



Tech Talk

# “Why Pure”



# Why data analytics with Pure?

Consolidating and Accelerating Data Analytics  
with Pure Storage

**Yifeng Jiang**

Principal Field Solutions Architect, Data Science

# Agenda

- Technology trend in data analytics and AI
- Data analytics and AI with Pure Storage
- Case study and summary

# Data analytics platform challenges and trend

Complicated and expensive data analytics & AI

- Large infrastructure, expensive software license
- Complicated platform leads to slow performance and low business value

Separating compute and storage

- Flexibility to scale independently like a service
- Deploy and scale easily and quickly
- Save cost



Photo by [Paul Skorupskas](#) on [Unsplash](#)

# Technology trends in data analytics

## S3: the true open data lake

- Single source of truth for big data
- In the cloud and on-premise

## Pluggable data lakehouse

- Same data, different engine, no data copy
- Open table format & storage API
- Choice of data lakehouses

## Kubernetes: key for as-a-service architecture

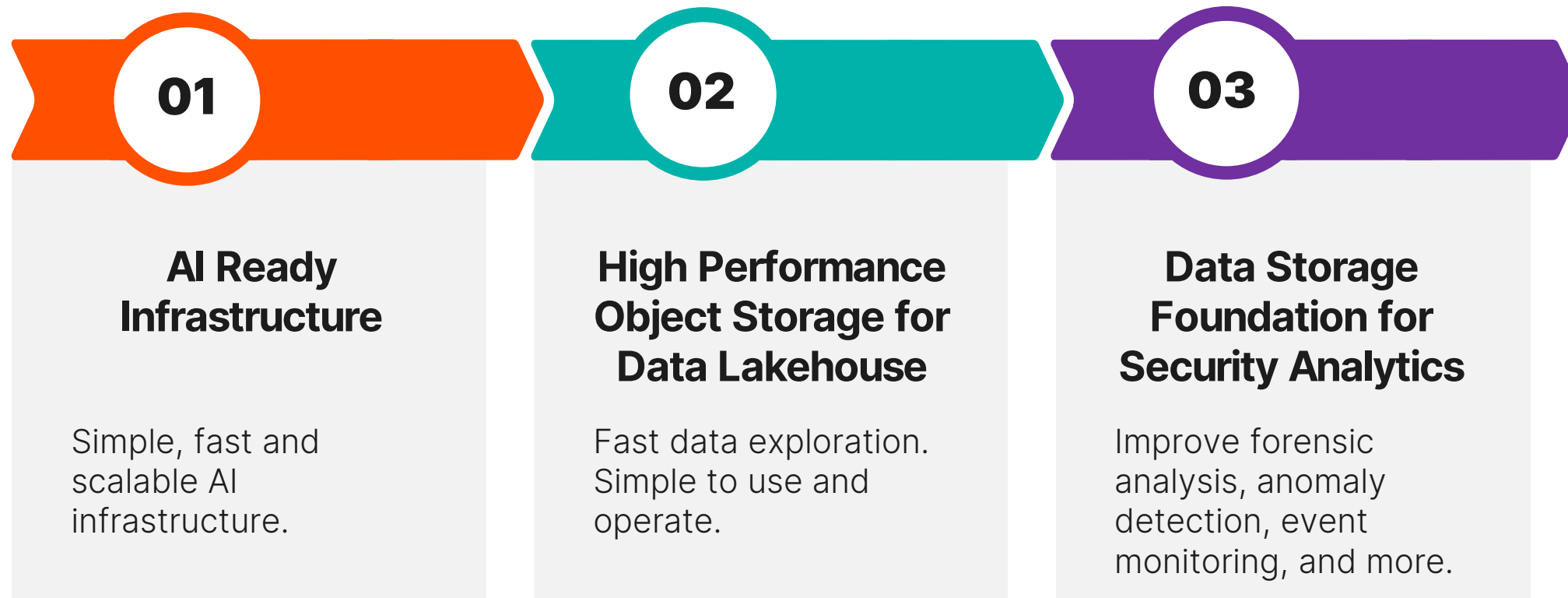
- Share and isolate cluster resource efficiently
- Streamline operation with standard API

## Distributed AI

- AI training across multiple GPU servers
- Your very-own ChatGPT

# Pure Storage's Solution

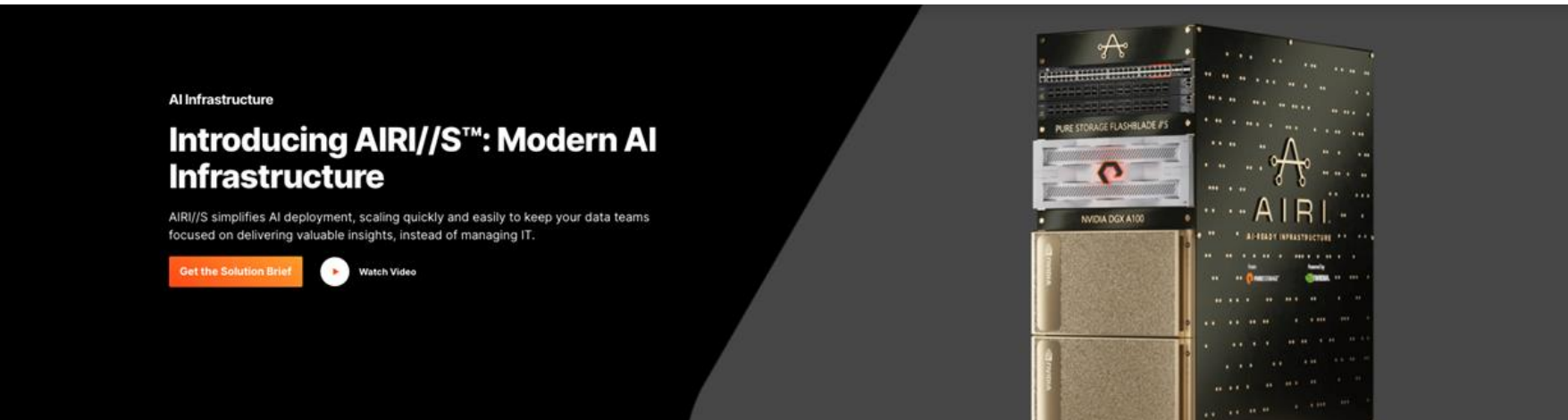
Consolidating and accelerating data and AI infrastructure





# AI ready infrastructure (AIRI)

AIRI//S - simple, fast and scalable AI infrastructure, designed by Pure Storage and NVIDIA



AI Infrastructure

**Introducing AIRI//S™: Modern AI Infrastructure**

AIRI//S simplifies AI deployment, scaling quickly and easily to keep your data teams focused on delivering valuable insights, instead of managing IT.

[Get the Solution Brief](#) [Watch Video](#)



REFERENCE ARCHITECTURE

## AIRI® for AI Workload Scaling

Get faster time-to-insights with Pure Storage® FlashBlade® and NVIDIA DGX A100.

Purity **FB**

FlashBlade® **FB**



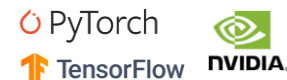
Modern  
Data  
Protection



Modern  
Analytics



AI Cluster



Modern  
Apps



Healthcare  
PACS



Cloud Ready



## A Rich Set of Solutions

Allows you to consolidate modern data applications onto a single scalable platform.

Eliminate complex and inefficient infrastructure silos.

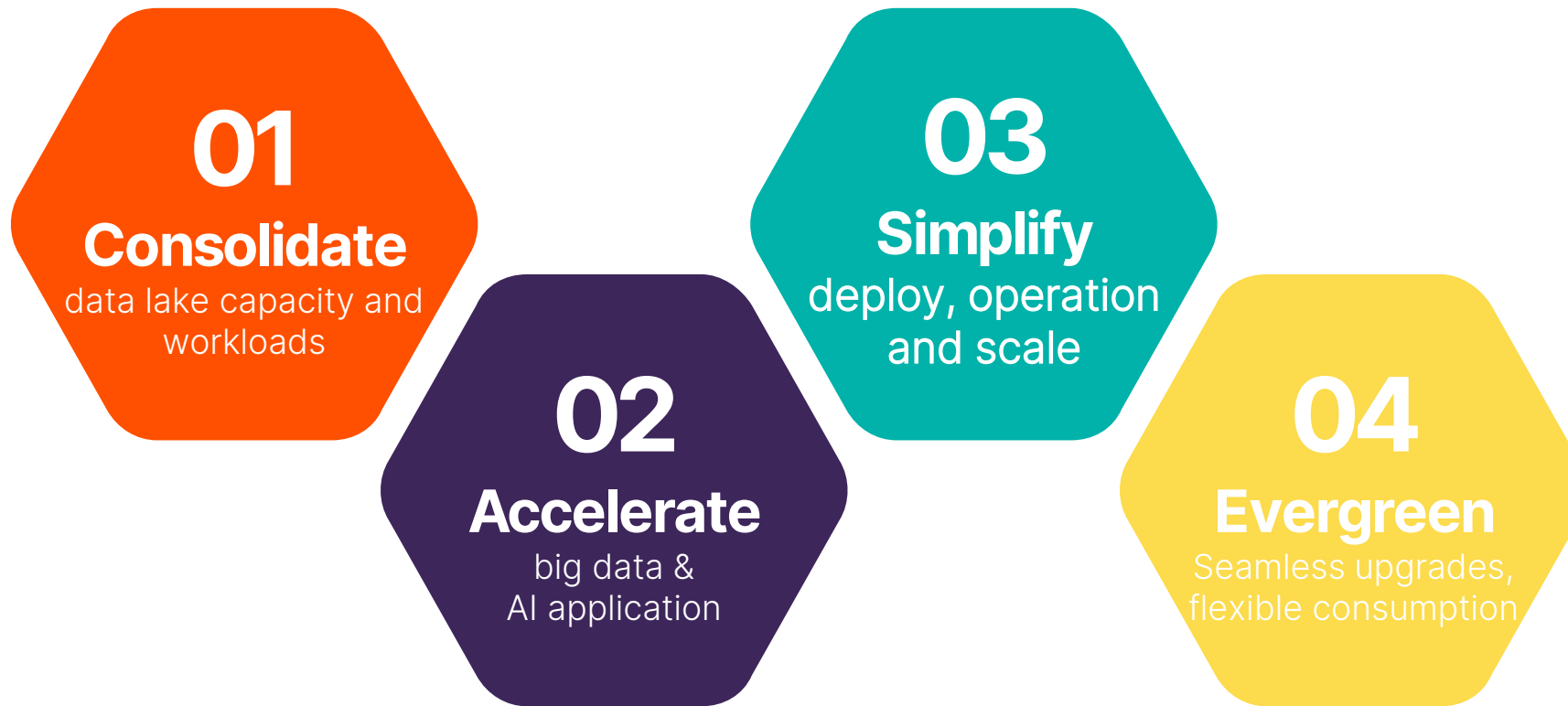
Achieve new levels of investment protection.

Flexible Deployment Models



# FlashBlade benefits for big data and AI

Unified Fast File and Object storage (UFFO)



# Capacity and workload consolidation with FlashBlade

Real sizing for a 10PB HDFS project

**Consolidate** datalake capacity with high density and efficiency

- 2 chassis (12RU) for 3.3PB usable capacity
  - 6x more efficient than HDFS (10PB HDFS equivalent requires 40 servers / 80RU)
- Reduce DC footprint, carbon emission and TCO.
- Size and scale compute and storage separately.

**Consolidate** big data and AI workloads

- Protocol and performance for mixed workloads, including fast S3 for datalake and fast NFS for AI/ML.
- Avoid data silos.

# FlashBlade//S at a glance

## POWER



**56TB** and  
**112 TB** blades

**7 to 100** blades in  
a single system

## SPEED



Up to **48M NFS IOPS**  
in a single system

Up to **30GB/s** bandwidth  
per chassis

## CAPACITY

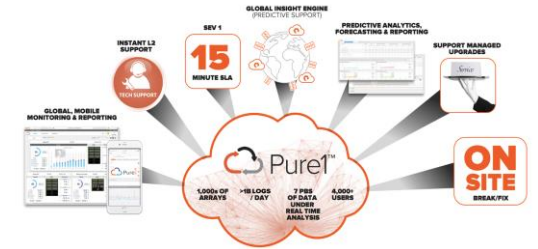


 **DirectFlash**

Up to **2 PB**  
usable capacity  
per chassis and **20PB** in a  
Single System

10 chassis, 52RU for 20PB.

## MANAGEMENT



Cloud-based Intelligent  
storage management  
platform

> NFS

> SMB

> OBJECT

> HTTP



# Accelerate big data and AI application

Real sizing for a 10PB HDFS project

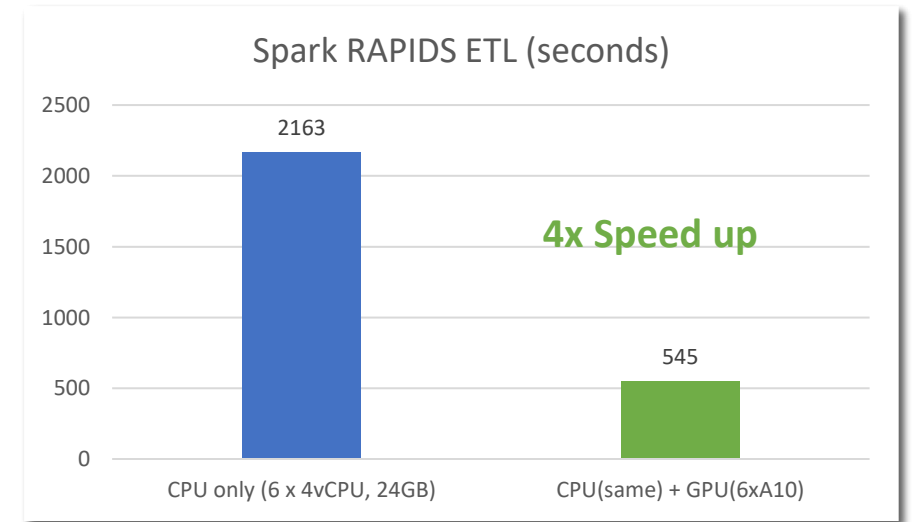
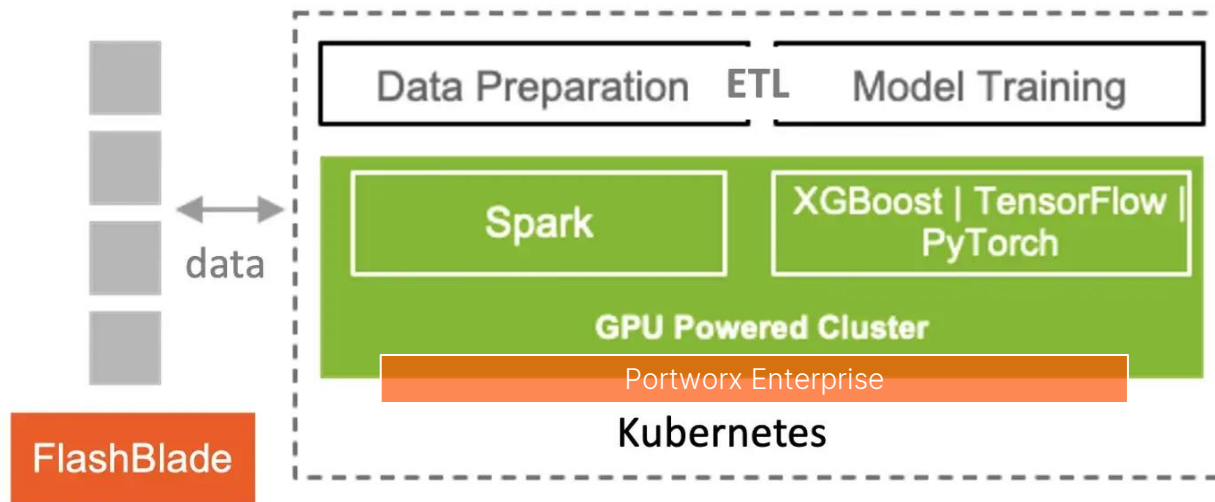
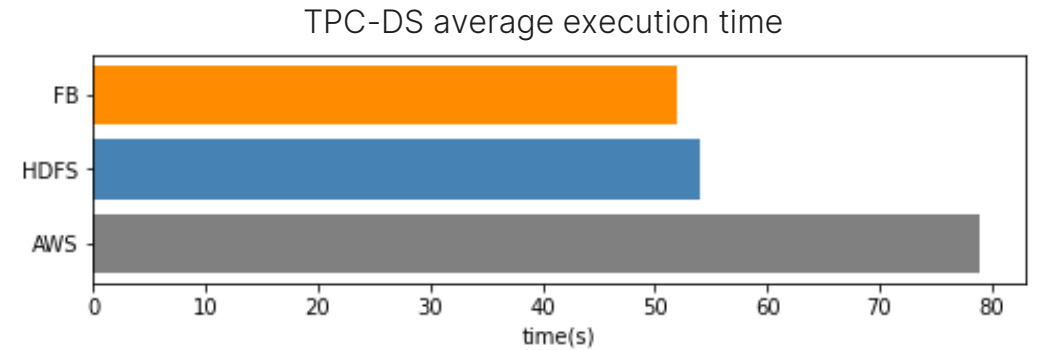
Accelerate big data and AI application with high performance

- 2 chassis (12 RU) for 18.4GBps S3 throughput. Even faster with NFS.
- 2 to 4ms latency most time
- High performance for big data & AI workloads. No tuning required.
- High performance for big & small files. No tuning required.

# Consolidating and accelerating data and AI infrastructure

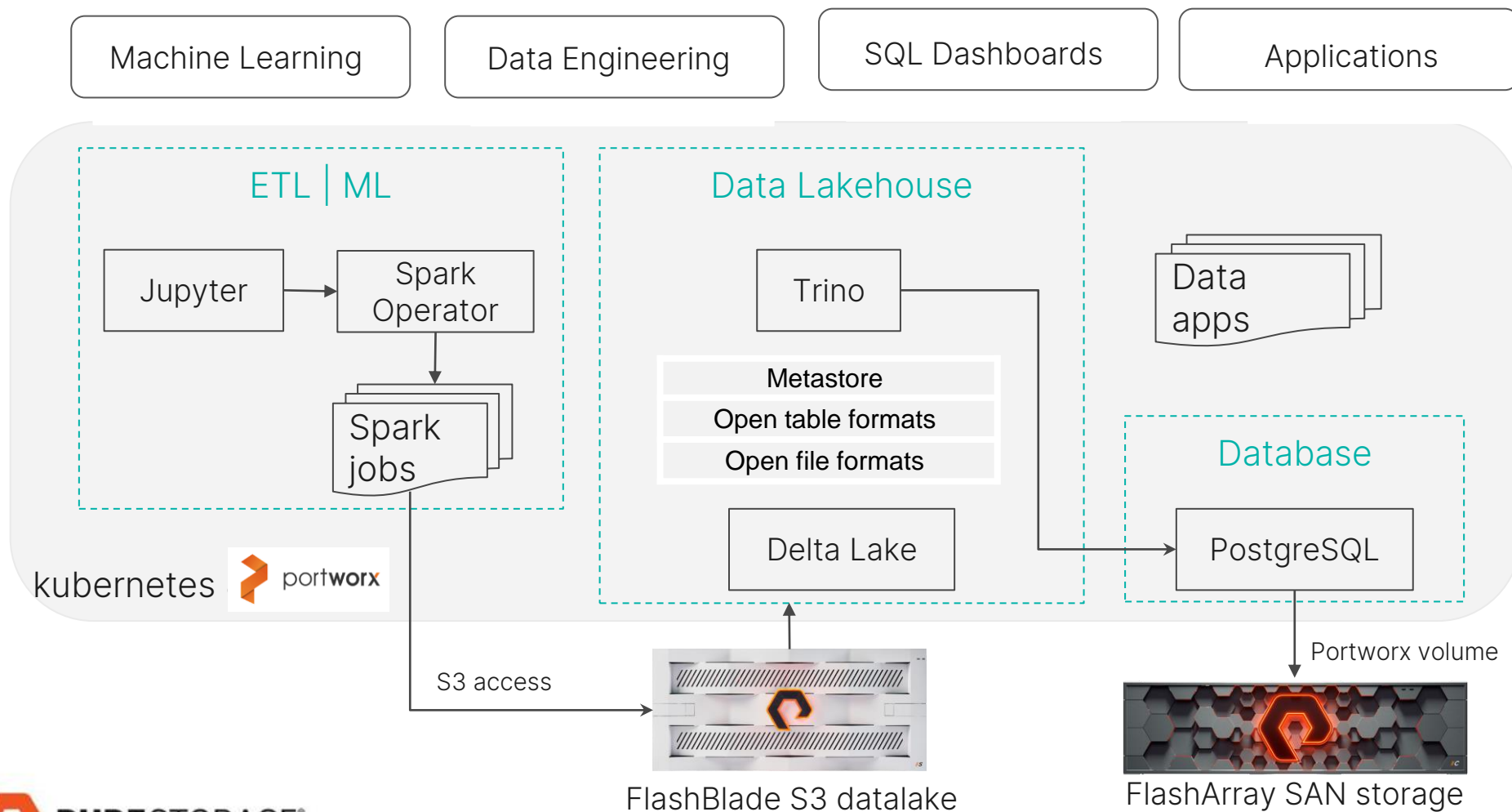
A single pipeline, from ingest to data preparation/ETL to model training and deployment.

Accelerate datalake and AI with fast file/object storage and GPU.



# Example: Open Data Lakehouse with Pure Storage

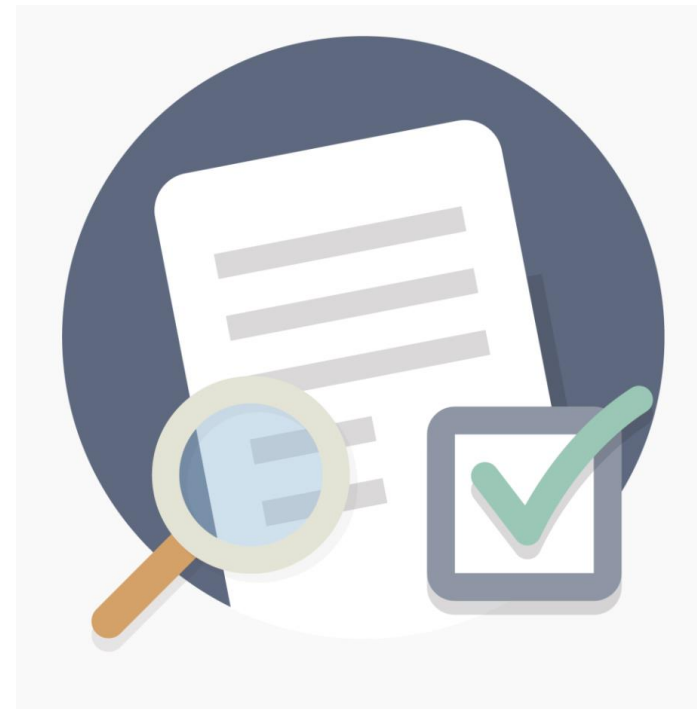
An example architecture for simple, fast and open data lakehouse with **Pure Storage**



- No lock-in
- Inexpensive
- Fast data exploration
- Simple
- Cloud ready



# Case Study and Summary



# Pure Storage for Meta AI Supercomputer

Meta AI

Research Publications People Events

## RESEARCH

### Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research

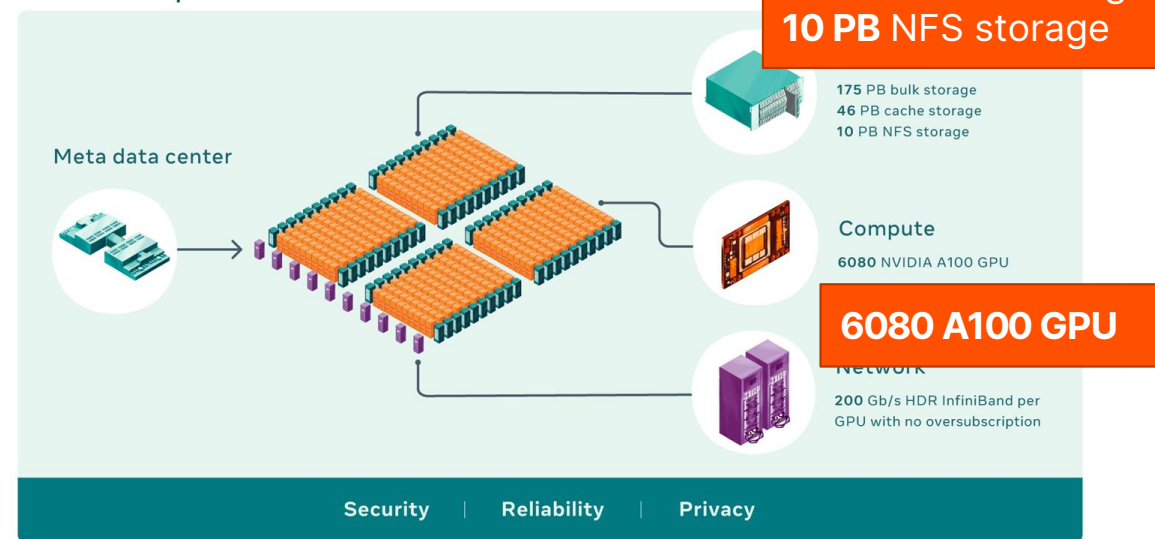
January 24, 2022



<https://ai.facebook.com/blog/ai-rsc/>



## AI Research SuperCluster Phase 1



## RESEARCH

### Facebook AI's Joelle Pineau receives Governor General's...

The award recognizes Canadian leaders for their groundbreaking innovations and positive impact on the quality of life in the country.

May 22, 2019

<https://www.youtube.com/watch?v=fZnykn1tDSE>

Pure Storage Tech Talk "Why Pure"

# Asia Cybersecurity Service Provider

## Challenges

- Legacy tech debt with Hadoop
- Expensive software license
- Needs of advanced analytics to quickly identify security threats

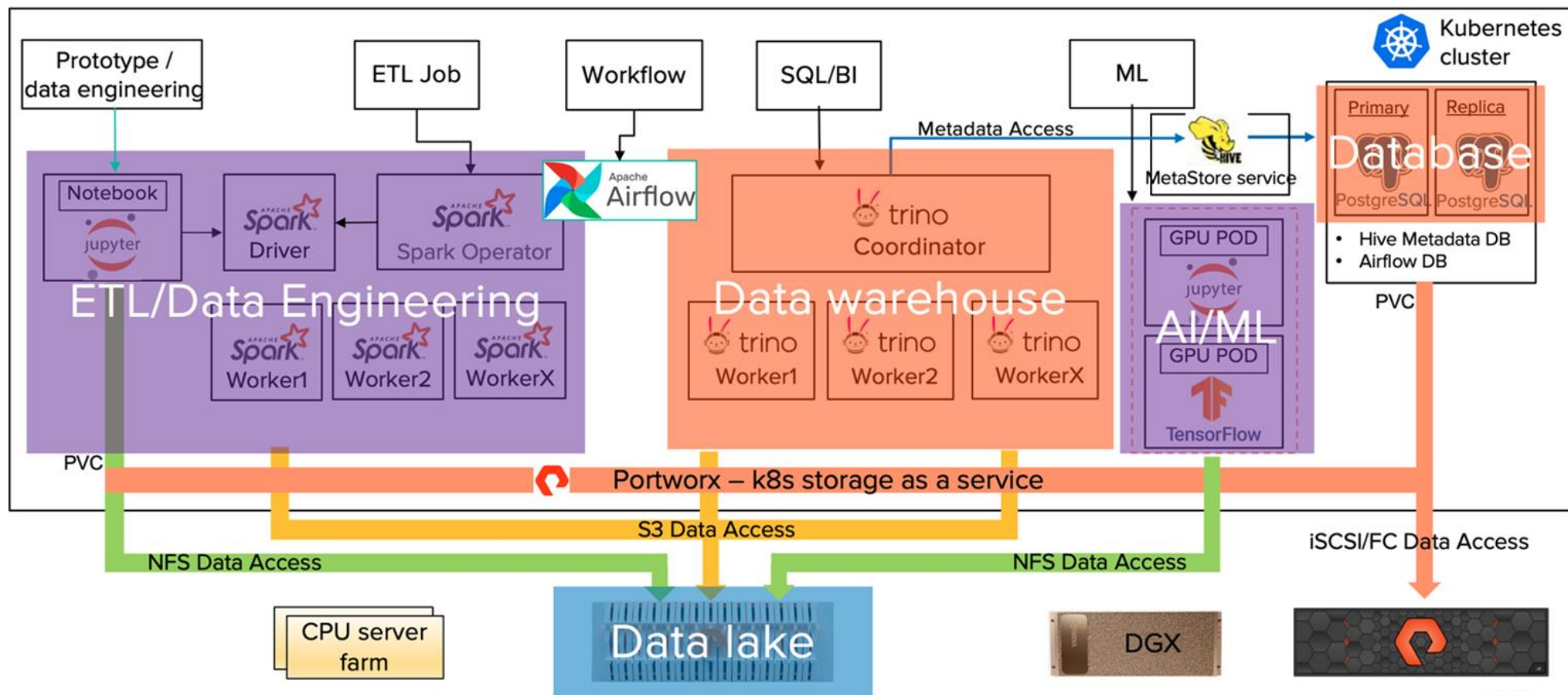
## Pure Storage's Solution

- A simple and fast data platform built with open-source software on FlashBlade, FlashArray and Portworx
- Big data and AI on Kubernetes, which is a key feature for the lean operation and apps team
- The consultative approach differentiated us from our competitors

## Customer Benefits

- A single pipeline, from ingest to data preparation/ETL/SQL to model training
- Low overall cost
- Big data without Hadoop complexity

# Asia Cybersecurity Service Provider



**Simple, fast, and open data lake, data warehouse, and ML  
with Pure Storage**

# Solution Review

Consolidate and accelerate data analytics and AI with Pure Storage

When you need simple and fast data platform

---

## FlashBlade for Data Analytics and AI

- Consolidate capacity and workloads for data analytics & AI
- Accelerate big data and AI with fast file & object storage
- Simplify deploy, operation and scale
- Seamless hardware and software upgrades and built-in compatibility with future technologies as part of Evergreen

## FlashArray **//X**



Business-critical applications  
Workload consolidation  
Virtualized environments

**Performance Optimized**

## FlashArray **//XL**



Most-demanding critical applications  
Extreme workload consolidation  
Large-scale virtualized environments

**Performance at Scale Optimized**

## FlashArray **//C**

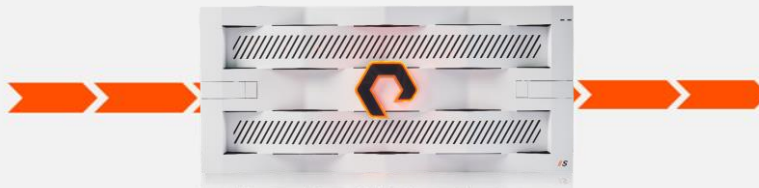


Less-demanding workloads  
Data protection and replication  
Large-scale content repositories

**Capacity Optimized**

**Optimize based on  
the needs of your  
business.**

## FlashBlade **//S**



**The Unified Fast File and  
Object Platform**

For high performance file  
and object workloads  
**Performance Optimized**

## FlashBlade **//E**



**The All-Flash Unstructured Data  
Repository**

For high capacity and price  
optimized file and object workloads  
**Capacity and Price Optimized**



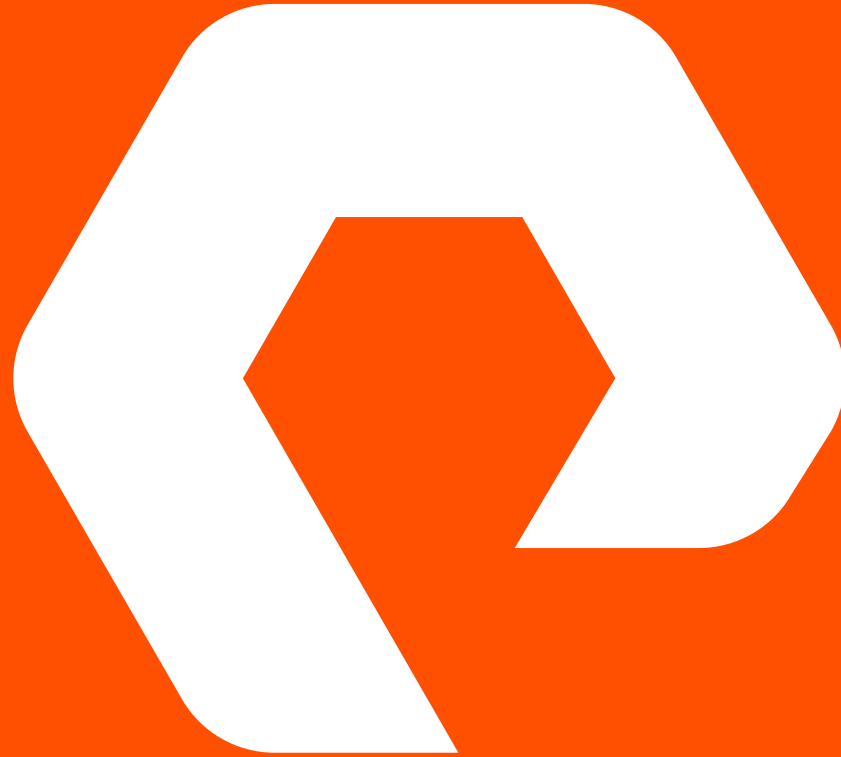
# Thank You!

홈페이지 : [www.purestorage.com/kr](http://www.purestorage.com/kr)

페이스북 : [www.facebook.com/purestoragekorea](http://www.facebook.com/purestoragekorea)

블로그 : [www.blog.naver.com/purestorage\\_korea](http://www.blog.naver.com/purestorage_korea)

유튜브 : [www.youtube.com/@PureStoragekr](http://www.youtube.com/@PureStoragekr)



Uncomplicate Data Storage, Forever