A Unified Data Platform for Big Data & Cognitive

Yong ZY Zheng(zhengyzy@cn.ibm.com)

IBM Spectrum Scale BDA

March 17, 2018
Agenda

• Challenges for Customers in Data Analytics
• Spectrum Scale for Big Data
• Spectrum Scale for Cognitive
• Unified Data Platform Architecture
• Q&A
Typical Challenges for Customers in Data Analytics

- **Inefficient data movements and many copies** due to poor interfaces and siloed infrastructure

- Different analytics workloads might need different interface
  - Traditional workloads (such as data warehouse, HPC etc) requires POSIX interface
  - Data ingestion clients need NFS/SMB interface
  - map/reduce jobs need HDFS interface

- Poor data accessing interfaces from storage system make customers have to set up siloed infrastructure

- Siloed data lakes bring inefficient data analysis (e.g. long time analysis because of data movement)

- Siloed data lakes bring inefficient storage space utilization (e.g. many copies on different storage systems)

- Multiple siloed storage systems bring further issues on cost, management and scaling.
Typical Challenges for Customers in Data Analytics

Efficient data lifecycle management should:

- Provide different tiers for different data (hot/warm/cold data) and could tell which files are hot for migration
- Leverage different disks from NVMe to Tape according to performance/cost
- Provide policy-based data movement automatically
- Easily move cold data back when needed
Typical Challenges for Customers in Data Analytics

- **Cross-site DR for business continuity**

What customers want for DR:

- DR for all data, not just for part of data
- Flexible options for different business continuity requirements
  - Different RTO(Recovery Time Object)/RPO(Recovery Point Object)
- Different options for low/medium/high cost for cross-site DR
Agenda

• Challenges for Customers in Data Analytics
• Spectrum Scale for Big Data
• Spectrum Scale for Cognitive
• Unified Data Platform Architecture
• Q&A
Big Data Analytics: Native environment Support

- Spectrum Scale completely transparent to Hadoop
  - HDFS Transparency makes GPFS transparent
  - Connector Works with 4.1.X/ 4.2.x/5.0.x
  - Shipped with Spectrum Scale 4.2.x/5.0.x
    - Download the latest from IBM developerWorks GPFS wiki

- HDFS Transparency and Spectrum Scale services could be managed from HortonWorks HDP Ambari GUI
Advanced Storage for Map/Reduce Data

Hadoop HDFS
- File number scaling limit from centralized NameNode
- Large block-sizes – poor support for small files
- Non-POSIX file system – obscure commands
- Difficulty to ingest data – special tools required
- Single-purpose, Hadoop MapReduce only
- Not recommended for non hadoop

IBM GPFS Advantages
- No single point of failure, distributed metadata
- Variable block sizes – suited to multiple types of data and data access patterns
- POSIX file system – easy to use and manage
- Rich data ingestion interfaces
- Versatile, Multi-purpose
- Enterprise Class advanced storage features
All-in-place Data Analytics

Traditional Analytics Solution

• Build analytics system from scratch, not only for compute but also for storage
• Add storage and compute resource at the same time no matter it’s required
• Native HDFS doesn’t support native POSIX
• Lacks of enterprise data management and protection capability

All-in-place Analytics Solution

• Can leverage existing Spectrum Scale storage
• Unified interface for File and Object analytics
• POSIX compatibility
• Mature enterprise data management and protect solutions derived from Spectrum Storage family and 3rd part components
All-in-place Data Analytics

NFS interface from HDFS

- HDFS doesn’t support random read/write, only append mode
- HDFS NFS Gateway has to write data from clients to the local disks first and then move it to HDFS to handle the out-of-order write requests
- No HA for NFS Gateway so far

In-place Analytics Solution

- Rich data ingest interface (SMB, NFS, HDFS, Swift/S3 etc)
- Spectrum Scale Protocol HA
- Random read/write support
- Efficient data ingest because of no data movement from local disk to Spectrum Scale
- Only one data copy and all data are visible immediately from all interface
Full data lifecycle management

- Data migration between different storage pool
- Policy-based Auto Tiering
- TSM+Tape based enterprise data backup for Hadoop

Client workstations

Hadoop

Users and applications

Single name space

Spectrum Scale

• Storage pool1
• Storage pool2
• Storage poolx
• External Storage poolx

Storage rich servers

IBM TSM/LTFS

Tape

Flash

Disk

External Storage

Storage pool1

Storage pool2

Storage poolx

Storage poolx

HDFS

NFS

POSIX

SMB

Swift/S3
Data DR from Hadoop community

- Hbase Cluster Replication
  - WAL-based asynchronous way
  - All nodes in both cluster should be accessible for each other
  - Both clusters could provide Hbase service on the same time
  - Only available for Hbase 0.96+, hbase-7709
  - Supported over HDFS Transparency + Spectrum Scale

- Hive metadata and data replication
  - Hive Replication v1 (available since Hive 1.2.0)
    - Lazy Async, non-blocking, out-of-order events
    - Primary-copy: only allow data copy from primary to replica
    - Disadvantages: slowness, full copy in staging directory (4x copy problem) etc
  - Hive Replication v2 (not all enhancements have been done yet)

- Leverage Hadoop Falcon
  - HDFS mirroring (distcp-based)
  - Hive mirroring (sqoop-based)
Data DR from Spectrum Scale

- **Spectrum Scale Active-Active DR**
  - 2 replica in production cluster; another replica in standby cluster
  - Dedicated network for two clusters (10Gb bandwidth)
  - Distance between two clusters is less than 100Km
  - Can achieve RPO=0 in DR

- **Spectrum Scale AFM based replication**
  - AFM based replication from production cluster (cache site) to standby cluster (home site)
  - Both sites should be Spectrum Scale cluster for Hadoop application failover
  - Only one cluster can provide Hbase service (conflict in assigning region servers if Hbase is up on both cluster)
Big Data Analytics Highlight

2+ Hadoop clusters over the same file system

Native 2+ file system support

Hadoop Platform
- MapReduce
- Zookeeper
- Hive
- HCatalog
- Pig
- YARN
- Spark
- HBase
- Flume
- Sqoop
- Solr/Lucene

HDFS Transparency
- hdfs://namenode:8020
- /fs2
- /data/fs1
- /data/fs2

Standard Apache Open-Source Components
- HDFS
- HBase
- Zookeeper
- Hive
- YARN
- Spark
- Pig
- Flume

Spectrum Scale (GPFS)
(Enterprise Data Platform for Analytics)
Big Data Analytics Highlight

Short Circuit Write

- All Hadoop nodes are Spectrum Scale nodes/HDFS Transparency nodes
- Short circuit write reduces the traffic between client and DataNode on the loop lo adapter
Big Data Analytics Highlight

Hadoop Storage Tiering

- HDP stack on Spectrum Scale side is optional
- Two typical scenarios
  - Take Spectrum Scale as ingestion tier (e.g. take ESS GS1S/GS2S with SSD as fast ingestion tier)
  - Take Spectrum Scale as archive tier (e.g. ESS GLxS)
- Will be available in HDFS Transparency 2.7.3-3 (scheduled to be released around the end of April)
Spectrum Scale BDA Focus in 2018

2017
- HortonWorks HDP 2.6.x certification
- Ambari integration
- HDFS Transparency 2.7.3-x

June
- Ambari integration remote file system support
- HDFS Transparency support 2+ file system
- Non root passwordless support for remote file system
- Short circuit write
- Fileset-based snapshot
- Hadoop Storage Tiering

2018
- Hadoop 3.0 support
- HortonWorks 3.0 support
- Ambari 2.7+ support
- Multiple standby namenode supports
- Native HDFS Federation support from Ambari
- Snapshot-based distcp
- IBM DSX local support
- IBM Big Replicate support
- Hadoop Performance data and performance tuning guide
Agenda

- Challenges for Customers in Data Analytics
- Spectrum Scale for Big Data
- Spectrum Scale for Cognitive
- Unified Data Platform Architecture
- Q&A
Cognitive Workloads/Deep Learning

Training (Research/Development)
- Data Cleansing
- Feature Engineering
- Modeling

Inference (Deployment/Production)
- Recognition, classification
- Application

Trained model

Data Hierarchy in Deep Learning
- AI/Deep Learning
- A/B Testing, ML Algorithms
- Analytics, Segment, Aggregates, Features, Training Data
- Cleaning, Anomaly Detection, Preparing
- Reliable Data Flow, Infrastructure, ETL, Structured and Unstructured Data Storage
- Instrumentation, Logging, Sensors, External Data, User Generated Content

Deep Learning Frameworks:
- TensorFlow (Apache)
- Caffe (BSD)
- Torch (BSD)
- Theano (free)
- CNTK (free)
- Neon (Apache)

IDE: IBM Power AI Enterprise/Spectrum Deep Learning Impact, Nvidia Digits

Spectrum Scale: POSIX for Power AI; need to evaluate performance

Transform/Prepare
- Machine Learning: IBM SPSS, IBM DSX, SAS

Spectrum Scale: SPSS works over GPFS POSIX/HDFS; SAS works over GPFS POSIX

Move/Store
- Platform: Hadoop, Spark, POSIX/NFS/SMB
- ETL: Talend

Spectrum Scale: POSIX/HDFS interface; need to evaluate performance

Collection
- Variable data ingestion end devices

Spectrum Scale: Protocol supports rich interfaces for data ingestion.
Cognitive Workloads IO Pattern Analysis (based on current engagements)

<table>
<thead>
<tr>
<th>Industry</th>
<th>Training Data Size per model</th>
<th>File size level</th>
<th>Workload &amp; IO Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security &amp; Public</td>
<td>Small(128GB data size is the large enough now)</td>
<td>• A lot of small files(e.g. 16KB picture, 1.2M pictures)</td>
<td>• Write once but read many times</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High IOPS</td>
</tr>
<tr>
<td>Telecom</td>
<td>Small(usually, tens of GB)</td>
<td>• Fetched structure data from database</td>
<td>• Write once but read many times</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High IOPS</td>
</tr>
<tr>
<td>Finance</td>
<td>Small(usually, tens of GB)</td>
<td>• Fetched structure data from database</td>
<td>• Write once but read many times</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High IOPS</td>
</tr>
<tr>
<td>EDA &amp; Heath Care</td>
<td>Usually huge( e.g. TB level)</td>
<td>• Large file(e.g. 1GB per file for Microscope-captured picture)</td>
<td>• Write once but read many times</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High throughput</td>
</tr>
</tbody>
</table>
Spectrum Scale for Deep Learning

- Internal SSD disks for entry level customers
- GSxS/SSD could be for customers that have large data size or whose data will increase in the near future

Advantages from Spectrum Scale
- Data read/write in parallel for performance
- Access the same data from all nodes
- SSD/NVMe

Requirements for Spectrum Scale
- High IOPS for small IO with low cost

What we are doing
- Best practice guide (in progress)
- Mmap() performance enhancement (done)
- Further IO optimization for write-once-but-read-many-times:
  - Prefetch/cache all data in LROC locally (in progress)
- Tuning profile for cognitive workloads
Spectrum Scale for Deep Learning

- Take LROC to prefetch and cache the training data
- Workloads read the training data multiple times from local LROC (IO read acceleration)
Solution Key Values:

- Support long-term rapid increasing big data with extremely scaling for file system
- Fast analytics results from in-place analytics without data movement
- Easy maintenance from centralized storage management for multiple Hadoop cluster
- Support internal disk based for entry level customer (less than 100TB data size) and scale to PB level in ESS
Customer Caffe POC

- Deep Learning Configurations
  - Take Caffe framework in POC (Note: LMDB is not involved in POC)
  - 16KB ~ 100KB per picture
  - 20GB totally (around 1M pictures)

- Spectrum Scale configuration
  - RDMA is not enabled yet
  - Sing SSD IOPS: 80K IOPS for 4K random read; 60K IOPS for 4K random write

- Requirements
  - 7000 pics/s for each single client with keeping GPU at full load

- POC Results
  - 7200 pics/s for single client
  - 50000 pics/s for 7 clients concurrently (only 40% Spectrum Scale disk bandwidth used)
Customer TensorFlow/Caffe POC

Cognitive analytics against data from native HDFS

- Existing HDFS customers
- Move data from native HDFS into Spectrum Scale for Cognitive workloads (posix-based)

Object Detection, Computer Vision, Voice Processing, Nature Language Processing…

- 40Gb or 100Gb Infiniband Network
- SSD with high IOPS or NVMe disks
- Hadoop + native HDFS
- ESS GS1S GS2S

HDFS Transparency

distcp
Power your All-in-place Analytics
Reference

• Download the latest HDFS Transparency package from IBM Spectrum Scale developerworks wiki here

• Download the latest Ambari integration mpack from IBM Spectrum Scale developerWorks wiki

• Refer the Big Data and Analytics Guide from IBM KC

• Any questions, mail to scale@us.ibm.com
Thank You