Workload	Container Orchestrator	Distributed compute engine	Training & inference framework	Outcomes
Last-mile data processing +model training	**Kubernetes**-based PinCompute platform	Previously Spark, now **Ray** due to streaming execution and heterogeneous CPU + GPU compute	**PyTorch**	Dataset-iteration wall-clock cut by 6x (from 90h to 15h). Dev cycles cut from days to hours. GPU utilization over 90%. Training throughput up 45% while per-job cost down 25%. ✧ (blog)
Offline / batch inference	**Kubernetes**-based PinCompute platform	Previously Spark, now **Ray**	**PyTorch, vLLM**	30x decrease in cost for search-quality jobs. 4.5x throughput increase. 4x reduction in job runtime for GPU inference jobs. Running around 1800 jobs per month. ✧ (blog, blog)
Large-scale training Especially recommender models	**Kubernetes**-based PinCompute platform	**Ray** on heterogeneous clusters	**PyTorch**	Running 5000+ training jobs per month. Transparent scaling. Greater developer velocity and interactive development. Enable heterogeneous resources to keep GPUs saturated. ✧ (blog, talk)