KYTHERA

DECIPHERING HEALTHCARE

Improving Data Quality Through the Medallion Architecture

Executive Summary

Kythera Labs implemented a Medallion Architecture to create specific, progressive improvements to large volumes of healthcare Real-World Data (RWD) as it moves through the Bronze, Silver, and Gold layers. We added an additional level, Platinum, with extended benefits for answering business questions faster.

Bronze: where we land data from all our data sources. Our data is sourced from various providers, each with its own format, delivery method, and relational model.

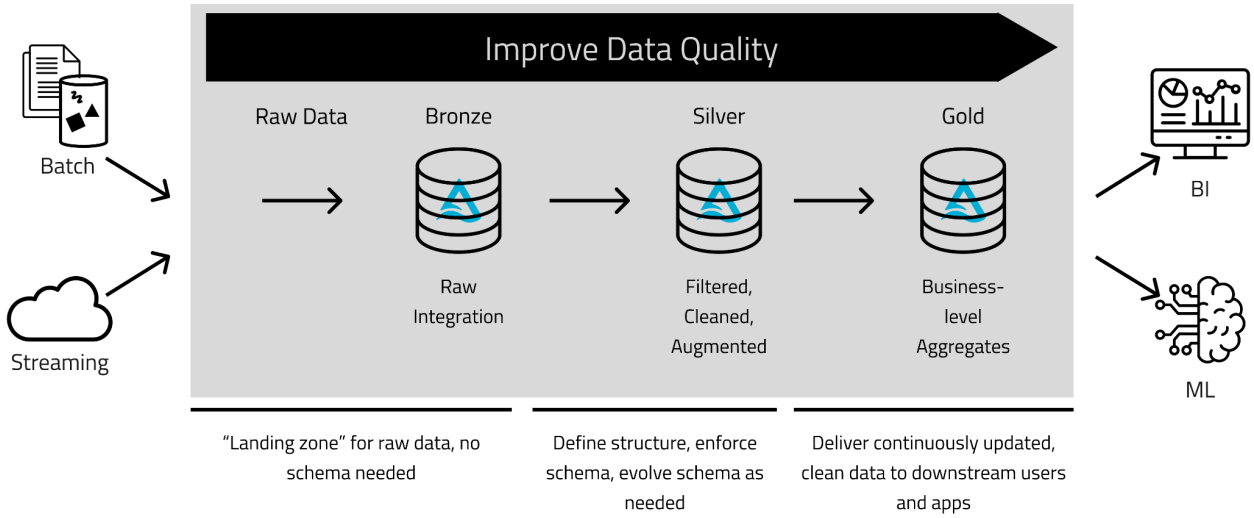
Silver: the ingested data goes through our automated "remastering" process, which includes matching, de-duping, cleaning, and format conforming, among others that provide a unified view of the data. At this stage, the data has been improved so the end user, whether an analyst, data engineer or data scientist, can use it confidently.

Gold: where data is organized in consumption-ready analytics for various projects such as customer analytics, product quality analytics, building patient cohorts, drug commercialization, market intelligence assessments, and more.

Platinum: Wayfinder is an environment to immediately access and analyze Kythera Labs' remastered and derived data assets using familiar SQL on a simple data lake model with Databricks' highly efficient Photon Serverless warehouses. Here Life Sciences and Healthcare Providers organizations can take advantage of our improvements in the Bronze, Silver, and Gold layers, including data cleaning, standardizing, de-identifying, uplifting, curating, and joining functionality to integrate data they supply and analyze robust, holistic, and unique datasets at scale.

The Medallion Architecture

Kythera Labs focuses on reducing uncertainty in healthcare data by empowering Healthcare and Life Sciences organizations to rapidly integrate, access, and analyze healthcare data with scale and speed. We work with big data (petabytes) from many sources and use data science to resolve quality issues inherent to healthcare data. In a nutshell, our process essentially begins with ingesting raw data, transforming it into usable formats (we call this "remastering") and ultimately developing use case-focused analytics for Life Sciences and Healthcare Provider organizations. It was critical for us to have a platform that could meet our unique data processing, storage, and sharing was necessary. We selected Databricks as the foundation for building our pipelines and data science platform and implemented a "Medallion" architecture. Databricks describes a Medallion architecture as a "data design pattern used to logically organize data in a Lake House with the goal of incrementally and progressively improving the structure and quality of the data as it flows through each layer of the architecture (from Bronze to Silver to Gold)."

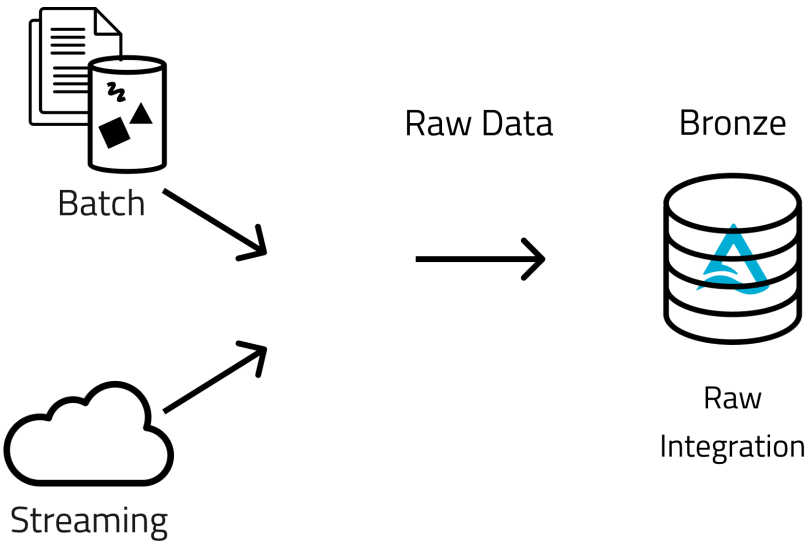


Since Kythera Labs is in the data quality business, taking an "improvement of data quality through the layers" was just the right approach for us. Our clients can quickly get answers to their business questions without having to scale the walls of data quality through the bronze, silver, and gold layers. And with Wayfinder, our white-labeled implementation of Databricks, we can leverage Delta Sharing to make the Silver and Gold layers available to our clients without copying large volumes of data.

Bronze Layer

Raw Data

The Bronze layer is where we land data from all sources. Our data comes from various sources, each with its own format, delivery method, and relational model. We are sourcing medical claims data, Electronic Health Records (EHR) data, pharmacy data, consumer data, and along with other data types. In the Bronze layer, there is no change to any of the values or structure of the data. However, since the data is so "big," the vendors usually deliver it in a series of "increments," basically many compressed CSV or parquet files. In this state, these files cannot be looked at all at once (they are not "spannable" in big data terms), so we combine all the similar files into a single Delta table containing all the data rows. And the data keeps on streaming in daily, so this process needs to be resilient and highly stable.

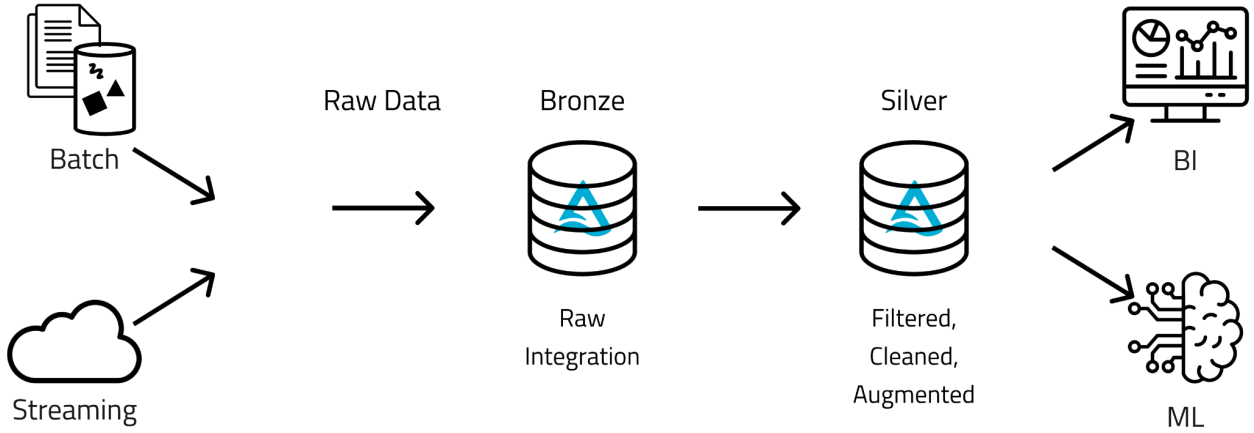


Silver Layer

In the Silver layer, the ingested data goes through our "remastering" process, which includes matching, de-duping, cleaning, format conforming, and other automated processes that provide a unified view of the data. At this stage, improvements to the data allow the end user, whether an analyst, data engineer or data scientist, to use it with a high degree of confidence.

Bronze to Silver Recap

In short, the Bronze data is still essentially unusable until it gets structured in a way that uncovers its real value. Since each vendor's input has a different structure and column names with their own idiosyncrasies and issues, analyzing the bronze data would be tremendously inefficient. In the process of moving to the Silver layer, these formats are made consistent, denormalized into a single data lake structure, values standardized, cleaned, augmented (with industry dimensional data), and documented. While this formatting or transforming of the data is critical, it's important to consider that while each ingested dataset is valuable, the data transformation, integration, and combination increase the value exponentially.



Since the Silver layer is where rules and logic are applied, it is critical that processing works at scale across the entire history of data received, not just future increments as they come in from vendors. We must reprocess the entire dataset with each new code release; this ensures the "Silver" assets are uniform, making it possible to identify and resolve new remastering opportunities. Next, a little deeper dive into what we call "remastering."

Remastering Healthcare Claims

Similar to how an audio technician can take an old recording from the past and improve the sound quality by removing static, we remove the "static" in the claims data to present a clearer picture of what is actually going on in the market. Basically, claims are used to pay for a healthcare service, like a physical exam or a knee replacement surgery. We think of claims as receipts, usually incomplete and inaccurate from an analytics perspective. A claim can be partially complete and yet still be approved and paid. As long as the claim includes required payment information, the other fields are less important, messy, far less accurate or even missing entirely. Consider, for example, an adjudicated

claim with a "service from date" of 2022-01-10 and a "service to date" of 2202-01-02. Easy logic tells you this is probably not a prediction of a long-running procedure that will finish 176 years from now. It is pretty safe to assume this is a typo, and 2022-01-02 was what was the intended "service to date."

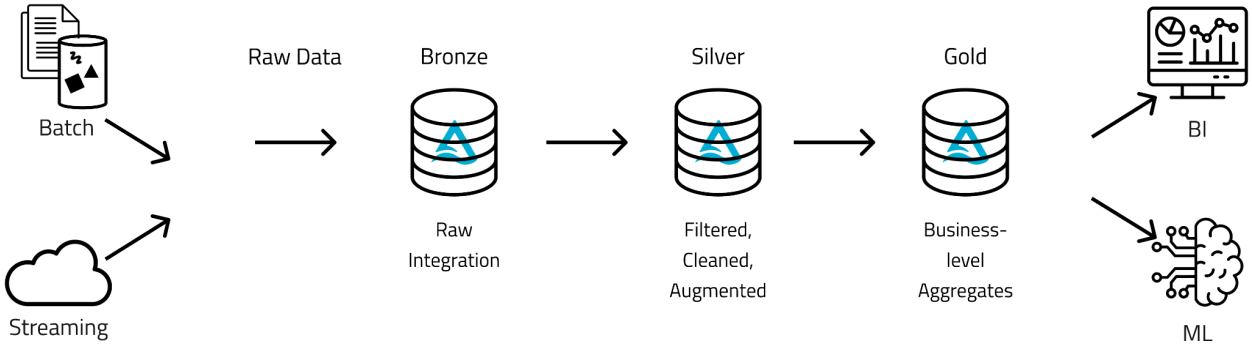
A more pervasive and complex example of a quality issue common to medical claims stems from the addresses on the claim. We have found an average of 5 different ways people can type in any single address. We leverage post office standardization software to make those uniform and add geolocation and other values to the address. Also, an address on a claim is often for a billing office rather than the site of service. So, if you are analyzing market share or patient migration, for example, without resolving the address to the actual site of service, the insights you reach would be unreliable and inaccurate. Using data science and advanced methodologies, we remaster claims data at scale to resolve issues like this and impute missing information.

Within the Silver layer, we join data to healthcare industry dimensional data to denormalize this relational information onto the data lake, which speeds up queries by reducing the number of expensive "join" operations in those varied analyses. Additionally, patient token transformation tasks are automatically performed at scale to conform with HIPAA privacy rules for healthcare data that needs to be de-identified.

For these data engineering tasks, we have many Airflows running scala/python notebooks that use JAR extensions and leverage Delta, Photon, and other features of the Databricks platform – resulting in our Silver layer of remastered source data assets. The Silver layer provides a solid foundation for moving to the next level of building our Patient Events Assets in the Gold layer.

Gold Layer

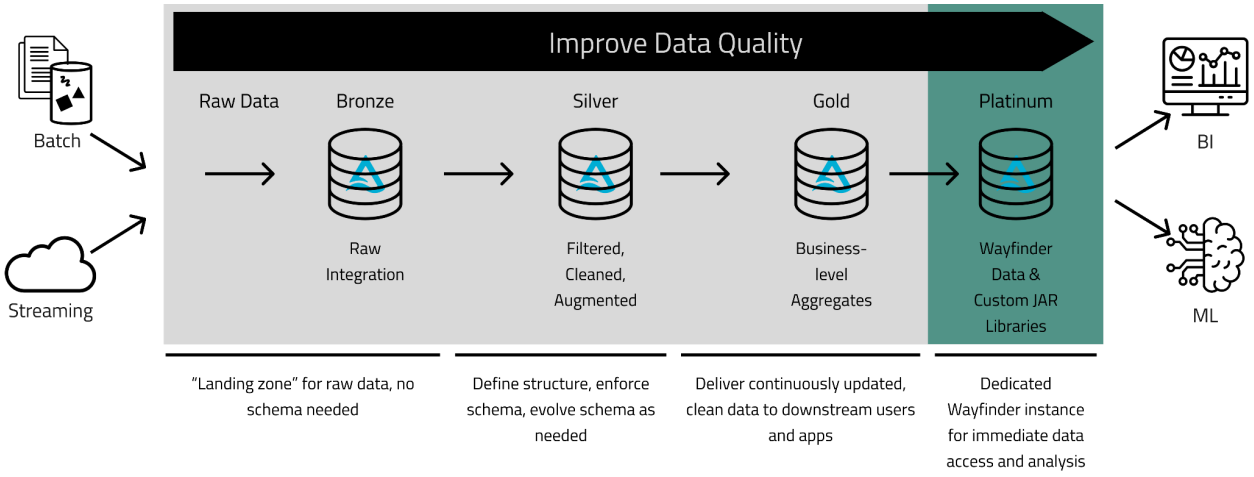
The Gold layer is where we organize data into consumption-ready analytics using the Silver layer denormalized and read-optimized data models with fewer joins. The final set of business logic is applied at the Gold layer, building output presentable for various projects such as customer analytics, product quality analytics, building patient cohorts, drug commercialization, market intelligence assessments, etc. Generally, the business logic at this phase leverages collections of rows to create new data sets focused on specific business concepts as opposed to single transactions/events. For example, patient events and episodes configured to illustrate a patient's journey throughout the continuum of care.



While the Medallion architecture stops at the Gold level, we like to point out it can go even further to deliver specific "insights." In keeping with the medallion nomenclature, we call this the Platinum level.

Platinum Level

After we built our Medallion pipeline, we discovered that while our clients valued being able to jump right over all the time-intensive data preparation steps, many struggled to utilize big data technologies for their research. The Gold layer output is still quite "big," approximately 2 to 4 terabytes (it's called "big" data for a reason). Our clients needed answers to their questions quickly and had little time to figure out how to deal with such large datasets. We realized they could benefit from the same infrastructure and platform we had for our processing.



To meet this need, we became involved with Databricks' "white label" program, where we could spin up Databricks workspaces for our clients. We automated this process to deliver a dedicated workspace or Wayfinder instance for each client minutes after contracting. With Databricks Unity

Catalog, we can instantly share our enormous data assets since no "ETL" is needed, saving our clients money and time.

Now our clients have a place to immediately get to work analyzing our data using familiar SQL on a simple data lake model using Databricks' highly efficient Photon Serverless warehouses. Users with developer skills or data scientists looking to build analytics or machine learning models have a place to build out even more refined sets of analytics output. The Platinum layer also includes a set of Kythera Labs' developed healthcare data tools for our pipeline that we knew our clients also needed. Since extending Databricks with our custom JAR libraries is simple, we can make those functions readily available in each client workspace, opening up opportunities to distribute those Platinum outputs. Now, we are far from a scratchy tune skipping on vinyl but offer a real "Platinum" selling hit.

Additional Platinum Level Benefits

Bring Your Own Data

The Platinum layer is not static, and we continue to add capabilities, features, and benefits. One such capability is a “Bring Your Own Data” (BYOD) approach where Life Sciences and Healthcare Provider Organizations with their own data (regardless of source: data vendors, consolidators, electronic health records, or others) can take advantage of the improvements we provide in the Bronze, Silver, and Gold Layers, such as data cleaning, standardizing, de-identifying, joining, and integrating data at scale. Through our Platinum layer, clients can bring their own data and explore more holistic datasets across sources and types to conduct precise, detailed, and longitudinal analyses. Reach out to talk to us about our medallion architecture and our approach to BYOD.

Conclusion

The Medallion architecture is additive, and you can analyze remastered data from the Silver layer and use derived data assets from the Gold layer to unlock market, patient, and practitioner insights. Using our Platinum layer (Wayfinder) compounds the value of the architecture through the efficiencies of Wayfinder that automate data processing, improve accuracy, and integrate disparate data sources to accelerate big healthcare data analysis and make it more cost-effective.

Want to learn more about how the Medallion architecture makes for more cost-effective healthcare big data analysis? Get in touch at seek@kytheralabs.com or connect with me on LinkedIn: Matt Ryan, Co-Founder and Head of Engineering at Kythera.