蚂蚁开源 ANT OPEN SOURCE

RAY

Ray History Server的历史任务可观测性实践

吴 昆

阿里巴巴高级研发工程师

2025/12/20

吴昆

阿里巴巴 高级研发工程师

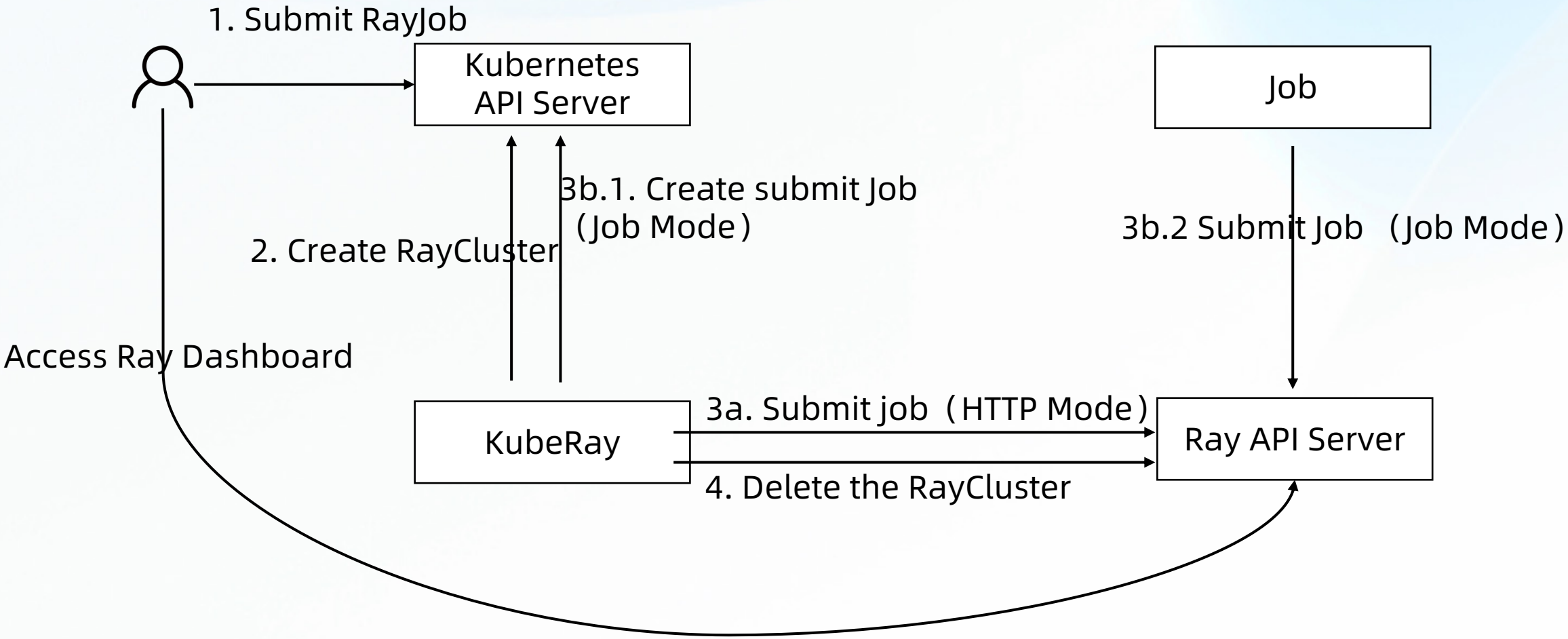
个人简介：吴昆就职于阿里巴巴容器服务团队，是容器服务调度领域负责人以及Ray on ACK研发，同时也是Koordinator，KubeRay，Kueue，Kubernetes等社区贡献者。

蚂蚁开源 ANT OPEN SOURCE

RAY

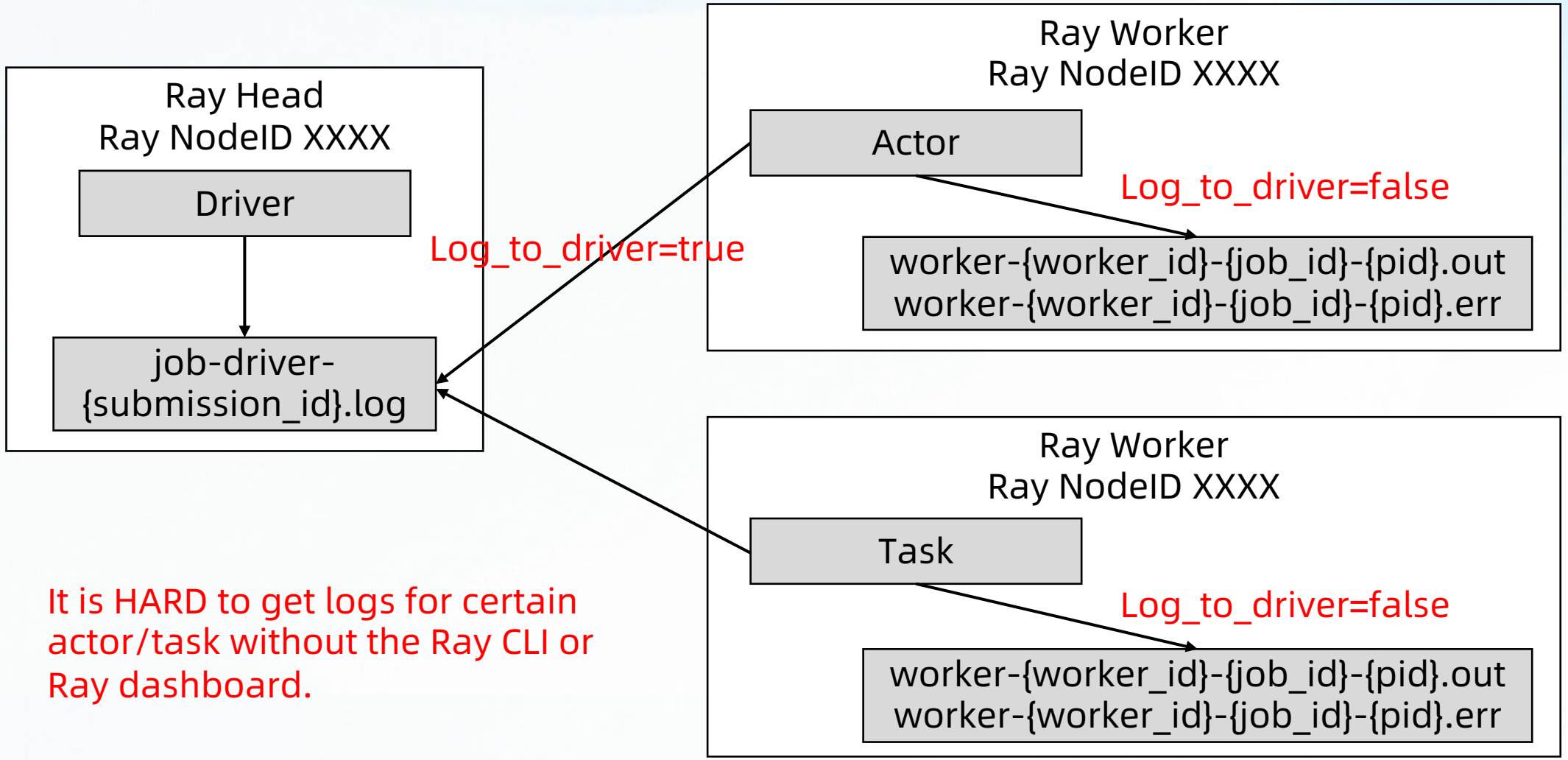
CONTENT
目录

RayJob on Kubernetes



We can not access the dashboard after the job complete

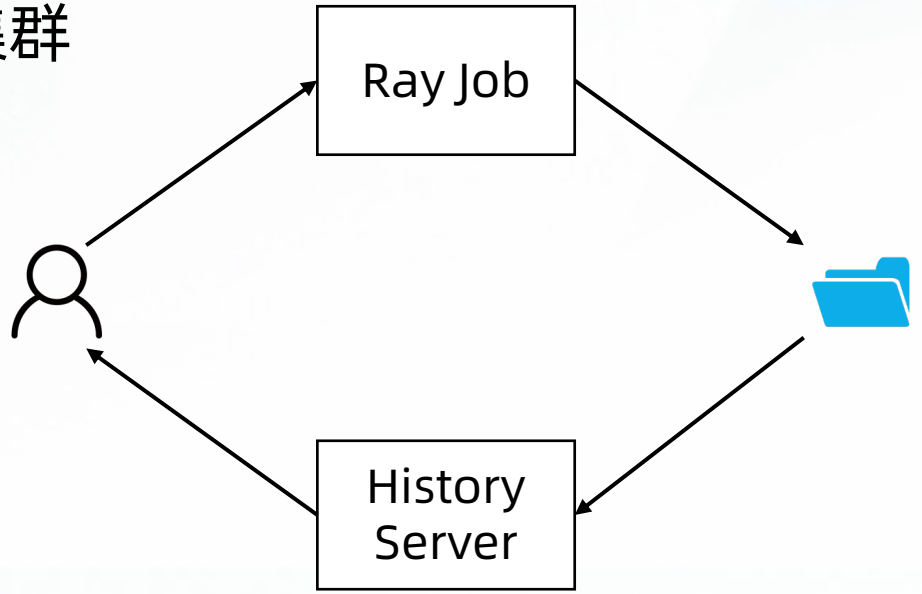
Ray的日志管理方式



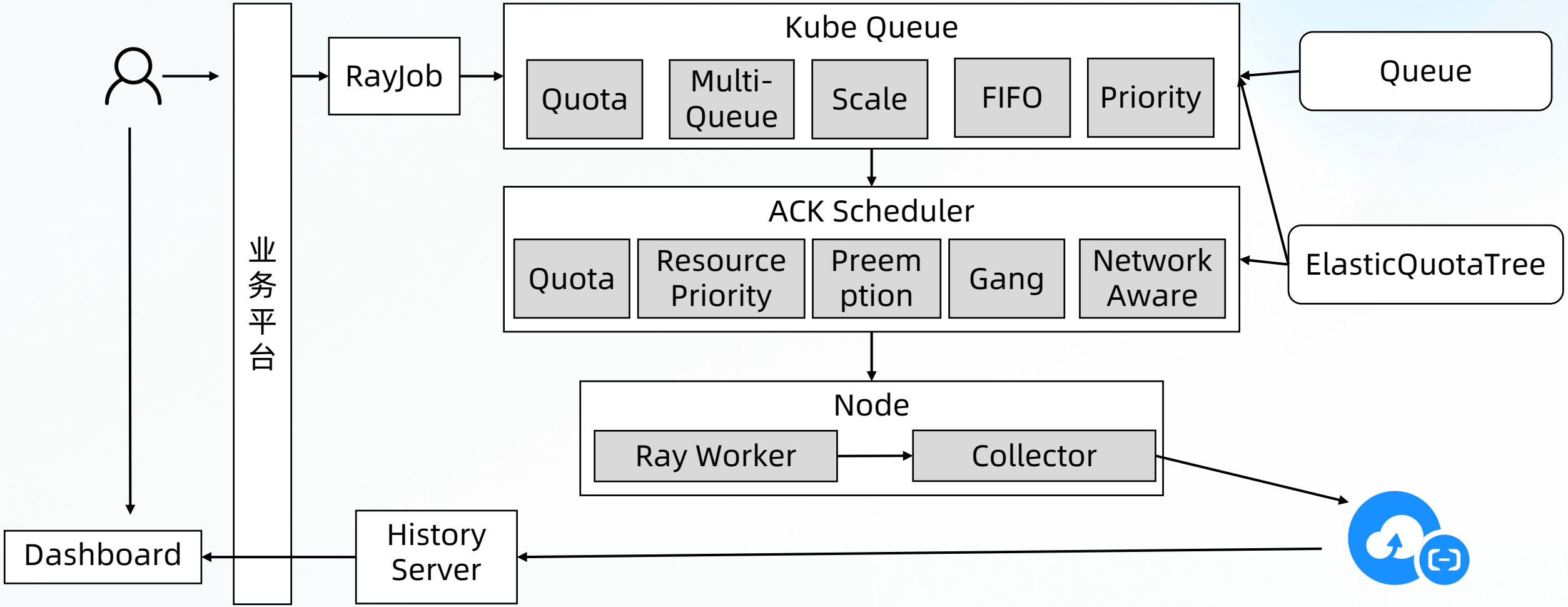
It is HARD to get logs for certain actor/task without the Ray CLI or Ray dashboard.

Why History Server ?

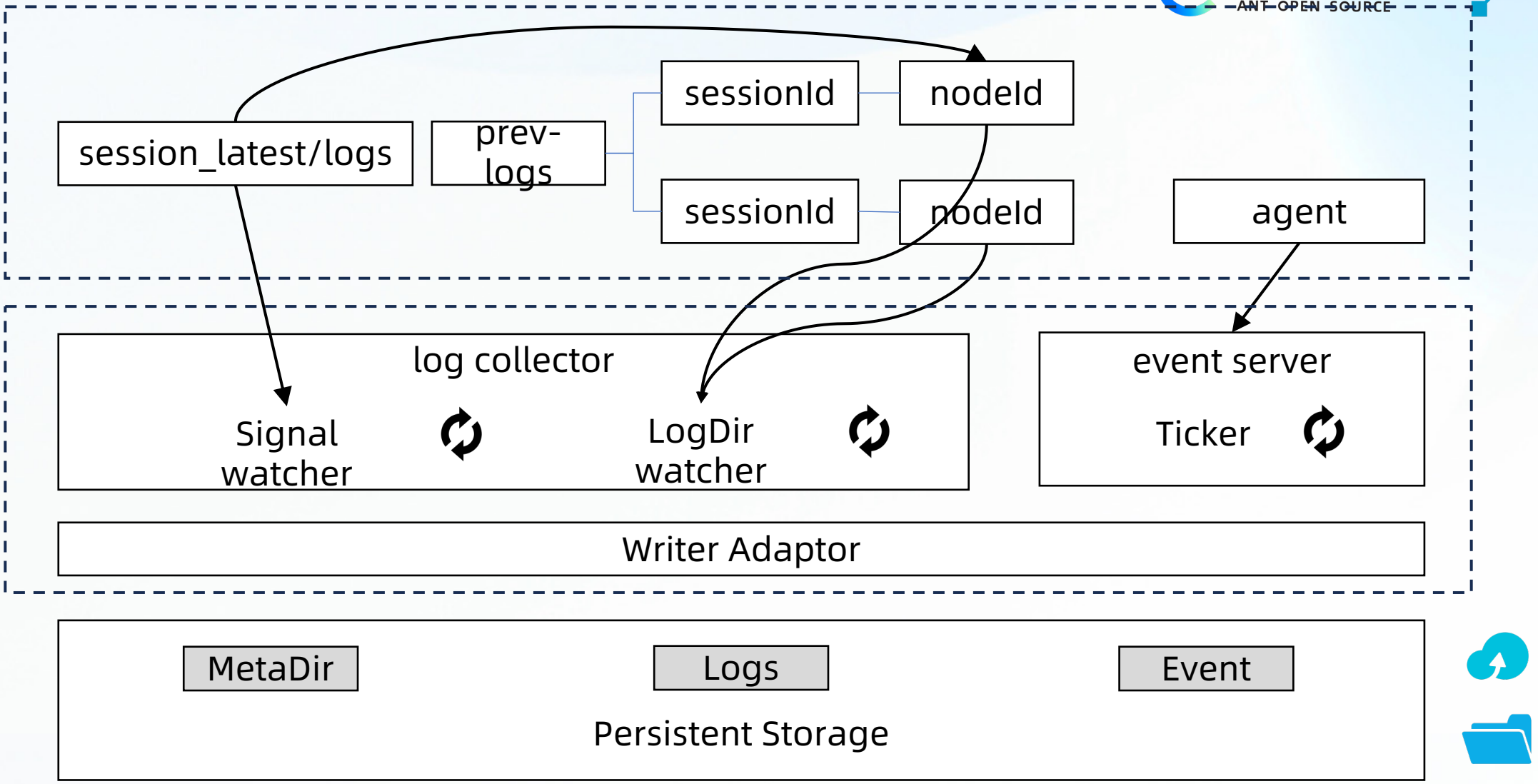
- 支持集群销毁后通过Ray Dashboard访问日志以及集群状态，保持一致的用户体验
- 日志以及集群数据存储在后端，避免部署多个Ray APIServer，节省集群资源
- 通过Event收集，HistoryServer可以支持集群状态回放等功能
- 通过Event收集以及重放，减少对Ray API Server的访问，提升Ray API Server性能



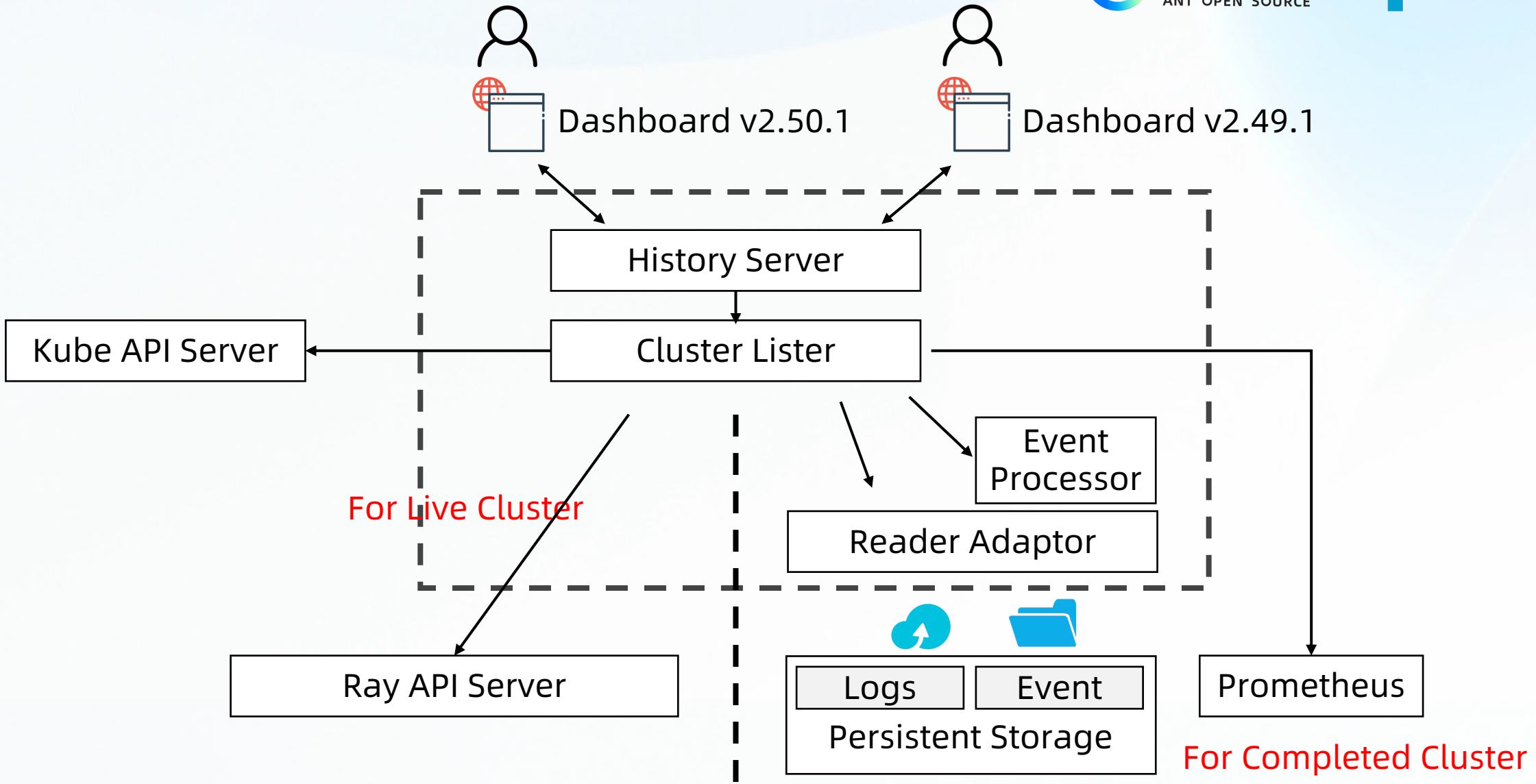
Ray on ACK使用场景



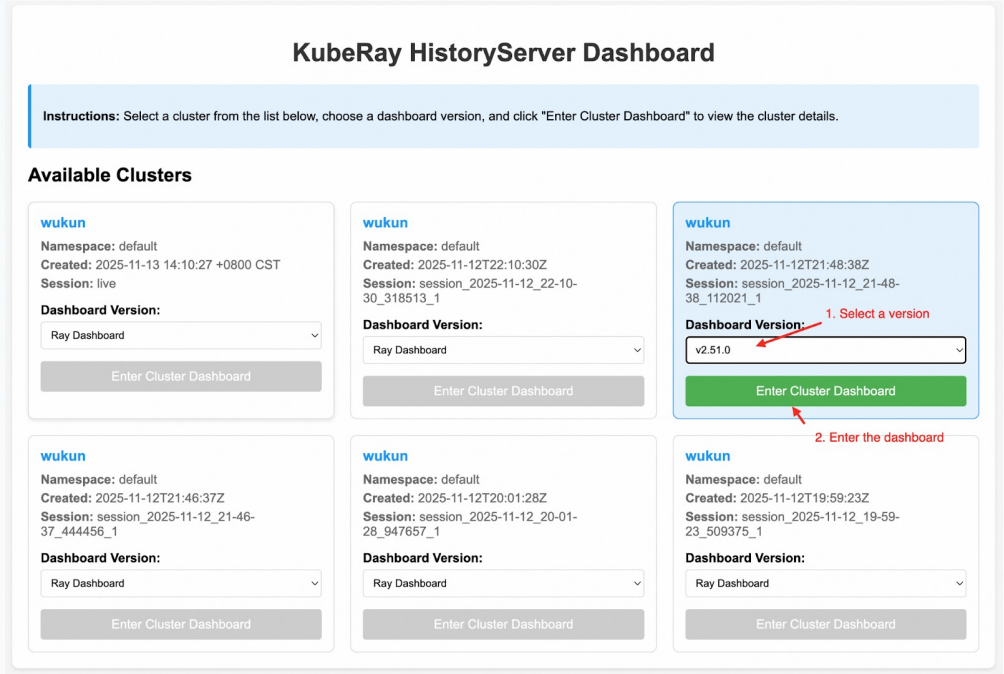
History Server 的设计与实现

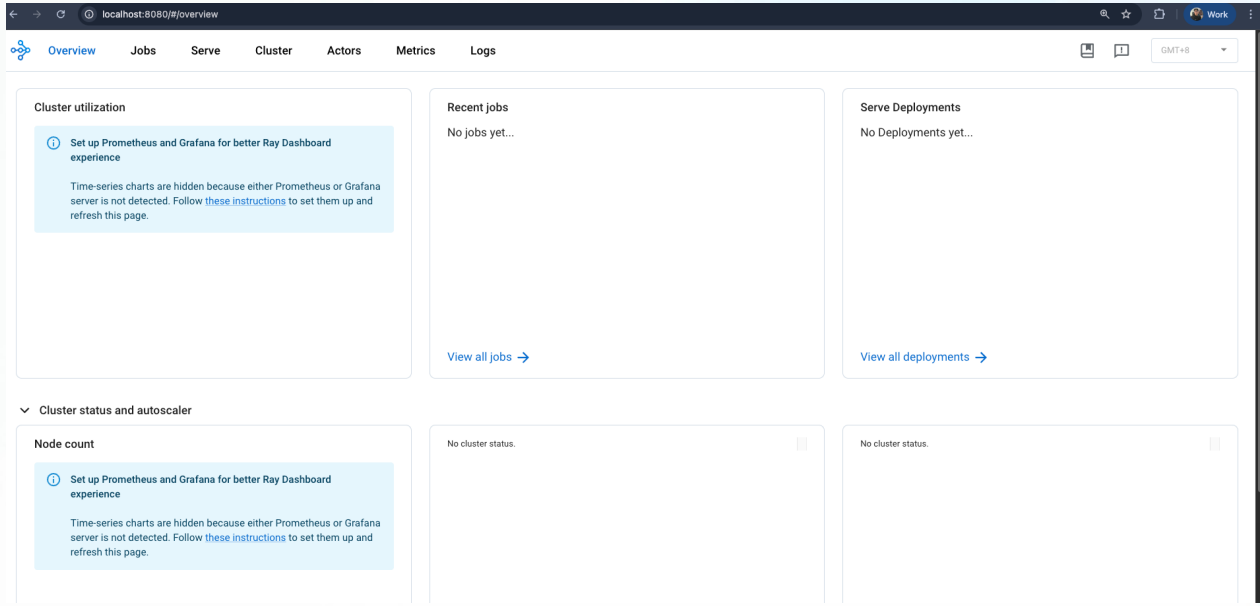


History Server 的设计与实现



History Server 的设计与实现

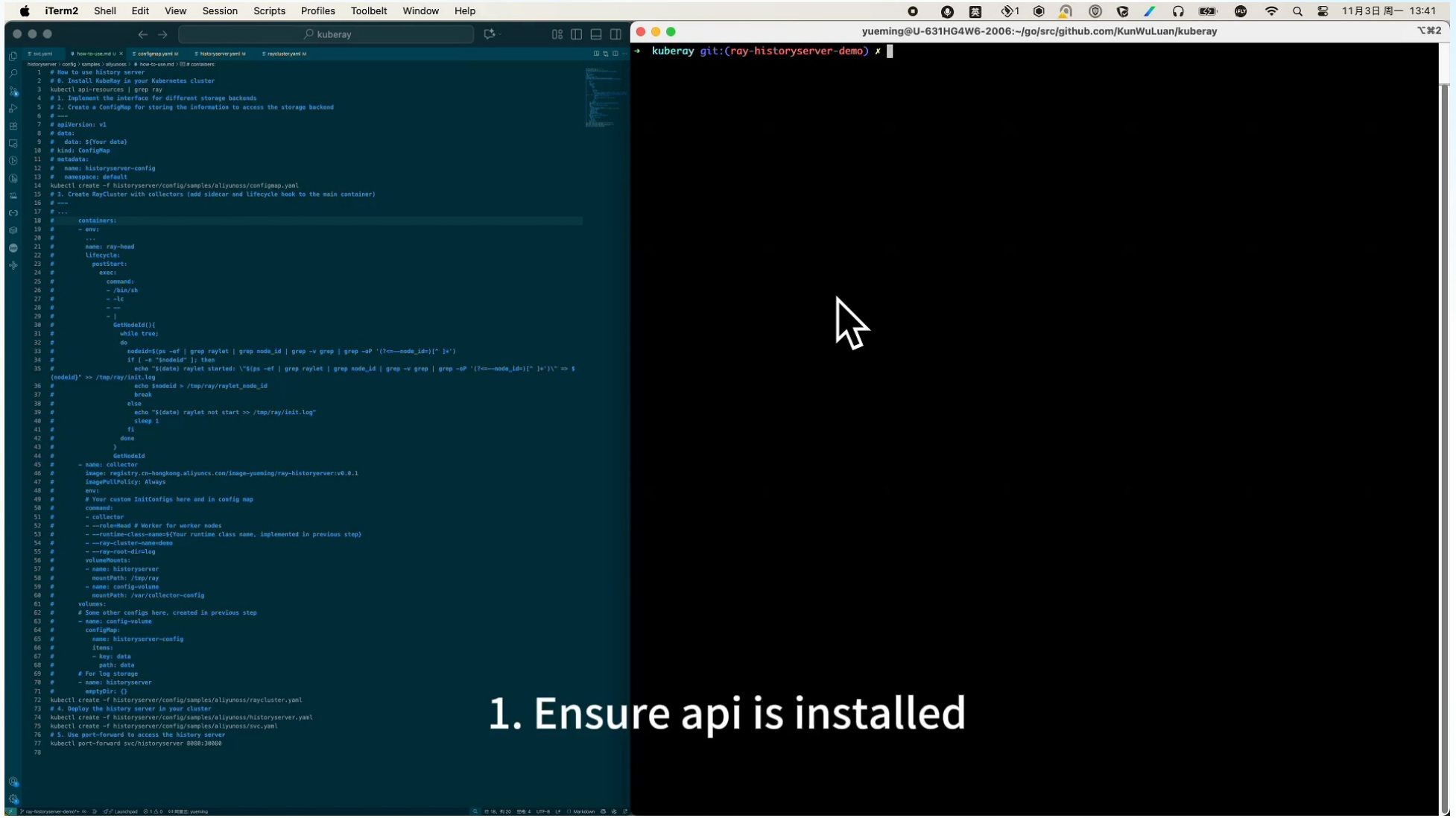




Demo视频

蚂蚁开源 ANT OPEN SOURCE

RAY



History Server RoadMap

Alpha Version
(push on write)
(api snapshot from API server)
(not support event)
(require Storage support Append)
(support only one dashboard)
Released in local branch

→

Beta Version
(push when pod is deleted)
(rebuild status from event)
(Any storage supported)
(Any version dashboard supported)
KubeRay v1.7

→

GA Version
KubeRay v1.8

What's Next

More storage
Authentication
Cluster state replay
...

蚂蚁开源 ANT OPEN SOURCE

RAY

谢谢

欢迎大家进群交流