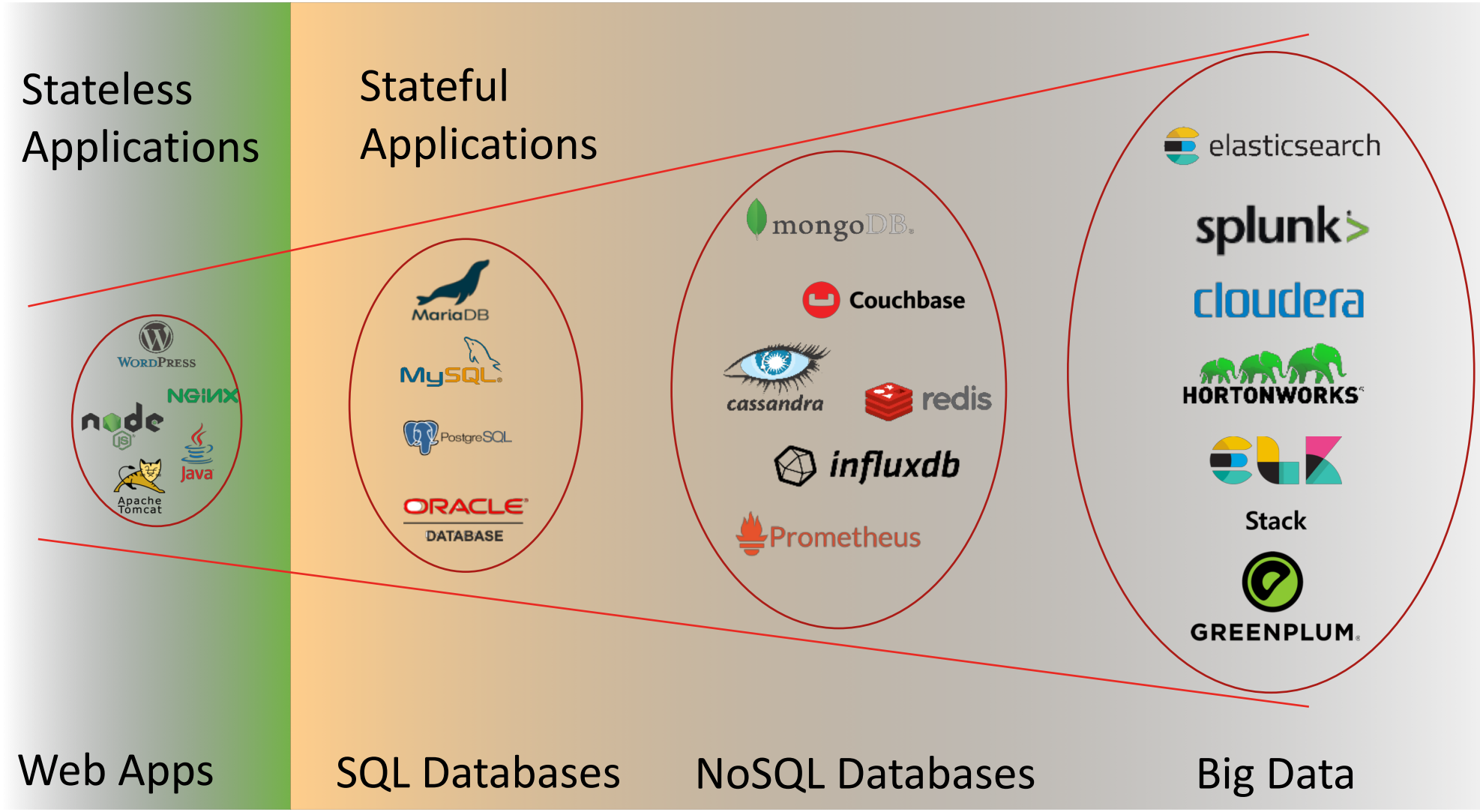
Data Protection for Application Running on Kubernetes

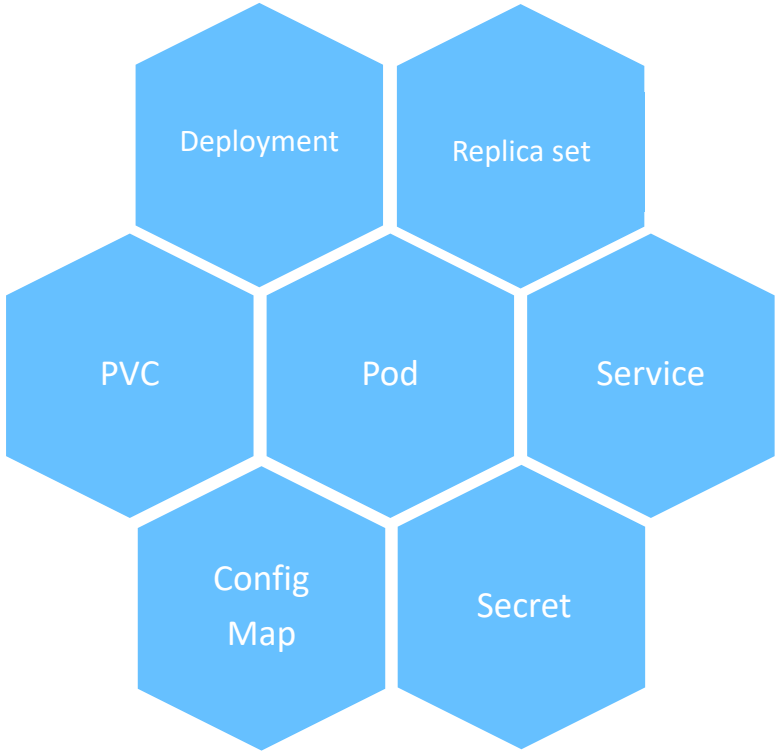
Ravikumar Alluboyina

Senior Product Architect, Robin.io

Spectrum of Applications

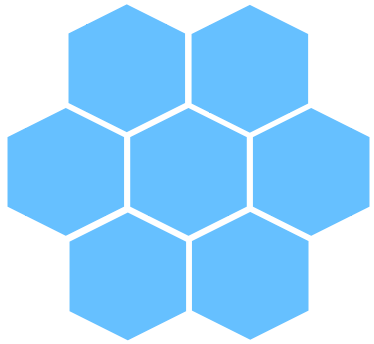


Application Composition

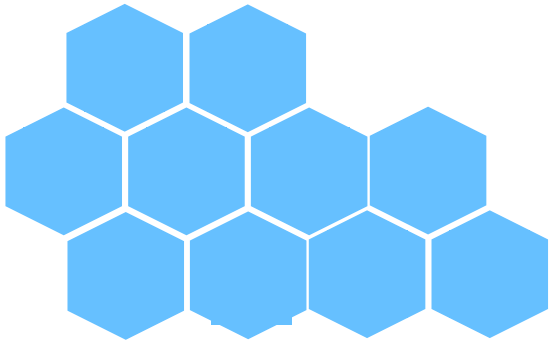


https://github.com/helm/charts/tree/master/stable/mysql

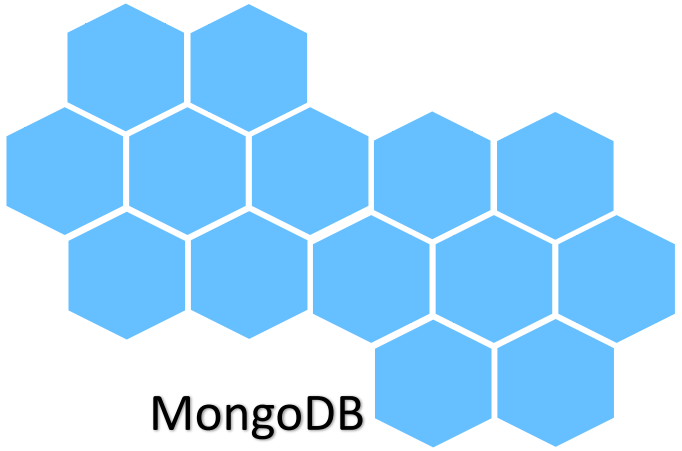
Application Composition .. The complexity



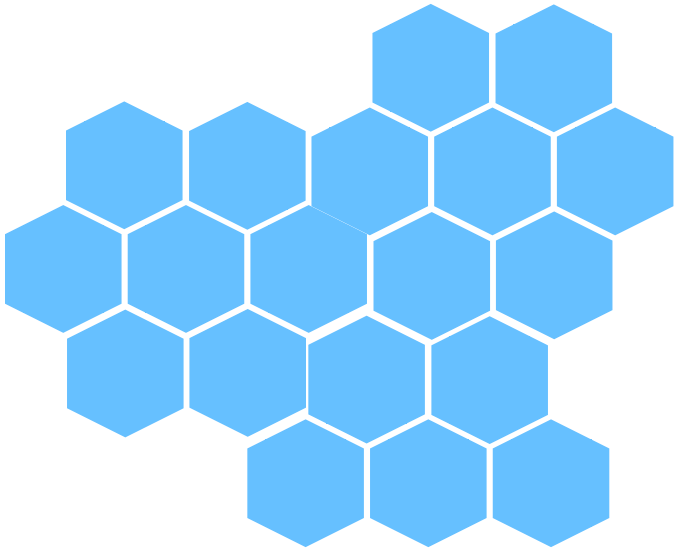
MySQL



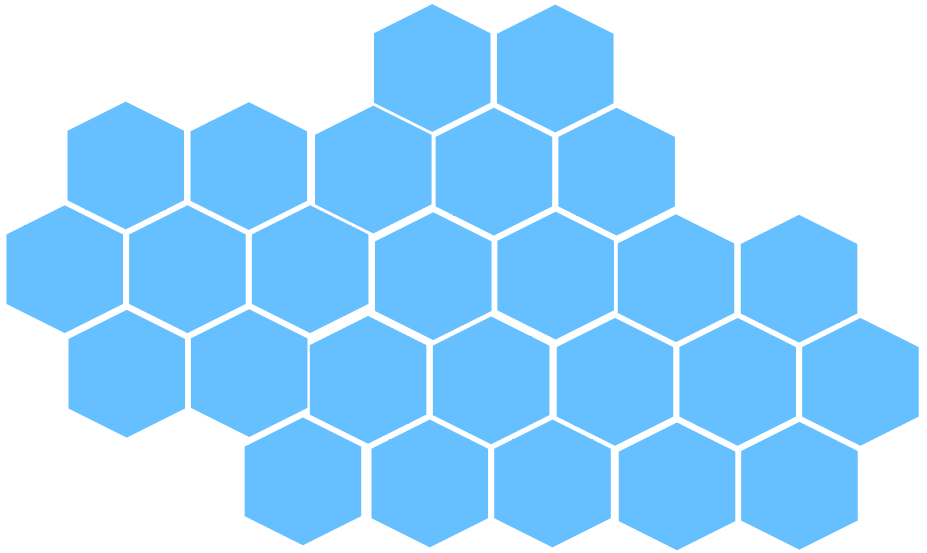
MariaDB



MongoDB



ElasticSearch



ELK Stack

https://github.com/helm/charts

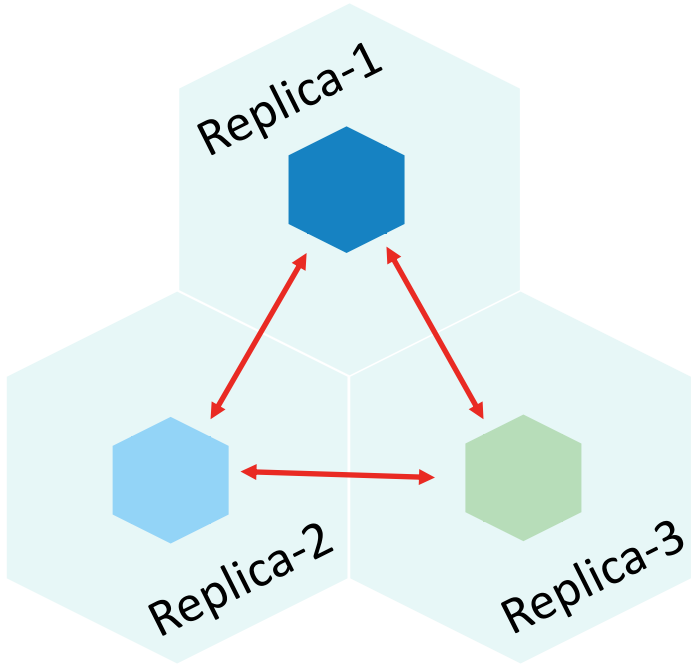
Data Protection

- Environment
  - Highly virtualized using containers
  - Highly consolidated
  - Multiple abstraction layers (Kubernetes, Docker, CRI, CNI, CSI)
  - Large scale
  - Multi Datacenter or Geo distributed
  - Distributed applications

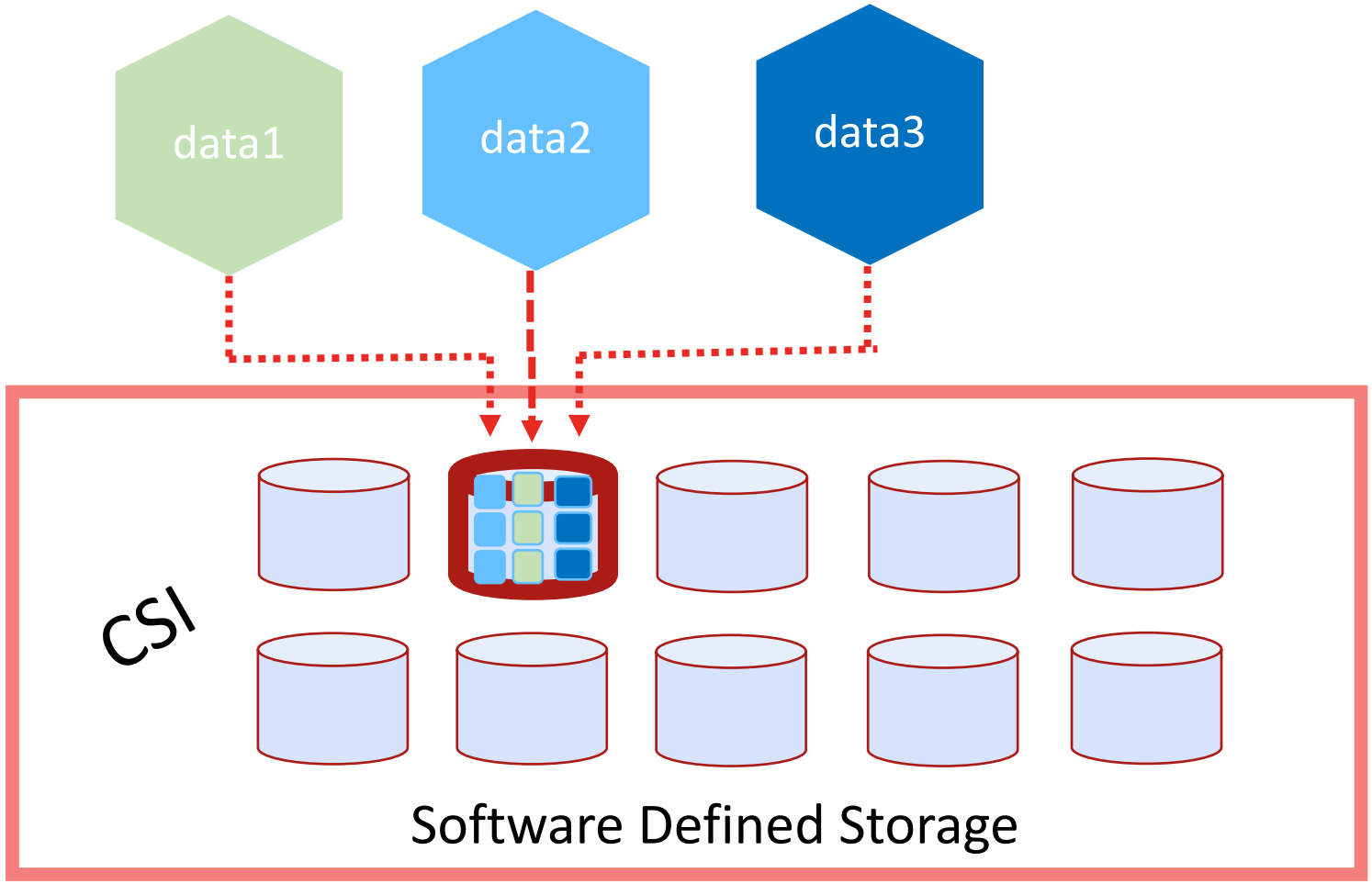
- Protect from
  - Poor resource planning
  - User errors
  - Hardware failures / Data center failures

Resource Planning

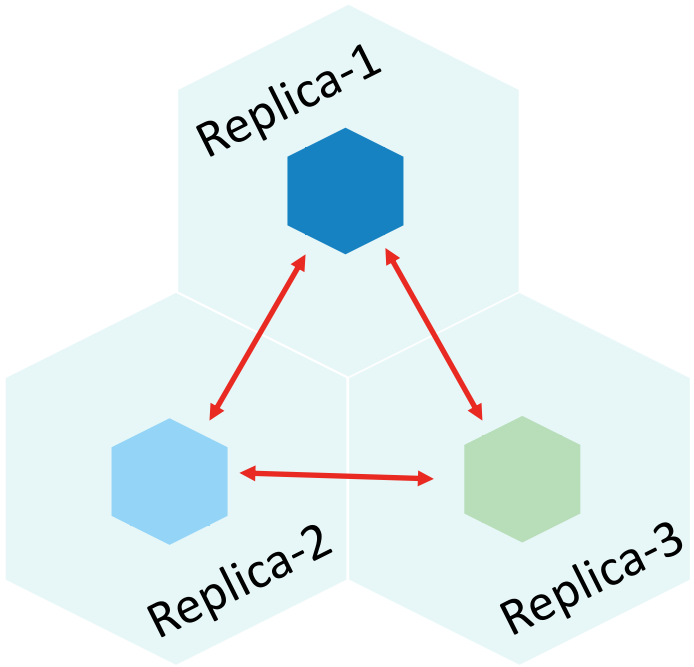
Cassandra Deployment

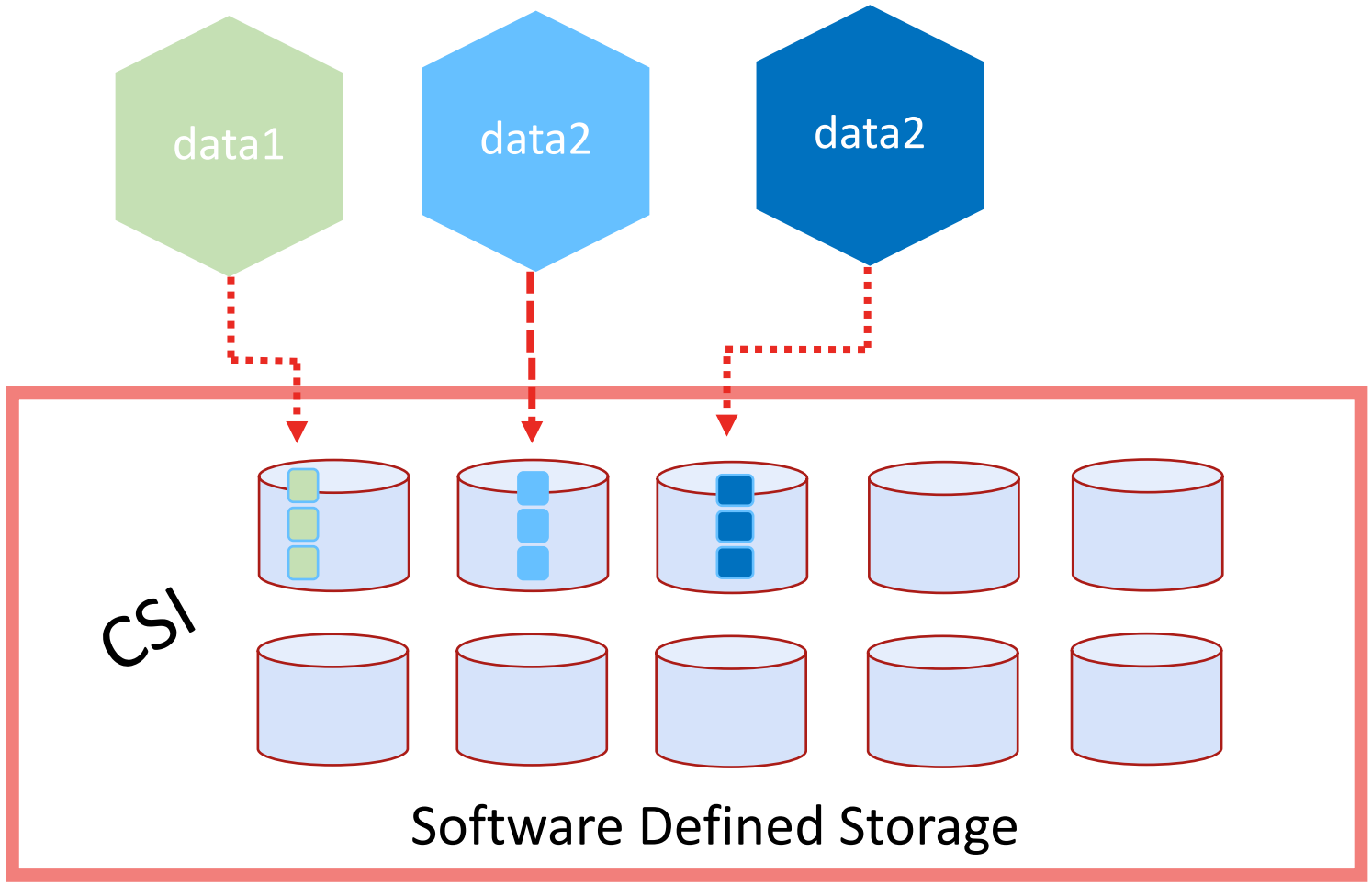


Still resilient to disk failure ???

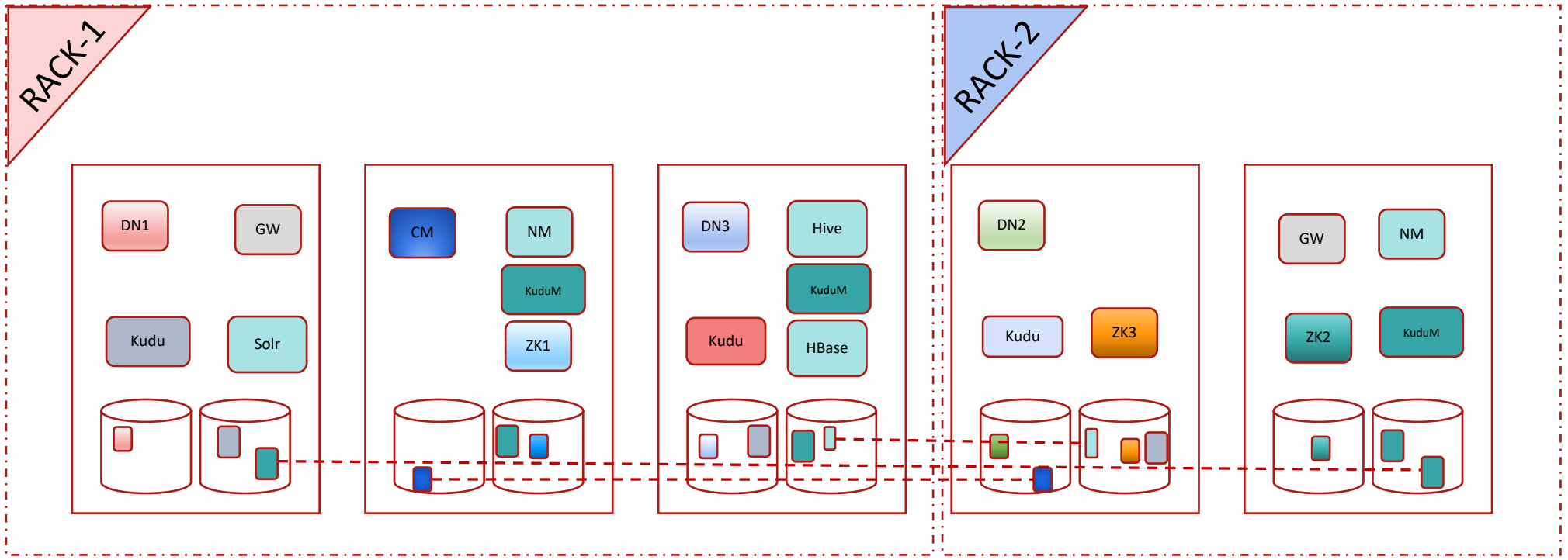


Let’s protect Cassandra …





Hadoop Deployment



Compute anti-affinity

Location Awareness Rack / DC

Storage & Compute Affinity

IO patterns QoS

High Availability

Application Planning Challenges

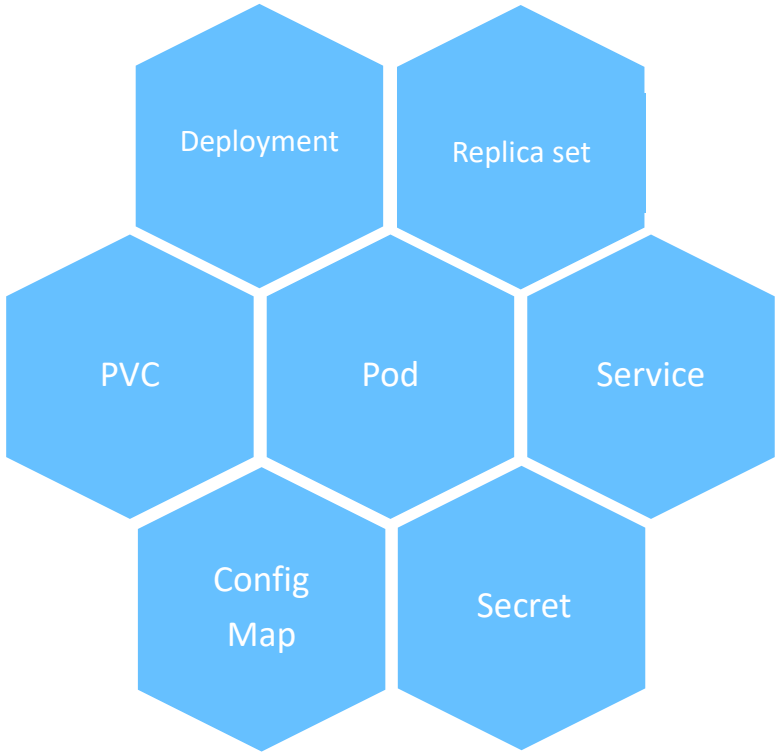
- Data-heavy applications deal with Multiple volumes
- Every volume will have different IO characteristics
- Consolidation (packing) makes the problem even harder
- Application Replication (Cassandra / Mongo) makes the allocation tricky

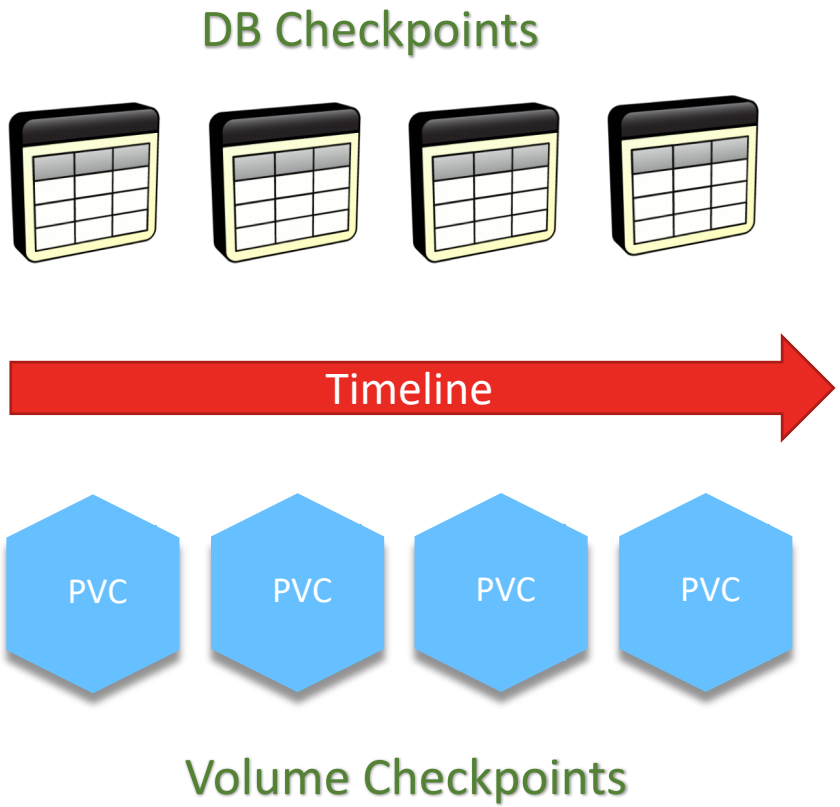
What are we looking for…..???

Application Aware Storage Provisioning

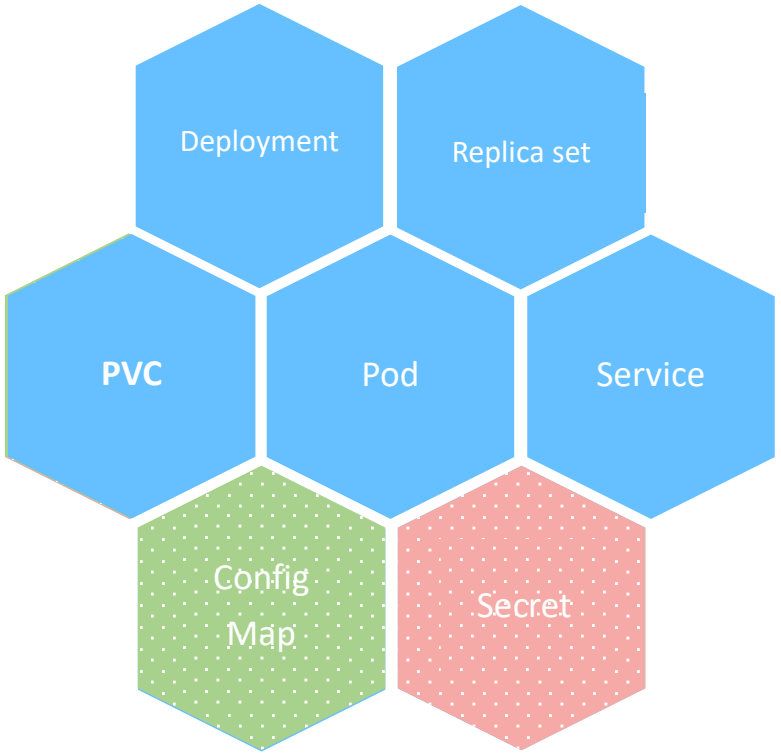
User Errors

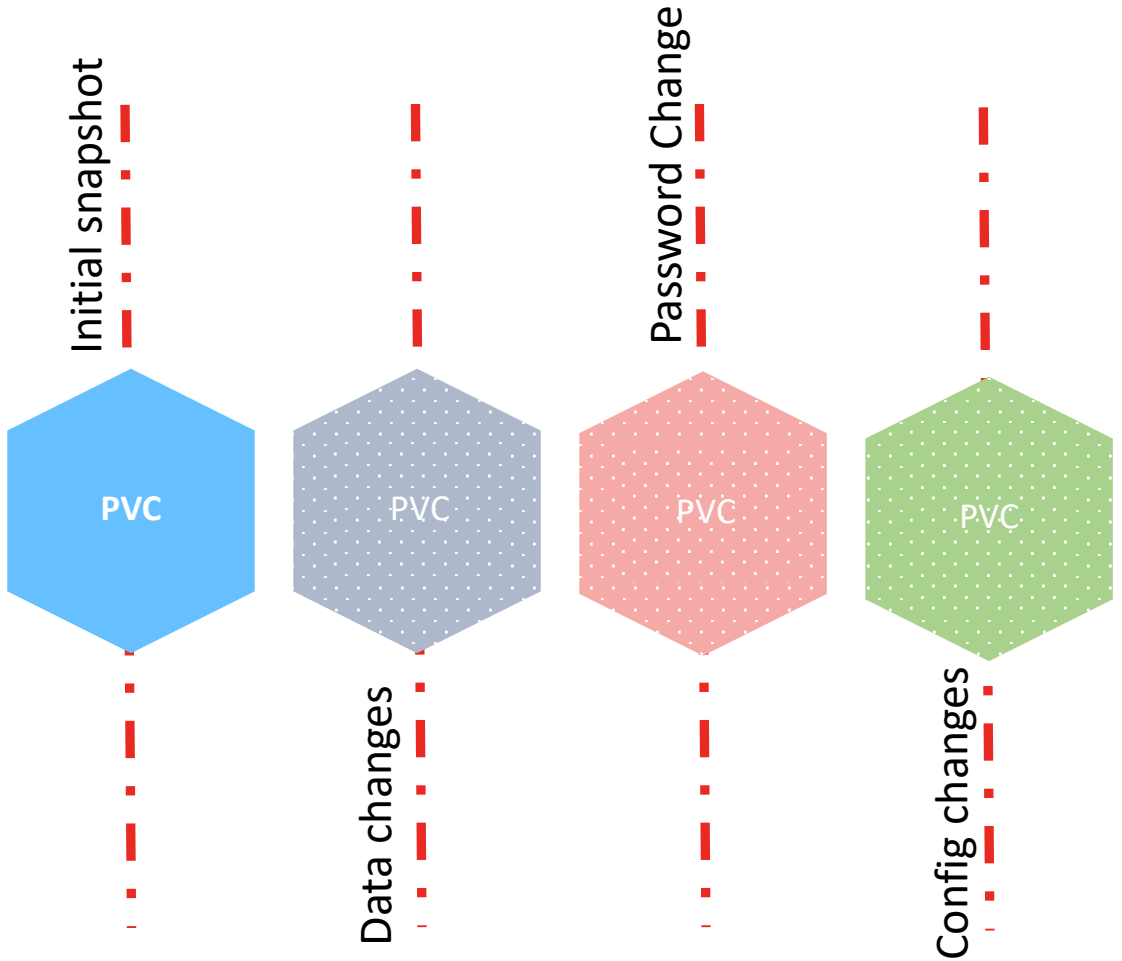
Let us talk Data Protection





Volume snapshots

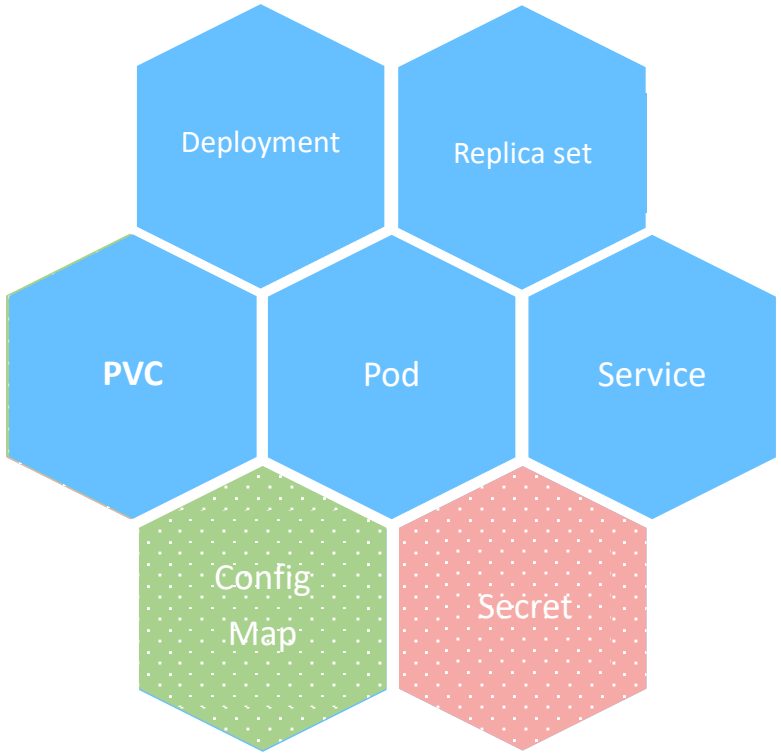




Rollback to this snapshot

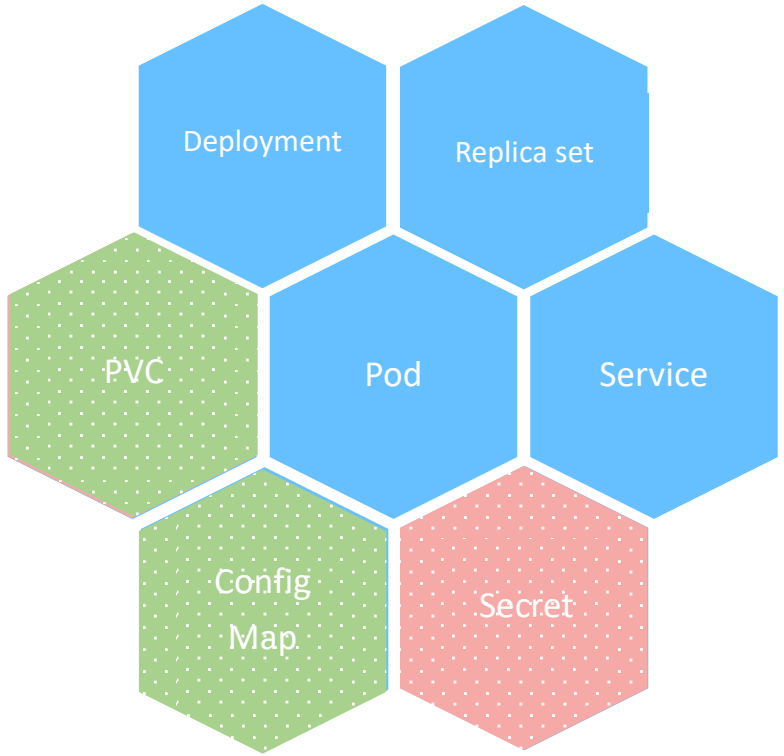
What is the problem here ?

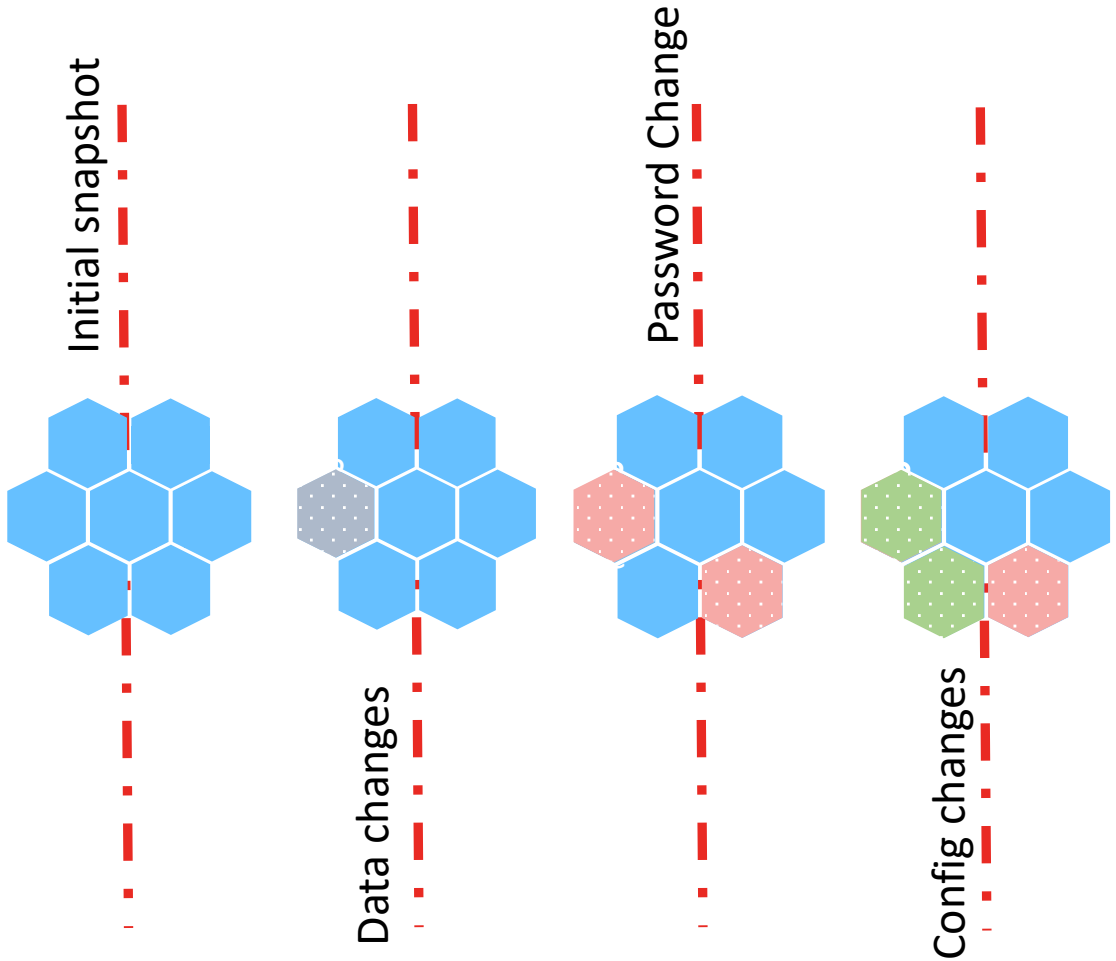
Volume snapshots



Config Drift !!!

Let us fix it …


Deployment
Replica set
PVC
Pod
Service
Config Map
Secret


Initial snapshot
Data changes
Password Change
Config changes

Recap (Data Protection)

- Snapshots and backups are not just data dumps
- Not all application have checkpoints and snapshots
- Data snapshots are prone to config drift issues
- Consistency group is a very critical construct
- Application buffers / FS page cache will need to be flushed to disk

What are we looking for…..???

Application Snapshots

Protect an entire Application, not just Storage Volumes

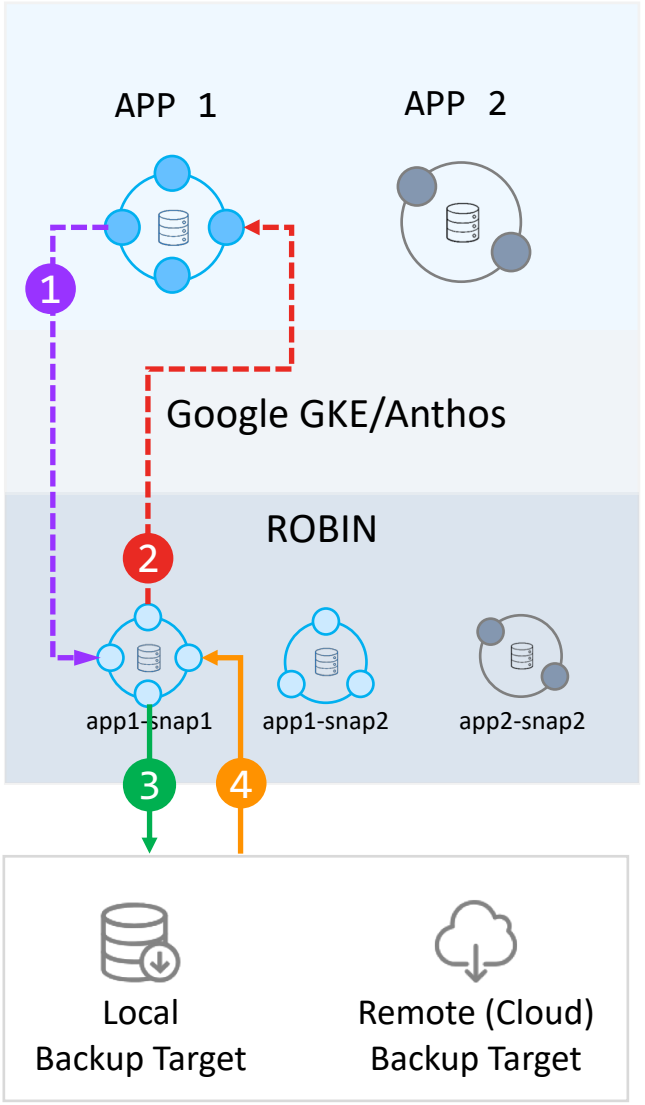
1 **Maintain periodic checkpoints of your entire app with data**

```
$ robin snapshot app1 snap1
```

1	DATA	PersistentVolumeClaims
2	CONFIG	ConfigMap, Secret, Labels, …
3	METADATA	Pods, StatefulSets, Services, …

2 **Rollback entire app+data to healthy state to recover from corruptions or user errors**

```
$ robin rollback snap1 app1
```



3 **Backup entire app+data as into external backup targets**

```
$ robin backup snap1 target
```

4 **Restore entire app+data to healthy state from catastrophic hardware and datacenter failures**

```
$ robin restore target snap1
```

› ROBIN Backups are fully self-contained

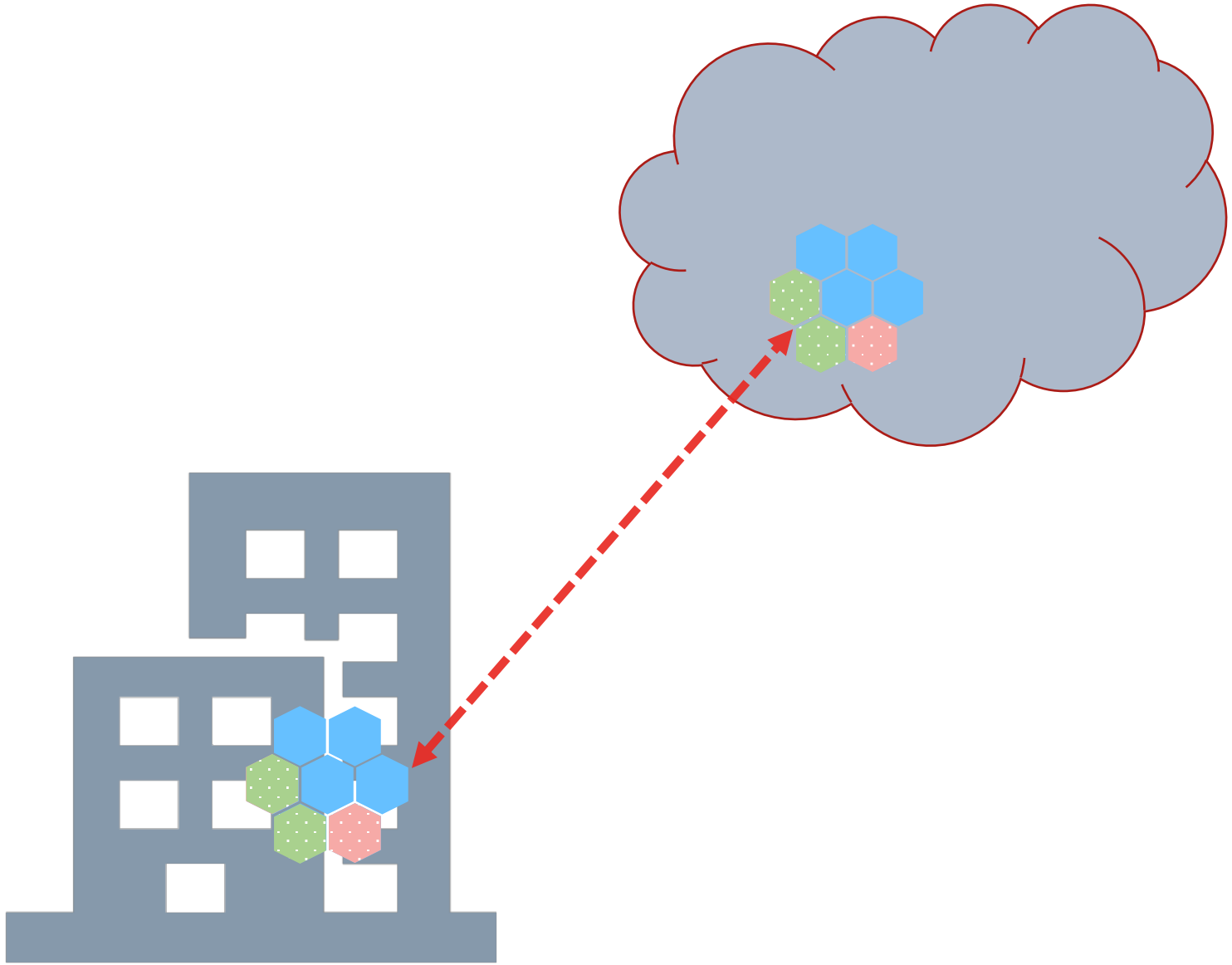
› Entire app resources can be restored in the same or different data center or cloud even if the source is completely destroyed

robin.io

Hardware / Site Failures

Application Backups

- Why do we need this?
  - Hardware refresh
  - Datacenter migration
  - Vendor lock-in
  - Performance
  - Test / Dev setups
  - Upgrade firedrills



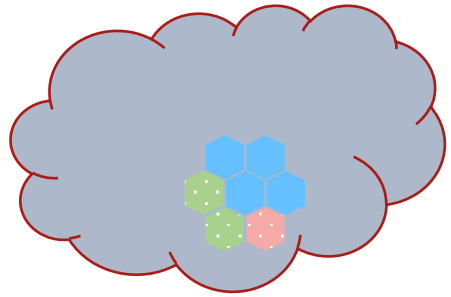
Application Backups

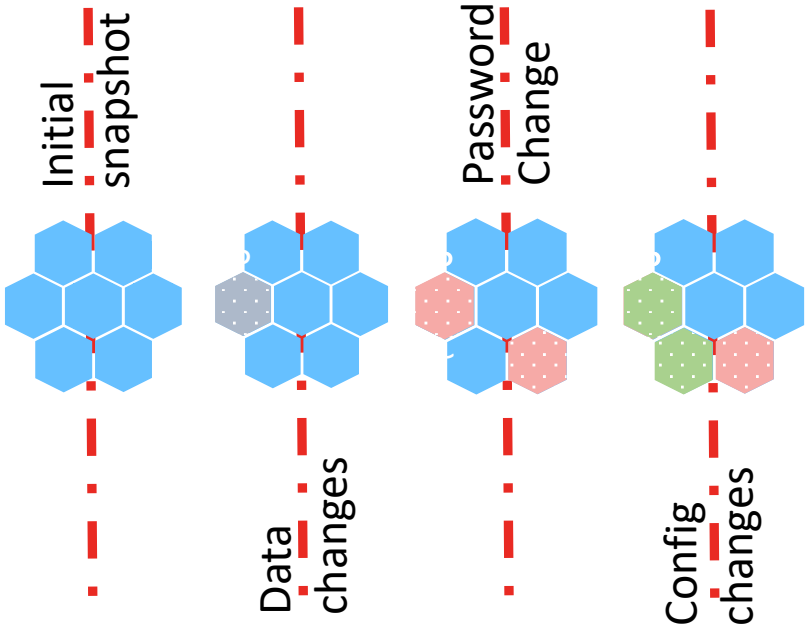
Time:

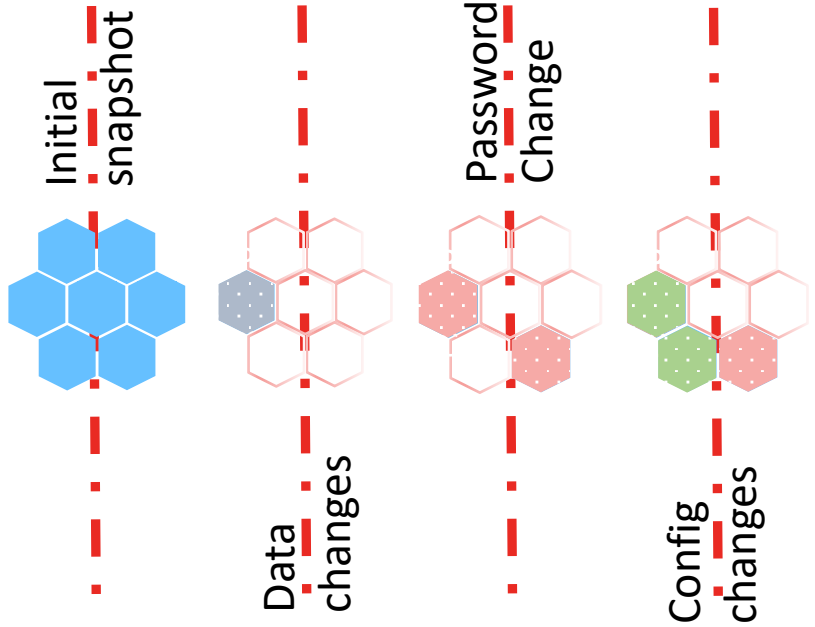
- Avoid full rehydration to Block
- Rehydrate on demand

Cost:

- Use Object store (Cheap)
- Send differentials

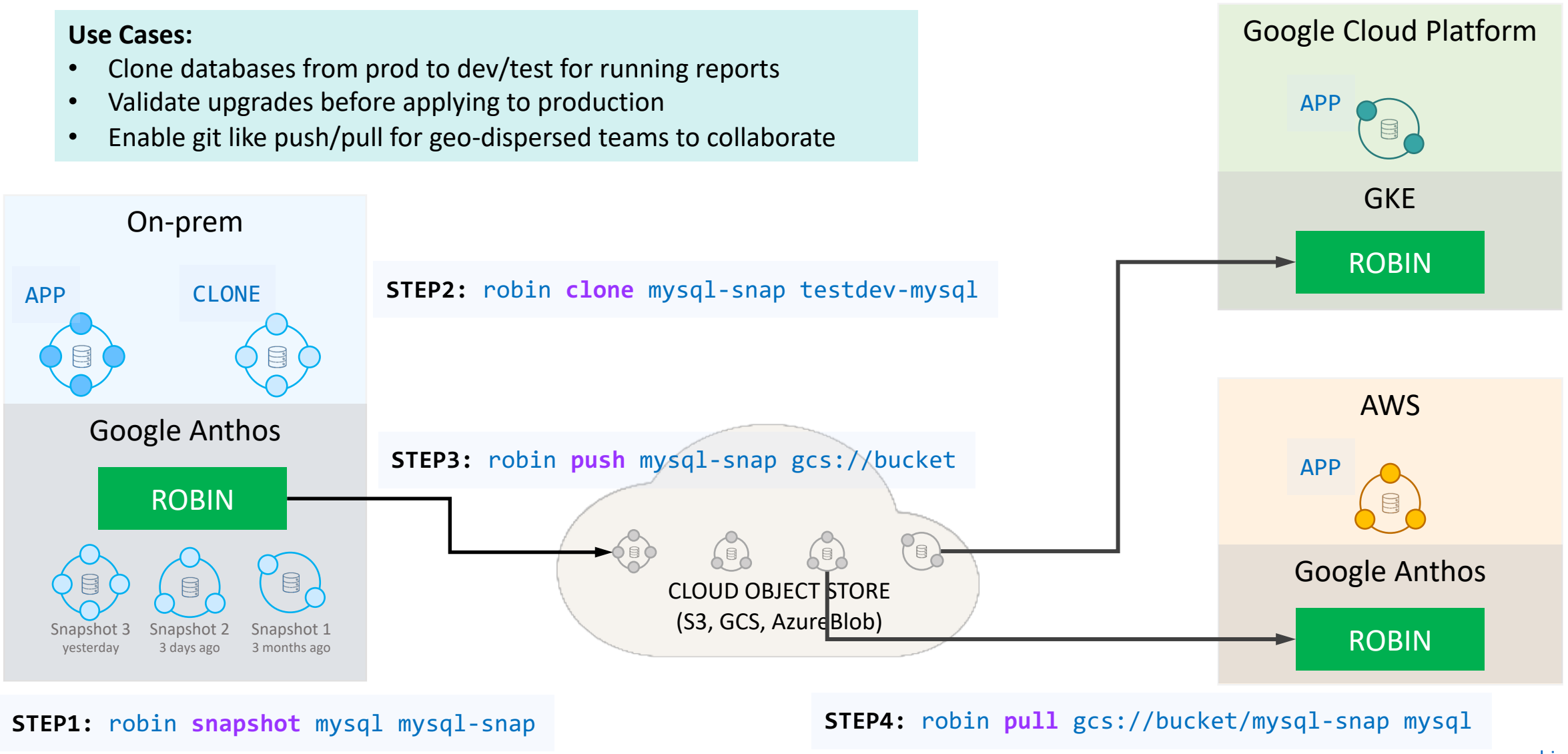






Collaborate on Applications using a Git-like workflow

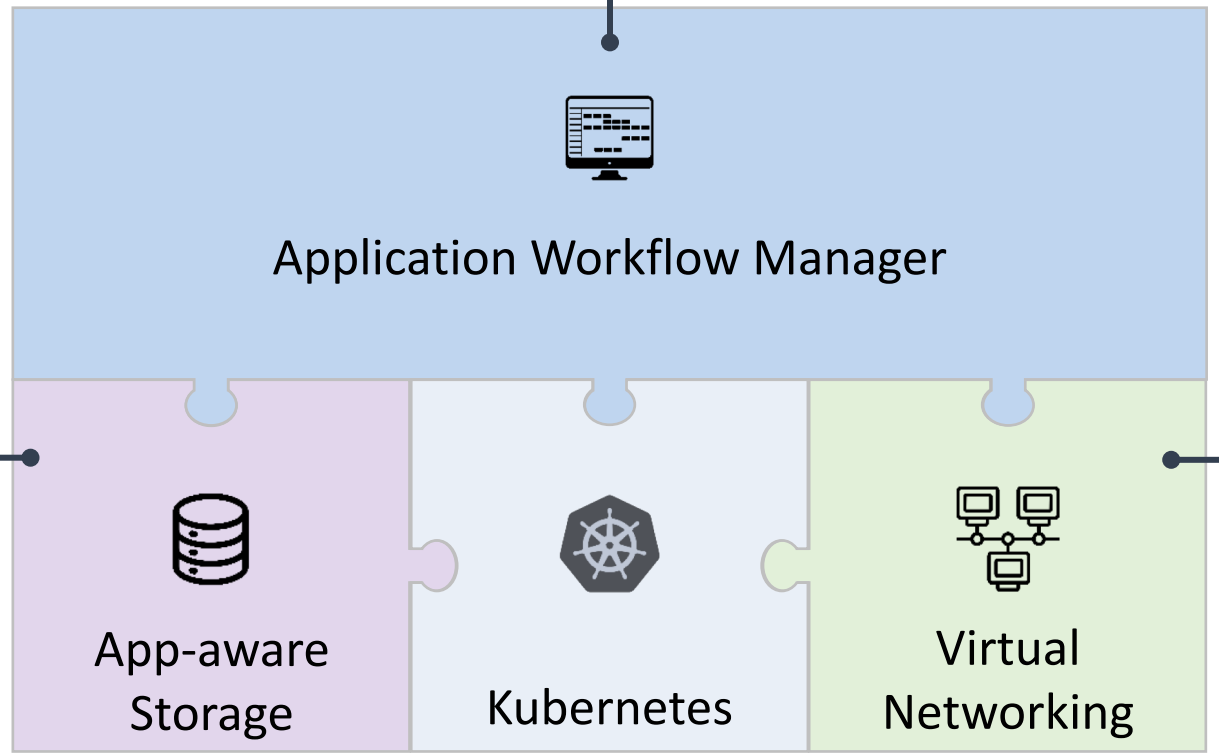
Use Cases:
- Clone databases from prod to dev/test for running reports
- Validate upgrades before applying to production
- Enable git like push/pull for geo-dispersed teams to collaborate



Robin Architecture Overview

1-click application Deploy, Snapshot, Clone, Scale, Upgrade, Backup

Application workflows configure Kubernetes, Storage & Networking



Robin's built-in enterprise-grade storage stack

Snapshots, Clones, QoS, Replication, Backup, Data rebalancing, Tiering, Thin-provisioning, Encryption, Compression

Built-in flexible networking

OVS, Calico, VLAN, Overlay networking, Persistent IPs

vm aws Azure Google Cloud Platform

Works any where

ROBIN software allows you run complex Big Data and Databases on Kubernetes

(Storage + Networking + Application Workflow Management + Kubernetes)

DEPLOYMENT PROOF POINTS

11 billion security events ingested and analyzed a day

(Elasticsearch, Logstash, Kibana, Kafka)

6 petabytes under active management in a single ROBIN cluster

(Cloudera, Impala, Kafka, Druid)

400 Oracle RAC databases managed by a single ROBIN cluster

(Oracle, Oracle RAC)

ROBIN.IO

Supercharge Kubernetes to Deliver Big Data and Databases as-a-Service

1-click Deploy, Scale, Snapshot, Clone, Upgrade, Backup, Migrate