MapR Guide to Big Data in Healthcare

Data Convergence in Healthcare

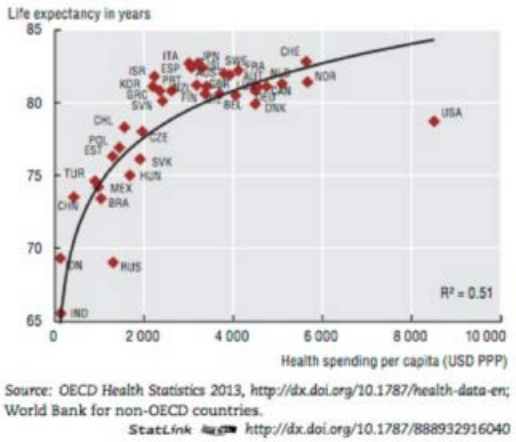
Data Convergence In Healthcare

Introduction

The healthcare industry, perhaps more than any other, is on the brink of a major transformation through the use of advanced analytics and big data technologies. Healthcare costs are driving the demand for big data-driven healthcare applications, especially in the United States. U.S. healthcare spending, which topped $3.2 trillion in 2015, has outpaced GDP growth for the past several decades and exceeds spending in any other developed country. It will grow by an estimated 15% in 2016. Yet technology decision makers in healthcare systems throughout the world cannot ignore the increased efficiencies, the attractive economics, and the rapid pace of innovation that can now be applied to delivering and paying for healthcare.

Many are finding that new standards and incentives for the digitizing and sharing of healthcare data—along with improvements and decreasing costs in storage and parallel processing on commodity hardware—are causing a big data revolution in healthcare with the goal of better care at lower cost. By one conservative estimate, applying big data analytics on a system-wide basis could reduce healthcare spending in the US by a whopping $300-450 billion annually.

Life Expectancy vs. Healthcare Spending



Source: OECD Health Statistics 2013, http://dx.doi.org/10.1787/health-data-en; World Bank for non-OECD countries.

StatLink http://dx.doi.org/10.1787/888932916040

But with these new innovations and potential for change come some significant challenges that must be addressed. There are reasons why healthcare has been slower to adopt advanced analytics than other industries. And hospitals have also been very slow to embrace standards for seamlessly sharing data, as have verticals such as manufacturing, retail, and financial services. The sheer growth of data volumes in healthcare along with the speed in which that data arrives, and the veracity and relevance of that data all present unique challenges to be overcome. Many organizations are finding that traditional relational database technologies can neither handle this new volume of data, nor the unstructured nature of new data—including images, documents, and telematics. New technologies and new thinking are needed, and both are now at hand.

Big Data Trends in Healthcare

There is a move toward evidence-based medicine, which involves making use of all clinical data available and factoring that into clinical and advanced analytics. Capturing and bringing all of the information about a patient together gives a more complete view for insight into care coordination and outcomes-based reimbursement, population health management, and patient engagement and outreach. Gaining this 360-degree view of the patient can also eliminate redundant, expensive testing; reduce errors in administering and prescribing drugs, and even avoid preventable deaths.

Also, it is certainly noteworthy that in today's healthcare environment, a clear majority of the data generated and therefore available for use—75% or more of the data by some estimates— is unstructured data. It emerges from sources like the rapidly-growing number of digital devices and sensors; emails; doctors' and nurses' notes; laboratory tests, and third party sources outside the hospital. It is the unstructured nature of this data along with the sheer enormity of the volumes generated that make healthcare data a perfect match for the MapR Converged Data Platform.The MapR Platform can acquire and store enormous masses of structured and unstructured data of any type, running on powerful, cost-effective hardware. Then with the overlay of advanced big data analytics, healthcare providers and executives can make great leaps ahead in terms of improving patient outcomes while lowering the costs of doing so.

Value-Based, Patient-Centric Care

A goal of modern healthcare systems is to provide optimal health care through the meaningful use of health information technology in order to:

- Improve healthcare quality and coordination so that outcomes are consistent with current professional knowledge
- Reduce healthcare costs, reduce avoidable overuse
- Provide support for reformed payment structures

Health payors such as insurers and public health systems (e.g., Medicare and Medicaid) are in the early stages of shifting from fee-for-service compensation to value-based data-driven incentives that reward high quality, cost-effective patient care and demonstrate meaningful use of electronic health records. This approach requires significant improvements in reporting, claims processing, data management, and process automation.

The focus on value-based care corresponds with an increased focus on patient centric care. By leveraging technology and focusing healthcare processes on patient outcomes, a continuum of care, doctors, hospitals, and health insurance need to work with each other to personalize care that is efficient and price conscious, transparent in its delivery and billing, and that is measured based on patient satisfaction.

Thus the goal now is to begin to move more decisively away from the long-standing fee for service practice by which payments are made to providers. In essence, providers get paid for seeing and treating patients. Currently there is little or no reward when and if providers improve quality of services, boost patient outcomes, or reduce costs. Thus fee for service has been a major roadblock in plans or desires to invest in digital solutions to, say, improve patient outcomes if the providers cannot recoup their investments. As one senior executive at KPMG put it, "Instead of rewarding leaders for transforming healthcare, our systems reward leaders for making narrow improvements within them."

Current thinking around longstanding, crucial payment practices is beginning to change, paving the way for a robust digital transformation of healthcare.

The Healthcare Internet of Things (IoT)

Also called the Industrial Internet, these terms refer to the rapidly increasing number of smart, interconnected devices and sensors and the tidal volumes of data they will generate and move between devices, and ultimately to people. Spending on healthcare IoT could top $120 billion in just four years, by some estimates. And most of the data created by the healthcare IoT is of the unstructured variety, again creating a major role for Hadoop and advanced big data analytics working within the Hadoop framework.

Today, a variety of devices monitor every sort of patient behavior, from glucose monitors to fetal monitors to electrocardiograms to blood pressure. Many of these measurements require a follow-up visit with a physician. But smarter monitoring devices communicating with other patient devices could greatly refine this process, possibly lessening the needs for direct physician intervention and maybe replacing it with a phone call from a nurse. Other smart devices already in place can detect if medicines are being taken regularly at home from smart dispensers. If not, they can initiate a call or other contact from providers to get patients properly medicated. The possibilities offered by the healthcare IoT to lower costs and improve patient care are almost limitless.

Reducing Fraud Waste and Abuse

The cost of fraud, waste, and abuse in the healthcare industry is a key contributor to spiraling healthcare costs in the United States, but big data analytics can be a game changer for health care fraud. The Centers for Medicare and Medicaid Services prevented more than $210.7 million in healthcare fraud in one year using predictive analytics. UnitedHealthcare transitioned to a predictive modeling environment based on a Hadoop big data platform, in order to identify inaccurate claims in a systematic, repeatable way and generated a 2200% return on their big data/advanced technology.

The key to identifying fraud is the ability to store and go back in history to analyze large unstructured datasets of historical claims and to use machine-learning algorithms to detect anomalies and patterns.

Healthcare organizations can analyze patient records and billing to detect anomalies such as a hospital's overutilization of services in short time periods, patients receiving healthcare services from different hospitals in different locations simultaneously, or identical prescriptions for the same patient filled in multiple locations.

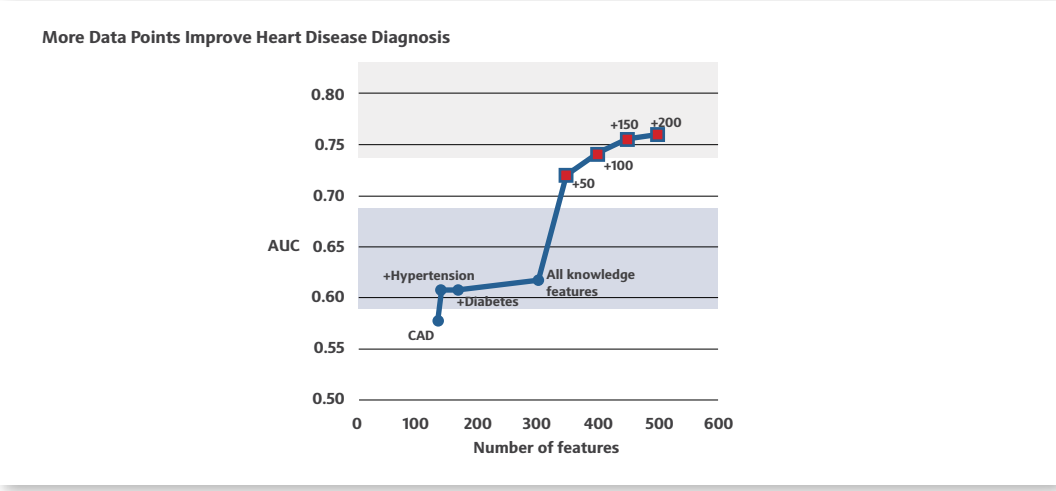
One major healthcare provider leveraged a data lake approach as it aggregated massive volumes of data as a data hub for various departments, including fraud prevention. As a result, the provider is on the way to capturing an incremental 20% of fraud, waste, and abuse in its claims department.

The Centers for Medicare and Medicaid Services uses predictive analytics to assign risk scores to specific claims and providers, to identify billing patterns, and claim aberrancies difficult to detect by previous methods. Rules-based models flag certain charges automatically. Anomaly models raise suspicion based on factors that seem improbable. Predictive models compare charges against a fraud profile and raise suspicion. Graph models raise suspicion based on the relations of a provider; fraudulent billers are often organized as tight networks.

Predictive Analytics to Improve Outcomes

Initiatives such as meaningful use are accelerating the adoption of Electronic Health Records (EHR) and the volume and detail of patient information is growing rapidly. The surge in the creation and broadening use of EHR was driven in part by a $30 billion federal government stimulus provided by the Health Information Technology for Economic and Clinical Health (HITECH) Act. The Act was designed specifically to provide incentives to adopt EHR and then encourage the sharing of patient information by clinicians everywhere in an attempt to lower costs, speed diagnosis, and improve patient outcomes. Being able to combine and analyze a variety of structured and unstructured data across multiple data sources, aids in the accuracy of diagnosing patient conditions, matching treatments with outcomes, and predicting patients at risk for disease or readmission.

Predictive modeling over data derived from EHRs is being used for early diagnosis and is reducing mortality rates from problems such as congestive heart failure and sepsis. Congestive Heart Failure (CHF) accounts for the most healthcare spending. The earlier it is diagnosed the better it can be treated avoiding expensive complications, but early manifestations can be easily missed by physicians. A machine learning example from Georgia Tech demonstrated that machine learning algorithms could look at many more factors in patients' charts than doctors, and by adding additional features, there was a substantial increase in the ability of the model to distinguish people who have CHF from people who don't.



Predictive modeling and machine learning on large sample sizes, with more patient data, can uncover nuances and patterns that couldn't be previously uncovered. Optum Labs has collected EHRs of over 30 million patients to create a database for predictive analytics tools that will help doctors make big data-informed decisions to improve patients' treatment.

Real-time Monitoring of Patients

Healthcare facilities are looking to provide more proactive care to their patients by constantly monitoring patient vital signs. The data from these various monitors can be analyzed in real time and send alerts to care providers so they know instantly about changes in a patient's condition. Processing real-time events with machine learning algorithms can provide physicians with insights to help them make lifesaving decisions and allow for effective interventions.

Wearable sensors and devices present the opportunity for caregivers to interact with patients in entirely new ways, making healthcare more convenient and persistent. Real-time monitoring changes the very nature of the relationship in that face-to-face care is not always a necessity. As an example, applications are being used for remote or in-home monitoring of patients with chronic obstructive pulmonary disease. Other monitors track the weight of patients battling obstructive heart disease to detect fluid retention before hospitalization is required. Still others track a child's asthma medication usage to be sure home caregivers and family members are aware of what needs to be administered, reducing visits to the ER. As is so often the case with new data volumes in healthcare, sensor data from wearable monitors is unstructured data that yields to the data acquisition and storage capabilities of Hadoop, as well as to the power and flexibility of advanced big data analytics.

Healthcare Data

Unstructured data forms about 80% of information in the healthcare industry and is growing exponentially. Getting access to this unstructured data—such as output from medical devices, doctor's notes, lab results, imaging reports, medical correspondence, clinical data, and financial data—is an invaluable resource for improving patient care and increasing efficiency.

Data Source	Description	Relevant to*
Claims	Claims are the documents providers submit to insurance companies to get paid. A key component of the Health Insurance Portability and Accountability Act (HIPAA) is the establishment of national standards for electronic healthcare transactions in order to improve efficiency by encouraging the widespread use of Electronic Document Interchange (EDI) between healthcare providers and insurance companies. Claim transactions include International Classification of Diseases (ICD) diagnostic codes, medications, dates, provider IDs, and the cost. Various efforts are underway to more accurately gather and assess claims data.	PA, PA, PR
EHR/EMR	Electronic Health/Medical Records data (EHR or EMR). Medicare and Medicaid EHR incentive programs were established to encourage professionals and hospitals to adopt and demonstrate meaningful use of certified EHR technology. EHRs facilitate a comprehensive sharing of data with other providers and medical applications. EHRs contain the data from the delivery of healthcare which includes diagnosis, treatment, prescriptions, lab tests, and radiology. Health Level Seven International (HL7) provides standards for the exchange, integration, sharing, and retrieval of electronic health record data. The federal government's HITECH act uses rewards and penalties to foster more meaningful use of electronic medical records.	PR, PA, PY
Genomics	Data is derived from DNA sequencing and bioinformatics to sequence and analyze the function and structure of a complete set of DNA within a single cell. Genomic data is increasingly found in the EHR/EMR where advanced big data analytics are leveraged to extract vital, unique patient information with significant predictive potential.	PA, PR, RS
IoT/Medical Devices	Medical device data comes from patient sensor data from the home or hospital. The potential of data analysis derived from the exploding number of these devices is significant when it comes to cost reduction and improved patient care.	PA, PR, RE
Healthcare and Pharma R&D	R&D comes from primary and secondary research from a wealth of sources. Data is generated from clinical trials data, university research, and patient sensors. Big data has the potential to revolutionize this sector.	PR, RS
Patient Behavior and Sentiment Data	This data comes almost exclusively as unstructured data from the broad social web (Facebook, Twitter, LinkedIn, etc.). This data can reveal what patients really think about various aspects of their healthcare. Efforts around analysis of patient behavior can lead to predictions of, say, the number of days a patient will be hospitalized in the year ahead.	PY, PR, PH
Public Data sources	http://www.healthdata.gov/ - Makes data available to just about anyone that can make good use of it https://www.nlm.nih.gov/hsrinfo/datasites.html - Offers selective links to high-interest healthcare information and data https://www.data.gov/health/ - Datasets, tools, and applications related to healthcare http://www.ahrq.gov/research/data/dataresources/index.html - Online searchable databases about trending healthcare topics	PA, RS, PH, RE

* **Key:** PA=Patient, PY=Payer, PR=Provider, PH=Pharma, RE=Regulators, RS=Researchers

Healthcare Industry Stakeholders

In healthcare, the discussion of stakeholders is centered around their vision of quality. For example, caregivers and healthcare providers think of quality in terms of offering the very best service, leveraging the most accurate, state-of-the-art testing and treatment protocols. Payers, on the other hand, want these providers to follow a prescribed, evidence-based diagnostic and treatment plan characterized by the fewest visits and tests. In fact, conflicts such as these can be mitigated at times by broader use of telehealth and telemedicine—so called virtual care programs—which are increasingly proving not only financially viable but also capable of boosting patient outcomes. Below is a quick look at the various stakeholders in healthcare and how big data solutions are impacting them now and in the near future.

Providers, including hospitals, physicians, clinics, etc. They want big data to provide them with as comprehensive a 360-degree view of the patient as they can possibly get. From a diagnostic perspective, they will benefit greatly from genetic data, not only from the patient but also from larger aggregate populations of patients with the same disease and presentation of symptoms. The decision making of physicians is increasingly being influenced by big data.

Patients. This key stakeholder group is leveraging digital technologies to more effectively partner with providers, select caregivers, and monitor and improve their own health. One study found that use of technology to measure fitness and health improvement goals is soaring, particularly among millennials. Almost all segments of the patient population are more aggressively searching online for healthcare information while using patient portals and physician performance scorecards. And using a smartphone, patients can snap a photo of a rash or bug bite to get a virtual diagnosis and treatment plan without any costly face-to-face office visits.

Payers, including insurance companies, government, employers, etc. Big data is playing a fast-growing role with these stakeholders both in fraud detection and expedited claims processing. The fraud potential is obviously sky high when billions of dollars pass through multiple groups, departments and individuals. Big data analytics has proven adept at analyzing nuances and anomalies, such as a patient's receiving services from multiple providers simultaneously or filling the same prescription at different places. To expedite claims, Hadoop works with Hive, Pig, and HBase to migrate claims workloads off creaky legacy systems to speed up claims processing and allow for deeper inspection and analysis of claims data.

Pharma. Developing ultra sophisticated drugs to treat highly complex diseases is one if not the most data-intensive effort. For example, data from genomic sequencing from large patient populations, key to some drug development, simply could not be undertaken on conventional database frameworks. Advanced big data analytics on Hadoop is the only viable alternative.

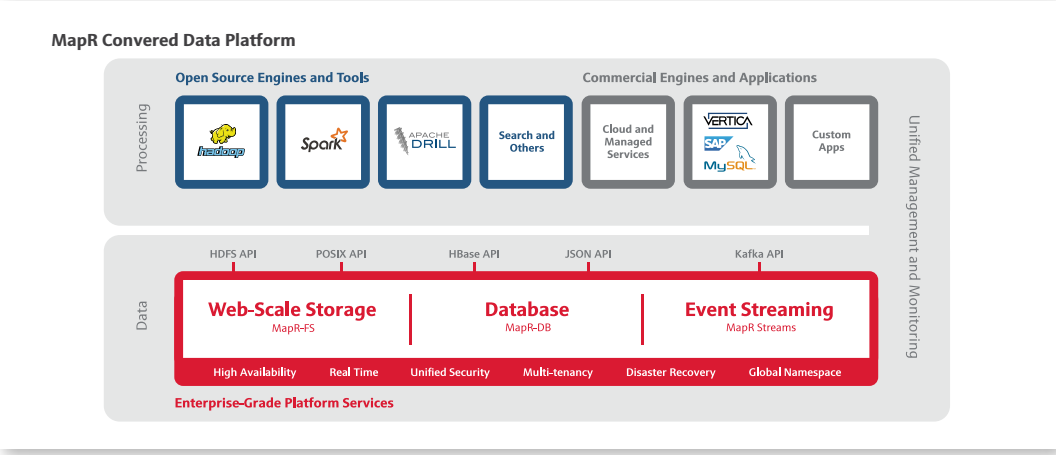
Researchers and universities. Not unlike pharma, these stakeholders utilize enormous amounts of data, often from all the stakeholders above. Properly organizing and then analyzing it is a job for big data analytics. Often not heavily funded, these stakeholders appreciate that big data analytics leverages commodity hardware to get the job done.

"With MapR I can spend less on outsourced resources and instead spend money on adding new features, analytics or visualization capabilities or acquiring new types of data. We can do things that truly matter to our customers."

Dan Blake
Chief Technology Officer
Valence Health

The MapR Converged Data Platform in Healthcare

By pursuing our data-centric vision for a new generation of applications, MapR has created an applications platform that converges the management of data of any size, speed, and format. It was for this work that we were recently awarded a patent (US9,207,930). This is the MapR Converged Data Platform.



Open Source Innovation on a Trusted Platform

The MapR Converged Data Platform is designed to deliver utility-grade data services and commercially supported open source innovations to development teams, IT operations, business analysts, and data scientists. Open source technology provides a fantastic creative force when looking to tackle the sophisticated new challenges that big data—and especially new data—can uncover.

Without a converged data platform, critical information can get stuck in "data silos" and an inefficient use of hardware resources can result in a costly "cluster sprawl" of under-utilized servers and storage. With the MapR Platform, businesses can enjoy real-time insights based on secure, protected, high-fidelity data.

Seamless Integration with Existing Enterprise Systems

One of the most profound design decisions made by MapR was to create an enterprise-grade file and storage system to house the data of the Hadoop ecosystem. The MapR File System, based on the trusted POSIX/NFS standard, makes it vastly easier to get data in and out of the MapR Platform using familiar enterprise tools. MapR also provides developer programmatic access to data with standard interfaces like SQL, HDFS, HBase, JSON, Kafka, and more.

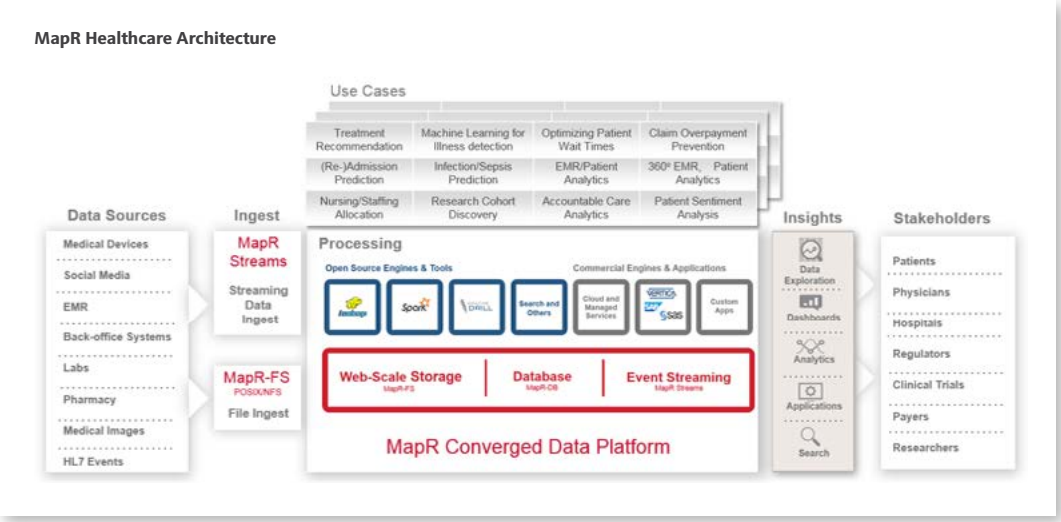
Continuous, Trusted Operations

With our consistent focus on the integrity of data, MapR has created a hardened, clustered platform that can withstand multiple hardware failures, data center outages, and malicious attacks and intrusions from cybercriminals. Many proven methods of data protection—such as failover, redundancy, and access controls—are built into the MapR Platform.

Big Data with Enterprise Stability

Game-changing big data applications and analytics will continue to rely on open-source software. As a company founded in and contributing to the open-source world of Hadoop and Spark, MapR continues to define enterprise requirements and best practices for successfully using the latest open source innovations. We deliver monthly updates to open source software packages to ensure you have the latest innovations.

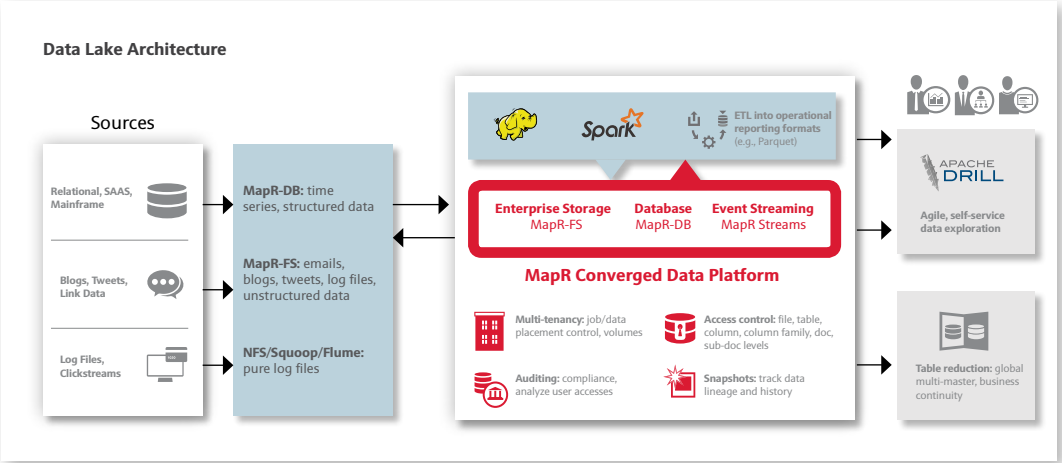
Healthcare Application Architecture



Production Examples

Valence Health Improving Outcomes and Reimbursements

Valence Health is using the MapR Converged Data Platform to build a data lake that is the company's main data repository. Valence consumes 3,000 inbound data feeds, with 45 different types of data, daily. These include critical data such as lab test results, patient health records, prescriptions, immunizations, pharmacy benefits, claims and payments, and claims from doctors and hospitals, and are used to inform decisions about improving both healthcare outcomes and reimbursement. The company's rapid client growth and the associated increasing volumes of data were straining its existing technology infrastructure.



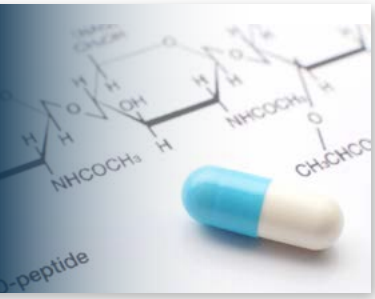
Prior to their MapR solution, if they received a feed with 20 million lab records, it would take 22 hours to process that data. MapR cut that cycle time down from 22 hours to 20 minutes, running on much less hardware. Valence Health is also now able to accommodate customer requests that were very difficult to address in the past. For example, a customer might call and say, "I sent you an incorrect file three months ago and I need you to take that file out." Their traditional database solution might take 3-4 weeks to get that data deleted. MapR snapshots provide point-in-time recovery that enables Valence to just roll back and remove that file in minutes.

UnitedHealthcare on Fraud, Waste, and Abuse

UnitedHealthcare provides health benefits and services to nearly 51 million people. The company contracts with more than 850,000 physicians and care professionals and approximately 6,100 hospitals nationwide. Their Payment Integrity group has the tough job of ensuring that claims are paid correctly and on time. Their previous approach to managing more than one million claims every day (10 TB of data daily) was ad hoc, heavily rule-based and limited by data silos and a fragmented data environment. UnitedHealthcare came up with a unique dual model strategy, which meant focusing on operationalizing savings, while at the same time pursuing innovation to constantly leverage the latest technologies.

Here's how they are doing it: in terms of operationalizing savings, the group is building a predictive analytics "factory" where they can identify inaccurate claims in a systematic, repeatable way. Hadoop is now the data framework for a single platform that's equipped with tools to analyze a slew of information from claims, prescriptions, plan participants, contracted care providers, and associated claim review outcomes.

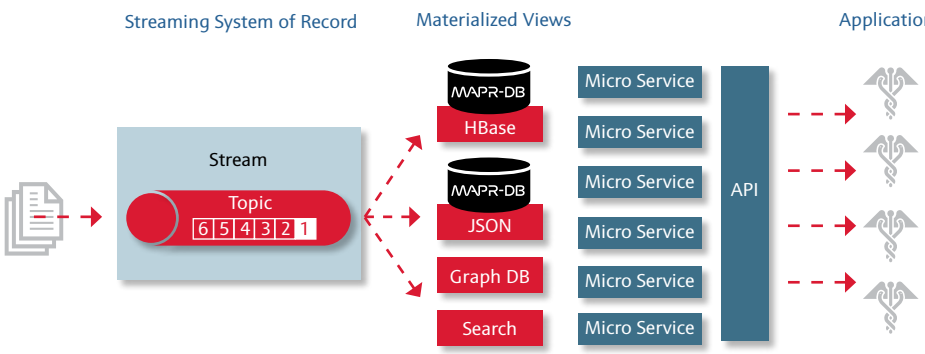
They integrated all this data from multiple data silos across the business, including over 36 data assets. And they now have multiple predictive models (PCR, True Fraud, Ayasdi, etc.) at their fingertips that provide a rank-ordered list of potentially fraudulent providers they can pursue in a targeted, systematic way.



Liaison Technologies Streaming System of Record for Healthcare

Liaison Technologies provides cloud-based solutions to help organizations integrate, manage, and secure data across the enterprise. One vertical solution they provide is for the healthcare and life sciences industry, which comes with two challenges—meeting HIPAA compliance requirements and the proliferation of data formats and representations. With MapR Streams, the data lineage portion of the compliance challenge is solved because the stream becomes a system of record by being an infinite, immutable log of each data change.To illustrate the latter challenge, a patient record may be consumed in different ways—a document representation, a graph representation, or search—by different users, such as pharmaceutical companies, hospitals, clinics, physicians, etc. By streaming data changes in real time to the MapR-DB, HBase, MapR-DB JSON document, graph, and search databases, users always have the most up-to-date view of data in the most appropriate format. Further, by implementing this service on the MapR Converged Data Platform, Liaison is able to secure all of the data components together, avoiding data and security silos that alternate solutions require.

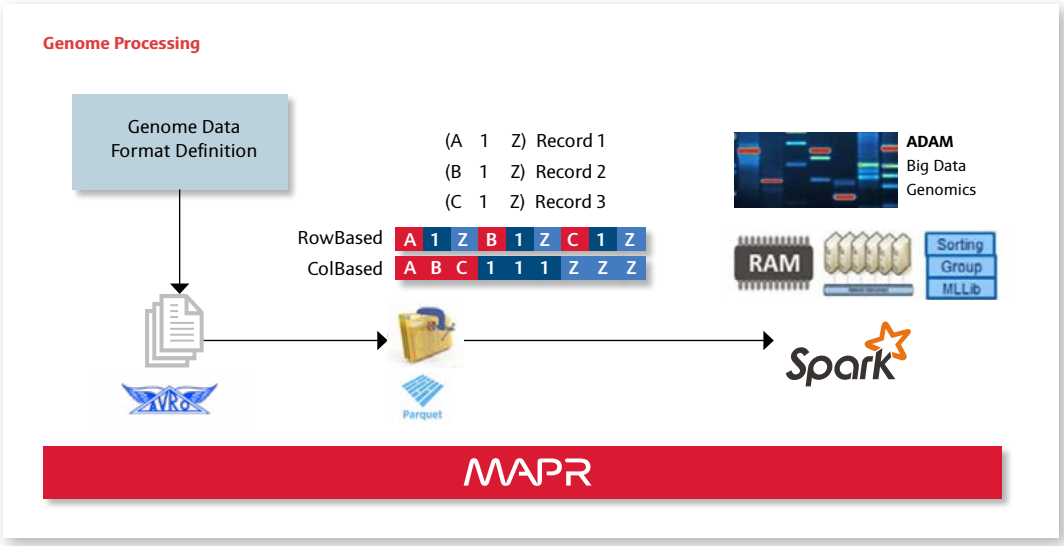
Streaming System of Record for Healthcare



Novartis Genomics

Next Generation Sequencing (NGS) is a classic big data application that deals with the dual challenge of vast amounts of raw heterogeneous data and the fact that best practices in NGS research are an actively moving target. Additionally, much of the cutting-edge research requires heavy interaction with diverse data from external organizations. It requires workflow tools that are robust enough to process vast amounts of raw NGS data yet flexible enough to keep up with quickly changing research techniques. It also requires a way to meaningfully integrate data from Novartis with data from these large external organizations—such as 1000 Genomes, NIH's GTEx (Genotype-Tissue Expression) and TCGA (The Cancer Genome Atlas)—paying particular attention to clinical, phenotypical, experimental, and other associated data.

The Novartis team chose Hadoop and Apache Spark to build a workflow system that allows them to integrate, process, and analyze diverse data for Next Generation Sequencing (NGS) research while being responsive to advances in the scientific literature.



Healthcare IoT Startup Working to Classify Heart Conditions Faster

The current heart rhythm analysis process is slow and classification is done manually. They do batch uploads from the devices into the analysis software machines to have medical analysts look at the classification data, and then submit a report to the doctors and hospital who then make medical decisions about the patients. The process takes over 24 hours, a long lag before doctors can access the patient data, increasing the risk of medical emergencies.

With MapR-FS, Telemed will now be able to ingest data from various medical devices directly via NFS into their cluster for real-time patient insight. This solution needed to be High Availability and also provide multi-tenancy (due to HIPAA) as they start hosting various hospital patient data and medical device company data. Being able to segment that data by their customers was really important.

With the help of MapR Professional Services, they have been able to build out a solution to hit their July 18th HIPAA review deadline, and provide an architecture that fits all the requirements in terms of HA, multi-tenancy, and provide real-time insights. The CEO has fulfilled his requirement and deadline to his investors and the company will be on track to start selling their SaaS solutions in Q3/Q4.

Conclusion

Improving patient outcomes at the same or even less cost is an extraordinarily tall order for any health-care provider, given overall costs of healthcare are rising in the US at a lofty 15% clip. Full- scale digital transformation is the key to making this goal a reality, with digitization, enhanced communications, and big data analytics being the legs to support the transformation effort. The many emerging use cases for big data analytics are intimately tied to the ability of Hadoop-based solutions to acquire and store massive quantities of disparate data—structured and unstructured—from just about any source and present it for in-depth analysis.

In selecting a big data platform and in particular a Hadoop distribution, be sure the platform is highly adept at handling the mix of data types in healthcare typically housed in silos, with clinical data in one silo; pharmaceutical data in another, and logistics information on hospital supplies in yet another. This platform should be flexible enough so that caregivers can use complex data like doctors' notes and imaging files for real patient analysis, not just for archiving.

Further Information

Solution Brief: Big Data and Apache Hadoop for the Healthcare Industry
Whitepaper: Next Generation Genome Sequencing
Video: Data Science for the Healthcare Industry with MapR Data Scientist Joe Blue
Blog: How Stream-First Architecture Patterns Are Revolutionizing Healthcare Platforms
Blog: How Big Data is Reducing Costs and Improving Outcomes in Health Care
Blog: Hadoop in Action: Using Hadoop to Detect Fraud, Waste and Abuse in Healthcare
Blog: Big Data and Genomics: At the Crossroad