WHITE PAPER

AI AND BIG DATA IN THE SAP WORLD

PREPARE YOUR SAP ENVIRONMENT FOR THE FUTURE WITH PURE

PURESTORAGE®

TABLE OF CONTENTS

INTRODUCTION

In today's world, more than one billion people are active on social networks, with the collective power to instantly make or break brands. The number of connected devices is expected to be 50 billion by 2020, and the data produced by those devices is staggering. Meanwhile, companies are in the midst of a perfect storm of technological innovation: the convergence of cloud, mobile, social, Big Data, and AI is reshaping the future of business and acting as a catalyst to empower individuals as employees, consumers, and as citizens to increase their reach and relevance. Unlocking the potential of this data presents breakthrough opportunities for businesses across a wide array of use cases, including Analytics, IoT, and Machine Learning.

Pure Storage data-centric solutions include SAP HANA certified enterprise data storage. Pure customers can run SAP products that are tightly optimized to solve modern-day data operations challenges. Pure enables these solutions to process massive volumes of data, to deliver high velocity streaming data, and to facilitate automated data movement and data visualization. Modern data management challenges also involve supporting the vast variety of data types – including structured data, semi-structured (or text-based) data, as well as unstructured data, such as image, audio, video, geo-spatial, and graph data. The Pure Storage SAP solution specifically addresses all these data management challenges from the ground up and at scale, partnering with the best infrastructure companies to ensure a complete solution architecture that delivers optimal application results.

ABOUT THIS DOCUMENT

In this paper, we will look at how one can manage and process large volumes of data within the enterprise computing framework (HANA) with data tiering on Pure Storage. We'll also look at how we can deliver enterprise-grade analytics to distributed frameworks such as Hadoop and beyond. Big Data is 'Big' for a reason: now, more than ever, customers are forced to consider methods for delivering large amounts of fast storage. Data warehousing requirements in the 50 PB range are not uncommon. Companies want scalability and reduced TCO – without compromising enterprise-grade analytic capability, security, compliance, and user-friendly consumption of data. In this document, we will look at how the Pure Storage SAP solution delivers these requirements, and also discuss how Pure has a vision to deliver these capabilities in distributed computing environments such as Hadoop. Finally, we will discuss how AIRI™, the new Pure Storage AI product in partnership with NVIDIA®, enables SAP customers to utilize and take advantage of new data types.

PURE STORAGE PRODUCTS

FlashArray

Pure Storage Flash Array is an SAP HANA certified enterprise storage array and can be deployed in hardware infrastructure environments following the SAP HANA tailored data center integration standards.

Pure Storage FlashArray//X is the world's first enterprise-class, all-NVMe flash storage array. It represents a new class of storage – shared accelerated storage (which is a term coined by Gartner) – that delivers major breakthroughs in performance, simplicity, and consolidation of mixed workloads. With latency as low as 250 µs, FlashArray//X's all-NVME architecture brings new levels of performance to mission-critical business applications. With built-in Purity ActiveCluster, more applications can now benefit from the simple to deploy, always-on business continuity provided via active/active metro stretch clustering. Pure DirectFlash™ with 100% NVMe also enables unprecedented performance density: //X currently supports ultra-dense 18.3TB DirectFlash modules, for maximum density with full performance. In addition, Purity's always-on and QoS features mean you can consolidate radically diverse applications without fear of I/O contention.

FlashStack™

Pure Storage FlashStack (in partnership with Cisco) is a flexible, all-flash converged infrastructure (CI) solution that brings the flash revolution to your data center, faster. It combines the latest in compute, network, and storage hardware, as well as virtualization software, into a single, integrated architecture that reduces time to deployment, lowers overall IT costs, and reduces deployment risk. Highly efficient components reduce the costs associated with power, cooling, and data center space. Based on 100% flash storage, FlashStack provides the performance and reliability mission-critical applications like SAP HANA need.

FlashBlade™

With FlashBlade, it's possible to efficiently store and analyze huge volumes of unstructured data on a simple, scalable platform that can significantly reduce your data center footprint in terms of space, power, and cooling requirements. More importantly, FlashBlade allows you to unlock new use cases and capabilities so you can take full advantage of your data.

FlashBlade redefines the unstructured data platform for the new stack. With FlashBlade, Pure Storage has once again turned the storage industry on its ear as entire racks of legacy systems may be consolidated into 4U of FlashBlade. The performance, density, and simplicity FlashBlade provides is a true game-changer. High performance has always been – and continues to be – the calling card for flash storage.

However, as flash technology enters the mainstream with increased capacity, density, and scalability, organizations are looking at other use cases. Unstructured data, which is growing at a much faster rate that structured data, is equal parts opportunity and challenge for SAP customers looking to integrate these types of data into their SAP systems. FlashBlade's highly-parallel architecture and massive bandwidth provide a perfect scale-out solution for Big Data, AI, and machine learning applications.

AIRI™

AIRI is the industry's first complete AI-ready infrastructure, architected by Pure Storage and NVIDIA® to extend the power of NVIDIA® DGX™ systems. Powered by FlashBlade storage and NVIDIA DGX-1 servers, AIRI offers enterprises a simple, fast, and future-proof infrastructure to meet their AI demands at any scale.

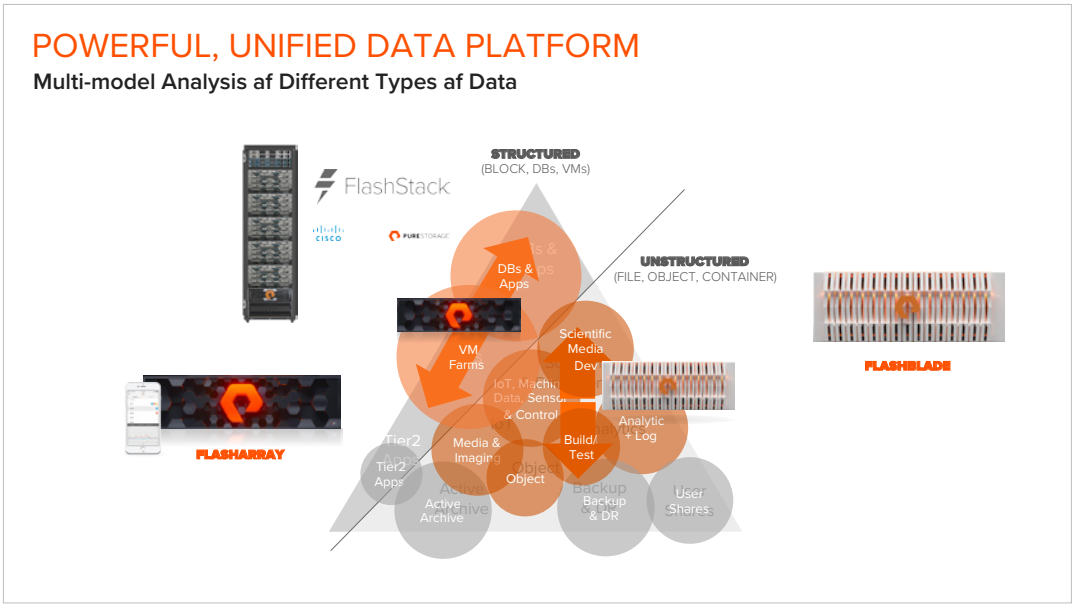

FIGURE 1. Pure Storage Data-Centric Architecture for different types of data

DATA TIERING

Data tiering (not to be confused with SAP HANA Dynamic Tiering, which is itself a data tiering strategy) is the assignment of data to various tiers of storage media based upon data type, operational usefulness, performance requirements, frequency of access, and security requirements of the data.[1] Tier 1 is essentially used to store mission-critical data used for processing and real-time analytics, while Tier 2 may be used to store older operational data which is less frequently accessed, and Tier 3 might store older voluminous data used to deliver predictive insight. As an example, Tier 1 could be used to store structured data (e.g., financial accounting documents, 0-24 months old); Tier 2, older structured data (e.g., 25-60 months old); and Tier 3, unstructured, semi-structured, and streaming data.

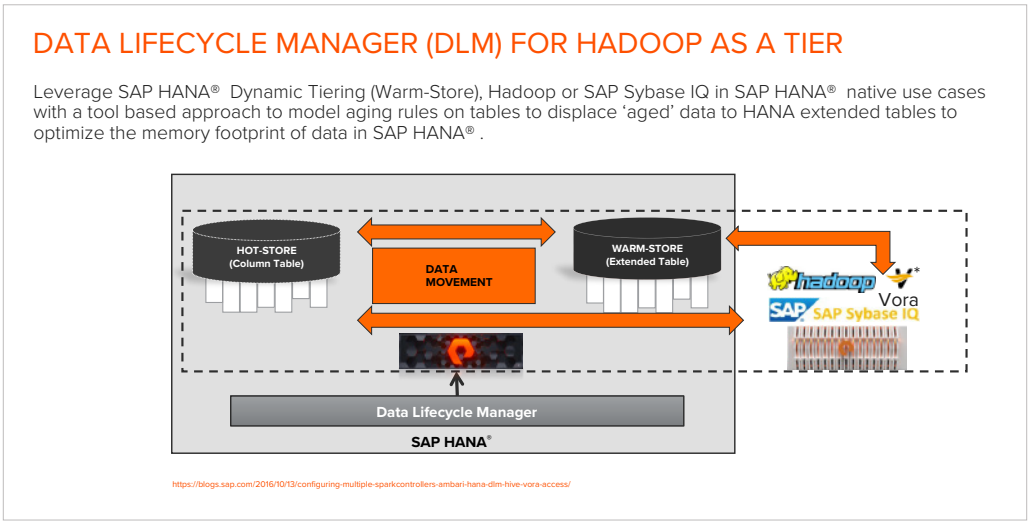

FIGURE 2. SAP DLM on Pure Storage

SAP HANA Dynamic Tiering

Dynamic Tiering is an add-on option to the SAP HANA database and is a native data-aging solution to manage larger volumes of less frequently accessed data. Dynamic Tiering allows for warm data to be moved to extended storage, which results in reduced size and better performance of the in-memory database.[2]

SAP continues to invest and develop its HANA Dynamic Tiering solution to help with data tiering. If you are interested in HANA Dynamic Tiering on Pure Storage, visit this SAP webpage. In this paper, we will be focusing on data tiering in the SAP world and how Pure Storage solutions deliver extreme performance on every data tier while enabling mixed workload consolidation and offering high ROI.

Why the Surge of Interest in Data Tiering?

As the variety and volume of data grows, more and more companies are facing increased infrastructure costs, with smaller budgets available for innovation. As a society, we create .5 quintillion bytes of data on a daily basis (2.5

[1] https://blogs.saphana.com/2018/03/20/whats-new-in-sap-hana-2-0-sps-03/

[2] https://www.purestorage.com/content/dam/flashstack/pdf/ReferenceArchitecture/Pure-Storage-FlashStack-for-IoT-Solution.pdf

followed by 18 zeros). With this deluge of data, it is imperative that we be able to store data cost-effectively, in order to unlock its value and facilitate predictive insight. The burden of managing voluminous data cost-effectively is shared among many SAP companies looking to expand and diversify their data-sets.

Data Tiering in the Context of Temperature

Designing a data tiering architecture based on temperature alone is usually not sufficient. This is particularly true in the case of an ERP platform, where additional logic is required to determine the 'business completeness of the data'. For instance, one could have invoices which are older than 1 year that remain unpaid; although these invoices may not have been accessed in quite some time, it does not make sense to move that data to another tier based on database optimization hints (statistics) exclusively. Additional business logic is required, which is typically delivered at the application layer.

Consideration for Regulatory/Legal Requirements

Careful consideration must be given to data tiering when one is subject to regulatory/legal requirements. Often an ERP system (e.g., SAP SoH, SAP S/4HANA) is perceived as the legal system of record. Thus, before one offloads data in an effort to reduce IT costs, one should consider the requirements around defensible deletion of data, and the ensuing ramifications in the event that defensible deletion cannot be demonstrated. However, for other types of applications, such as a data warehouse (which is not perceived as the legal system of record due to the fact that the data is sourced from various applications), data tiering could be performed using DB optimization hints/time-slices.

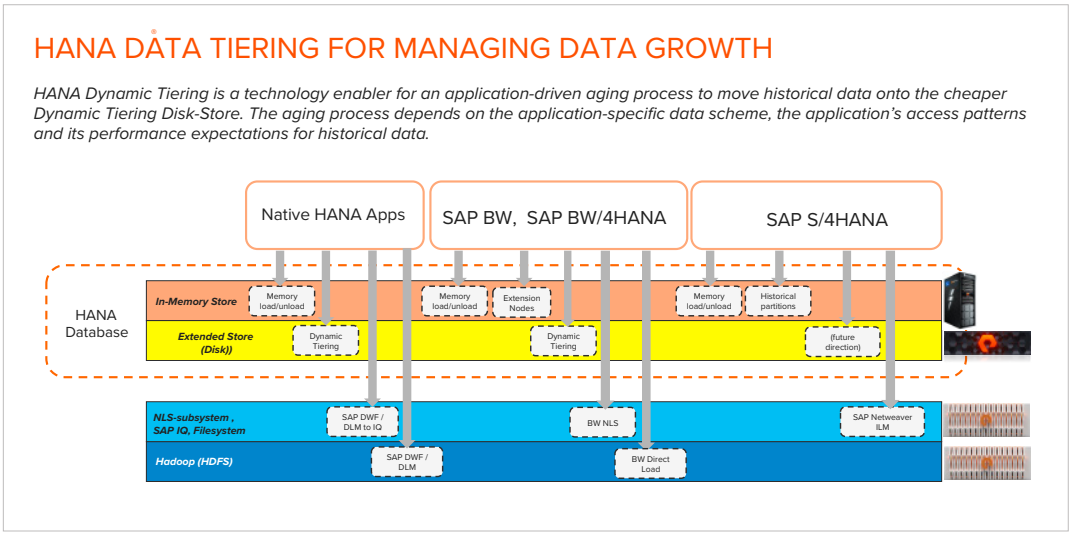

FIGURE 3. HANA Dynamic Tiering and Pure Storage

DESIGNING AN EFFECTIVE DATA TIERING STRATEGY

To design an effective data tiering/distributed computing strategy which has consideration for IT, business, and legal requirements, one should look at a variety of different factors, as follows:

DATA VOLUME – IS THE DATA RANGE...

- 1-100TB ?
- 100 TB+ ?

DATA TYPE – IS THE DATA...

- Streaming?
- Unstructured?
- Semi-structured?
- Structured data?

DATA VALUE

The type and value of the data should determine the storage solution to be used and also where the data should begin its lifecycle. Is it...

- High value low volume, required for real-time analytics?
- Medium volume, medium value?
- Is it high volume, low value (from a real-time analytics perspective?)
- Is the data required for regulatory requirements? And if so, for how long?
- Does the data need to be segregated as per a corporate records retention policy?
- Is the data required to be preserved for legal hold, etc.?

When one can answer these questions, the decision as to where to store and process the data becomes clearer. Business and legal requirements should be the primary driver on where to store the data.

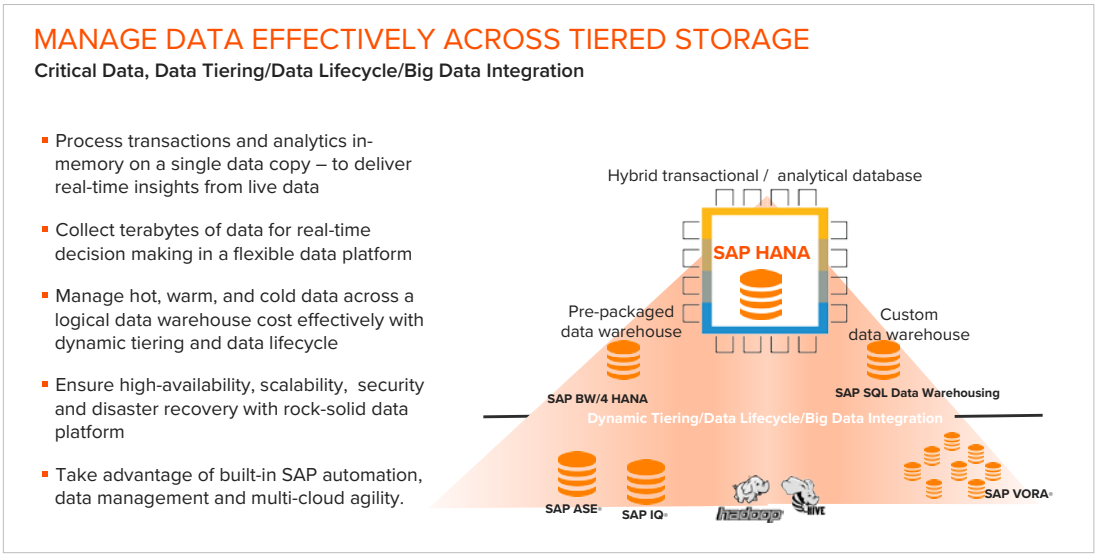

FIGURE 4. Manage SAP data across tiered storage

Perception vs Reality

SAP HANA may often be perceived as expensive by companies, and thus they look at what initially one might consider a cheaper storage tier, such as Hadoop, for the management of data. What is often overlooked is the fact that the HANA platform offers a multitude of capabilities which are currently beyond the reach of many competitors and customers. SAP HANA already offers data reduction due to columnar storage and compression and a wide variety of data management capabilities by means of data tiering. Pure Storage solutions complement SAP's advantages, and even expand on them, with industry-leading data reduction and total compression rates that other all-flash vendors cannot touch. Pure Data-Centric Architecture frees SAP data from limitations – such as running only in Hot Data or siloed in one cloud; instead, you might take advantage of different clouds based on business needs. For more information click here.

HADOOP – DATA LAKE OR MURKY SWAMP?

Open Source

Hadoop is delivered by the Apache Software Foundation (ASF), a charitable organization, funded by individual donations and corporate sponsors. Apache is comprised of an all-volunteer board which oversees more than 350 leading open source projects with a mission to provide software for the public good. Hence, Hadoop can be used with distributed storage for processing of very large data sets on computer clusters built from commodity hardware. Hadoop software is generally available for 'free', and what one pays for is support on that software.

Hadoop – The Dream

In 2003, Google published a seminal paper titled "The Google File System". This paper, in combination with another on MapReduce, inspired the creation of Hadoop and the Hadoop File System (HDFS), which spawned a multi-billion dollar industry and forever changed the course of history for Big Data. For the first time, the world was given a recipe to tap into massive amounts of data for insight using a relatively simple programming model.

Hadoop – The Reality

With Hadoop beckoning customers with attractive features for storing and analyzing data, why would companies not want to offload their data from Tier 1 storage? Let us take a closer look at some of the challenges around Hadoop deployments:

LACK OF ENTERPRISE-GRADE ANALYTICS

The distributed (granularity of storage) nature of the data stored within Hadoop often makes it difficult, if not impossible, to provide enterprise features and functions, such as drill-down capabilities, which companies are accustomed to in the traditional RDMS world.

SECURITY/DATA GOVERNANCE CONCERNS

Security, data breaches, and compliant destruction of data are major concerns for enterprise companies, especially those operating in regulated environments. The Apache communities, as well as distributors, have picked up the pace to address these concerns with open-source projects such as Project Atlas (DGI initiative), Apache Knox (Access control and authentication), and Apache Sentry.

HIGH AVAILABILITY ISSUES

Data in Hadoop is distributed across multiple nodes, but there is only one NameNode in the cluster (Hadoop's metadata server) that knows the data distribution. All applications have to go through this single NameNode to access data. This makes NameNode both a performance bottleneck and a single point of failure.

DISTRIBUTED COMPUTING CHALLENGES

Hadoop stores three copies of everything. Since traditional storage scaled linearly, there is a risk of over-provisioning compute capacity. This can result in poor CPU utilization.

SPECIALIZED PROGRAMMING SKILLS REQUIRED

To unlock the value of data stored in Hadoop, specialized programming skills are required, and currently demand outstrips supply.

CHALLENGING PLATFORM TO MANAGE

Large environments can run to the TB-PB range, with hundreds of servers, plus the myriad tools required to support administration, monitoring, security, and data governance.

IS COMMODITY HARDWARE THE SOLUTION?

The fact that Hadoop leverages commodity hardware instead of certified appliances draws immediate attention to it as a platform for managing and processing infinite volumes of data. However, commodity hardware by itself will not allow Hadoop to leverage Big Data for data tiering as well analytics, AI, and the other new innovations required for the digital enterprise. At Pure, we focus all our engineering effort and IP around being the leaders in all-flash technology. Through industry-leading software and hardware, we provide solutions that deliver unprecedented workload consolidation and effortless scale to meet the needs of the modern enterprise.

In summary, while Hadoop certainly has attractive capabilities (scalability, flexibility, etc.), simply offloading data to Hadoop may not meet all business and legal requirements. At Pure Storage, we understand the challenges of operating an effective data lake and bring the necessary insight and expertise to clear the murky swamp and deliver an optimized data lake. But this is only half of the story. SAP customers can take advantage of FlashBlade as a target-system for data-tiering as well as a source-system for Analytics and AI.

MODERN DATA CHANGES EVERYTHING

The modern data gold rush, fueled by a surge of real-time applications and AI, is just getting started. While Hadoop was the main widely available analytics tool for Big Data a decade ago, data scientists today have a plethora of tools at their disposal. Apache Spark is a real-time streaming framework that's simpler and more powerful than Hadoop. It is becoming the de facto standard for Big Data processing and analytics, and there are already integration capabilities between Spark – an in-memory data processing engine – and SAP Vora and SAP Data Hub. SAP Vora extends Apache Spark by providing additional business functions such as hierarchy support, currency translation, and advanced SQL push-down. Kafka is a real-time messaging tool for any file size, small or large. Hive offers a SQL-like interface that results in random, not sequential, accesses. And the list goes on. All these new data types are becoming more critical to SAP companies looking to gather data from new sources and run analytics on them. While SAP HANA is very successful at running mission-critical analytics, these new data types allow you to gather data rapidly and run powerful analytics.

FIGURE 5. Data is now different

The driving force behind the shift in Big Data is simple. The explosion of new technologies, industries – and perhaps the fourth industrial revolution – is fueled by novel tools built around the singular fact that unstructured data is indeed unstructured: it takes on many different and ever-changing forms. Big Data is no longer large, sequential, batched, and fixed. For machine learning training, the assumptions about data are the polar opposite of those underlying the

traditional HDFS and DDAS architecture. Training data must be accessed randomly, not sequentially, and files are often small in size.

> Data is now truly dynamic. Thus a new class of data platform is needed, one built for the needs of modern, dynamic data – and architected for the unknown.

What if you could simply deploy a new type of storage infrastructure for SAP that is able to handle the dynamic nature of modern data and analytics? Such a storage system has never been invented before, simply because architecting for the unknown is an extremely difficult proposition.

FlashBlade for New Data Types

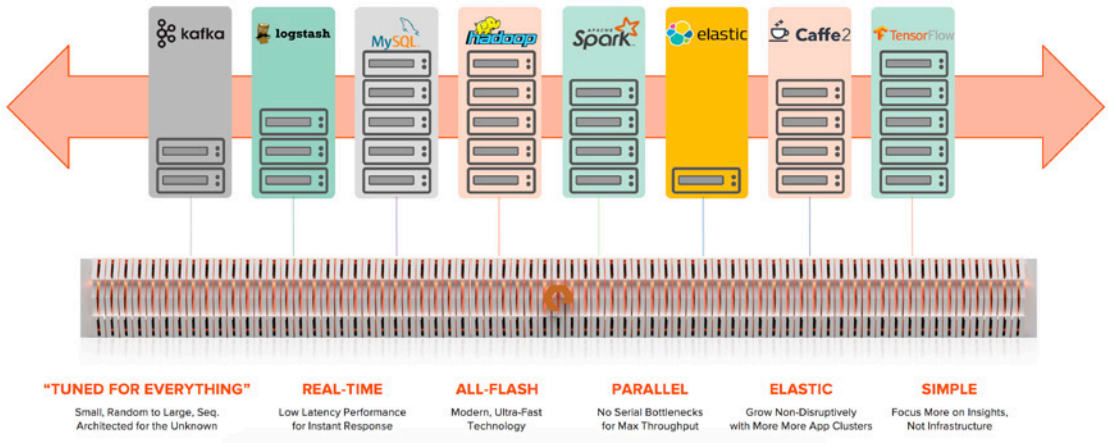

FIGURE 6. FlashBlade handles new data types

FlashBlade is tuned to handle a variety of types of data: from small, metadata-heavy to large files, with random or sequential access patterns, accessed by up to 10s of thousands of clients, all requiring real-time response, without the need to constantly retune the storage platform. From the ground-up, hardware to software, FlashBlade is built with a massive parallel architecture to deliver maximum performance for any modern workload today – and into the future.

FlashBlade allows data pipelines to grow dynamically with the needs of data science and engineering teams. Building on the Pure Storage for SAP modernization story that started with FlashArray for SAP HANA, FlashBlade connects ERP (S/4HANA) and analytics (BW/4HANA) to fast-moving Big Data, AI, and the IoT world that digital enterprises are looking at to automate business and manufacturing processes, remotely monitor and control operations, optimize supply chains, and conserve resources.

While the attractiveness of Big Data for enterprises is well understood, storage vendors have not caught up to the fact that it demands a new approach and a new architecture to properly take on large projects. When Pure Storage introduced FlashBlade, the goal was not just to deliver transformative IO performance, but also metadata performance, massive linear scalability, and economics.

SCALABLE

FlashBlade was built to scale effortlessly and linearly from small deployments to very big deployments, with great care taken to ensure that every dimension of the system scaled elegantly and linearly. IO performance, bandwidth, metadata performance, NV-RAM, protocols, user connections – everything had to scale with the system as it grew.

COST-EFFECTIVE

Storage of data can be expensive, especially when data resides in memory. From a business and economics perspective, it does not make sense to store all data within memory. Hence companies with large data footprints are often attracted to flash storage, where data temperature (i.e., ease of access in the storage tier) can still be relatively warm without storing it in main memory. FlashBlade starts with hardware architected for highly-parallel and high bandwidth data access – and a particular focus on simplicity.

FLEXIBLE

Pure Storage FlashBlade allows SAP companies to tap into a variety of data sources to gain valuable business insights from data sources such as social media, log, clickstream data, etc. Digital enterprises running SAP need to integrate data from many different sources to make correct decisions at a faster rate. While SAP S/4HANA continues to be the core and backbone of these organizations, they must make sure they process and integrate different types of data rapidly. FlashBlade is built for the era of Big Data and cloud, and able to support a plethora of real-time applications and data types.

RESILIENT TO FAILURE

FlashBlade workloads are protected with service assurance: high mixed workload performance, and full performance even through failures and upgrades. No more tuning of anything – ever. This is highly important in the SAP world, where what can seem to be insignificant workloads eventually become critical parts of various data integrations.

FLASHBLADE DELIVERS SAP DATA TIERING

SAP HANA combines database, data processing, and application platform capabilities in-memory. The platform provides libraries for predictive, planning, text processing, spatial, and business analytics. It delivers the most integrated and sophisticated platform on the market for intelligent data tiering and distributed computing. The capabilities are available from data inception to final destruction. At Pure Storage, we understand that **not all data is equal**, and that is how we designed and delivered FlashArray, FlashBlade, FlashStack, and AIRI.

CATEGORIZING DATA IN ACCORDANCE WITH OPERATIONAL RELEVANCE

Pure Storage offers an intelligent, cost-effective, tiered approach for managing data from terabyte to petabyte scale, allowing for distributed computing across various tiers. HANA data management capabilities can take advantage of this storage model and be leveraged independently of each other, based on business needs.

BUSINESS BENEFITS

- TCO reduction
- Integrated Big Data platform
- Faster processing of data in accordance with SLAs
- Segregation of data based on business value/legal requirements
- Access to data when needed
- Significant footprint reduction
- Ongoing control of infrastructure costs
- Distributed computing
- Predictive insight

The following is an overview of the HANA multi-temperature storage options with consideration for data use cases:

Data Temperature	Product Feature/ Integration option	Technology	Sample Use Cases	SAP BW on HANA	SAP Business Suite on HANA	SAP HANA Native	Data in the context of Temperature	Possible actions
	SAP HANA In-Memory	SAP HANA in-Memory	Ideal for real-time analytics/streaming data e.g. Stock tick data streamed into **SAP HANA** for immediate price fluctuation analysis and trading actions	✓	✓	✓	Hot data = Active/ operationally-relevant data stored within HANA memory. Hot data is frequently accessed and has higher performance requirements.	Write, Read, Update, Delete
	SAP HANA Dynamic Tiering	Tightly Integrated disk-based columnar technology	Big Data/Petascale extension - Best suited for data which does not require high speed RAM processing e.g. Historical stock price data stored in **HANA extended tables** for trend analysis and portfolio management	(✓1)	*	(✓2)	Warm data = = Active data Integral to the operation of the platform. Warm data may be older and not queried often, but is still online and available for update.	Write, Read, Update, Delete
	Data Aging	Tightly Integrated disk-based columnar technology	Financial accounting data which is closed/cleared and older than three years	x	✓3	○	Data which is closed/cleared and is moved to cold partitions on disk	Write, Read, Delete
	Near-line Storage (NLS)	Separate disk-based storage	Older data stored in DSOs/Infocubes. The concept is partition the data in time-slices. Excellent Data compression	✓	x	x	Ready only data stored which is infrequently accessed	Traditional usage includes Write/Read. Planned support in BW 7.40 SP11 for exceptional inserts, updates and deletes for NLS will be supported for SP9/SP10. Please see Notes section
	Data Archiving (ADK)	Separate disk/file-based storage	Data which is traditionally retained to fulfill statutory/legal requirements. Excellent data compression. This functionality is complemented with SAP ILM which is used to manage the lifecycle of data from inception in the DB to final destruction from the hardware	✓	✓	x	Read only/infrequently accessed	Write, Read Destroy function available using SAP Information Lifecycle Management (ILM)

FIGURE 7. Multi-temperature storage options with SAP HANA (courtesy of SAP)

Data Temperature	Product Feature/ Integration option	Description	SAP BW on HANA	SAP Business Suite on HANA	SAP HANA Native	Removal	Storage
	Dynamic Tiering	• Warm data management capability, ideal for Big Data Use cases when size/cost constraints prohibit an all in-memory solution • Integral part of the HANA platform to optimize RAM utilization • High performance and efficient compression • Transparent for all operations. No changes required for BW operations • Utilizes disk backed, smart column store • Excels at queries on structured data from terabyte to petabyte scale	(✓1)	•	(✓2)	✓	✓
	Data Aging	• The displacement of data based upon application-level instructions/algorithms • The idea is to have a temperature column within the table which determines if the data is hot (operationally relevant, needed for day to day business transactions, visible and updatable by default, or cold (historic – no longer required for operational use • There is limited Data Aging Functionality delivered with SFIN on HANA and SoH	x	✓3	○	✓	✓
	BW NLS on SAP IQ	• Optimize persistence in system landscape by relocating infrequently accessed data to Sybase IQ • Less frequently accessed data is archived in time partitions • The data in NLS partitions is static and used primarily used for read access • Data in near-line storage resides in a highly compressed state in cost-efficient storage with fewer backups to reduce operational costs • Lower SLA requirements	✓	x	x	✓	✓
	Data Archiving (ADK)	• Standard delivered functionality to archive, delete and retrieve data within the Business Suite e.g. ERP, BW, CRM etc • No license required • When one leverages this type of archiving, one makes a copy of the data, compresses it by up to 80% and stores the data in an SAP-proprietary format. Archiving includes the deletion of the stored data from the database, so that the copy turns conceptually into a move • A number of standard SAP display transactions can be used to read the archived data leveraging Archive indices/ABAP API	✓	✓	x	✓	x
	SAP Information Lifecycle Management (ILM)	SAP Information Lifecycle Management (ILM) manages the entire lifecycle of data from inception in the DB to final destruction from the hardware. As ERP is generally perceived as the legal system of record, ILM is used primarily for ERP environments. Using SAP ILM, we can: • Segregate data as per business/legal requirements • Find data during litigation (e-Discovery) • Isolate and preserve data during audit/litigation • Destroy data with a full audit trail • Compliantly decommission SAP & non-SAP systems	x	✓	x	✓	Optional Storage available with SAP ILM Store
	OpenText	The OpenText product suite offers complementary functionality to the SAP platform in the context of content management • SAP Document Access by OpenText provides a 360° view of SAP-related content • SAP Extended ECM by OpenText is an holistic and enterprise grade product for managing SAP & non-SAP related content	x	✓	x	x	✓
	Hadoop	Hadoop is traditionally positioned for batch processing of raw, unstructured data i.e. audio, video etc. With SAP HANA Vora one can deliver in-memory computing capabilities directly on the data stored in Hadoop, which has previously posed significant challenges for companies. SAP HANA Vora can bridge the gap between SAP data and data stored in a Data Lake.	Consumption via SDA		x	x	✓

FIGURE 8. Data removal & data storage overview (courtesy of SAP)

Deploying Intelligent Data Tiering

In order to define and deliver an intelligent data tiering solution, categorize the data prior to ingestion as follows:

1. Structured data required for real-time analytics
   - **Hot Data:** stored in HANA (memory)
2. Older structured data (3-5 years old)
   - **Warm Data:** stored on FlashArray under HANA control

     (Dynamic Tiering in the case of BWoH, Native HANA; Data Aging in the case of S4HANA)
3. Older data required for legal/regulatory requirements
   - **Cold/Frozen Data:** IQ/Hadoop/OpenText etc.
   - **Unstructured/Semi-Structured/Streaming Data, Raw Data:** Hadoop.

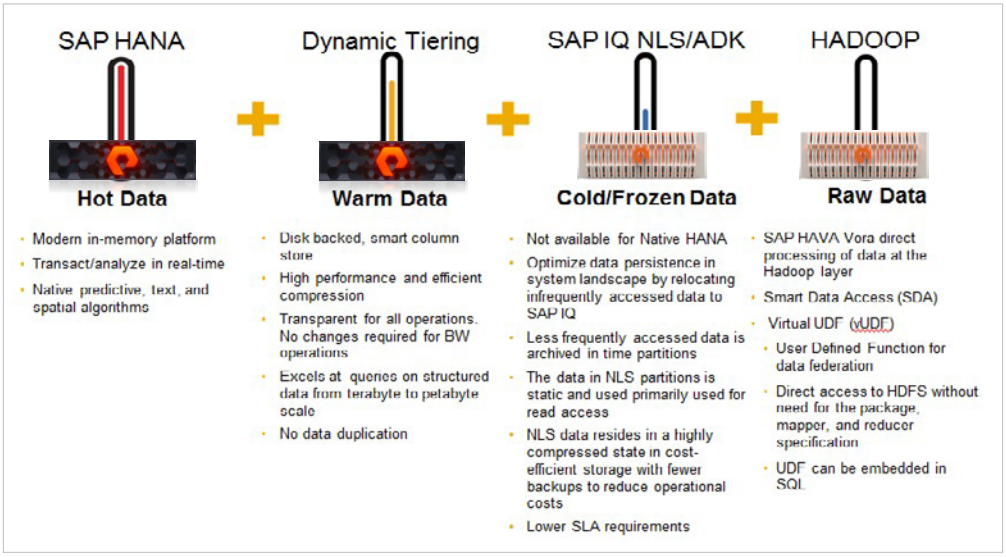

FIGURE 9. SAP HANA combines strengths of different processing domains

Delivering Data Governance to Data Stored Within Hadoop

As mentioned earlier in this document, enterprises are concerned about data governance, compliant retention, and destruction of data owhen storing data in Hadoop. While there are ongoing projects in the open-source community, such as Project Atlas, to deliver these capabilities to the Hadoop ecosystem, SAP has been delivering compliant data retention and destruction for many years. SAP Information Lifecycle Management (ILM) with integration to Hadoop on FlashBlade can offer the following capabilities:

- Segregation of data at the business object level based on country/company-specific requirements
- E-Discovery capability in the context of litigation
- Legal hold functionality to isolate and preserve data during litigation (while destroying all other data that has fulfilled its legal requirements)
- Compliant destruction of data in accordance with legal requirements
- Provisioning of an audit trail to support defensible deletion of data
- Compliant decommissioning of legacy systems

DELIVERING AI TO SAP ENTERPRISES

For several years, SAP has prioritized building AI capabilities into traditional SAP business applications and creating new solutions that deliver AI features and benefits. Most recently, SAP introduced AI capabilities in SAP ERP to automate business processes. SAP understands that data is machine learning's fuel and is capitalizing on the fact that 70% of the world's business transactions run through SAP systems currently. Companies running SAP are increasingly looking to infrastructure that is able to integrate, connect, and optimize data associated with AI . It's no wonder that Gartner classifies smart machines a top five investment priority for more than 30% of CIOs.[3] Pure Storage is keenly aware of the market interest in AI and is working directly with strategic alliance partners to create new end-to-end AI and Big Data infrastructure solutions.

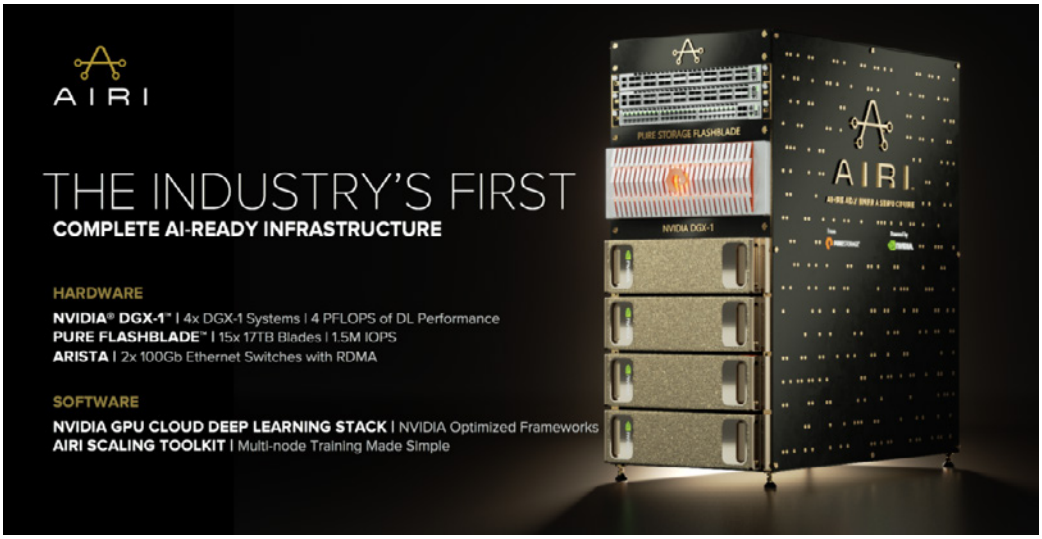

FIGURE 10. Pure Storage AIRI, in partnership with NVIDIA

Built on Pure Storage FlashBlade™ and NVIDIA® DGX-1™, **AIRI**™ makes AI-at-scale simple and accessible for all enterprises. Engineered as a fully integrated software and hardware solution by Pure and NVIDIA, AIRI offers a converged infrastructure approach that addresses the challenges associated with rapidly scaling-up AI computing capacity in the data center. AIRI is powered by FlashBlade, the industry's first storage platform architected for modern analytics and AI, and four NVIDIA DGX-1 supercomputers, delivering four petaFLOPS of deep learning performance. (Note that there is also a version of AIRI available that employs Cisco switches.)

AIRI software is built on the NVIDIA GPU Cloud Deep Learning Stack and the AIRI Scaling Toolkit, enabling seamless scale and maximized performance for multi-node training. AIRI's simplified deployment model lets data scientists jumpstart their AI initiatives in hours, not weeks or months. AIRI was inspired by our customers demanding infrastructure that's built for the AI-powered industrial revolution.

[3] https://www.gartner.com/newsroom/id/3763265

SAP DATA HUB ON FLASHBLADE – BENEFITS

1. Create powerful data pipelines

Incorporate powerful third-party libraries such as TensorFlow **(AIRI!!!)**

2. Build agile, data-driven applications

Establish data lakes, data streams, and create powerful data pipelines – all on FlashBlade

3. Achieve centralized visibility & governance

Gain a system-wide view of data sources, endpoints, and quality

Establish and meet service levels for the diverse landscape

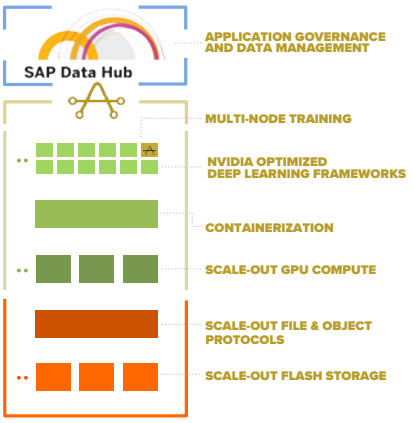

FIGURE 11. Benefits of SAP and FlashBlade

SAP DATA HUB ON FLASHBLADE AND AIRI

SAP Vora is an enterprise-ready, easy-to-use, in-memory distributed computing solution to help organizations uncover actionable insights from Big data.[4] Vora was purposely designed to deliver enterprise-grade features to Big Data environments, allowing application users to unlock the potential of their data and remove their dependency on data scientists. Vora seamlessly integrates with HANA and provides one logical view across all data. SAP's software development philosophy embraces container-based software deployment and management based on microservices architectures. SAP Vora and SAP Data Hub are examples of these offerings, and Pure Storage supports and empowers these new applications by delivering the best performing data management solutions for high performance and scalability. Pure's FlashBlade array is well suited for these new architectures. When running SAP Data Hub on FlashBlade, customers can experience new performance enhancements and have the SAP Data Hub as close as possible to their data lake for additional efficiency.

SAP customers following this strategy can take advantage of FlashBlade in multiple scenarios:

- **Data Tier SAP Data**
- **Running Data Lakes and Data Warehouses on FlashBlade**

  Typical data lake and data warehouse infrastructures tend to be complex to manage and maintain. Each application has a separate data warehouse, each copying data back and forth from the data lake. For lines of business relying on this data, this complexity is challenging to utilize, and for IT it can be a nightmare to manage. Pure's FlashBlade array is the industry's first data platform engineered for a wide range of modern workloads, from instant restore to AI, machine learning, and software test and developent. It is built for **any type** of unstructured data. By definition, unstructured data means unpredictability – data can

4 https://news.sap.com/sap-delivers-live-insights-from-big-data-to-customers/

take any form, size, or shape, and can be accessed in any pattern. FlashBlade is capable of accelerating any data, small or large, random or sequential.

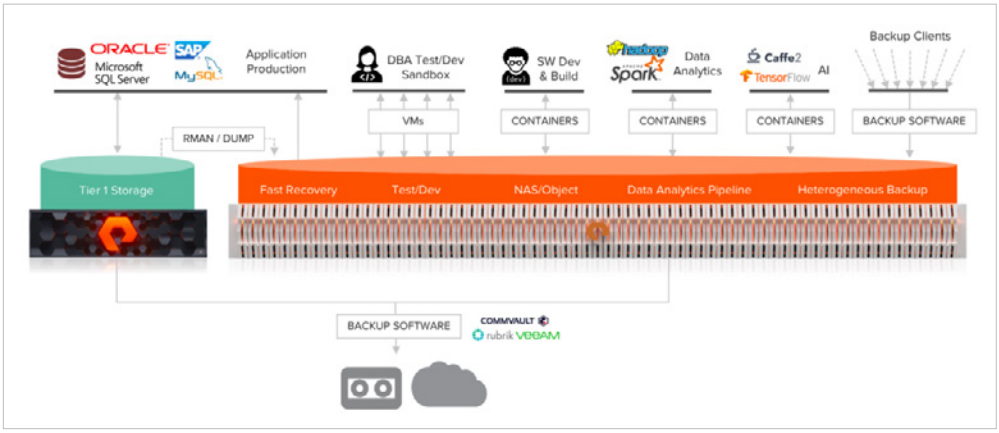

FIGURE 12. Modern infrastructure, consolidating silos into two storage systems

- **Use FlashBlade as a Data Source for AI and Data Streaming**

  A deep neural network is a massively parallel model, with millions to billions of neurons loosely connected together to solve a single problem. A GPU is a massively parallel processor built with thousands of compute cores that offer the optimal architecture for performing the computations that are core to deep learning, delivering the power of hundreds of traditional CPU servers in a single node. FlashBlade, powered by Purity software, is also a massively parallel platform capable of delivering high performance access to billions of objects and files for 10s of thousands of clients in parallel.

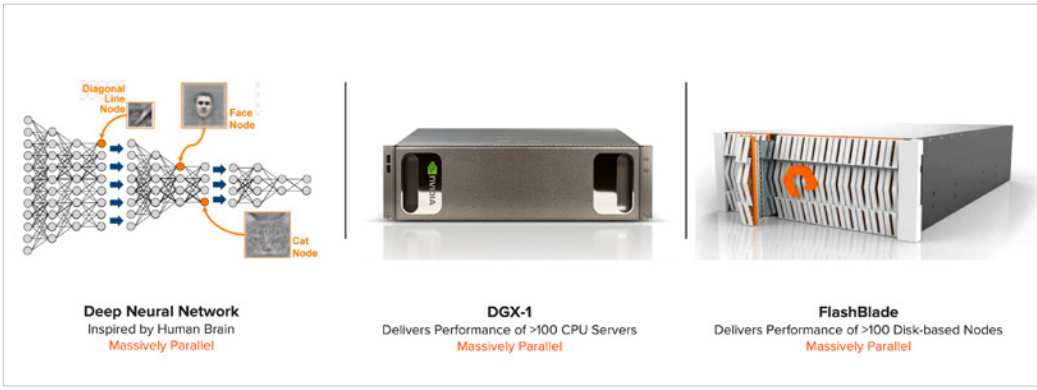

FIGURE 13. Any AI platform must be massively parallel at the core of its design

Modernizing your data using this solution unifies data and processing, despite being strewn across multiple platforms. Users can choose the best approach for a data workload or analytics project, and offload certain datat sets from the data warehouse to the data lake and vice versa. FlashBlade is uniquely powered to help solve the challenges of data warehousing, management, and analysis.

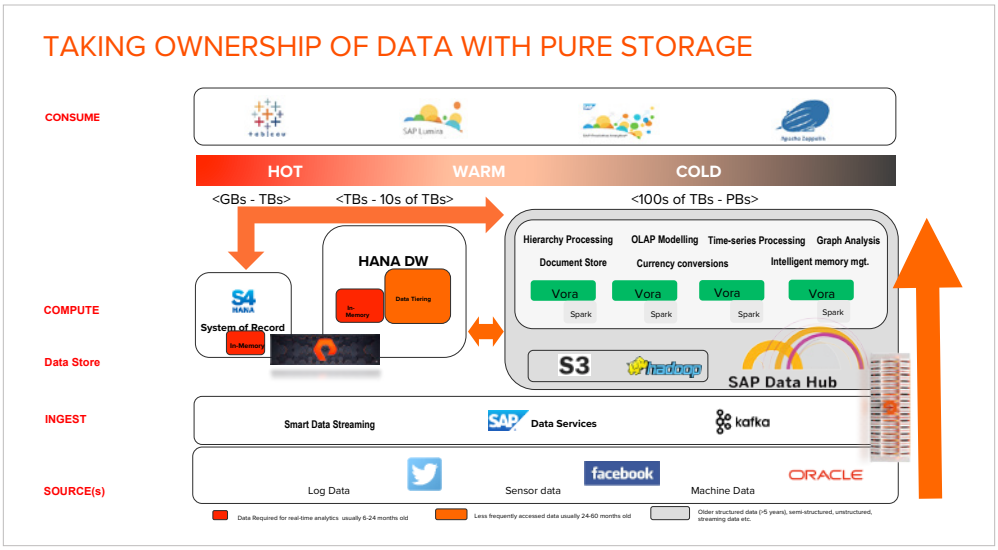

FIGURE 14. Pure Storage Data-Centric Architecture for SAP, Big Data, and AI

CONCLUSION

The evolution of web-scale application architectures has driven a significant change in requirements for enterprise customers. As figure 15 illustrates, Pure Storage provides SAP customers with solutions that help them run their applications efficiently and capitalize on all sorts of new data types and sources. This is achieved by running primary SAP applications, such as SAP S/4HANA and BW/4HANA, in a private, secure, and easy-to-scale cloud. Pure's hybrid cloud solutions, built on top of Pure CloudSnap™ (not available at time of publication; expected availability 2H 2018), makes taking advantage of the public cloud easy when necessary. Other types of data from different sources can run with extreme performance on FlashBlade and AIRI, giving enterprises the edge they need to exploit their data and take action with unprecedented speed. Pure infrastructure can be utilized with all these different data processes while flexibly and continuously supporting data tiers for the appropriate data type and infrastructure.

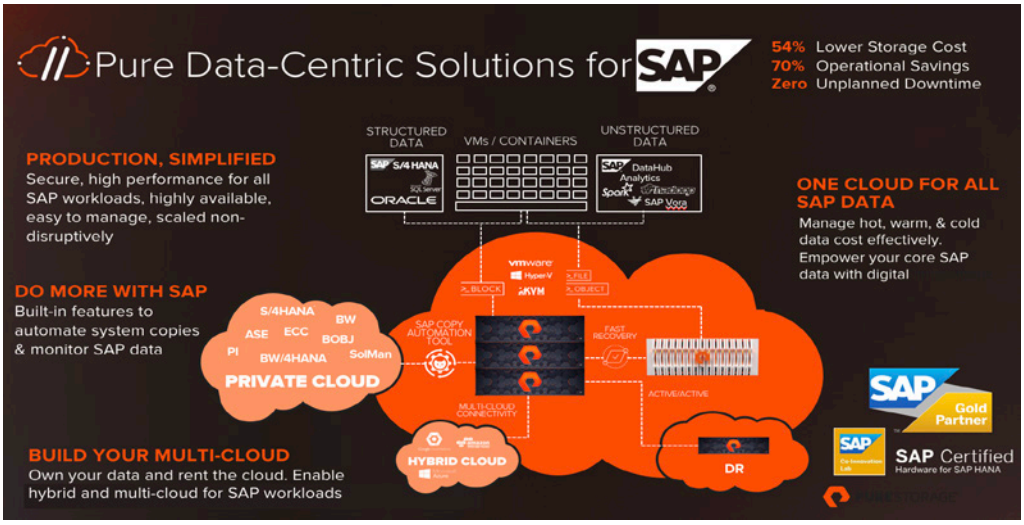

FIGURE 15. The Pure Storage SAP solution

ABOUT THE AUTHOR

Samer Kamal is the head of SAP solutions at Pure. He has over 15 years of consulting experience in SAP basis and security, SAP architecture in cloud and on-premises, and SAP integrations. In addition to his intimate knowledge with SAP's portfolio, Samer has also built his career by working for several Fortune 500 companies. His experience is coupled with empathy and solid understanding of running an enterprise IT shop.

ps_wp21p_ai-and-big-data-with-sap-and-pure_ltr_01