Section 2: Week 4: Theory to Practice: Health Care

Nate Bachmeier

TIM-7020: Database and Business Intelligence

December 29th, 2019

North Central University

Data Management in Healthcare

The global growth of data has increased from 4.4 to 44ZB between 2013 to 2020 and continues to double every two years (Mansouri, Nadjaran, & Buyya, 2017). This new era of Big Data impacts all industries, as they need to evolve data management strategies to align with the high volume, velocity, and variety of their data sources. The health care industry, in particular, stands to benefit significantly from this transformation, yet they remain one of the laggards due to regulation and compliance limitations. These limitations impact multiple participants in the health care ecosystem, ranging from patients to doctors and insurers. Until removing these constraints, the health care industry will be slow to innovate and encounter higher operational and capital expenditures (OPX/CAPX) relative to their peer industries.

# What is the current state of Health Care

Consider all of the different data sources that the health care system needs to consume and correlate today. These sources include low-resolution manual entry sources, such as schedules and visitation notes, along with high-resolution images, such as magnetic resonance imaging and ultra-sound video. Then factor the secondary layers that exist to support these systems, like insurers and pharmaceutical companies. Outside of the medical facility, IoT devices, such as personal fitness trackers and outpatient care monitors, stream continuous feeds of high volume unstructured data to disparent third-party servers. Each of these siloed systems could benefit from having insight into their peers. For instance, researchers cannot always access sufficient data to prove or disprove a treatment works. Doctors do not always have the full picture of their patients, leading to misinformed decisions. Patients need to manage multiple online profiles and cannot easily access a ‘single pane of glass’ that encompasses the entire state of their health.

The health care industry contains three distinct communication flows across primary care, secondary care, and administration (Choo, 2016). Choo differentiates them as a primary care provider is notified of any new medical record versus secondary are specialists that receive a subset of information. Administrators need to orchestrate any billing and scheduling of the patients. Each of these roles has been transitioning away from paper-based health records, though many offices reside in rural areas that might lack technical expertise. Many are looking towards cloud-integrated solutions to get over this hurdle (Singh, Kumar, & Khatri, 2019), though specific business cultures are slower to adopt change (Marbury, 2019). According to Marbury, provider adoption of cloud-technologies is ad-hoc and lacks a coherent story. For instance, a medical facility might use the cloud to store database backups and host the patient scheduling portal but leave all other aspects on their on-premise infrastructure.

# Why do these challenges exist

Regulation and compliance requirements force many health care systems to remain in private data centers and out of the public cloud. Applications within a private data center are less agile than those in the public cloud, precisely because of a lacking of instantaneous provisioning of elastic resources.

## The criticality of OnDemand resources

For example, connecting two private data centers requires buying proprietary Virtual Private Network Appliances (VPN) and secondary hardware to support the scenario. Purchasing and installing these components could take several months, compared to the public cloud, which can provision a VPN in a matter of minutes. Similar experiences exist in data platform tooling, like though provided by Microsoft Azure and Amazon Web Service. These modern solutions can glue together multiple built-for-purpose data management technologies that address specialized scenarios. Consider the distinction between a surgeon performing open-heart surgery and machine learning practitioner seeking disease correlations. The prior is substantially more time-critical and could benefit in-memory databases versus the later where slower and cheaper batch processing stores are acceptable. However, there are scenarios where the inverse is also true, such as the practitioner needs interactive business intelligence. In these scenarios, the public cloud allows for costly high-performant stores to become hydrated economically for a matter of hours, then released when no longer needed. Private data centers also lack access to cloud-native solutions such as Storage as a Service (StaaS), Machine Learning as a Service (MLaaS), and other Software as a Service (SaaS) offerings. These technologies remove aspects of the learning curve, accelerate innovation by removing boiler-plate efforts, and allow the engineering teams to focus on the core competencies of the specific business goals.

## Influence of Regulation

A pivotal hindrance to adopting modernized platforms comes from the legal requirements of health care professionals. The Health Insurance Portability and Accountability Act (HIPAA) the Genetic Information Nondiscrimination Act of 2008, Health Information Technology for Economic and Clinical Health Act (HITECH), and similar laws seek to protect the patients from discrimination and secure their information against negligence (Hofstra University, 2019) (HSG, 2018). These health care administrators deliver these requirements through a combination of encryption and network segmentation strategies. However, these objectives are hard to accomplish in practice due to the many-to-many relationship between Electronic Health Records (EHR) and decentralized consumers of the data (Virtru, 2019).

## The distrust of public cloud

Celebgate (Owen, 2019), misconfigured storage resources (Robinson, 2018), and other media sensations, cause the public to distrust of cloud security, despite continuous and substantial investments from primary public cloud providers into cybersecurity (Patterson, 2018) (Business Insider, 2017). Few, if any, health care providers have security budgets that rival the scale of AWS or Azure. Both service providers support the notion of a virtual private cloud (VPC) that operates as a resource island with multiple levels of physical security (e.g., single-tenant hosts, geo-fencing, and offloaded management services) and virtual security (e.g., AI-based monitoring, custom hypervisors, and software-defined networking). Though irrespective of the quality, regulatory requirements place artificial constraints on where health care data can reside. Azure Stack and AWS Outpost, preconfigured network rack solutions, are attempting to address these limitations by bringing public cloud capabilities into the private data centers. These racks address many core scenarios but lack MLaaS and similar secondary systems.

# Who does this impact

Health care data resides in private silos that cannot be easily correlated, which introduces challenges for multiple parties.

## Patients

Users prefer a single and familiar portal to access all of their data and be able to ‘slice and dice it’ however they see fit. Today instead, patients need to extrapolate a snapshot of their health state by cross-reference decoupled web portals owned by health insurance providers, medical facilities (e.g., hospitals), medical encyclopedias (e.g., WebMD), personal fitness monitors, and third-party laboratories. The decentralized nature is overly complicated, and the inconsistent experience leads to knowledge gaps.

## Medical Facilities

Hospitals and similar establishments are required to perform services upfront and collect payment afterward. Their optimal scenario requires billing the patient insurance first and then the insurance provider charging the patient second. However, a double-blind situation can exist that causes charges to become stuck in a limbo state for months to years. Annodiately a full reconciliation for my child’s birth took 18 months. During this time, the facility cannot commit those funds to new projects, collect interest, and is encounters non-payment risk. A more open and transparent data sharing portal could enable these conversations to happen in more real-time.

## Researchers

Many data science algorithms, such as neural networks, require enormous amounts of example data to train models efficiently. This lack of training data prohibits investigation into specific niche diseases and limits researchers' ability to save lives. Consider the story of Henrietta Lacks, whose HeLa cells became the defacto data set for cancer research (Skloot, 2010). Potentially dozens of similar stories exist today and are waiting for discovery, which will not happen until there is more open access to anonymized medical result files.