Introduction and Updates to Spectrum Scale HDFS Transparency

Spectrum Scale German User Meeting 2022
Köln, Germany – October 19/20, 2022
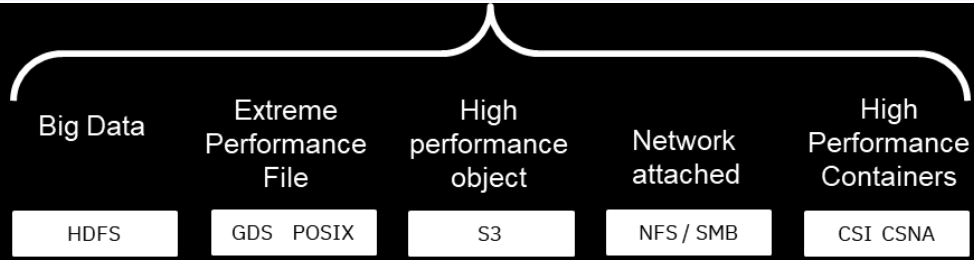
Dr. Qais Noorshams (IBM)

IBM's Global Data Platform for File & Object Data

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

4 **Data Security Services**

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

Identify

2 **Data Caching Services**

Global Data Platform
(powered by Spectrum Scale)

Protect

Local Cache   Local Cache   Local Cache   Local Cache

Detect

Respond

Investment protection   Object Storage   File Storage   NextGen workloads

Recover

File & Object Storage
(NetApp, PowerScale, etc)

IBM COS

Spectrum Scale

Spectrum Fusion

3 **Data Management Services**
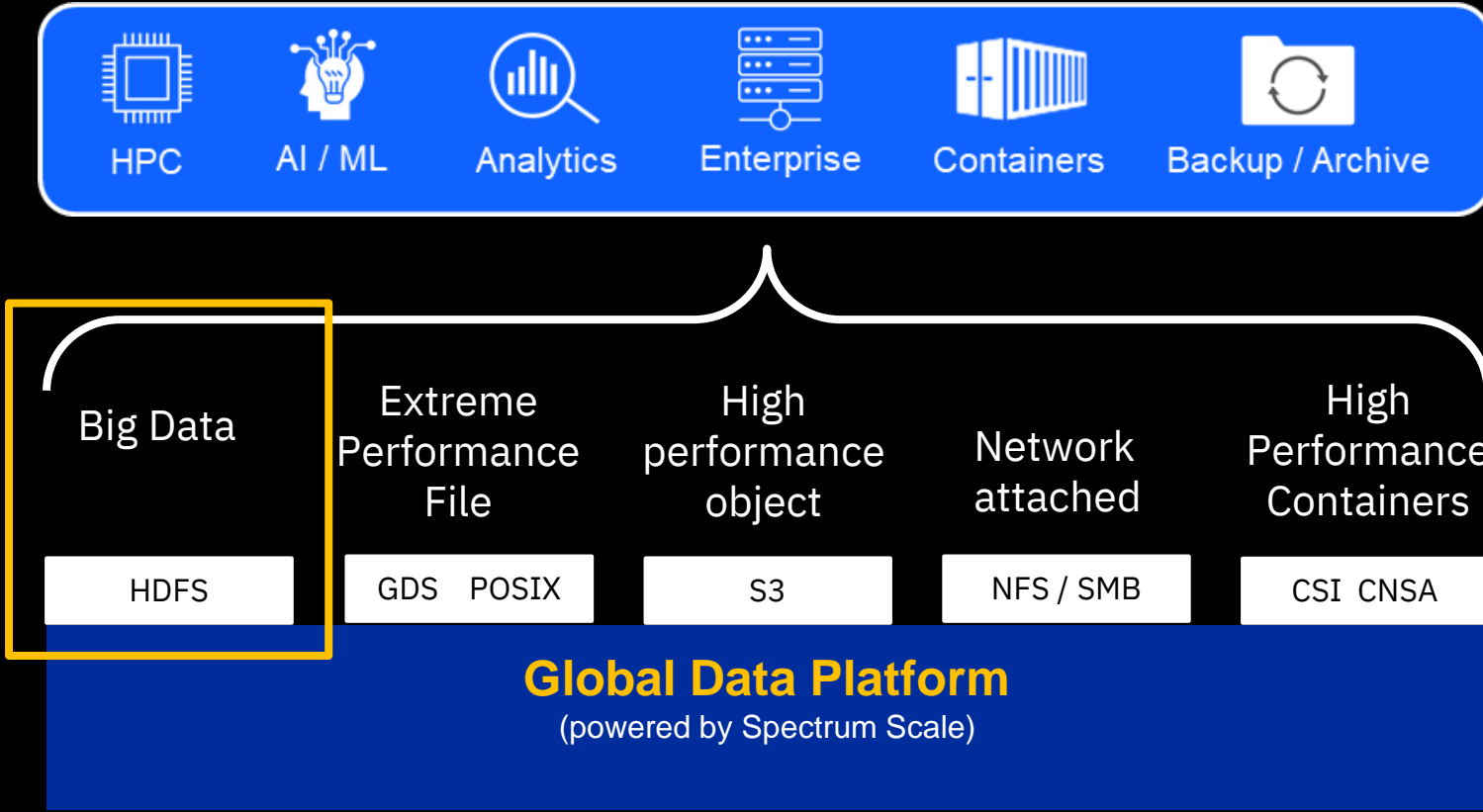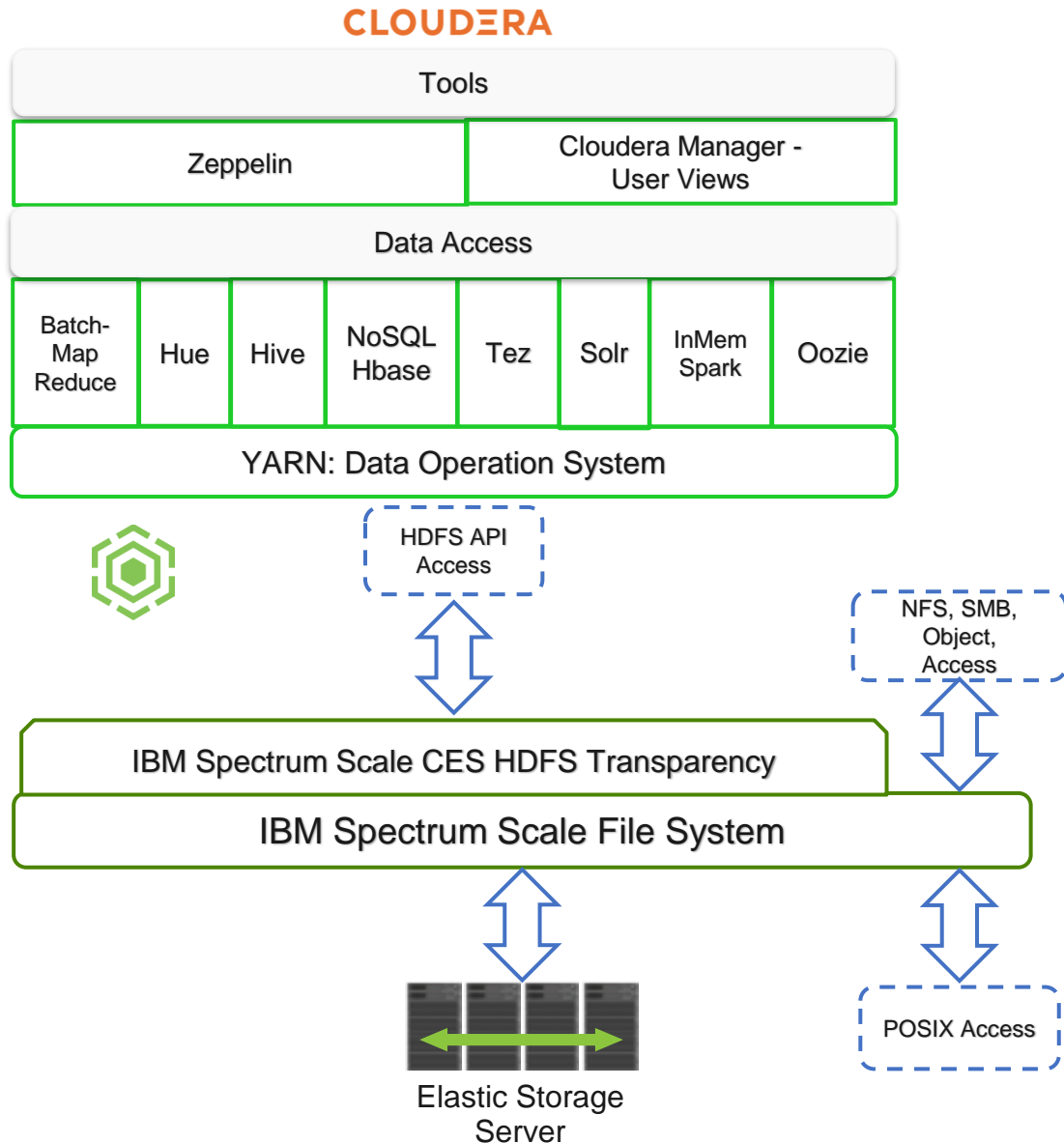
IBM Spectrum Scale HDFS Transparency

- 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

CLOUDERA

Tools	
Zeppelin	Cloudera Manager - User Views

Data Access							
Batch-Map Reduce	Hue	Hive	NoSQL Hbase	Tez	Solr	InMem Spark	Oozie

YARN: Data Operation System

HDFS API Access

NFS, SMB, Object, Access

IBM Spectrum Scale CES HDFS Transparency

IBM Spectrum Scale File System

POSIX Access

Elastic Storage Server

- 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, 5.1.5

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 is not affected

Highlights in 3.1.0-10 efix-3 (220705.102333)

- Optimized parallelism for HDFS lease handling for significant performance improvement

Highlights in 3.1.0-10 efix-3 (220705.102333), 3.1.1-10

- Added security fix for CVE-2022-23305, CVE-2022-23307, CVE-2022-23302, CVE-2020-9488

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

CDP 7.1.7 SP1 & 7.1.8

CDP support for IBM Spectrum Scale

- CDP Private Cloud Base 7.1.7 SP1 is certified with IBM Spectrum Scale starting from v5.1.2.2

- CDP Private Cloud Base 7.1.8 is certified with IBM Spectrum Scale starting from v5.1.4 on x86

- Cloudera has designated CDP 7.1.7 as Long Term Support (LTS) release supported for 4 years

- CDP 7.1.8 will be supported for 18 months. Customers on 7.1.7 LTS can upgrade to 7.1.8, however, will detach from the 4-year LTS support

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

Support Stack

CDP Private Cloud Base	Cloudera Manager (CM)	Cloudera Runtime (CDH)	IBM Spectrum Scale	HDFS Transparency	RHEL x86_64	RHEL ppc64le
7.1.8	7.7.1	7.1.8	5.1.4.0+	3.1.1-9+	7.9, 8.4, 8.6	–
7.1.7 SP 1	7.6.1	7.1.7.1	5.1.2.2+	3.1.1.8+	7.9, ~~8.2~~ (EOS), 8.4	7.9, ~~8.2~~ (EOS), 8.4

Summary

- IBM HDFS Transparency enables HDFS protocol on IBM Spectrum Scale for Big Data workloads

- Hadoop distributions
  - CDP Private Cloud Base is certified with IBM Spectrum Scale
  - Apache BigTop opensource alternative

- Latest HDFS releases include performance & memory improvements and security fixes

- **Join the HDFS breakout session tomorrow!**

Thank you for using
IBM Spectrum Scale!