DIVING IN: NAVIGATING A DATA LAKE FOR PREDICTIVE CARE
Patient Data Intelligence for Next-Generation Care Delivery

ABSTRACT
The transition from fee-for-service to value-based care requires healthcare providers to proactively manage patient health risks across the care continuum. To achieve this transformation, providers need to make analytics a primary component of their IT strategy. A data lake pools data from multiple sources and applies analytical models to provide a new approach to information management, reporting, and predictive analytics to help create advanced analytic insights, deploy evidence-based care strategies, and improve patient engagement outcomes. This white paper explains how healthcare organizations can build and develop their data analytics infrastructure, data science skills, and data governance processes necessary for a high-performing data lake. With unprecedented levels of patient data intelligence at their fingertips enabled by a data lake, caregivers can drive informed, data-driven decisions at the point-of-care.

August, 2015
To learn more about how EMC products, services, and solutions can help solve your business and IT challenges, contact your local representative or authorized reseller, visit www.emc.com, or explore and compare products in the EMC Store.

Copyright © 2015 EMC Corporation. All Rights Reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided "as is." EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

VMware, VMware vCloud Suite, and VMware vCloud Hybrid Service are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. All other trademarks used herein are the property of their respective owners.
TABLE OF CONTENTS

EXECUTIVE SUMMARY ........................................................................................................... 4
TURNING GROWING DATA INTO ACTIONABLE INSIGHTS FOR FUTURECARE .................................................................................................................. 4
DATA LAKE: A PATHWAY TO PREDICTIVE CARE ............................................................... 5
BUILT FOR THE RIGORS OF ADVANCED HEALTHCARE ANALYTICS ....................... 6
CHART A COURSE FOR REAL-WORLD HEALTHCARE USE CASES ......................... 6
JUMPSTARTING THE JOURNEY: THE FEDERATION BUSINESS DATA LAKE ............ 9
HOW TO SET SAIL ON A HEALTHCARE DATA LAKE .............................................. 11
CONCLUSION: EMBRACE THE FUTURE OF PREDICTIVE CARE .................................. 12
EXECUTIVE SUMMARY

As healthcare providers transition from fee-for-service to value-based care models, they are discovering new and forward-looking analytic models to identify high risk patients, deploy evidence-based medicine, and reduce adverse events and infections. New pay-for-performance models will require healthcare organizations to improve quality and reduce costs by proactively classifying at-risk members of the community, identifying gaps in care, and improving patient engagement.

In many cases, providers will need new methodologies to tailor treatment plans that risk-stratify patients, reduce the need for emergency treatments, and prevent hospital readmissions. Addressing these challenges requires near real-time analytics at the point-of-care to guide the way to informed diagnosis and improved outcomes.

At the same time, health systems are continuing to look for ways to be more efficient in their day to day business. Analytics can also provide insight to optimize staffing schedules, better manage IT resources, and identify trends in services that can be used for marketing.

A data lake offers a technology infrastructure that assembles information generated across the health system, including data imported from outside sources and services. The data lake can reveal actionable insights about an organization’s performance indicators and impacts of patient care interventions to better manage risks and deliver affordable, higher quality care. Using an enterprise hybrid cloud framework, a data lake is layered with information-rich data sources, analytic tools, and data science best practices that enable providers to link and correlate information in completely new ways.

This white paper describes IT components that comprise a data lake and provides examples of use cases for patient care delivery, clinical research, population health management, and security analytics. The paper also includes EMC recommendations for how to get started with a data lake, including deploying a technology infrastructure, establishing a data science practice, and enacting proper data governance. Through professional services, Big Data workshops, and education, EMC can help healthcare providers gain the technology, people skills, and processes to implement a data lake that aligns with their strategic goals and enable a business value-driven roadmap.

TURNING GROWING DATA INTO ACTIONABLE INSIGHTS FOR FUTURECARE

As healthcare reform continues, providers will be reimbursed for pay-for-performance based on patient outcomes rather than a fee for services such as doctor visits, hospital stays, and tests. The Centers for Medicare and Medicaid Services (CMS) is moving aggressively to adopt a value-based payment structure, giving providers two to three years to transform their business model.¹

With this increased focus on accountable care, providers are using personalized patient monitoring tools to help prevent recurrences and possible hospital readmissions as they deliver safer, appropriate, and cost-effective care. With hospitals facing potential penalties if their rates of readmission are too high, caregivers are also working to gain additional insight into patients' medical histories, lifestyles, and demographic profiles. Additionally, providers are making investments that expand their meaningful use of the electronic medical record (EMR) and are looking to analyze a broader set of clinical and business data points, including PACS images, genomics, lab results, clinician’s notes, cost analysis, and claims in an effort to deliver predictive, targeted care and precision medicine.

Growing at 48 percent per year through 2020, healthcare data is one of the fastest-growing segments in the digital universe, as reported in IDC’s study, The Digital Universe Driving Data Growth in Healthcare.² Clinical applications, compliance requirements, genomic and proteomic sequencing, and FutureCare-enabling technologies³ for cloud, Big Data, mobile, and social media are among the largest drivers of this data growth.

Yet, today’s traditional IT infrastructures often lack the scalability, performance, and analytic capability to support today’s rapidly expanding Big Data requirements. To move forward with evidence-based practices, providers are deploying new approaches to assemble, manage, and analyze data. This is especially important as they risk stratify patient populations to identify high-risk cohorts and provide targeted interventions to better manage chronic conditions.

³ Source: http://www.meritalk.com/FutureCare
A data lake acts as the technology enabler to capture maximum value from all of the data being created across the continuum of care. Simply put, a data lake provides an IT environment that incorporates structured, semi-structured, and unstructured data from trusted external and internal sources and ultimately improves effectiveness and quality of critical business and clinical practices. In addition, a data lake applies advanced analytics to produce actionable insights that enable timely interventions to prevent adverse health events and ultimately, elevate overall population wellness.

To move this strategy forward, EMC has developed a fully engineered, enterprise-grade data lake paired with professional services to help healthcare organizations launch this transformational journey. Known as the EMC Federation Business Data Lake, this solution simplifies and accelerates data analytics deployments, positioning IT as a strategic partner to ensure the organization can make more timely and informed business and clinical decisions. EMC can also work with your team to help define early-win use cases and plan a long-term data lake roadmap to harness Big Data analytics today and advance new value-based care models.

**DATA LAKE: A PATHWAY TO PREDICTIVE CARE**

A data lake meets rapidly evolving business and clinical requirements by quickly and efficiently analyzing new combinations of data from multiple sources across the health system. Traditionally, healthcare providers have invested substantial time and effort to extract, transform, and load (ETL) data from its original format into data warehouses purpose-built for business intelligence. A data lake strategy simplifies storage, management, and analysis of Big Data by consolidating data in real-time, near real-time or in batch from disparate sources and across multiple protocols. (Figure 1)

![Figure 1. Evolution of Big Data Analytics](image)

A data lake provides a unified location for all relevant data generated by the healthcare system, serving as a repository for structured data drawn from traditional databases and unstructured data, such as patient images, lab reports, pathology, genomics, clinical notes, and social media activity. Instrumentation, sensor, and telemetry data can also be streamed and acted upon in near real-time.

To execute predictive care analytics at scale and in near real-time, providers can create analytic “sandbox” environments and narrow the source and scope of the data set, helping to reduce time to insight. The data lake opens opportunities to find correlations across vast stores of data they previously were not able to query or examine.

Traditional business intelligence environments typically provide retrospective rear-view mirror reporting while data lakes enable the ability to run predictive analytics at scale. Through data science and predictive analytics, healthcare organizations can identify potential outcomes and explore “what if” scenarios—an essential building block of clinical predictive analytics and precision medicine.

With such future-focused insights, healthcare providers can advance accountable care initiatives, creating a new realm of data science for uncovering trends, patterns, relationships, correlations, and discoveries that impact integrated patient care. Data can also be consolidated from outside resources, including payers, genomic research centers, biobanks, and social media feeds.
Data lakes also open up possibilities for integrating information from wearables, fitness devices, appliances built on the Internet of Things (IoT) such as heart monitoring implants, for personalized, real-time care delivery. Caregivers can employ advanced analytics to use data generated by these devices to help reduce in-hospital complications and unnecessary readmissions, deliver personalized medicine, identify genetic markers, improve clinical trial safety, and much more.

**BUILT FOR THE RIGORS OF ADVANCED HEALTHCARE ANALYTICS**

At the foundational level, a well-designed and effective data lake uses intelligent storage and compute resources to analyze data in real-time with modern data analytics tools, such as Hadoop, HAWQ (Hadoop with Query), SPARK, in-Memory ingest/processing, Cassandra, No-SQL, and mongoDB. Highly scalable storage solutions, for example, allow the data lake to expand to petabytes of data without impacting performance of applications. Data services, such as tagging, metadata management, governance, encryption, and data protection, provide additional value to the data lake environment by ensuring the provenance and veracity of the data in the lake.

Data lakes provide a data analysis environment for ad-hoc data sciences teams. The data lake provides a centralized data hub that can feed the traditional legacy enterprise data warehouse, (which is focused on production reporting and SLAs) while supporting ad hoc query and discovery on the raw, atomic data. Traditionally, there has been a natural point of friction between data science and data warehouse teams since data science tools, such as SAS, can negatively affect data warehouse performance. With a data lake, the data science team can freely access and analyze the data without affecting data warehouse SLAs.

Virtualized, open-source analytics platforms comprise another component of a data lake infrastructure. For instance, Hadoop provides an open-source platform for distributed processing of very large data sets. It serves as a clearinghouse for all types of Big Data that data scientists can access, organize, and analyze using advanced analytics databases, applications, and tools. These critical building blocks enable healthcare organizations to derive intelligence and value from all Big Data ingested and amassed in the data lake.

A well-designed and effective data lake architecture delivers these important capabilities:

- **Ingest & Store:** Stores structured and unstructured data for all types of analytics from numerous sources, blending capacity and performance as needed for the analytics use case.
- **Analyze:** Supplies modern data management and analytics tools for all types of analytics, including Hadoop-based, In-Memory No-SQL, and Scale-out MPP.
- **Surface & Act:** Provides users and applications with data that will enable real-time changes in outcomes and influence critical decisions.

**CHART A COURSE FOR REAL-WORLD HEALTHCARE USE CASES**

EMC has worked with several healthcare providers that have begun their journey to a data lake across various uses cases, including patient care delivery, clinical research, population health management, and security analytics. This section highlights how healthcare providers have achieved preliminary benefits from the data lake, such as better clinical outcomes and protocols, reduced costs, faster and more accurate clinical research studies, and improved data security.

**PREDICTIVE ANALYTICS TO PROACTIVELY MANAGE PATIENT HEALTH**

By unifying data across the healthcare enterprise, providers can perform patient data intelligence analytics to help improve care coordination and management to help reduce costs. For this use case example, providers feed the data lake from multiple clinical sources, including the EMR, laboratory, PACS, patient generated data, and reported outcomes. (Figure 2) Retrospective data is combined with target-rich, high-value data that is accessible and available in real-time for advanced analytics. Analysts and data science teams can then conduct specific queries of data to further improve treatment outcomes and care protocols.
EMC recently worked on a project with a hospital CFO to help identify potential factors that were contributing to organizational risk. Analyzing data residing in the health system’s data lake, the CFO sought to reveal how, where, and when patients contract hospital-acquired conditions, such as sepsis. Data lake analytics uncovered that a higher rate of sepsis was occurring on a certain floor when a particular clinician was on duty. To resolve the issue quickly, the provider retrained the clinician in hand-washing procedures, helping to reduce such incidents in the future.

Data lake analytics also can help predict which patients or cohorts are at highest risk for re-admittance due to acute myocardial infarction, heart failure, and pneumonia. For this use case, EMC worked with data analysts and scientists at an integrated delivery network (IDN) to examine patient characteristics, outside influences, and population statistics. The IDN analyzed this data to score patients from 0 to 100—the higher the score, the higher the risk of that patient being readmitted within 30 days. For patients who scored above 50, the IDN developed a high-touch, post-discharge treatment plan, including remote monitoring for medication adherence, to help prevent re-admittance.

Integration and linkage of data is also critical as healthcare organizations consider predictive staffing models. At another health system, EMC worked with analysts to review community event data, weather conditions, flu progress, and other factors for the purpose of determining the likelihood of a surge in patient volume. By mapping insights from this analysis to internal staffing schedules, the health system now can ensure sufficient resources are on hand even when needs change rapidly. This is important not only to ensure proper care but also to improve staff satisfaction and loyalty.

Ultimately, a data lake helps healthcare organizations run their operation “as a business.” Real-time insights and predictive models provide the key to increasing competitive advantage while delivering better clinical outcomes, improving productivity and efficiency, reducing risk, and lowering costs.

**TARGET PRECISION MEDICINE WITH CLINICAL RESEARCH**

As healthcare organizations work to improve disease treatment and prevention, a data lake can be used by multiple constituencies for clinical research projects managed by healthcare systems, academic medical centers, clinical research organizations (CROs), and independent research laboratories. Data lakes can simplify access to information from EMRs and other clinical sources, as well as data generated by outside medical institutions, research facilities, and government agencies.

As new research is conducted, information, knowledge, and insights from those projects become part of the collective intelligence of the data lake to support more advanced studies in the future. The potential impacts can extend from pathology to pharmacovigilance to genomics. With detailed insight into the patient’s genomic profile, combined with routine and molecular-level pathology, oncologists can create a highly targeted, individualized therapy. The clinical research use case for the data lake becomes the pathway to precision medicine.
EMC is currently collaborating with a large healthcare system that is working to bring together the large amounts of data being generated across the organization through separate research activities into a data lake to enable its medical community to explore and develop new insights into human disease to improve diagnostics, treatment, and the lives of patients. With the goal to speed clinically relevant research into the clinical setting, the healthcare system’s data lake environment is being built with products and solutions from across the EMC Federation including EMC, VMware, Pivotal and RSA. For additional information, go to http://www.emc.com/about/news/press/2014/20141117-01.htm.

Using a data lake to provide Clinical Research as a service allows researchers to focus their high-level skills and domain knowledge on making discoveries instead of wrestling with technology. No longer do individual research teams need to acquire and deploy their own IT infrastructures. Instead, IT can allocate a research environment from the data lake and present it as-a-service to researchers with all the performance, data protection, and security required, improving research productivity and enabling faster time to insights.

In addition, data lakes can help reduce cost when conducting clinical trials for new drug efficacy. It allows the healthcare organization to perform cohort analysis across the patient population and identify candidates that most precisely match the profiles needed for the clinical trial. Consequently, the organization can set up the trial faster, avoid false starts that require requalifying participants, and ultimately gain more accurate results that help improve patient treatments.

**IMPROVE POPULATION HEALTH MANAGEMENT**

To further minimize risk, providers need to work proactively to manage health and wellness across the entire population served by the hospital or integrated delivery network (IDN). Using a data lake that incorporates individual patient records, regional health statistics, demographics, and other vital data, the organization can create a population patient registry to track health trends and classify patients according to defined risk profiles. This enables the health system to risk stratify the patient population, identify individuals at highest risk, and then implement proactive programs that target conditions that drive the highest cost.

For this use case example, a profile is created that is lifestyle-based such as a smoker with a chronic condition such as diabetes. Other profiles could include people with comorbidities, or individuals with a family history of heart disease or certain cancers. With detailed insights from the data lake, outreach and care strategies then can be ultimately developed for each patient profile.

Depending on clinical protocols and patient classifications, patients would then receive custom discharge plans. Consider a patient who recently underwent open heart surgery, lives on a third floor without an elevator, and lacks convenient access to a pharmacy. The discharge plan may include a visiting nurse for the first month to monitor the person’s condition and medications.

EMC recently provided subject matter expertise to a health system that wanted to improve community wellness management for diabetic patients to ensure they received all appropriate screenings, performed regular blood sugar tests, and maintained a healthy diet. When overdue screenings were identified, a nurse navigator was prompted to call the person and make any necessary appointments. In this case, the use of analytics is contributing to more positive patient outcomes along with reducing the overall cost of care.

**SECURITY ANALYTICS TO PROTECT HEALTH INFORMATION**

Controls like firewalls and authentication may not be sufficient enough to protect healthcare organizations from today’s increasing cybersecurity threats from internal and external sources. Healthcare providers need to quickly identify and analyze behaviors of people or programs that have or appear to have valid credentials to view specific and sensitive patient information. An organization that consolidates all system logs and network activity into a data lake can get better and faster at spotting anomalies, which leads to quicker and more targeted responses. IT management then can close the loop by using intelligence generated from the data lake to build predictive models of when and where problems are most likely to occur.

Healthcare providers can use solutions such as RSA Security Analytics to analyze historical data, and any other business and clinical context information in the data lake. This environment helps providers recognize issues that would not be detected by formal detection and remediation methods, as well as new insider threats and vulnerabilities specific to your infrastructure configuration.

With RSA Security Analytics, healthcare organizations gain the following benefits:

- **Visibility**: Spot advanced attacks with complete visibility from the endpoint to the cloud. Eliminate blind spots where threats can take root with visibility across logs, networks, endpoints and cloud data.

- **Analysis**: Detect and analyze even the most advanced attacks in real time. Begin finding incidents immediately with out-of-the-box reporting, intelligence, and rules.

- **Action**: Prioritize investigations and streamline multiple analyst workflows, enabling immediate incident response and escalation that provides the advantage of time to the defender. Understand exactly what is happening, what to do about it, and how to prioritize workflow.
JUMPSTARTING THE JOURNEY: THE FEDERATION BUSINESS DATA LAKE

To simplify and accelerate deployment of a data lake, EMC has developed the Federation Business Data Lake reference architecture to help healthcare providers start and accelerate their journey. This fully engineered data lake incorporates pre-configured building blocks from EMC, Pivotal, RSA, VCE, and VMware, as well as an open source ecosystem for building new capabilities using your existing infrastructure and third-party solutions. (Figure 3)

By removing the challenges of integration across multiple vendors, the Federation Business Data Lake allows healthcare providers to start building analytic models for a variety of healthcare use cases in as little as one week rather than months. Providers also can use automated tools to aggregate, manage, and analyze data created across the healthcare system to gain more holistic patient care insights.

EMC is currently working with a healthcare organization that is leveraging the Federation Business Data Lake solution, which is built on a secure hybrid-cloud solution that consolidates patient healthcare information into a data lake incorporating 9.2 billion medical events. This advanced approach to delivering technology is comprehensive in that it provides for real-time data capture and extremely rapid analytical processing that seamlessly facilitates the redistribution of healthcare data. As a result, patient data from thousands of disparate sources is combined to provide the analytical insights to inform, enable and empower the point of care. For additional information, go to http://pulseblog.emc.com/2015/03/18/inovalon-and-emc-federation-redefine-healthcare-through-massive-data-lake/

AGILE, HIGHLY SCALABLE STORAGE FOR FAST-GROWING HEALTHCARE DATA LAKES

At the foundation of the data lake, EMC storage solutions offer a highly effective strategy to consolidate your storage infrastructure into a simple, efficient, scalable and flexible shared storage environment that is easy to set up, manage, and extend as your data and applications grow. For example, EMC Isilon scale-out network-attached storage (NAS) provides a simple, scalable, and efficient platform to store massive amounts of unstructured data and enable applications such as medical imaging to create a scalable and accessible data repository without the overhead associated with traditional storage systems.

EMC Isilon enables providers to build a scale-out data lake where they can store their current data and scale capacity, performance, or protection as their business and clinical data grows. The scale-out data lake helps lower storage costs by efficient storage utilization, eliminate islands or silos of storage, and lower management costs of migration, security, and protection. For more information on the key capabilities of the EMC Isilon Scale-Out Data Lake, go to http://www.emc.com/collateral/white-papers/h13172-isilon-scale-out-data-lake-wp.pdf

EMC Elastic Cloud Storage (ECS) further strengthens the data lake architecture with a new set of features and capabilities, including geo-scale analytics that support HDFS, flexible deployment options, and multi-cloud APIs to seamlessly connect to public clouds.
And, VCE Vblock Systems integrate compute, network, and storage technologies as converged infrastructures that simplify deployment, maintenance, and scalability. Vblock Systems are flexible and can be configured with any EMC storage technology. In addition to Isilon and ECS, Vblock Systems also can be configured with EMC XtremIO all-flash storage arrays, which provide increased performance for high-demand Hadoop, high-speed offload ingest, and high-speed block technologies where the lowest latency and highest throughputs are required. XtremIO storage enables data scientists to use new analytics tools that require enhanced read/write throughput and deliver new levels of insights.

As the value of use cases leveraging the data lake continues to grow, providers need to incorporate enterprise-grade data protection and disaster recovery capabilities. EMC offers several solution options for data lake protection including Hadoop Distributed Copy for deployments where compute and storage have been integrated and Isilon snapshots managed by the EMC Networker Snapshot Management for deployments where compute is separate from storage and storage is shared. For more information on EMC Data Protection Solutions for the Business Data Lake, go to http://www.emc.com/collateral/white-papers/h13932-data-lake-protection-tech-review.pdf

The EMC Clinical Archiving solution helps healthcare organizations to increase the availability of historical data from legacy EMRs and other clinical and business systems to enable new tributaries to be created which feed the data lake. The liquidity of this type of data is essential for supporting analysts and decision makers who need predictive models to navigate emerging business challenges while supporting clinical innovation for the delivery of precision medicine. The solution provides the ability to dynamically siphon patient information in a single, universal repository using XML, an open data format. For more information on the EMC Clinical Archiving Solution, go to http://www.emc.com/collateral/solution-overview/h13023-clinical-archiving-so.pdf

AN ARRAY OF OPEN SOURCE ANALYTICS SOLUTIONS

The analytics layer is completely virtualized using VMware running on Vblock Systems with predefined analytics use cases, automated provisioning, and a configuration comprised of the Pivotal Big Data Suite. This suite includes PivotalHD, featuring the SQL-on-Hadoop engine (HAWQ). It also provides enterprise-class SQL for seamless integration and interoperability with top analytics platforms such as SAS, Tableau and others.

As another solution in the Pivotal Big Data Suite, Pivotal Greenplum provides a massively parallel database for Big Data analytics, while Pivotal HAWQ offers a fully compliant SQL-on-Hadoop analytics engine. With its open architecture, the data lake allows data scientists to use their choice of advanced analytics solutions, including BDS offerings—Pivotal Gemfire and Pivotal Cloud Foundry—as well as open source solutions such as Redis and RabbitMQ.

Hadoop provides an open-source software framework for distributed storage and distributed processing of very large data sets. It is the core of a Big Data analytics environment. As part of EMC’s participation in the Open Database Platform (ODP) initiative, the data lake architecture supports a choice of Hadoop distributions, including Pivotal HD, Hortonworks, and Cloudera.

These technologies form the basis for building analytics models to address a range of healthcare use cases. Each model can draw from a mix of structured and unstructured data in clinical and business environments, as well as publicly available data, trusted private third-party data, and social media feeds.

AN OPEN-ENDED TOOLBOX FOR HEALTHCARE ANALYTICS

The open analytics toolbox provides data scientists with numerous options for creating special-purpose analytics environments and building stochastic, prescriptive models, as well as other types of models. Data scientists choose the products they are most comfortable and familiar with from the toolbox. Some examples include MongoDB, Alteryx, Cassandra, Flume, Tableau, and SAS.

Using these tools in the cloud-based data lake, data scientists can spin up production analytics environments or sandboxes in minutes. This dramatically accelerates time-to-insight compared with traditional analytics infrastructures that can take four to six weeks to create. The tools also provide a rich set of capabilities for reporting analytic insights to support real-time decisions affecting everything from predictive staffing to prescriptive discharge plans to precision medicine.

SIMPLIFIED ACCESS TO A CATALOG OF CLINICAL AND BUSINESS INSIGHTS

The data and analytics catalog makes insights derived from the data lake easily consumable across the health system. Through the catalog, business and clinical leads can conduct a federated query—asking a Google-like question—and obtain intelligent answers from the data lake.

For example, an orthopedic surgeon may want a list of patients who had knee replacements and were assigned certain ICD-9 codes because a clinical research study showed a pattern of post-surgery emergency room visits among this group. By quickly identifying these patients, the healthcare provider can proactively reach out with prescriptive measures to prevent complications, saving staff resources and improving the long-term outcomes for these patients. For additional information on the Federation Business Data Lake, go to http://www.emcfederation.com/solutions/business-data-lakes.htm.
HOW TO SET SAIL ON A HEALTHCARE DATA LAKE

EMC offers educational and professional services, as well as best practice recommendations, to help healthcare providers launch their planning and deployment of a data lake.

BUILD A DATA LAKE PATHWAY AND FOUNDATION

As a starting point, EMC offers a Big Data Strategy Methodology to guide stakeholders with defining their goals, identifying legacy investments, technology and organizational, and collaboratively creating a future-state Big Data architecture. (Figure 4) The methodology starts with stakeholder interviews and business evaluations to establish readiness, includes a Big Data Vision Workshop to identify high value use cases, builds consensus around technical feasibility, and culminates in a preliminary roadmap for achieving meaningful outcomes from one or multiple use cases.

During the Big Data Vision Workshop, EMC Big Data experts work side by side with clinical, business, and IT leadership over several weeks to map out goals and strategies for the provider’s Big Data initiative. The EMC team captures initial data sources and completes preliminary analytic modeling to explore “the art of the possible” with stakeholders at the client site.

For organizations who have an identified high value analytic use case, EMC offers the Proof of Value Service. This engagement demonstrates the ROI of the target Big Data use case in a working analytics environment using real data collected from the healthcare provider to construct a working pilot. It is a compelling way to understand the potential impact of data science on the organization, and the skills needed to fully extract its value.

EMC’s Technology Onboarding Service also is available when the healthcare provider is ready to deploy a healthcare data lake. With this service, EMC consultants install and deploy the data lake infrastructure, optimize the analytics environment for target use cases, and configure and customize analytics applications to the provider’s specific requirements.

DEVELOPING DATA SCIENCE EXPERTISE WITHIN THE HEALTH SYSTEM

While technology is a critical aspect of a data lake, developing strong data science capabilities within the healthcare organization is equally important. To help your organization get started, EMC provides training and data science practitioners to help teach you and your staff how to fish. The mentorship approach enables your organization to execute a high value use case while mentoring your staff.

EMC offers a range of educational services to help everyone from business and clinical leaders to aspiring Big Data practitioners and seasoned data scientists to increase their effectiveness in using Big Data. EMC training courses enable leaders in the organization to develop a baseline understanding of data science and Big Data. This helps them identify opportunities for integrating Big Data into their business strategies and decision-making processes. For team members with greater experience in Big Data, EMC also offers advanced-level courses for specific methods and tools with EMC Proven Data Science Certification.
For additional resources, healthcare organizations can engage EMC Consulting resources to jumpstart burgeoning data science practices through in-residence mentoring and skills training. Providers also can learn from Pivotal Labs, a division of EMC that teaches “extreme” programming for data scientists looking to extract even more value from data lake technologies.

**OPERATIONALIZE PROCESSES FOR DATA GOVERNANCE AND SECURITY**

As with any new practice, data science requires formal processes to ensure proper governance over information access and analytics reporting. With this mind, healthcare organizations must identify and verify the quality of all data sources flowing into the data lake. To help with this, EMC recommends the health system develop policies and procedures to establish when, where, and how data can be shared inside and outside the data lake.

In addition, department leaders across the organization must understand the role of data science and how to best integrate analytic insights into their clinical and business processes. This will likely require training, and potentially further organizational transformation. EMC offers data governance acceleration engagements, training and mentorship services to help healthcare providers jump start their governance programs to address the challenges of data quality, privacy, and governance.

**CONCLUSION: EMBRACE THE FUTURE OF PREDICTIVE CARE**

The complexities of achieving improved patient outcomes, regulatory compliance, and strategic business goals require healthcare organizations to become more proactive in the care of their patients. Technology advancements in storage and data science tools can provide a vital set of predictive assets to meet these growing demands. The availability of lower-cost, scale-out architectures, converged infrastructures, and advanced security and analytics solutions is enabling IT departments to deploy analytics at scale.

With a data lake, healthcare organizations can consolidate legacy systems and data with new data sources, including those flowing from cloud services, social media, or third parties, such as payers. Data lakes enable healthcare providers to execute data analytics across disparate systems running databases, data warehouses, and structured or unstructured data sets, without impacting day to day operations or access to data.

Ultimately, healthcare providers can use the data lake to shorten their analytics time-to-value as they make the transition to value-based care. Moving forward, the use of advanced analytics for next-generation care delivery will be a key differentiator for providers in a competitive marketplace. With data science based on predictive analytics, providers can deliver better patient outcomes, experience fewer complications, less unnecessary emergency room interventions, and higher levels of wellness across the population—and all at a lower cost.