

# Efficiently Manage your Hadoop and Analytics Workflow with IBM Spectrum Scale

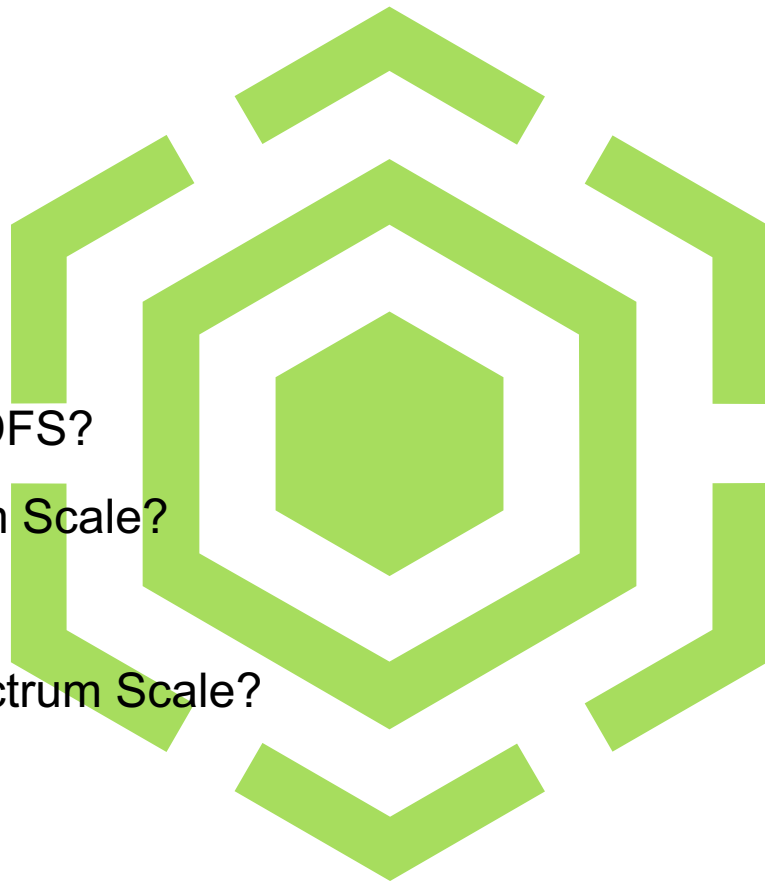
---

Andreas Koeninger  
IBM Spectrum Scale Big Data and Analytics



# Questions

- Who runs Spectrum Scale?
- Who runs an ESS?
- Who runs a native HDFS cluster?
- Who runs HDFS on Spectrum Scale?
- Who runs NFS Gateway or S3 on native HDFS?
- Who runs SMB, NFS or Object on Spectrum Scale?
- Who runs Kubernetes or OpenShift?
- Who runs Kubernetes or OpenShift on Spectrum Scale?

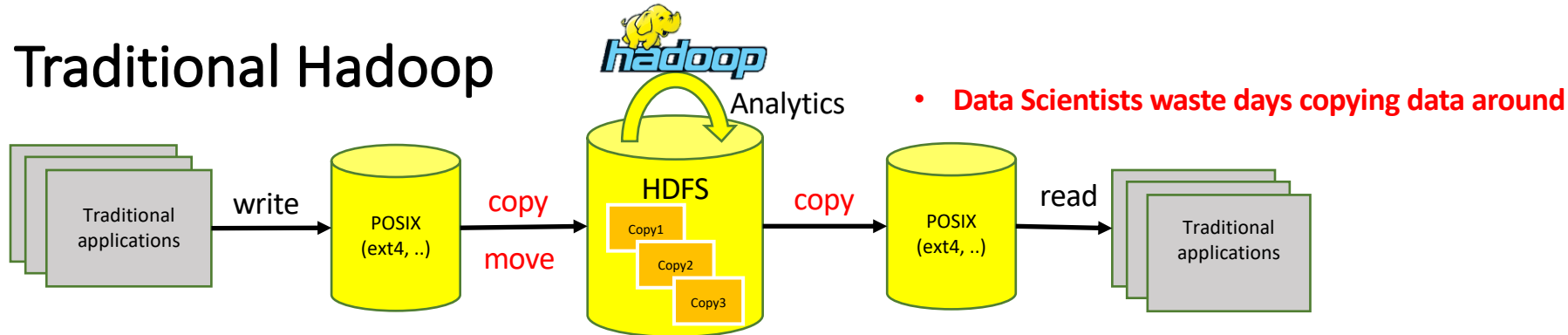


# Outline

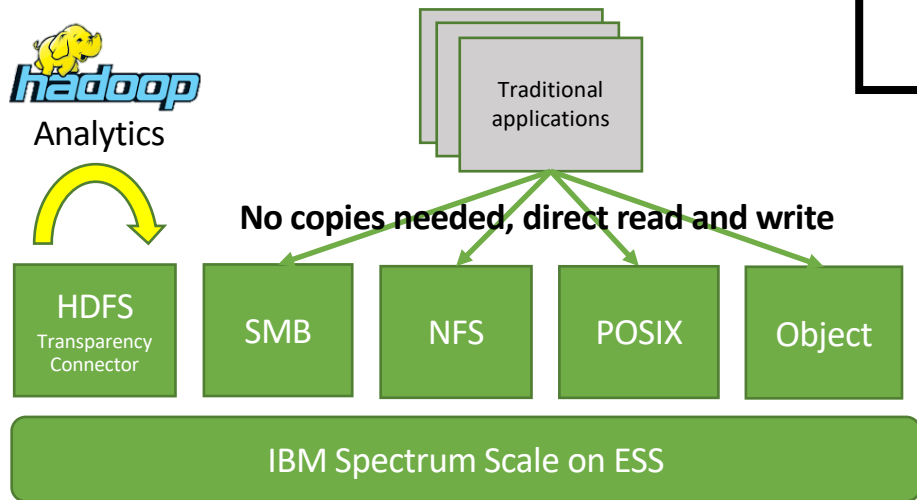
- Traditional Hadoop vs. Spectrum Scale
- Use Case 1: HDFS on Spectrum Scale
- Use Case 2: HDFS Storage Tiering & Federation
- Use Case 3: HDFS Backup
- Use Case 4: Spectrum Scale as Ingest Tier
- Use Case 5: Next-gen workloads
- Use Case 6: Disaster Recovery
- Spectrum Scale HDFS integration into CES



# Traditional Hadoop



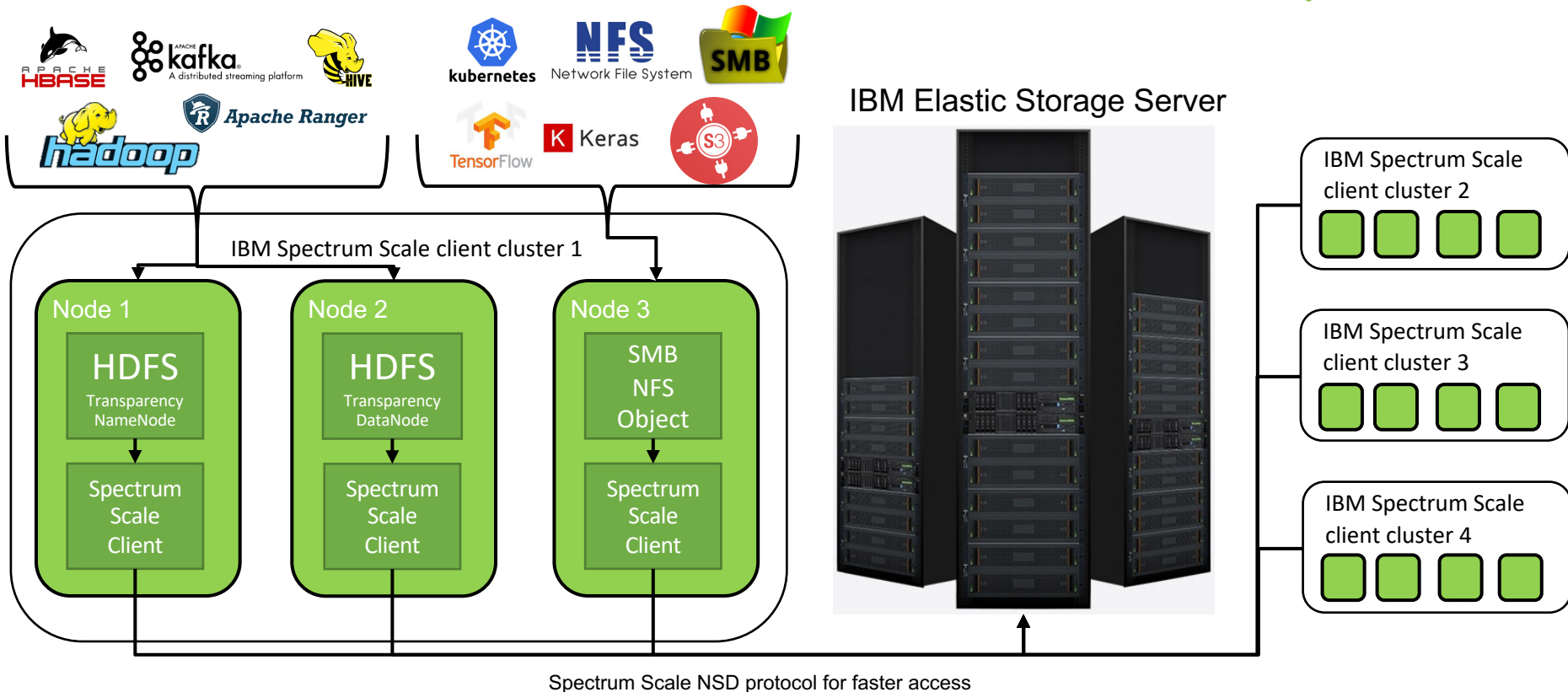
## HDFS with Spectrum Scale



- **Multiple copies** = Analytics based on stale data
- **Costly data protection:** 3-way replication  
→ 5 PB data = 15 PB storage  
(Erasure Coding in HDFS has limitations, e.g. append not supported)

- **Multi-protocol access** to the same data  
→ No copies, direct read, one version
- **Software RAID** eliminates the need for expensive 3-way replication  
→ Only 30% overhead  
→ 5 PB data = 6.5 PB storage
- **Stateless NameNode:** Fast failover, low memory footprint

# Use Case 1: HDFS on Spectrum Scale



# Use Case 1: HDFS on Spectrum Scale



## All data and metadata is stored in ESS

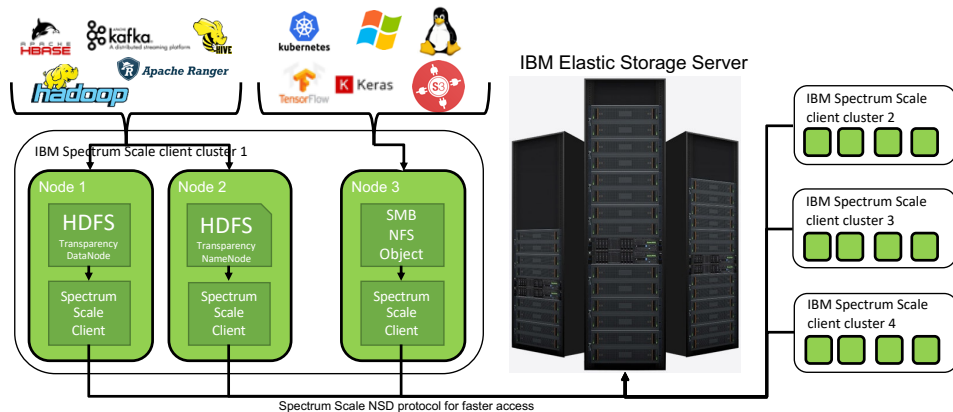
- **No additional storage** needed on DataNodes or NameNodes
- **Low memory footprint** for NameNodes which allows faster failover (only Kerberos tickets are stored in shared edits log)

## Multi-protocol access

- **Single source of truth:** Access the same data through HDFS, NFS, SMB, Object and POSIX without any copying

## IBM Spectrum Scale on ESS

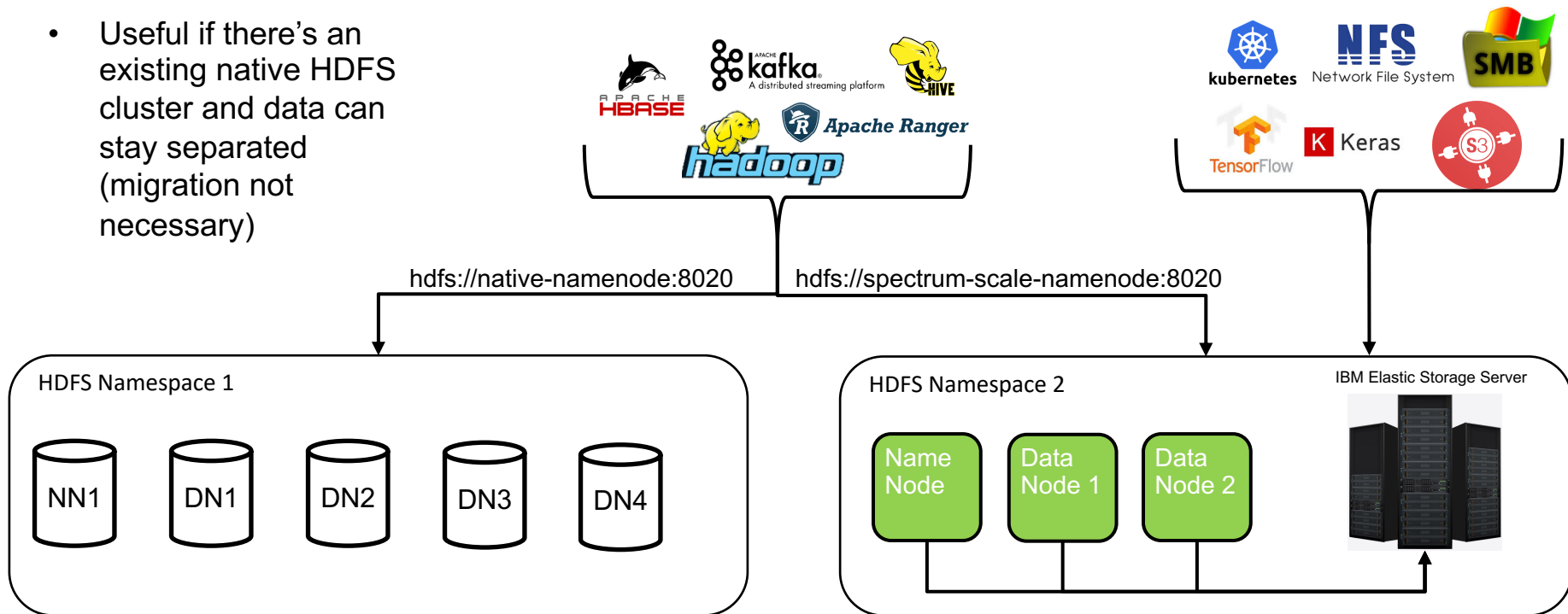
- **Erasure Coding:** Lower storage footprint, higher performance since no replication needed
- **Easy to scale:** Add more building blocks if you need more storage or bandwidth
- **Easy to manage:** GUI and REST API available, single storage system instead of hundreds of storage nodes



# Use Case 2: HDFS Storage Tiering & Federation



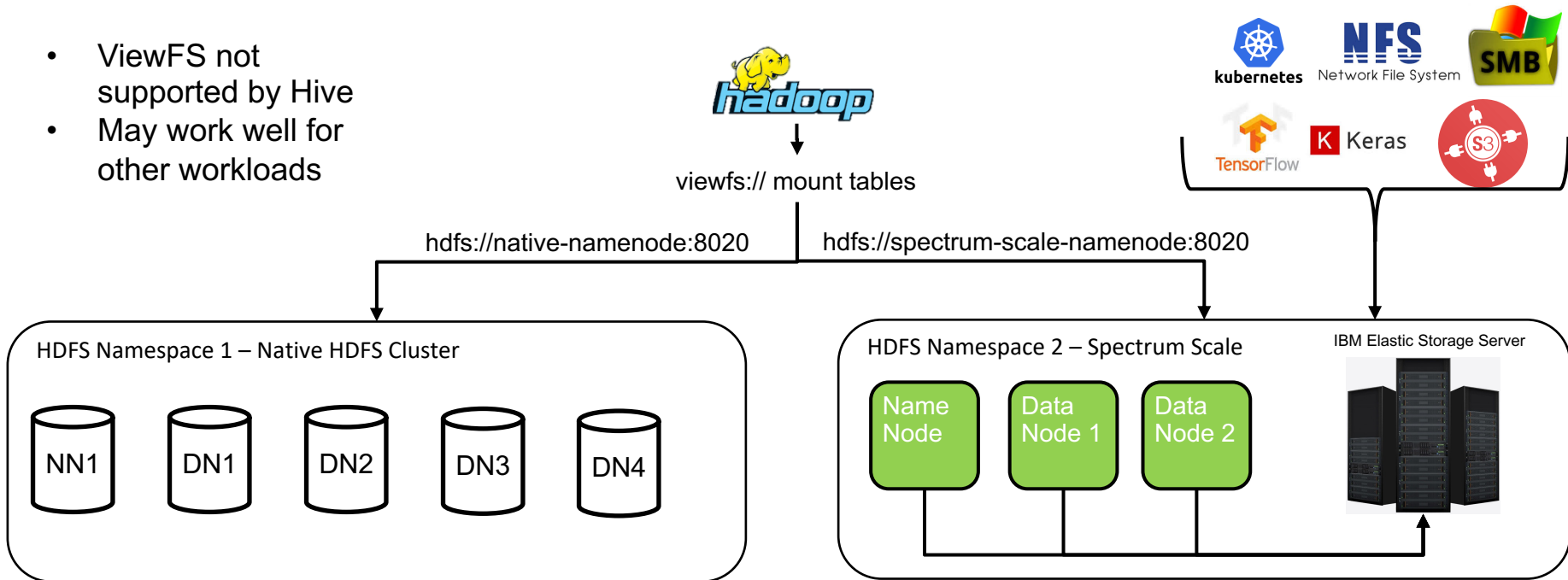
- Useful if there's an existing native HDFS cluster and data can stay separated (migration not necessary)



# Use Case 2: HDFS Storage Tiering & Federation



- ViewFS not supported by Hive
- May work well for other workloads



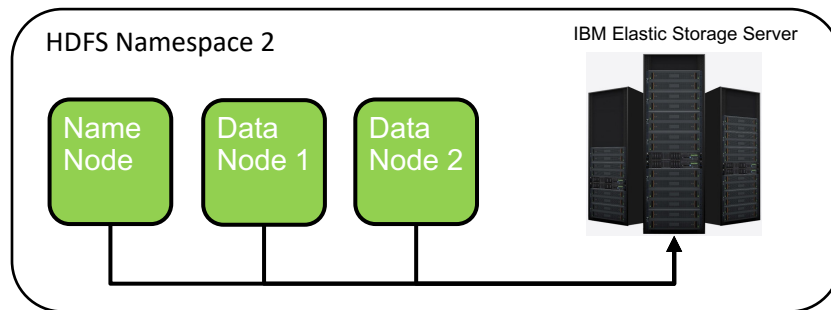
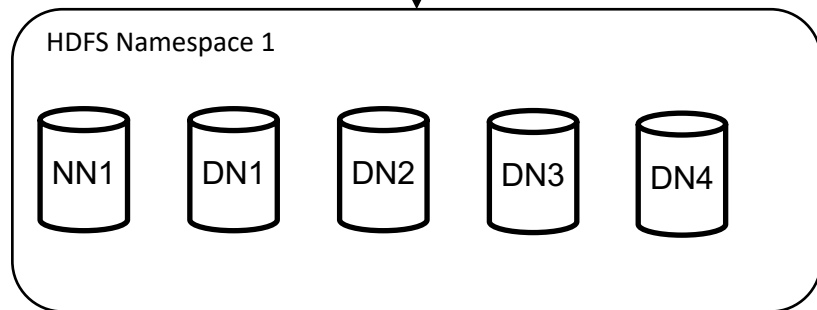


# Use Case 3: HDFS Backup



## Sample Backup Flow:

1. Applications write HDFS data to native Cluster
2. Admin uses distcp to copy HDFS data to Spectrum Scale  
(Note: distcp runs on top of YARN, so a client cluster is needed)

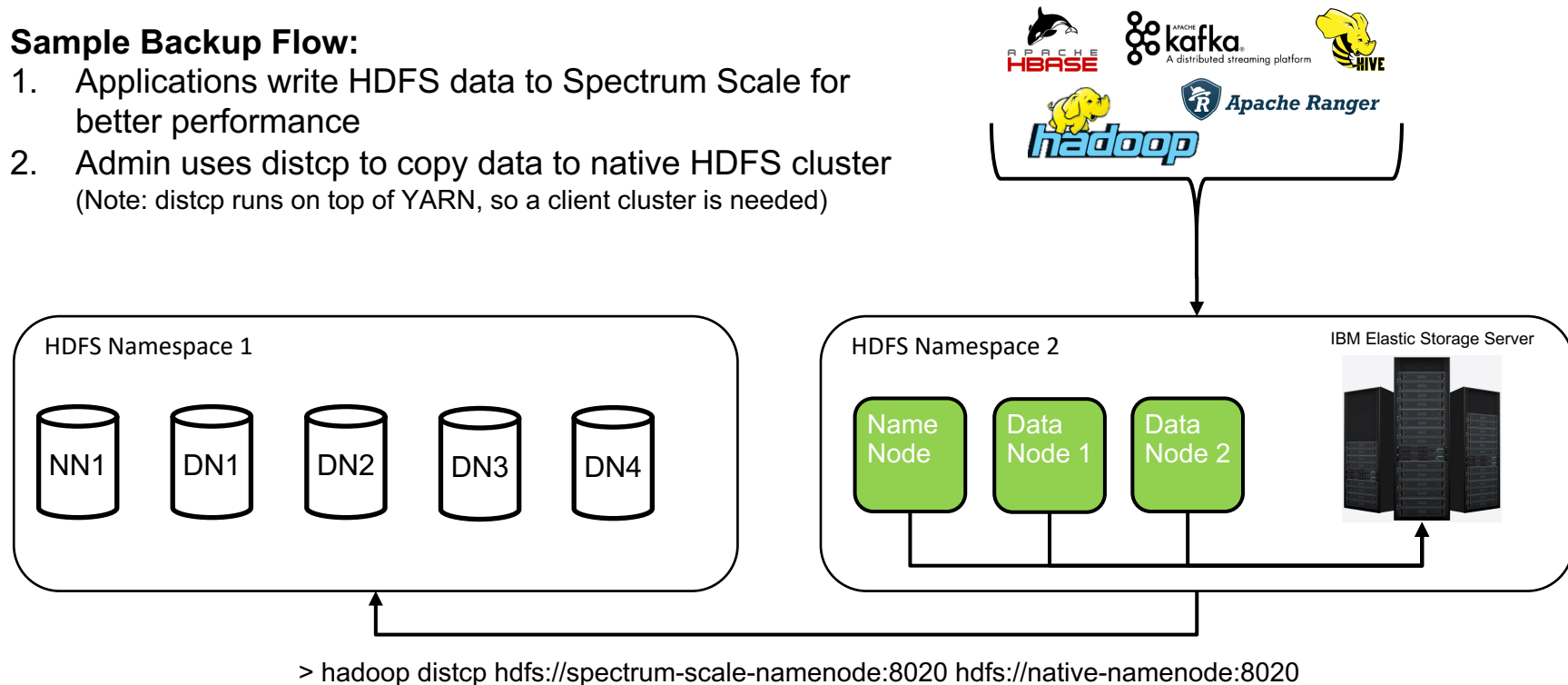


```
> hadoop distcp hdfs://native-namenode:8020 hdfs://spectrum-scale-namenode:8020
```

# Use Case 3: HDFS Backup

## Sample Backup Flow:

1. Applications write HDFS data to Spectrum Scale for better performance
2. Admin uses distcp to copy data to native HDFS cluster  
(Note: distcp runs on top of YARN, so a client cluster is needed)

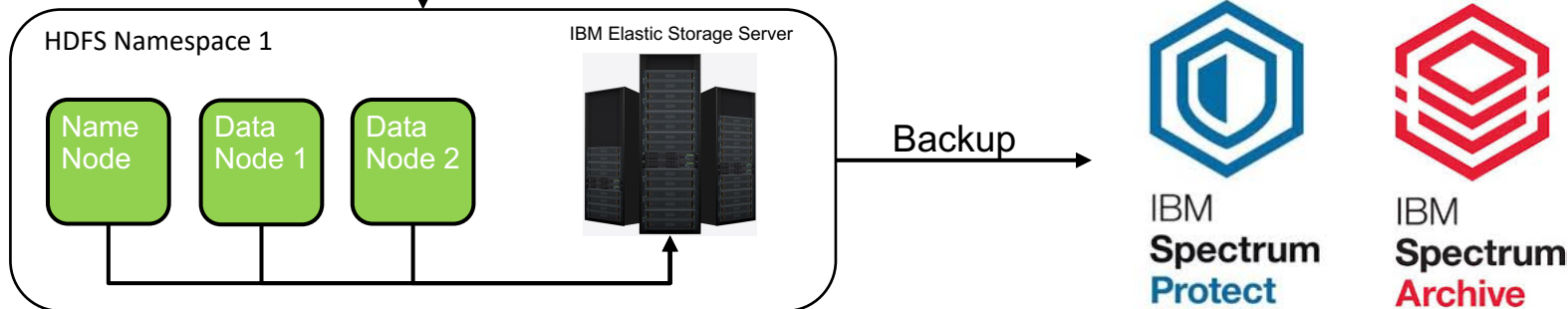


# Use Case 3: HDFS Backup

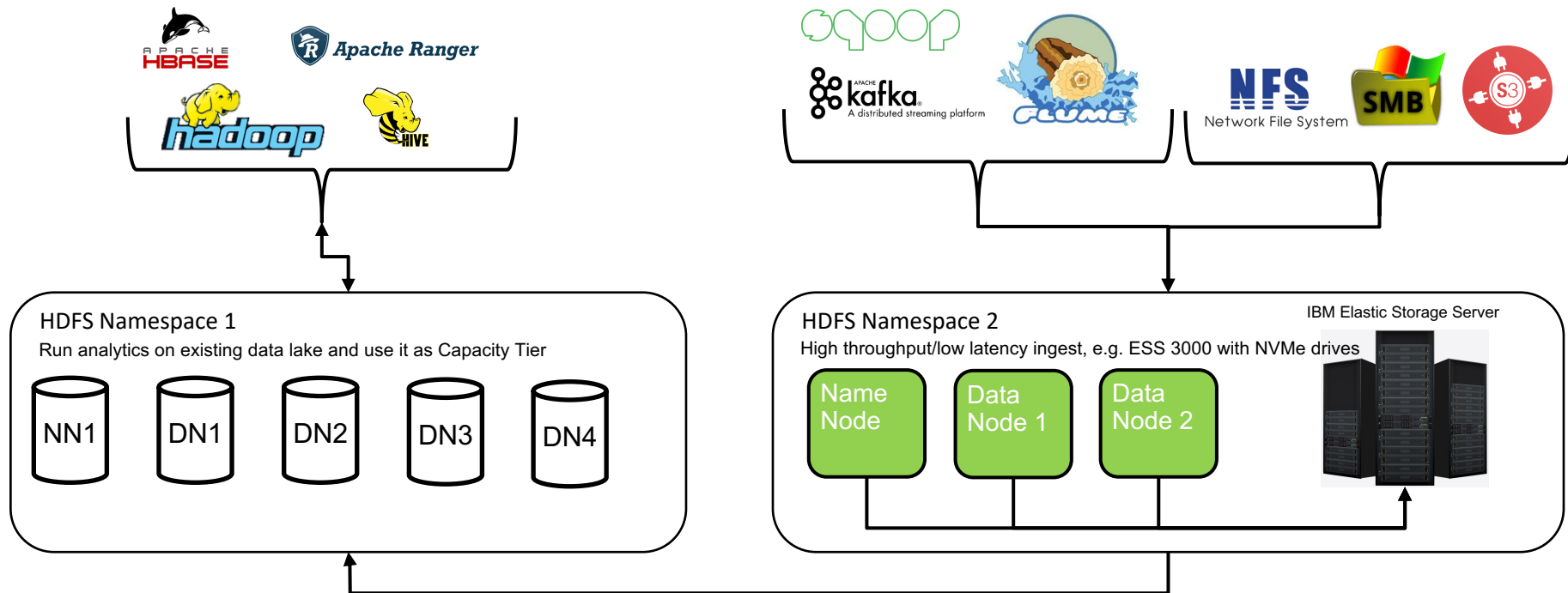


## Sample Backup Flow:

1. Applications write HDFS data to Spectrum Scale
2. Admin dumps databases storing metadata (e.g. Hive Metastore, Ranger DB, ...) to Spectrum Scale
3. Admin creates a Spectrum Scale Filesystem snapshot
4. Snapshot is backed up using Spectrum Protect or Spectrum Archive

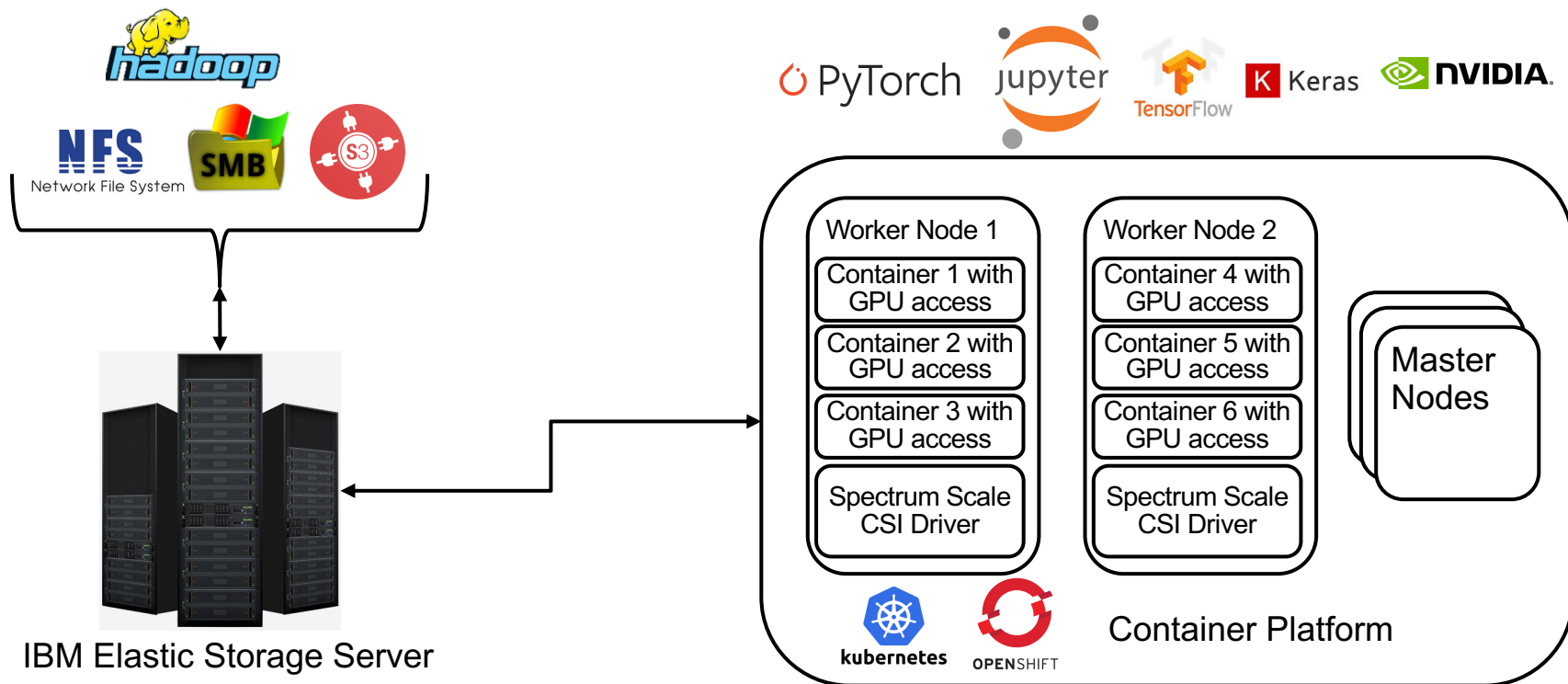


# Use Case 4: Spectrum Scale as Ingest Tier

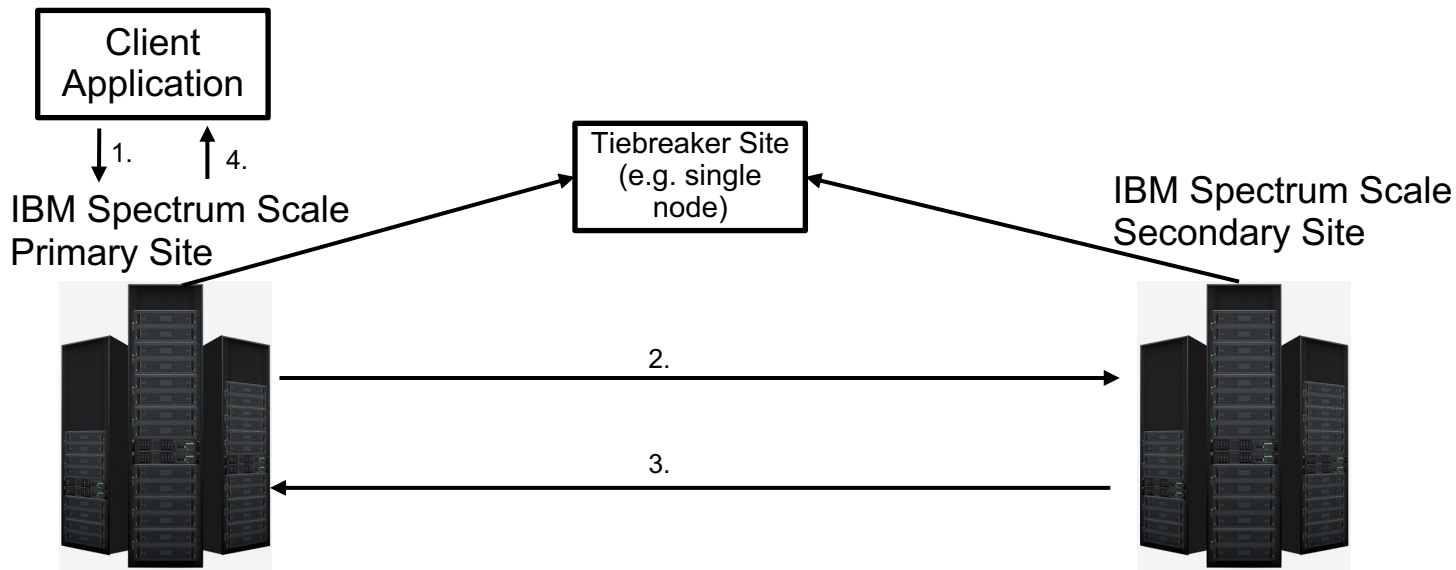


```
> hadoop distcp hdfs://spectrum-scale-namenode:8020 hdfs://native-namenode:8020
```

# Use Case 5: Next generation workloads



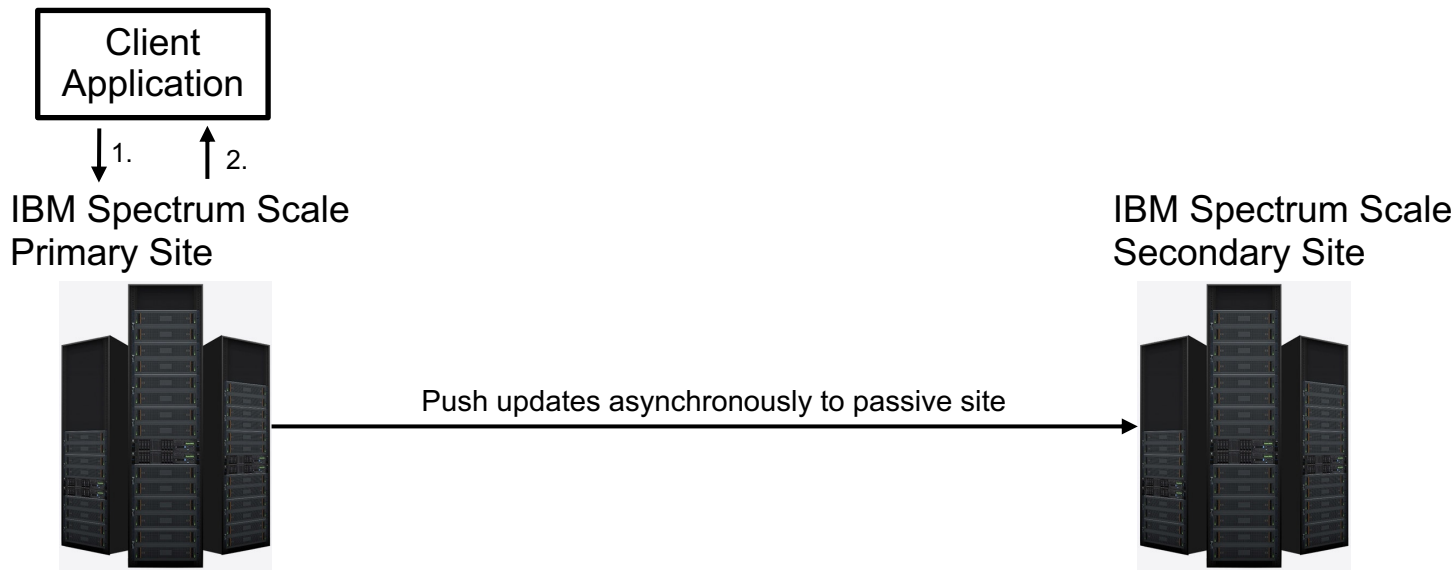
# Use Case 6: Disaster Recovery and Fault Tolerance IBM Spectrum Scale



## Active/Active: Stretch Cluster configuration

- All nodes are members of one Spectrum Scale cluster
- Network between the sites is critical for performance (< 100km distance recommended)
- Tiebreaker Site needed so that the remaining site can continue the operation

# Use Case 6: Disaster Recovery and Fault Tolerance IBM Spectrum Scale



## **Active/Passive: Async DR with Active File Management**

- Two separate Spectrum Scale clusters
- Network between the sites is not critical (e.g. high latency, low bandwidth, WAN-like)
- Snapshots can be used as Recovery Point Objective (RPO)

# Spectrum Scale HDFS integration into CES

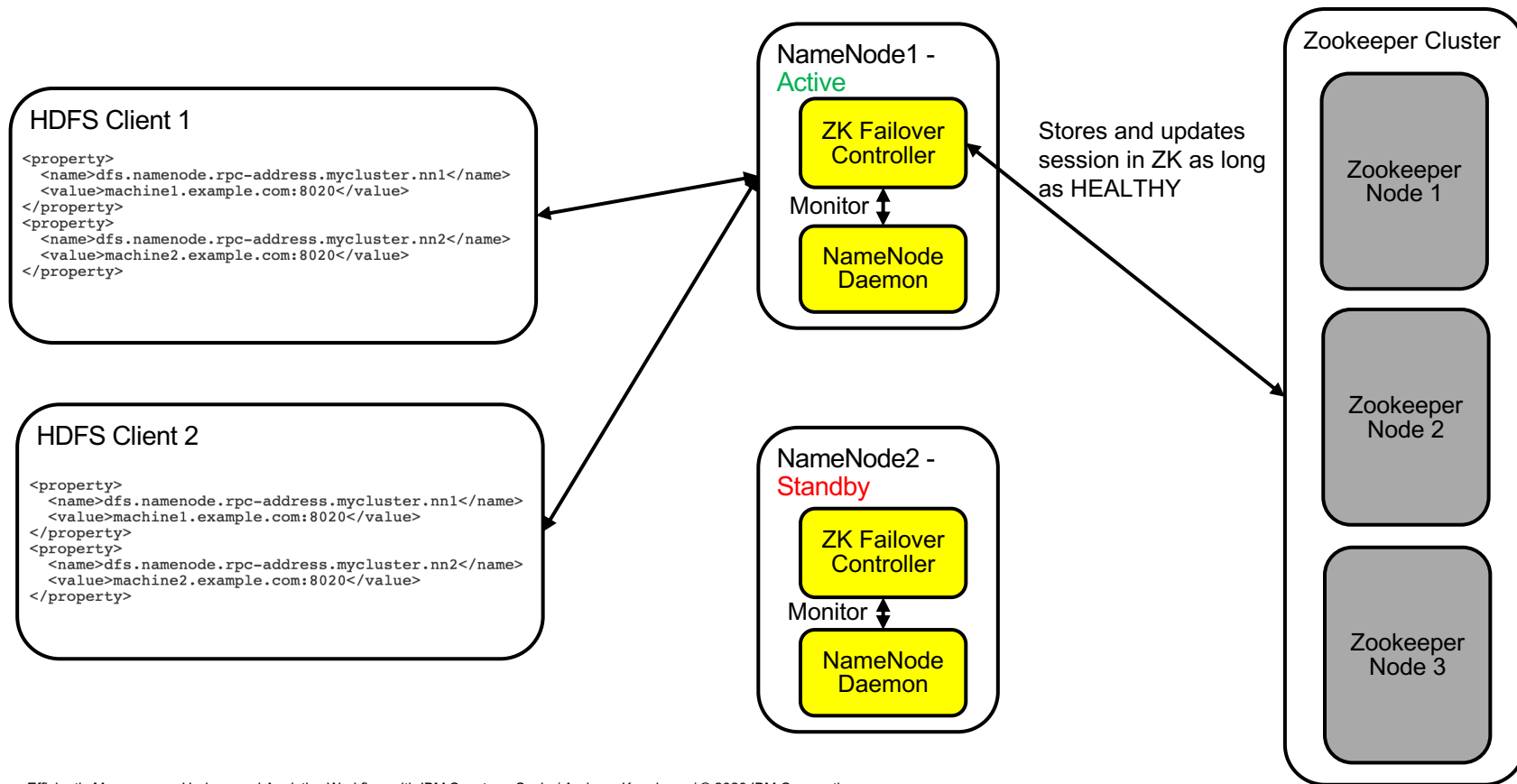


# Spectrum Scale HDFS integration into CES



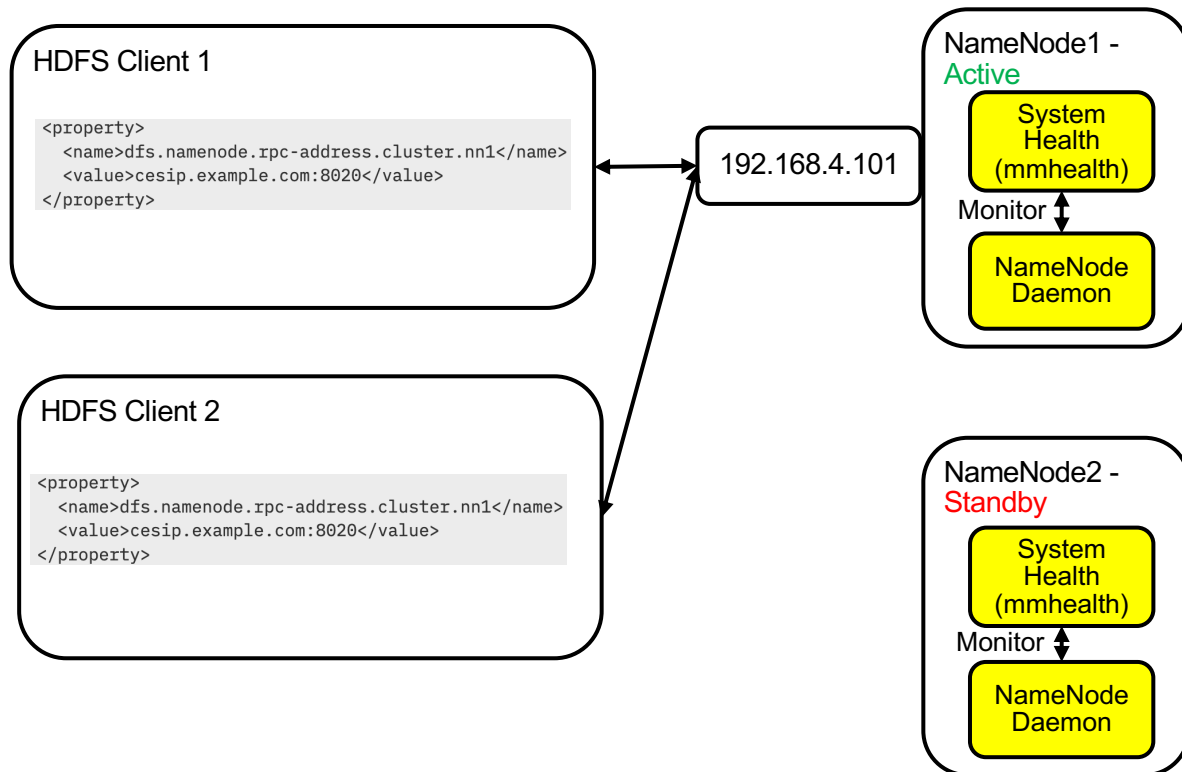
- Available in:
  - Spectrum Scale 5.0.4.2 and later
  - HDFS Transparency 3.1.1.0 and later
  - Supported with Open Source Apache Hadoop only
- Only new installations supported for now (no upgrade support)
- HDP customers stay on HDFS Transparency 3.1.0-x
- CDP will be based on the CES HDFS model for light integration between IBM Spectrum Scale and CDP
- Fully integrated in Spectrum Scale Installation Toolkit

# Traditional Hadoop/HDFS Transparency <= 3.1.0.x



# Spectrum Scale CES/HDFS integration

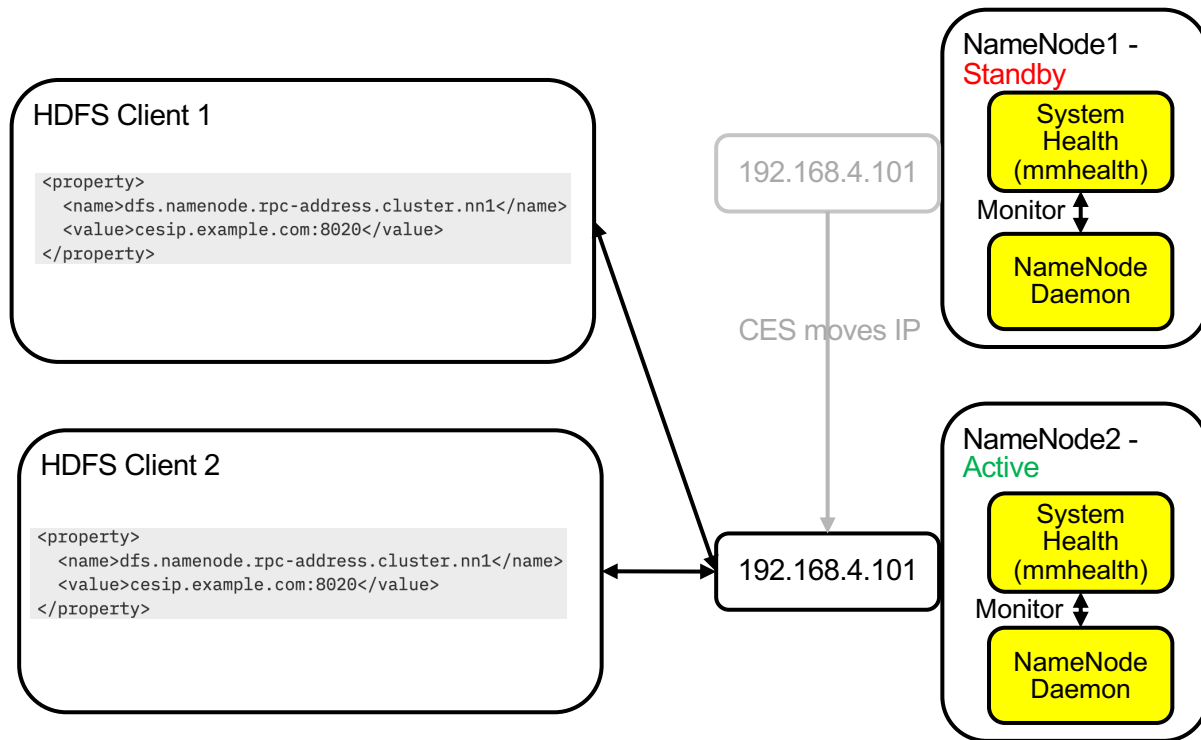
## HDFS Transparency >= 3.1.1.0



- CES IP is assigned to the active NameNode
- HDFS clients always talk to this single CES IP
- SystemHealth monitors the active NameNode
- If something goes wrong CES moves the CES IP to another NameNode and sets the new NameNode to active and the old NameNode to standby

# Spectrum Scale CES/HDFS integration

## HDFS Transparency >= 3.1.1.0



- CES IP is moved to the new NameNode
- New NameNode is active, old NameNode is standby
- HDFS clients retry connection during failover

# Spectrum Scale CES/HDFS configuration



- Two HDFS clusters in the same Spectrum Scale cluster:
  - hdfsmycluster1
  - hdfsmycluster2
- Active NameNodes:
  - hdfsmycluster1: ak-43
  - hdfsmycluster2: ak-44
- Other CES IPs are available for other protocols (SMB, NFS, Object)

```
[root@ak-44 ~]# mmces address list
```

Address	Node	Ces Group	Attributes
192.168.4.100	ak-43.localnet.com	hdfsmycluster1	hdfsmycluster1
192.168.4.101	ak-44.localnet.com	hdfsmycluster2	hdfsmycluster2
192.168.4.102	ak-42.localnet.com	none	none
192.168.4.103	ak-42.localnet.com	none	none
192.168.4.104	ak-42.localnet.com	none	none
192.168.4.105	ak-43.localnet.com	none	none
192.168.4.106	ak-44.localnet.com	none	none
192.168.4.107	ak-43.localnet.com	none	none
192.168.4.108	ak-44.localnet.com	none	none
192.168.4.109	ak-42.localnet.com	none	none
192.168.4.110	ak-43.localnet.com	none	none
192.168.4.111	ak-44.localnet.com	none	none

# Spectrum Scale CES/HDFS configuration



Active protocol services:

```
[root@ak-42 ~]# mmces service list -N cesNodes
Enabled services: SMB NFS HDFS
ak-43.localnet.com:  SMB is running, NFS is running, HDFS NameNode is running
ak-44.localnet.com:  SMB is running, NFS is running, HDFS NameNode is running
ak-42.localnet.com:  SMB is running, NFS is running, HDFS NameNode is running
```

NameNode state of mycluster1:

```
[root@ak-42 ~]# /usr/lpp/mmfs/hadoop/bin/hdfs haadmin -checkHealth -scale -all
-----
Cluster Name:          mycluster1
Namenode Service ID:   nn1
Health State:          OK
Service State:         STANDBY
Error Message:
-----
-----
Cluster Name:          mycluster1
Namenode Service ID:   nn2
Health State:          OK
Service State:         ACTIVE
Error Message:
-----
```

# Spectrum Scale CES/HDFS monitoring

Query NameNode state using mmhealth:

```
[root@ak-41 ~]# mmhealth node show -N cesNodes
```

Node name: ak-42.localnet.com  
Node status: DEGRADED  
Status Change: 6 days ago

Component	Status	Status Change	Reasons
GPFS	TIPS	19 days ago	gpfs_maxstatcache_low, gpfs_pagepool_small
NETWORK	HEALTHY	19 days ago	-
FILESYSTEM	HEALTHY	19 days ago	-
DISK	HEALTHY	19 days ago	-
CES	FAILED	6 days ago	hdfs_namenode_process_down(mycluster1), hdfs_namenode_failed(mycluster1), hdfs_namenode_unknown_state(mycluster1)
HDFS_DATANODE	FAILED	2 days ago	hdfs_datanode_process_down
PERFMON	HEALTHY	19 days ago	-
THRESHOLD	HEALTHY	19 days ago	-

Node name: ak-43.localnet.com  
Node status: TIPS  
Status Change: 6 days ago

Component	Status	Status Change	Reasons
GPFS	TIPS	19 days ago	gpfs_maxstatcache_low, gpfs_pagepool_small
NETWORK	HEALTHY	19 days ago	-
FILESYSTEM	HEALTHY	19 days ago	-
DISK	HEALTHY	19 days ago	-
CES	HEALTHY	6 days ago	-
PERFMON	HEALTHY	19 days ago	-
THRESHOLD	HEALTHY	15 days ago	-

Node name: ak-44.localnet.com  
Node status: TIPS  
Status Change: 6 days ago

Component	Status	Status Change	Reasons
GPFS	TIPS	19 days ago	gpfs_maxstatcache_low, gpfs_pagepool_small
NETWORK	HEALTHY	19 days ago	-
FILESYSTEM	HEALTHY	19 days ago	-
DISK	HEALTHY	19 days ago	-
CES	HEALTHY	6 days ago	-
HDFS_DATANODE	HEALTHY	14 days ago	-
PERFMON	HEALTHY	19 days ago	-
THRESHOLD	HEALTHY	19 days ago	-

# Spectrum Scale CES/HDFS monitoring



Drill down into component HDFS\_NAMENODE:

```
[root@ak-41 ~]# mmhealth node show HDFS_NAMENODE -N cesNodes
```

Node name: ak-42.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	FAILED	6 days ago	hdfs_namenode_process_down(mycluster1), hdfs_namenode_failed(mycluster1), hdfs_namenode_unknown_state(mycluster1)

Event	Parameter	Severity	Active Since	Event Message
hdfs_namenode_process_down	HDFS_NAMENODE	ERROR	6 days ago	HDFS NameNode process for hdfs cluster mycluster1 is down
hdfs_namenode_failed	HDFS_NAMENODE	ERROR	6 days ago	HDFS NameNode health for hdfs cluster mycluster1 failed
hdfs_namenode_unknown_state	HDFS_NAMENODE	WARNING	6 days ago	HDFS NameNode service state for hdfs cluster mycluster1 is UNKNOWN

Node name: ak-43.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	HEALTHY	7 days ago	-

There are no active error events for the component HDFS\_NAMENODE on this node (ak-43.localnet.com).

Node name: ak-44.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	HEALTHY	7 days ago	-

There are no active error events for the component HDFS\_NAMENODE on this node (ak-44.localnet.com).



# Spectrum Scale CES/HDFS monitoring

Display HEALTHY events for component HDFS\_NAMENODE as well:

```
[root@ak-41 ~]# mmhealth node show HDFS_NAMENODE -v -N cesNodes
```

Node name: ak-42.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	FAILED	2020-01-29 14:33:44	hdfs_namenode_process_down(mycluster1), hdfs_namenode_failed(mycluster1), hdfs_namenode_unknown_state(mycluster1)

Event	Parameter	Severity	Active Since	Event Message
hdfs_namenode_process_down	HDFS_NAMENODE	ERROR	2020-01-29 14:33:44	HDFS NameNode process for hdfs cluster mycluster1 is down
hdfs_namenode_failed	HDFS_NAMENODE	ERROR	2020-01-29 14:33:43	HDFS NameNode health for hdfs cluster mycluster1 failed
hdfs_namenode_unknown_state	HDFS_NAMENODE	WARNING	2020-01-29 14:33:43	HDFS NameNode service state for hdfs cluster mycluster1 is UNKNOWN

Node name: ak-43.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	HEALTHY	2020-01-28 17:31:54	-

Event	Parameter	Severity	Active Since	Event Message
hdfs_namenode_active	HDFS_NAMENODE	INFO	2020-01-29 12:49:51	HDFS NameNode service state for hdfs cluster mycluster1 is ACTIVE
hdfs_namenode_ok	HDFS_NAMENODE	INFO	2020-01-28 17:31:54	HDFS NameNode health for hdfs cluster mycluster1 is ok
hdfs_namenode_process_up	HDFS_NAMENODE	INFO	2020-01-28 17:31:26	HDFS NameNode process for hdfs cluster mycluster1 is ok

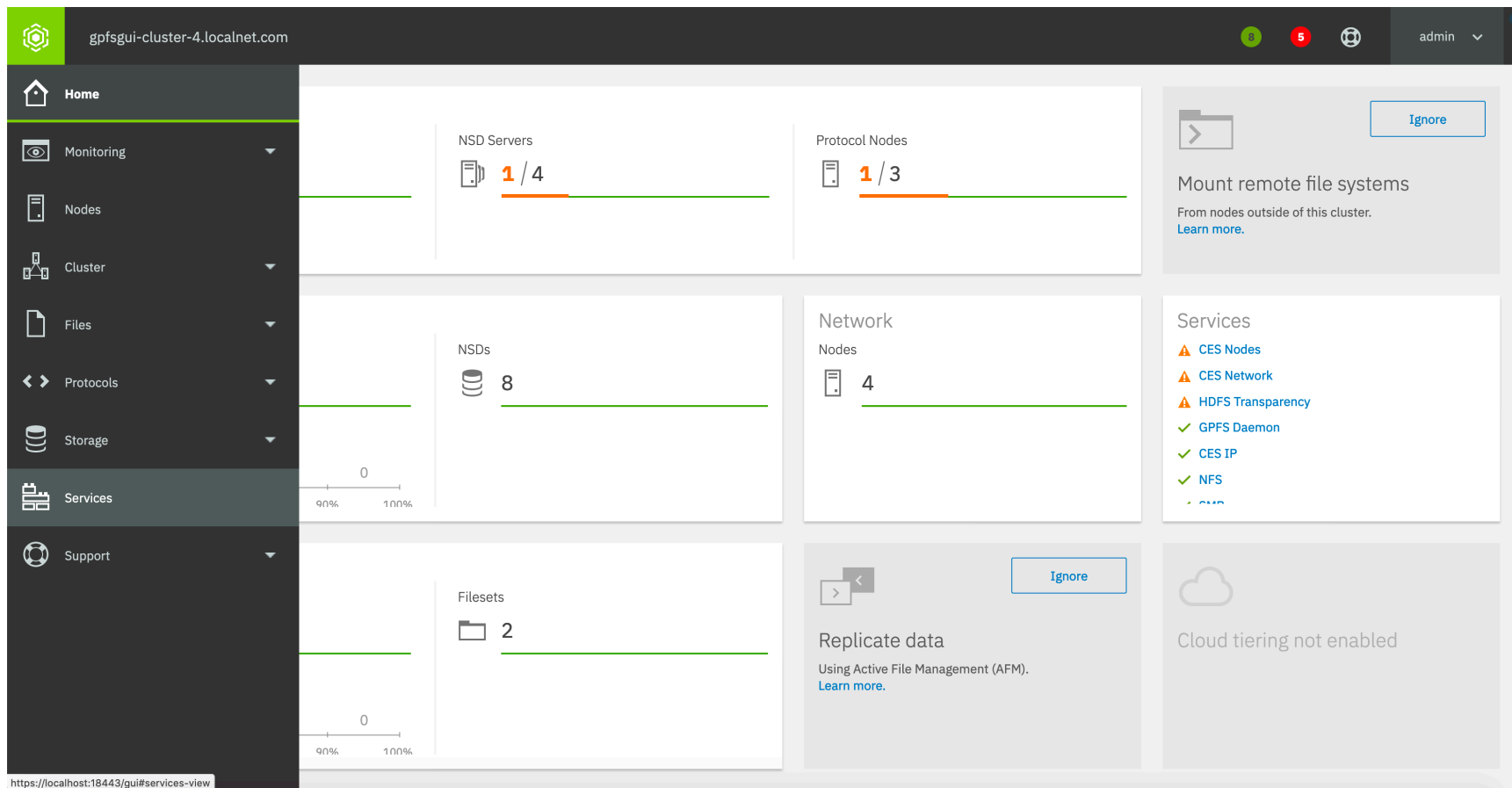
Node name: ak-44.localnet.com

Component	Status	Status Change	Reasons
HDFS_NAMENODE	HEALTHY	2020-01-28 17:33:09	-


  












Event	Parameter	Severity	Active Since	Event Message
hdfs_namenode_active	HDFS_NAMENODE	INFO	2020-01-28 17:33:09	HDFS NameNode service state for hdfs cluster mycluster2 is ACTIVE
hdfs_namenode_ok	HDFS_NAMENODE	INFO	2020-01-28 17:32:10	HDFS NameNode health for hdfs cluster mycluster2 is ok
hdfs_namenode_process_up	HDFS_NAMENODE	INFO	2020-01-28 17:31:42	HDFS NameNode process for hdfs cluster mycluster2 is ok

# Spectrum Scale CES/HDFS monitoring





# Spectrum Scale CES/HDFS monitoring


 gpfsui-cluster-4.localnet.com





Services


**CES Nodes**  
One out of 3 nodes is in the "failed" state.


**CES Networking**  
One out of 3 nodes is in the "degraded" state.


**HDFS Transparency**  
One out of 3 nodes is in the "failed" state.


**GPFS Daemon**  
Manages the I/O and buffer management operations on the IBM...

**NFS**  
Provides NFSV3 and NFSV4 access to the files stored in the IBM Spectrum...


**SMB**  
Provides access to data through Microsoft Windows-style shares.


**GUI**  
Manages the graphical user interface and the REST-based API that are us...

**Performance Monitoring**  
Collects performance and capacity data from various services.


**File Authentication**  
Authenticates the NFS and SMB users.


**CES Nodes****Addresses**EventsSettings

 All defined CES IPs are assigned

 Add CES IP

Actions ▾



 Refresh












III


<input type="checkbox"/>	Address	Node	Group	Attribute
<input type="checkbox"/>	192.168.4.100	ak-43.localnet.com	hdfsmycluster1	hdfsmycluster1
<input type="checkbox"/>	192.168.4.101	ak-44.localnet.com	hdfsmycluster2	hdfsmycluster2
<input type="checkbox"/>	192.168.4.102	ak-43.localnet.com		
<input type="checkbox"/>	192.168.4.103	ak-43.localnet.com		
<input type="checkbox"/>	192.168.4.104	ak-43.localnet.com		
<input type="checkbox"/>	192.168.4.105	ak-43.localnet.com		
<input type="checkbox"/>	192.168.4.106	ak-44.localnet.com		
<input type="checkbox"/>	192.168.4.107	ak-44.localnet.com		
<input type="checkbox"/>	192.168.4.108	ak-44.localnet.com		
<input type="checkbox"/>	192.168.4.109	ak-44.localnet.com		
<input type="checkbox"/>	192.168.4.110	ak-43.localnet.com		


# Spectrum Scale CES/HDFS monitoring


 gpfsgui-cluster-4.localnet.com 6 5  admin ▾





Services


 **CES Nodes**  
One out of 3 nodes is in the "failed" state.


 **CES Networking**  
One out of 3 nodes is in the "degraded" state.


 **HDFS Transparency**  
One out of 3 nodes is in the "failed" state.


 **GPFS Daemon**  
Manages the I/O and buffer management operations on the IBM...

 **NFS**  
Provides NFSV3 and NFSV4 access to the files stored in the IBM Spectrum...

 **SMB**  
Provides access to data through Microsoft Windows-style shares.



 **GUI**  
Manages the graphical user interface and the REST-based API that are us...












 **Performance Monitoring**  
Collects performance and capacity data from various services.

 **File Authentication**



**Name Nodes** Data Nodes Events













↻ Refresh

Search  


<input type="checkbox"/>	Node Name	↑	CES Status	CES Network Status	HDFS Health State	CES Node Group	Active Name Node	CES Network
<input type="checkbox"/>	<a href="#">ak-42.localnet.com</a>		 Failed	 Degraded	 Failed	hdfsmycluster1		
<input type="checkbox"/>	<a href="#">ak-43.localnet.com</a>		 Healthy	 Healthy	 Healthy	hdfsmycluster1		6
<input type="checkbox"/>	<a href="#">ak-44.localnet.com</a>		 Healthy	 Healthy	 Healthy	hdfsmycluster2		6


# Spectrum Scale CES/HDFS monitoring


 gpfsui-cluster-4.localnet.com 1 5  admin ▾





Services


**CES Nodes**  
One out of 3 nodes is in the "failed" state.


**CES Networking**  
One out of 3 nodes is in the "degraded" state.


**HDFS Transparency**  
One out of 3 nodes is in the "failed" state.


**GPFS Daemon**  
Manages the I/O and buffer management operations on the IBM...

**NFS**  
Provides NFSV3 and NFSV4 access to the files stored in the IBM Spectrum...

**SMB**  
Provides access to data through Microsoft Windows-style shares.

**GUI**  
Manages the graphical user interface and the REST-based API that are us...

**Performance Monitoring**  
Collects performance and capacity data from various services.

**File Authentication**

Name Nodes

**Data Nodes**

Events

Refresh

Search

Q

III

<input type="checkbox"/>	Data Node	↑	Data Node Status
<input type="checkbox"/>	ak-41.localnet.com	✓	Healthy
<input type="checkbox"/>	ak-42.localnet.com	✗	Failed
<input type="checkbox"/>	ak-44.localnet.com	✓	Healthy

# Spectrum Scale CES/HDFS monitoring



gpfsui-cluster-4.localnet.com

Services

**CES Nodes**  
One out of 3 nodes is in the "failed" state.

**CES Networking**  
One out of 3 nodes is in the "degraded" state.

**HDFS Transparency**  
One out of 3 nodes is in the "failed" state.

**GPFS Daemon**  
Manages the I/O and buffer management operations on the IBM...

**NFS**  
Provides NFSV3 and NFSV4 access to the files stored in the IBM Spectrum...

**SMB**  
Provides access to data through Microsoft Windows-style shares.

**GUI**  
Manages the graphical user interface and the REST-based API that are us...

**Performance Monitoring**  
Collects performance and capacity data from various services.

**File Authentication**  
Authenticates the NFS and SMB users

Name Nodes

Data Nodes

Events

Actions ▾

Current Issues ▾

↻ Last Updated: 10:53 AM

Export

Q

III

Severity ↑	Event Time	Reporting Node	Event Name	Message
Warning	1/29/20 2:33:43 PM	ak-42.localnet.com	hdfs_namenode_unknown_state	HDFS NameNode service state for hdfs cluster mycluster1 is UNKNOWN
Error	2/2/20 6:41:19 PM	ak-42.localnet.com	hdfs_datanode_process_down	HDFS DataNode process for hdfs cluster mycluster1 is down
Error	1/29/20 2:33:44 PM	ak-42.localnet.com	hdfs_namenode_process_down	HDFS NameNode process for hdfs cluster mycluster1 is down
Error	1/29/20 2:33:43 PM	ak-42.localnet.com	hdfs_namenode_failed	HDFS NameNode health for hdfs cluster mycluster1 failed

# Useful links



**IBM Spectrum Scale Big Data and Analytics support in IBM Knowledge Center:**

[https://www.ibm.com/support/knowledgecenter/STXKQY\\_BDA\\_SHR/bl1adv\\_kc\\_bigdata\\_analytics\\_kclanding.htm](https://www.ibm.com/support/knowledgecenter/STXKQY_BDA_SHR/bl1adv_kc_bigdata_analytics_kclanding.htm)

**Native HDFS to IBM Spectrum Scale HDFS migration:**

<https://developer.ibm.com/storage/2019/01/18/migrating-data-from-native-hdfs-to-ibm-spectrum-scale-based-shared-storage/>

**CES HDFS support Blog Post:**

<https://developer.ibm.com/storage/2020/02/03/ces-hdfs-transparency-support/>

**CES HDFS support in IBM Knowledge Center:**

[https://www.ibm.com/support/knowledgecenter/STXKQY\\_BDA\\_SHR/bl1bda\\_ceshdfs.htm](https://www.ibm.com/support/knowledgecenter/STXKQY_BDA_SHR/bl1bda_ceshdfs.htm)

**Contact:**

Andreas Koeninger <andreas.koeninger@de.ibm.com>

# Thank you!

## Provide Feedback



Tell IBM What You Think

Let us know what you think about IBM Spectrum Scale. It takes only a couple of minutes for you to help us improve our service. [IBM Privacy Policy](#)

Not Now

 Provide Feedback

Please help us to improve Spectrum Scale with your feedback

- If you get a survey by email or a popup in the GUI, please respond
- We read every single reply

**Disclaimer:** All product plans, directions and intent are subject to change or withdrawal without notice. References to IBM products, programs or services do not imply that they will be available in all countries in which IBM operates. IBM, the IBM logo, and other IBM products and services are trademarks of the International Business Machines Corporation, in the United States, other countries or both. Other company, product, or services names may be trademarks or services marks of others.