AI & Analytics 트랜드

“Open Analytics as-a-Service for your Spark and Data Warehouse”

Yifeng Jiang

APJ Principal Solutions Architect, Data Science, Pure Storage

PURESTORAGE®

PURE//ACCELERATE® techfest 22 seoul

Session Agenda

- Trend in data lake and data warehouse
- Spark, data lake and warehouse with Pure
- Demo: Simplifying data prototype and data engineering
- Application-aware backup and DR for big data

Spark, data lake and warehouse

- Every big data user has a data lake, Spark and data warehouse use case
  - Data lake with Hadoop (HDFS and MapReduce)
  - Fast processing with Spark
  - Everyone likes SQL, business intelligence and data warehouse
- Embracing as-a-service model and architecture
  - Cloud-native, hybrid cloud or even on-premise
  - Separating compute and storage

Photo by Paul Skorupskas on Unsplash

Trend in data lake and warehouse

S3: the true open data lake

- Simple to use and operate
- Highly available and scalable
- Open standard for many use cases

Pluggable data warehouse

- Same data, different engine, no data copy
- Open table format & storage API
- Choice of Trino, Vertica and many data warehouses

Kubernetes: key for as-a-service architecture

- Share and isolate cluster resource efficiently
- Streamline operation with standard API
- Scale up and down with one command

Latest and greatest analytics software

- Avoid vendor lock in, best analytics tool, open source or commercial
- Fast innovation in open community
- Fast unified data lake, open architecture

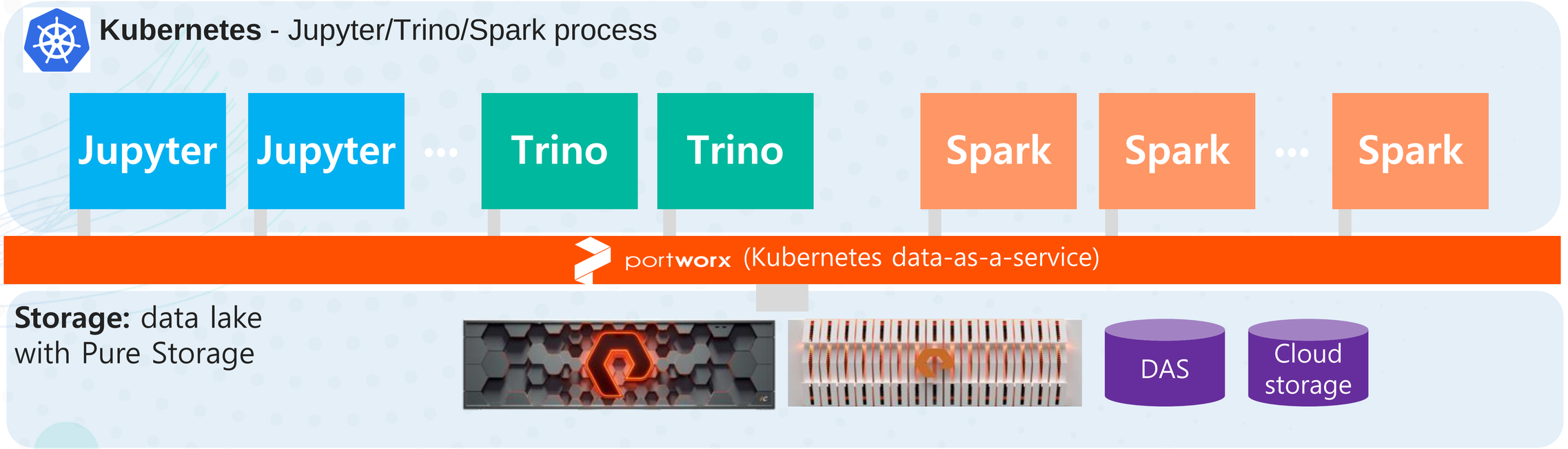
Spark, Data Lake and Warehouse with Pure Storage

Simple, reliable and low-cost data analytics at scale

PURESTORAGE®

PURE//ACCELERATE® techfest22 seoul

Data lake and warehouse with Pure



- Scale storage and compute independently
- Analytics anywhere, bare metals, VMs and containers on fast unified data lake
- Deliver and operate like a service

- Pathway to hybrid cloud
- Storage as a service to Kubernetes applications
- Application-aware backup, DR

Analyze 1PB data in HDFS vs. FlashBlade®

	HDFS with DAS	FlashBlade S3
Storage Required	3PB	0.5-1.3PB*
Number of Data Node	60 (3+ racks)	0
Number of Name Node	3	0
Number of Compute Node	60 (co-exist /w DN, 2RU physical server)	Much less (sizing by compute, 1RU physical server, VM or container)
Rack Space	63 (up to 4 racks)	1~2 rack most likely
DC & Operation Cost	$$$	$

Benefits

- Reduce TCO.
- Flexible expansion – scale compute & storage independently
- Easy to deploy, operate and upgrade

Assumptions

- 1PB raw data
- HDFS 3x replication
- 50TB per Data Node /w DAS

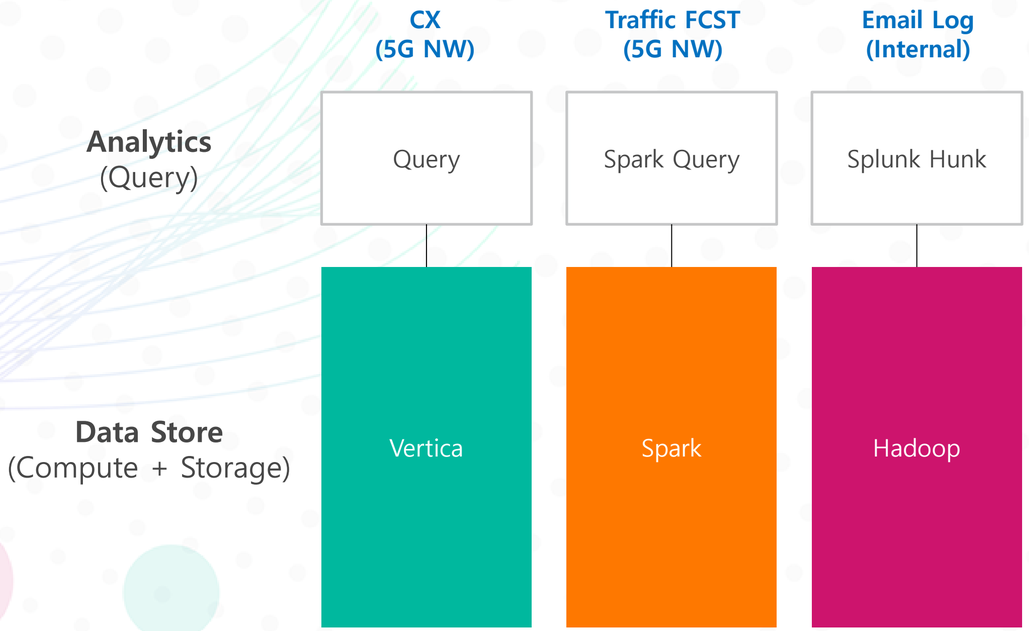
PURESTORAGE®

* Depending on data compression ratio

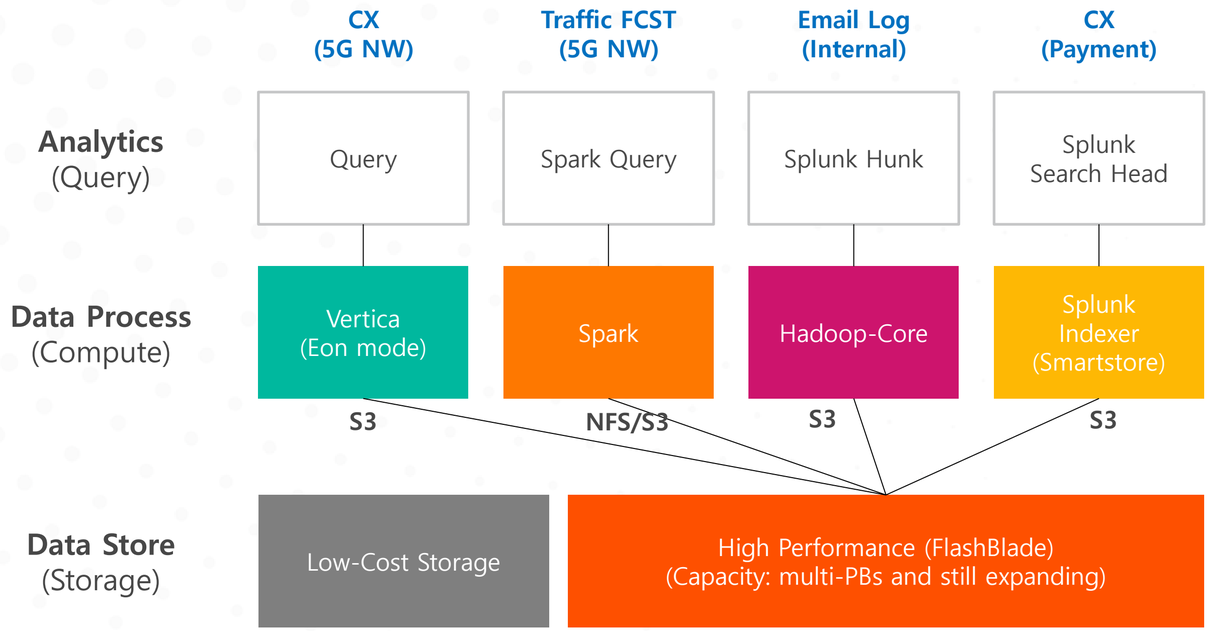
PURE//ACCELERATE® techfest22 seoul

Case study: unified data lake at large telecom

Before



After



CHALLENGES

- Performance and scalability reached its limits due to DAS architecture
- Struggled with complex operation and failure handling
- High performance and scalable storage was required as data grows

BENEFITS

- PB scale, high performance system achieved by separating compute & storage
- Simple and effortless operation with FlashBlade and Pure1
- Consolidate multiple analytics applications to single platform. Avoid silos.
- Future-proof unified data lake with Evergreen program

PURESTORAGE®

PURE//ACCELERATE® techfest22 seoul

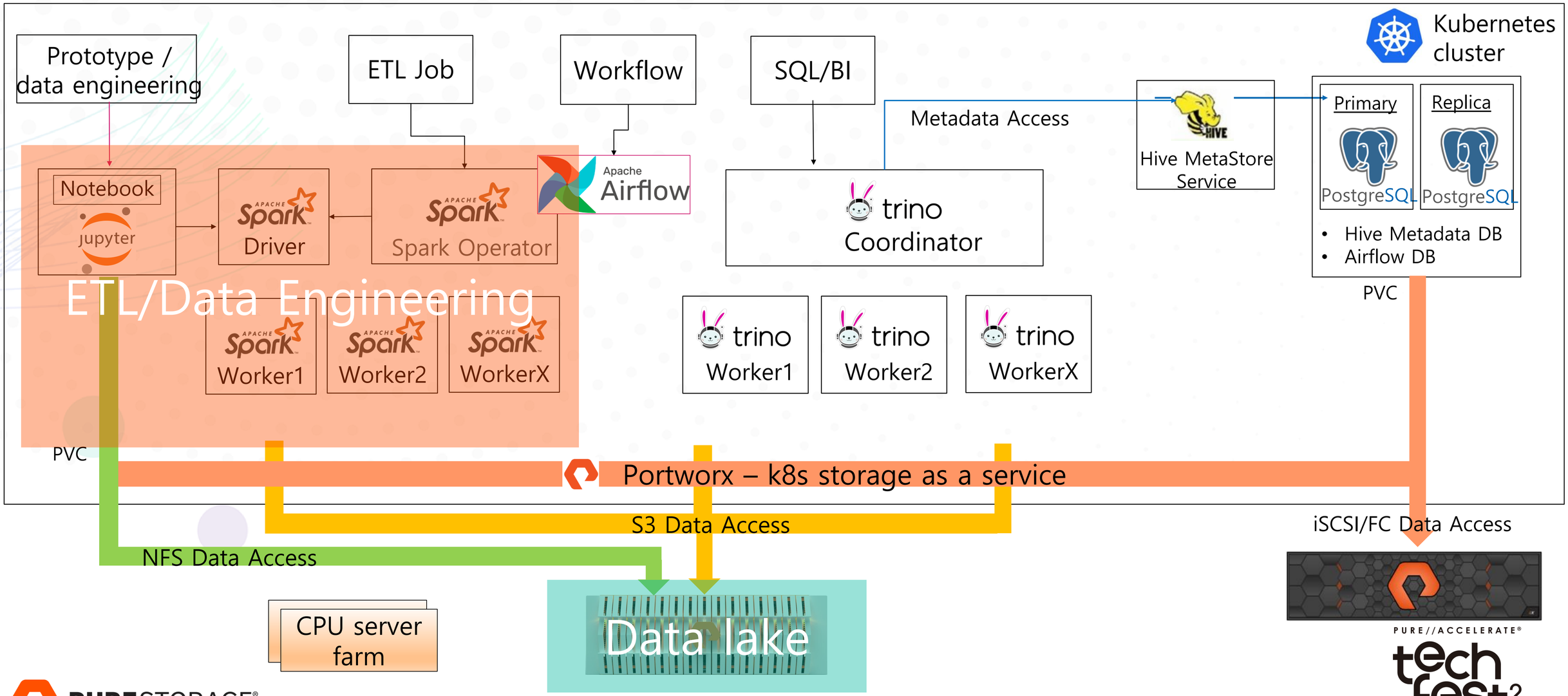
Architecture and Demonstration

PURESTORAGE®

PURE//ACCELERATE® techfest22 seoul

Simple & scalable Spark and data warehouse with Pure

Jupyter, Spark, and Trino on Kubernetes with Pure for prototype, data engineering and workflow



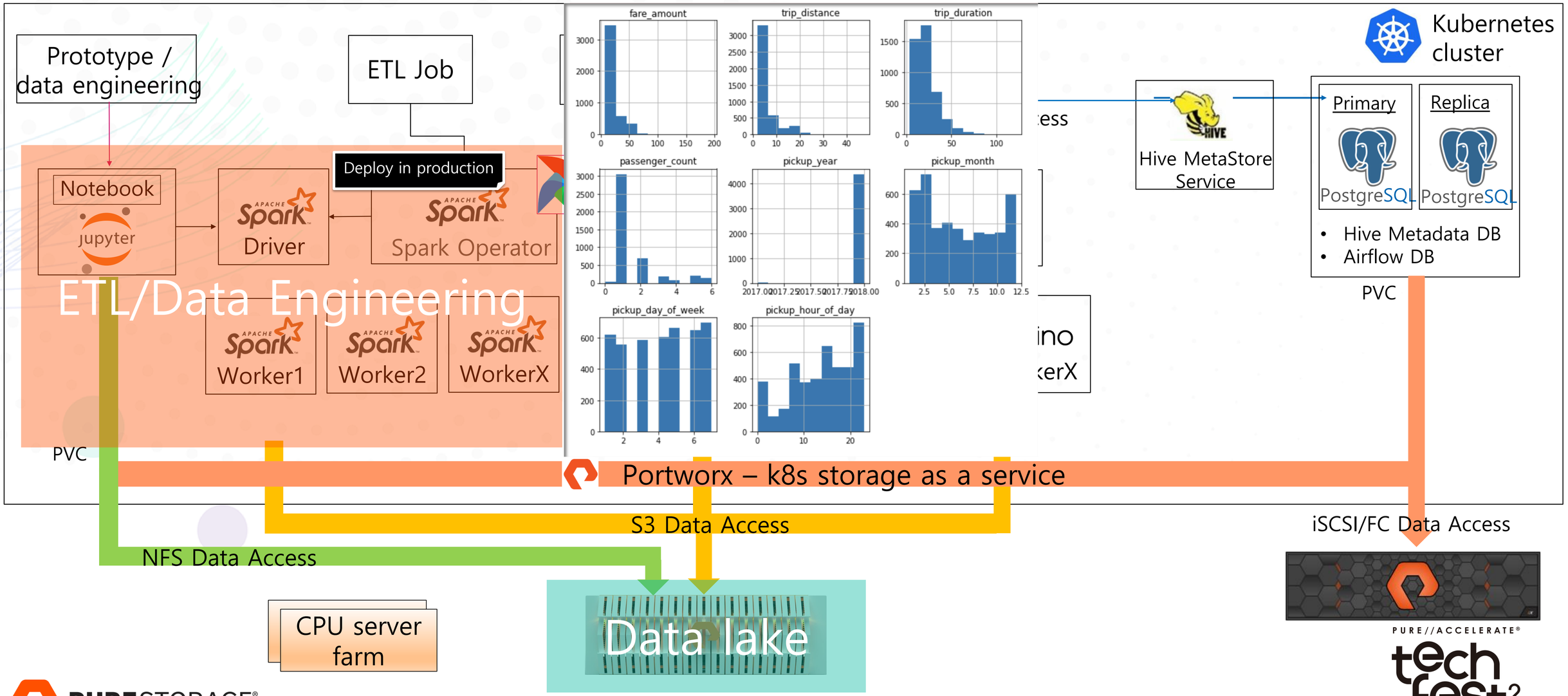
PURESTORAGE®

PURE//ACCELERATE®

techfest22 seoul

Data exploration and prototype with Jupyter on Pure

Jupyter, Spark, and Trino on Kubernetes with Pure for prototype, data engineering and workflow



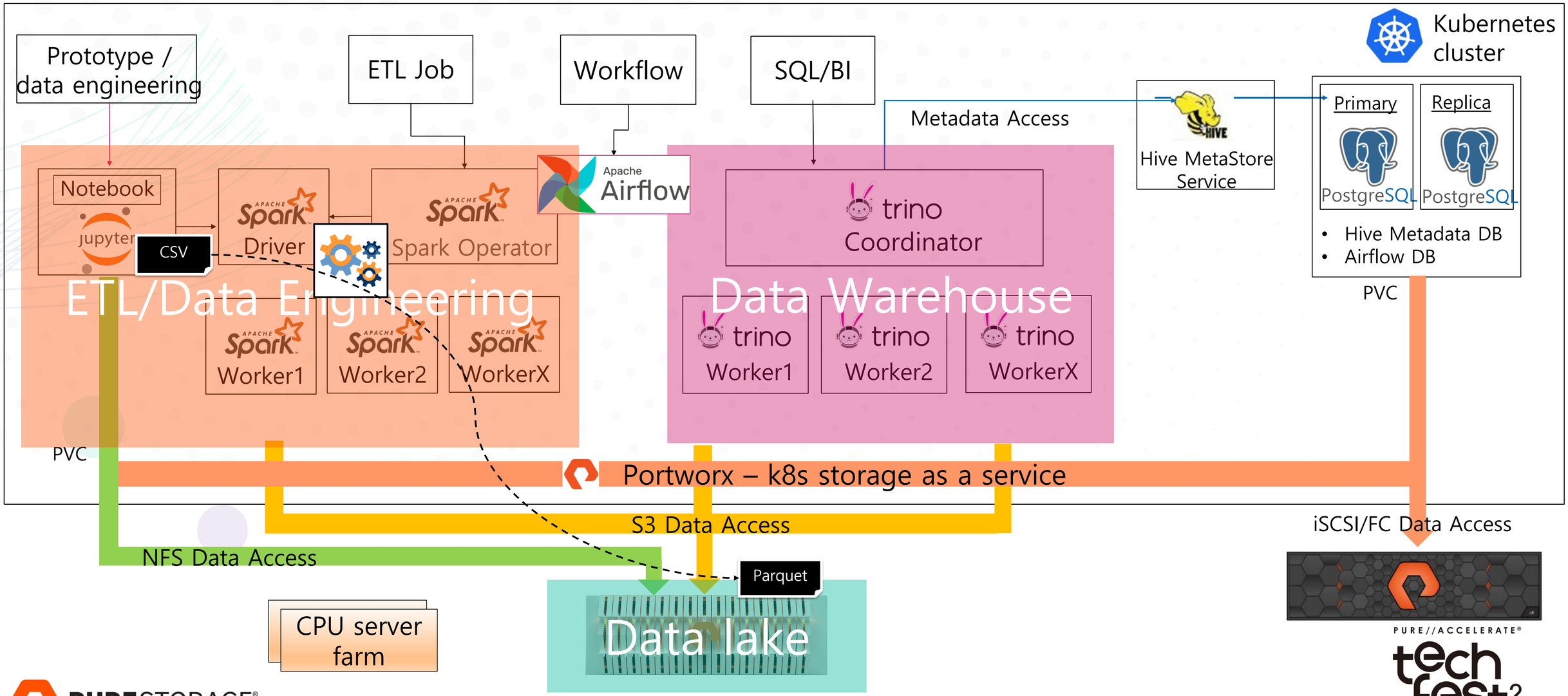
PURESTORAGE®

PURE//ACCELERATE®

techfest22 seoul

Simple & scalable Spark and data warehouse with Pure

Jupyter, Spark, and Trino on Kubernetes with Pure for prototype, data engineering and workflow



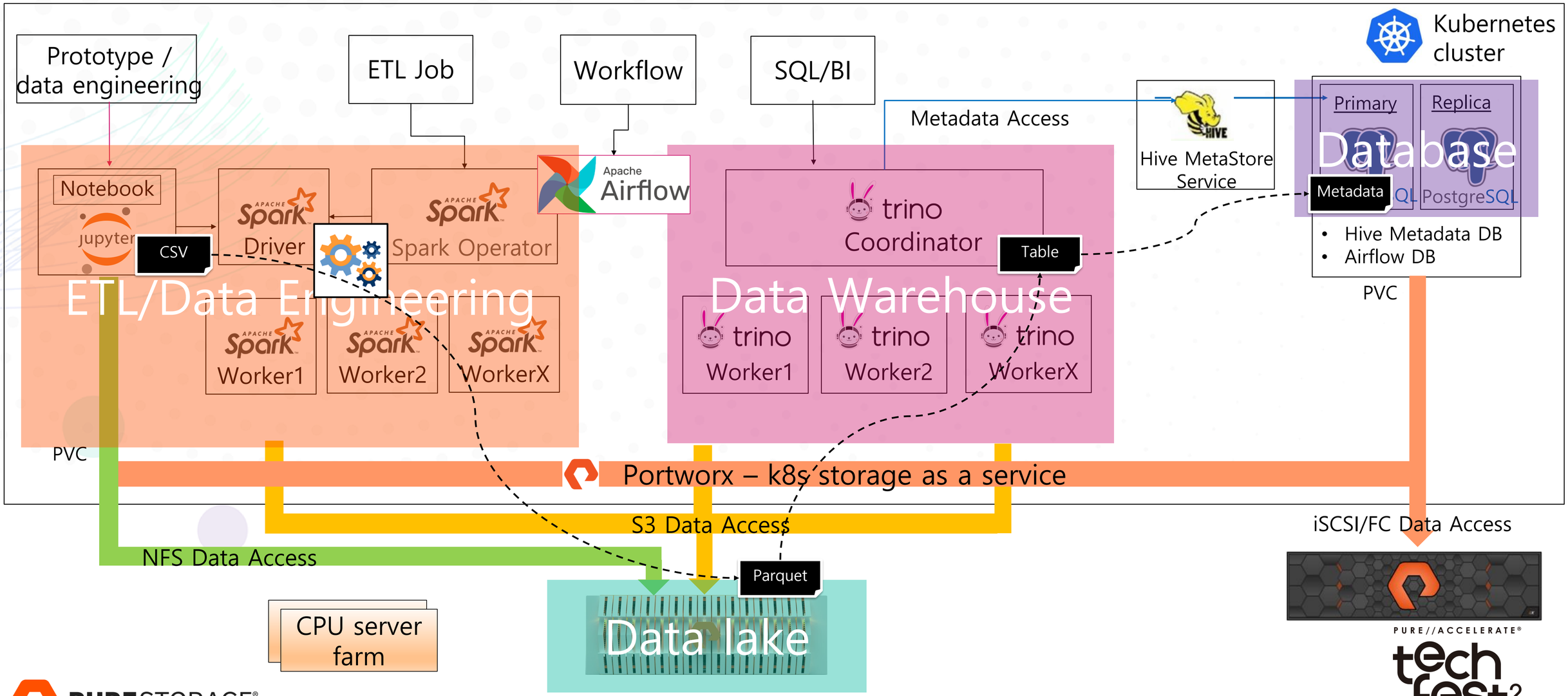
PURESTORAGE®

PURE//ACCELERATE®

techfest22 seoul

Simple & scalable Spark and data warehouse with Pure

Jupyter, Spark, and Trino on Kubernetes with Pure for prototype, data engineering and workflow



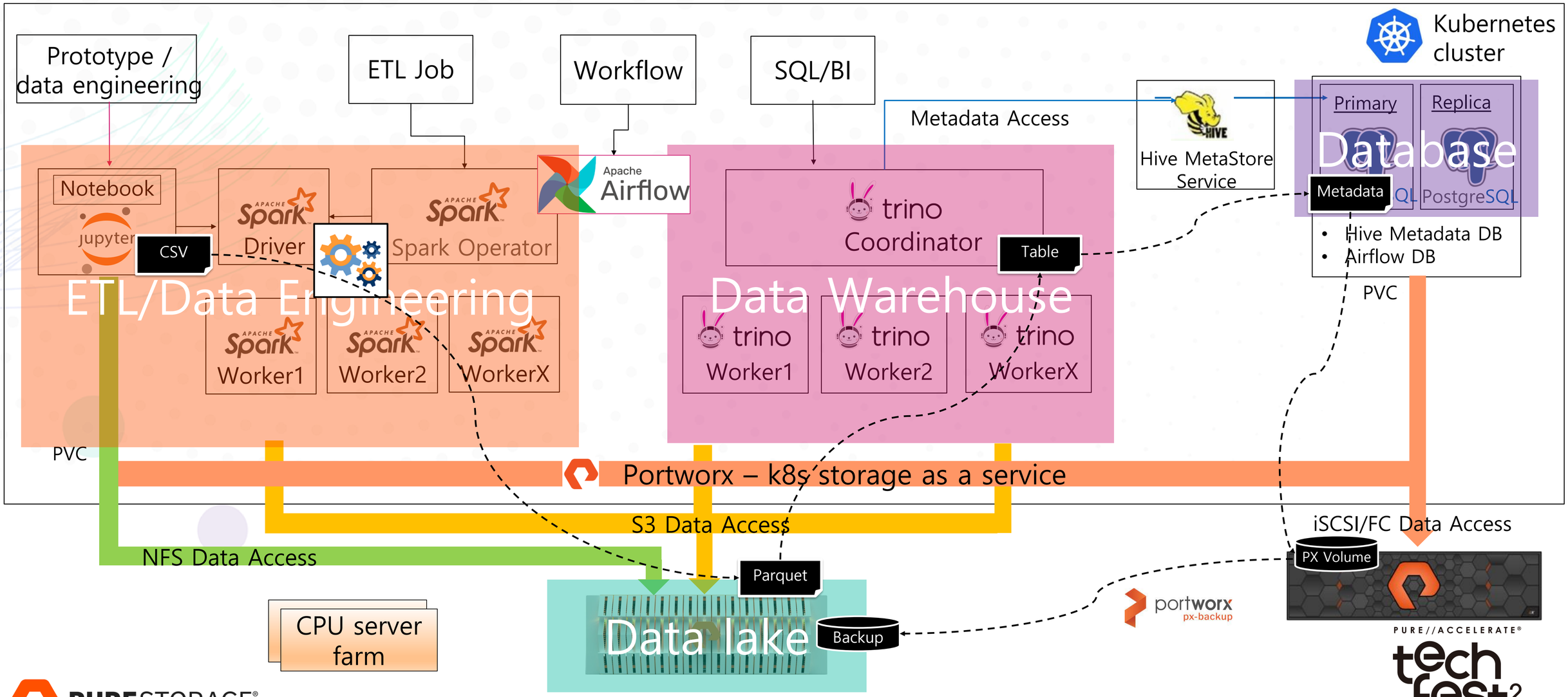
PURESTORAGE®

PURE//ACCELERATE®

techfest22 seoul

Simple & scalable Spark and data warehouse with Pure

Jupyter, Spark, and Trino on Kubernetes with Pure for prototype, data engineering and workflow

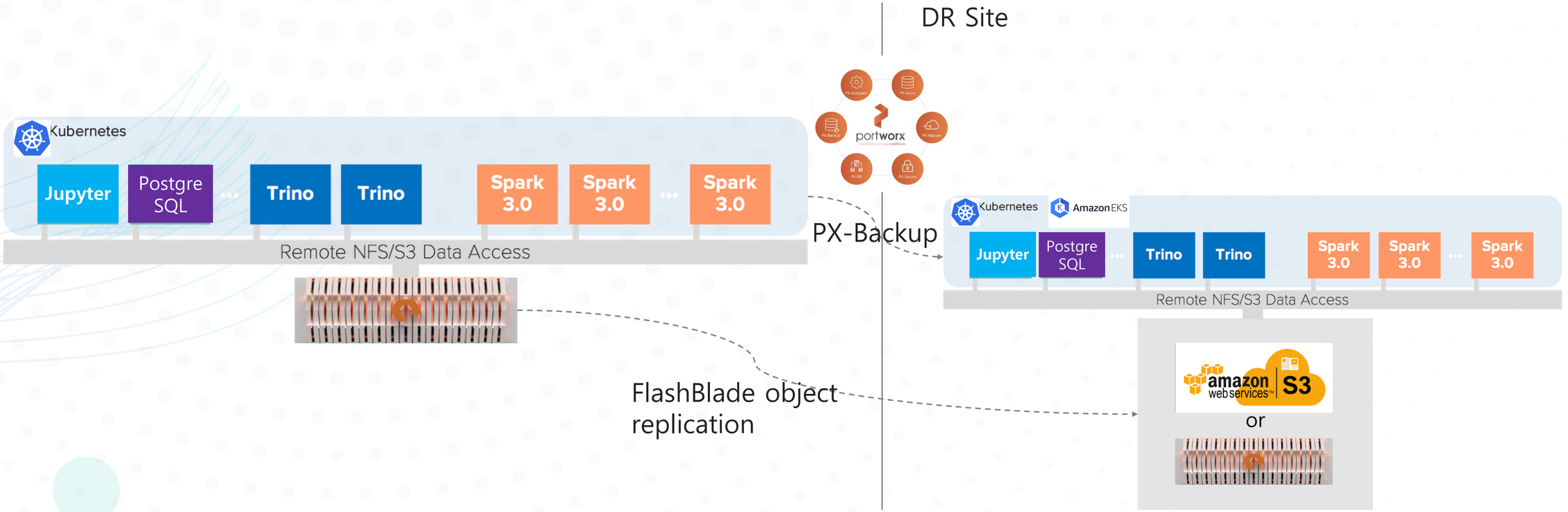


PURESTORAGE®

PURE//ACCELERATE® techfest22 seoul

DR & Hybrid cloud with FlashBlade and Portworx®

Automated, secure and low-cost application-aware backup and DR for big data



Hybrid cloud made easy

- Replicate data to remote FlashBlade or Amazon S3 for DR using FlashBlade replication.
- Backup Kubernetes apps, restore on cloud hosted Kubernetes services using Portworx PX-Backup.

PURESTORAGE®

PURE//ACCELERATE® techfest22 seoul

Key Takeaways

- Data engineering simplified with Pure – a single environment for prototype, production and workflow.
- FlashBlade S3 is fast and unified data lake.
- Portworx delivers storage-as-a-service, application-aware backup and enterprise storage features to Kubernetes.

PURE//ACCELERATE®
tech fest 22
seoul

Thank you

PURESTORAGE®