FlashArray Now Certified with Hortonworks Data Platform v3.0.0

Today is a double whammy with two great announcements, first one is Pure Storage FlashArray with is a now certified storage with Hortonworks Data Platform v3.0.0 and the second one is we are excited to launch DirectFlash Fabric which uses NVMe-oF RoCE protocol. Our NVMe-oF protocol first implementation uses RoCE which stands for RDMA over Converged Ethernet (RDMA = Remote Direct Memory Access). FlashArray//X with DirectFlash technology delivers a great performance which exceeds SAS SSDs.

FlashArray Now Certified with Hortonworks Data Platform

Before I get into what is the importance of this certification and how we can change the big data world, here is the brief history of Hortonworks and their Hortonworks Data Platform(HDP).

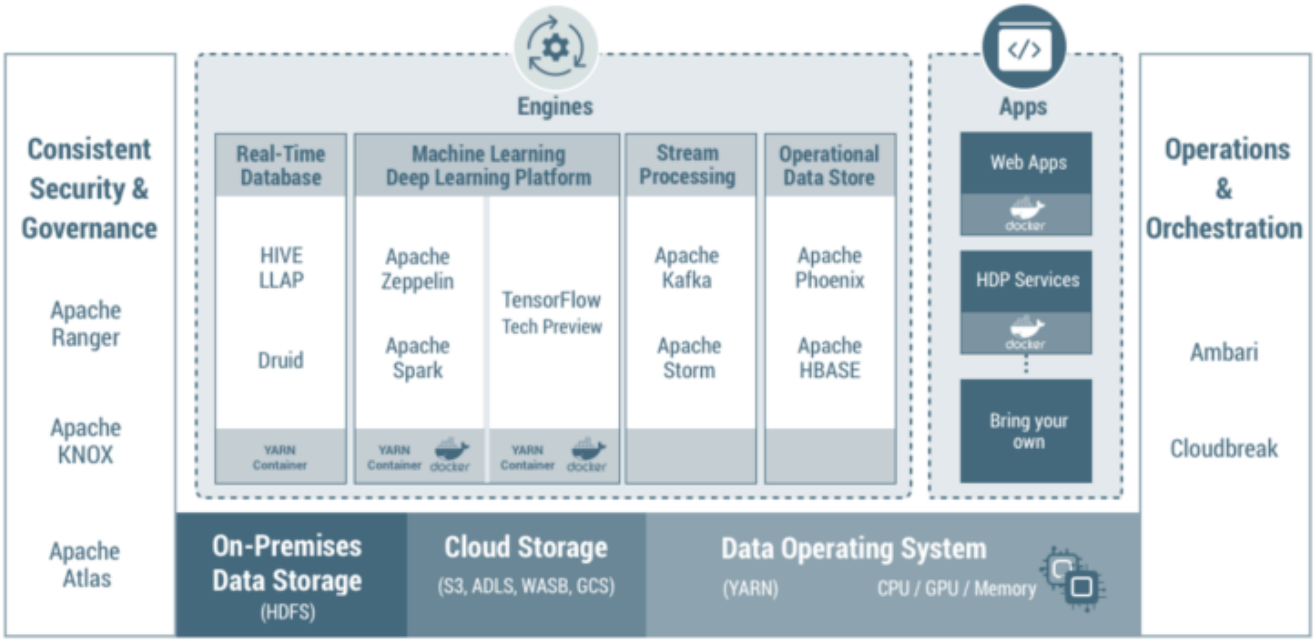
Hortonworks:

Hortonworks was formed in June 2011 as an independent company, funded by $23 million venture capital from Yahoo! and Benchmark Capital. The company employs contributors to the open source software project Apache Hadoop. The Hortonworks Data Platform (HDP) product includes Apache Hadoop and is used for storing, processing, and analyzing large volumes of data. The platform is designed to deal with data from many sources and formats. The platform includes Hadoop technology such as the Hadoop Distributed File System, MapReduce, Pig, Hive, HBase, ZooKeeper, and additional components. Now Hortonworks completed its merger with Cloudera in Jan 2019.

Hortonworks Data Platform:

HDP is a very popular big data platform or at the very least it was one of the two most popular on-prem Hadoop platforms. Hadoop is a collection of open-source applications which uses many servers to solve problems involving lots of structured and unstructured data. It is a framework based on distributed storage (HDFS) and the processing engine called MapReduce. HDFS basically breaks files into large blocks and distributes them across the data nodes in the cluster. HDFS by It then manipulates the data in each node in parallel and reduces the data set by sending the data to the Reducer program to get the final aggregated result. HDP is enterprise-ready, open-source Apache Hadoop distribution based on a centralized architecture. HDP addresses the complete needs of data at rest, powers real-time customer applications, and delivers robust big data analytics that accelerate decision making and innovation. The

latest version HDP delivers new capabilities for the enterprise to enable agile application deployment, new machine learning/deep learning workloads, real-time data warehousing, & security and governance.

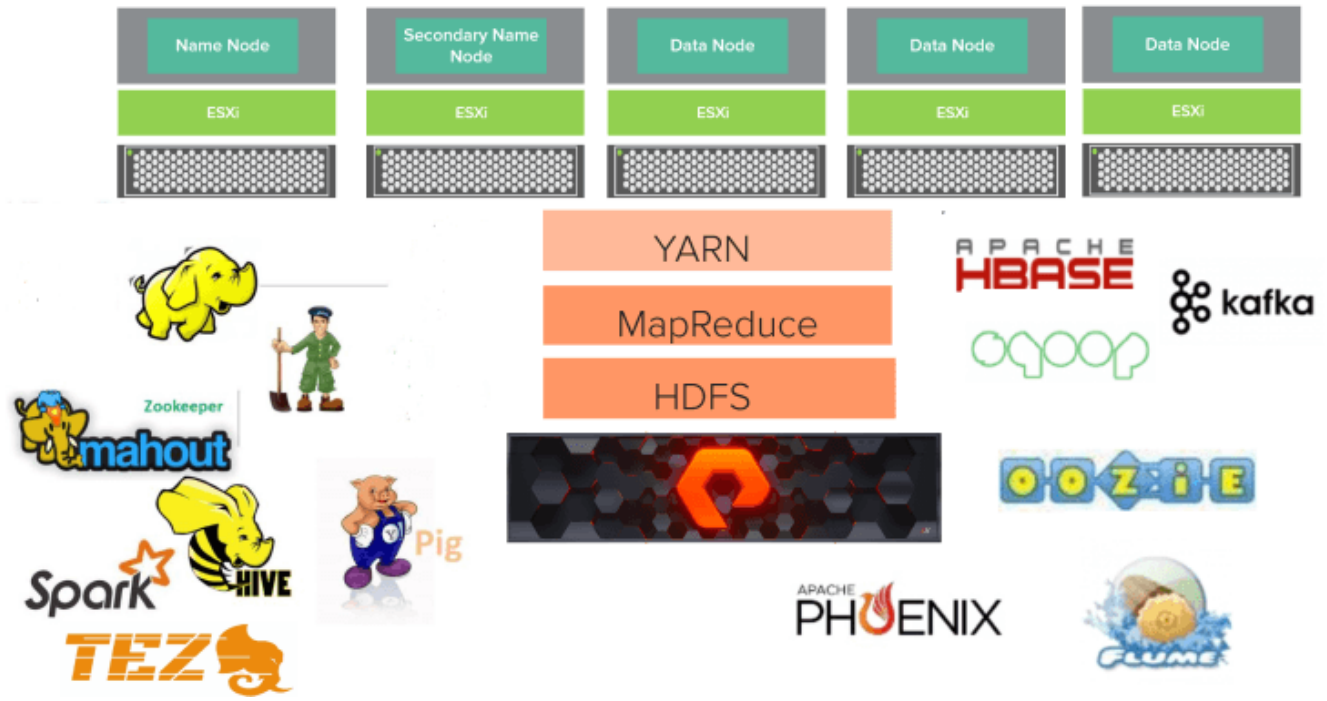


Challenges with Big Data and DAS:

One of the biggest challenges of working with big data is data growth. Data grows constantly and often at a very high velocity. This is where a SAN architecture in which you separate compute from storage performs better than traditional DAS based Hadoop systems. Traditional DAS systems are often notoriously difficult to scale and manage. In contrast, SAN is extremely easy to manage and scale. In particular, the Pure Storage FlashArray is able to scale from one a basic model to higher capacity and performance model non-disruptively. Now with FlashArray//X DirectFlash Fabric, its performance exceeds SAS DAS SSDs performance, in terms of latency and throughput. This claim of better performance of FlashArray//X DirectFlash Fabric is based on the tests we have done comparing Apache Cassandra, MongoDB, MariaDB on FlashArray//X DirectFlash Fabric and SAS SSD DAS. For Apache Cassandra we found out that writes latency and operations per second were 30% better for Apache Cassandra based on FlashArray//X DirectFlash Fabric than Cassandra cluster on SAS SSD DAS. This makes FlashArray//X a very good choice for the Hortonworks Data Platform.

Reference architecture:

For certification purposes, we have built a 5 node cluster as shown below with Pure Storage FlashArray//X with DirectFlash Fabric. The nodes were virtualized with VMware VSphere 6.5 and with Red Hat 7.x operating system. The HDP 3.0.0 was installed on this five node cluster and QATS was used for certification/testing. Basically, QATS (Quality Assured Testing Suite) is a product integration certification program designed to rigorously test Software, File System, Next-Gen Hardware and Containers with Hortonworks Data Platform (HDP).



The QATS suite validates the following aspects for the HDP components:

1. Complete functionality
2. High Availability
3. Security aspects
4. Kerberos
5. Ranger

For eg: For MapReduce it will test the following aspects:

Functional	HA	Kerberos	Ranger	Wire Encryption	Transparent Data Encryption
YES	NA	YES	NA	NO	NO

Functional scenarios **:**

- NN Bench Tool
- MR Bench Tool
- Terasort Job: Generate input with TeraGen. Sort with TeraSort.
- JHS: for a different state of Jobs
- Compression: For wordcount job, different types of Compression Codec for different Compression Types
- Kerberos: Above Functional scenarios are run in a Kerberos enabled environment.

Hortonworks certified components:

The following components of HDP are certified to run on FlashArray//X.

HDP 3.0.0 Certified components

HDFS	MapReduce	Yarn	Zookeeper	Tez
Hiveserver2	Hiveserver2-llap	Hiverserver2concurr	HBase	Phoenix
Phoenix-queryserver	Phoenix-qs-concurr	Spark2	Spark-hive	Sqoop
Kafka	Oozie	Pig	Knox	Atlas
XASecure	Storm	Accumulo	Superset	Druid

Below is each component version and description:

Components	Version	Description
HDFS	3.1.0	Apache Hadoop Distributed File System
YARN	3.1.0	Apache Hadoop NextGen MapReduce (YARN)
MapReduce2	3.1.0	Apache Hadoop NextGen MapReduce (YARN)
Tez	0.9.1	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
Hive	3.1.0	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
HBase	2.0.0	Non-relational distributed database and centralized service for configuration management & synchronization
Pig	0.16.0	Scripting platform for analyzing large datasets
Sqoop	1.4.7	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
Oozie	4.3.1	System for workflow coordination and execution of Apache Hadoop jobs.
ZooKeeper	3.4.6	Centralized service which provides highly reliable distributed coordination
Storm	1.2.1	Apache Hadoop Stream processing framework
Accumulo	1.7.0	Robust, scalable, high performance distributed key/value store.
Infra Solr	0.1.0	Core shared service used by Ambari managed components.
Ambari Metrics	0.1.0	A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster

Atlas	1.0.0	Atlas Metadata and Governance platform
Kafka	1.0.1	A high-throughput distributed messaging system
Knox	1.0.0	Provides a single point of authentication and access for Apache Hadoop services in a cluster
Log Search	0.5.0	Log aggregation, analysis, and visualization for Ambari managed services. This service is Technical Preview.
Ranger	1.0.0	Comprehensive security for Hadoop
Ranger KMS	1.0.0	Key Management Server
SmartSense	1.5.0.2.7.0.0-846	SmartSense – Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
Spark2	2.3.1	Apache Spark 2.3 is a fast and general engine for large-scale data processing.
Zeppelin Notebook	0.8.0	A web-based notebook that enables interactive data analytics. It enables you to make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
Druid	0.12.1	A fast column-oriented distributed data store.
Kerberos	1.10.3-30	A computer network authentication protocol which works on the basis of ‘tickets’ to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner.
Superset	0.23.0	Superset is a data exploration platform designed to be visual, intuitive and interactive. This service is Technical Preview.

Summary:

FlashArray//X, now with DirectFlash Fabric, using NVMe over RoCE, is going to be a game changer in the Big Data world. FlashArray//X is very simple to use and deploy, which in-turn eases implementation and encourages adoption. It scales non-disruptively as your data grows and accelerates your project timelines. FlashArray//X with DirectFlash Fabric is certified with Hadoop Data Platform 3.0.0.