Spectrum Scale Hadoop Integration and Support for HortonWorks HDP

Ted Hoover
Manager
Spectrum Scale and High Performance Analytics Solutions
Challenges with the Early Big Data Storage Models

- Ingest data at various end points
  - More data source than ever before, not just data you own, but public or rented data

- Move data to the analytics engine
  - It takes hours or days to move the data!

- Perform analytics
  - Can’t just throw away metadata due to regulations or business requirement

- Key Business processes are now depend on the analytics

- It’s not just one type of analytics

IBM Storage & SDI

- It takes hours or days to move the data!

- Can’t just throw away metadata due to regulations or business requirement
Big Data Platform requires a “Data Ocean” Approach
 Fortune 2000 Enterprise

1. Bring analytics to the data
2. Single Name Space to house all your data
3. Geographically dispersed management of data including disaster recovery
4. Unified data access with File, Object and Block
5. Encryption for protection of your data
6. Optimize economics based on value of the data
Spectrum Scale: Unleash new storage economics on a global scale

IBM Spectrum Scale
Automated data placement and data migration

4000+ customers using Spectrum Scale as data plane for HPC and analytics workload
Hortonworks announced the planned availability of Hortonworks Data Platform (HDP®) for IBM Elastic Storage Server (ESS) and IBM Spectrum Scale

Las Vegas, NV (IBM PartnerWorld) - 14 Feb 2017: IBM (NYSE: IBM) and. The agreement with Hortonworks will lead to certification of Hortonworks HDP on Power with IBM Spectrum Scale and Hortonworks HDP on x86 with IBM Spectrum Scale.

With the recent announcement of availability of Hortonworks HDP on POWER8 platform, IBM clients already have increased choice when selecting the platform for their Hadoop distribution. This new agreement will enable IBM clients to also leverage their existing and future investments in IBM storage in deploying Hadoop based big data applications. Additionally, this agreement will provide existing Hortonworks HDP customers an enterprise-class storage alternative with IBM Spectrum Scale for their Hadoop and Spark workloads.

The agreement represents the first IBM Storage offering and industry first enterprise Software Defined Storage solution certified for Hortonworks.

When completed, the HDP certifications will provide a significant layer of confidence to existing IBM and future clients. Those users can now run Hadoop applications on the leading software-defined storage solutions. IBM clients will have the benefits of enterprise storage to analyze data in place with Hortonworks analytics applications. With the choice of centralized or distributed deployments, organizations can improve business efficiency with the data management, backup, security and hybrid cloud storage.

Why are we partnering with Hortonworks?

- Hortonworks is one of the key distributions of Hadoop
  - Pure play 100% open source distribution
  - Hortonworks is #1 Apache Hadoop committer
  - Has 1000+ customers
  - ODPI compliance
  - Apache Spark is part of HDP distribution

- Spectrum Scale is already certified and supported with IBM’s Hadoop distribution (IBM IOP/BigInsights)

- Hortonworks certifying Spectrum Scale/ESS will allow clients to confidently deploy Spectrum Scale/ESS for Big Data Analytics use cases powered by Hadoop
What value does Spectrum Scale provide to Hadoop?

- The default storage for Hadoop is HDFS
- HDFS is Hadoop Distributed File System which runs on storage rich servers (storage internal to servers)
- Spectrum Scale provides a HDFS connector and allows existing Hadoop applications to run directly on Spectrum Scale
- This enables customers to create complex analytics workflows, minimize data movement & copies and speed up time to insight
HDFS Shortcomings addressed by Spectrum Scale

- HDFS is a shared nothing architecture, which is very inefficient for high throughput jobs (disks and cores grow in same ratio)
- Costly data protection:
  - uses 3-way replication; limited RAID/erasure coding
- Works only with Hadoop i.e weak support for File or Object protocols
- Clients have to copy data from enterprise storage to HDFS in order to run Hadoop jobs, this can result in running on stale data.
Key Deployment Models

**Shared Storage Architecture**
- Compute Nodes
  - IBM Power
  - X86
- InfiniBand / 40 GigE / 10 GigE

**ESS based deployment**

**ESS based deployment**

2017 enablement focus

**IBM Storage & SDI**

**Shared Nothing Architecture**
- Storage-Rich Servers
  - IBM Power
  - X86
- 10 GigE / InfiniBand

Storage rich server based deployment
Why Spectrum Scale

- Reduce the datacenter footprint
  Spectrum Scale provides **in-place analytics** for your Hadoop data eliminating the need for multiple copies of the same data or large migrations of data between HDFS and the POSIX file system.

- Control cluster sprawl
  Spectrum Scale delivers **federation across clusters**, both Scale and HDFS, turning isolated data lakes into a data ocean while still maintaining needed separations.

- Make HDFS access enterprise-ready
  IBM Elastic Storage Server (Pre-integrated Spectrum Scale based system) adds **storage functions necessary to the enterprise** (e.g. Encryption, DR, SW RAID) to your Hadoop setup.
Client Use Case: Life Sciences with HPC and Hadoop/Spark

IBM Storage & SDI

LSF Job 1 → LSF Hadoop Job → File A Event on shared pool

LSF Job 2 → File B on shared pool

LSF Spark Job → LSF Job 5 → File E on FPO pool → File F on FPO pool → File F replicated to a remote Spectrum Scale Server

File C on FPO pool

File D Event on FPO pool

HPC Compute Cluster

ESS based shared storage cluster

Storage

Storage

Storage
Client Use Case: HPC and Data Analytics

Ingest

HDFS

POSIX & Object

Analyze

HDFS

Analyze

HDFS

HPC Compute Cluster

ESS based shared storage cluster

Simulate

POSIX

iterate

IBM Storage & SDI
Client Use Case: Unified Analytic/Workflow Pipelines

Ingest → Analyze → Export → Visualize

POSIX → HDFS → NFS → Object → POSIX

RDBMS → Analyze → Dashboard → Report → Share

POSIX → HDFS → POSIX → SMB → Object

Data Lake

Structured data warehouse

Hadoop on Shared Nothing Cluster (FPO)

Warehouse extension for unstructured data

DB2 on Shared Nothing Cluster (FPO)

Compute Cluster

SAS Analytics

Dashboard & Reporting

ESS

Storage
Federate Spectrum Scale Filesystem with Existing HDFS Filesystem

Extending the Filesystem

- Run analytics across multiple HDFS and/or Spectrum Scale clusters
  - No need to move the data
  - Applications can still access data that has been moved seamlessly

Commands:
$ hadoop distcp viewfs://clusterX:/hadoop/hdfs/file1 viewfs://clusterY:/hadoop/gpfs/file1
$ Hadoop fs rm viewfs://clusterX:/hadoop/hdfs/file1
Spectrum Scale Makes HDFS Enterprise Ready

Protecting Business Data

- Use ESS warm data tier with Spectrum Archive to tape
- Powerful policy engine
  - Information Lifecycle Management
  - Fast metadata ‘scanning’ and data movement
  - Automated data migration to based on threshold
- Users not affected by data migration
  - Single namespace
- Optionally Spectrum Protect and Spectrum Archive can be used directly with Spectrum Scale FPO

viewfs://clusterX:/hadoop/hdfs/file1
viewfs://clusterY:/hadoop/gpfs/file1

IBM Spectrum Scale HDFS Transparency Connector

hdfs://nn1.node.net:8020/hadoop/hdfs
hdfs://nn2.node.net:8020/hadoop/gpfs

IBM Spectrum Protect
IBM Spectrum Archive
Hadoop Data Analytics with IBM Spectrum Scale

IBM Spectrum Scale is a proven enterprise ready alternative of HDFS for Hadoop data analytics

Completely compatible with HDFS client API and Hadoop applications

Unified File, Object and HDFS access interface

Mature enterprise features make it easy to manage and protect data
Spectrum Scale HDFS Transparency Connector

- Integrates with HDFS – reuses HDFS client and implements NameNode and DataNode RPCs
- Stateless
- Free download – [link](#)
- Supports Spectrum Scale 4.1.x and 4.2.x
- Currently supported with IOP 4.1 and 4.2
- Ambari integration available for both IOP 4.1 and 4.2
Spectrum Scale Ambari Integration
Ambari Integration with HDFS Transparency: Overview

- Spectrum Scale is added as a new service after full IOP install with HDFS
- Spectrum Scale service “integrates” with HDFS
- Supports “un-integrate” capability
  - Flip back and forth between HDFS & GPFS
  - Will not move data back and forth between HDFS & GPFS
- Will simplify future upgrades
HDFS Transparency NameNode and DataNodes
Big Data Oceans extending HDFS across Clusters

Unified Data Repository, Support Multiple Analytics
- HDFS Transparency Connector

Federate ESS with Existing HDFS Filesystem
- HDFS Transparency Connector

Federate ESS with Existing Spectrum Scale Filesystem
- Single Name Space

Extending the Filesystem
- Run analytics across multiple HDFS and/or Spectrum Scale clusters
  - No need to move the data
  - Build Data Oceans on demand

Improve Hadoop Cluster Utilization
- Manually move less frequently accessed data to an ESS tier
  - Applications can still access data that has been moved seamlessly

Expand use of Shared Nothing Clusters
- Simplicity of Storage Rich Servers with enterprise features
- Advanced Routing (AFM), encryption, QoS, compression
- Mix cluster types
- Backup and Archive Support
Spectrum Scale Wiki Links

Spectrum Scale wiki – Analytics

Spectrum Scale wiki - FPO

Spectrum Scale wiki – Hadoop

Spectrum Scale wiki – HDFS Transparency connector
Thank You.
IBM Storage & SDI