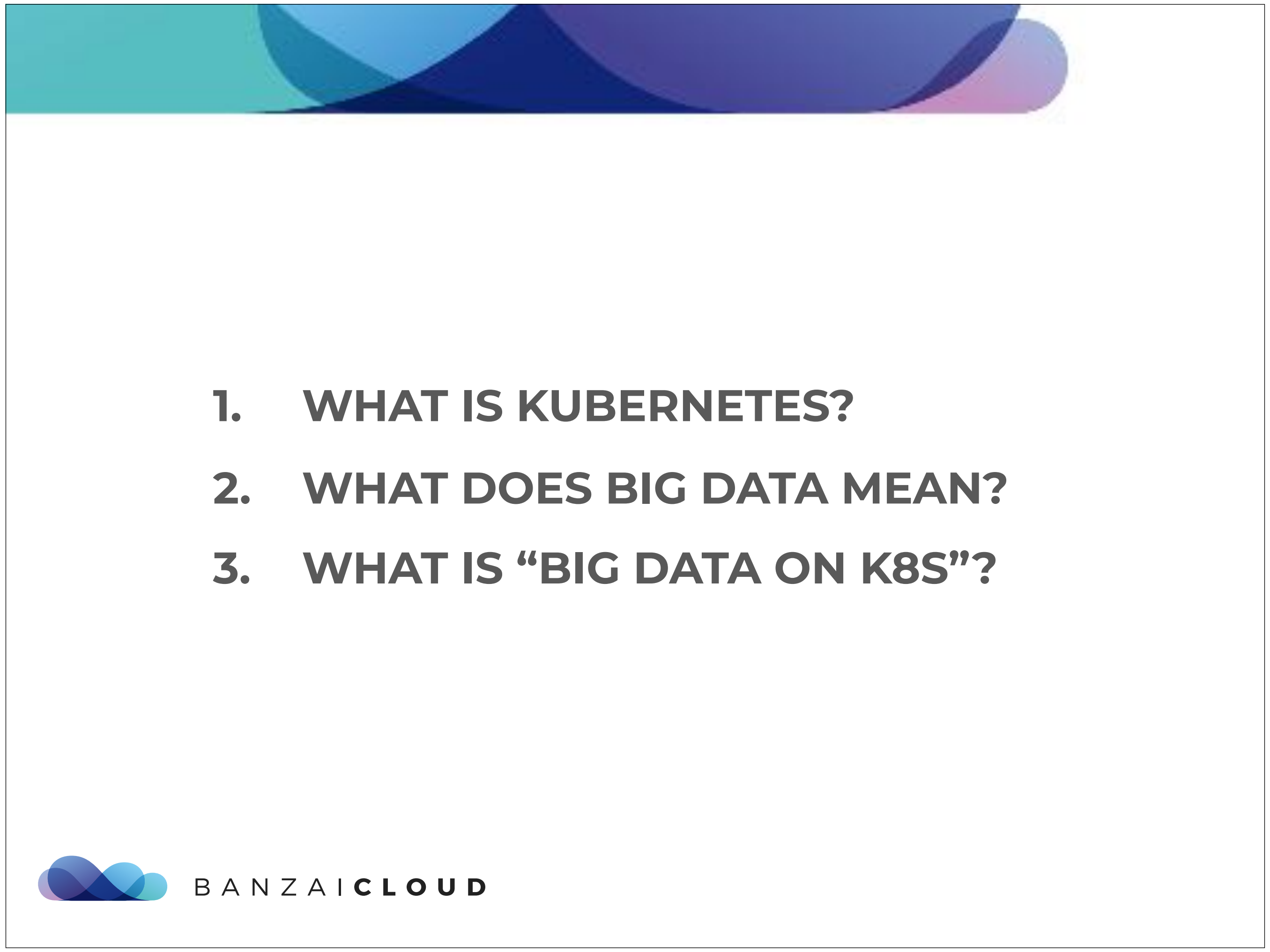
Big Data Workloads on Kubernetes

intro

Marton Sereg

BANZAICLOUD

whoami and whoisbanzai



1. WHAT IS KUBERNETES?
2. WHAT DOES BIG DATA MEAN?
3. WHAT IS “BIG DATA ON K8S”?

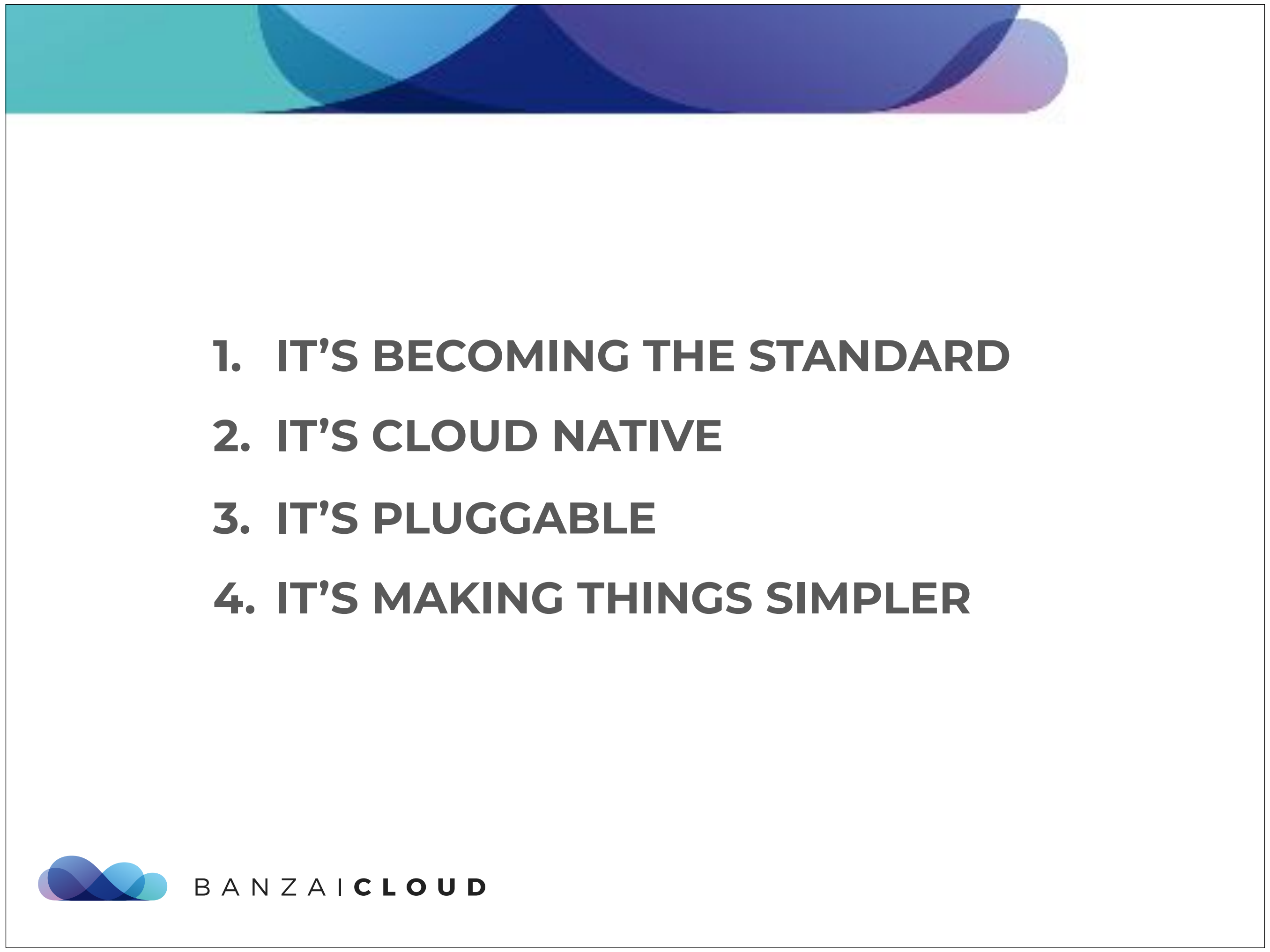
BANZAICLOUD

1. Q? Container orchestration - deployment, scaling, management of containers / good community / becomes a standard / CNCF landscape - standards, working groups for CNI, serverless events,
2. Hadoop ecosystem / Spark (data processing) / Zeppelin (notebook) / Kafka (distributed streaming, messaging) - why these? Most popular, not attached too tightly to core hadoop/yarn Q?
3. Running Spark, Zeppelin and Kafka natively on k8s, without any unnecessary “glue”

WHY IS IT GOOD TO RUN THESE WORKLOADS ON K8S?

BANZAICLOUD

Main question
See next slide



1. **IT’S BECOMING THE STANDARD**
2. **IT’S CLOUD NATIVE**
3. **IT’S PLUGGABLE**
4. **IT’S MAKING THINGS SIMPLER**

BANZAI**CLOUD**

1. Not a technical advantage / companies need containers and therefore orchestration systems / kubernetes is becoming the de-facto standard / they will use it anyway / use it for big data - only one kind of infrastructure to manage
2. Designed from the ground up with cloud (resiliency, service-discovery, hybrid, multi-tenant, upgrades, isolation, unified interfaces) in mind / standards for containerd, CNI, container storage interface, grpc / the CNCF landscape / monitoring (prevention, downtime, cost - Prometheus), logging (unified, collect - Fluentbit/d), service mesh (reliability, security, monitoring, canary testing - Istio, Conduit)
3. The model is turned upside down - cloud providers need to add their extensions for persistent volumes, autoscaling vs ambari/yarn/cloudbreak), / custom schedulers / custom resources (e.g. Spark)
4. Yarn - not needed, kubernetes scheduler is fine / Zookeeper - not needed, etcd is working just fine / Twitter Heron - vs Storm / infrastructure code is separated from app logic

CHALLENGES?

BANZAICLOUD

Existing projects need code modifications - Still the beginning of the road
Data locality, external storage
Managing state - stateful sets, cloud object store (Ozone vs Minio,Rook), container storage interface,
logging, monitoring, etc still need to be set up properly
Optimized schedulers - heron - own scheduler (multiple schedulers can run as well)
Legacy vendors
Mindset - open source community

Thank you
Q & A

BANZAICLOUD