

[Data Insight & Security Summit 2023]

# AI분석을 위한 초고성능 데이터 레이크

효성인포메이션시스템  
데이터사업팀 강준범 컨설턴트

# Dall-E & ChatGPT

- AI 알고리즘을 기반으로 한 서비스 → 모델 학습을 위한 방대한 양의 데이터와 분석 환경 필요

## 서비스

발라

AI 분석에 필요한 사항은?

발라

AI 분석을 위해서는 다음과 같은 사항들이 필요합니다.

1. 데이터 수집 및 전처리: AI 분석을 위해서는 대량의 데이터가 필요합니다. 이 데이터를 수집하고, 필요한 형태로 가공하여 모델 학습에 사용합니다.

2. 모델 개발: AI 분석을 위해서는 모델 개발이 필요합니다. 이를 위해서는 알고리즘 개발, 모델 학습 및 검증 등이 필요합니다.

3. 컴퓨팅 자원: 모델 개발 및 학습에는 많은 컴퓨팅 자원이 필요합니다. 대규모 데이터를 처리하고, 고성능 모델을 학습시키기 위해 GPU 등 고성능 컴퓨팅 자원이 필요합니다.

4. 인력: AI 분석을 위해서는 AI에 대한 전문적인 지식과 기술이 필요합니다. 이를 위해 전문가나 엔지니어, 데이터 분석가 등이 필요합니다.

5. 보안: AI 분석을 위해서는 데이터의 보안과 개인정보 보호가 필요합니다. 따라서 보안적인 측면에서도 충분한 대책이 필요합니다.

발라

AI 분석 환경을 알려줘

발라

AI 분석 환경은 데이터 분석 및 모델 학습을 위한 최신 기술과 하드웨어를 포함하는 환경입니다. 일반적으로 다음과 같은 구성 요소로 구성됩니다.

1. 데이터 수집 및 저장: AI 분석 환경은 대량의 데이터를 수집하고 저장할 수 있는 고성능 데이터 저장 시스템을 필요로 합니다. 이를 위해서는 대용량 파일 시스템, 데이터베이스, 데이터 레이크 등을 활용하여 데이터를 저장하고 처리할 수 있어야 합니다.

2. 컴퓨팅 리소스: AI 분석을 위해서는 대규모의 컴퓨팅 리소스가 필요합니다. GPU, TPU 등의 고성능 하드웨어를 활용하여 모델 학습 및 추론을 수행할 수 있습니다.

3. 분산 처리 시스템: 대규모 데이터를 처리하기 위해서는 분산 처리 시스템이 필요합니다. 이를 위해서는 Hadoop, Spark 등의 분산 처리 시스템을 활용할 수 있습니다.

4. 머신 러닝 및 딥 러닝 프레임워크: AI 분석 환경에서는 TensorFlow, PyTorch, Apache MXNet 등의 머신 러닝 및 딥 러닝 프레임워크를 활용하여 모델을 학습하고, 추론할 수 있습니다.

## Insight

새롭게 생각하고 싶은 그대에게  
삶의 목적을 찾는 45가지 방법

올 챗GPT · 번역 AI · 일러스트 시터스톡 AI



인간 출판 기획자와 AI가 펴낸 최초의 책!

인간이 개발한  
기획자의 자카르타  
참고용 작성한  
챗GPT

전문가 교육  
영역인 번역제와  
자리를 대체한  
번역 AI

장르적 영역으로  
인사받은 일러스트로  
작 표지로 담은  
시터스톡 AI

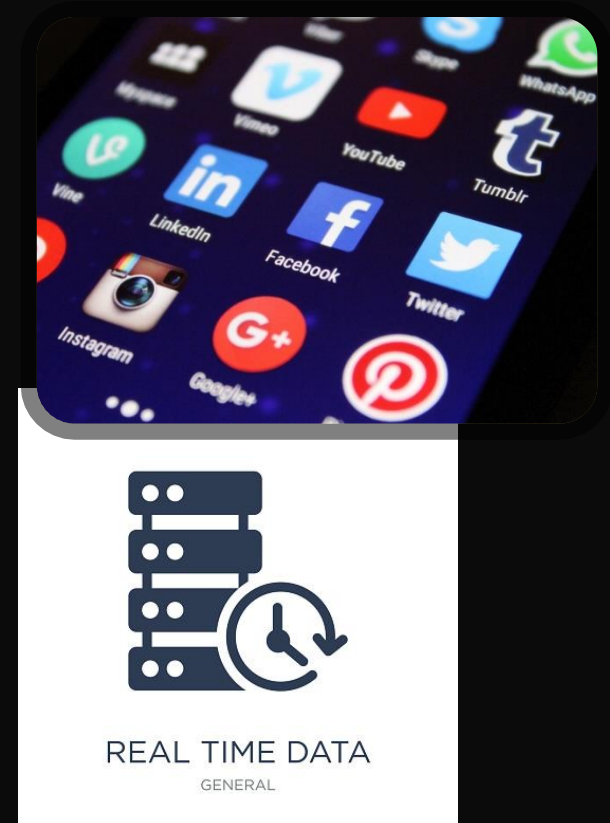
챗GPT까지 AI가  
대신한 승객적  
결과물

시는 인간을 이해하고 감정을 전달하는, 인생의 동반자. 인간에게 가장 가까운 친구이자, 가장 소중한 존재. 인간에게 가장 가까운 친구이자, 가장 소중한 존재. 인간에게 가장 가까운 친구이자, 가장 소중한 존재.

모든 책의 인제는 30%의 번역, 일러스트, 편집, 디자인, 마케팅, 유통으로 이루어집니다.

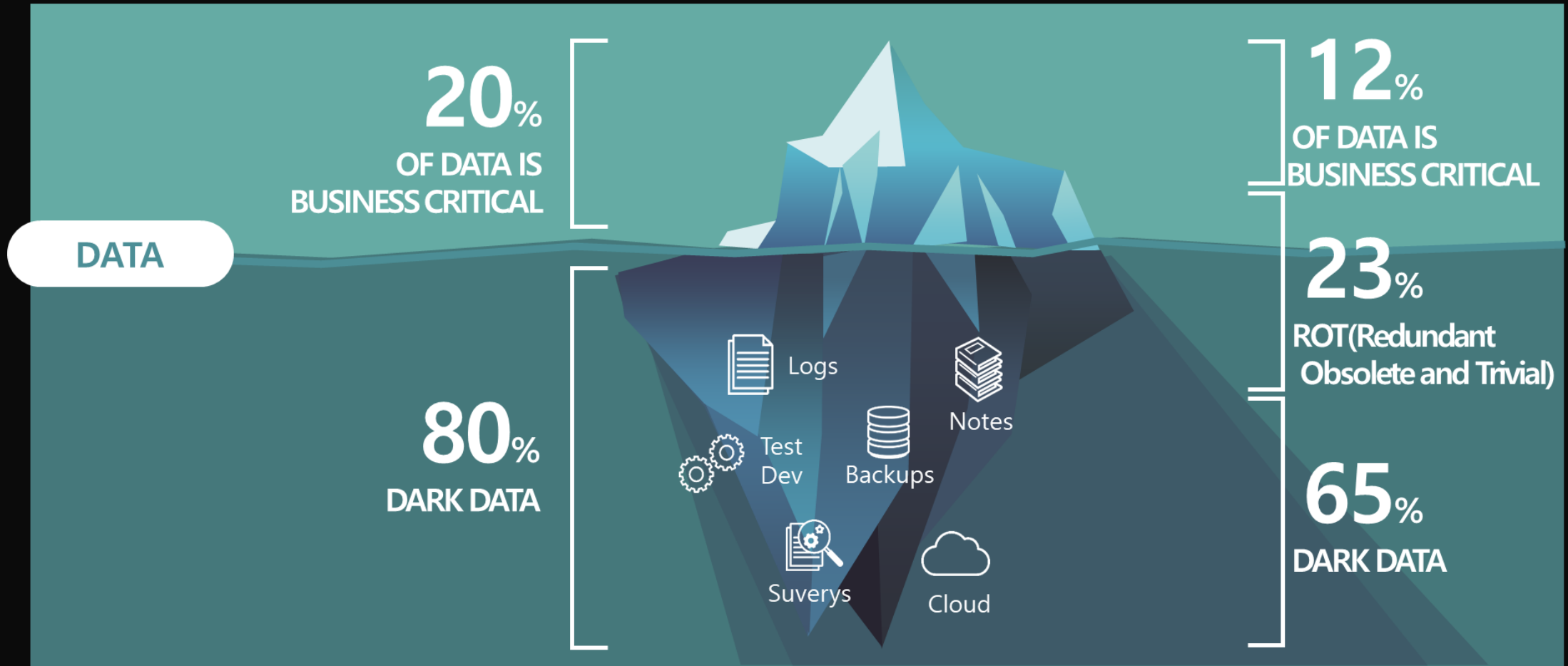
# 데이터의 발전

- 데이터의 **형태 발전**과 기하급수적 **용량 증가**



# Dark Data

- 다크 데이터, 수집 및 분석 가능한 **도구 부재!** 너무 많은 데이터! 불완전한 데이터!

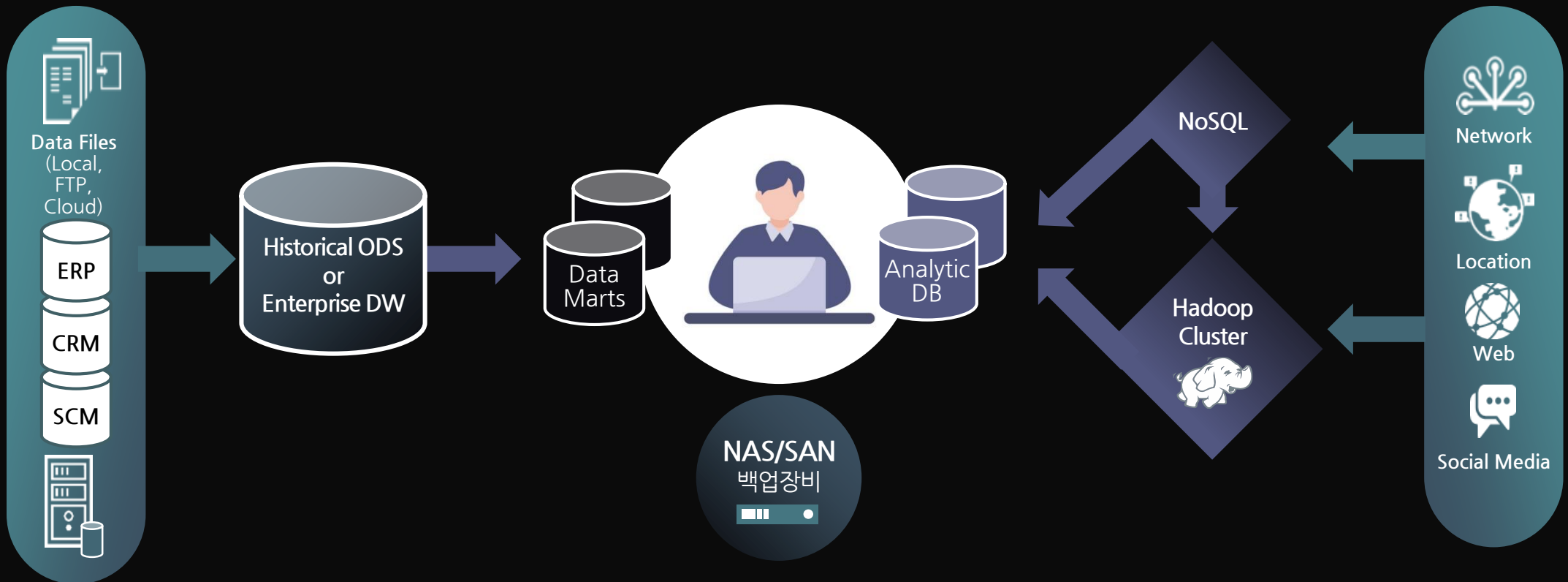


Source : Gartner, IBM, Datumize

# Hadoop기반 초창기 데이터 레이크 아키텍처

- DW 정보계 시스템, 초창기 Data Lake 시스템

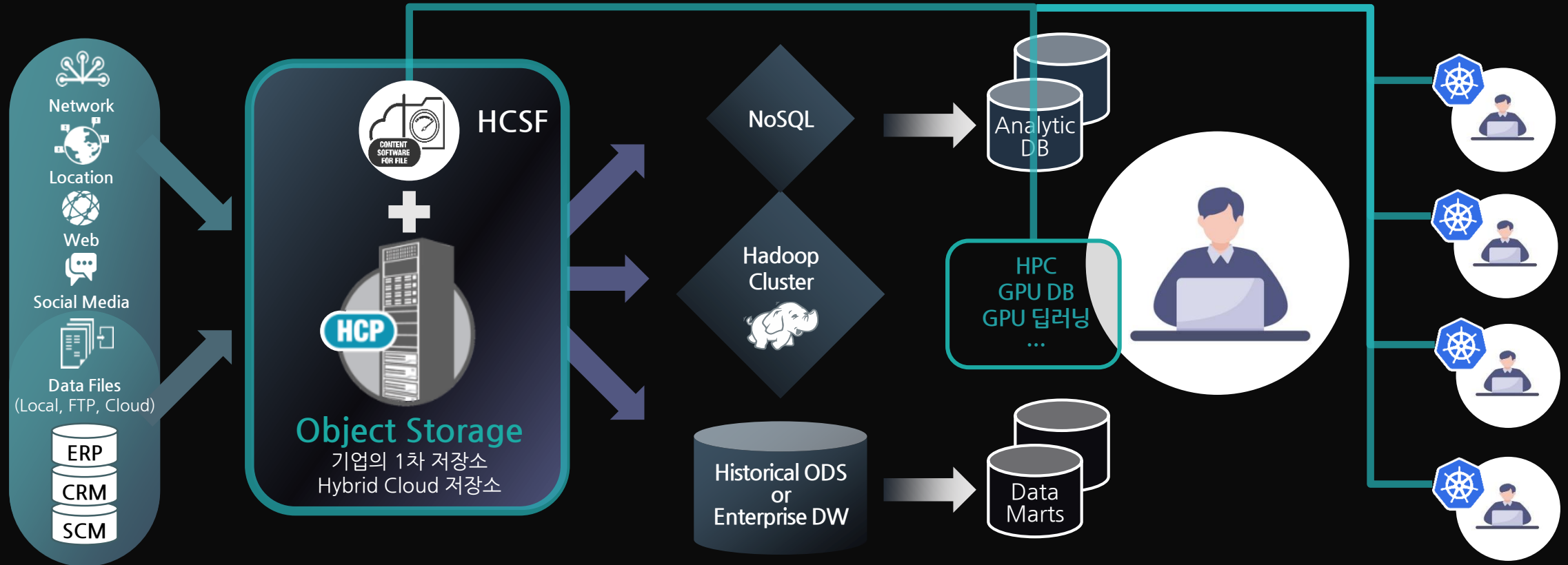
- ✓ 데이터를 모으기 위한 다양한 저장소 발생, 데이터 중복 등 **사일로 및 연관관계 도출의 어려움**
- ✓ 모든 데이터를 데이터 타입에 상관없이 저장하기 위해 **데이터 레이크 필요** (1차 저장소)
- ✓ 분석 환경 및 분석 엔진 추가 도입 시 **추가 저장소** 발생





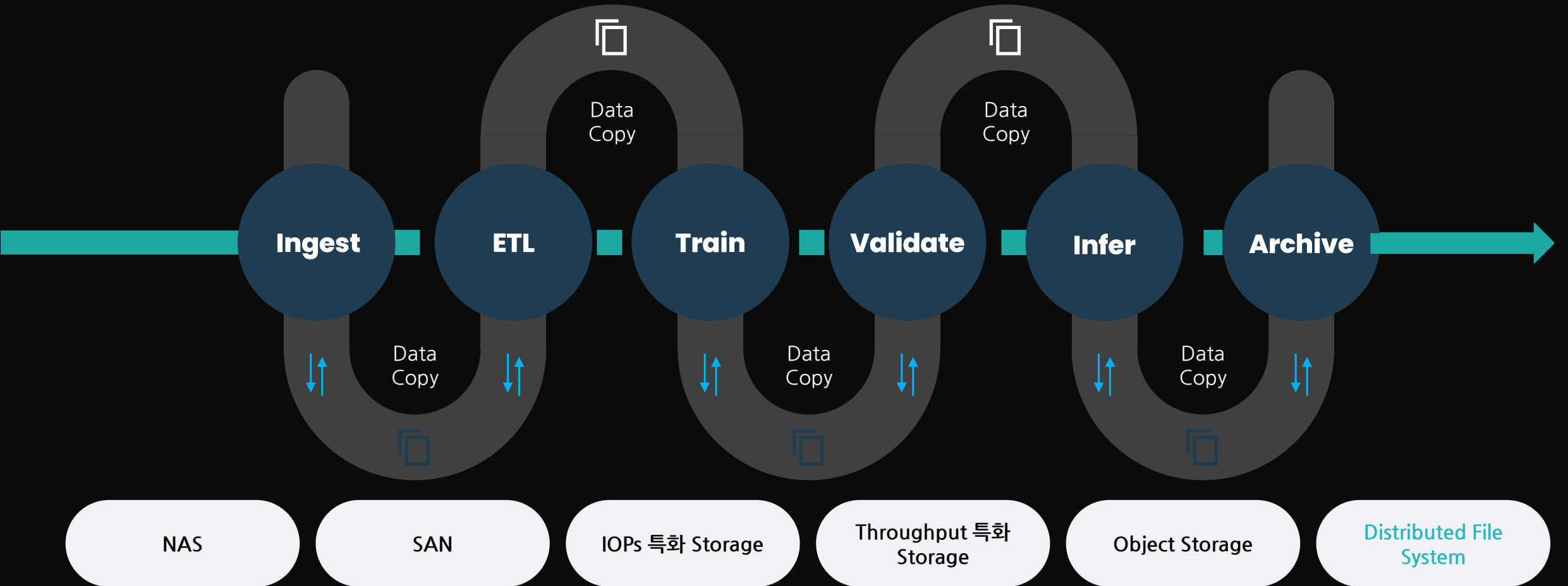
# 차세대 고성능 데이터 레이크 아키텍처

- 단일저장소 기반 Data Lake / 고성능 분석 도구(GPU) / 고성능 스토리지
  - ✓ 새로운 저장소가 발생할 때 마다 레거시 시스템에 연결하지 않고 오브젝트 스토리지로 연결 - 진정한 데이터 레이크 활용 가능
  - ✓ 오브젝트 스토리지와 오토 티어링 가능한 고성능 분산 스토리지 - 고성능 분석 도구의 스토리지 병목 현상(bottleneck) 해결



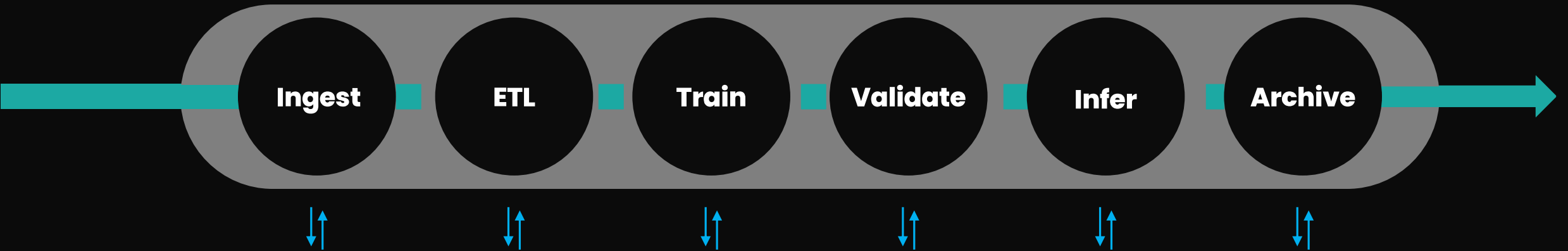
# AI Data Pipelines의 저장소

- ✓ Pipeline 수행도구(수집/학습/검증/추론/저장)를 위한 Data Copy 필요
- ✓ 성능의 목적 및 지원 프로토콜에 의한 다양한 저장소 필요



# AI Data Pipelines의 저장소

- ✓ 별도의 튜닝 없이 높은 IOPs, Throughput 보장
- ✓ 멀티 프로토콜 지원을 통한 다양한 분석엔진과의 유기적 연동
- ✓ 무제한 확장성, 노드 증가에 따른 선형적 성능 향상



**Unified Single File System**

Zero Copy, No Tuning Architecture



# Deep Learning IO

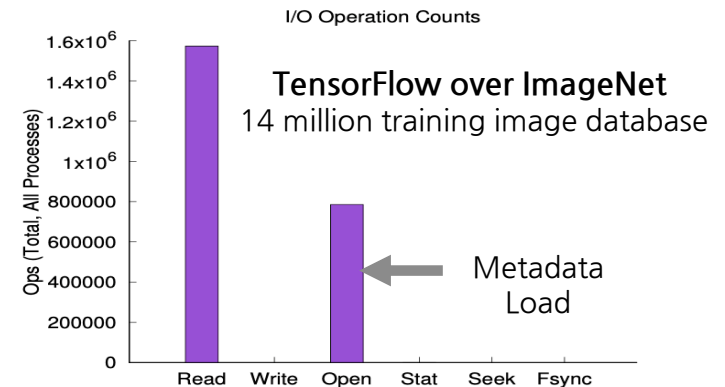
- 많은 Small Files에 대한 빈번한 Random Read 요구

## Deep Learning IO 프로세스

- **Mini batch** - 학습 데이터에서 임의의 하위 집합 반복  
각 mini batch에서 확률적 경사 하강법 수행  
(Stochastic Gradient Descent: SGD)
- **Epoch (반복)** - 전체 데이터 세트에 대해 Random 처리

## Deep Learning IO 특성

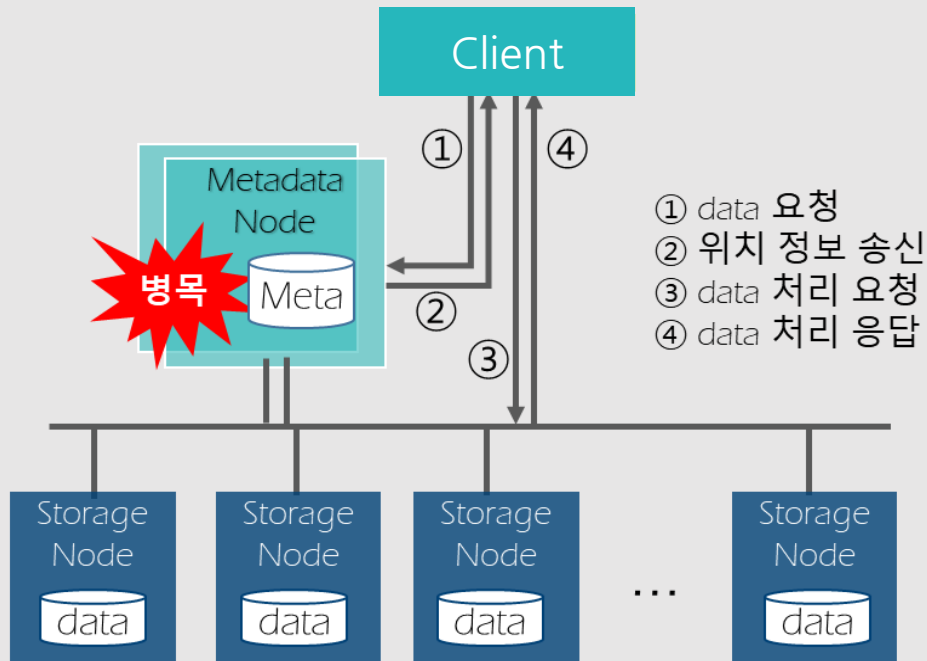
- 매우 빈번한 small IO 요청
- File을 찾기 위한 분산 파일 시스템의 Metadata overhead
- IOPs와 Throughput이 중요



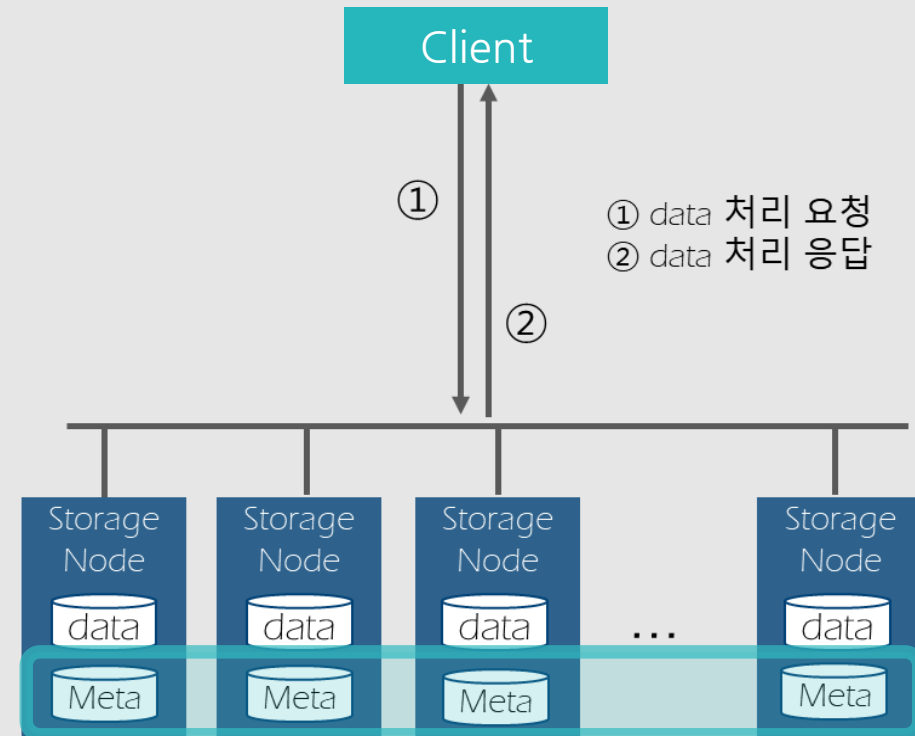
# AI / HPC 환경을 위한 고성능 데이터 레이크 저장소 - MetaData 관리

- AI/ML 환경에서의 스토리지

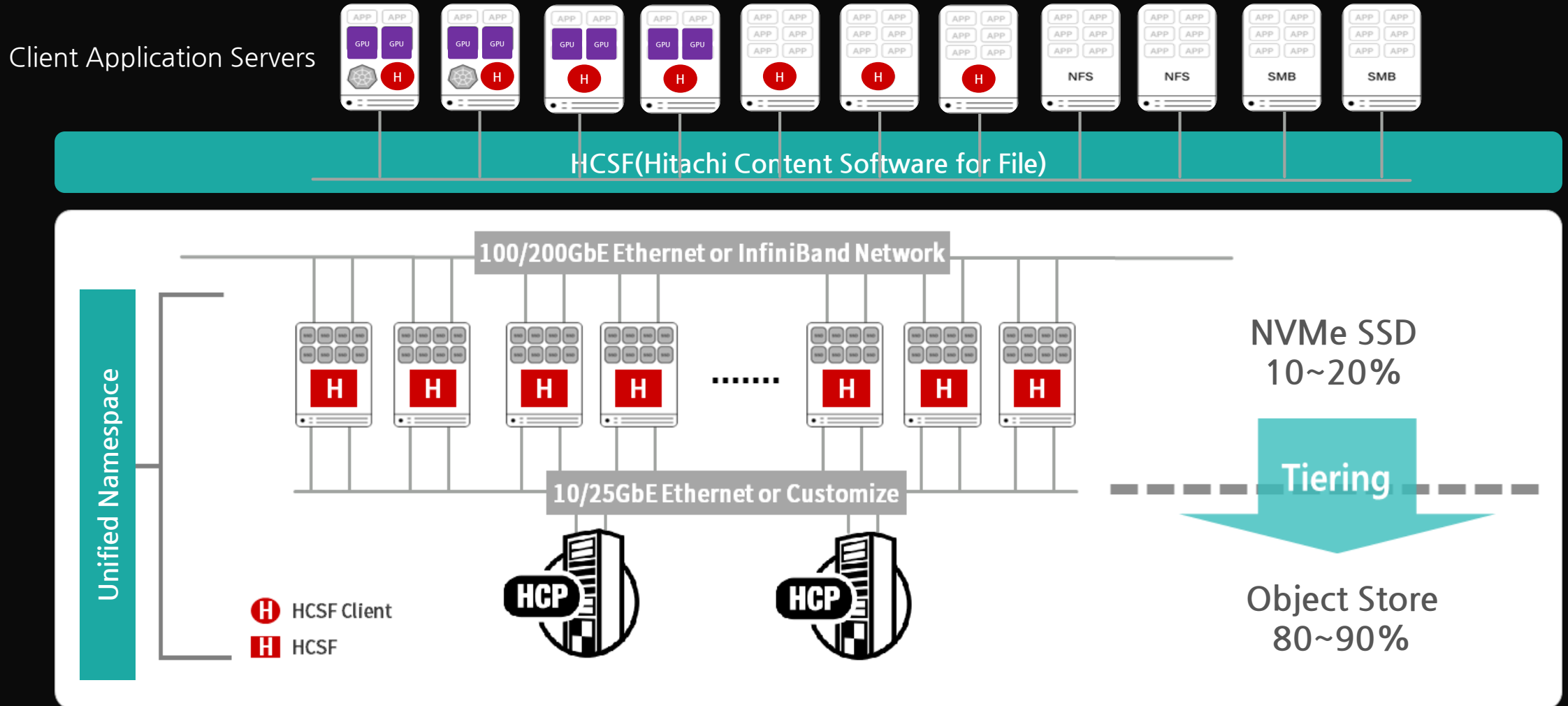
## 일반적인 병렬 파일시스템



## HCSF 병렬 파일시스템

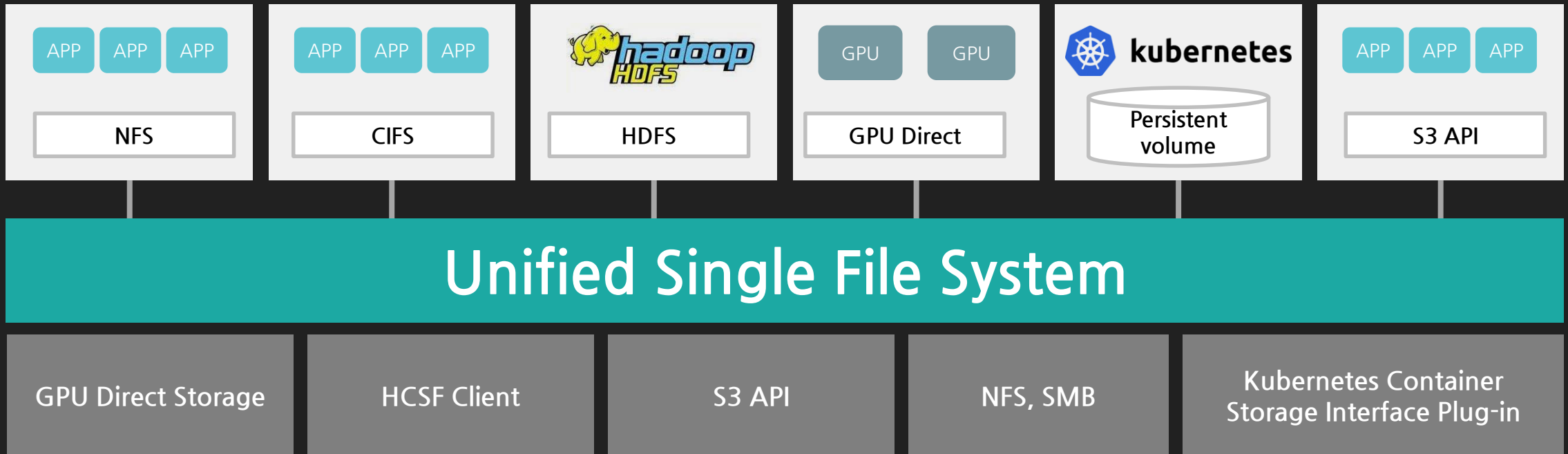


# AI / HPC 환경을 위한 고성능 데이터 레이크 저장소 - Architecture



# AI / HPC 환경을 위한 고성능 데이터 레이크 저장소 - Multi Protocol

- ✓ HCSF Client (전용 Client)
- ✓ GPU Direct Storage
- ✓ NFS, SMB, S3
- ✓ Kubernetes CSI 지원

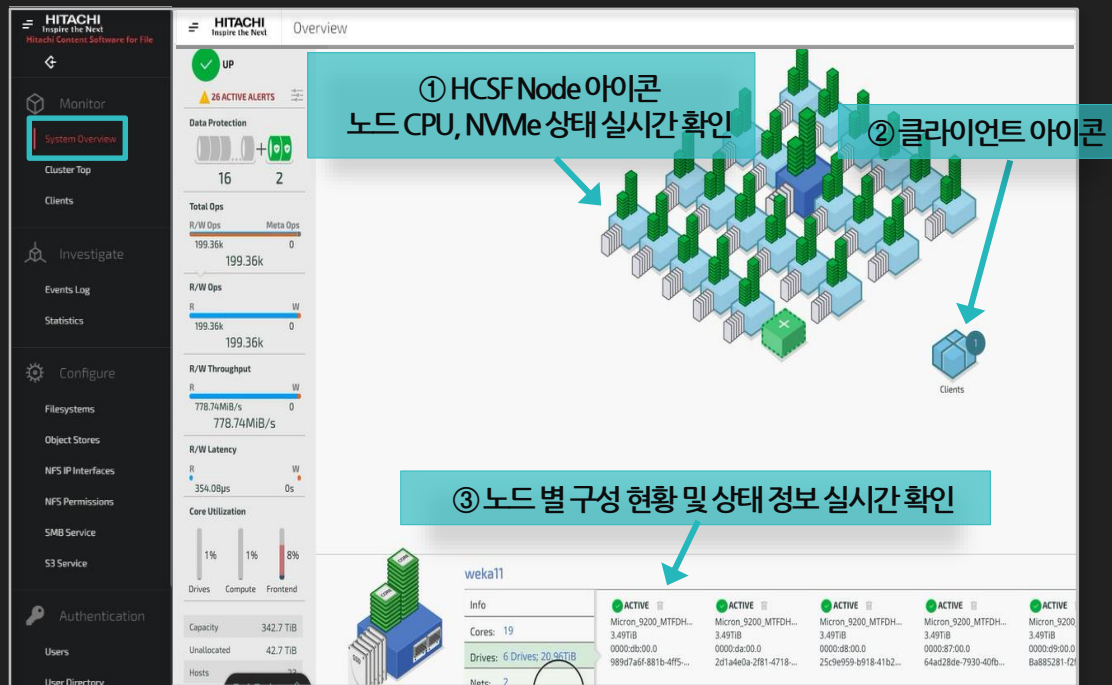


# AI / HPC 환경을 위한 고성능 데이터 레이크 저장소 - 시스템 관리 모니터링

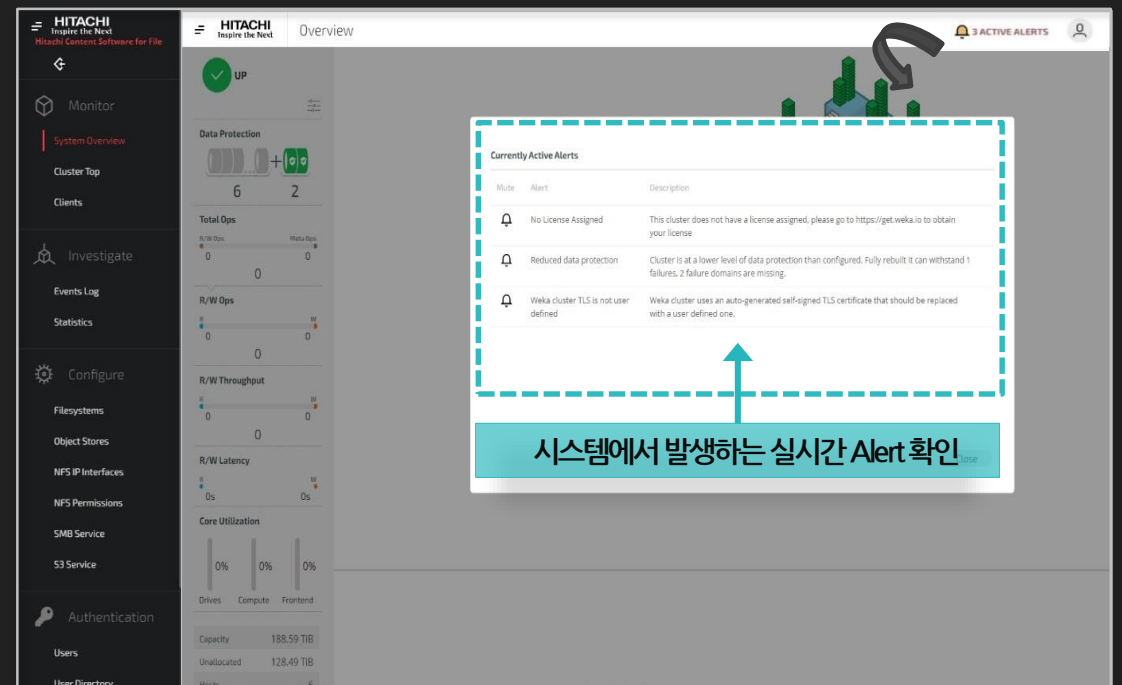
- CLI와 GUI 기반 관리를 기본으로 지원

- ✓ 하드웨어 구성, 장애 및 모니터링, 리소스 사용률, 데이터 처리량, 스토리지 전체 성능 실시간 모니터링

## HCSF Overview 화면



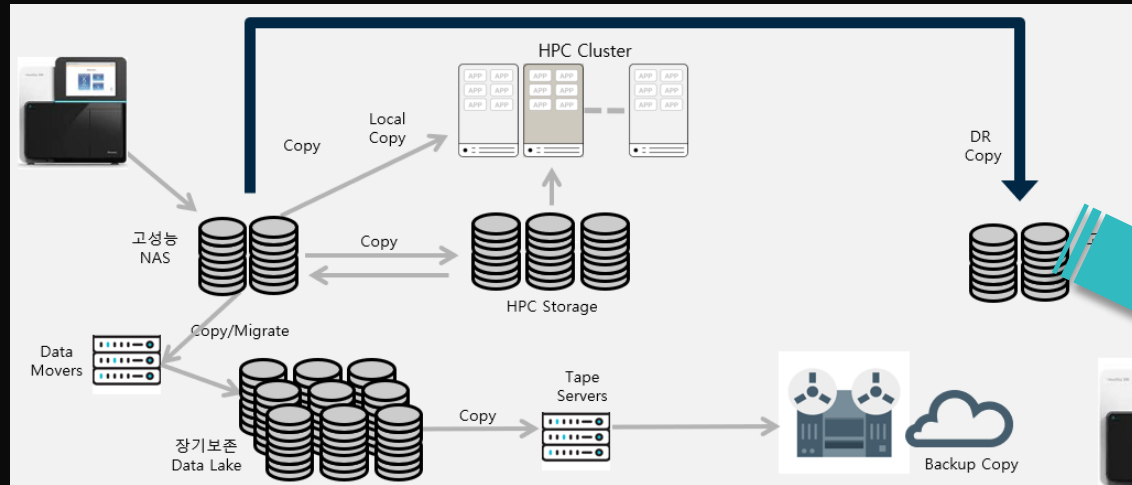
## HCSF Currently Active Alerts 화면



# A 글로벌 기업 고성능 데이터 레이크 사례

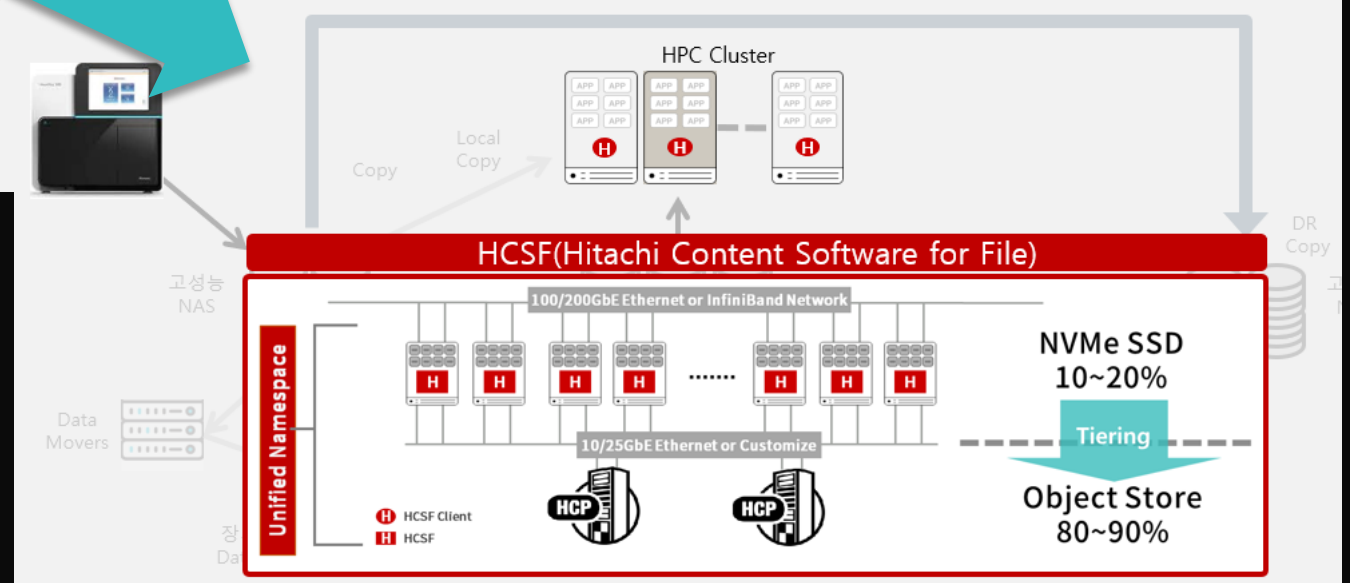
## • 글로벌 A 기업

- ✓ 데이터 중복 및 silo를 제거하고 관리를 용이
- ✓ 복잡한 스토리지 환경을 단순화



## Zero-Copy 아키텍처

- 단일 시스템으로 축소/통합되어 인프라 비용 최대 75% 절감
- 실행 가능 데이터로 전환되는 소요 시간 95~98%까지 단축
- 단일 네임스페이스에서 14EB의 거대한 확장성
- 관리 오버헤드 20배 감소

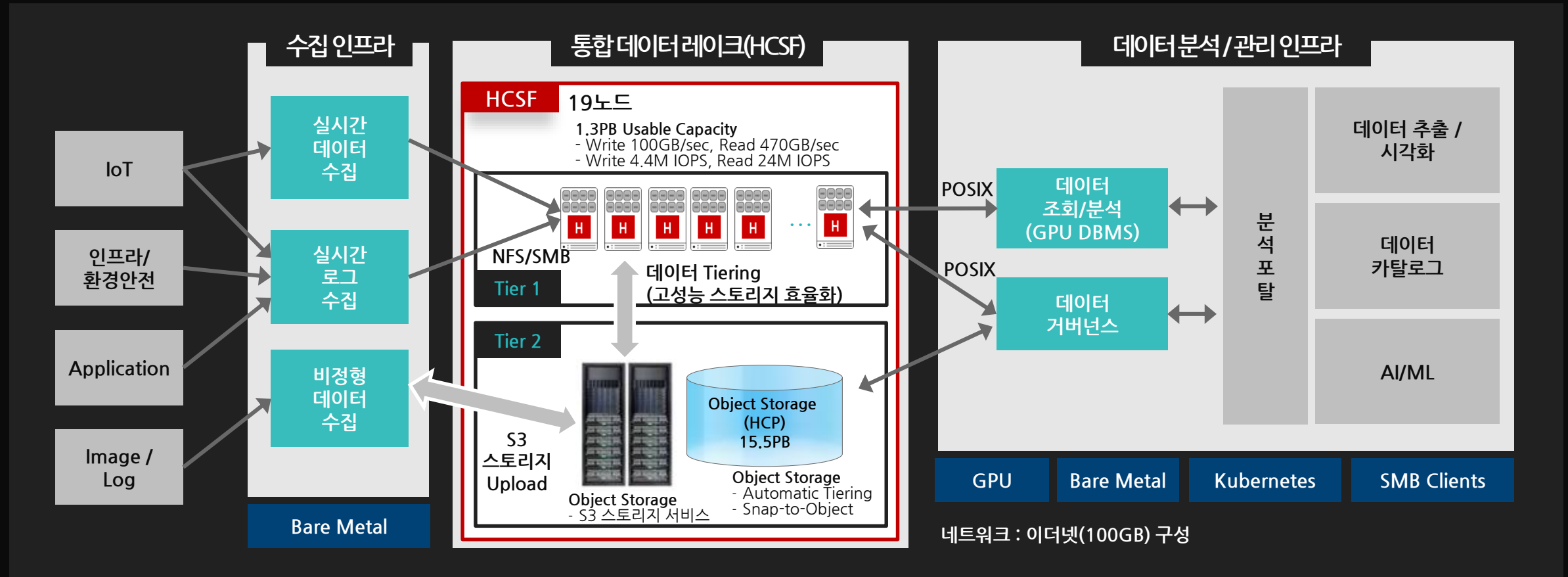




# B 제조사 고성능 데이터 레이크 사례

## • Local B 제조사

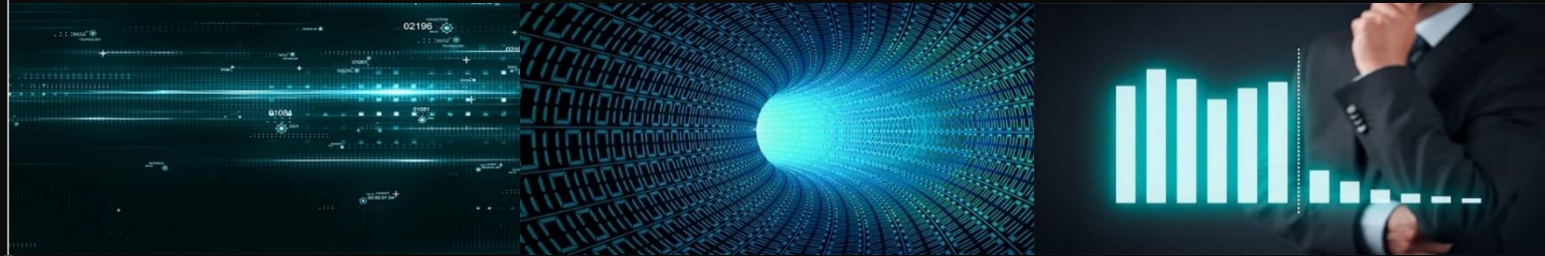
- ✓ DW / Hadoop 데이터 분석 체계 운영 고객사의 차세대 전사 데이터 분석 체계의 통합 저장소 구축
- ✓ 고성능 데이터 분석 기반과 매우 다양한 데이터 운영 환경을 모두 수용



# AI / HPC 환경을 위한 고성능 데이터레이크 저장소 - HCSF

## HCSF?

## Hitachi Content Software For File



### 1 슈퍼컴퓨터/HPC를 위한 고성능 병렬 분산 스토리지 (SC19 IO500 1위)

### 2 AI/ML 처럼 IO 집약적인 워크로드에 적합

- ✓ 높은 처리량, 높은 IOPs, 매우 짧은 대기 시간이 동시에 필요한 혼합 워크로드에 특화
- ✓ DPDK, GDS(네트워크 패킷 처리 기술)
- ✓ EB규모의 확장 및 선형적 성능 향상 및 파일 크기에 따른 성능 제약 없음

### 3 zero-copy (진정한 Data lake storage)

- ✓ POSIX, NFS, SMB, S3, CSI 및 GPUDirect® Storage(GDS) 지원, 데이터는 모든 프로토콜 간에 완전히 공유
- ✓ Auto tiering은 정책 기반으로 별도 솔루션 없이, 성능 영향없이 자동 동작

### 4 MSA 기반 아키텍처로 손쉬운 노드 확장과 Metadata 성능저하 방지

- ✓ 타사 대비 확장 간편성
- ✓ 타 분산파일시스템 대비 데이터 증가에 따른 Metadata 성능 병목 현상 없음
- ✓ 통합관리 포털 및 손쉬운CLI를 통해 관리의 편리성 제공

### 5 혼합 워크로드에서 초고성능 분석 환경을 위한 스토리지 통합 솔루션

- ❖ GDS : NVIDIA GPU Direct Storage 의 약어. GPU - Storage 병목 구간 고속 처리 기술
- ❖ DPDK : Data Plane Development Kit의 약어. 데이터 패킷 네트워크 고속 처리 기술



## Design Your DX Journey! DX센터 시즌3

HYOSUNG  
효성인포메이션시스템

Accelerate Data-Driven Innovation.

1

디지털 코어 현대화

SDDC부터 PaaS 플랫폼까지  
최적화된 클라우드 인프라 환경

2

데이터 옴스 최적화

데이터 패브릭 실현, 데이터 레이크를  
통한 대용량 데이터 저장/분석

3

AI / HPC 플랫폼

GPU 기반 데이터 연산부터 AI/ML옴스  
솔루션 및 성능최적화를 위한 AI 인프라 설계





감사합니다.