storage

윤건호 부장
Sr. Systems Engineer
Pure Storage

차경환 상무
NVIDIA Partner Engineer
BNINC

Powering 100s of leading AI projects



©2023 Pure Storage

BNINC 회사 소개

㈜ 비엔아이엔씨

비엔아이엔씨는 NVIDIA Elite 파트너로서, 최고의 기술력과 경험을 바탕으로 신뢰할 수 있는 IT 전문 기업입니다.

주요 사업 분야

AI & HPC System

- AI & HPC를 위한 GPU 시스템 & 플랫폼 구축
- 초 고성능 서버, 스토리지, 네트워크 클러스터 컨설팅 및 구축

Solution & Consulting

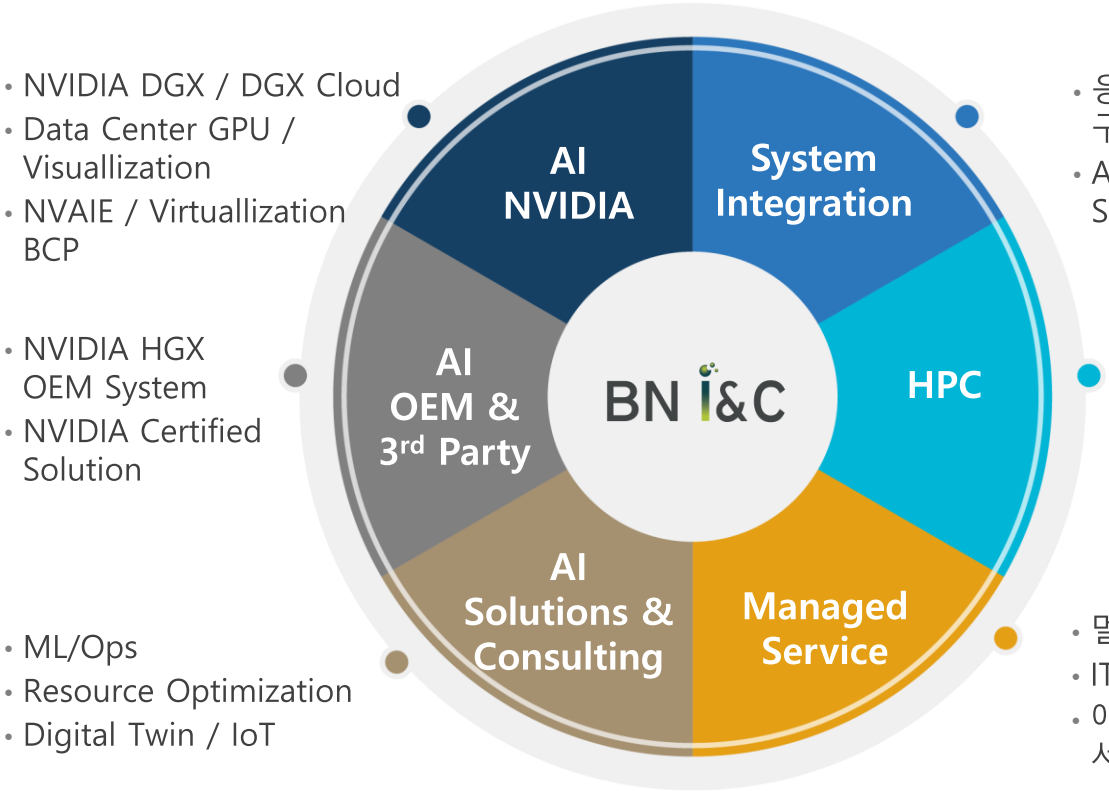
- NVIDIA AI/딥 러닝 S/W 플랫폼 컨설팅 및 구축
- 3rd AI 솔루션 컨설팅 및 구축
- IoT, DT 플랫폼 개발, 컨설팅

System Integration

- 통합 시스템 구축 및 솔루션 개발, 컨설팅
- MSA 컨설팅 및 구축

Managed Service

- 고객 IT 인프라의 안정적인 유지 관리



- NVIDIA DGX / DGX Cloud
- Data Center GPU / Visuallization
- NVAIE / Virtuallization BCP

- NVIDIA HGX OEM System
- NVIDIA Certified Solution

- ML/Ops
- Resource Optimization
- Digital Twin / IoT

- 응용 프로그램 및 인프라 구축 컨설팅
- Application Managed Service

- 금융 HPC Application 컨설팅 / 지원

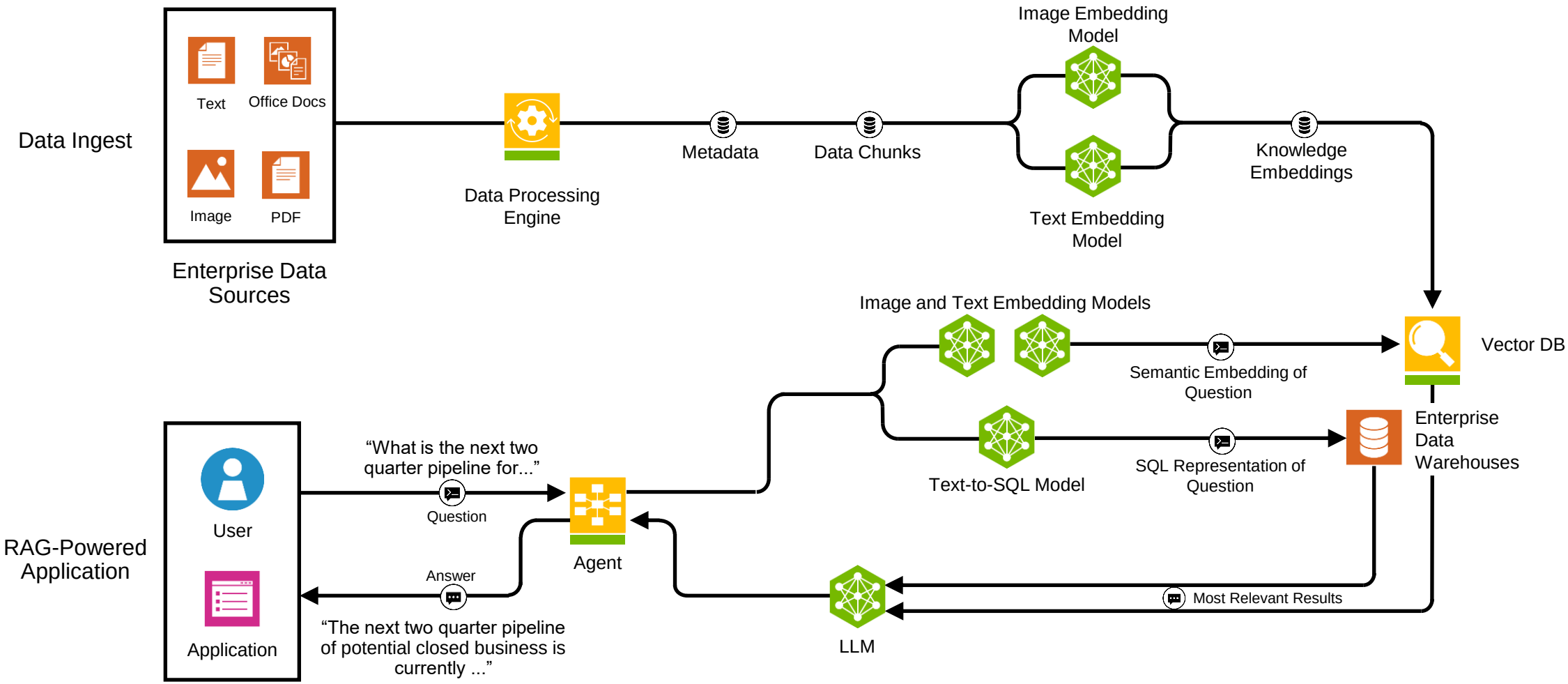
- 멀티 벤더 서비스
- IT 인프라 운영 관리 서비스
- 애플리케이션 운영 관리 서비스

BN I&C

Enterprise RAG with Networked All-Flash Storage

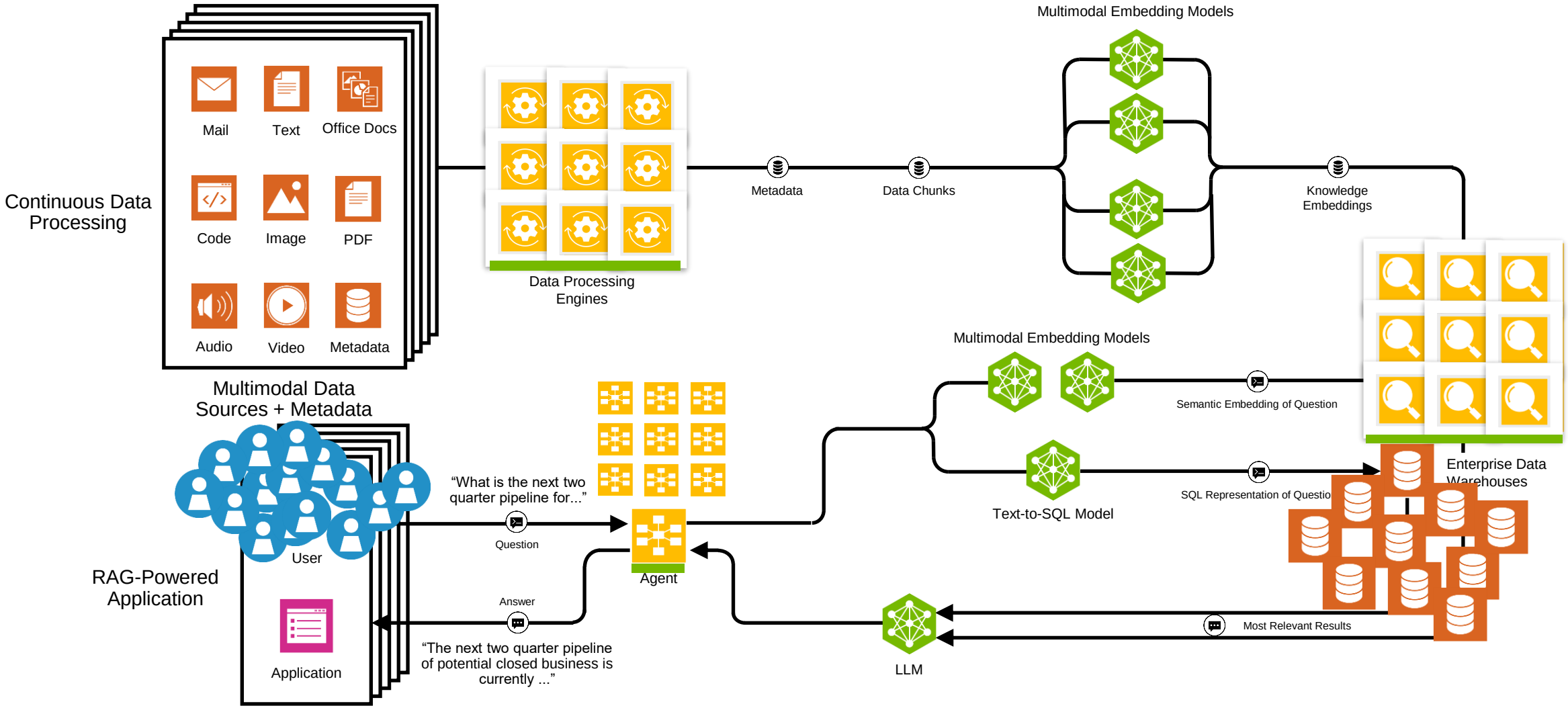
Retrieval Augmented Generation

Enable LLMs to provide up to date, proprietary, and domain specific answers



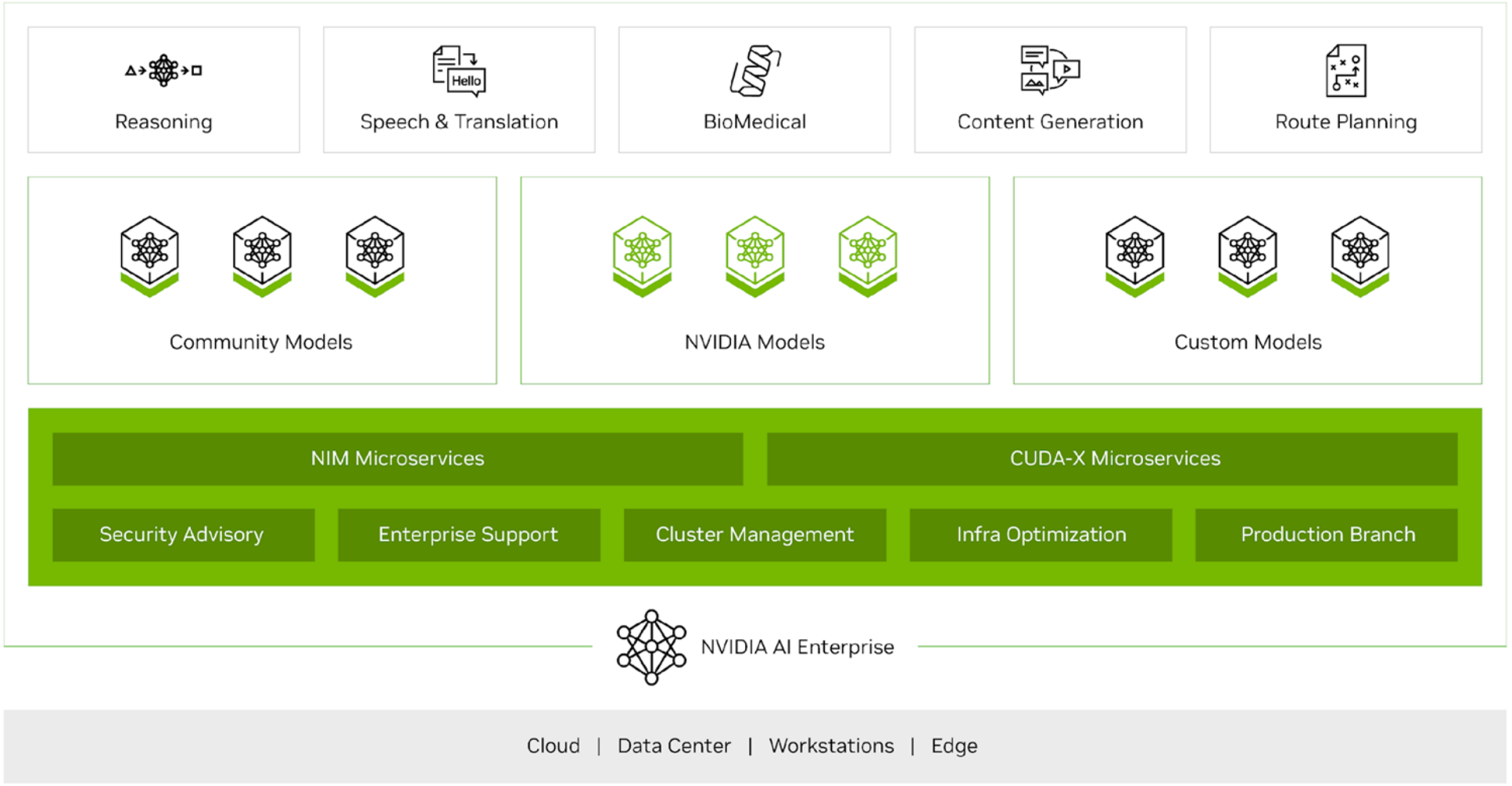
Enterprise Scale Retrieval Augmented Generation

Multimodal data sources with millions of users



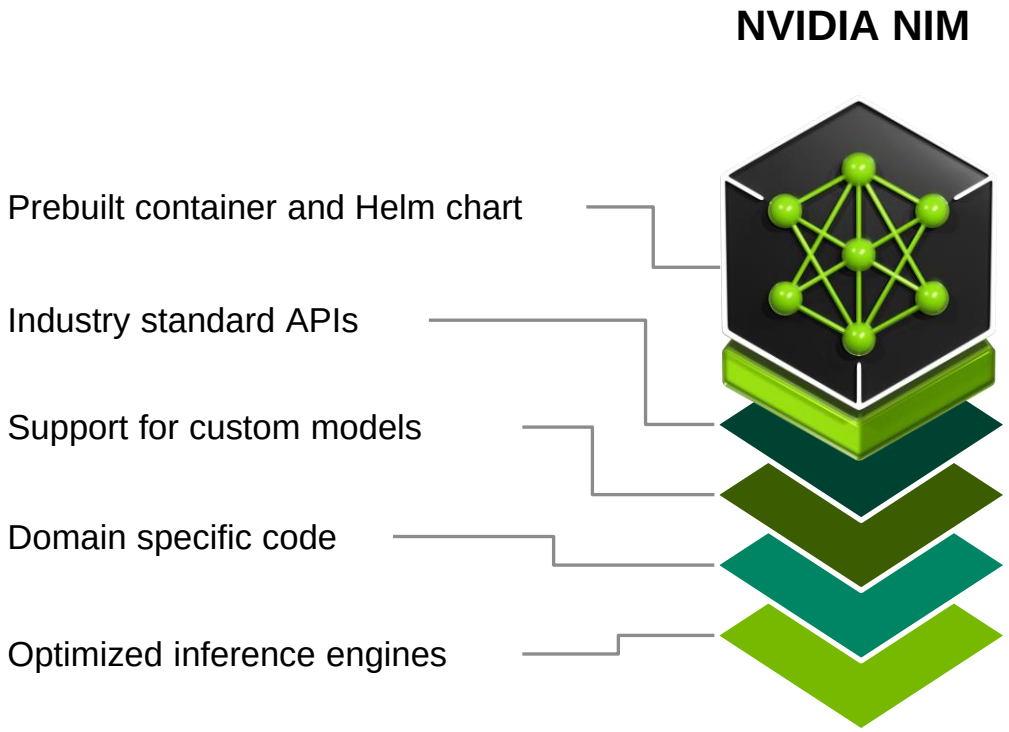
NVIDIA AI Enterprise

High Performance and Efficient Runtime for Generative AI



NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI



Deploy anywhere and maintain control of generative AI applications and data

Simplified development of AI application that can run in enterprise environments

Day 0 support for all generative AI models providing choice across the ecosystem

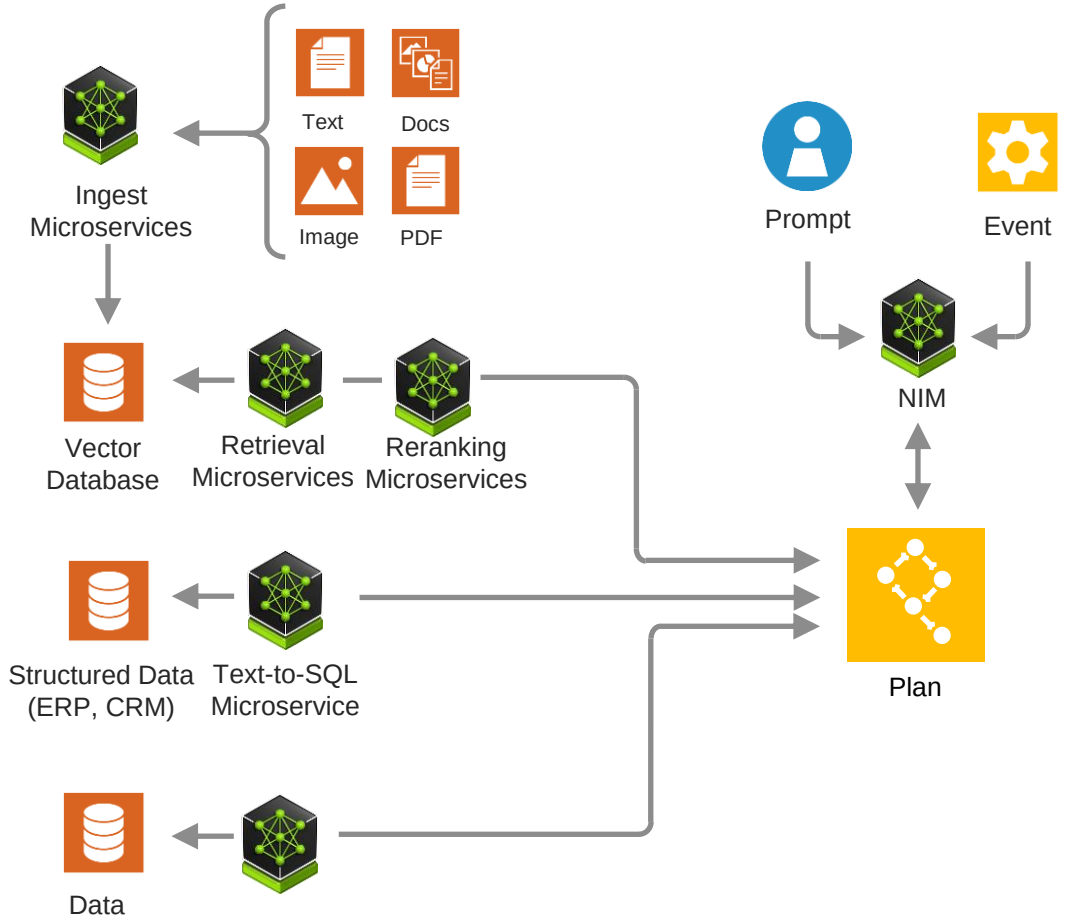
Improved TCO with best latency and throughput running on accelerated infrastructure

Best accuracy for enterprise by enabling tuning with proprietary data sources

Enterprise software with feature branches, validation and support

NeMo Retriever Supercharges RAG Applications

World Class Accuracy and Throughput



2X World-class accuracy with nearly 2x fewer incorrect answers

7X Faster embedding inference throughput

Optimized Inference Engines

World class models and community model support

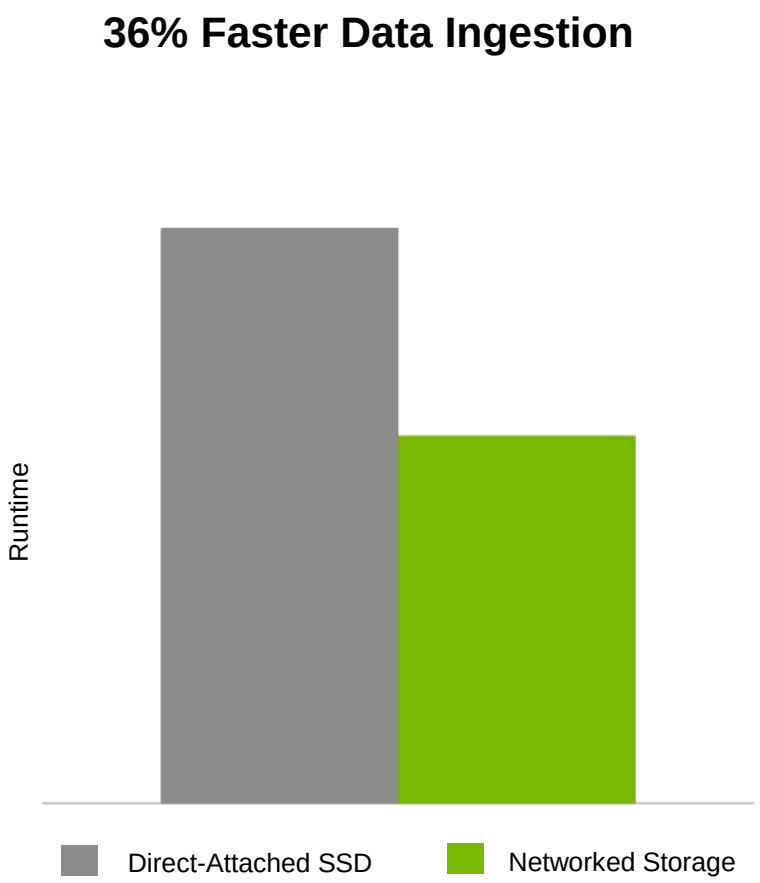
Flexible and modular deployment

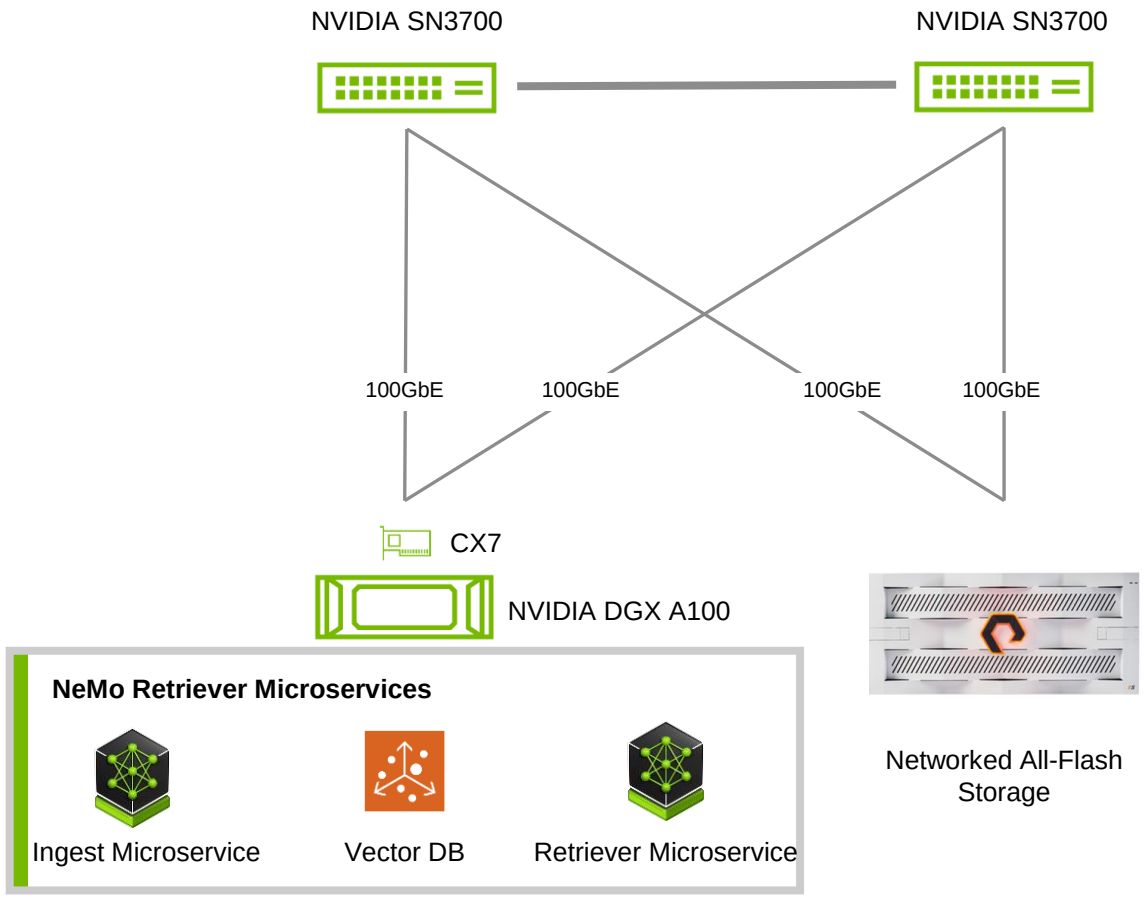
Customizable models and pipelines

Production Ready

Data Ingestion with Networked Storage

Improved Performance

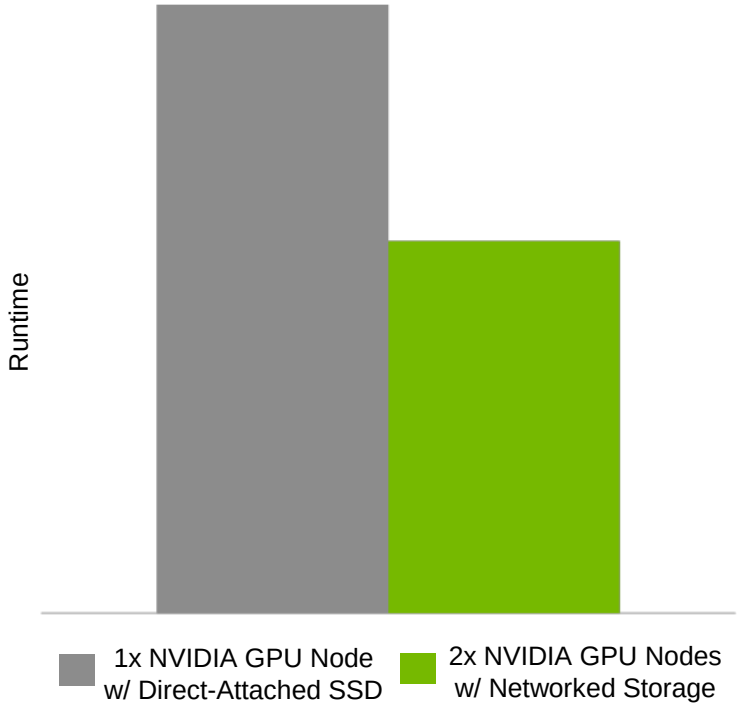


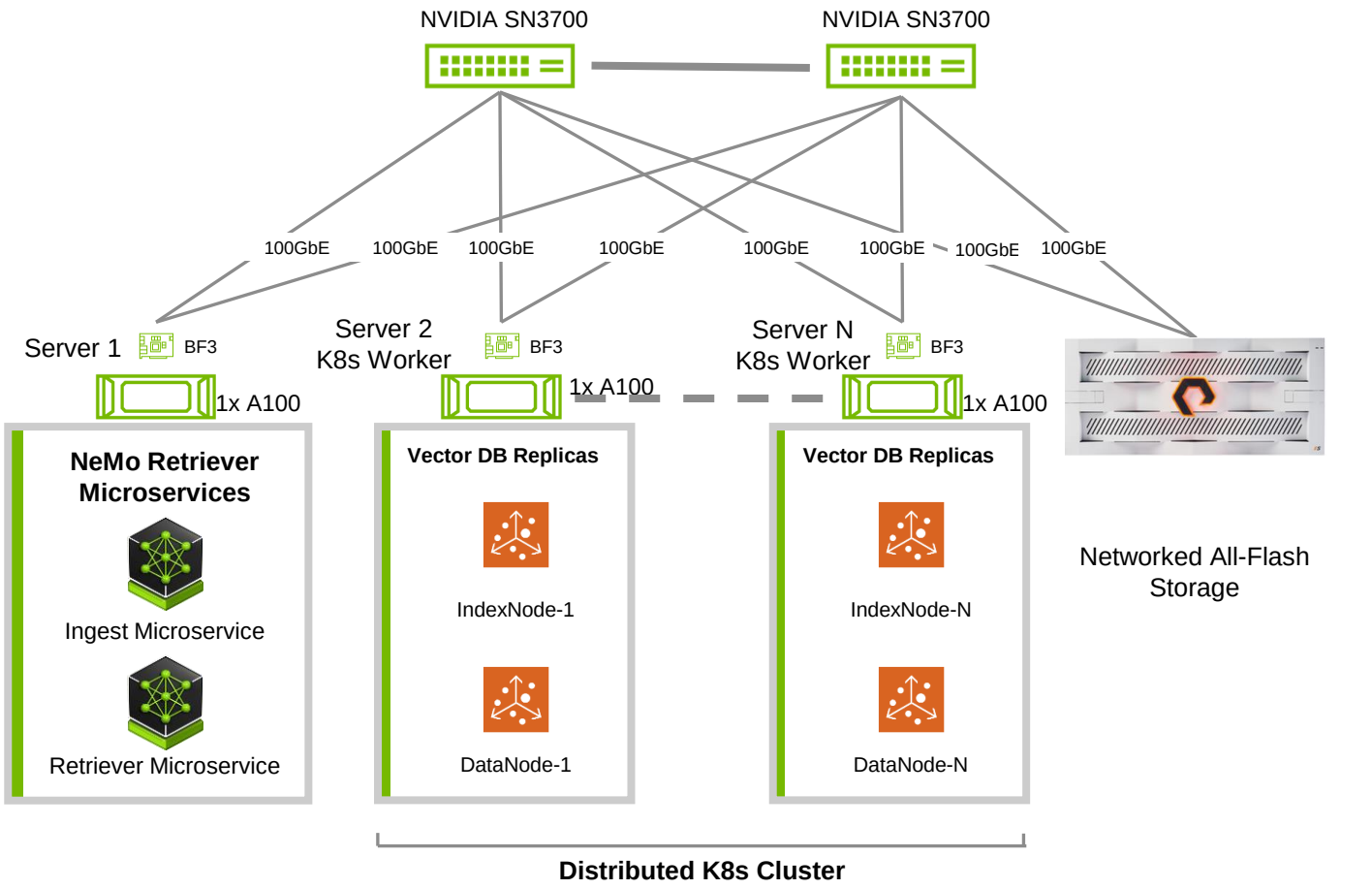


Multi-node Data Ingestion Scale Out

Optimized Networking & Storage to Scale Out Embedding and Indexing Performance

Scale Out Data Ingestion

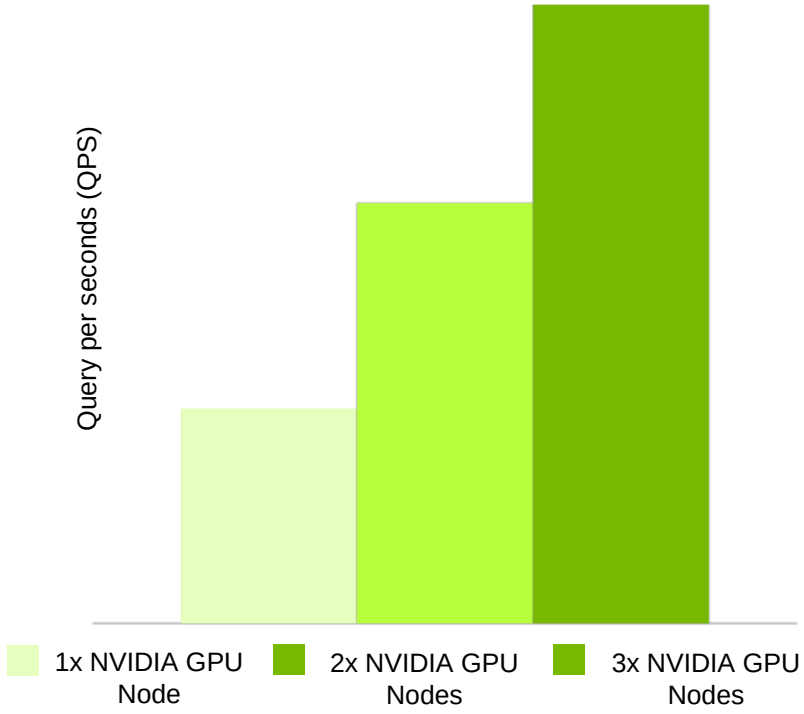


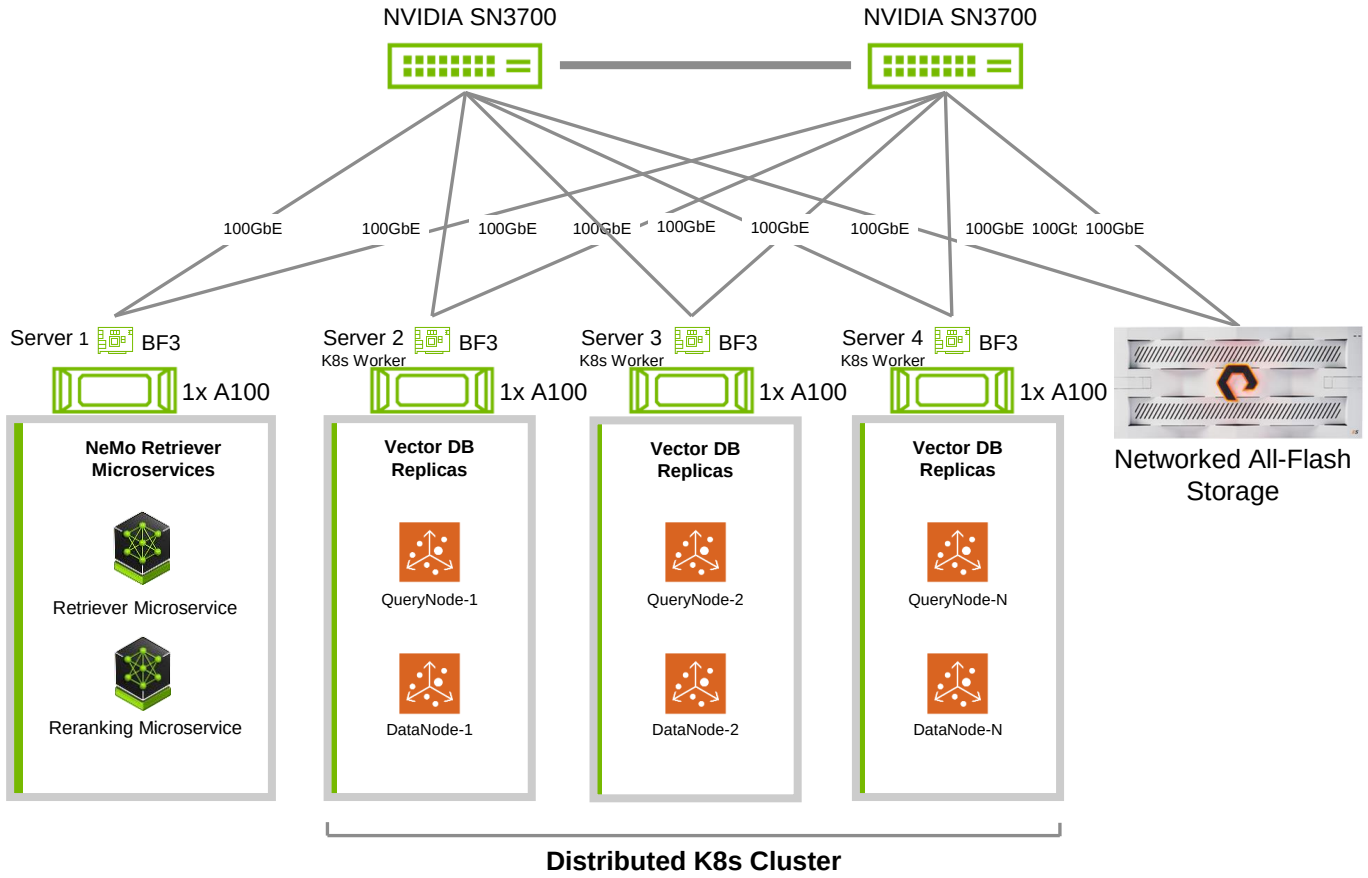


Multi-node Query Scale Out

Optimized Networking & Storage to Scale Out Query Performance

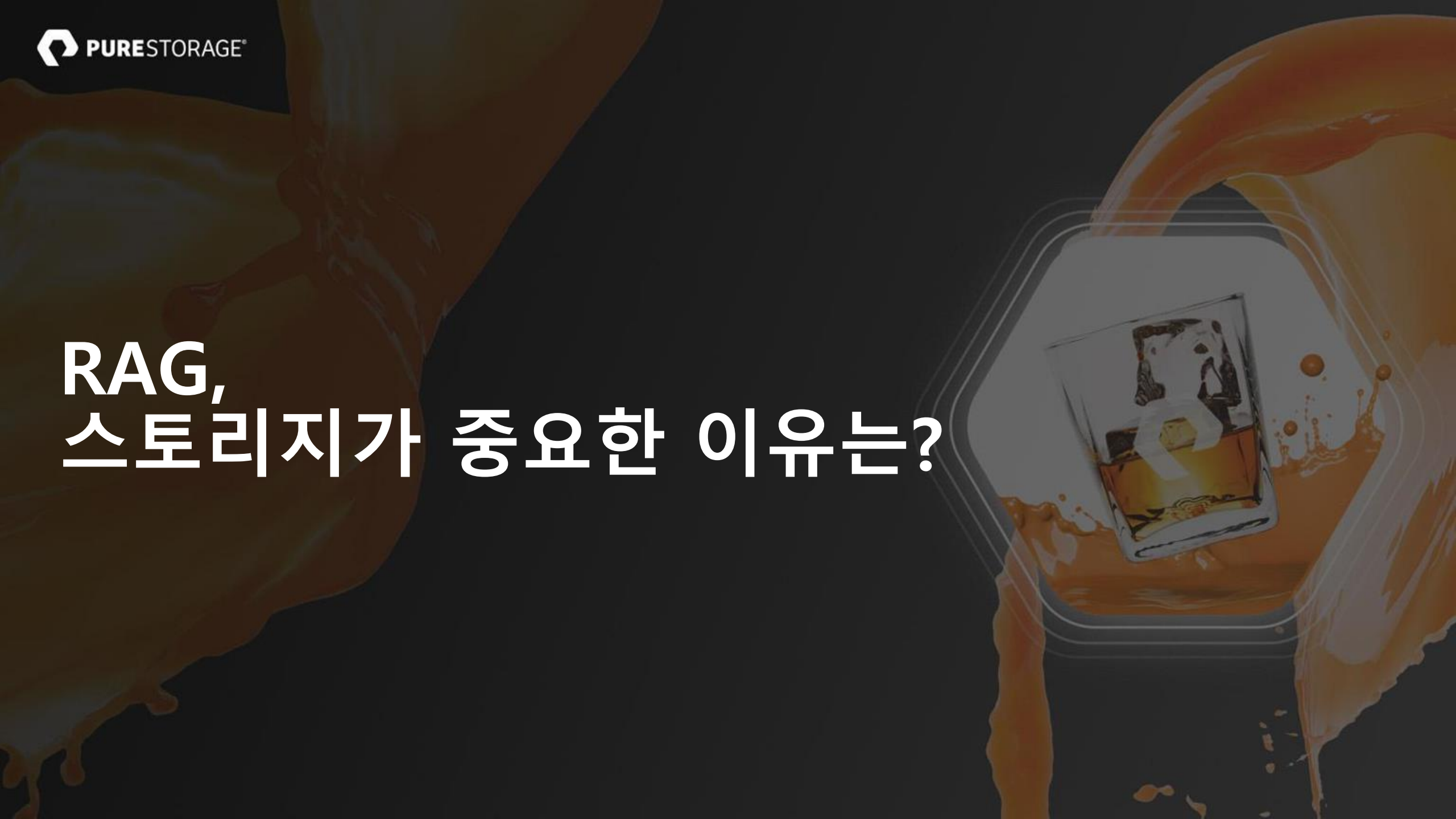
Scale Out Data Query





PURESTORAGE®

RAG, 스토리지가 중요한 이유는?



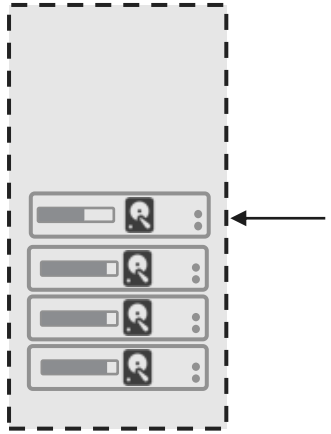
스토리지와 컴퓨팅 자원을 전문화

©2024 Pure Storage

스토리지 및 컴퓨팅의 독립적인 확장

레거시 DAS 접근방식

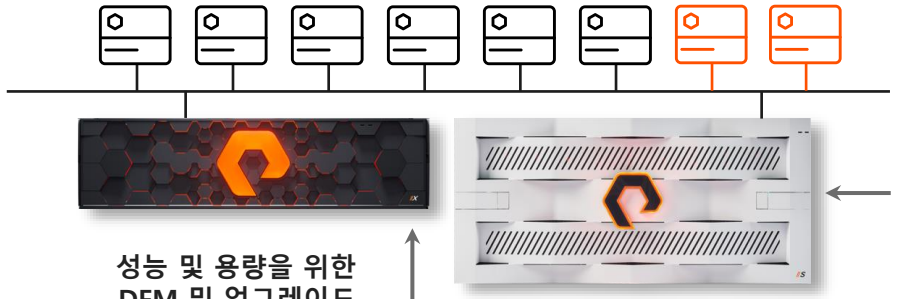
데이터 재분배
클러스터의 데이터를 재분배 합니다.



서버 추가
추가 스토리지 또는 컴퓨팅에는 동일한 구성을 가진 새 서버 노드가 필요합니다.

전문적인 분산 접근방식

서버
컴퓨팅 요구 사항을 충족하기 위해 동적으로 서버를 추가하거나 제거합니다
.



성능 및 용량을 위한 DFM 및 업그레이드 컨트롤러 추가
데이터 이동이나 재분배 없음, 무중단

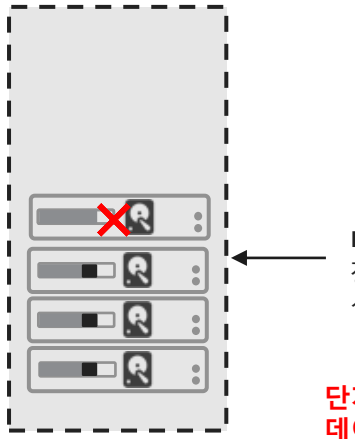
성능 및 용량을 위한 블레이드 및 DFM 추가
데이터 이동이나 재분배 없음, 무중단

독립적으로 필요에 따라 스토리지를 확장하고 컴퓨팅 자원을 추가합니다.

©2024 Pure Storage

향상된 노드 장애 복구

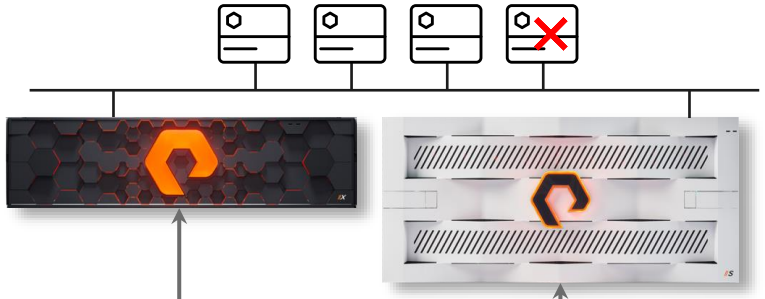
레거시 DAS 접근방식



데이터 재구성
장애발생한 노드의 데이터가 재구성되었습니다.

단지 노드가 장애났을 뿐인데 데이터 처리성능이 매우 떨어지게 됩니다.

전문적인 분산 접근방식



더 빠른 복구
더 빠른 재분배 또는 재구성을 통한 영구 데이터

노드장애에도 성능 저하가 없어요.

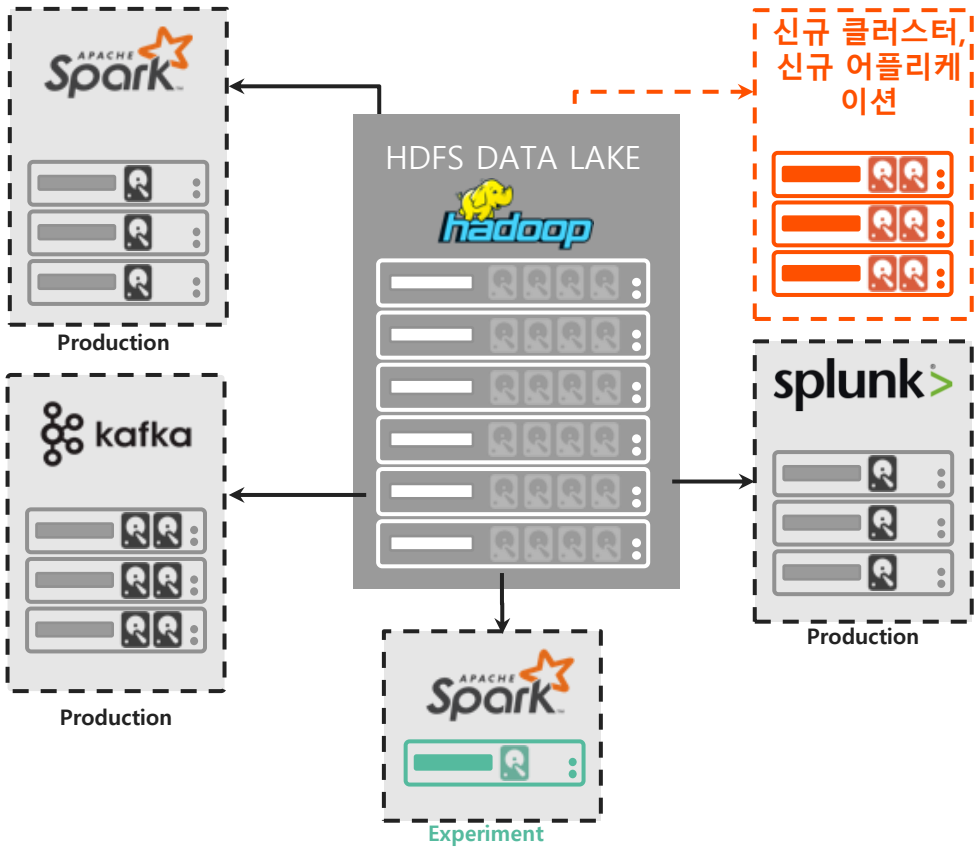
퓨어스토리지는 장애시에도 성능 저하를 최소화해요.

노드 장애 시 데이터 재구성 및 재조정을 제거합니다.

©2024 Pure Storage

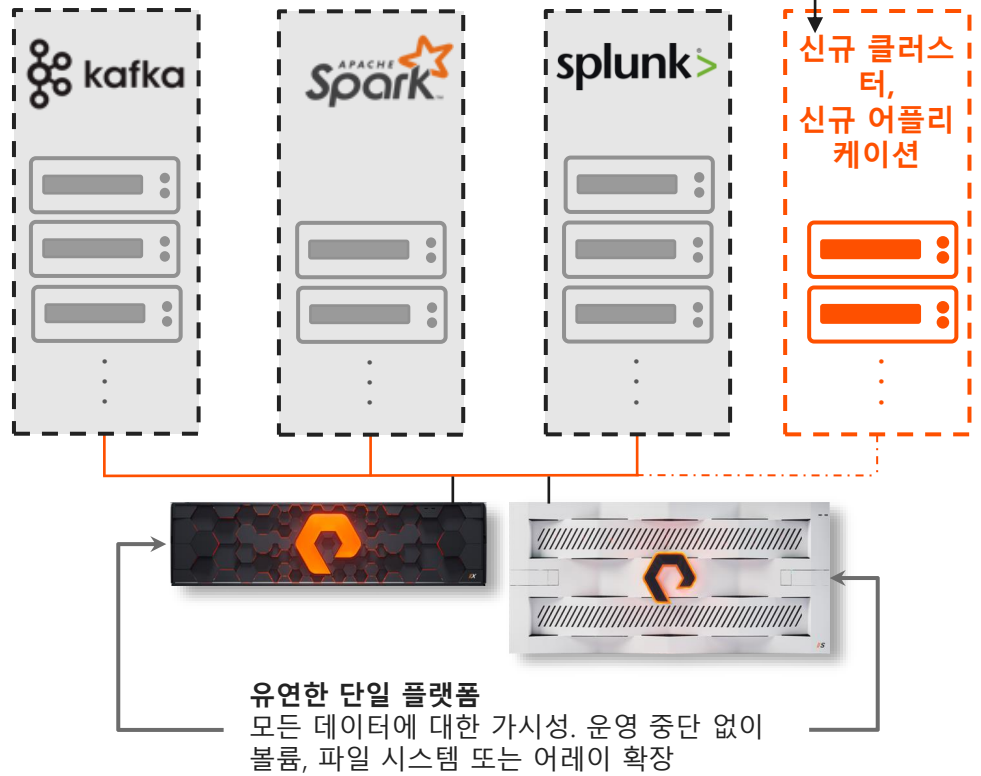
프로비저닝 단순화

레거시 DAS 접근방식



전문적인 분산 접근방식

필요에 따라 새 클러스터 프로비저닝
몇 분 안에 준비되며 간단하게 확장 가능

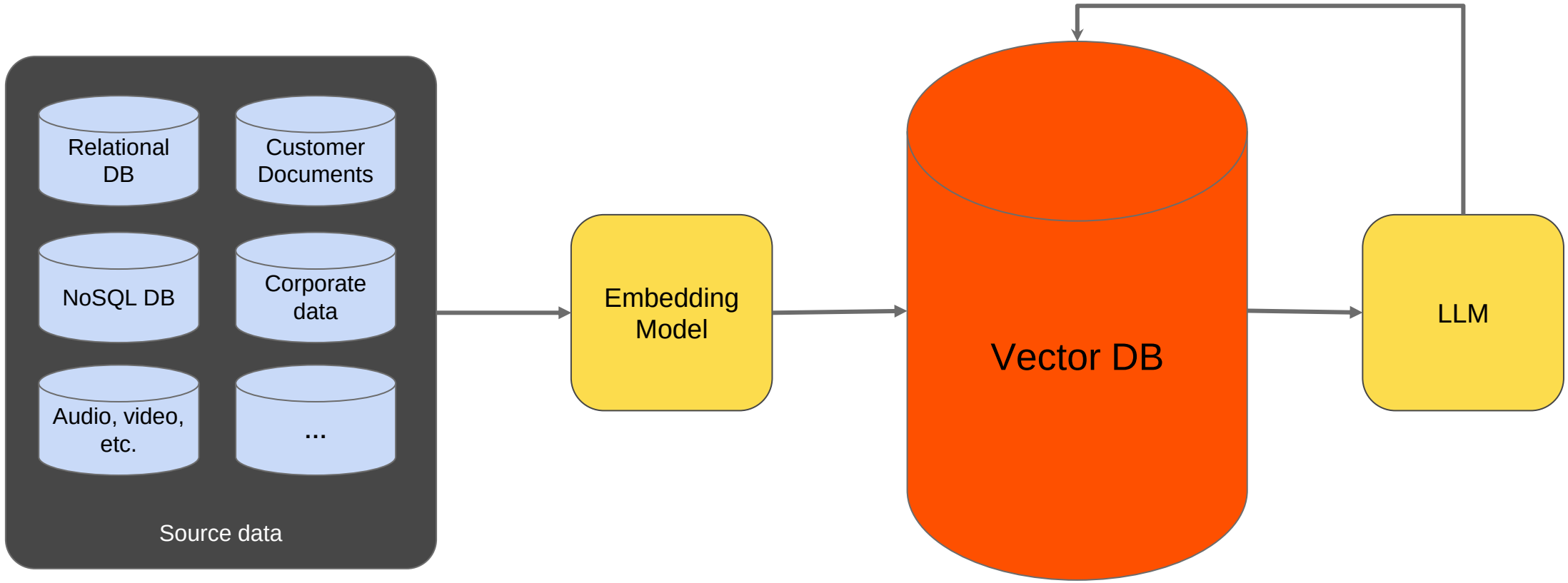


유연한 단일 플랫폼
모든 데이터에 대한 가시성. 운영 중단 없이 볼륨, 파일 시스템 또는 어레이 확장

Silo에 있는 추가 데이터의 비효율성을 제거합니다. 몇 분 만에 새로운 클러스터를 간편하게 프로비저닝할 수 있습니다.

©2024 Pure Storage

데이터 스토리지 요구 용량을 10배 증가 시키는 RAG 환경



e.g. 100 TB

~1PB

©2024 Pure Storage

FlashBlade//S

//Speed
//Simplicity
//Scalability
//Sustainability

- 차세대 All QLC 플래시
- 혁신적인 모듈식 분산 아키텍처
- 입증된 Purity//FB 소프트웨어
- 에버그린 민첩성과 혁신을 위해 설계

More than: **2x** 집적도
성능
전력 효율성

© 2022 Pure Storage, Inc.

시중의 다른 Scale-out 스토리지 솔루션과 비교할 수 없는 AI 성능 효율성

105% RU당 사용 가능한 TB 증가*

48% 더 적은 전력 필요*

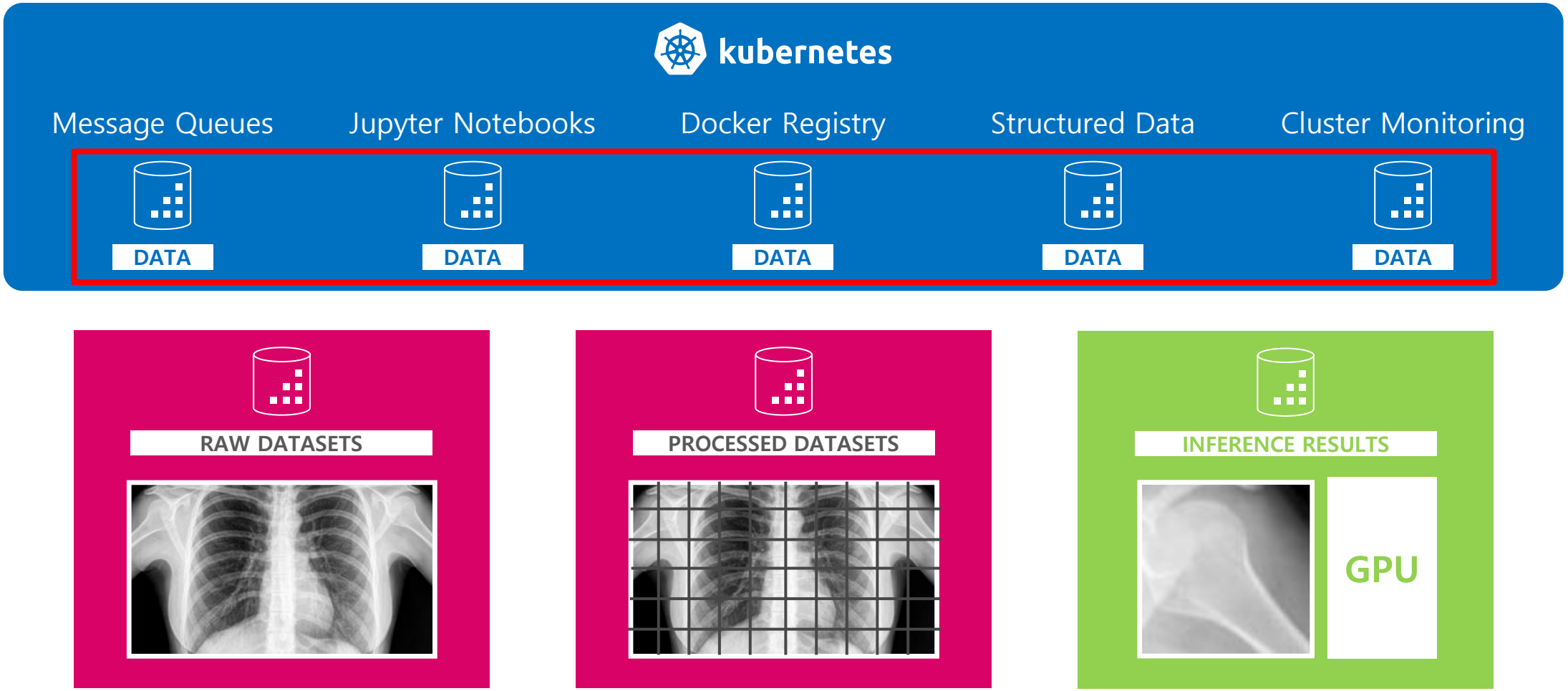
28% 냉각 필요량 감소 *

1.3 Watt per TB

*vs. Legacy NAS

기존 AI 데이터 관리 환경

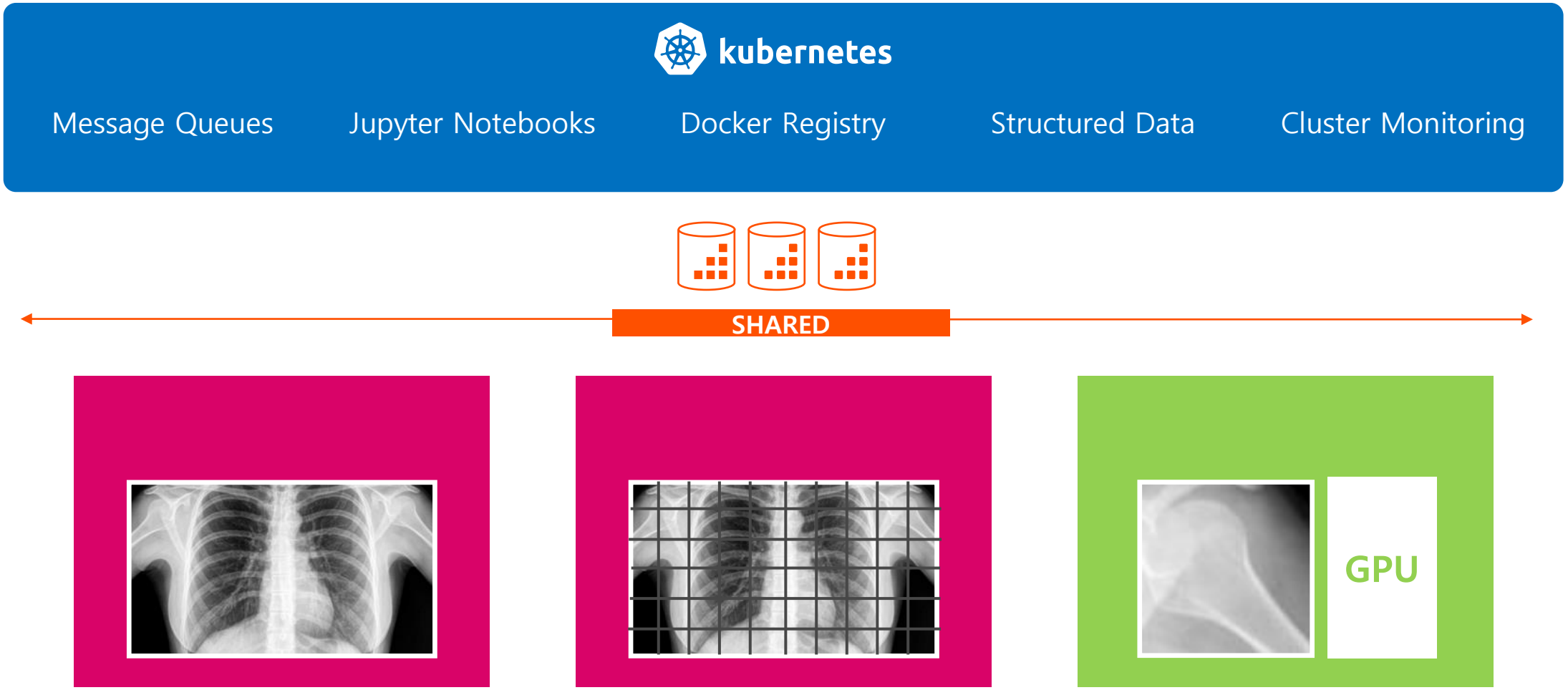
기존의 인퍼런스 파이프라인은 데이터의 불필요한 이동, 중첩된 데이터의 비효율적 보관을 할 수 밖에 없습니다.



©2024 Pure Storage

이상적인 AI 데이터 관리 환경

데이터허브 인퍼런스 파이프라인은 데이터를 중앙저장, 모든 컴퓨팅자원이 공유해서 사용합니다.



©2024 Pure Storage

IOPS? Throughput? 그 이상의 것을 제공

몇 달에 걸친 모델 학습이 며칠만에 끝날 수도 있습니다.

any job | any protocol | any size | any object count | any processing type

Ingestion | Persistence | Processing | Training | Inference

AI 워크로드에는 Multi-dimension 성능이 필요합니다

예측 가능한 확장

필요에 따라 세부적으로 성능 확장

복잡하지 않은 성능

별도의 복잡한 튜닝 불필요

전문 지식 없이 성능 향상

필요한 아키텍처는 이미 제품 내 포함

업계 최고 성능 당 전력

에너지 소비량 85%, 상면 80% 감소

©2024 Pure Storage

단순 스토리지 하드웨어? 그 이상의 것을 제공

Full Stack 데이터관리로 GPU 활용 개선과 AI 모델 서비스 단순화가 가능합니다.

AI INFRASTRUCTURE

Kubeflow, TensorFlow, RAY, jupyter, cassandra, NVIDIA, mlflow, Weaviate, neo4j, Milvus, Pinecone

PORTWORX
KUBERNETES
DATA MGMT

Automate | Protect | Unify

KUBERNETES

OpenShift | EKS | AKS | GKE | Others

AI STORAGE & FULL STACK INFRA

FlashBlade | FlashArray | AIRI DGX BasePOD | NVIDIA-Certified OVX Servers | FlashStack for AI

NEW PEOPLE & PROCESS

- Machine Learning Pipelines
- Model Inference
- Retrieval Augmented Generation
- Model Training
- MLOps
- Data Scientist, AI Engineer

©2024 Pure Storage

PortWorx : 클라우드 네이티브 환경 통합

유연한 확장성, 높은 가용성 및 자동화 서비스를 통해 애플리케이션을 배포하고 컨테이너 데이터를 보호합니다.

portworx Enterprise

Developer-Ready 스토리지

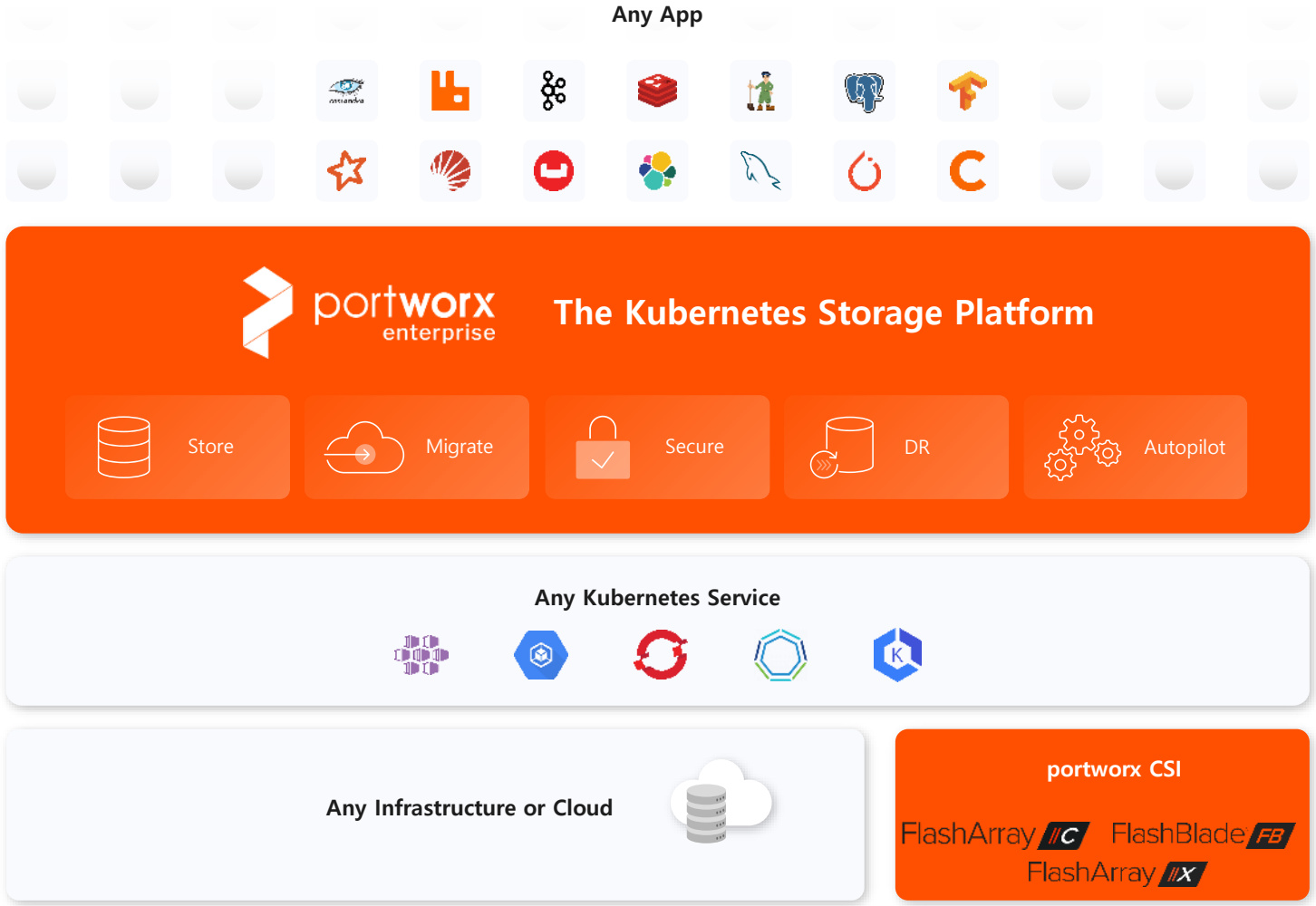
✓ 인프라의 유연성(On-Prem/Cloud)을 통한 개발자의 민첩성 향상
✓ 모든 스토리지 또는 데이터 서비스, 그리고 클라우드 데이터 이동성에 대한 셀프 서비스

애플리케이션 성능의 극대화

✓ 최적의 성능과 온디맨드 확장성을 통한 운영 환경에서 미션 크리티컬 애플리케이션 실행

비즈니스 연속성 보장

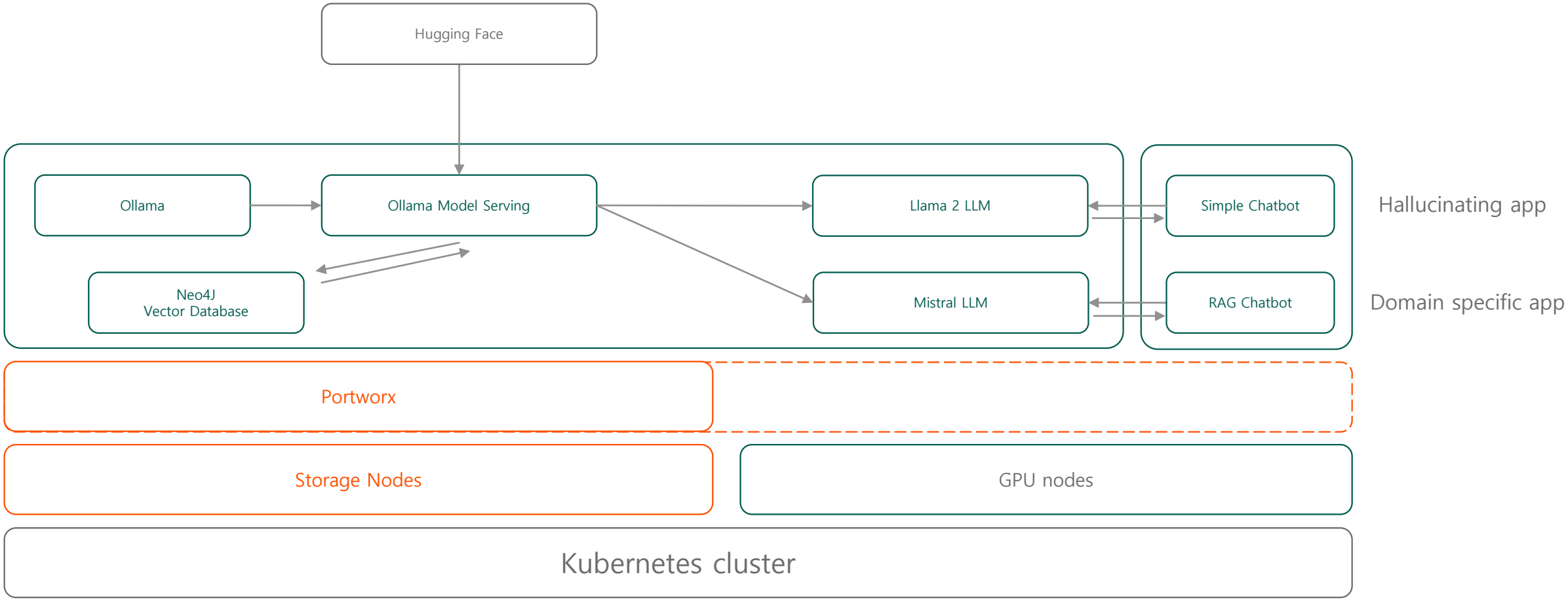
✓ 데이터센터 전반에 걸쳐 재해 복구 시간 단축
✓ 쿠버네티스 운영 환경에 대한 고가용성 및 내결함성 및 복원력



©2024 Pure Storage

클라우드네이티브 + AI워크로드

RAG 모델에 적용하여 기존 아키텍처 개선이 가능합니다.



©2024 Pure Storage

Networked Storage is the Optimal Data Platform for Generative AI

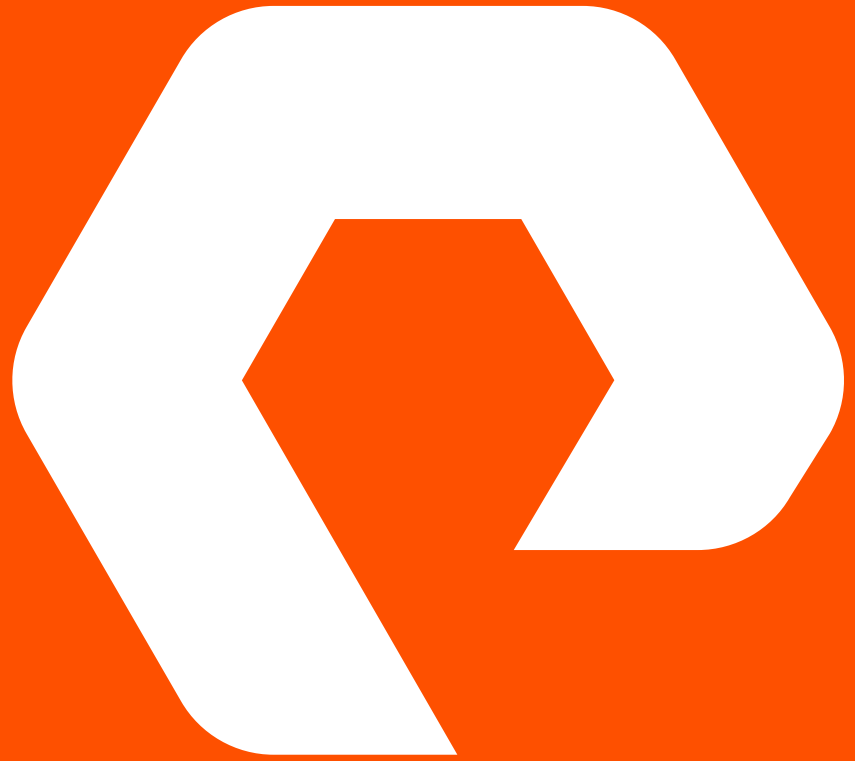
Linear Scaling
높은 처리량, 낮은 대기 시간
제공성능 및 용량 확장
수십 개의 AI 서버 지원

Peak Utilization
고립된 로컬 스토리지 제거
서버와 GPU 간에 데이터 공유
AI 워크플로우의 여러 단계 지원

Data Protection
신뢰할 수 있고 단순화된 보호,
랜섬웨어 완화 및 복구, 암호화,
백업 및 신속한 복구

Composable Storage
신속하고 유연한 프로비저닝
파일 및/또는 객체 스토리지,
클라우드 지원

Pure Storage FlashBlade



Uncomplicate Data Storage, Forever