

Containing an Elephant

Hadoop/HBase on Kubernetes

Dhiraj Hegde, Architect at Salesforce
@dhirajtxt | dhegde@salesforce.com







Immutability of Containers (Pods)



- Immutability of Containers (Pods)
- Availability and Scalability

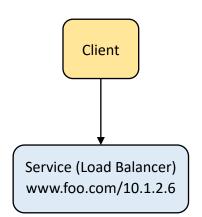


- Immutability of Containers (Pods)
- Availability and Scalability
- Consistency across clouds

Kubernetes and Stateless Applications







HTTP Server

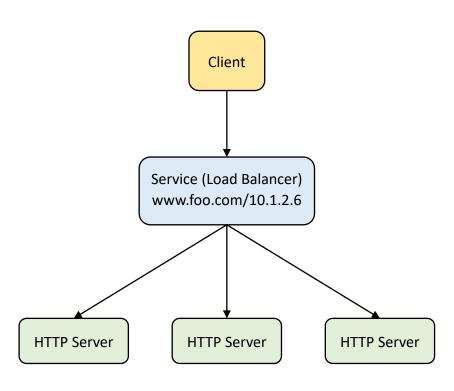
HTTP Server

HTTP Server

Kubernetes and Stateless Applications



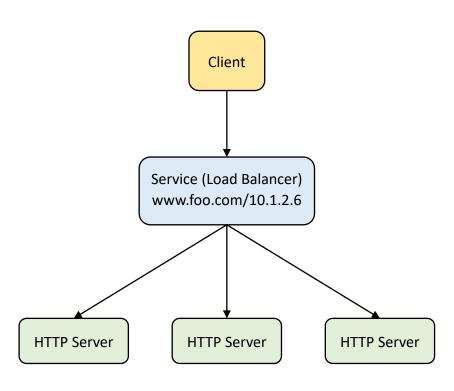




Kubernetes and Stateless Applications





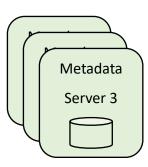


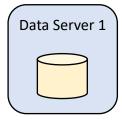


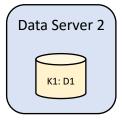
Client

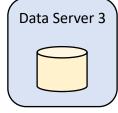
DNS

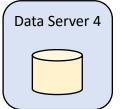
Data Server 1: 10.2.3.4 Data Server 2: 10.4.5.6



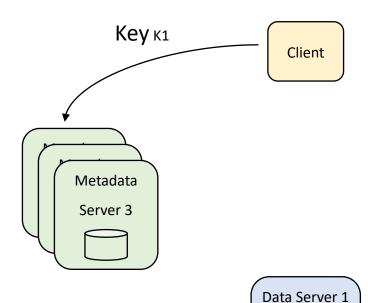










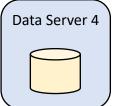


DNS

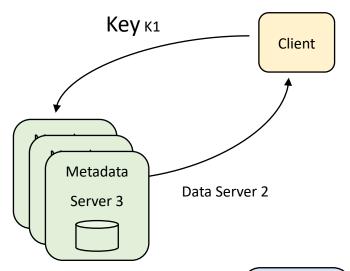
Data Server 1: 10.2.3.4 Data Server 2: 10.4.5.6

Data Server 2 Data Server 3

K1: D1





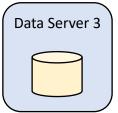


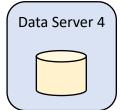
DNS

Data Server 1: 10.2.3.4 Data Server 2: 10.4.5.6

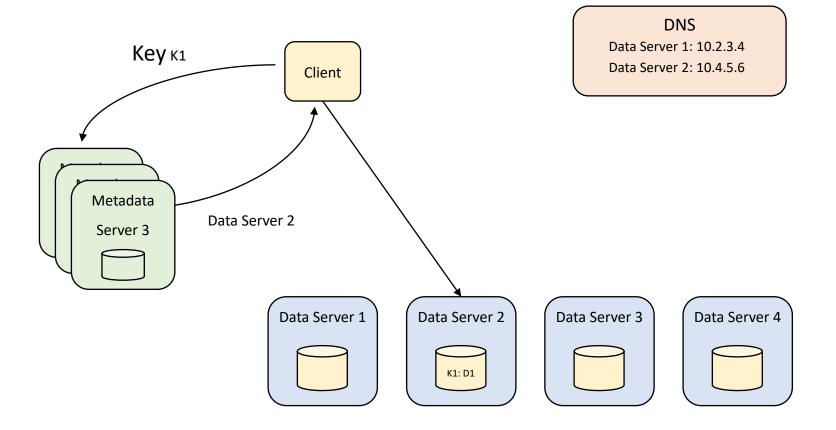
Data Server 2 Data Server 1

K1: D1

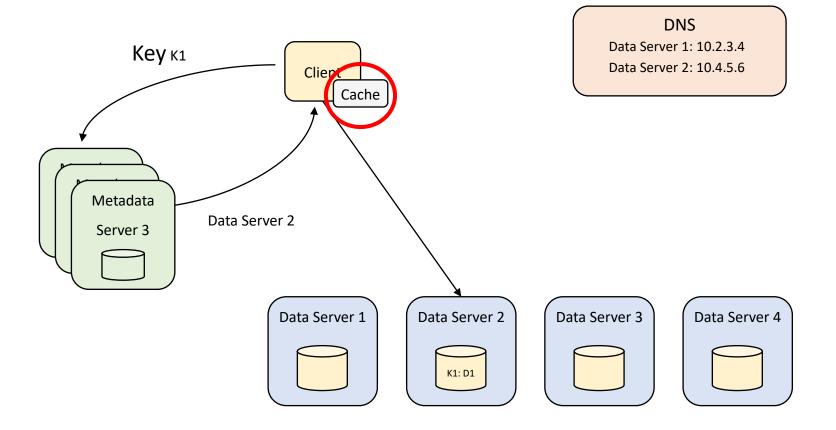




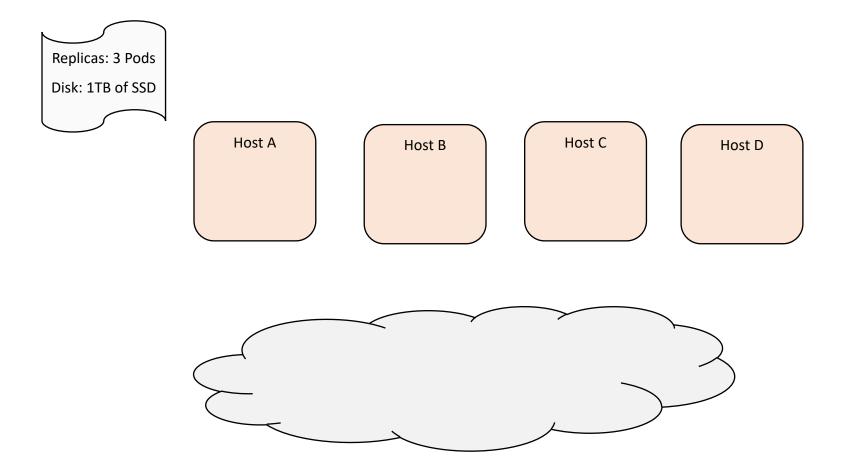




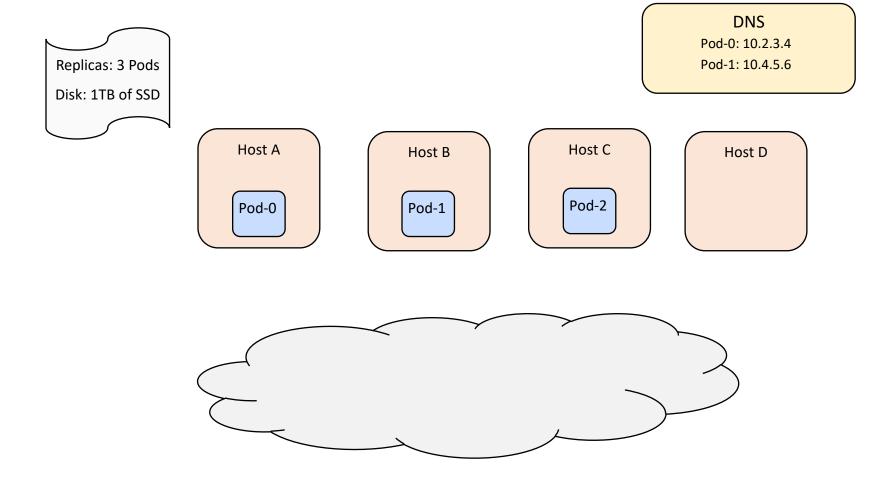




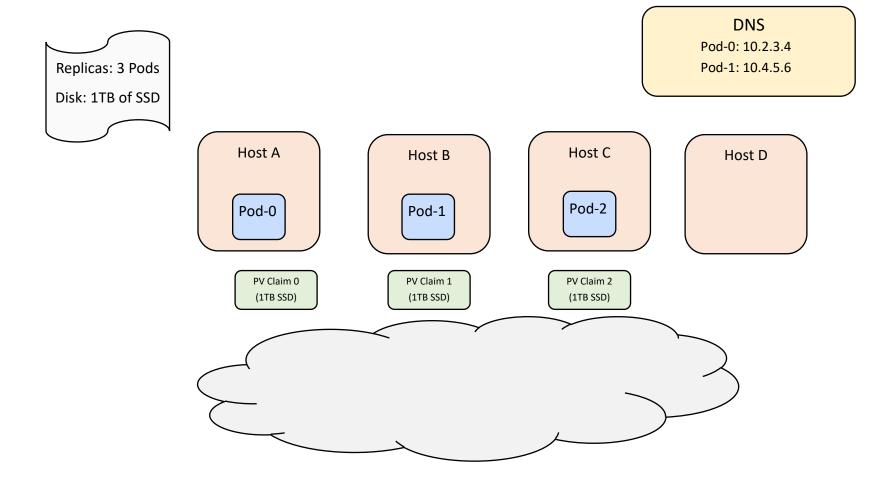




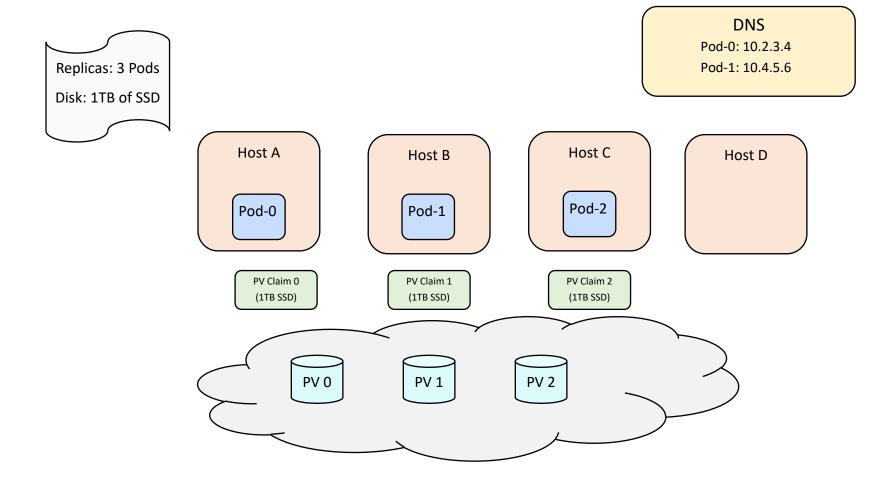










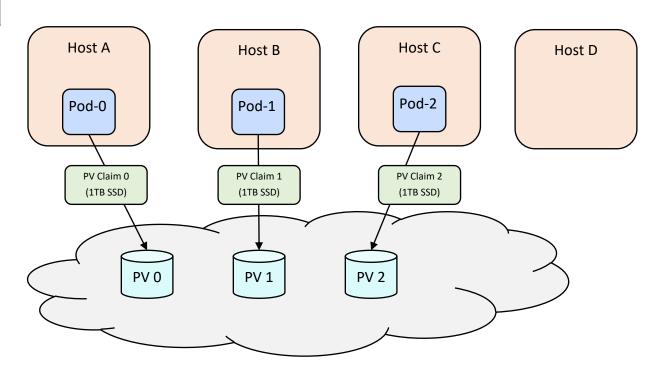




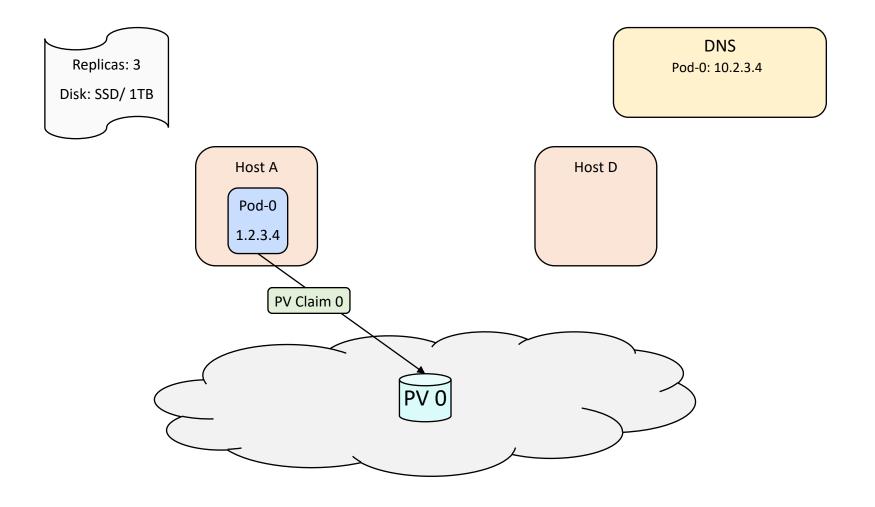


DNS

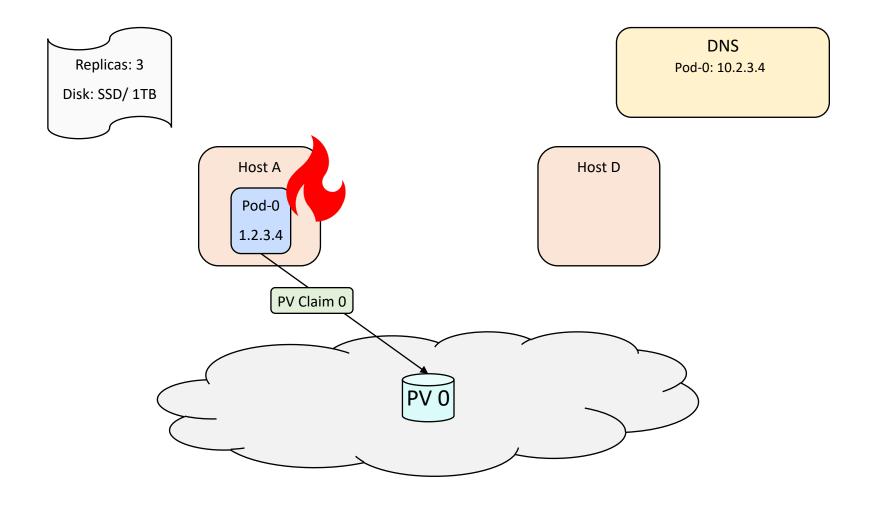
Pod-0: 10.2.3.4 Pod-1: 10.4.5.6



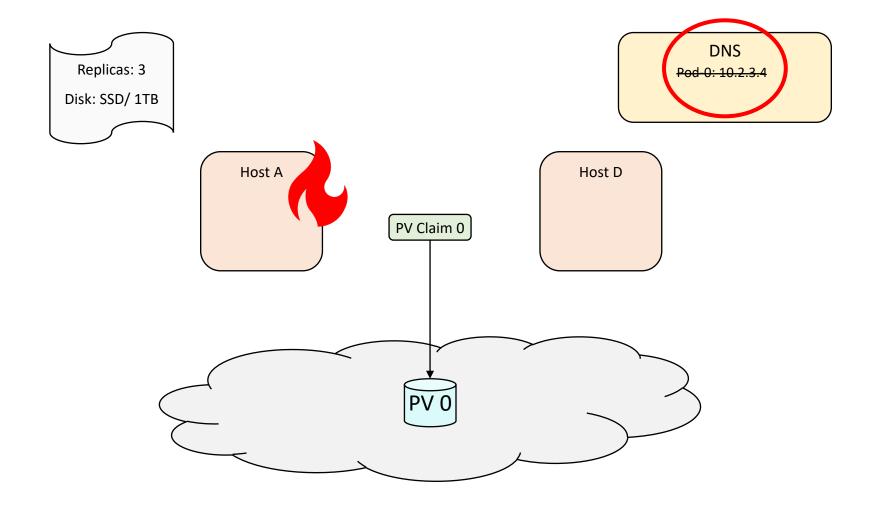




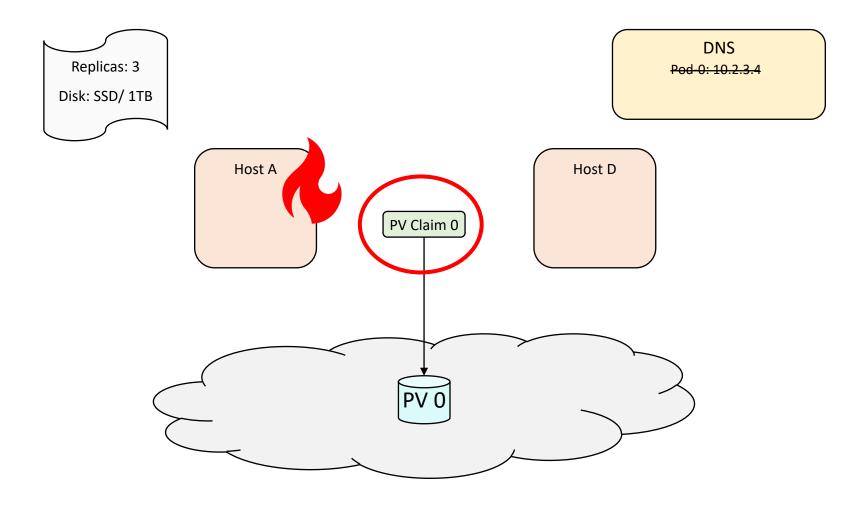




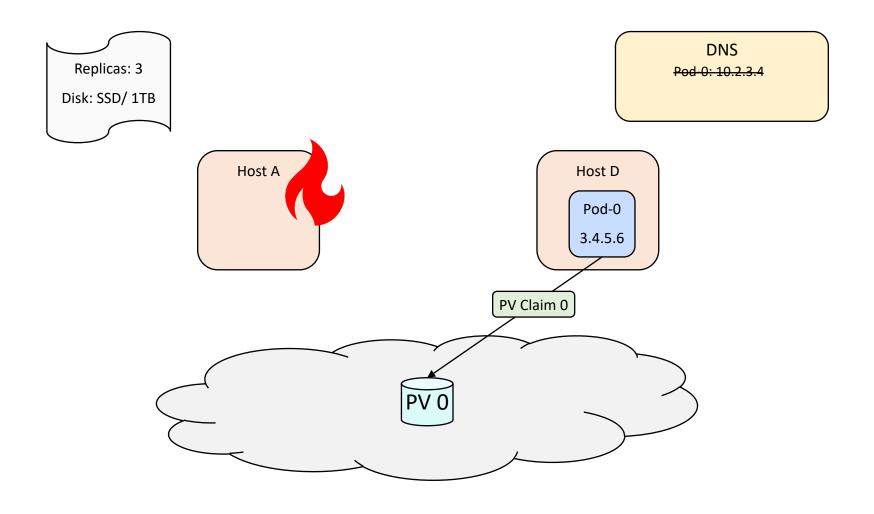




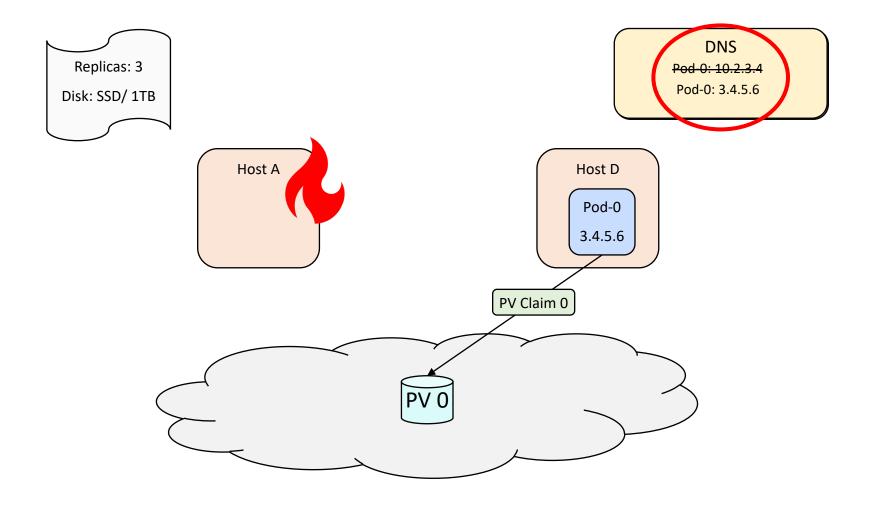




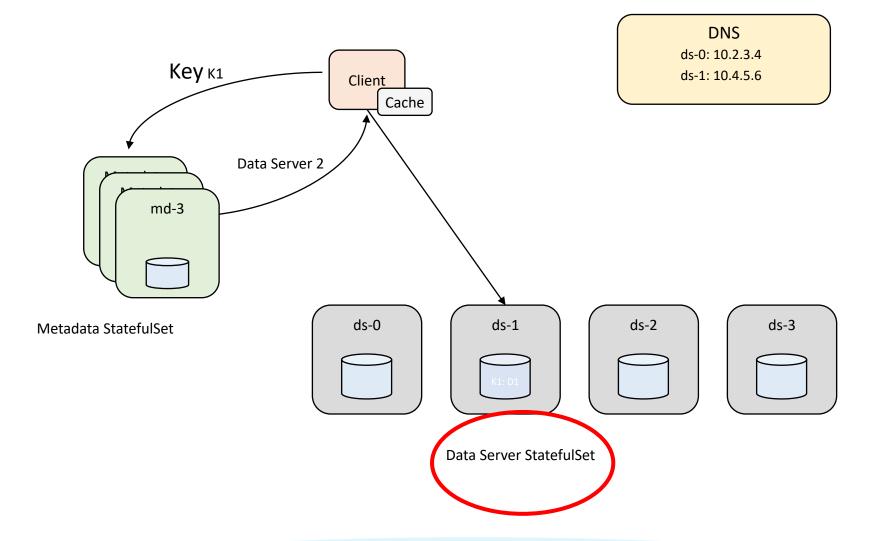




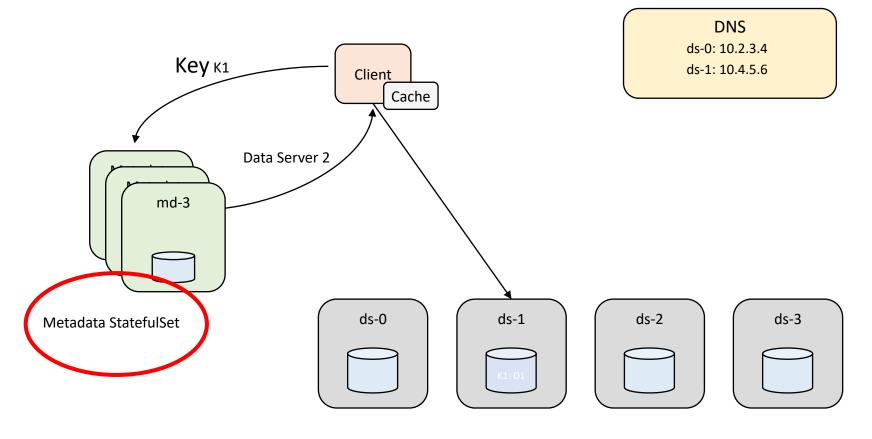












Data Server StatefulSet



STS Controller

Pod-5

Pod-4

Pod-3 v1 Pod-2 v1

Pod-1 v1



STS Controller

Pod-5 v2

Pod-4

Pod-3 v1 Pod-2 v1 Pod-1 v1



STS Controller

 $\begin{array}{c}
\mathsf{Pod-5} \\
\mathsf{v2}
\end{array}
\left(\begin{array}{c}
\mathsf{Pod-4} \\
\mathsf{v2}
\end{array}\right)$

 $\begin{array}{c}
\mathsf{Pod-3} \\
\mathsf{v1}
\end{array}
\qquad
\begin{array}{c}
\mathsf{Pod-2} \\
\mathsf{v1}
\end{array}
\qquad
\begin{array}{c}
\mathsf{Pod-1} \\
\mathsf{v1}
\end{array}
\qquad
\begin{array}{c}
\mathsf{Pod-0} \\
\mathsf{v1}
\end{array}$



STS Controller

 $\begin{array}{c}
\mathsf{Pod-5} \\
\mathsf{v2}
\end{array}
\left(\begin{array}{c}
\mathsf{Pod-4} \\
\mathsf{v2}
\end{array}\right)
\left(\begin{array}{c}
\mathsf{Pod-3} \\
\mathsf{v2}
\end{array}\right)$

 $\begin{array}{c}
 \text{Pod-2} \\
 \text{v1}
 \end{array}
 \begin{array}{c}
 \text{Pod-1} \\
 \text{v1}
 \end{array}
 \begin{array}{c}
 \text{Pod-0} \\
 \text{v1}
 \end{array}$



STS Controller

Pod-5

Pod-4 v2 Pod-3 v2

Pod-2 v2

Pod-1

v1



STS Controller

Pod-5

Pod-4 v2 Pod-3 v2 Pod-2 v2 Pod-1 v2



STS Controller

Pod-5

Pod-4 v2 Pod-3 v2 Pod-2 v2 Pod-1 v2

Custom Controller - Batching



STS Controller

Pod-5 v1 Pod-4 v1

Pod-3 v1 Pod-2 v1 Pod-1

v1

v1

Pod-0

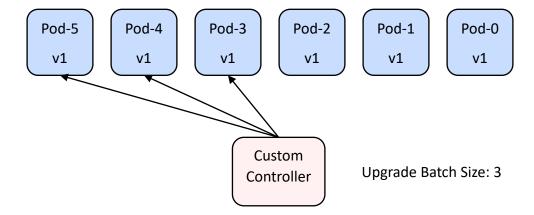
Custom Controller

Upgrade Batch Size: 3

Custom Controller - Batching



STS Controller



Custom Controller - Batching



STS Controller

Pod-2

v1

Pod-1

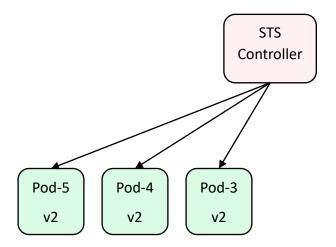
v1

Pod-0 v1

Custom Controller

Upgrade Batch Size: 3





Pod-2

v1

Pod-1

v1

Pod-0 v1

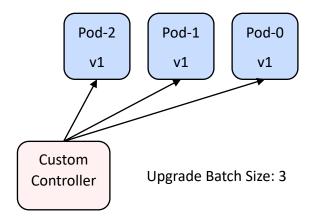
Custom Controller

Upgrade Batch Size: 3



STS Controller

 $\begin{pmatrix} \mathsf{Pod}\text{-}\mathsf{5} \\ \mathsf{v2} \end{pmatrix} \begin{pmatrix} \mathsf{Pod}\text{-}\mathsf{4} \\ \mathsf{v2} \end{pmatrix} \begin{pmatrix} \mathsf{Pod}\text{-}\mathsf{3} \\ \mathsf{v2} \end{pmatrix}$





STS Controller

Pod-5

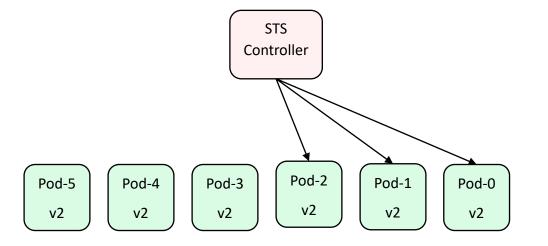
Pod-4 v2

Pod-3 v2

Custom Controller

Upgrade Batch Size: 3





Custom Controller

Upgrade Batch Size: 3



STS Controller

Disruption Budget: 1



Pod-4



Pod-2 v1



Pod-0 v1





Pod-5 v1 Pod-4 v1 Pod-3 v1 Pod-2 v1 Pod-1 v1

Pod-0



STS Controller

Disruption Budget: 1



Pod-4 v1







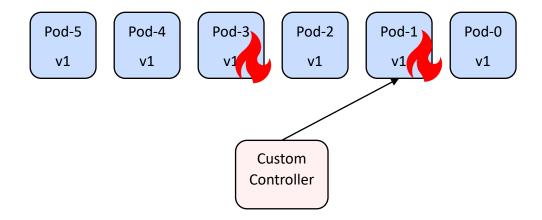
Pod-0 v1

Custom Controller





Disruption Budget: 1







Disruption Budget: 1



Pod-5 v1



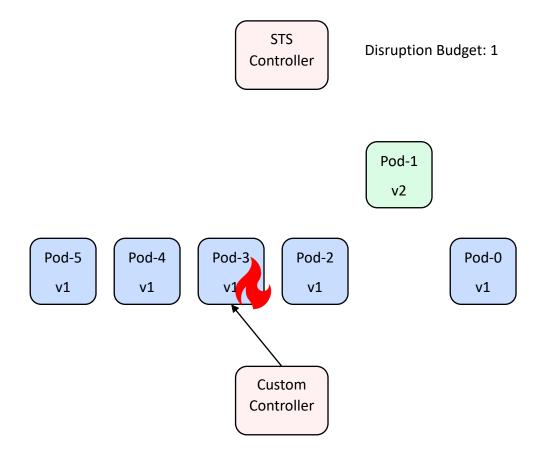




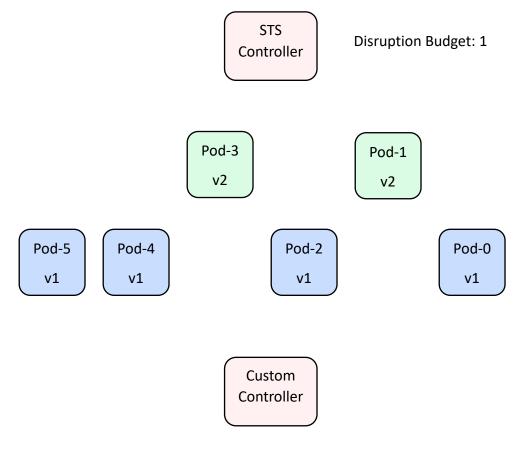


Custom Controller

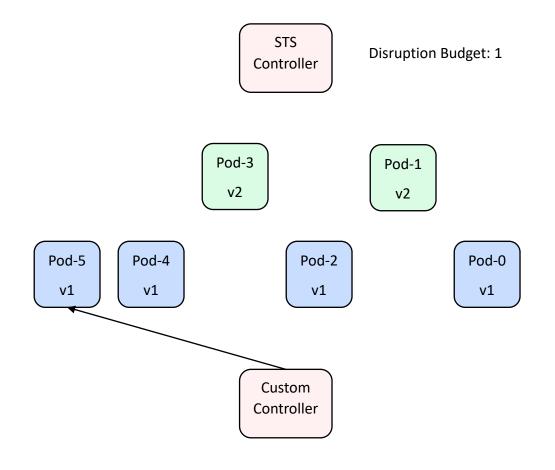




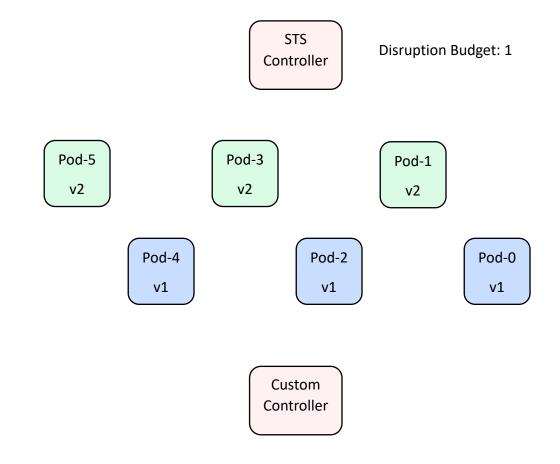














STS Controller

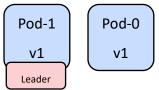
Pod-2 v1 Leader Pod-1 v1

Pod-0 v1



STS Controller

Pod-2 v2





STS Controller

Pod-2 v2 Pod-1 v2

> Pod-0 v1 Leader

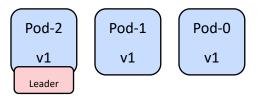


STS Controller

Pod-2 v2 Pod-1 v2 Pod-0 v2

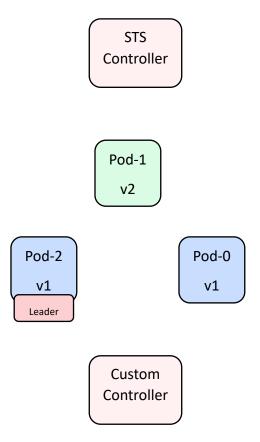


STS Controller



Custom Controller







STS Controller

Pod-2 v1 Leader

> Custom Controller



STS Controller

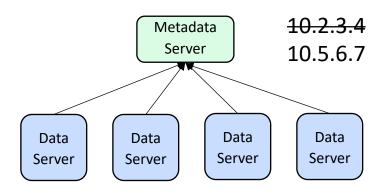
Pod-2 Pod-1 Pod-0 v2 Leader

Custom Controller

Hadoop/HBase DNS Caching Bugs



- As Pods are recreated, DNS records keep changing
- Hadoop/HBase code caches DNS resolutions, causing failures
- Impact is most severe for metadata servers



Hadoop Bugs: Interim Solution



Client

Service-0 10.1.2.3 Service-1 10.1.2.4 Service-2 10.1.2.5

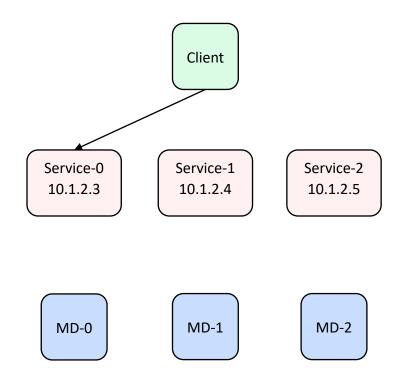
MD-0

MD-1

MD-2

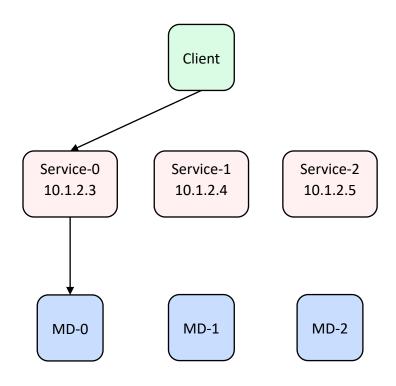
Hadoop Bugs: Interim Solution





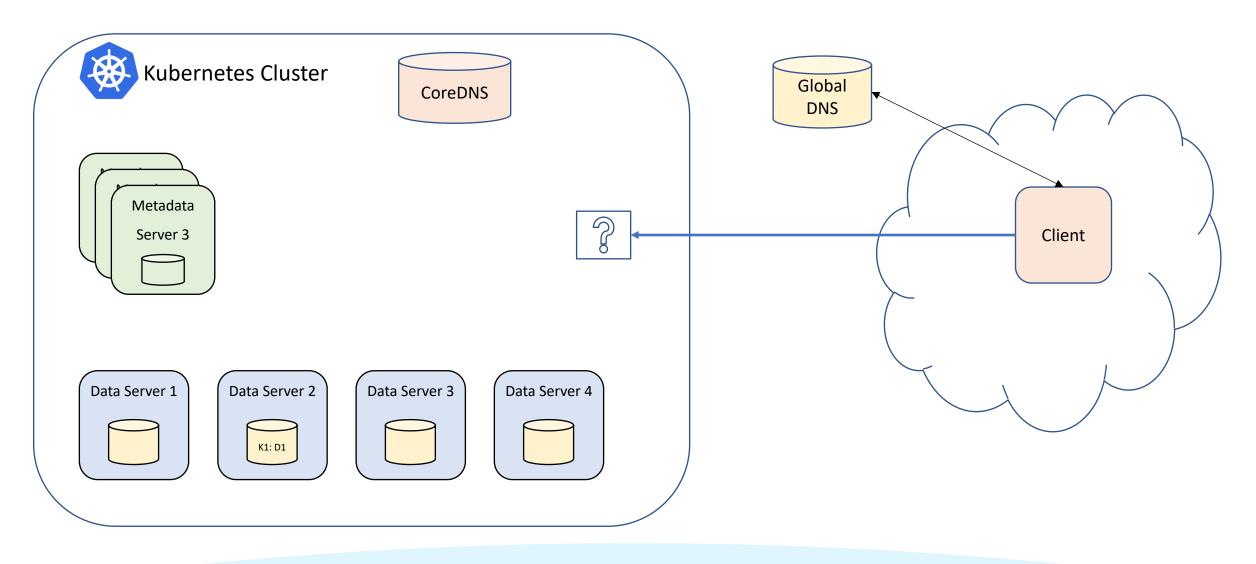
Hadoop Bugs: Interim Solution





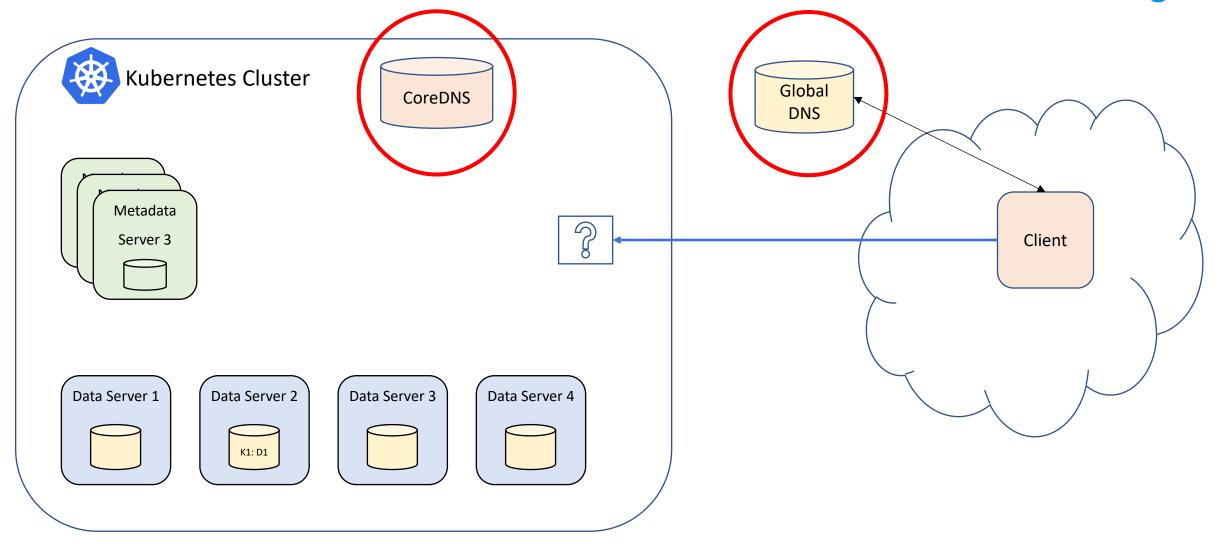
DNS Access





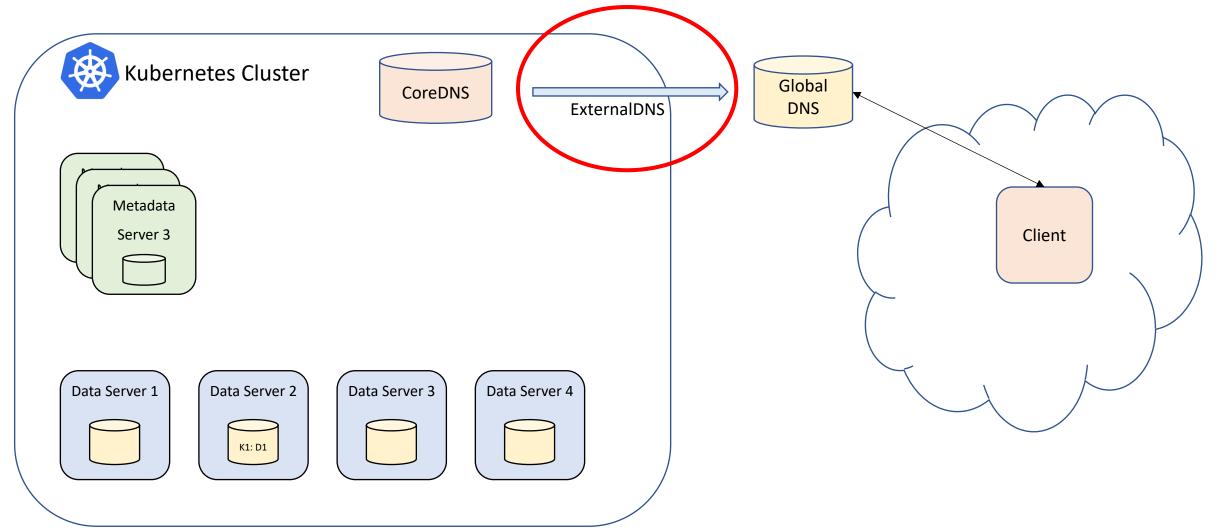
DNS Access





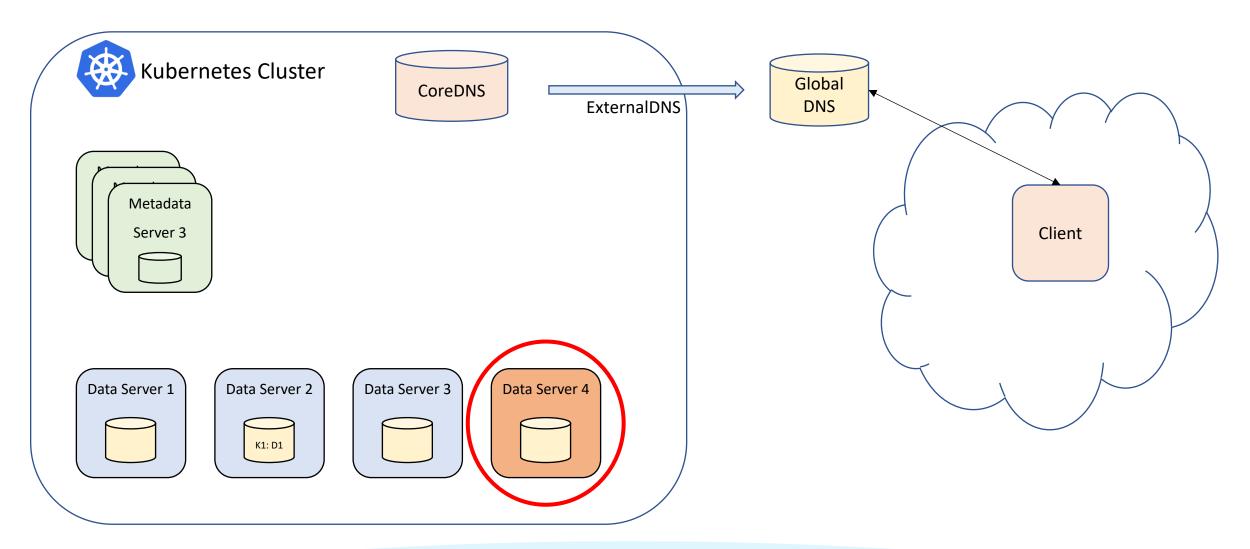
DNS Access





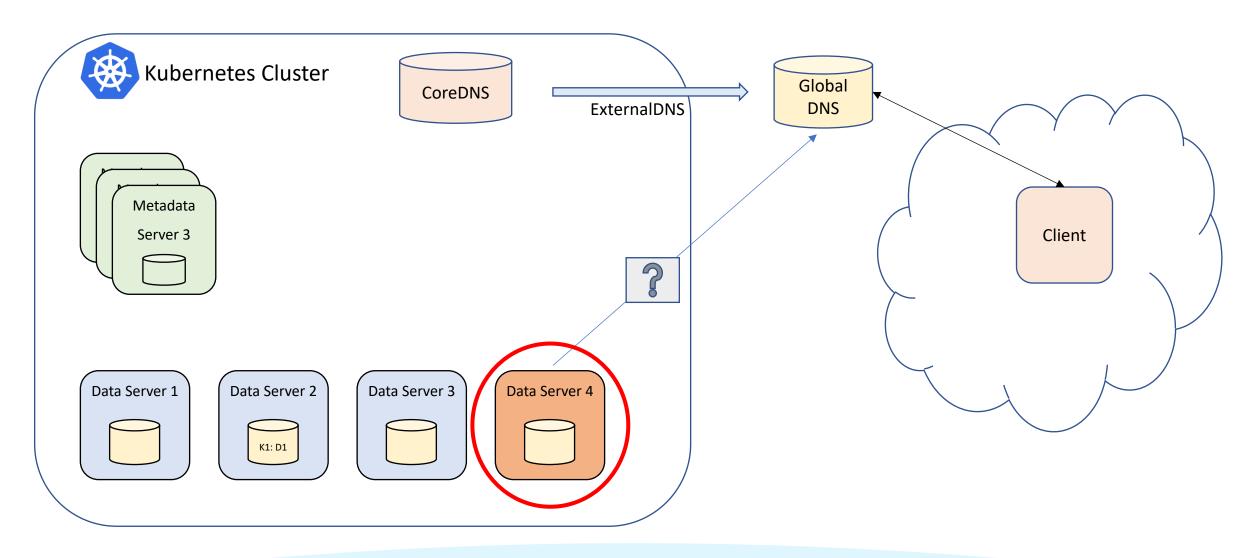
DNS Validation





DNS Validation

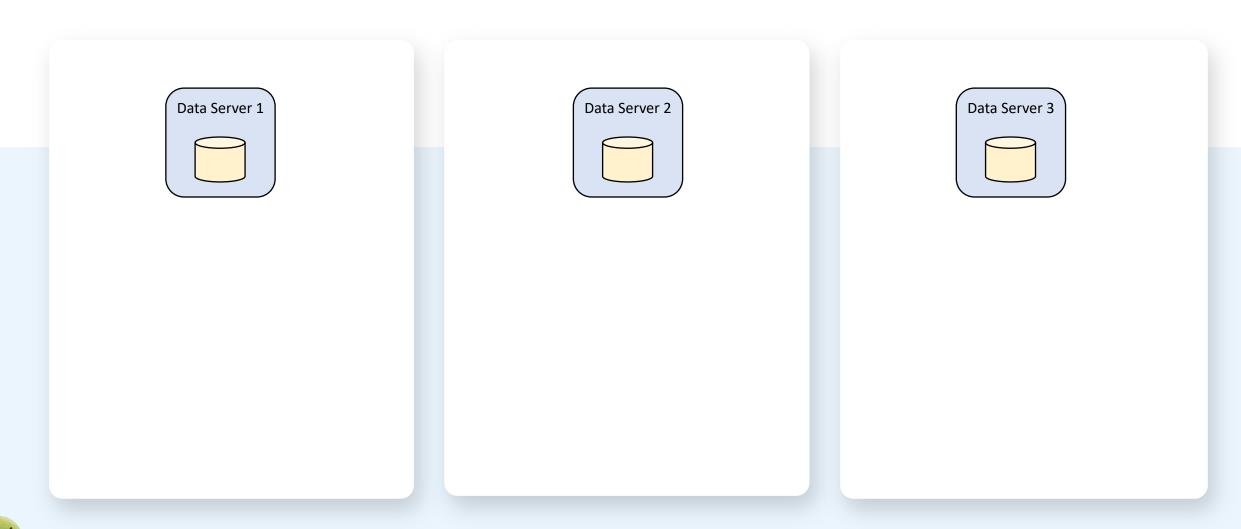




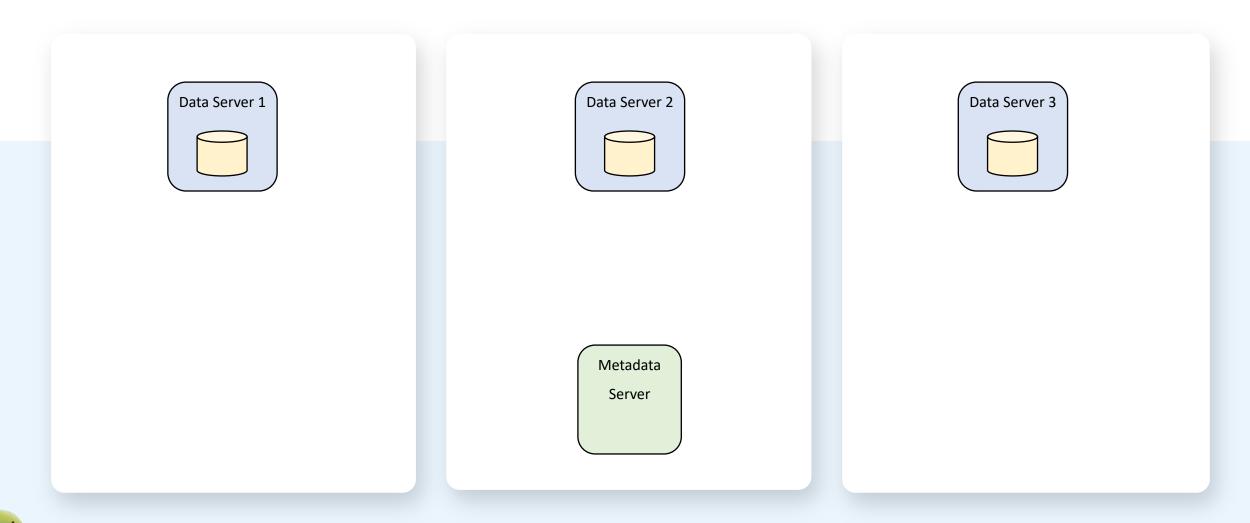


AZ 1 AZ 2 AZ 3

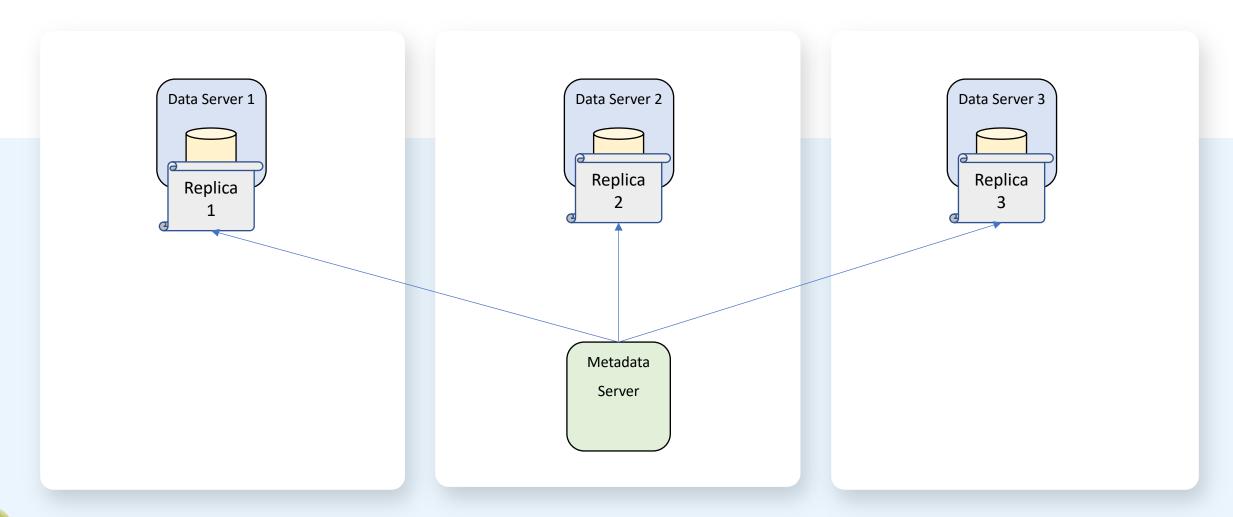






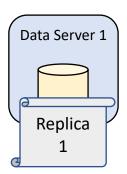




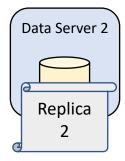




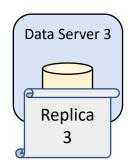
AZ 1 AZ 2 AZ 3



Metadata Server 1



Metadata
Server 2



Metadata Server 3



