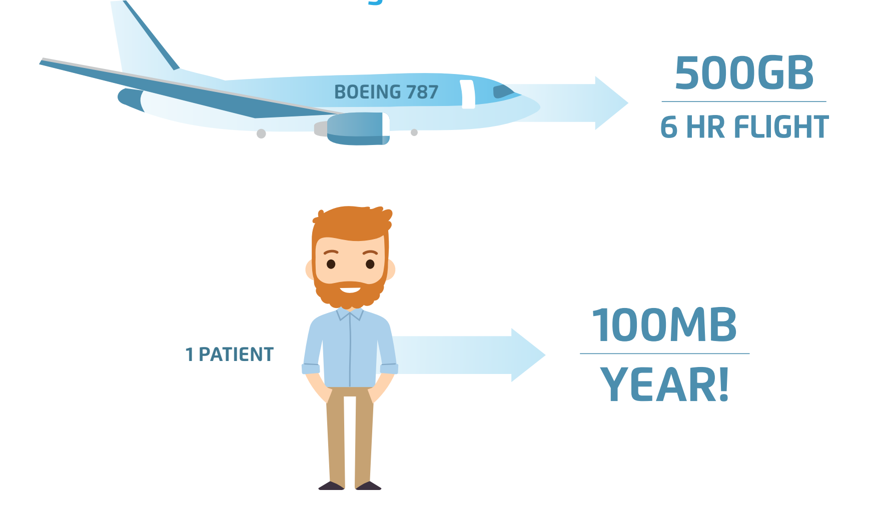
Hadoop in Healthcare: Getting More from Analytics

by Sean Stohl, SVP, Product Development and Bryan Hinton, SVP, Platform Engineering

Healthcare IT professionals are no strangers to the term big data, but, considering the larger data landscape, healthcare has only scratched the surface of the available technology and capabilities of big data. To understand our position on the big data spectrum, consider healthcare in comparison to a legitimate big data field, the airline industry: An EMR for one patient contains 100 megabytes (MB) per year, while one 6-hour flight delivers 500 gigabytes (GB).

With one GB equal to 1,000 MB, healthcare certainly has room to grow in the volume side of big data and is poised to do so (discussed more in the next section). Our current analytics infrastructure won't be able to handle this momentous increase. The key is to be ready for that

We Are Not "Big Data" in Healthcare Yet

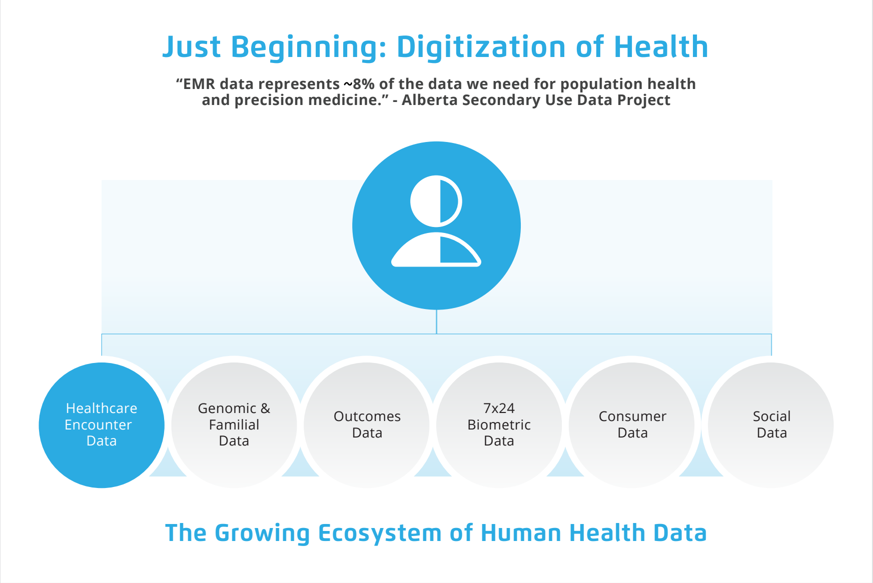


growth now by understanding the capabilities and organizational requirements of big data technology, such as Hadoop, and being fully prepared to leverage it.

IT'S TIME FOR BIG DATA IN HEALTHCARE

According to the Alberta Secondary Use Data Project, "EMR data represents [approximately] 8 percent of the data we need for population health and precision medicine." This leaves a significant amount of data to add. As the chart below describes, health data stands to grow to include five more data sets:

1. Genomic and familial data
2. Outcomes data
3. Biometric data
4. Consumer data
5. Social data


Just Beginning: Digitization of Health
"EMR data represents ~8% of the data we need for population health and precision medicine." - Alberta Secondary Use Data Project
Healthcare Encounter Data
Genomic & Familial Data
Outcomes Data
7x24 Biometric Data
Consumer Data
Social Data
The Growing Ecosystem of Human Health Data

As this additional information enters healthcare data systems, the industry will edge increasingly closer to the big data threshold—the dimensions that qualify large data as big data. Gartner analyst David Laney has identified three parameters of big data, or the "three Vs":

- **Volume**—how much data you have (from hundreds of terabytes to petabytes)
- **Velocity**—speed at which the system brings in data (from hundreds of gigabytes up to a terabyte per day)
- **Variety**—diversity of data in the system and a range of sources (from hundreds to thousands to millions of source systems, including structured, text, tagged, images, and video)

Healthcare has yet to hit the three Vs of big data, and while these parameters are a good guide to understanding big data, they don't mean that an industry can't move forward before reaching this threshold. In fact, given what we know about increasing data demands in healthcare (as explained in the previous graphic) and the potential speed of IT innovation, healthcare can (and in some cases, should) make steps toward big data now.

Moore’s Law: A Lesson in IT Preparedness

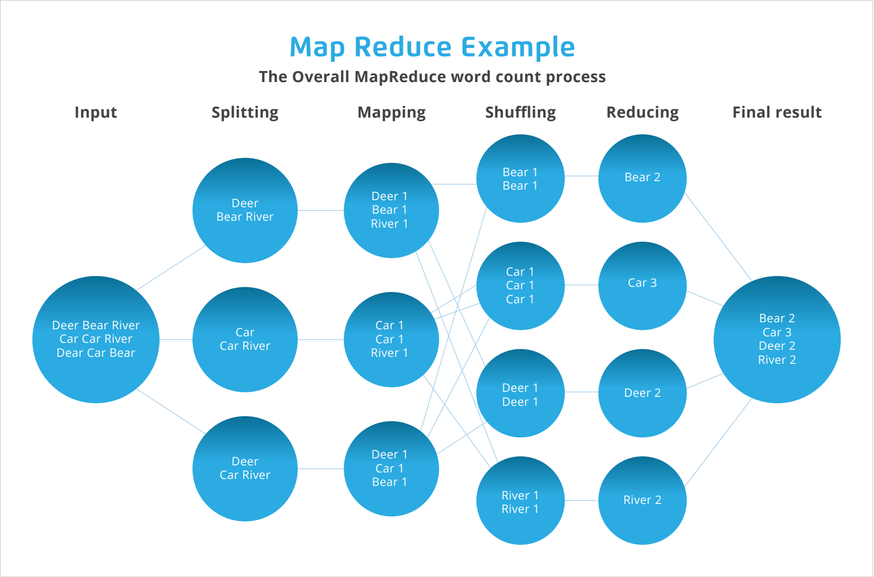
History of technology can help predict how likely (and quickly) healthcare will evolve toward big data—or to the point where the industry must use big data solutions, such as Hadoop. According to Moore’s Law, Intel cofounder Gordon Moore’s 1965 prediction, the number of transistor per square inch on a CPU chip had doubled every year since the technology’s introduction and would continue to do so for the immediate future. As this growth progressed, the tech industry would start to hit limits unless they scaled up. This means that they’d have to adopt more IT assets to support increasing demands on CPU chips. So even without volume, velocity, and variety in health data, Moore’s Law show us why it’s time to move toward big data solutions in healthcare. In other words, we need to scale up now, or we will eventually hit limits on our data capabilities.

SCALING UP FOR BIG DATA IN HEALTHCARE: HADOOP

Doug Cutting and Mike Cafarella of Yahoo introduced Hadoop in 2005. Named for Cutting’s son’s toy elephant, Hadoop is an open source software framework that uses commodity hardware to get rapidly to the data and generate answers. Cutting and Cafarella built Hadoop on two models:

- **Map Reduce**: A programming model that allows for implementation for processing and generating large data sets across hundreds or thousands of servers in a Hadoop cluster. It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.
- **HDFS** (Hadoop Distributed File System): A system that can store very large data sets reliably and stream those data sets at high bandwidth to user applications.

This simple word count chart shows how Map Reduce works to identify and group together the numbers of certain words in one type of data:



Map Reduce Example

The Overall MapReduce word count process

WHY DO WE NEED BIG DATA AND HADOOP IN HEALTHCARE?

In simple terms, we need big data and Hadoop in healthcare to prepare for the evolving data-driven needs in the industry. As mentioned earlier, we've only scratched the surface of the data we need for population health and precision medicine (we're at about 8 percent in, according to the Alberta Secondary Use Data Project). Even if we haven't hit the three Vs of big data, we're very likely heading toward more data with more complexity.

Tools for More Types of Data

A challenge in many data-heavy industries is getting different forms of data into a RDBMS (relational database management system). Structured data is in a relational format and ready to be stored in a RDBMS, but two other forms of data—semi structured and unstructured—are not in a relational format. Semi-structured data includes CSV, XML, X12 (835/837), HL7, and JSON files, as well as doctor notes with template-generated sections; unstructured data includes emails, text messages, Word documents, videos, and pictures, as well as doctor notes in free-form sections.

The Opportunity for Hadoop in Semi-Structured Data

A real opportunity for Hadoop in healthcare lies in semi-structured data. This type of data has some structure (or schema) from which to pull data (this is schema-on-read, whereas schema-on-write is structured). Doctor notes developed with template-generated sections are an example of semi-structured data, or schema-on-read.

Hadoop can be a great asset with semi-structured data because data in this format has some flexibility, and users can define their own data types and work with data of different types, shapes, and structures. In addition, you can store schema-on-read in its entirety, meaning that you don't need to decide (or necessarily know) which information will be important over time. Once this diverse data enters the HDSF, you can use it for varying purposes.

A LOOK AT VALUE: THREE USE CASES FOR HADOOP

1. **Archiving:** Hadoop adds value as an archiving tool because it can handle increasing amounts of data, including historical data from different storage tools. As a result, you can leverage historical data for your current data-driven decisions and predictions.

2. **Streaming:** In addition to historical data, you need access to real-time data (streaming). Streaming with Hadoop allows for different streaming technologies—such as Spark Streaming, which can handle both batch processing (collects a group of transactions over time and produces

batch results) and real-time processing (continual input, process, and output of data over a shorter period).

3. **Machine learning:** Machine learning is a type of artificial intelligence that enables computers to learn in response to new data without being specifically programmed to do so. And while it can also be done outside of Hadoop, Hadoop offers important advantages in machine learning, particularly when it comes to large volumes of data and data real-time streaming. Technology within Hadoop enables you to get more directly to the predictive and prescriptive capabilities of machine learning—information that helps identify what your patients need ahead of time and get those interventions into the workflow.

MEETING AND OVERCOMING THE CHALLENGES TO ADOPTION

According to a 2015 Gartner survey on the challenges of Hadoop adoption, personnel (finding people with the right skillset) and determining how to get value from Hadoop were leading concerns. Building on Gartner's information, we've broken down adoption challenges into four areas:

1. Organizational
2. Buying
3. Administering
4. Using

1. Organizational Challenges

When it comes to adopting new technology, we often see two main camps: One will gravitate towards the "shiny new thing" (in this case, Hadoop and big data), while the other is "stuck in the mud" and reluctant to veer from established technologies. Both camps present unique challenges: Those excited by Hadoop's newness and promise may be easy to get on board, but enthusiasm itself doesn't guarantee success; that excitement needs to tie into business value if Hadoop is going to be successful. The less eager group will be used to the technology they've been using; if it works and is bringing value, they'll be tougher to convince to move to Hadoop.

There isn't a simple answer to these organizational challenges. Your best strategy may be to acknowledge these mindsets in your workforce and take time learning where your team members land on the spectrum. This way, you'll understand more about your challenges and be better prepared to navigate them—both by getting people on board and keeping them focused on value.

2. Buying Challenges

The challenge associated with investing in Hadoop is determining how (and if) you'll get value from it. Your organization will be more likely to put resources toward Hadoop with a clearly mapped out explanation of value.

Investments in healthcare IT and EMR conversions to new systems aren't guaranteed to succeed (to return value and serve their intended purpose). There are documented cases, for example, of costly EMR conversions that haven't delivered value in the treatment setting (which means there's also no business value).

Ensure that your organization is set up for Hadoop success a strategy for understanding and realizing value. The Cloud offers a great way to start experimenting with Hadoop and understanding its business value before you make a large investment. With pay-per-use tools (such Google Compute Engine, Amazon Web Services, and Windows Azure), you can start learning how Hadoop will benefit your organization without having to buy a large Hadoop cluster (including multiple servers and a lot of RAM). With these Cloud tools, you can pay as you use them to determine Hadoop's value without spending thousands of dollars on Hadoop infrastructure before you know if it's worthwhile.

3. Administration Challenges

We sum up the administration challenges of Hadoop in five issues:

1. Fewer experienced people: Whereas you can find people with a wealth of experience in other systems (Oracle, Teradata, SQL), there are just fewer people currently available with expertise in Hadoop; and those with the knowledge tend to be expensive.

Invest in your people. This includes building a learning culture (as opposed to one-off training), as you will always need to be learning with big data and Hadoop. Some large-scale online courses provide opportunities learn piece by piece and to relearn—making learning part of the culture. These courses include Coursera, Udacity, Pluralsight, and EDX.

2. Lack of best practices: We haven't yet established best practices for Hadoop the way we have for the systems mentioned above. There's rarely a set of practices and procedures defined for us, so we're creating and modifying them as we go along.

In keeping the culture of learning we discuss above, best practices in Hadoop will be part of the learning process. Your workforce is not going to learn Hadoop or optimal ways to use it just once. This area and technology is going to be evolving for the foreseeable future, so we'll be continuously finding our way.

3. Myriad of tools: It's not simple. Many tools are at work in Hadoop, all requiring operational knowledge and management.

Packaged solutions can ease some of the challenges of administering Hadoop. These include Hortonworks, Cloudera, and MAPR. They provide a much better assembly and implementation experience than downloading a system and putting it together outside of a package. A packaged solution puts all the tools together for you, so you know everything is compatible and will run with the same technology.

4. Open Source: The tools are open source, but require assembly—adding another layer to the process.

The packaged solutions described directly above will also help with the challenges of open source tools (namely, assembly).

5. Security: You're putting an enormous amount of data into this system—security and segmentation are naturally a concern.

Security will likely always be somewhat of a concern, but Cloud vendors are doing an increasingly better job about getting certified and standardizing practices

4. Use Challenges

The basic tools of Hadoop have presented their own using challenges due to the variety of lesser-known programming languages they've employed. This issue isn't unique to healthcare—it also affects the broader data market. Developers have had to know Scala, Java, or Python to work in Hadoop, whereas SQL is a much more widely known programming language.

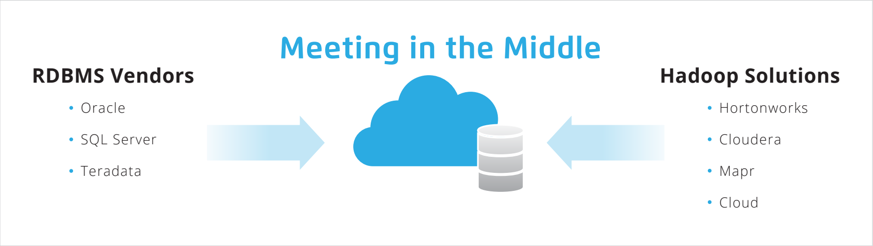
In response, the IT industry has invested heavily in SQL on Hadoop with a goal to get more users in the Hadoop ecosystem. You now have several options from which to choose (the next challenge, consequently, will be choosing a programming framework). There are four significant options for SQL on Hadoop:

1. Hive
2. Impala
3. Spark SQL
4. Apache Drill

IMPLEMENTING HADOOP: MEET IN THE MIDDLE

Instead of a rip-and-replace approach to implementing Hadoop (one where you completely replace existing systems with Hadoop), you may be better served with a convergence approach. This way, you meet in the middle between existing tools and what you're introducing with Hadoop.

You'll find value with Hadoop and big data with the types of work for which they're suited, but you may still find use for established RDBMS for certain workloads. The graphic below shows how these two types of systems can work together—or converge.



The middle ("convergence") is your EDW environment. This is where you run programming languages, including SQL, Spark, Hive, R, Python. These will also need to run in your analytics environment at some point. Your source marts may be in Hadoop, HDFS, or relational. There's an integrated layer where the Hadoop and your relational system and your analytics engine work together. So, it's an additive approach, where your traditional EDW and Hadoop can work together.

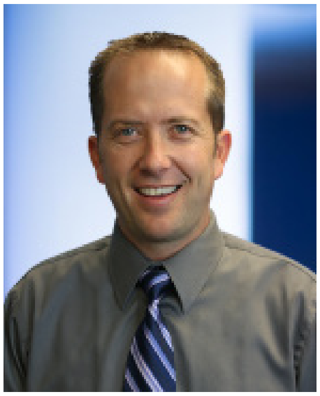
FOUR WAYS TO MAKE THE MOST OF HADOOP IN YOUR BIG DATA STRATEGY

As we've discussed throughout this report, Hadoop is loaded with capability as part of a big data strategy. You'll determine the framework's real potential, however, by how you deploy it. Keep in mind these four approaches as you introduce you Hadoop into your data operations:

1. Let use cases determine the need to implement Hadoop. (Be pragmatic.)
2. Think additive.
3. Invest in people now.
4. In general, The Cloud will give you the most flexibility in deploying Hadoop.

BIG DATA IS COMING TO HEALTHCARE—BE READY WITH HADOOP STRATEGY

We know that demands on healthcare data technology are growing, and will continue to do so for the foreseeable future. Earlier in this report, we referenced Moore's Law and how it helps forecast monumental growth in healthcare data. Our current data strategies won't be able to keep up with this expansion and will fail to turn information into valuable insights and informed medical decisions. In response, we're looking to the agility, efficiency, and scope of Hadoop to prepare for big data and fully leverage its insights to improve patient care and reduce costs.



Sean Stohl
Senior Vice President, Product Development

Sean Stohl started with Health Catalyst in 2012. He oversees the Data Acquisition Services Team in Product Development that works with Health Catalyst's clients to bring their source system data into their Enterprise Data Warehouse. Sean also works on the Health Catalyst Analytics Platform and enhancing it to bring in new sources of data into the EDW. Prior to joining Health Catalyst, Sean worked at Goldman Sachs in the Private Wealth Management Technology group and Intel Corporation. Sean holds a MS in Information Systems Management and a BS in Business Management from Brigham Young University.

Bryan Hinton
Senior Vice President, Software Engineering

Bryan joined Health Catalyst in February 2012. Prior to joining the Catalyst team, Bryan spent 6 years with Intel and 4 years with the LDS Church. At both places he spent a lot of time working with data and has always loved seeing what stories the data of an organization has to tell. At Intel he was on teams responsible for Intel's factory reporting systems and equipment maintenance prediction. At the LDS Church he led the .NET Development Center of Excellence and was responsible for the Application Lifecycle Management (ALM) processes and tools used for development at the Church.

Bryan graduated from BYU in Computer Science and loves all things BYU. He and his wife, Noel, have three children and live in Riverton, Utah. They love spending time together, watching sports, reading, and being outside.

HealthCatalyst
ignite change

ABOUT HEALTH CATALYST

Health Catalyst is a mission-driven data warehousing, analytics, and outcomes improvement company that helps healthcare organizations of all sizes perform the clinical, financial, and operational reporting and analysis needed for population health and accountable care. Our proven enterprise data warehouse (EDW) and analytics platform helps improve quality, add efficiency and lower costs in support of more than 50 million patients for organizations ranging from the largest US health system to forward-thinking physician practices.

For more information, visit www.healthcatalyst.com, and follow us on Twitter, LinkedIn, and Facebook.

3165 East Millrock Drive, Suite 400
Salt Lake City, Utah 84121
ph. 800-309-6800