Connecting the Healthcare Dots to Enable Precision Medicine

Enabling SAP Health services with Dell EMC Isilon Data Lake in a Hadoop environment

ABSTRACT

Dell EMC offers an end-to-end solution for the personalized medicine data analysis work stream that leverages the capabilities of the SAP® Health services, utilizing SAP HANA®: powered by Dell EMC Ready Solutions, Dell EMC Isilon® network attached storage (NAS), and the Apache™ Hadoop® platform. This paper provides a high-level overview of this solution, including its key components, architecture and top use cases.

August 2017

TABLE OF CONTENTS

Published in the USA 8/17.

AN UNPRECEDENTED OPPORTUNITY IN HEALTHCARE

Schleidgen et al define precision medicine as a field that seeks “to improve stratification and timing of health care by utilizing biological information and biomarkers on the level of molecular disease pathways, genetics, proteomics as well as metabolomics.”[1] A key enabler for precision medicine, genomic sequencing is becoming an adjunct to standard of care, and will soon be pervasive in many areas of healthcare. For example, we are entering an era in which all cancer patients can have their genomes sequenced to enable precision treatments tailored to the individual and the type and stage of disease.

With the advent of genome sequencing, the cornerstone test method for DNA, RNA and the microbiome, researchers and physicians now have a critical resource to help them understand the underlying molecular mechanics and pathways of disease. Armed with fresh analytic insights, researchers can accelerate their development of personalized, patient-centered treatments and innovations, including drugs, therapies, devices and diagnostic tests. At the same time, data-driven genotypic insights integrated with phenotypic or medical records can also help physicians deepen their understanding of their patients and get patients more involved in their own healthcare, to improve prevention, diagnosis, treatment and care.

This is the upside of the story. From an IT transformation perspective, there are significant challenges that arise with the era of the genome. For starters, genomic data alignment and assembly require an enormous amount of computational power and temporary data storage capacity, and archival storage is orders of magnitude higher. A single genome can total hundreds of gigabytes, and one day physicians and researchers might want to compare the genomes and health records of millions of patients to identify causes and treatments for diseases. It is a huge challenge to get insights out of such an overwhelming amount of data that is growing exponentially — notwithstanding the new data rush from stream-based medical devices and the Internet of Things (IoT).

This is where the SAP Health platform comes into play. Using the SAP Health platform, organizations can combine and analyze traditional patient data sources with clinical, research, molecular, image, social information and more. Drawing on the power of SAP HANA® and its standard-setting in-memory database, the SAP Health platform is extremely effective at generating rapid insights from huge datasets. This platform for personalized-medicine applications enables processing and real-time analysis of big medical data from various sources, all in a single system driven by SAP HANA.

AN END-TO-END SOLUTION FOR CLINICAL AND RESEARCH DATA ANALYSIS

SOLUTION OVERVIEW

Dell EMC helps organizations solve the genomic data puzzle in their journey toward precision medicine with an end-to-end solution for the genomics data analysis work stream that leverages SAP Health. This solution is based on a Dell EMC reference architecture that brings together the combined capabilities of SAP Health, a Dell EMC Isilon data lake, the Cloudera distribution of Apache™ Hadoop® and SAP® Vora™ into SAP HANA by Dell EMC Ready Solutions.

---

[1] BMC Medical Ethics, December 2013, 14:55.

The Dell EMC solution overcomes some of the key challenges in the genomics data analysis work stream, including low storage efficiency, multiple copies of data, controller-centric upgrades and balancing, and low availability. The solution provides:

- Scalable, performant storage for unstructured data
- A single storage container for integration and interoperability — via a Lambda+ clinical architecture designed to handle massive quantities of data
- Consistency of data as it scales — data availability across metadata and container types
- A single platform for stream, file and object data
- Integration of SAP HANA data with private, hybrid and public clouds

In combination with SAP Leonardo, SAP's business innovation platform, and the SAP-certified Dell Edge Gateways, this enables a personalized medicine workflow from the IoT edge with patient monitoring to the core with SAP HANA.

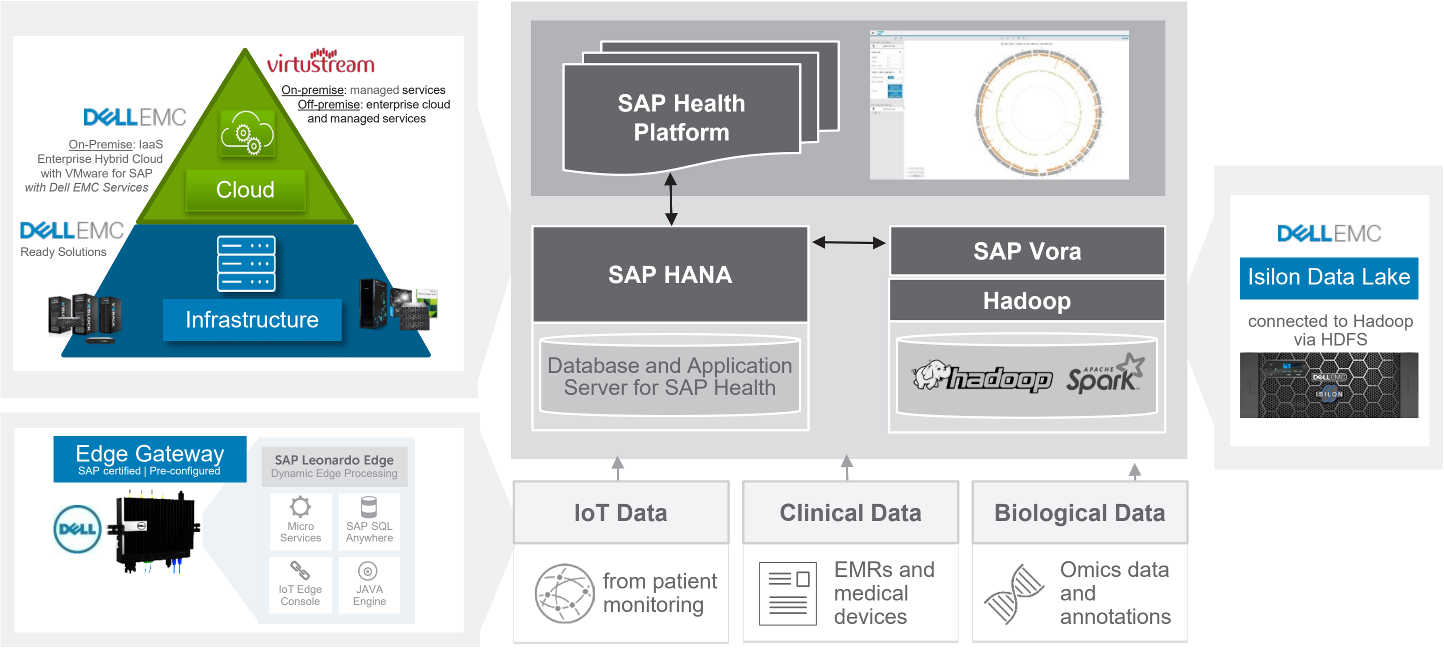


Figure 1. Comprehensive Solution Overview

KEY SOLUTION COMPONENTS

SAP HEALTH PLATFORM

With SAP HANA at its heart, SAP Health serves as a platform for precision-medicine applications. SAP Health enables the processing and real-time analysis of big medical data from various sources in a single system.

SAP HANA

SAP HANA is a database designed to handle big data by processing transactions and analytics in-memory on a single data copy. It combines powerful data transformation with analytical functionality to deliver real-time insights from live data or analytical insights beyond the obvious.

CLOUDERA DISTRIBUTION FOR HADOOP

The Cloudera Distribution for Hadoop (CDH) serves as an enterprise-grade compute cluster that shares memory and disks across commodity servers for pre-calculating data via HDFS, the Hadoop distributed filesystem using MapReduce, a scalable, distributed compute paradigm. As Hadoop clients run MapReduce jobs, the clients access the data stored on the Isilon cluster through an HDFS interface. With this unique architecture, the Isilon OneFS operating system becomes the HDFS file system for MapReduce clients.

SAP VORA

SAP Vora, an in-memory, distributed computing solution, serves as a gateway to SAP HANA, via a collection of databases designed to communicate with SAP HANA. In the Dell EMC solution, SAP Vora is deployed as a plugin for Hadoop. It can use HDFS and can also use the functionalities of EMC Isilon to store its own data.

SAP Vora enables native, multi-protocol access, including HDFS, to SAP HANA via Dell EMC Isilon. Relational, graph and object data are accessed in-memory for SQL, time series, graph and JavaScript Object Notation (JSON) with specialized algorithmic access to data formats.

DELL EMC INFRASTRUCTURE

DELL EMC ISILON FOR HADOOP AND UNSTRUCTURED STORAGE IN A DATA LAKE

The solution leverages Dell EMC Isilon network attached storage (NAS) in an Apache Hadoop cluster to support SAP HANA. The Isilon Infinity F800 all-flash storage array provides the foundation for a “data lake” for tiered distributed storage for unstructured files with the following summary specifications per 4U chassis:

- 16-core Intel® 2697A v4 CPU with up to 256GB RAM
- Backend network QDR IB or 40GbE; Frontend network 10GbE or 40Gbe
- Three capacity options: 96TB, 192TB or 924TB in a 4RU chassis (35-inch standard depth)

Each 4RU chassis with four nodes provides up to 250,000 IOPs and 12 GB/s (for streaming reads) with a density of up to 924TB. A single cluster can scale up to 100 chassis or 400 nodes with an unprecedented total capacity of 90PB. This performant system that powers the data lake is shown in Figure 2.

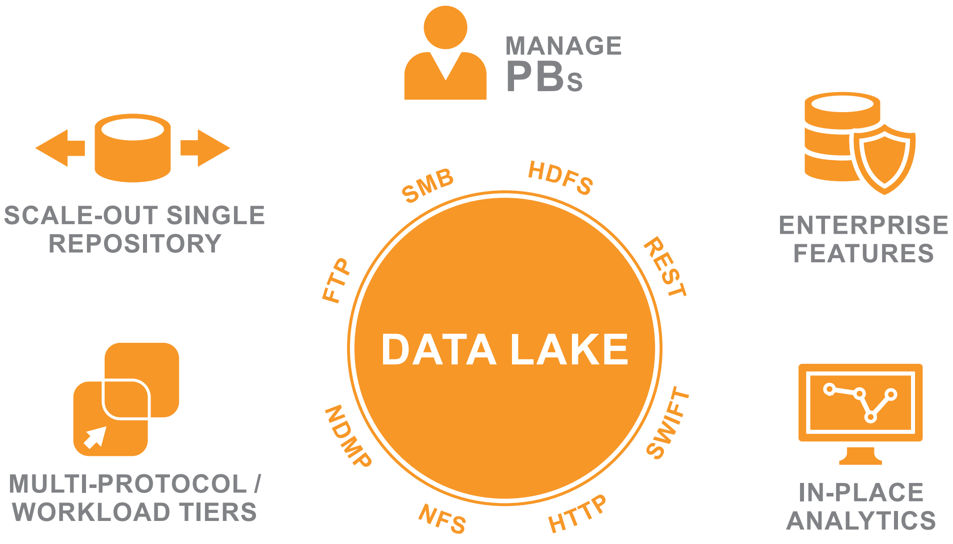


Figure 2. A Dell EMC Isilon data lake for tiered distributed storage.

Powered by the OneFS operating system, the Dell EMC Isilon all-flash array uses a revolutionary modular architecture to provide a powerful yet simple scale-out storage platform to speed access to massive amounts of unstructured data, while dramatically reducing cost and complexity. The Dell EMC Isilon OneFS distributed file system creates a cluster with a single file system and single global namespace.

The Dell EMC Isilon scale-out NAS platform provides Hadoop clients with direct access to big data through an HDFS interface. Powered by the distributed Dell EMC Isilon OneFS® operating system, a Dell EMC Isilon cluster delivers a scalable pool of storage with a global namespace.

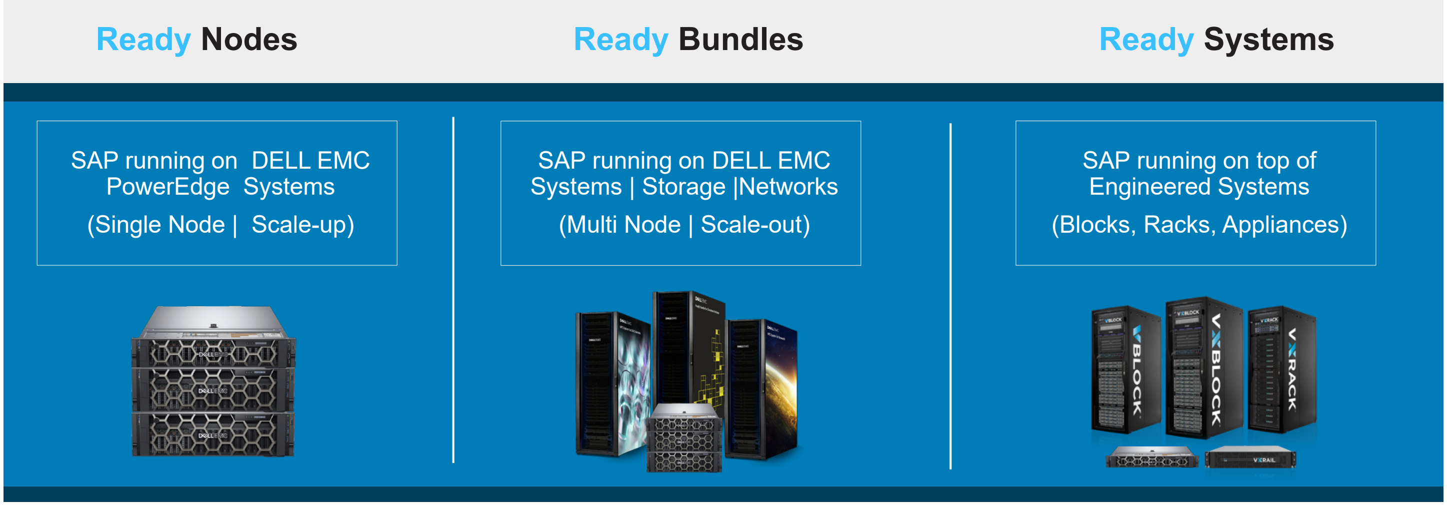
Isilon storage offers multiple advantages for organizations leveraging SAP Health. These include:

- The high efficiency of scalable NAS
- High availability of data
- Self-balancing, self-healing and self-encrypting capabilities
- Tiering from scratch or temporary storage near compute to the archive based on throughput and performance needs
- Density at scale — from terabytes to petabytes of data in a single cluster
- A globally coherent cache and an architecture that puts one cache closer to compute (L3 cache)
- Operational advantages, in the form of storage that is simple to install, add and manage

DELL EMC READY SOLUTIONS FOR SAP

The SAP HANA platform in this solution is based on the Dell EMC PowerEdge™ R940 server with Intel® Xeon® processors, the most powerful enterprise server platform in the Dell EMC portfolio. This platform is built for speed and scalability while offering value-added features that enhance management and reliability. With 48 DDR4 DIMM slots and 24 hard drives (including up to 12 Express Flash NVMe PCIe SSDs), the system scales to handle the most demanding workloads.

To streamline and accelerate deployment, the Dell EMC infrastructure is available in Ready Solutions that are optimized for SAP HANA.



Ready Nodes

SAP running on DELL EMC PowerEdge Systems

(Single Node | Scale-up)

Ready Bundles

SAP running on DELL EMC Systems | Storage |Networks

(Multi Node | Scale-out)

Ready Systems

SAP running on top of Engineered Systems

(Blocks, Racks, Appliances)

Figure 3. Ready Solutions

- **Ready Nodes** — Dell EMC offers Ready Nodes dedicated to SAP HANA. Built on Dell EMC PowerEdge™ servers with Intel Xeon processors, these Ready Nodes are pre-sized, pre-built and delivered with SAP HANA software pre-loaded.
- **Ready Bundles** — Dell EMC Ready Bundles bring together SAP-certified Dell EMC servers, storage and networking, including support for SAP HANA Tailored Data Center Integration (TDI).
- **Ready Systems** — These pre-built systems deliver the convenience of an appliance and the flexibility of TDI, including options that incorporate Dell EMC V[x]Block, VxRail and VxRack.

DELL EMC HYBRID AND CLOUD OPTIONS

In addition to on-premises deployments, Dell EMC offers a choice of hybrid cloud and off-premises platforms for SAP HANA environments.

- **Hybrid cloud** — The Dell EMC Enterprise Hybrid Cloud solution combines hardware, software and services from Dell EMC and VMware into a platform built on converged systems to deliver the foundation for infrastructure-as-a-service. Dell EMC and VMware have spent thousands of hours of engineering time designing, testing and proving the platform in a lab setting, so organizations deploying cloud solutions don’t have to expend that effort.
- **Off-premises managed cloud for SAP** — Virtustream, a Dell Technologies business, offers an unparalleled platform for running SAP in the cloud. The platform was purpose built for handling complex, mission-critical enterprise applications. Virtustream experts have extensive experience working directly with SAP applications. They have the expertise to seamlessly migrate SAP deployments to the cloud.

DELL EDGE GATEWAY

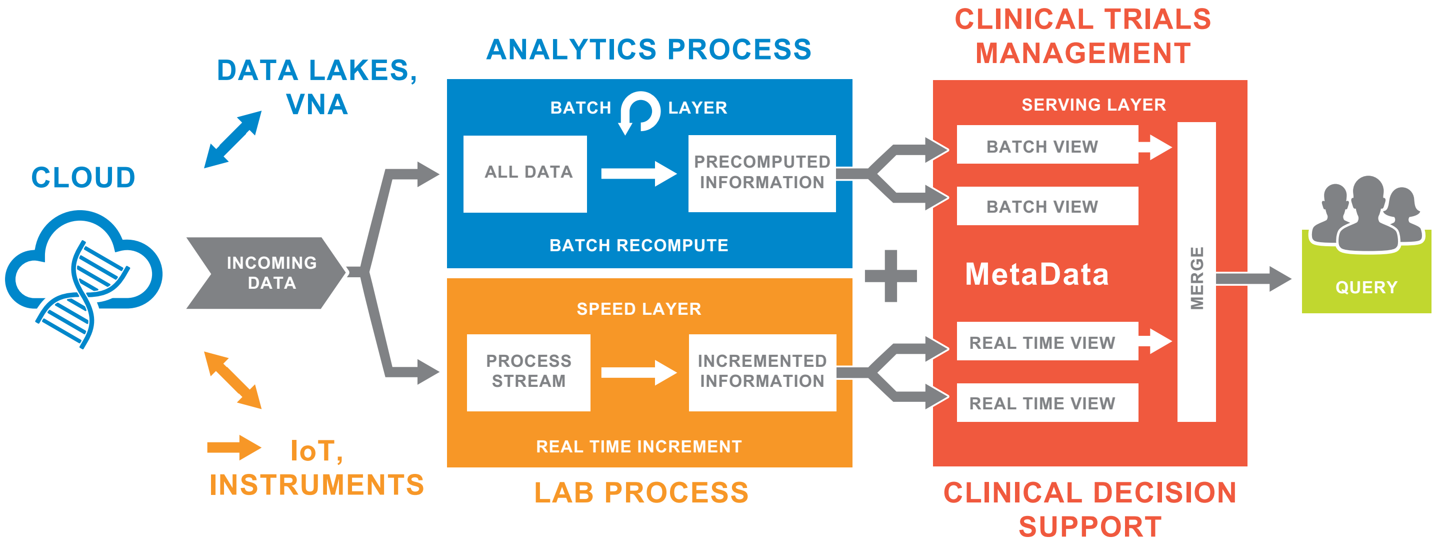
The SAP-certified Dell Edge Gateway, with powerful dual-core Intel® Atom™ processors, connects varied wired and wireless devices and systems, aggregates and analyzes the input, and sends it on to the SAP Health solution running on the SAP HANA platform or to the Hadoop Cluster, depending on the customer requirements. Because the gateway sits close to your devices and sensors, it sends only meaningful data to the cloud or control center, saving you expensive bandwidth. Dell Edge Gateways are tested and validated, pre-configured and pre-installed with SAP Leonardo Edge software which streamline and accelerate deployments while reducing costs and risk to deliver the right business outcomes.

ARCHITECTURE GOALS

Uploading complete datasets of genomics research data in to SAP HANA often is not possible due to the time required for data transfer and the costs of SAP HANA infrastructure. Therefore, using Hadoop to preselect data before uploading it to SAP HANA is helpful. Using SAP Vora as a Hadoop plugin provides connection between these two environments, as shown in Figure 1. It allows users to preselect data in the Hadoop environment and upload only the needed data into SAP HANA. This can also be used to separate personal identifiers for data privacy and data protection reasons.

BATCH AND STREAM PROCESSING WITH SEAMLESS CONNECTIVITY TO IN-MEMORY ANALYTICS

The Lambda+ clinical architecture at the heart of the Dell EMC solution takes advantage of both batch-processing and stream-processing technologies to improve throughput for huge amounts of data while reducing latency. This architecture is built around three layers: a speed layer that processes streaming data, a batch layer that processes slower-moving batch data, and a serving layer that merges the batch view and the real-time view to enable comprehensive queries critical to the integrative nature of precision medicine, as shown in Figure 4.



Derived from: Merz N & Warren J, "Big Data", 2013, Manning

Figure 4. An enhanced three-layer architecture with speed, batch, serving layers and automated tiering with metadata movement.

The Lambda+ architecture overcomes key challenges with the original Lambda architecture. The original architecture integrated stream architectures with a distributed file system (originally Hadoop) and an application serving layer. This approach was taken to keep data raw as it passed from the device layer via the stream tier to the application layer and finally to the archive tier. However, this approach brought multiple challenges, including code maintenance, "data refraction," and integration and interoperability issues.

The real improvement with the current Lambda architecture comes when the file-level, object-level, stream-level and application-level metadata are available at each of the three Lambda layers — stream, batch and serving — and are tiered automatically. The "+" in the Lambda+ derivative is this single data "fabric" for these three layers, which mitigates the metadata issues. The data fabric has four central themes: data persistence, data services, data distribution and data security. All these themes are essential for healthcare.

USE CASES

Some of the top use cases for the Dell EMC solution for SAP Health include:

EMBEDDED ELECTRONIC MEDICAL RECORDS (EMR)

The SAP HANA solution enables embedded EMR via in-memory SQL database models for medical records. The EMR data becomes columnar and secondary tables that are hash-partitioned for faster parallel access of cross-tabular data during analytics. A "data-vault" modeling approach helps ensure logging, auditing and monitoring, thereby making the data model more extensible to multiple dimensions in the healthcare variable space. Stable variables (like business information) are separated from contextual or rapidly changing variables, making the system performant for in-memory analytics. The contextual separation also maintains the relationship between data in memory. This design philosophy is critical to dynamic data and makes population-scale health analytics a reality.

GRAPH ANALYTICS, MACHINE AND DEEP LEARNING

The Dell EMC solution for SAP Health is ideal for enabling in-memory graph analytics for genomics, microbiomics, transcriptomics and computational graph theory algorithms. An in-memory infrastructure provides the best architecture to down-sample large, multidimensional datasets and perform multivariate analytics. Providing a distributed "data movement" infrastructure from the various tiers of data to the SAP HANA in-memory analytics engine and to object or cloud systems provides a complete solution. The machine learning (ML) and deep learning (DL) libraries within HANA enable the analysis of images from various modalities (described below) to be integrated within a deep learning environment.

MODALITY INTEGRATION

The insights from the data for precision medicine currently reside separately in various modalities, including chemistry lab tests, Pathology, Radiology and Genomics. As Patient Reported Outcomes (PRO) gains more traction and more "N-of-1" studies are approved by regulatory agencies, newer modalities like medical devices and IoT will add to the complexity of integration and security with current infrastructures for Electronic Medical Records.

The Dell EMC solution for SAP Health helps organizations address these challenges. The solution can be used to integrate various modalities (radiology, pathology, genomics, microbiomics and IoT) in a "data lake" and move the integrated data into "in-memory analytics and machine/deep learning" to take analytics to population scales.

PATIENT MONITORING USING IOT TECHNOLOGIES

Through the Dell EMC IoT infrastructure using Edge Gateways, patients can be monitored via personal and medical devices that capture and aggregate data. These devices analyze the input and send only the meaningful data to the SAP Health platform. This enables a 360-degree view of all relevant patient information in a single-pane-of-glass view.

THE DELL EMC ADVANTAGE

Dell EMC is ideally positioned to deliver this solution. Dell EMC delivers the cost advantages of storing unstructured data via the Isilon and Elastic Cloud Storage (ECS) platforms, along with best practices cultivated in SAP HAHA deployments around the world.

Storing data in an Isilon scale-out NAS cluster instead of HDFS clients streamlines the entire analytics workflow. Isilon's HDFS interface eliminates the need to extract the data from a storage system and load it into an HDFS file system.

Isilon's multiprotocol data access for HDFS, along with SMB and NFS, eliminates the need to export the data after it has been analyzed. The result is that you can not only increase the ease and flexibility with which you analyze data, but also reduce capital expenditures and operating expenses.

By scaling multi-dimensionally to handle the exponential growth of big data, a Dell EMC Isilon cluster connects natively with Hadoop to provide the best of both worlds: data analytics and enterprise scale-out storage. This combination helps you adapt to fluid storage requirements, non-disruptively add capacity and performance in cost-effective increments, reduce storage overhead and exploit your data through in-place analytics.

GETTING STARTED

To help your organization get started with your analytics solution, Dell EMC provides Ready Solutions for SAP with a wide range of resources to accelerate deployment and time to value.

Using the Dell EMC Global SAP Center of Excellence, your IT leaders can consult with experts, explore demonstrations of the SAP Health platform, and access test systems to size their analytics projects. These activities can help your organization reduce project risks and enable analytic success.

Located near SAP's worldwide headquarters in Walldorf, Germany, the Dell EMC Global SAP Center of Excellence can be accessed through scheduled engagements virtually or on site. The Dell EMC Isilon eLab can also be accessed to conceive and test data lake infrastructure for specific applications and use-cases.

KEY TAKEAWAYS

- The Dell EMC solution for SAP Health provides an end-to-end solution for the personalized medicine data analysis work stream. This solution overcomes key IT challenges in the work stream, including the need to increase storage efficiency and availability.
- The Dell EMC solution is based on a proven reference architecture that brings together the complementary capabilities of SAP Health, a Dell EMC Isilon data lake, the Cloudera distribution of Apache Hadoop via SAP Vora into SAP HANA powered by Dell EMC Ready Solutions.
- The Dell EMC Global SAP Center of Excellence provides a wide range of resources to help your organization explore the capabilities of the SAP Health Platform and develop a closer understanding of your specific needs.

To learn more, visit: DellEMC.com/sap or contact your Dell EMC representative for a one-on-one conversation about your needs and goals.