Deep Science Publishing
https://doi.org/10.70593/978-81-988918-5-3_8

Chapter 8: Building end-to-end data engineering solutions for medical research and analysis

8.1. Introduction to Data Engineering in Medical Research

"This capability is really a set of skills required to get data to a ready state for analysis where focus is on the building of reproducible, manageable, reliable, enduring, data payloads for analysis and delivery. Data engineering is different from data science but serves as a precursor skill for high quality delivery of data and some level of creation of data products from mainly structured data . Data Engineering has an intense development activity going through data product enablement, scale and sustainability as a function of the service. The capabilities of data technology for data engineering within the enterprise is at scale but limited to initial ingest activities while the capabilities of Information and Data Management Organizations lag farther behind on supportive technology enabling strategies of automating and doing this work at scale (Amer-Yahia et al., 2022; Ates et al., 2022; Chamari et al., 2023). Healthcare, especially Medical Research has made only small improvements in roads at enabling data engineering to support effectively through Data Science Solutions, high quality management and delivery services on data for Data Science solution product creation. Low hanging fruit such as Electronic Medical Record and Patient Health Information have been addressed in initial constructs. Small healthcare organizations have very few resources for Data Engineering Services and Solutions building and primarily operate in a physical model for Data Organization, Management and Delivery. Larger organizations operate at scale for initial ingest building pipelines and storage platform constructs but very few enable scalable model management capabilities or data pipeline reduction management processes for easy, self-guided exploration and validation of data for data product creation involving Data Science Solutions or any other type of discipline and therefore they have low levels of Data Product Creation Work Quality (Amer-Yahia et al., 2022; Ates et al., 2022; Chamari et al., 2023)."

We apply our general, data-centered model-building, boosting, and analysis approaches in the areas of genomics, proteomics, and image analysis. The models that we build represent quite diverse functions and semantics: for example, multiscale signal processing, contiguous pattern detection and characterization, nonnegative matrix factorization, similarity-based pattern recognition, joint modeling, filter design, and I/O modeling. These signal processing and pattern recognition models can be classified into four general application types, each with biomedical examples (Shah et al., 2021; Yu & Lv, 2021; Dolgui & Ivanov, 2022).

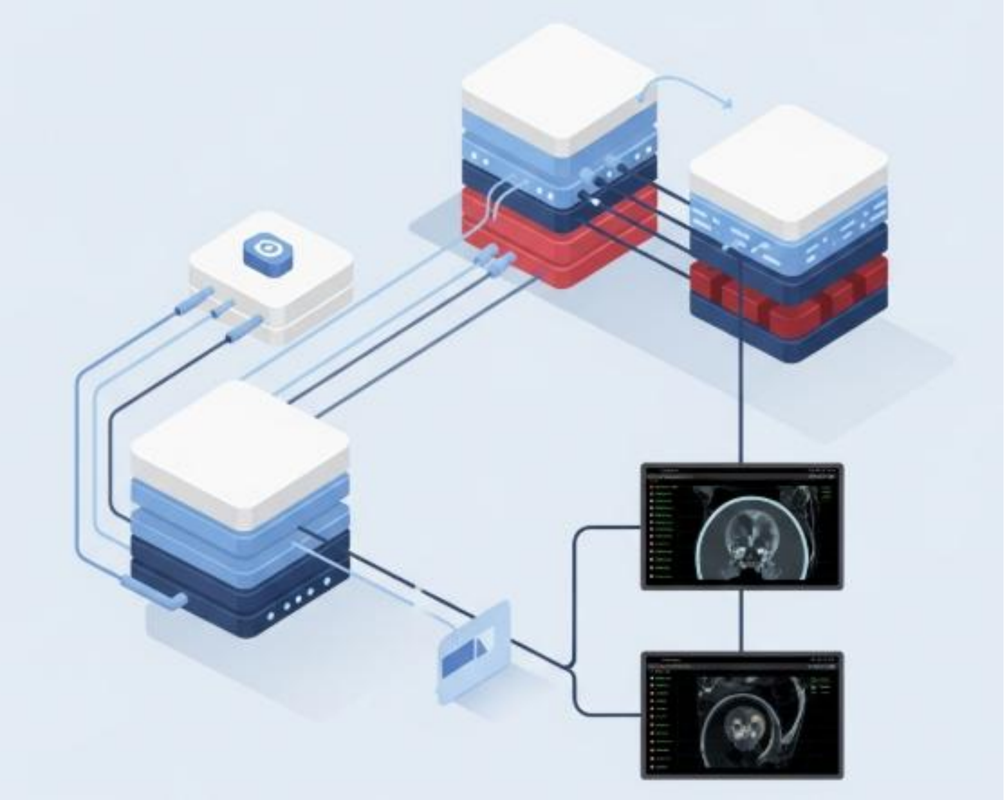

Fig 8.1: Ata Engineering Solutions for Medical Research and Analysis

8.1.1. Background and Significance

Diverse biomedical sub-disciplines and applications analyzed the high-dimensional data produced by many modern technologies including genomics, proteomics, metabolomics, and medical imaging. A few general issues arise from these endeavors. First, most medical research has operated from a "data-rich, information-poor" perspective, attempting to extract pathology or information using assumed patterns from expertise rather than techniques that could yield unique solutions for unique questions. Second, and more generally, most clinical analyses have not been performed from the end-to-end information engineering approach common in the computer research sub-disciplines such as computer vision, speech comprehension, and natural language processing. Data engineering consists of the processes of consolidating, cleaning and structuring data into informative, actionable forms for input into model-building, machine- and/or human-driven analysis solutions. Data engineering may be most important when the volume and/or dimensionality of data to be modeled is very large or when the intended automation of the remaining steps in the analysis pathway is high. When performed upstream from modeling and analysis, data engineering can provide prepared, informative data representations for a widely diverse set of issues and end-users.

8.2. Understanding Medical Data Types

Structured data, characterized by its fixity and adherence to defined data models, is heavily relied upon by various organizations in the healthcare and biomedical domain to solve numerous research problems. A considerable portion of structured data originates from Electronic Health Record systems and includes administrative, clinical, and resource management details like demographics, medical histories, medications, allergies, laboratory test results, radiological images, vital signs, progress notes, provider information, and discharge summaries. Researchers may use such data to predict risk-related conditions, which would require additional surveillance and funding. Predictive modeling would necessitate dataset preparation, which includes structured data cleaning and normalization. Structured data is also easily accessible from clinical local databases for creating cohort data, modeling, and making predictions on specific phenomena. For instance, a review of unique conditions in hospitals revealed a cohort of patients with Atrial Fibrillation who underwent spontaneous conversion, which is approximately a percentage of all discharged patients with AF. Modeling on the unique cohort revealed that in-hospital admissions with cardioversion for AF share significant comorbidities and complication rates leading to longer lengths of stay and higher readmission rates. Such predictive studies would further assist in transitional design in patients with AF.

Contrary to structured data, unstructured data, defined by the lack of a specific data model, is becoming increasingly voluminous over the last couple of decades, with sources like citation databases, repositories, social media, scientific literature, clinical trials, and biomedical journals containing lifesaving protocols and findings that demand data processing and specialized algorithm-ready accessible content. Although unstructured data poses difficulties during processing due to its semi-structured or unstructured nature, researchers have found ways to extract critical insights and obtain desired results from unstructured data resources. For example, Natural Language Processing was employed to clean clinical notes and identify patients facing uncertainties regarding their prognosis; a custom generative language model was proposed to achieve ambitious few-shot and zero-shot performance on various social network tasks.

8.2.1. Structured Data

Data used in medical research can typically be organized in two types: structured and unstructured. In this section, we discuss the characteristics of both types and highlight how they differ in their construction, usage, and meaning. This is an important aspect of building data engineering solutions since these differences are visible at the data infrastructure and pipeline levels. The two types require additional considerations when working with them, especially in processing, organizing, and analyzing data. Planning for both types from an early stage of solution building can lead to optimized workflows in a data pipeline or an infrastructure whose ingestion, storage, and access are modeled after the specific type's needs.

Structured data requires very little to no processing after intake from a data source. In most cases, the interactions with the data are through database queries where researchers can build their own queries and need to know SQL or a similar querying language. The queries return result tables with values organized by rows and columns. Patients represent rows, with one for each individual in the dataset. Columns represent specific medical observations or laboratory tests. The vast majority of research activities over structured data are related to cohort building, which may involve filtering rows by applying rules over a single or multiple columns, mostly based on logical operators, and operations to handle missing values. Although cohort building is mostly done on structured data, it is also common to first analyze specific populations by extracting from the entire dataset only the rows that correspond to the particular attributes of interest. Then, results such as aggregate values or frequency distribution plots for the attributes of interest and possibly some other columns are generated, and the research is published.

8.2.2. Unstructured Data

Within the medical field, unstructured data is defined as information that is either hidden or not well labeled and or difficult to access. This class of information accounts for roughly 80–90% of medical data currently available, which includes: typed or handwritten physician or nurses notes; discharge summaries; consultation notes; pathology reports; radiology and electrocardiogram reports; pertinent negatives and findings; telephone and interview notes; and so forth. The size of an individual note may vary from a couple of sentences to several pages; notes are often annotated with medical codes to signify important findings, but their unstructured nature makes automated processing extremely difficult.

Free text clinical notes appear to be the most useful for clinical decision making, patient outcomes, and other clinical research; to illustrate this point, several key data extraction algorithms for diagnosis and transfer notes show that, while guidelines and standard operations do not contain clinically relevant guideline elements, the standard notes do. Other studies have also shown that semi-structured data (data that has some inherent structure but probably isn’t in a database or lacks external standards) sources contain important process elements. Interestingly, while semistructured and structured data have started corresponding in vertical and temporal aspects of volume distributions, such as ED admissions, unstructured data is still the primary data source for these process elements.

8.3. Data Collection Techniques

Having a sound data collection is one of the main requirements of having a successful study. The data that has been collected must fulfill the purpose of the investigation. The initial research design uses data collection protocols such as qualitative methods which includes unstructured, semi-structured, and structured interviews as well as workshops or focus groups and collection of text documents. These protocols then later on are supported by quantitative data collection methodology using surveys or questionnaires as well as obtaining clinical trials data. The qualitative data will provide the basis to design the quantitative data collection and the quantitative data will be provided for a more reliable and valid conclusion of the results. For large sample sizes, quantitative research is much cheaper and easier to analyze due to the use of closed questions.

Surveys and Questionnaires

The survey methodology has a long established history in health services research. Standardized survey instruments capture processes, pathways, and outcomes across populations, typically in larger groups, in a way that can be quantified, analyzed across sub-groups and compared in a systematic way both across time and across diverse geographic regions. Patient surveys provide comparative benchmark data for practices, institutions and payers. Diagnostic surveys can assist in the identification of illnesses and diseases. Questionnaires with closed and open questions have been designed to elicit information regarding patterns of diagnosis, procedure use, severity, treatment choices, satisfaction, outcome expectations, recovery complications, degree of recovery, the specific nature of post recovery restrictions, and when restrictions were removed.

Clinical Trials Data

Data is also obtained from ADaM and SDTM files which are shared around a particular clinical study via a clinical research site. The process by which the data is generated, processed, validated, and ultimately submitted for approval and licensing of designer therapies is called a clinical trial. Clinical trials are often the largest or the only clinical evidence upon which critical differences among therapies are evaluated.

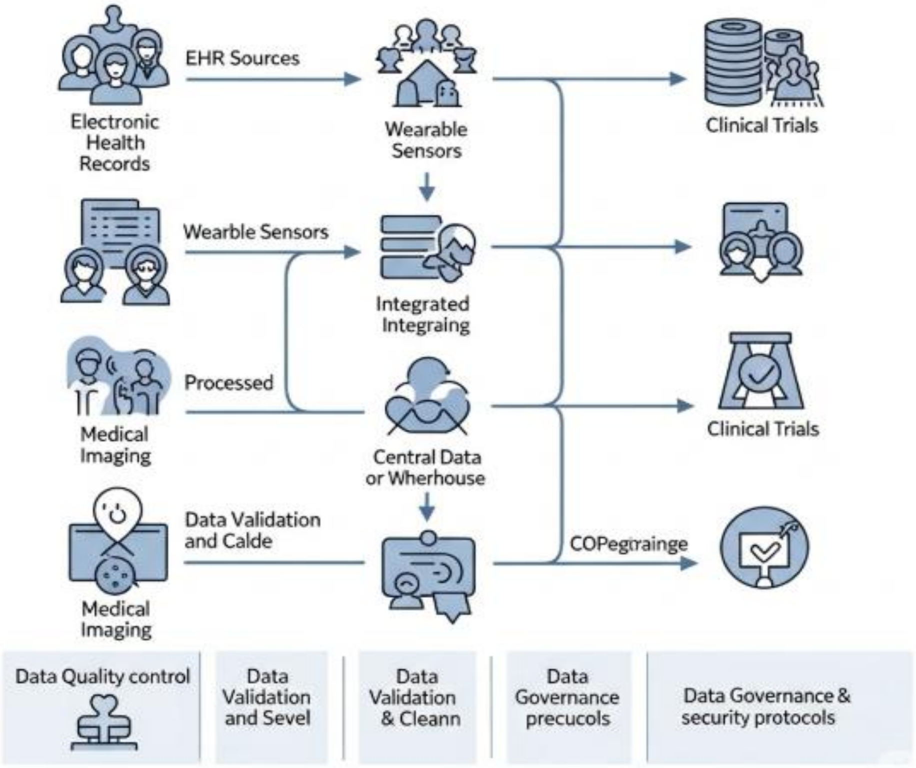

Fig 8.2: Data Collection Techniques

Almost by definition, the batch data processing pipelines that we have focused on in this section exhibit some latency – collecting data for some period of time and then processing and serving it to users. If that latency is after the business event happens, how can users monitor the metrics and inputs that matter and be alerted for situations that demand attention before the fact? One answer is to process rows individually one at a time as they come in from their data sources continuously. With this approach, input and output sources are easily-integrated cable systems that support a simple request-reply mechanism. One should not dismiss the possibility that the incoming stream of records for some data category representing some particular business process is powered by a batch process – a periodic job that computes that data by querying other sources and then pushes the computed results out.

Data storage systems are commonly differentiated into databases and data lakes. The fundamental difference between those two types of solutions is how data is stored and the logic behind its query language.

8.3.1. Surveys and Questionnaires

Surveys and questionnaires offer researchers a unique qualitative perspective, augmenting quantitative data obtained through clinical trials and other methodologies. These tools gather invaluable patient data, including comparative drug use or exposure, responses to questions on the quality or length of life, and even risk factors unexplored in clinical trials. Moreover, the possibility of acquiring some of this patient information in the few minutes it takes to complete a survey or questionnaire has led to a worth-its-weight-in-gold collection of health information from the nearly 40% of US players who answer. At its most basic, a survey or a questionnaire is a form – on paper or electronically – that uses questions (open-ended or closed-ended) as data collection points. Once you have collected responses from a sufficiently large representative sample of your population, the answer(s) you get to each question can be considered the "data." While the sample may represent only the few dozen people you asked, the responses can be viewed as an average for whatever demographic group the sample represents; or for the entire demographic group of interest if your sample size was sufficiently large, and you applied a random sampling method. When querying a large sample of players, the opinion, health experience, or health complaint expressed can be weighted and expressed as common to the group as a whole. For example, given a large enough representative sample of women aged 30 or older, one might say, for the purposes of formulating future drug use protocols, that "19% of the patients surveyed… had experienced breast cancer."

8.3.2. Clinical Trials Data

Clinical trials are the default models for validating the effects of either drugs or treatment protocols. Actually, every product to be consumed by humans introduced into the global market goes through a series of trials before being endorsed as finalized for consumer use. In the medical world, clinical trials are extremely critical, taking as factors involved the therapy to eliminate or diminish the suffering of patients with acute or chronic diseases, and the financial resources demanded by such experiences. Ethically it cannot be debated if they should be done, as human lives are at risk in untested new therapy protocols, yet it would be prudent to ask how they are done and under what guidance have they to be executed. Governments have instituted regulatory bodies that have to approve any protocol to be followed during clinical trials, both in terms of the parties involved – pharmaceutical companies, hospitals, healthcare professionals, or any combination of them who are either funding or running the trials – and of the patients who will participate in the studies. Safety must be guaranteed and stopping the execution of trials is always an option if the set goals cannot be achieved, putting at risk either the patients' health or physicians' ethical responsibilities. Complex processes are set in place so that, through a number of stages validated studies, the trials produce accurate results. In the case of trial data collection, we will have data resulting from pre-market as well as post-market trials.

The difference in the data derived from the various stages of the trials processes is in the involvement of volunteers, either healthy or sick with the disease that is targeted by the trial. In the pediatric population, the role of parents is also important during trials since all authorities are extremely careful with utilizing children in such initiatives.

8.4. Data Storage Solutions

Data can be persisted in several types of storage solutions. Choosing the appropriate solution for the intended purpose is a task requiring attention, as the design of such systems is fundamental for the proper flow of data. To document the variety of existing solutions and designs, we will use the data storage pyramid normally used to explain the design of storage systems to practitioners.

This difference has consequences on how data can be persisted in terms of structure and the data transformation steps necessary before data can be queried. However, despite being commonly accepted that data lakes are defined as storage systems holding raw data without any specific pre-processing or organization, and that databases make use of some logical design for structuring data, it is difficult to define the strict boundaries of such a classification. For instance, data may be added to data lakes pre-structured as database tables. Databases can also be used storing some pre-structured data.

Databases power the vast majority of solutions in the market for storing medical data. Currently, around 90% of the submissions are made in relational databases. Due to their history with evident use in clinical settings, associations have released specifications defined by the community, which define logical schemas for SQL and NoSQL databases that are aimed explicitly at health care data.

8.4.1. Databases for Medical Data

Studies that correlate medical records with outcome variables, such as prescription adherence, mortality, or disease recurrence, are growing in number. The non-structured or semi-structured nature of medical notes makes conventional structured databases not very useful for many medical applications. In addition to these non-traditional usage needs, a lot of patient data is also flowing from wearables and other devices used in healthcare. This data, which is usually time series data, makes regular databases even less useful and necessitates the popularity of specialized time series databases. Additionally, the risks of regulating patient medical records are undoubtedly serious, and database security is another important aspect that has to be considered. While on-premises setups of databases can be effectively secured, it is crucial that solutions are compliant with relevant regulations. The choice of database technology must also fit into the expected future scale of usage.

The explosion of patient data is pushing the creation of new databases for research purposes advancing the semantics of sensitive data. For example, the initial federal deposit of databases has paved the way for a connection between cancer therapy and genomic data for researchers. The variant database architecture, developed for a gene-centric approach to rare genetic diseases of unknown etiology, has gone through the efforts to harmonize the underlying allele severity score against the severity scores from the thousands of existing disease-specific databases. As another example, in the public repository, the acceptance of new seed sequence submissions would facilitate the translation of newly found miRNAs into clinical practice.

8.4.2. Data Lakes

Data lakes take advantage of cloud-related capabilities such as elastic storage size and increasingly larger blob service (semi-structured) and native object storage optimizations to allow storage for cold, unstructured, and semi-structured data, as well as structured data. As such, data lakes allow users to store “any shape of data” for analysis and usage with machine learning engines in their raw form. Native object stores combining the above characteristics with rapidly decreasing costs per stored byte are becoming increasingly compelling solutions for medical research.

Atileo is a data lake for precision medicine, providing elastic and secure storage for the growing volumes of unstructured, semi-structured, and structured data used by Precision Medicine. Atileo allows for seamless proximity of analysis-powered virtual machines with data held in native object storage, providing a fast and affordable option for both batch and interactive analysis. By supporting the TeleMeta catalog, Atileo allows easy query exploration and programmatic usage through a comprehensive API and SQL query interface. In addition to data lakes, the organization currently supports general-purpose and data-processing clusters for traditional Big Data analysis.

A key principle of cloud computing applied by data lakes is the separation of the resources used to store data in persistent form and the resources used to execute ad hoc queries or batch analysis tasks infrequently. This allows the flexible horizontal adaptation of both resources without the requirement to adapt the unspecified service level of the service to users.

8.5. Data Processing Frameworks

Often, the main benefit of having data in a single place in a known schema is that it can be joined with other internal datasets, transformed, manipulated, and utilized to produce new insights and products. The two immense, broad categories for how data is processed at this stage are interactive dashboarding frameworks and scalable data processing frameworks. The former category includes familiar products. Scalable data processing frameworks provide more complex computation capabilities and, in most instances, serve as the backend computation engines behind internal dashboards or other types of end-user products. Additionally, because these products provide more complex capabilities, they tend to be popular with data teams when answering ad-hoc one-off questions. For prototyping particular, more unique visualizations and computations, it can also be more prudent for analysts to write the logic and computation in the language and framework used for exploration.

One of the key characteristics of data analytics that distinguishes it from other common uses of software systems is that analytics is inherently an interactive computing process. An analyst uses exploratory data analysis to iterate many times over augmented focus swaps between examining a slice of the data, visualizing it in different ways by applying ideas from a deep understanding of the relevant business processes, and iteratively refining their mental model.

8.5.1. Batch Processing

Batch processing refers to a processing model where a computer program is developed to run without human intervention, taking a set of data input files and producing processed data output files. Input/output mechanisms that are common in batch processing are file-based, although databases can also be used. One of the first batch processing systems was the one developed for the IBM 7094 and run on it by a research institution, which involved a disk-to-tape transfer controlled by the computer to facilitate batch operation. A file system structure, together with mass storage resources like disk packs, cassette tapes, magnetic tapes, and card files that held standard programs, enabled effective and economical operation of this early batch processing system.

Current batch processing engines leverage distributed processing of data to perform calculations on big datasets. In the early days of these platforms, disk partitions had to fit on a single machine, but clusters were not common. As a result, it was attractive to use both disk partitions and distributed libraries to speed up the batch processing. The concept of Job Tracing allows queries to record the Job Tracing information while executing sub-queries, indicating which specific join type was being executed, along with its partition. Sub-queries later access the trace files corresponding to the original query, allowing each sub-query to know which type of partition file is being generated and how it should process it, especially if the join is performed for each partition.

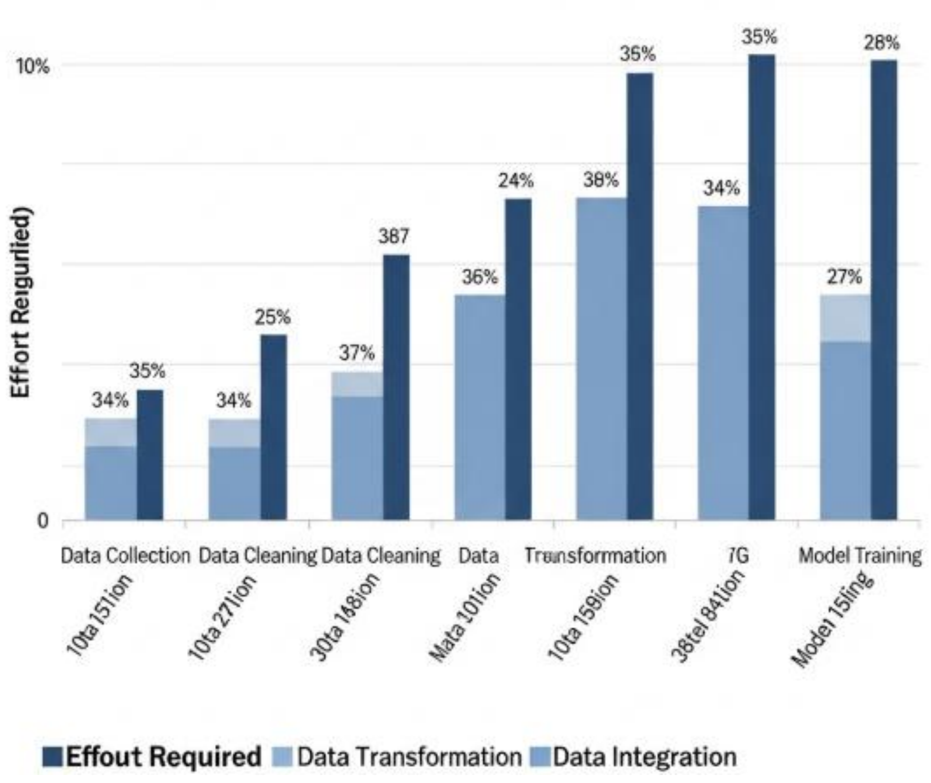

Fig : Building End-to-End Data Engineering Solutions

8.5.2. Stream Processing

Stream processing is a form of computing that allows for the processing of data in motion. In stream processing we typically work with localized state, where state is usually keyed by some finite set as is useful for performing aggregation. This localized state is sometimes paired with a small amount of global state, but often the global state is just a side-effect of the localized aggregations. Stream processing gives us the ability to echo every incoming item or generate a new incoming item every second based on the incoming items without wrapping anything in the traditional begin-processing-at-this-time-effort-and-finish-when-some-assumption-is-done barrier. As such, the cycle of creating files, generating records within the files, and finally running batch jobs to process those files is replaced with a single continuous pipeline of transformations. This gives us added flexibility to quickly react to problems or to notifications, either by retraining the model or asking the model for a score. Given that some models need to be retrained frequently for accuracy, the continuous processing pipeline gives us the necessary capability. Furthermore, being able to cry out and say, "I just saw an item!", "I just scored an item!", "I just retrained my model!", or "My model is dead!" gives an added robustness to monitoring than the more traditional solutions.

8.6. Conclusion

A distributed data pipeline is an essential component for supporting the workflow of medical research and analysis. The deployment of a large number of well-defined medical vocabularies with hundreds of thousands of codes in a non-availability model, and the supporting performance within a few minutes that we achieved is an essential feature for the user to be able to explore answers for their research within a few minutes. Other services such as frequent exporting of the data warehouse content into end-users' designed to enable machine learning and training has been enabled. With the growing need for electronic health data, especially after the pandemic, the open-source software solutions presented can empower system and data engineers to deploy well-defined medical data pipelines similar to the deployment described with modest resources.

8.6.1. Emerging Trends

Over the years new technologies, paradigms, platforms, and solutions are emerging in the area of data engineering supporting medical research and analysis applications. Some of the emerging trends are self-service cloud-based data engineering scalable and elastic solutions accessible to different kind of users, data as a product concept providing the ability to easily publish and consume curated and peer-reviewed regulated data products, using large multi-modal available synthetic data facilitating the creation of artificial intelligence models to manifest generalization properties, abstracting services into workflows for complex adaptive data engineering processes; decentralized trusted architecture to support reproducible research being able to register use cases making the connection to the data, verifiable datasets to add checkable assertions, regulations automating the burden of compliance, oracles solving the trusted data exchange problem backing up sensitive operations requiring real data values; native frameworks to integrate the capability of being able to efficiently query and leverage huge decentralized networks of datasets; trusted algorithms allowing to train artificial intelligence models with simulated data supporting privacy by construction and security by design properties, discovering actionable knowledge from streaming simulated data capable of creating high impact events; generalizable artificial intelligence models trained on huge amounts of synthetic data able to consistently drive measurable desired outcomes in varying real environments governing the associated business processes. Providing services at scale for three key pillars of advanced data-powered products and capabilities common to the healthcare and life sciences domains, data-centric engineering taking advantage of the company growing network of computed datasets, predictive, prescriptive and experimental-driven biology providing support for the research scientist community to assist in their critical and most pressing moments.

References

Chamari, L., Petrova, E., & Pauwels, P. (2023). An end-to-end implementation of a service-oriented architecture for data-driven smart buildings. Ieee Access, 11, 117261-117281.

Amer-Yahia, S., Koutrika, G., Braschler, M., Calvanese, D., Lanti, D., Lücke-Tieke, H., ... & Stockinger, K. (2022). INODE: building an end-to-end data exploration system in practice. ACM SIGMOD Record, 50(4), 23-29.

Zeydan, E., & Mangues-Bafalluy, J. (2022). Recent advances in data engineering for networking. Ieee Access, 10, 34449-34496.

Ates, H. C., Nguyen, P. Q., Gonzalez-Macia, L., Morales-Narváez, E., Güder, F., Collins, J. J., & Dincer, C. (2022). End-to-end design of wearable sensors. Nature Reviews Materials, 7(11), 887-907.

Yu, W., & Lv, P. (2021). An end-to-end intelligent fault diagnosis application for rolling bearing based on MobileNet. IEEE Access, 9, 41925-41933.

Shah, M. A., Szurley, J., Mueller, M., Mouchtaris, T., & Droppo, J. (2021). Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks.

Dolgui, A., & Ivanov, D. (2022). 5G in digital supply chain and operations management: fostering flexibility, end-to-end connectivity and real-time visibility through internet-of-everything. *International Journal of Production Research*, *60*(2), 442-451.