THE ROLE OF DATA LAKES IN HEALTHCARE

PERFICIENT®

vision. execution. value.

According to information technology research and advisory company Gartner, there is a market shift towards hybrid approaches to data management, as companies look to solve complex challenges with a mix of alternative and traditional deployments.[1]

Healthcare providers and health plans are asking the question, "Do we need an enterprise data warehouse, a data lake, or both as part of our overall data architecture?" This question is top of mind with healthcare executives who are challenged to improve medical outcomes for patients, drive patient and member engagement, and bend the cost curve. In order to do so, these organizations are dependent upon the ability to rapidly ingest and analyze large volumes of data in batch or real-time from an extensive range of sources in a variety of formats.

This guide takes a deeper dive into the role data lakes play in healthcare and will:

- **Explore** the data lake concept and how it differs from a more conventional enterprise data warehouse approach
- **Discuss** how a data lake typically co-exists with an enterprise data warehouse to enable advanced analytics that are difficult to attain with a traditional enterprise data warehouse architecture
- **Present** use cases that are best served by a data lake environment, and the typical starting points for a data lake effort and associated architecture

WHAT IS A DATA LAKE?

DATABASE SCHEMA

A data lake is not just Big Data; it is a collection of various data assets that are stored within a Hadoop ecosystem with minimal change to the original format or content of the source data (or file). Thus, the data lake lacks a formal schema-on-write. Access to information contained within the data lake uses various tools which apply "schema-on-read."

An enterprise data warehouse (EDW) is not a data lake. By definition, an EDW is "an integrated, subject-oriented, time-variant and centrally managed database system intended for mixed workload management and large-query processing."[2] The EDW's schema is predefined in that as data is written to the EDW it must conform at write (a.k.a. "schema-on-write").

There are more differences between an EDW and a data lake, but the most significant is that the structure of the data is not fully known nor always enforced in a data lake.

TIME TO DELIVERY

Healthcare organizations are attracted to the concept of a data lake because it does not have the same EDW requirements of extracting, transforming, and loading (ETL) data in a conforming, pre-defined manner. An EDW may take several months of modeling, mapping, ETL development, and testing to move source data into a dimensional schema. While the end result is consistent, validated data before query, the time-to-delivery can impact the value of the EDW to the end user.

USER EXPERIENCE & KNOWLEDGE

Data lakes provide information in its raw format, along with a specific method for accessing data that applies the schema upon read. Generally, users of the data lake are experienced analysts, accustomed to data wrangling techniques which apply schema upon read or interpret data content from unstructured formats. Less-experienced users will struggle without significant search tools and data extraction automation.

That doesn't mean the data lake lacks metadata, nor rules governing its usage, security, or management. It's quite the opposite. A successful data lake will organize its data in a manner to promote better and more efficient access, and will re-use data management processes or introduce new tools that improve search and general knowledge of the data content.

HEALTHCARE USE CASES FOR A DATA LAKE

Traditional EDW approaches are built with the intention of attracting a large number of self-service business users with an interest in day-to-day clinical, financial and operational reporting that draws from structured and pre-processed data. On the other hand, data lakes are designed as highly agile, configurable alternatives for answering complex questions, leveraging all available data sources.

We will discuss two basic use cases for a data lake in healthcare:

1. **Predicting healthcare costs**
2. **Evidence-based care**

PREDICTING HEALTHCARE COSTS

Typical analysis of healthcare data comes from reports or dashboards developed against the EDW, which had extracted data from electronic health records (EHR) that are commonly found in electronic medical records (EMR) systems or in an EDW. The records (or facts) are cast against dimensions such as time, place, diagnosis, or treatment category where observations are measured performances that have a velocity or rate of change based on past measurements.

With the advent of a data lake strategy, providers and health plans are attempting to enrich their data and predict patterns of risk or greater cost using an expanded data set. These new predictive measurements can show a higher degree of relevance with added values. Advanced users incorporate the new models into actionable change with factors that are not always found in an EHR, for example patient-generated data or personal health records (PHR).

PHR is a health information record related to the care of a patient that is maintained by the patient. The purpose of adding PHR to analysis is to provide a complete and accurate summary of an individual's medical history. A PHR record has variant formats that include textual patient-reported data, family medical history, exercise or diet regime, and data from smart devices that measure heart rate and movement.

There are two issues at hand. First, an EHR does not maintain nor contain PHR to the extent that basic BI reports or dashboards from an EHR/EDW show any relevance. Second, the extreme effort required to create a schema (where no PHR standard exists) – plus the effort of collecting, extracting, managing, and providing PHR data into the EDW – poses a considerable time-to-value challenge.

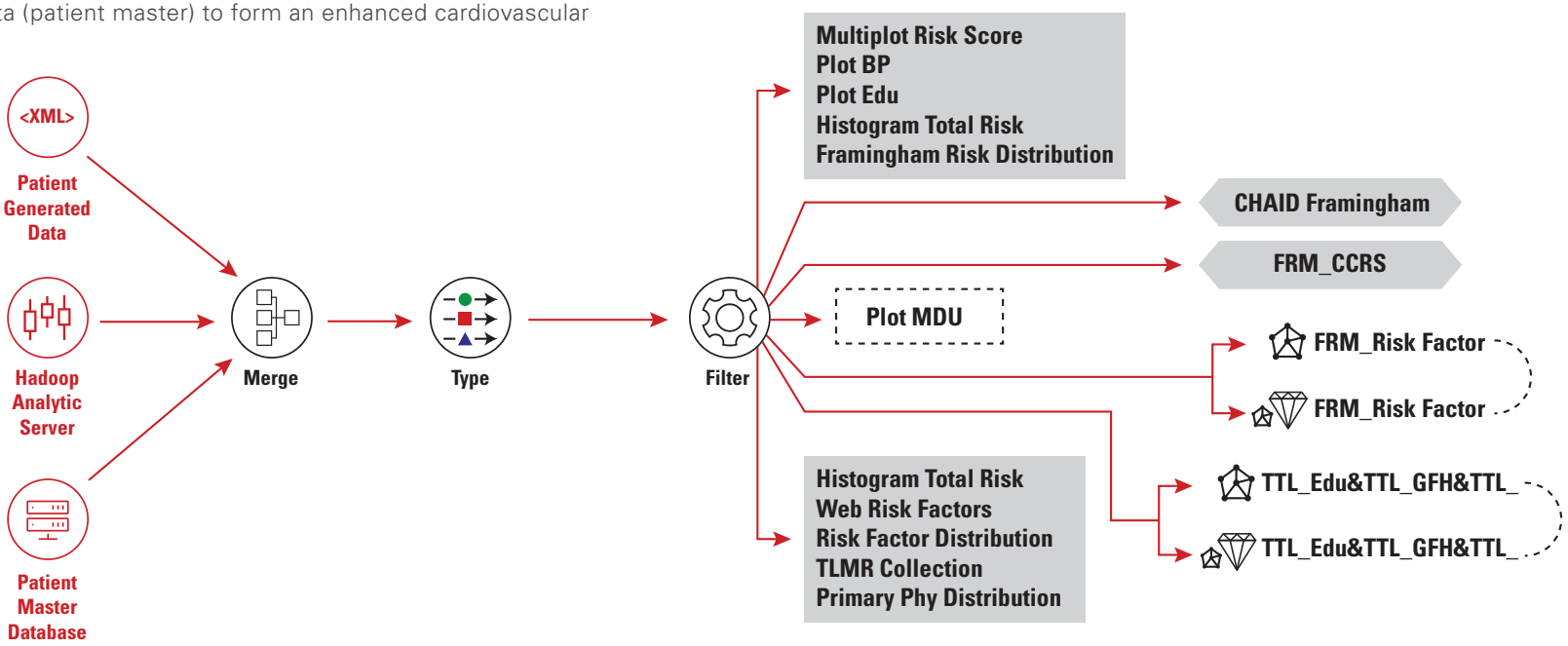
The time-to-value proposition makes a strong case for a data lake. PHR records can be stored in original or near-original format, pushing off schema until read. PHR records can be searched, catalogued, and even tagged with meta-fields garnished from keywords or word pairs to improve knowledge of the data content. Once PHR factors are understood they can be included in predictive measurements and ranked by their relevance.

THE FRAMINGHAM HEART STUDY AND PHR PREDICTORS

The Framingham Heart Study is the origin of the term "risk factor." With more than 1000 published medical papers related to the study, it is one of the most influential evidence-based studies that shows how heart disease is effected by both measured factors (such as blood pressure, smoking, cholesterol, age and gender) as well as lifestyle, environmental factors, and genetics.[3]

Figure 1 shows an example of how to use a data lake (PHR Data) in concert with EHR/EDW data (patient master) to form an enhanced cardiovascular patient risk model. This model relies upon factors extracted from PHR records stored as XML data types. Processing merges PHR/XML fields to patient structured data, calculating risk scores and determining relevancy factors. Reports are then generated on the standard heart health scores and compared to new risk models using a broader range of variants.

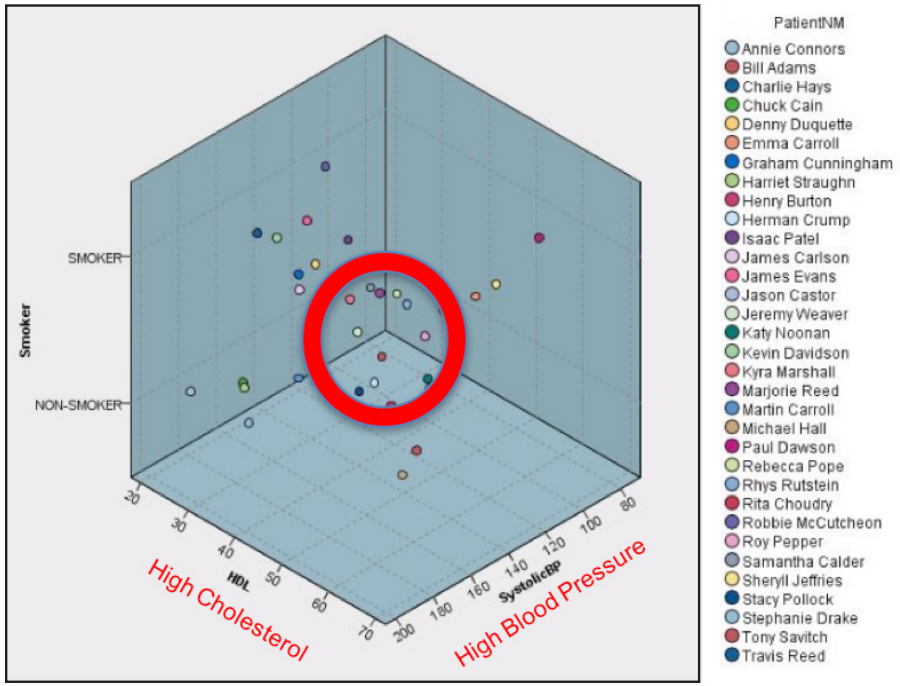
Figure 1 - Cardiovascular Patient Risk Model



Once analysis is reviewed by a healthcare specialist the data can be promoted via applications or dashboards and shared with patients to determine what types of preventative measures should be taken prior to the patient being discharged.

Published studies have shown how heart health impacts the cost of healthcare, and by identifying patients with higher risk, providers can adopt better treatment and care programs at a more effective cost/benefit ratio.[4]

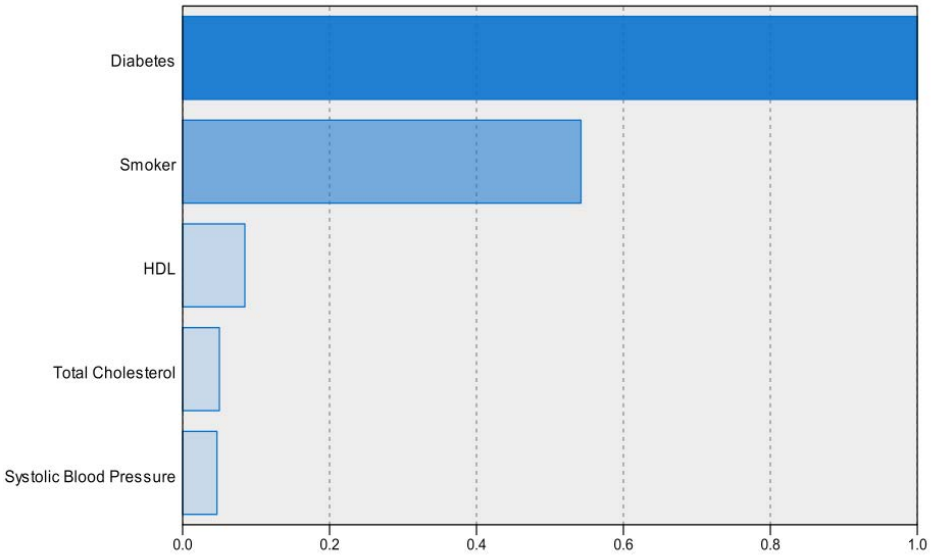
Figure 2: Basic Framingham Analysis



** Patient names have been modified to protect their identify*

When examining the basic data elements found in Framingham Heart Study (cholesterol, smoking habit, blood pressure, sex) we can begin to see patterns from the analysis.

Figure 3: Predictor Importance



Further analysis detects the most important predictors for heart disease and can help identify patients with the greatest risks.

DATA LAKES AND EVIDENCE-BASED CARE

Evidenced-based care or practice (EBP) is the integration of clinical expertise, patient values, and the best research evidence into the decision-making process for patient care. Clinical expertise refers to a clinician's cumulative experience, education and clinical skills. The patient brings to the encounter his or her own personal preferences and unique concerns, expectations, and values. The best research evidence is usually found in clinically relevant research that has been conducted using sound methodology.[5]

A data lake represents a cost-effective platform for EBP. The term "evidence- based" implies that some types of evidence are sufficient to guide general analysis but fail to offer the desired level of clinical rigor required for a broader approach to multi-variant analysis, drawing on data from several areas outside of the EHR. Resources of evidence include EHR/EMR systems (reflecting internal-based records), clinical or trial published research (datasets, publications, and sometimes programs), government assets (surveys, publications, collective libraries), genomic research (GEO datasets and profiles), and of course PHR (patient-reported data, family medical history, exercise or diet regime, data from smart devices).

The body of evidence is broad, deep, and expansive. Furthermore, it would be a struggle for any EDW to incorporate such vast amounts of information with such tremendously variable content, so Big Data – and more specifically, the data lake – provide an optimal platform.

Professor Dhavendra Kumar, a Consultant in Clinical Genetics at the University Hospital of Wales, Cardiff University, United Kingdom proposed that practice of evidence-based medicine (EBP) would include analysis of genetic and genomic profiles and its success would depend upon the robustness of translational research.[6] One of the most successful applications of this kind of data has been in the characterization of human cancers, including the ability to predict clinical outcomes.

Utilizing the data lake approach, we can clearly see how it would support the multi-variant analysis of combining GEO data sets with PHR (family history) and EHR (patient treatment) to develop evidence-based care models. Data sets from the GEO libraries are copied into the data lake in original format based upon study criteria (assuming only a subset of the GEO library is required). Modeling tools (SAS, SPSS, R, Python or Scala) then analyze statistical relevance between genome marker, patient information, treatment and outcomes. End models that adhere to repeatable data science methods are now made available to clinical staff for validation and verification.

CANCER AND GENETICS

Originally published in 1988, Erich Segal's novel "Doctors" makes the case that of the thousands of human afflictions and ailments, only a handful have a cure, or even a treatment that is based on understanding. The rest are treated by trial and error. Cancer is a prime example of the trial-and-error approach, and according to oncologist Siddhartha Mukherjee's Pulitzer-prize-winning best selling novel, "The Emperor of All Maladies," cancer is inevitable; if anyone lives long enough, they will have cancer.[7] This is because cancer cells are a byproduct of mutation, and that process in turn is partly accelerated by the shortening of telomeres, which is linked to aging.

The correlation of cancer with genetics is becoming increasingly evident today thanks in part to Big Data. We know that certain types of cancers run in families, some affect certain groups of people, and others seem to be triggered by certain genes that manifest under the appropriate conditions. In this context, one must mention the role that 23andMe (www.23andme.com) is playing. Through a blood or saliva sample, the company can analyze your genetic profile not only for international DNA ancestry connections, but more importantly to identify a relative 'genetic predisposition' for all types of cancer. For the individualized cancer treatment this could mean the difference between life and death.

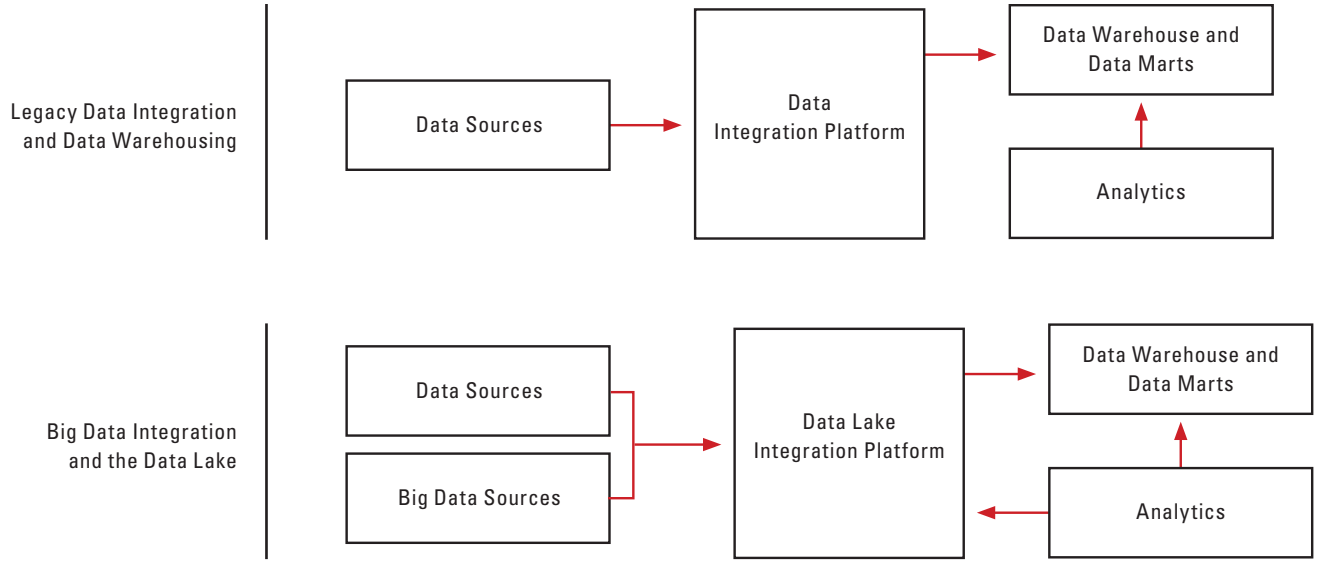
THE BIG DATA REVOLUTION AND CANCER

The top minds in the world now have access to vasts amounts of data; an ocean of information from which to draw up patterns that could potentially lead to cures for cancer and other diseases. However, in order to draw out meaningful correlations from these patterns, we need to embrace all the number-crunching power of today's supercomputers and the ever-advancing field of Big Data.

These new technologies enable us to ask questions of cancer in ways that were simply not possible before. In order to draw a big picture view of cancer, the cancer genome first needs to be broken down into 100 base-pair long fragments and then put back together. This process, called gene sequencing, typically requires the sequencing of hundreds of millions of such pieces, and hundreds of these tests have to be performed in order to make an observation that can be a candidate for a pattern. That is the scale of data that we must embrace.

While traditional EDW solutions often necessitate multiple data transformations, in the case of analyzing patterns for cancer it is imperative that the data is not accidentally contaminated in this process of data manipulation – a key feature of data lakes.

Figure 4 Transitioning to the Data Lake



THE DATA LAKE PLATFORM OF HEALTHCARE INFORMATION SYSTEMS

A data lake is a powerful architecture with the potential to transform healthcare by providing a singular repository for structured, semi-structured, variable-format, internal, and external data. It enables data scientists and healthcare analysts to mine data that is scattered across data warehouses, data marts, operational systems, transactional systems, and external data sources.

The value of a data lake is realized in its power to share data and support for rapid exploration and discovery processes. Data science teams use these tools to uncover variables and metrics that better predict business performance and support decision making. A data lake enables predictive and prescriptive analytics necessary to support healthcare use cases and initiatives.

Figure 5: Provider Data Lake

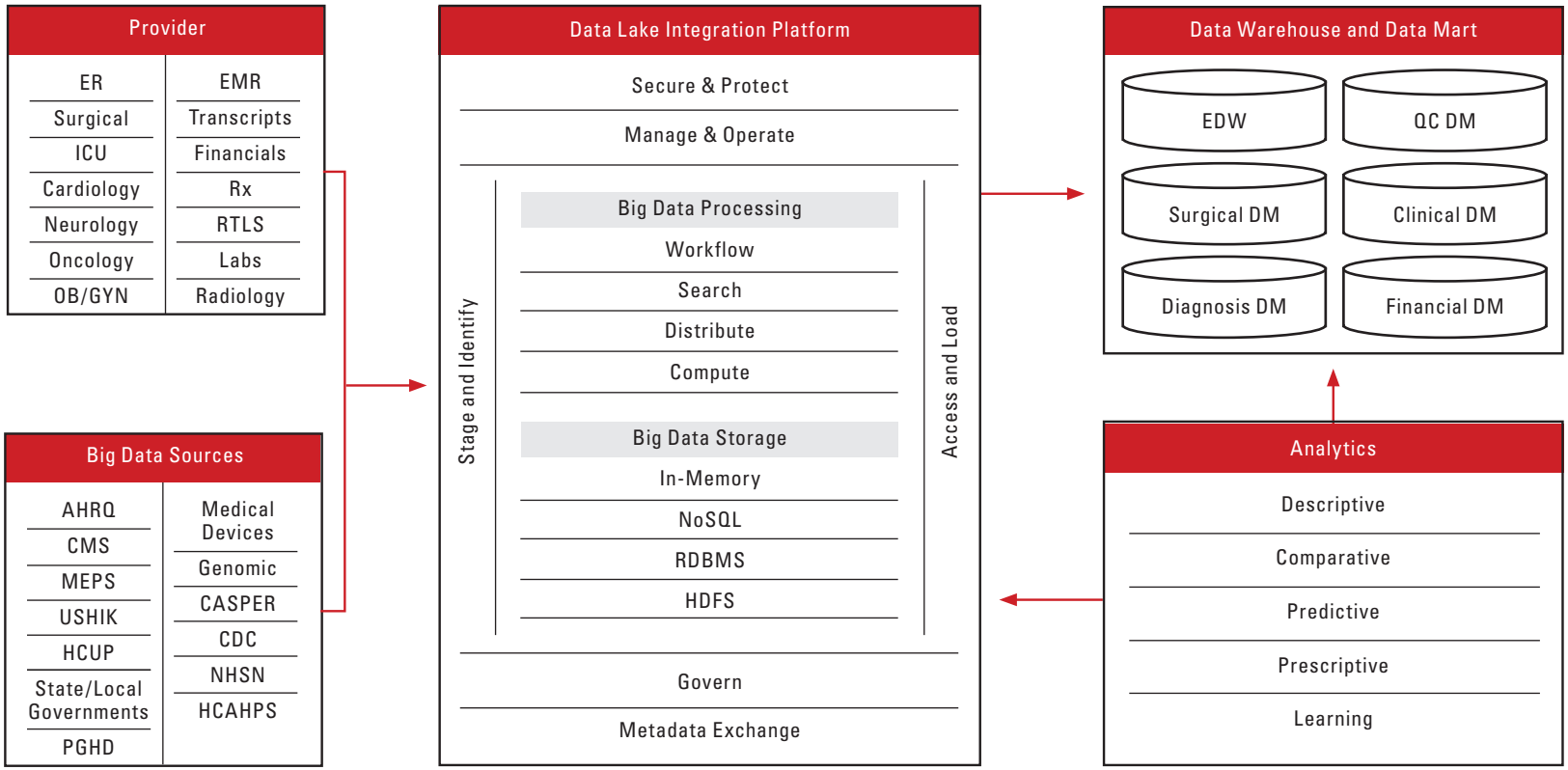
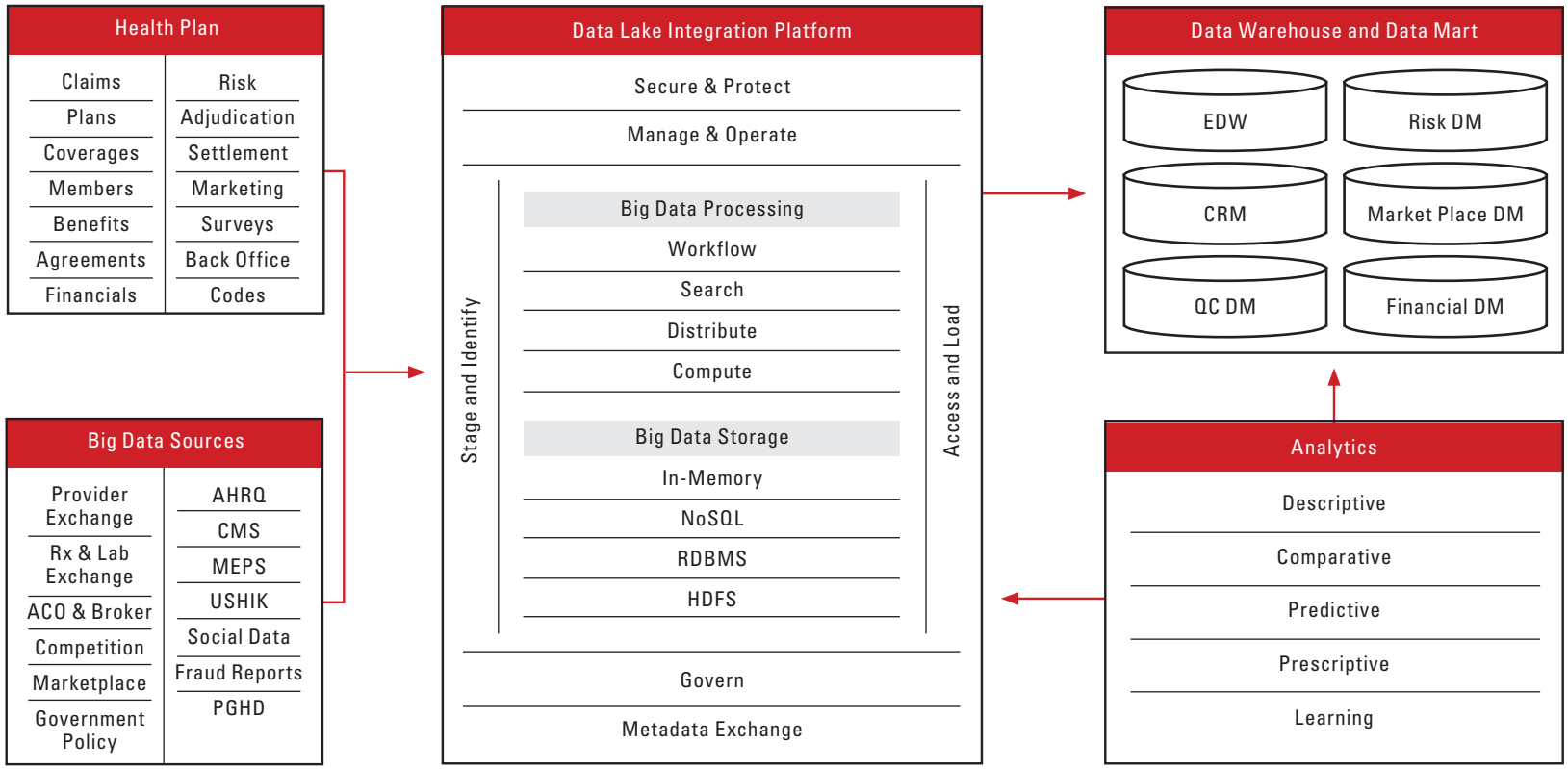


Figure 6: Health Plan Data Lake



In summary, a data lake has the following characteristics:

- Centralized data integration, with a schema-at-read, indexed-for-search optimized store that is based on usage, as well as historical archive and key value data retrieval
- Optimized storage with relevant latency, redundancy, isolation, and durability (memory, file system, NoSQL, or database)
- Mechanism for rapid ingestion of data, distributed compute, and workflow control
- Ability to integrate and map data across multiple data types and sources
- Provides access to users, EDW, data marts, and analytical applications
- Incorporates metadata exchange and governance
- Cataloged and indexed for rapid search and data retrieval
- Ability to manage security, permissions, and provide data masking on sensitive patient information
- Operationally managed and centrally controlled
- Supports self-provisioning of compute nodes, data, and analytic tools without IT intervention

Success Story

CEDAR GATE TECHNOLOGY

Cedar Gate Technology (CGT) provides analytical software solutions for the healthcare industry. It offers value-based healthcare, prescriptive analytics, and enterprise management solutions.

The Affordable Care Act requires healthcare organizations to do more than ever before – to cut fixed and variable costs, manage capacity and risk, and grow revenue while improving outcomes across patient populations. All of this requires a robust analytics solution. We worked with CGT leadership to perform a comprehensive review of data sources, data types, and analytical needs, and recommended Hortonworks as the Big Data platform.

We assisted in the installation and configuration of Hortonworks across multiple environments (development, test, production), and worked with subject matter experts to distill the desired user experience and metrics into business use case requirements for the CGT data lake. We took an agile approach using incremental, iterative work sequences known as sprints, electing to build portions of the application that deliver the most business value and functional product rapidly.

Apache Pig, a platform for analyzing large data sets, was leveraged as the data manipulation tool for ingesting client data. Refined data was ultimately stored in Apache Hive, the Hadoop Data Warehouse where the client would use Elasticsearch search engine for quick retrieval and manipulation of data to provide meaningful insights.

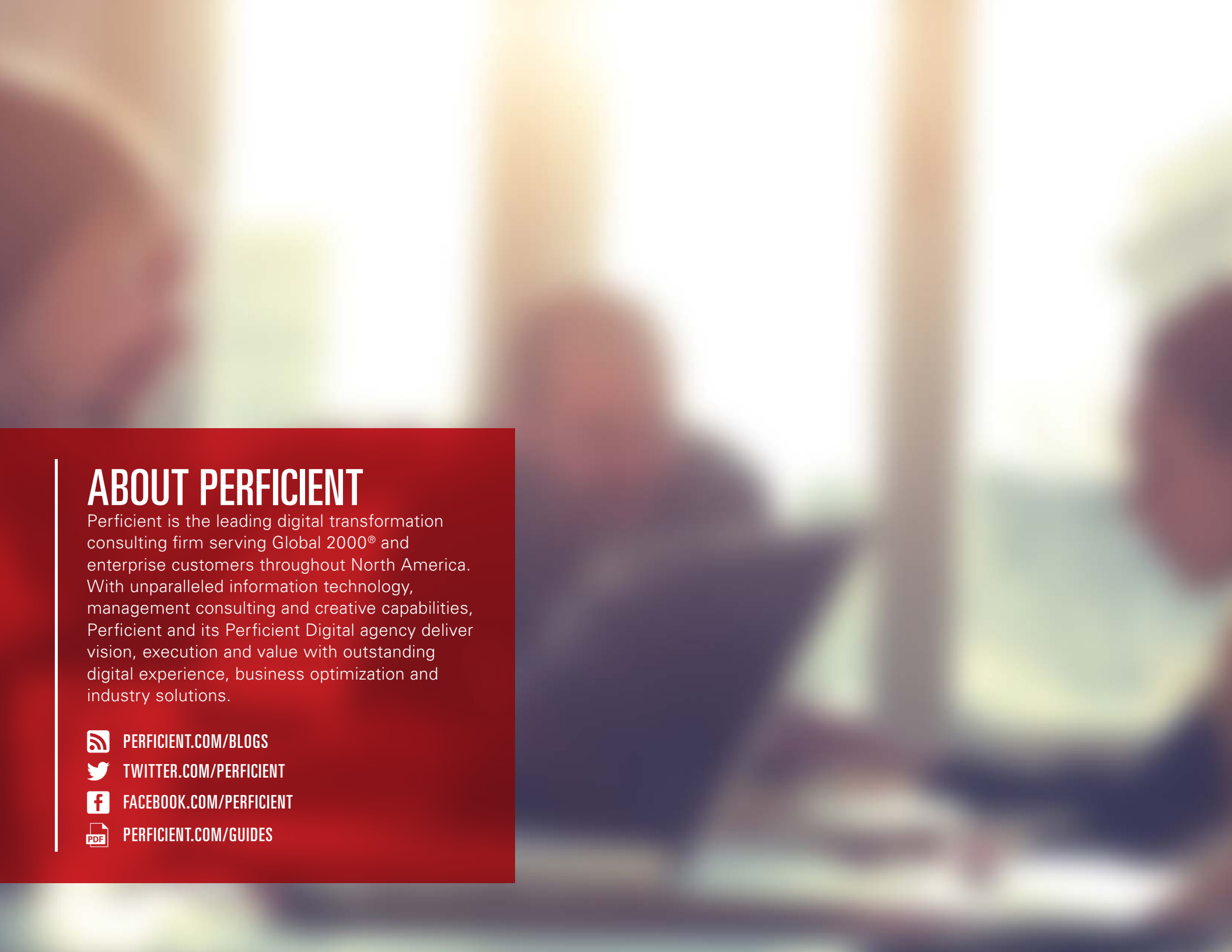
THE DATA LAKE JOURNEY AND HOW PERFICIENT CAN HELP

Perficient is recognized as one of the largest healthcare IT consulting firms in the United States. We understand the complexities of the healthcare industry and the unique challenges healthcare organizations face. Our healthcare practice delivers strategic business and technology consulting insights that help our clients transform with today's digital consumer experience demands. This strategic guidance is then transformed into pragmatic technology solutions that improve clinical, financial and operational efficiency.

We have extensive experience with Big Data, data warehousing, business intelligence and analytics, and work with our clients to transform data into timely and actionable insights.

Our client engagements begin with setting user expectations and identifying business use cases. When implementing a data lake, once we set user expectations and identify business use cases, we move into our initial discovery scrums. During discovery we identify existing or potential data sources and usage patterns. Upon completion of our discovery scrums we rapidly build a data lake ecosystem that includes information governance, data security and patient record protection, data validation, and data management.

We provide data lake roadmaps, rapid deployment pilots, and full life-cycle service engagements. Our main objective is to provide a better "time-to-value" delivery model using a data lake in conjunction with or as a replacement to an EDW and to provide best practices to ensure the data lake doesn't turn into a "dumping pond."



ABOUT PERFICIENT

Perficient is the leading digital transformation consulting firm serving Global 2000® and enterprise customers throughout North America. With unparalleled information technology, management consulting and creative capabilities, Perficient and its Perficient Digital agency deliver vision, execution and value with outstanding digital experience, business optimization and industry solutions.

PERFICIENT.COM/BLOGS

TWITTER.COM/PERFICIENT

FACEBOOK.COM/PERFICIENT

PERFICIENT.COM/GUIDES