S22548

Unlocking Business Transformation with an AI Center of Excellence

NetApp, in partnership with NVIDIA

Santosh Rao, NetApp; Tony Paikeday, NVIDIA;

April 2, 2020

Agenda

1) AI Use Cases in 2020
2) Why AI Infrastructure Matters
3) Enabling AI with an Integrated Data Pipeline
4) How to Get Started

NetApp

AI Use Cases in 2020

MORE AI INDUSTRY APPLICATIONS GAIN TRACTION IN 2020


Robotics


Quality & Inspection

CONSTRUCTION


Military Applications

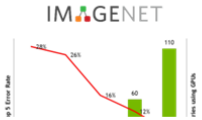

Research

GOVERNMENT


Intrusion Detection


Cyberwarfare

CYBERSECURITY



Cancer Research


Pharma R&D

HEALTHCARE


Financial Forecasting


Fraud Detection

FINANCE


Autonomous Cars


Vehicle Safety

AUTOMOTIVE

Online Advertising


Recommendation Engines

MARKETING


Call Center


Sentiment Analysis

CUSTOMER SERVICE


Shopping Assistance


Inventory Analytics

RETAIL


Crop Management


Disease Identification

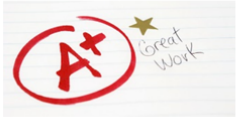
AGRICULTURE


Security


Energy Efficiency

SMART CITIES

Cognitive Tutoring



Grading

EDUCATION

CREATING SMARTER, SAFER CITIES

AnyVision builds safer cities with high-speed, real-time recognition from surveillance video streams and the ability to detect 115M individuals in 0.2 seconds.

SMARTER INSPECTION SERVICES

Avitas Systems predicts its solution will decrease annual inspection costs by 25%, reduce maintenance downtime by 15%, and increase inspection turnaround time by 25%.

The robots can handle the heat and use infrared cameras and chemical and other sensing technologies to inspect assets under dangerous conditions and keep production running.

Image credit: Avitas Systems

OPTIMIZED LOGISTICS MANAGEMENT

To analyze and optimize routes of 200,000+ mail carriers in real-time the USPS used GPU-accelerated data analytics to optimize routes—using fewer trucks, handling more deliveries, and narrowing delivery windows.

Why AI Infrastructure Matters

DIFFERENT ROLES, SAME GOALS

Everyone wants the best AI tools—nobody wants to design/build it

Data Scientists

Data Engineers

IT Management

PLAN / CODE / BUILD / TEST / DEPLOY / OPERATE / MONITOR

What if you could iterate on models much faster than today?

What if you could automate reproducibility?

What if your DL/ML projects could get deployed 3 months faster?

THE CASE FOR AN (IT-LED) AI INFRASTRUCTURE

The Benefits of AI Centers of Excellence (CoE)

PEOPLE

1. Build communities of practice
2. Centralize AI talent pipeline

PROCESS

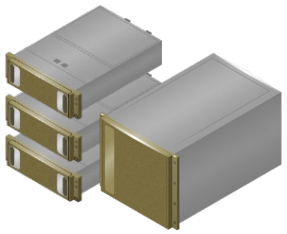
3. Consolidate experience in going from (AI ideas) → (PoCs) → (Production)
4. Reduce CapEx and accelerate ROI through infrastructure centralization

TECHNOLOGY

5. Democratize AI across the company, accessible to every product, service and supply chain

DO I BUILD IT ON PREM OR IN THE CLOUD?

Understanding how each supports development workflow



CLOUD

- Early exploration
- Limited access to capital/budget
- Modest datasets already native/local to cloud provider
- Fewer experiments / slower pace of experimentation

ON-PREM

- “Deep learning enterprise”
- Requires “GPU-ready” data center
- Large datasets local to on-premises
- Frequent experiments (often in parallel), rapid pace

FACTORS TO WEIGH

Data Gravity, sovereignty and security

Maintaining lowest cost per training run

Ensuring ability to fail fast, learn faster

THE 3RD OPTION YOU NEED TO CONSIDER

Bringing AI training closer to the nexus of clouds and data

IF YOU:

- Don't have an **AI-ready** data center
- Are **budget-challenged** in updating yours
- Need an **affordable OpEx** model for AI-optimized facilities

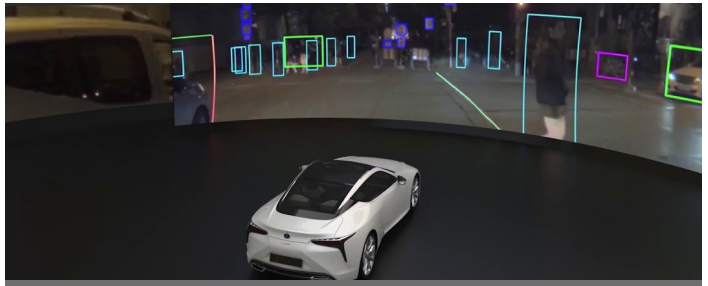
Colocation services for AI infrastructure

BENEFITS:

- Already optimized for AI-infrastructure
- Many already have their data sets residing at colo
- Faster deployment, less CapEx
- Low-latency, high BW, direct-connect to major clouds
- Cloud-like ease, performance of on-prem

“A-HA” MOMENTS IN AI INFRASTRUCTURE

ONTAP AI: design insights gained from deep learning data centers



Example:

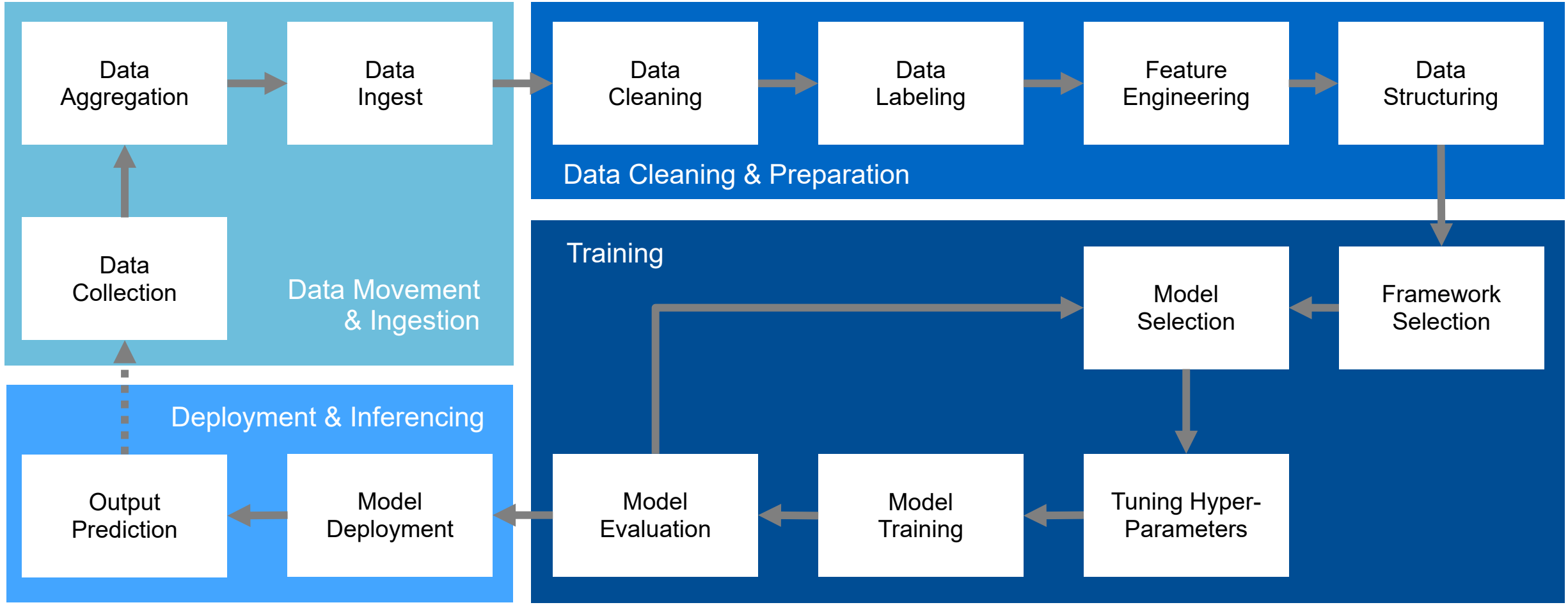
- Autonomous vehicle = 1TB/hr
- Training sets up to 500 PB
- RN50: 113 days to train
- Objective: 7 days
- 6 simultaneous developers

= 97 node cluster

Rack Design	Networking	Storage	Facilities	Software
• DL drives close to operational limits • Similarities to HPC best practices	•100G EDR or 100GbE preferred • High-bandwidth, ultra-low latency	• Datasets range from 10k’s to millions objects • Terabyte levels of storage and up	• Assume higher watts per-rack • Higher FLOPS/watt = DC less floorspace required	• Scale requires “cluster-aware” software

What Is an AI Pipeline?

Four phases for AI and ML pipelines

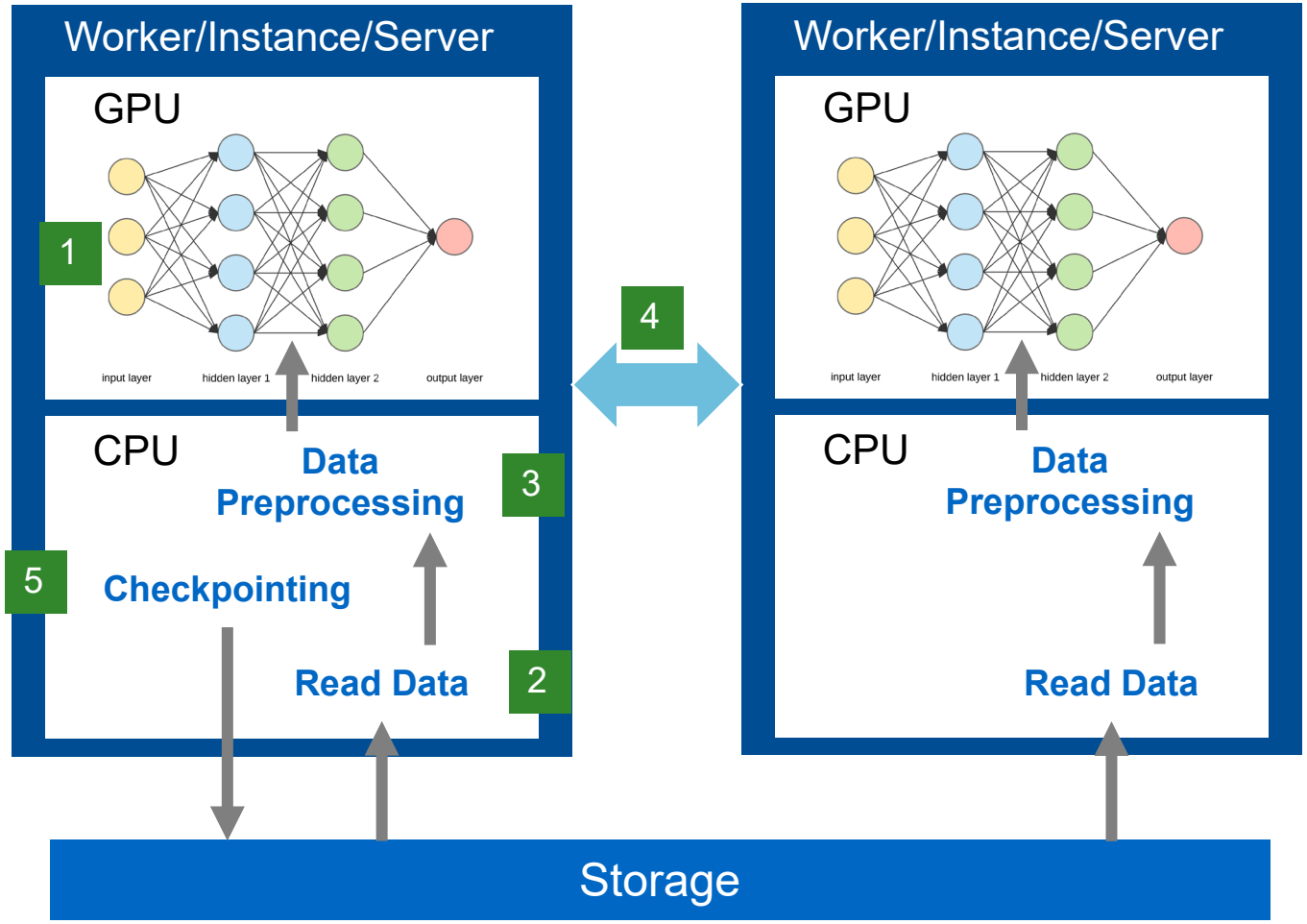


NetApp

Training is the dominant phase in the pipeline

A closer look at the training phase

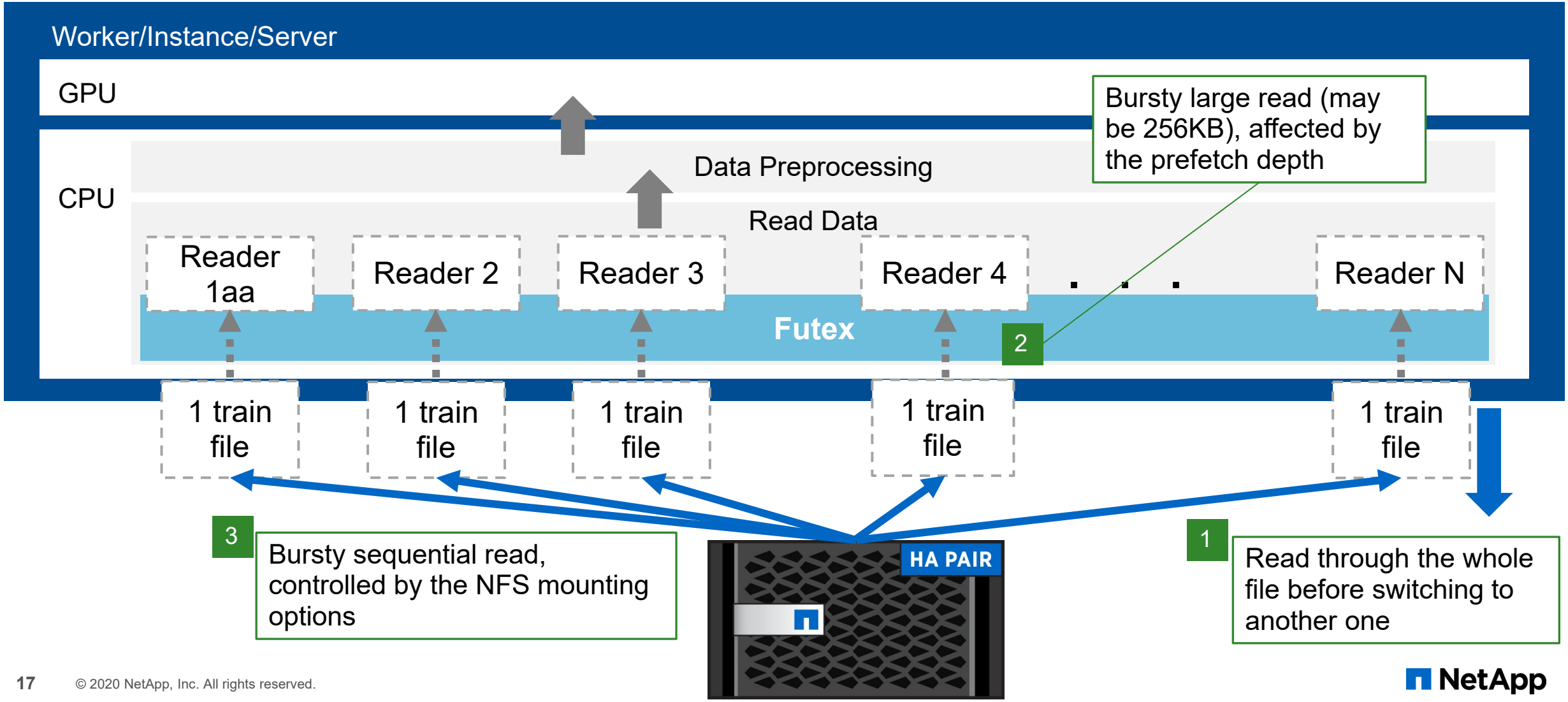
1. Select a model and a combination of Hyper-parameters
2. Read a batch of examples from the storage
3. Process the example before feeding into GPU
4. Sync the learned model after processing a batch of examples
5. Save the latest learned model for a certain period of time
6. Repeat from step 2 until accurate



NetApp

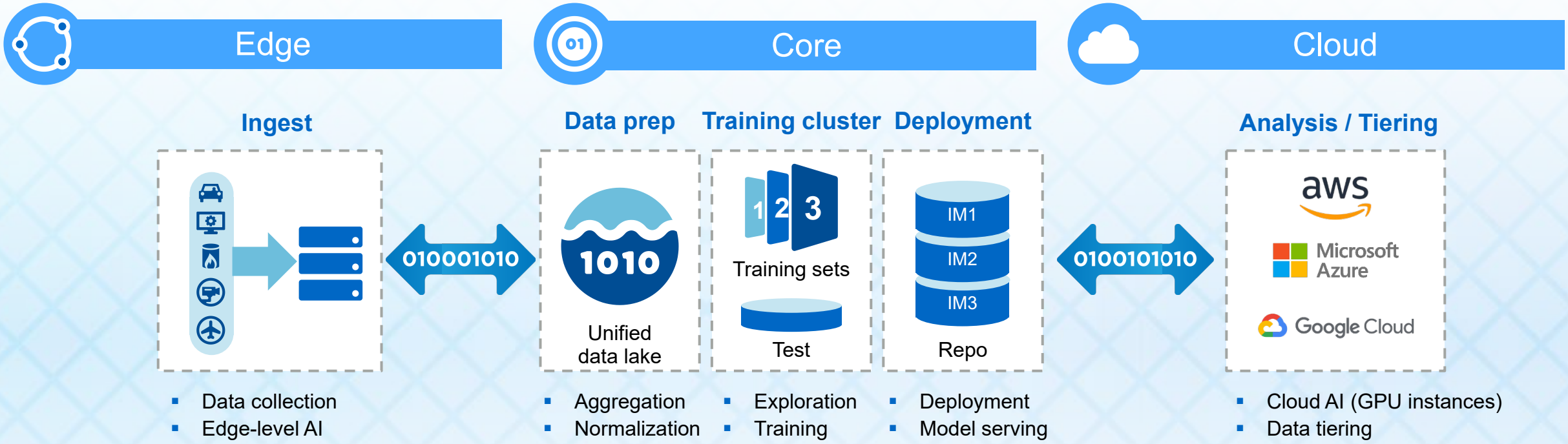
Data reading from the storage

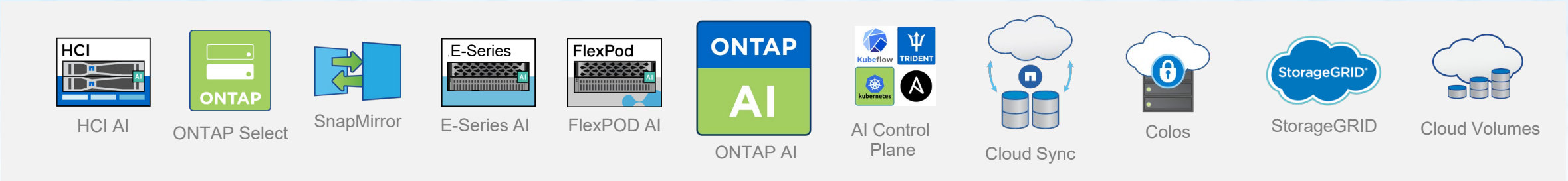
Using Tensorflow as an example



NetApp

Edge to core to cloud data pipeline for AI





NetApp

Automate Your AI Data Pipeline

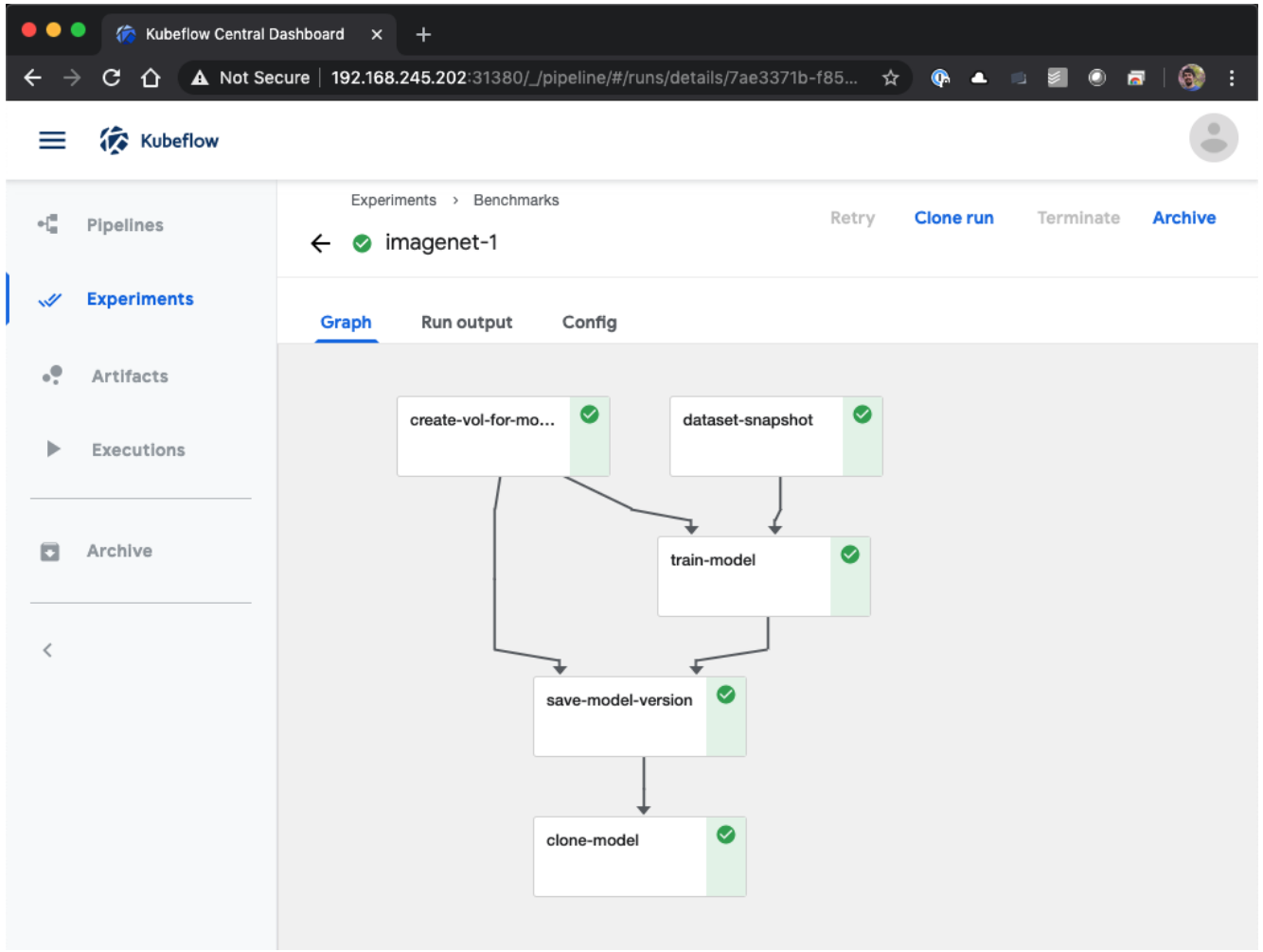
Control Plane Solutions

NetApp

Automated data prep, training, Dev/Test, and deployment

Automate your AI data pipeline

- Kubeflow Pipelines: Platform/standard for defining and deploying portable and scalable AI/ML workflows
  - Python SDK–familiar and comfortable for data scientists
- Example pipeline steps
  1. Create new volume for storage of model
  2. Create NetApp® Snapshot™ copy of existing dataset volume (for traceability)
  3. Execute containerized AI/ML training job
  4. Create NetApp® Snapshot™ copy of model volume (versioned model baseline)
  5. Create clone of model volume for testing

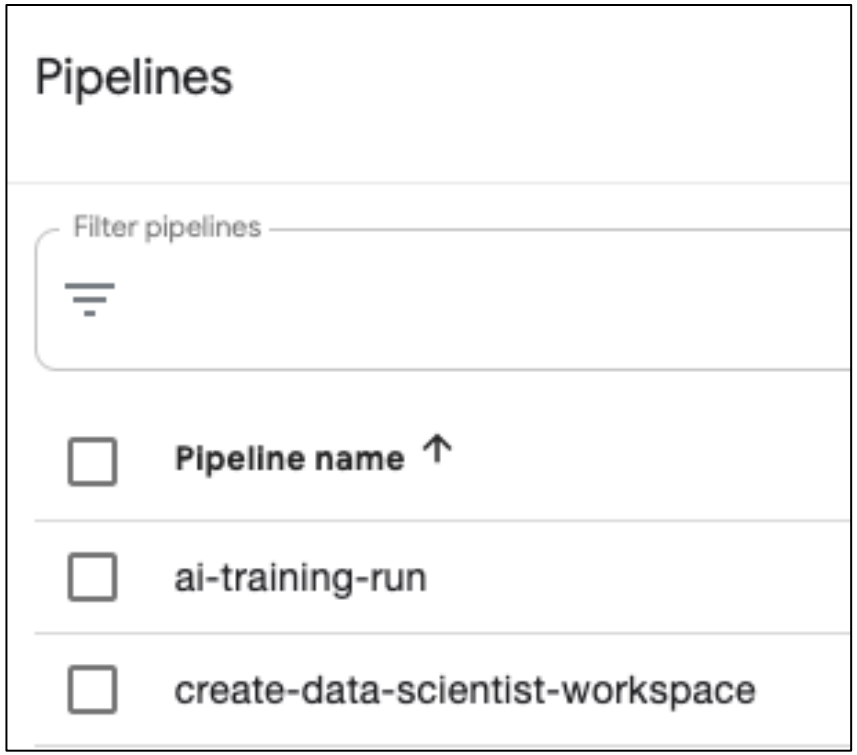


NetApp

Example pipelines

Plug and play workflow automation

- AI training run with automatic traceability and versioning
  - Plug in data prep, training, and validation commands, then press play!
- Create an exact copy of production data for a development workspace
  - Create near-instantaneous, space-efficient copy of production dataset(s)
  - Experiment without fear of “messing up” production

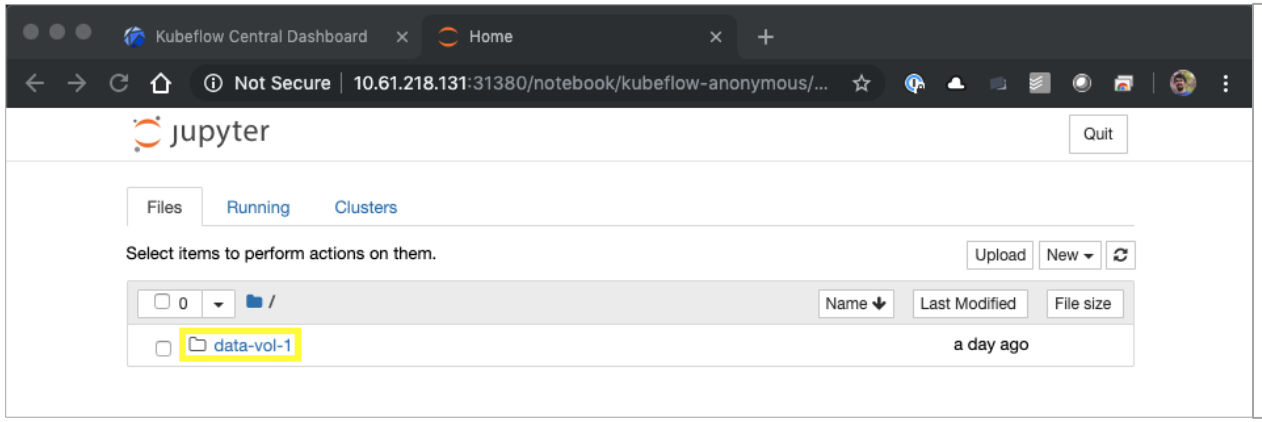


NetApp

On-demand data scientist/developer workspaces

Instant data accessibility; facilitating rapid experimentation

- Jupyter Notebooks
  - Wiki-like documents that contain live code and descriptive text
  - Widely used in AI/ML community as means of document, storing, and sharing projects
- NetApp AI Control Plane implementation
  - Petabytes of data accessible from within familiar interface
  - Protect production data while still making it accessible to Data Scientists for experimentation
  - Trigger NetApp® Snapshot™ copy creation from within notebook for dataset/model versioning/baselining



Create snapshot

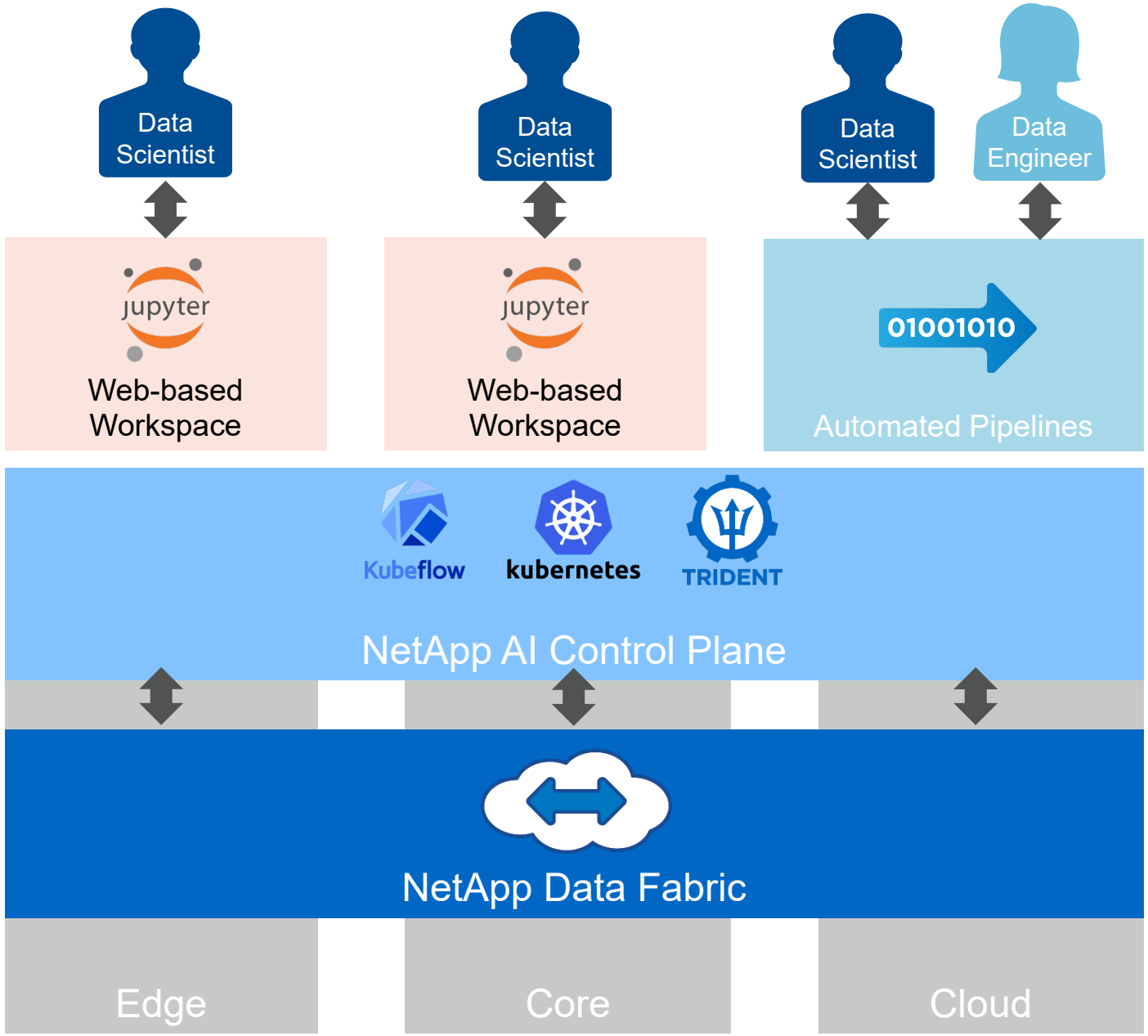
In [5]:
```
## Enter the name of SVM that Trident is configured to use
##   (i.e. the SVM that you specified when creating your Trident backend)
tridentSVM = 'trident'
##

volume = Volume.find(name = volumeName)
timestamp = datetime.today().strftime("%Y%m%d_%H%M%S")
snapshot = Snapshot.from_dict({
    'name': 'jupyter_%s' % timestamp,
    'comment': 'Snapshot created from within a Jupyter Notebook',
    'volume': volume.to_dict()
})
response = snapshot.post()
print("API Response:")
print(response.http_response.text)
```


NetApp

The bigger picture

Full-stack AI data and workload management

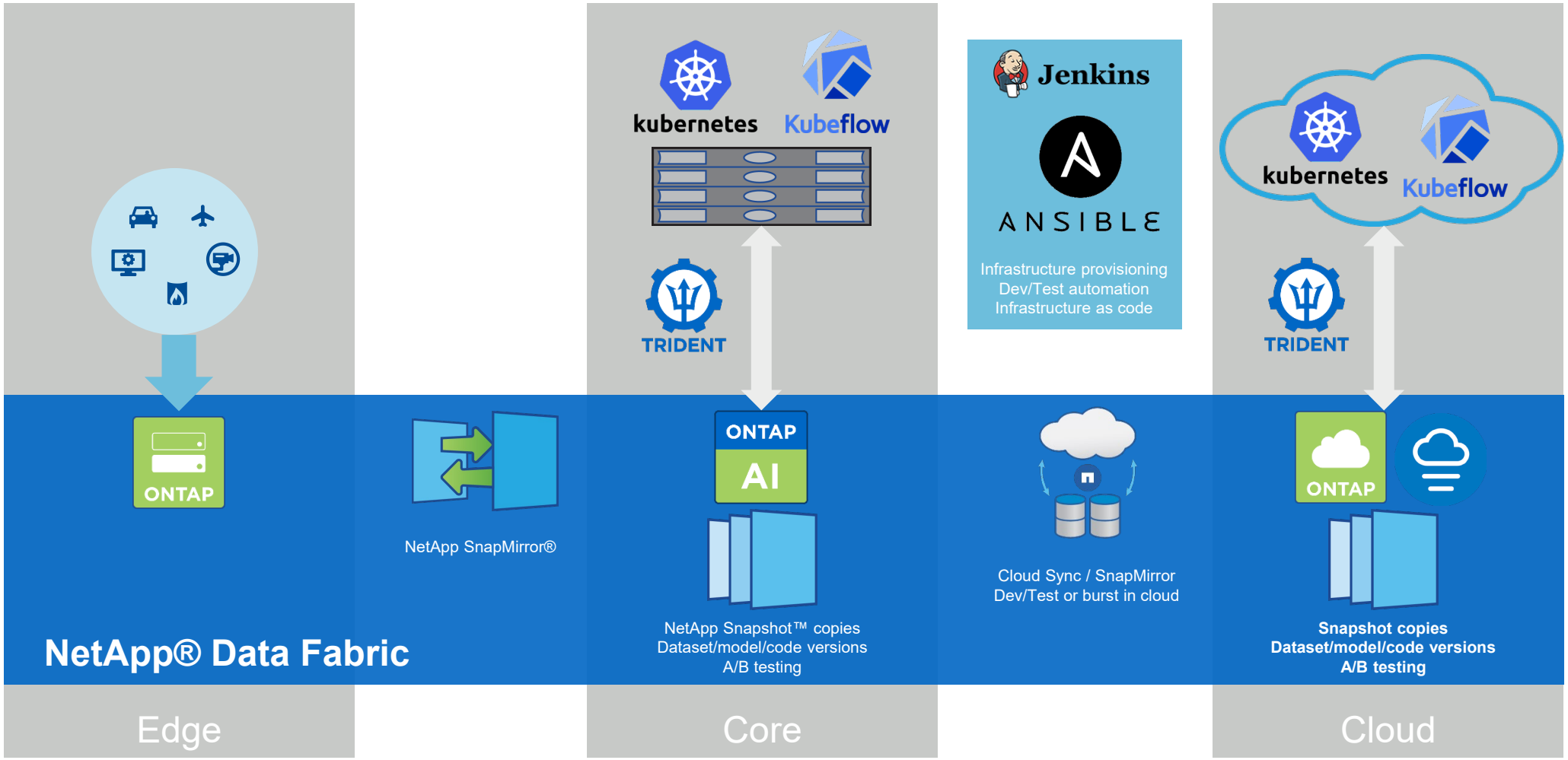
- On-demand Jupyter Workspaces
  - Full access to production datasets
- Automate data prep, training, and deployment workflows
- Workspaces and workloads can span edge, core, and cloud
  - Choice of any compute and/or cloud
  - Cross-site Data Scientist collaboration
- Built-in versioning
  - Full dataset to model traceability
  - Seamlessly switch between model versions for dev/test, A/B testing, etc.



NetApp

Flexible architecture

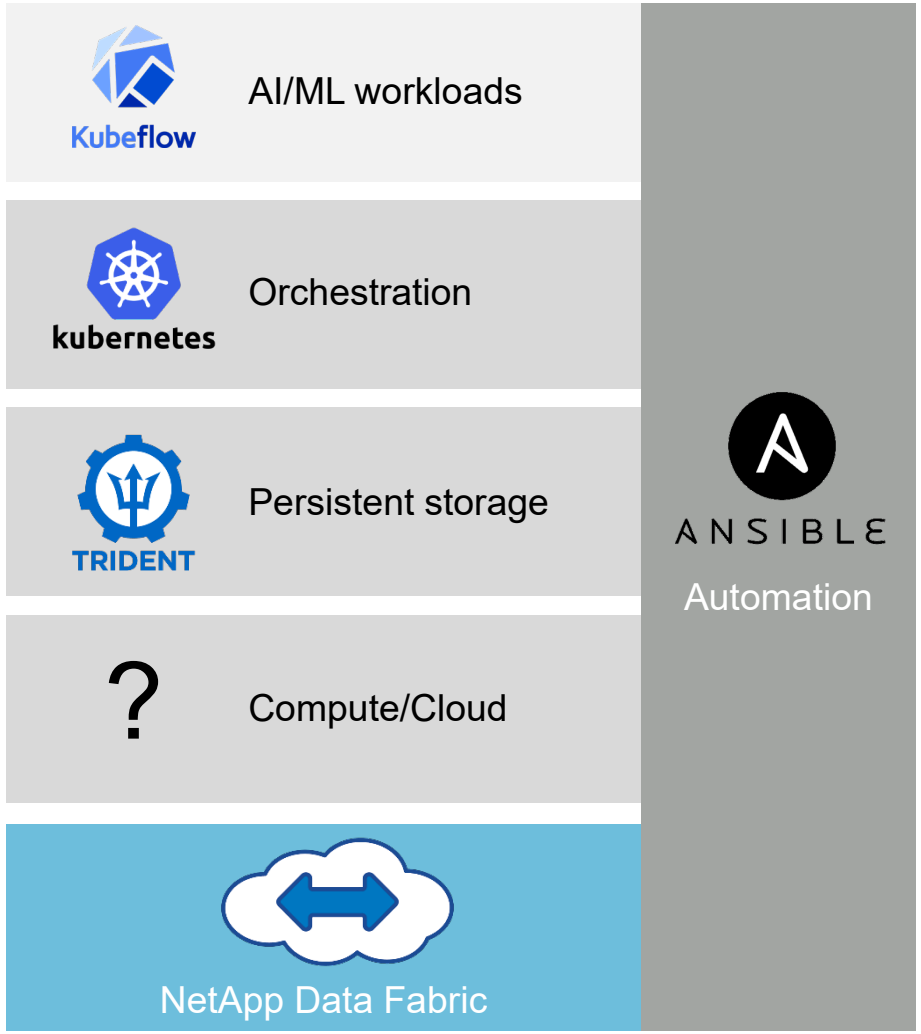
Data and workloads are available whenever and wherever they are needed



Fully open stack

Enabling portability, scalability, automation, and simplicity

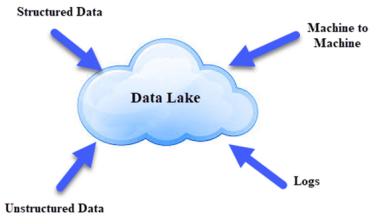
- Kubeflow = an AI/ML toolkit for k8s
  - Standard open-source platform for deploying AI/ML workloads
- Kubernetes (k8s) = container orchestration
  - Industry-standard, open-source container platform
- NetApp® Trident = storage provisioner for k8s
  - Enterprise-class storage presented in Kubernetes-native format
- Choice of compute and/or cloud
- NetApp Data Fabric = data portability and protection
  - Edge to core to cloud data movement
- Ansible = deployment automation, infrastructure as code



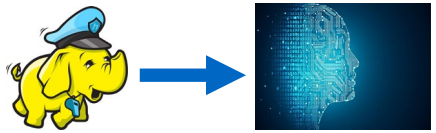
NetApp

Data Mover Solutions

Customer requirements and challenges

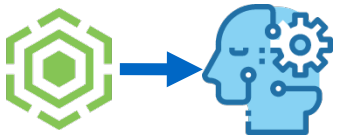


Data in data lake



Data lake data into AI

Data sync between HDFS/MapRFS and NFS

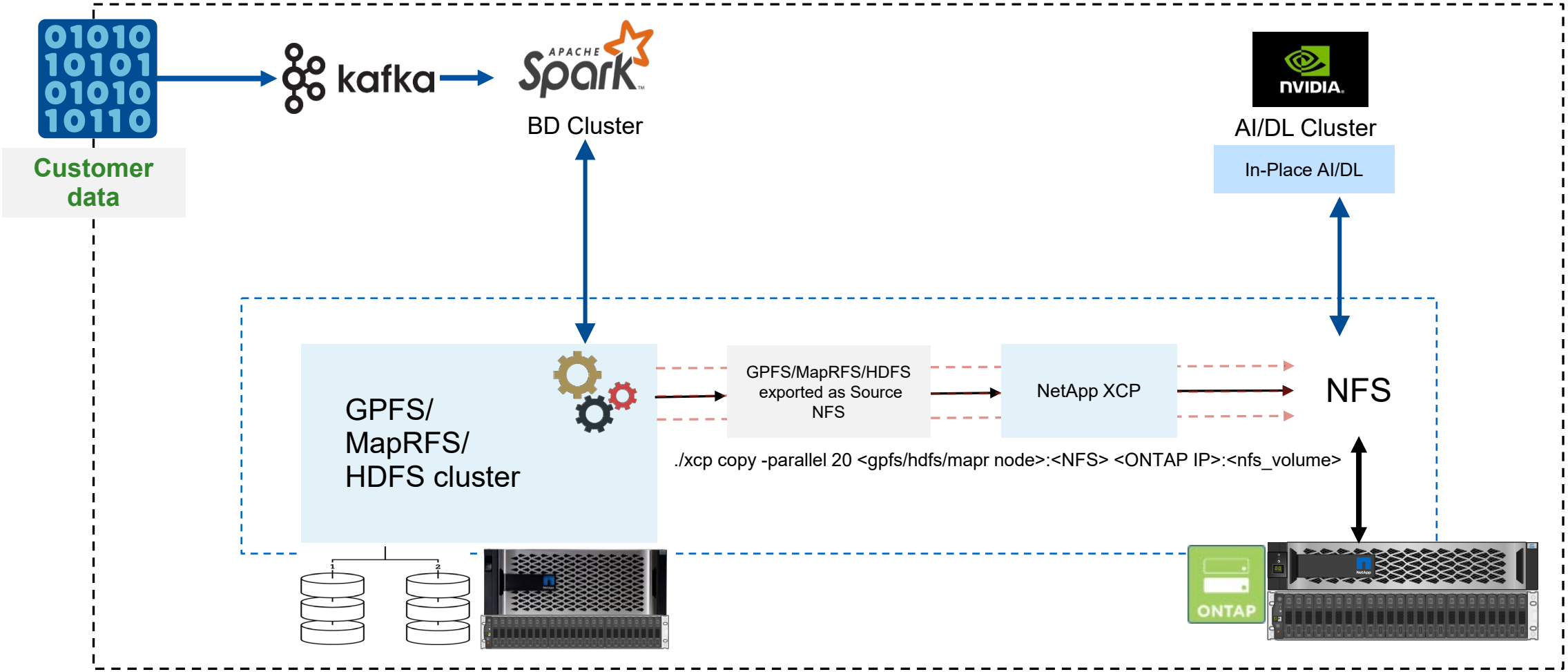


HPC (GPFS) data into AI

NetApp

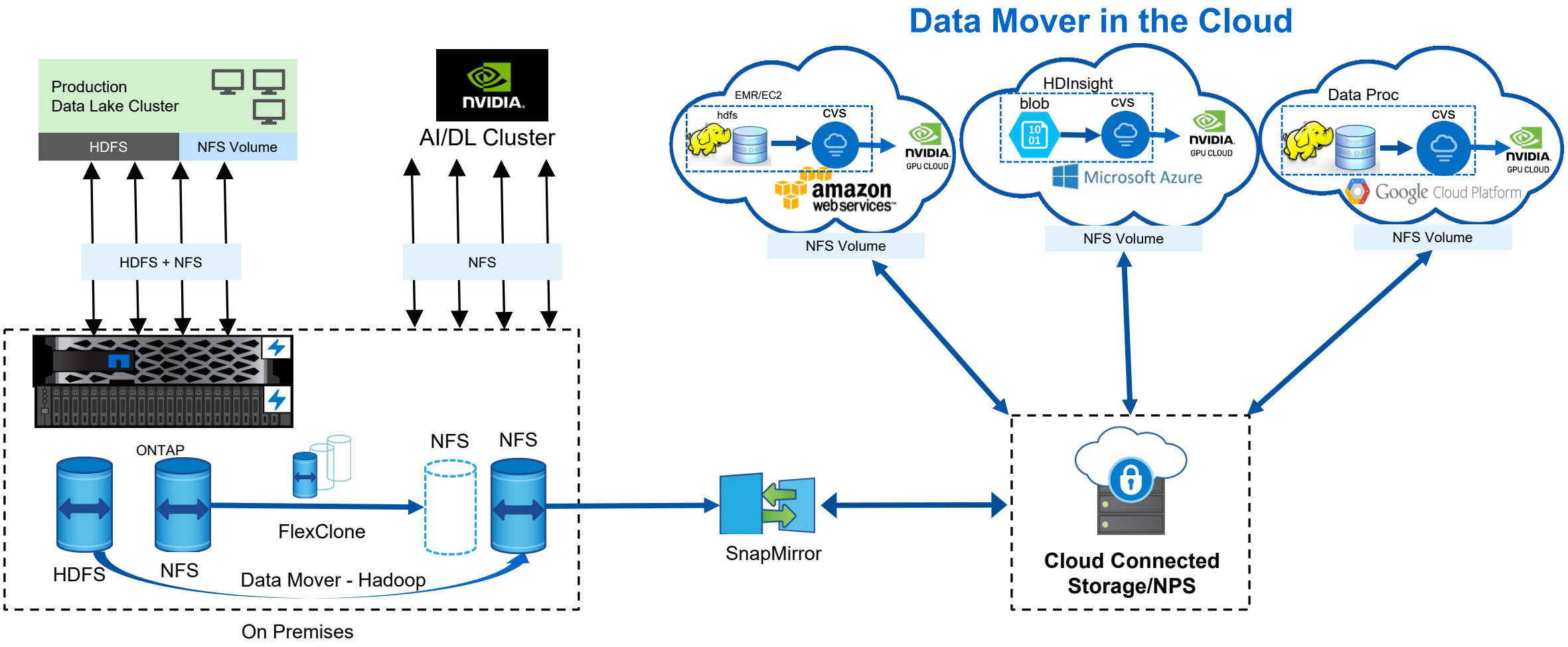
NetApp Data Mover Solution for AI – XCP

Extends from edge to core to cloud; federates data sources and GPUs for AI processing



NetApp

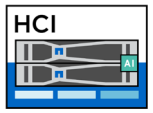
Data Mover Solution



NetApp

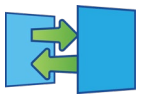
Getting Started

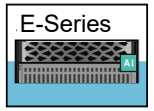
Chart your path to success in 2020



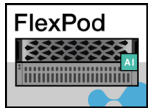
HCI AI

ONTAP Select


SnapMirror



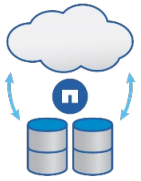
E-Series AI



FlexPOD AI

ONTAP AI

AI Control Plane


Cloud Sync


Colos

StorageGRID


Cloud Volumes

NetApp

NetApp ONTAP AI

Simplify, accelerate, and integrate your data pipeline for deep learning with NetApp, NVIDIA and Mellanox

- **Proven architecture for DL**
  - Powered by NVIDIA DGX-1 and DGX-2 systems and NetApp cloud-connected all-flash storage with Mellanox fabric
- **Simple to deploy**
  - Eliminate design complexity and guesswork
  - Speed innovation and experimentation
- **Deliver performance and scalability**
  - Accelerate results
  - Start small and grow non-disruptively
- **Build an integrated data pipeline**
  - Intelligently manage your data from edge to core to cloud
  - Backed by AI expertise and single point of contact support

NetApp

Don’t let infrastructure delay your ROI for AI

For organizations who:

- **Want to deploy AI infrastructure** built on ONTAP AI
- **Lack the CapEx budget** to build a data center for GPU-computing
- **Need access to AI-ready data center** facilities **-** NOW
- **Need an affordable OpEx model** for hosting their ONTAP AI investment

No data center?
No problem.

ONTAP AI-Ready Data Center Program Partners

Aligned Energy
colovore
CORE SCIENTIFIC
Cyxtera
CyrusOne
data dock
DIGITAL REALTY
edgeconnex
EQUINIX
FLEXENTIAL
FUJITSU
IDCFrontier
interxion
NTT Communications
Transform. Transcend.
QTS
ScaleMatrix
switch
VANTAGE DATA CENTERS
VERNE GLOBAL
LEADERS IN HIGH PERFORMANCE COMPUTING

ONTAP AI-Ready Data Center

A suite of AI infrastructure solutions from partners

- Get a world-class AI-ready data center now
- De-risk deployment with a Test Drive
- Leading-edge infrastructure without CapEx

No data center? No problem!

ONTAP AI Hosting

Co-location services for customer-owned ONTAP AI infrastructure

Try it now! Buy it if you love it!

ONTAP AI Test-Drive

Kick-the-tires using colo-provided ONTAP AI infrastructure

Cloud-like ease + on-prem performance

ONTAP AI as-a-Service

Rent ONTAP AI infrastructure from colo provider

Learn more, contact: testdriveprogram@nvidia.com

NetApp

Get started on your AI journey

- Schedule a time to talk to our AI experts
- Work with us to identify and prioritize use cases
- Participate in a tailored workshop
- Visit our Briefing Centers
- Sign up for a Customer Proof of Concept (CPoC)
- Learn more at www.netapp.com/ai

NetApp

NetApp

Thank You