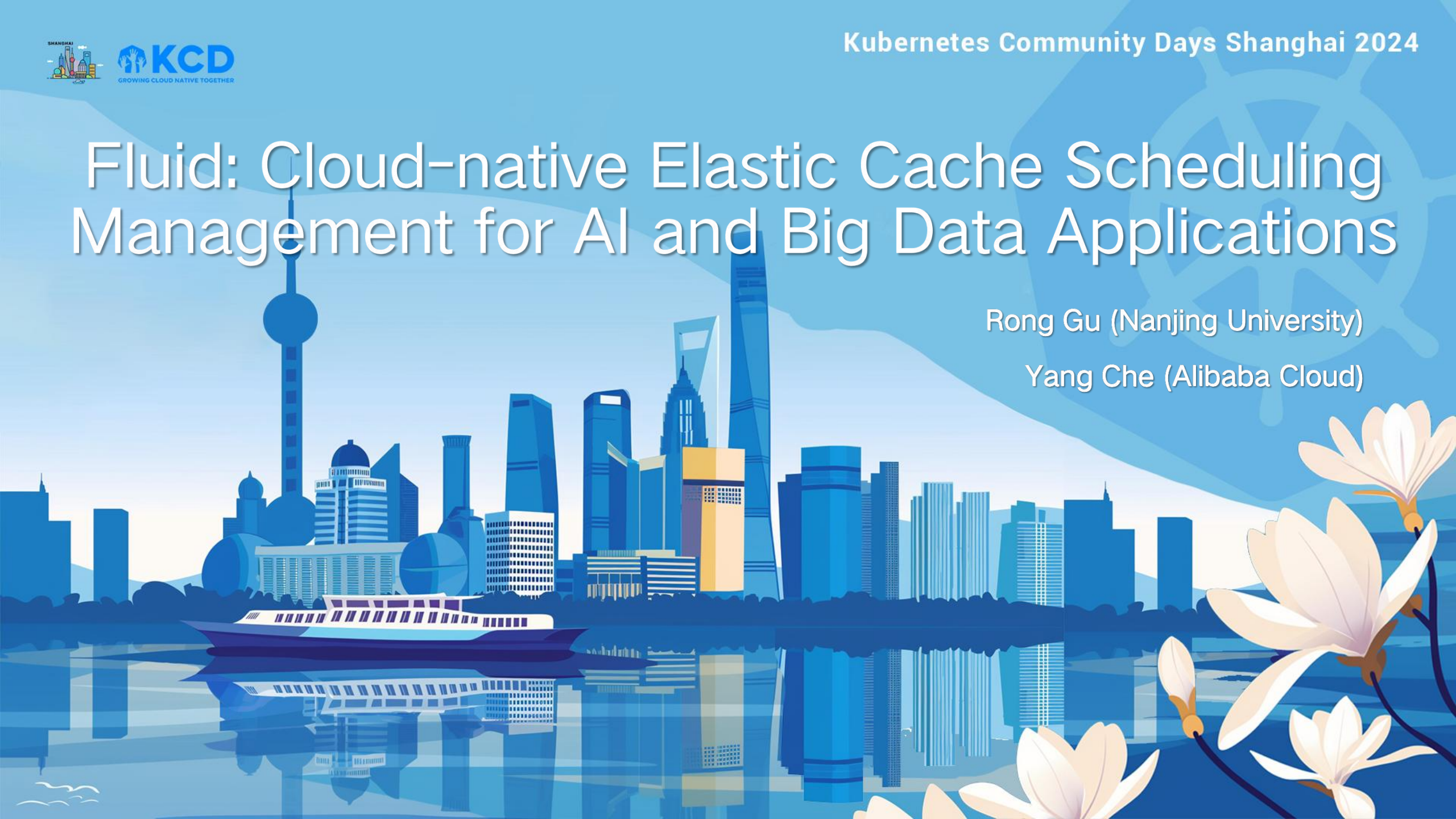
Fluid: Cloud-native Elastic Cache Scheduling Management for AI and Big Data Applications

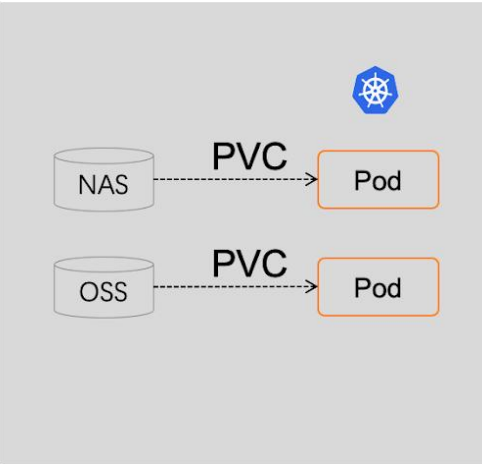
Rong Gu (Nanjing University)

Yang Che (Alibaba Cloud)

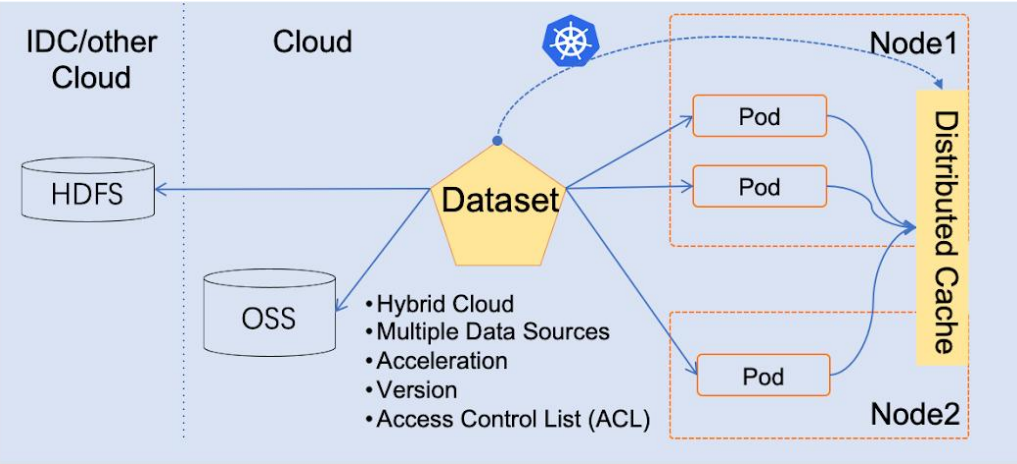


Abstract the “process of using data on Kubernetes”

The perspective of CSI



The data usage perspective of Fluid



Dataset Abstraction and Cache Management

- Allow users to create distributed data caches from datasets that are stored in different types of remote distributed storage, speeding up data access for your applications running in Kubernetes.

Data Cache Aware Scheduling

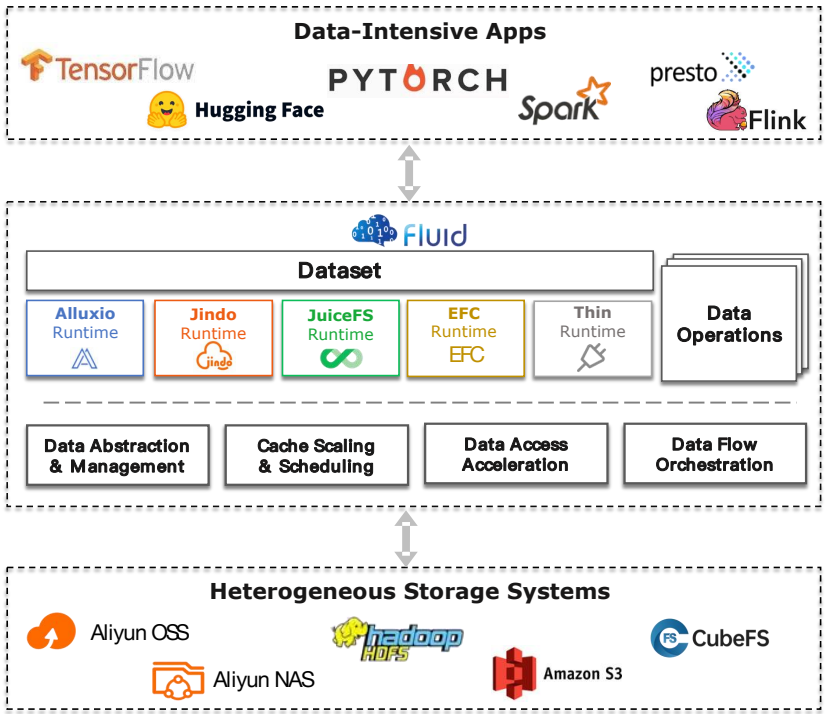
- Make workloads using dataset become aware of the data cache locality in Kubernetes and schedule them onto the nodes based on the distance of data transmission, such as within the same node, rack, availability zone, or region.

Elastic Cache

- Provide the ability to elastically scale the distributed cache up and down. This includes manual, timed, and automatic elastic scaling based on metrics. It’s according to the needs of the workload, which leads to high I/O throughput or cost saving.

Fluid — Cloud Native Data Orchestration and Acceleration

Fluid is a CNCF sandbox project focusing on AI/ML Data and Task Orchestration in Kubernetes, helping users to manage data, accelerate data access, and simplify data processes.



Standardized – K8s Native APIs for dataset management and orchestration

Performance – Accelerate data access via elastic distributed cache

Extensible – Runtime plugins for different cache and storage backends

Elasticity – Data and task co-aware scheduling and auto scaling

Run Anywhere – Support CSI and sidecar mode for different Kubernetes environments

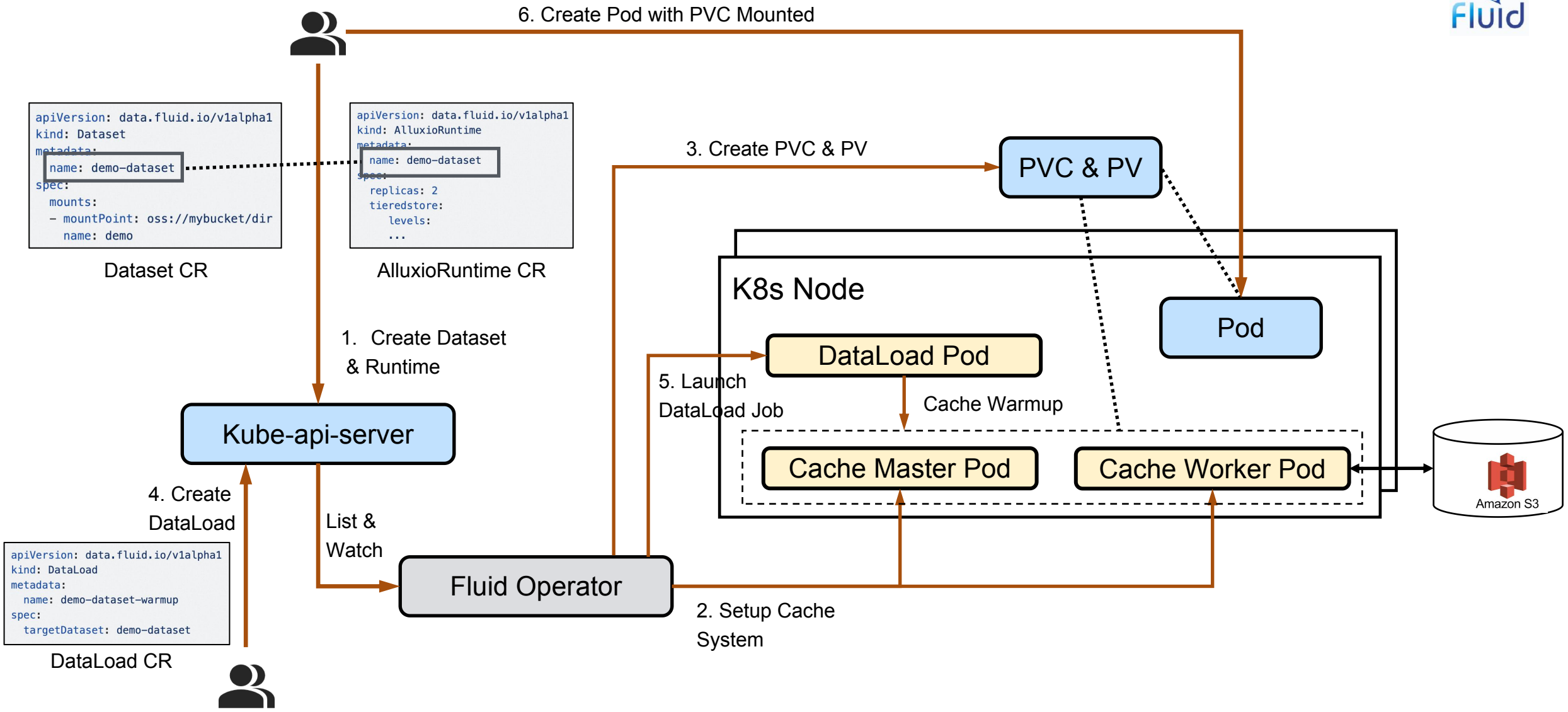
Joint launched by Nanjing University, Alibaba Cloud and Alluxio

https://github.com/fluid-cloudnative/fluid

More than 500 contributors and 30 enterprise adopters.

https://github.com/fluid-cloudnative/fluid/blob/master/ADOPTERS.md

How to use Fluid



Fluid Recent Update

- Configurable tiered data locality scheduling capability to optimize affinity between App pod and the data cache.
- Support three kinds of data operations with different modes: once, onEvent, Cron, including: DataLoad, DataMigrate and DataProcess.
- Support defining dataflow for data operations
- Support Python SDK for data scientists and operators to interact with Fluid control plane.
- A new runtime for sharing in-memory immutable data: VineyardRuntime is supported in Fluid.
- Security Hardening: Define a more restricted minimum necessary cluster role permissions for Fluid components, including eliminating all the secret-related and some create/update/delete privileges.
- CSI Plugin and the FUSE Recovery feature are optimized.

Contact us


Fluid开源社区技术交流群
575人
此二维码365天内有效（2025年04月16日前）
钉钉扫一扫群二维码，立即加入群聊

Thanks.