Introduction and Updates to Spectrum Scale HDFS Transparency

Spectrum Scale UK User Group Meeting 2022
London, UK – June 30th, 2022

Dr. Qais Noorshams (IBM)

IBM’s Global Data Platform for File & Object Data

HPC | AI / ML | Analytics | Enterprise | Containers | Backup / Archive

1 Data Access Services

Big Data	Extreme Performance File	High performance object	Network attached	High Performance Containers
HDFS	GDS POSIX	S3	NFS / SMB	CSI CSNA

2 Data Caching Services

Global Data Platform
(powered by Spectrum Scale)

Local Cache | Local Cache | Local Cache | Local Cache

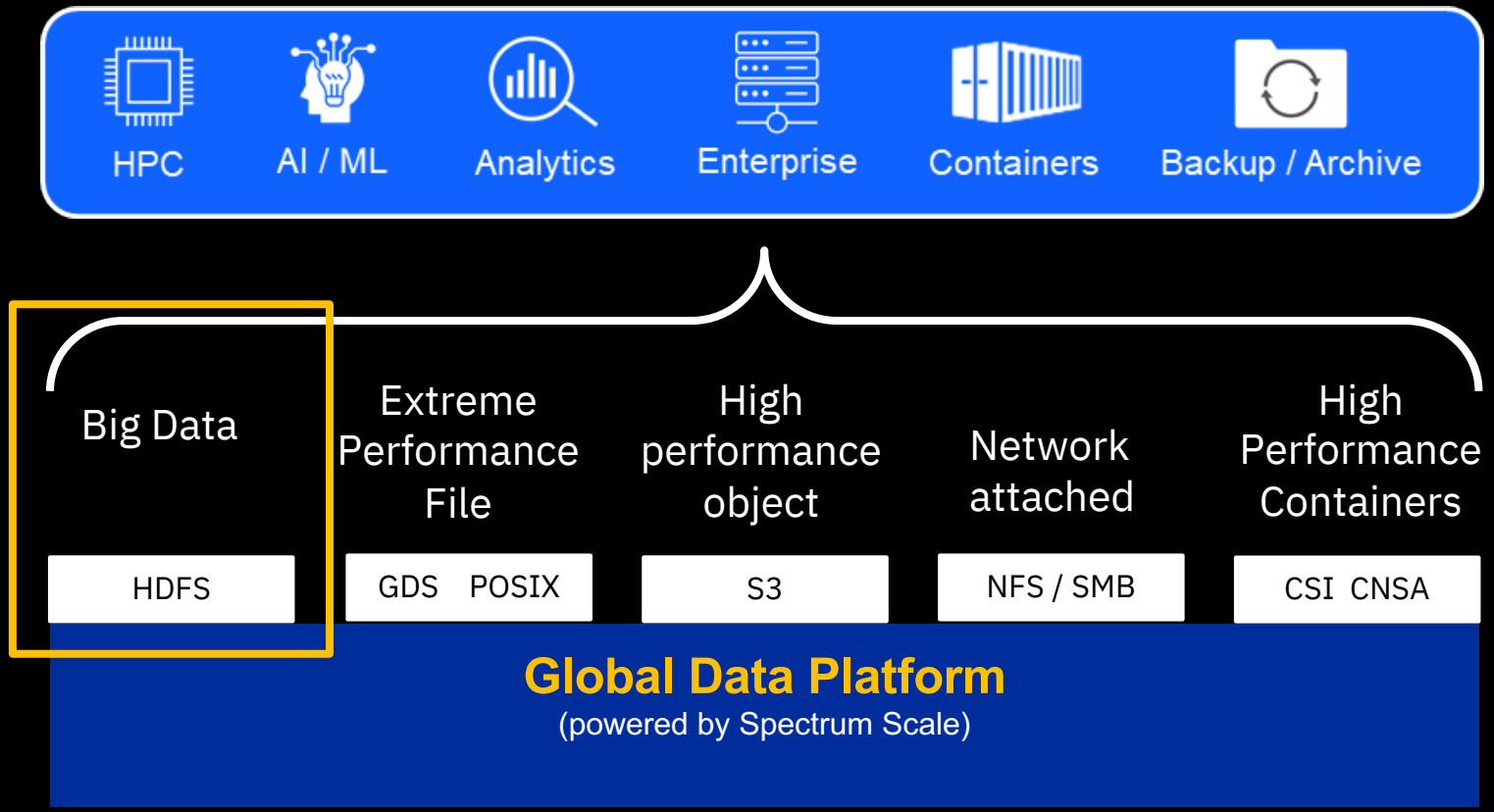
3 Data Management Services

Investment protection	Object Storage	File Storage	NextGen workloads
File & Object Storage (NetApp, PowerScale, etc)	IBM COS	Spectrum Scale	Spectrum Fusion

4 Data Security Services

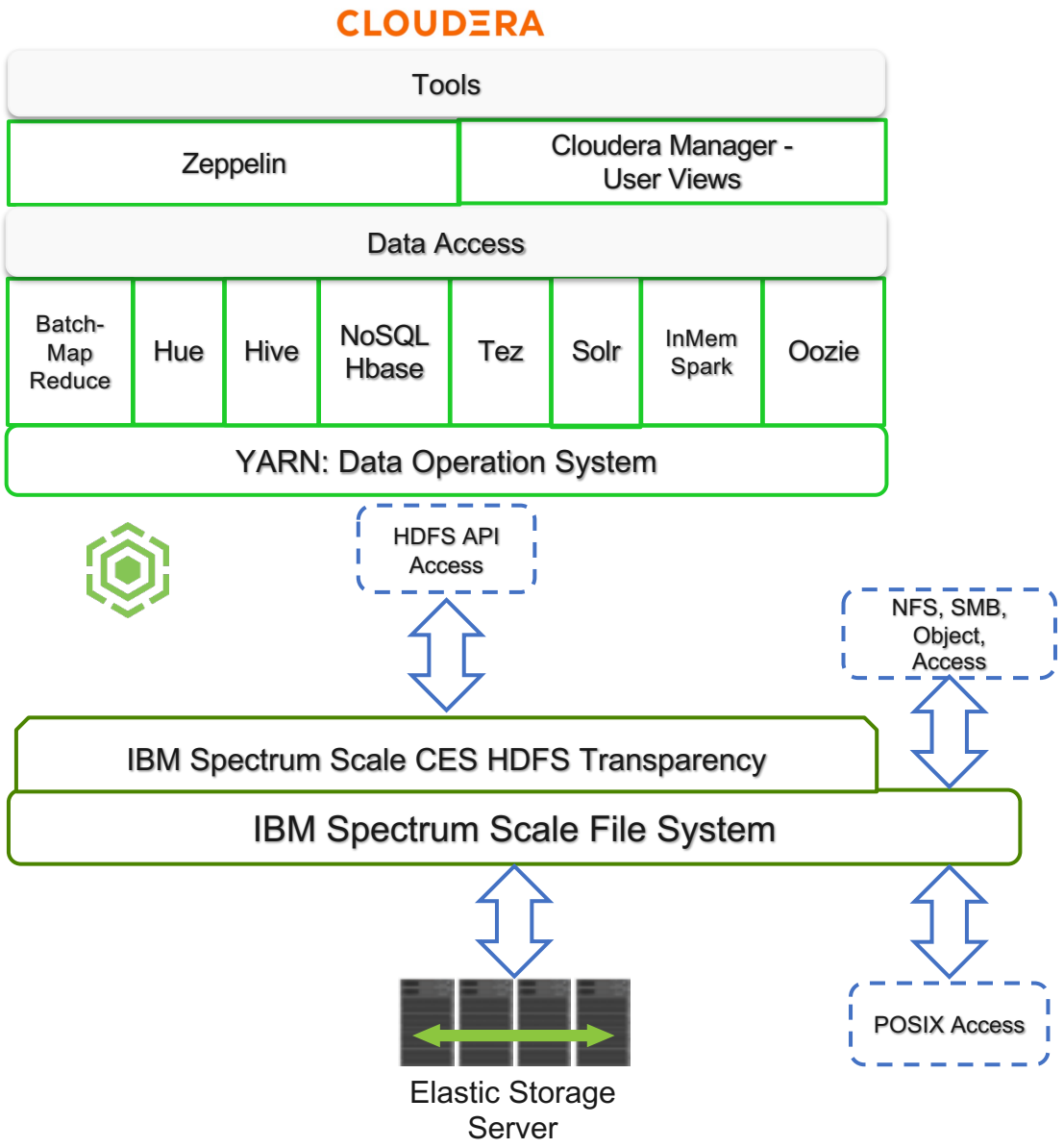
- Identify
- Protect
- Detect
- Respond
- Recover

IBM Spectrum Scale HDFS Transparency



- Analytics and data insights:
  Run AI & ML workloads seemlessly over HDFS on IBM Spectrum Scale
- Best of both worlds:
  Exploit the open source and the IBM Spectrum Scale ecosystem
- Save space, time, and money:
  No data copying, no redundant data replication, built-in data protection

Hadoop distribution: Cloudera



- Provides on premise Hadoop distribution with IBM Spectrum Scale HDFS Transparency using Cloudera Manager (CM) to deploy, manage and monitor the Hadoop cluster
- Industry-wide trend to decouple Hadoop workloads from underlying storage layer
- IBM and Cloudera collaboration in certification with Cloudera Data Platform (CDP) Private Cloud Base distribution

Hadoop distribution: BigTop – Opensource alternative

- BigTop provides deployment, integration and tested certification of the leading open-source big data components
- BigTop is supported by the opensource community
- BigTop may be an attractive alternative for customers who wants to control and manage their Hadoop environment without the need to pay the Cloudera license and support fees

UPDATES TO HDFS TRANSPARENCY IN SPECTRUM SCALE 5.1.3 & 5.1.4

https://www.ibm.com/docs/en/spectrum-scale-bda?topic=summary-changes

HDFS Transparency

Highlights in 3.1.0-9, 3.1.1-6

- Optimized parallelism for DataNode request processing for the performance improvement exclusive to HDFS on IBM Spectrum Scale

Highlights in 3.1.0-10, 3.1.1-8

- Added security fix for CVE-2021-4104 and CVE-2019-17571
- *Note that HDFS Transparency 3.1.0-10 is the last release in the 3.1.0-x stream*

Highlights in 3.1.0-10 efix, 3.1.1-9

- Optimized internal metadata data structures for the NameNode for improved memory efficiency (30%)

Highlights in 3.1.0-10 efix-2 (220316.150114)

- Potential duplicate file names occurring in HDFS Transparency, the POSIX interface of IBM Spectrum Scale is not affected

Highlights in 3.1.1-7

- Support for Java 11

Highlights in 3.2.2-1

- Supports HDFS Transparency 3.2.2-1 for Apache BigTop distribution on RHEL 7.9 on x86

Note:
- HDP support in 3.1.0-x HDFS Transparency
- CDP support in 3.1.1-x HDFS Transparency
- BigTop support in 3.2.2-x HDFS Transparency

CDP 7.1.7 SP1

CDP 7.1.7 SP 1 support for IBM Spectrum Scale

- CDP Private Cloud Base 7.1.7 SP1 is certified with IBM Spectrum Scale starting from IBM Spectrum Scale 5.1.2.2
- Cloudera has designated CDP 7.1.7 as Long Term Support (LTS) release / stable version to be supported for 4 years

Support Stack

- CDH Runtime 7.1.7.1 & Cloudera Manager 7.6.1
- IBM Spectrum Scale 5.1.2.2+, Spectrum Scale CSD unchanged at 1.2.0-0
- HDFS Transparency 3.1.1.8+
- RHEL 7.9, 8.4 on x86 & Power (RHEL8.2 EOS on 4/30/22)

Updates from Cloudera

- There is no in-place upgrade option for HDP to CDP on Power. Only side by side migration is available. Cloudera recommendation is always side by side migration anyway even for x86
- For CDP HDFS to CDP with Spectrum Scale: Only side by side migration is available

Thank you for using
IBM Spectrum Scale!