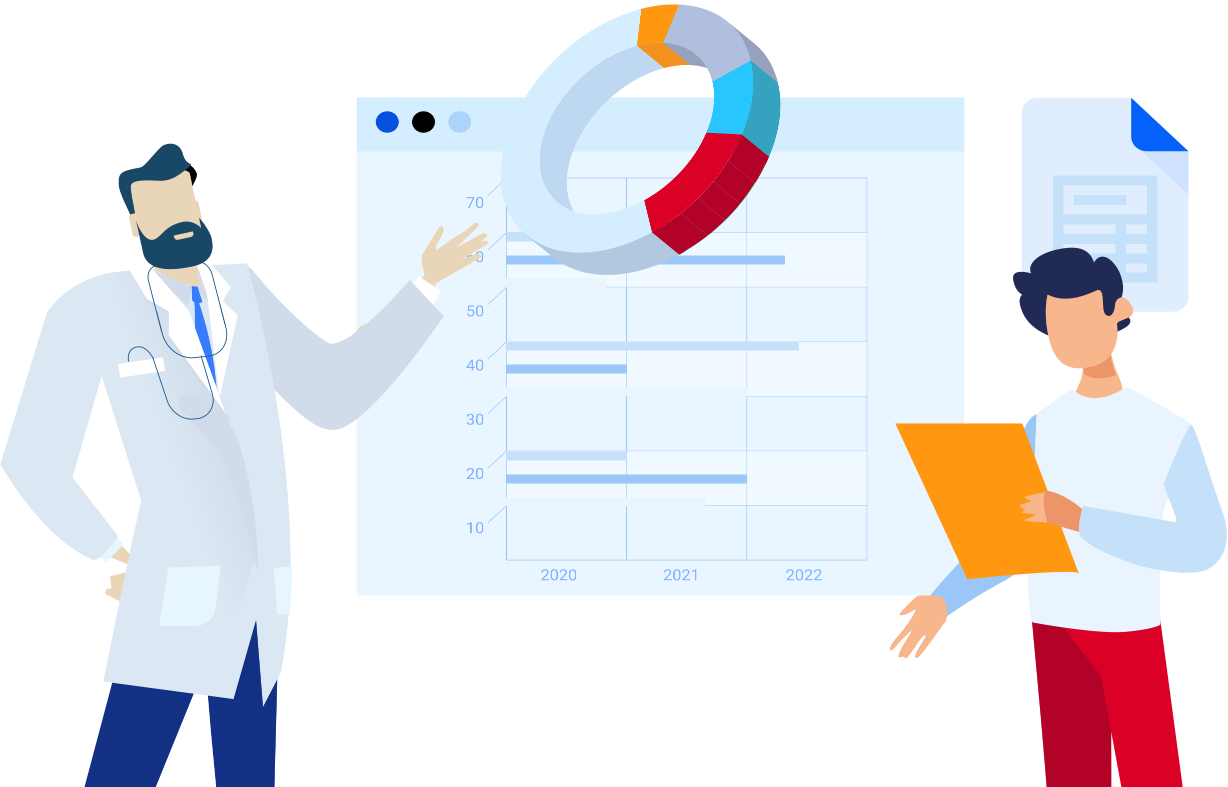
WHITEPAPER

Clinical Data Fabric

2022

keyrus
life science

CONTENTS



PART I

The challenges involved with life science data

THE ABUNDANCE OF LIFE DATA

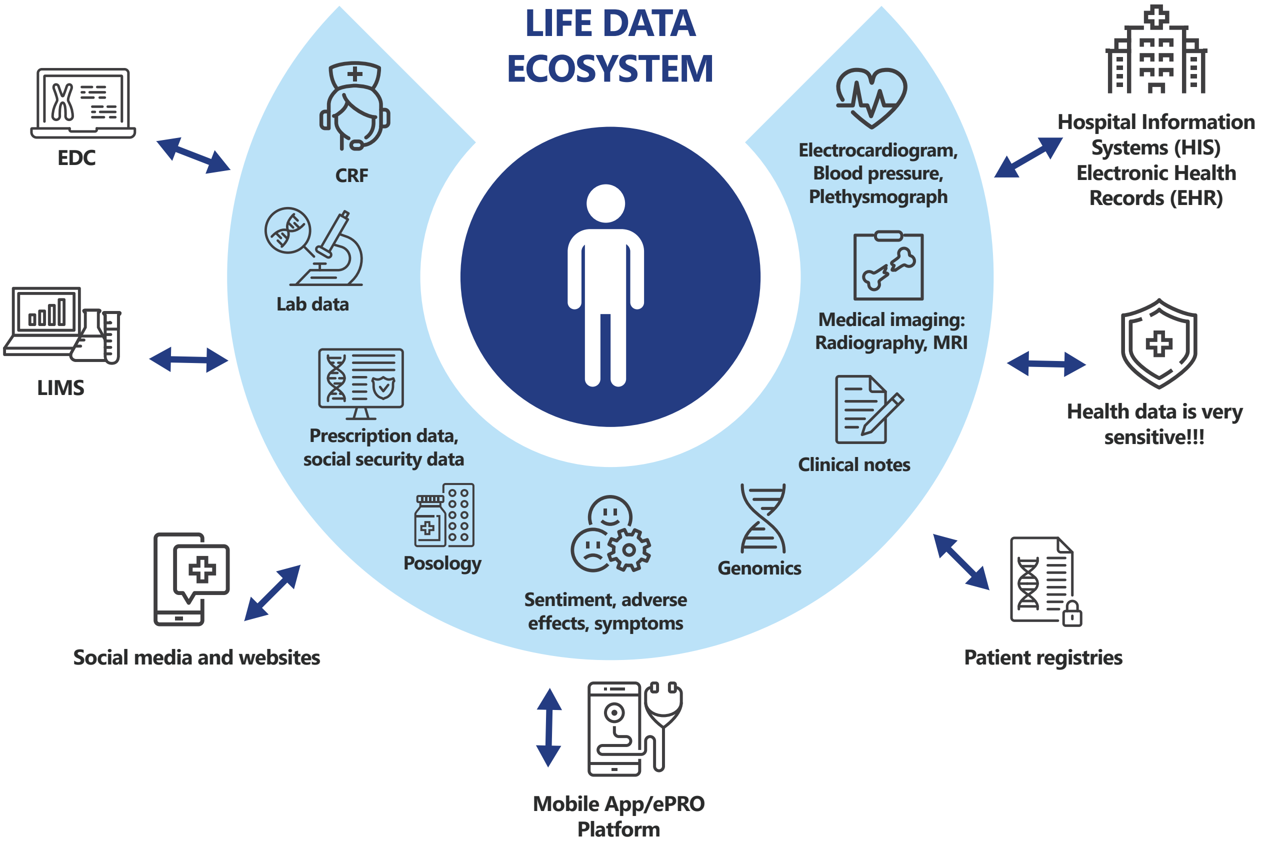
Nowadays, all companies active in the life science ecosystem consider data as a crucial asset to their success. Pharma companies are collecting more and more information during their clinical trials or commercial and medical activities. Clinical laboratories have to deploy complex data infrastructures to better serve their patients and clients, and Biotech and MedTech companies need to gather and analyze a lot of scientific and technical data (literature reviews, chemical property databases, gene databases, ...). To tackle these challenges, life science companies have to be equipped with the right technological stack enabling them to collect, import, manage, store, analyze, and share the required data both internally and with their partners.

To answer those needs, the ultimate solution is the construction of a tailor made clinical data fabric (CDF), which is, to put it simply, a data management design architecture that connects the life data sources with the consumers of the data (applications, reporting tools, AI/ML algorithms, etc.). However, a CDF needs to be considered with care and conceived with a good level of expertise in order to be efficient. Indeed, on the one hand, the number, complexity and format of data sources are increasing; and on the other hand, the data consumers (applications, data scientists, statisticians, clinical project managers) are willing to combine more and more data sources for operational and analytical purposes to create new insights, new drugs and improve people's lives. This puts a raising pressure on the shoulders of data architects, data engineers, and clinical data managers; which without the proper toolkit, will have difficulties to deal with the pace of new data initiatives and products, while implementing regulation rules and maintaining security at the highest level. Developing an effective clinical data fabric is therefore becoming essential these days for most organizations.

In this white paper, we first describe the critical challenges in the life data ecosystem and derived the data processing needs to properly address them. Then, we introduce the concept of a clinical data fabric as the answer to these needs and describe its key components. Finally, we illustrate the concept on concrete use cases and discuss the methodological components involved to implement a clinical data fabric.

KEY CHALLENGES IN THE LIFE DATA ECOSYSTEM

The surge of life data along with metadata and their applications


LIFE DATA ECOSYSTEM
EDC
CRF
Lab data
LIMS
Prescription data, social security data
Posology
Sentiment, adverse effects, symptoms
Genomics
Clinical notes
Medical imaging: Radiography, MRI
Electrocardiogram, Blood pressure, Plethysmograph
Hospital Information Systems (HIS) Electronic Health Records (EHR)
Health data is very sensitive!!!
Patient registries
Social media and websites
Mobile App/ePRO Platform

The original source of life data are the people, may that be patients or healthy individuals.The figure above displays examples of life data and their applications as well as the following instances:

- In a clinical trial (CT) context, we first think about the clinical notes generated by physicians during the clinical visits, i.e., the clinical report forms (CRF). Those are considered as the main source of data during a clinical trial. It generally leads to highly structured tabular data following CDISC-SDTM data standards for submission to authorities like the FDA or EMA, or CDISC-ADaM formating for bio-statistical analysis, thanks to clinical data managers.

- Increasingly, during CTs, data reported directly by the patients (sleep quality, nutrition, adverse effects...) are collected using a mobile application or a web-application, the so-called electronic Patient Reported Outcomes (ePRO). These data are often the answers to standardized questionnaires, specific to a therapeutic area, and lead to tabular data or sometimes free texts.

- When combined with convenient (non-intrusive) wearable devices like heart rate or blood oxygen saturation sensors producing signal data, ePRO can really provide interesting insights on a patient's daily life.

- Clinical laboratory data play a more and more important role as well. These are the blood, urine, marrow or any other biopsy analyses. Although the use of images is more and more common, they usually create structured tabular data, but they are not really standardized across the industry. Thus, during a CT, since we often need to use the services of several laboratories and hospitals to cover a country, some data normalization/standardization steps are required. Therefore, clinical data managers spend a lot of time integrating clinical reports from different laboratories and from different formats to create the SDTM or ADaM data outputs used by biostatisticians for further analysis.

- For research purposes or patient pre-screening, it is necessary to extract data from clinical notes that physicians share within the hospital to track the patient's treatment. Those notes are stored in the hospital electronic health record (EHR) system. It is usually a free text combining measurement results and physician comments. These primary data are collected for operational purposes and need more integration efforts to extract relevant pieces of information for secondary uses.

- Medical imaging data are now mainstream. We can think about radiographies or magnetic resonance imaging (MRI), but it can also be a picture of a biopsy slice, a picture of cells, or of the skin. Their analysis can be automated using machine learning algorithms to detect/count specific cells or detect skin diseases. These image data often require specific storage systems, and it is crucial to store all the related metadata (the data about the data) providing some context information. These metadata describe the lab that conducted the experiment, the measurement device, the normalization methodologies, and should be linked to a full description of the patient's status at the time the image was recorded. This context information is essential for regulatory and analytical purposes. For instance, this is particularly important to assess if two MRI images can be compared using machine learning algorithms.

- The cost of sequencing DNA or RNA has dropped significantly during the last decades. It is now a common analysis to determine genetic variations that increase the risk of developing certain diseases or not responding to a treatment, and to create personalized cancer treatments. These analyses produce a lot of data and metadata; including the patient's status, the information about the biopsy/sample collected, the specification of the sample preparation steps, the details about the sequencing methodology and the data normalization, and finally the sequence of nucleotides with quality scores e.g. in FASTQ format.

- Social media and patient communities/registries are a source of real-world evidence data (RWE). They can provide information about how drugs are actually consumed, their actual posology, the potential adverse effects and competitor treatments. They can also be used to fine tune trial protocols by assessing the feasibility of a study (e.g., adapting eligibility criteria). The data coming from social media is often free text, from which knowledge should be extracted using natural language processing techniques. For rare/orphan diseases, patient registries constitute an essential source of information to push research further, and it is promoted by European initiatives like the "European rare disease platform", aiming at improving interoperability of rare disease registries.

- Companies are also collecting data issued from the commercialization of their therapeutic specialties, such as pharmacovigilance data, patient information systems and hotline data, sales representative reported data, which exponentially fuels daily their data warehouse and is issued from multiple origins all around the globe.
- Finally, prescription data and drug store data can be collected, anonymized and aggregated to provide more detailed sales figures to pharma companies, create patient pathways, and get insights on e.g., drug effectiveness. These tabular data are collected by drug store associations and social security organizations and are subject to strong regulatory requirements.

The life data ecosystem is thus evolving towards an increasing demand from data consumers, due to the diversity of new applications and use cases requesting life data; along with an increase of heterogeneous data (tabular, text, images, signals...) and their metadata collected to support these new applications. Unfortunately, the data produced by source systems can seldom be directly consumed by applications like reporting tools or machine learning algorithms. The data must be stored and transformed beforehand. Furthermore, the combination of different data sources is often required to create valid inputs to those applications, as well as data chaining with external databases. This amounts to having to create and manage complex data pipelines between sources and applications. Moreover, since those pipelines are becoming numerous, it is also important to facilitate their creation by providing a centralized view and access point on all the data available.

The needs:

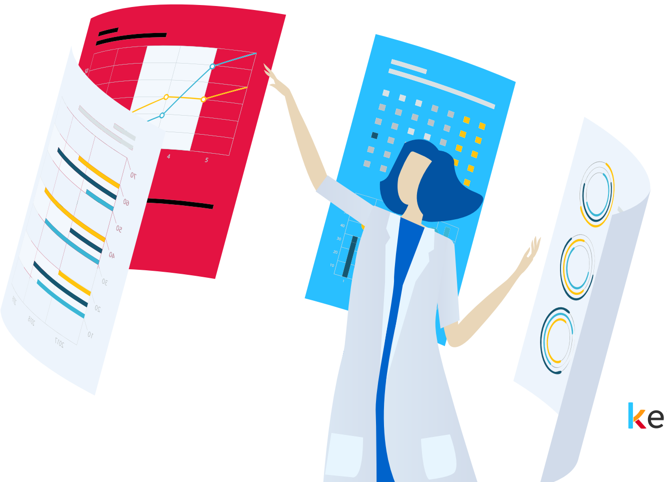
- Storage capabilities of a high volume of heterogeneous data
- Efficiently store and manage metadata to provide context information
- Data transformation capabilities
- Create aggregated data views combining different data sources
- Bridge with external data warehouses to enrich generated or existing data
- Centralized view on all the data available
- Make the data available from one single point and manage access rights

Diversity of operational systems platforms to collect the data

With the large amount of data and metadata that need to be collected for operational and analytical purposes, comes along a lot of software platforms and data repositories, creating a scattered data landscape. Pharma, BiotTech, MedTech, and clinical lab companies have to set up, maintain and integrate all these systems with their current infrastructure. This often means creating a lot of complex data pipelines to deal with the myriad of specific data standards and formats.

Clinical data platforms are now under rapid development. They combine an Electronic Data Capture system (EDC) capable of creating the forms to collect the data during patient visits (CRF); a Clinical Trial Management System (CTMS) allowing for patient visit scheduling, managing team resources and budget; and a Trial Master File (TMF), a searchable document store containing all the trial documents for regulatory purposes. Besides, a myriad of ePRO providers offer a large scope of services (web-app or mobile-app, connectivity with wearables, chatbot assistant, patient telemonitoring etc.) depending on the therapeutic area and the data that need to be collected during the clinical trial. These platforms often have their own data storage systems hosted on the cloud or on private servers and provide data exchange capabilities via APIs. Many clinical trials adopt a hybrid design combining data collection via CRF (so using EDC systems) with decentralized data collection points using ePRO and wearable devices. So, to get the full patient picture, it is necessary to combine those sources. It is important to note that the platforms mentioned above do not always provide the right framework to create and store a combined view of all these data sources, to support reporting and further analysis. This is where a CDF can be valuable to facilitate data interoperability between platforms and reporting systems, and to capitalize on all the data collected.

Research laboratories have to deal often with a myriad of data sources hosted by different public and private organizations like the NCBI-GenBank, NCBI-PubChem or EMBL-EBI for bio-chemical and genetic data. Additionally, many university laboratories make some of their data publicly available to promote collaboration. Sometimes, these data are accessible via a web-portal and can be queried via APIs. However, the data are often extracted manually and end up in a myriad of flat files. A CDF capable of extracting and synchronizing relevant data from all these sources and exposing them to data consumers would save researcher time, promote re-usability and increase research efficiency.



The needs:

- Connectivity via APIs to enable data exchange with applications and existing databases
- Make sure the data remains up-to-date/synchronized
- Store data independently from the applications
- Exposing prepared data to reporting tools and external data science environments

Implementation of regulation requirements

Clinical data are very sensitive and thereby subject to strong regulation rules. These rules are promulgated by organizations like the FDA and EMA, by country authorities and specific contracts. They ensure the safety of the patient by making sure drug development follows strict guidelines; and they also protect the privacy of the patient. Meanwhile, it is important to gather enough data to get more insights and develop new drugs and treatments.

Regulation rules are numerous: the general data protection regulation (GDPR), the «ICH E6 (R2) good clinical practice» (ethical and quality standards for conducting clinical trials) or the "FDA 21 CFR part 11". These sets of rules are usually translated into concrete constraints and guidelines by data governance teams, and materialized into concrete checks on data pipelines, anonymization transformations, codification or data access control. Let us take the example of data integrity, often defined by the ALCOA guidelines, introduced by the FDA. The data should be:

- **Attributable** - need to identify the person or computer that generated the data;
- **Legible** - the data should be readable, in the digital world, it means in a format and with descriptions that are common enough to be understood in the future;
- **Contemporaneous** - a data generated time stamp must be recorded;
- **Original** - keep track of the source data before any transformation;
- **Accurate** - accuracy checks are performed to make sure the data is correct (data is within range or calibration).

To implement all these rules, it is crucial to keep track of all the metadata providing information about the context in which the data has been collected; to store the raw data as well as the transformed data; and to put in place relevant data validation rules. Furthermore, to assess if those rules are properly implemented, we need to have a holistic view on all the data pipelines/ transfers across all company activities, making all the possible transformations traceable in each data flow.

The needs:

- Full compliance with regulation rules to collect, process, store (in which country), and consume sensitive patient data
- Centralized view on all the data available
- Make the data searchable based on keywords
- Data pipeline descriptions across applications to maintain traceability
- Data transformation capabilities
- Define who can get access to what and when

Dealing with constantly evolving technologies & the need for flexibility

It takes on average 12 years for a drug to go from the preclinical phase to the 'go to market' phase. This is a long period during which technology may evolve. Thus, one could be willing to use different EDC, CTMS or TMF systems, providing extra features or embedding new technologies to collect data across the different phases of the study, while keeping all the data available. Which, later on, will be particularly important to leverage past clinical trial data to optimize new trial operations. Thus, in addition to data sources and operational systems, a third party capable of connecting and exchanging data with all your existing applications and creating an independent but connected data layer, can ease decommissioning (old) and commissioning (new) of applications. And besides, making sure that the data collected remains available in your organization and avoiding any kind of vendor lock.

Of course, the evolution of technologies and data requirements may also affect the CDF itself. This is the reason why it must be composed of exchangeable technological components.

The needs:

- Connectivity via APIs to enable data exchange with applications and existing databases
- Store data independently from the applications
- Exchangeable technological components

Support reporting and A.I. initiatives

Machine learning algorithms and reporting tools are big consumers of life data. We can think about applications like patient adverse event monitoring, protocol optimization based on past study data, heart attack predictions based on wearable sensor data, cancer cell detection from biopsy images, or drug posology extraction on clinical notes. To build these applications, we not only need to create beautiful interactive dashboards and use advanced machine learning libraries, but we also need to select, gather and prepare enough data.

Bio-statisticians & data scientists, when assessing the feasibility of a predictive model or a reporting dashboard, often start to conduct a data audit (availability assessment, quality assessment, data comparability etc.). When the data are scattered across different storage systems and when the metadata are scarce, this task is tedious. Therefore, a centralized view of all the data available along with their metadata would definitely be helpful.

Furthermore, data must be prepared to be consumed by reporting and machine learning tools. It is often assumed that 80% of the workload of a data science project is in data preparation. Actually, the data preparation step here encompasses data ingestion, data cleaning, data validation and eventually data preparation for modelling or reporting. Therefore, data scientists and bio-statisticians could gain a significant amount of time if the data was already pre-ingested, cleaned and validated and if the reuse of prepared data sets was made easier, enabling collaboration between teams.

Finally, to create value, data science models must often be deployed in a production environment. This environment should not only launch the model based on specific triggers, but also make sure that the required data are available from all the source systems, and eventually transmit the predictions to their consumers.

Thus, a centralized data layer providing advanced connecting capabilities; and embedding a relevant set of technologies to facilitate data ingestion, to promote reuse of existing validated data sets, and to push model in production; would definitely be an enabler for data analytic initiatives.

The needs:

- Centralized view on all the data available
- Data transformation capabilities
- A secure environment to train and test machine learning models connected to relevant data sources
- Exposing prepared data to reporting tools and external data science environments
- Maintain data traceability and security when it is consumed by applications or analysts

Make data ready for analytics, available faster & processed automatically

It can take several days from the moment a data point is generated to be available for consumption by reporting systems. In many applications, like adverse event monitoring, this is too long. Usually, these delays in the data pipeline are due to manual interventions as the data has to be manually transferred from one application to another, or the data must be validated manually.

These interventions can often be automated by leveraging data communication between applications using APIs as well as automation tools. However, a data transformation step might be needed, or some extra data should be appended to enable the latter application to execute its tasks. A third party, capable of orchestrating data pipelines between applications is thus required.

The needs:

- Connectivity via APIs to enable data exchanges with applications and existing databases
- Automation capabilities: trigger data processing tasks based on events or a schedule
- Data transformation capabilities

Leverage the value of unstructured data

Unstructured data are numerous but often left behind because they are much harder to turn into valuable insights. The clinical notes and all the CT documents and reports (text, Word, PDF or handwritten) in the TMF are of that kind.

To deal with these unstructured data, we first need to collect all the relevant documents in the same place with all the metadata information available (source system, creation date, document type etc.) and store them. Then, specific treatments should be applied like optical character recognition (OCR) and natural language processing techniques (NLP). The latter extracts pieces of information like: protected health information (PHI), diseases, drugs and posology. Those techniques are based on machine learning algorithms. They can be developed on premises, when training data is available, or external solutions like Amazon Comprehend Medical or Google Healthcare natural language APIs can be used. In any case, the outputs of those algorithms have to be stored, filtered, combined with other data and exposed to consuming systems, and this is what a CDF can do.

The needs:

- Connectivity via APIs to enable data exchange with applications and existing databases
- Ingestion technologies to extract data and metadata from free text
- Storage capabilities of a high volume of heterogeneous data

Make FAIR data a reality

Making the data findable, accessible, interoperable and reusable (FAIR) is difficult to achieve with current data management approaches. Usually, you can only achieve those objectives locally within an application like a clinical trial platform or within a specific department data warehouse, provided that the relevant data governance tools are in place. However, to get FAIR data at a larger scale it is key to have a connected platform promoting interoperability between applications and data sources, making metadata searchable, and exposing the data in different ways to foster reusability.

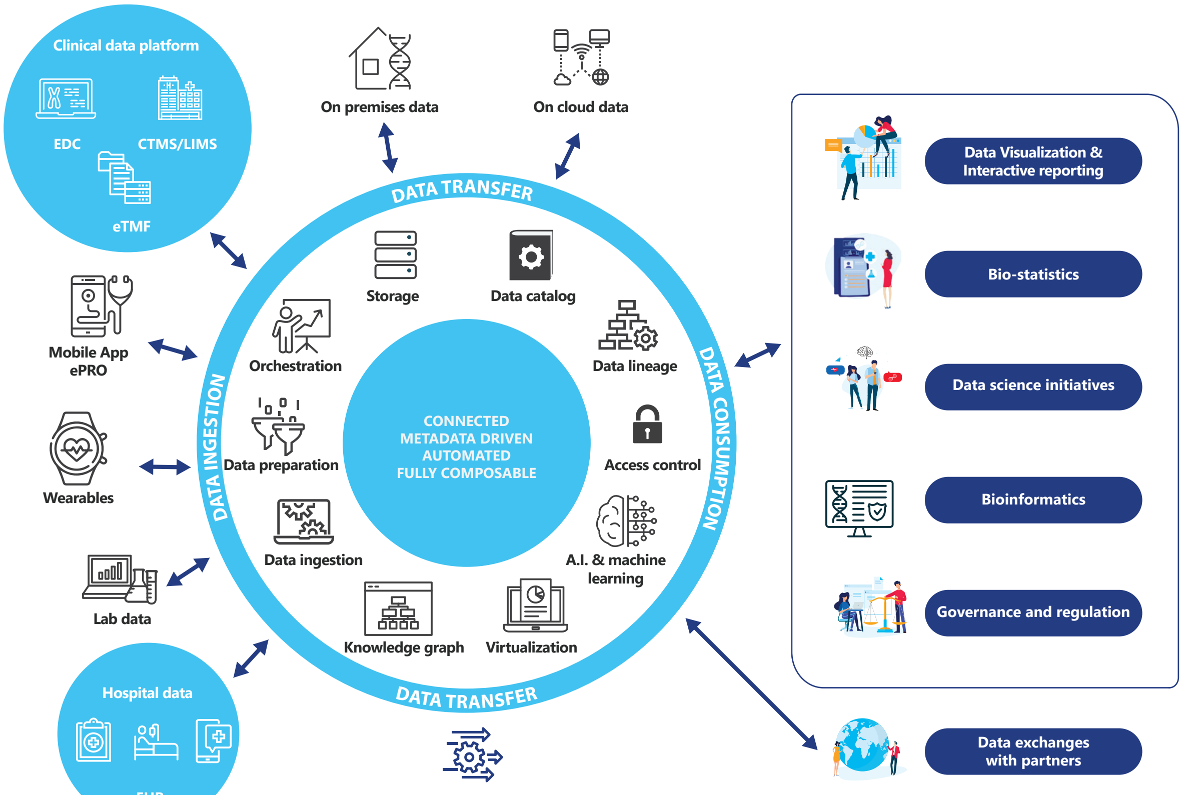
The needs:

- Centralized view on all the data available
- Make the data available from one single point and manage access rights
- Efficiently store and manage metadata to provide context information

PART II

The Clinical Data Fabric (CDF)

INTRODUCTION TO THE CLINICAL DATA FABRIC



THE CLINICAL DATA FABRIC

A data fabric is an emerging **data management design architecture** that consists of an **integrated layer (fabric) of data and connecting processes**. The fabric presents a centralized enterprise-wide coverage of data across **multiple data sources** and applications that is future-proof and not constrained by any single platform or tool restrictions. It is not meant to replace existing systems, but to make them **interoperable**.

The data fabric follows the following guidelines/philosophy to meet the challenges described in the previous section:

- A CDF has **advanced connecting capabilities** to interact (in-bound and out-bound) with all data storage systems and applications/platforms no matter where they are located (on-premise, private or public clouds).

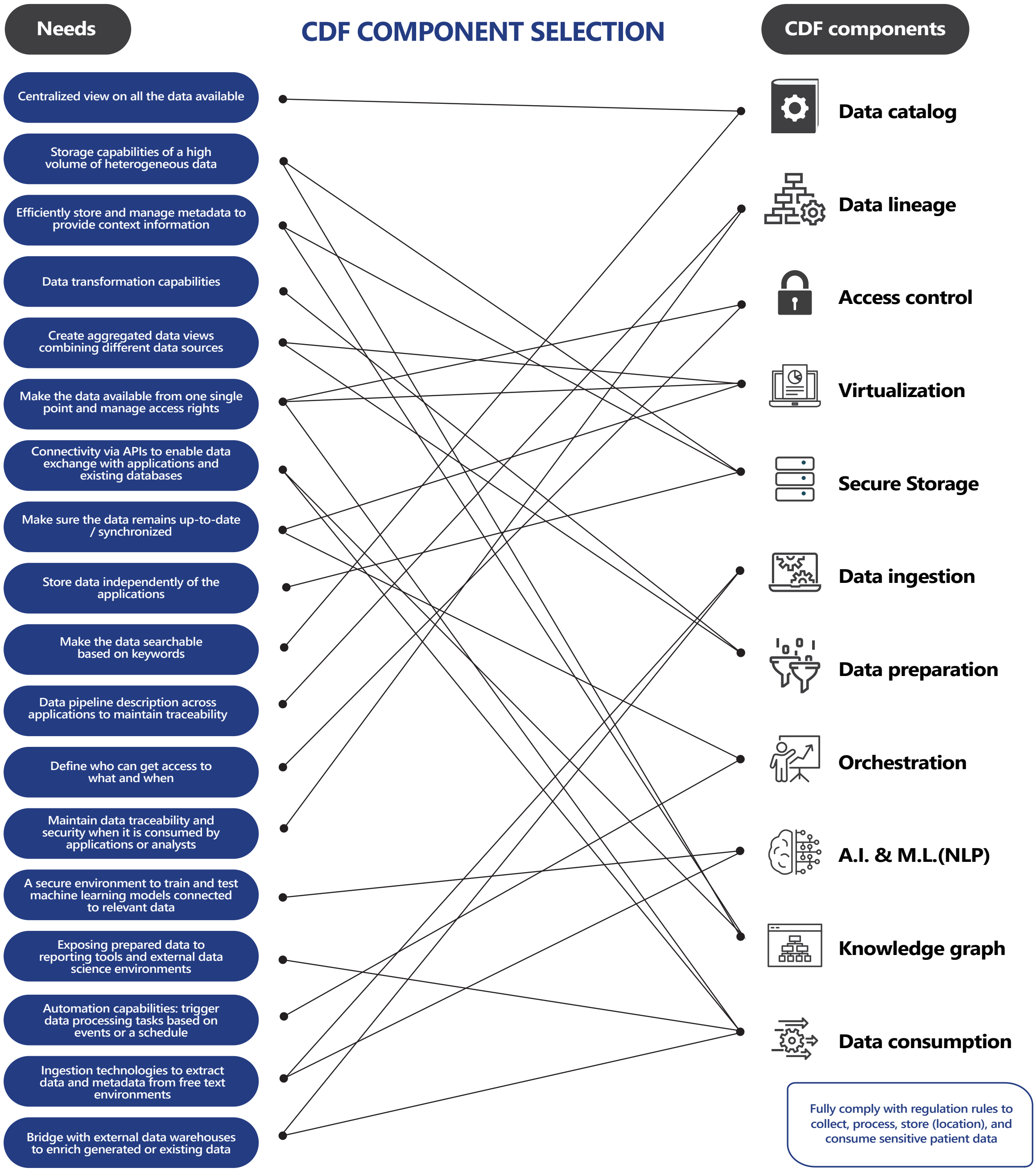
- A CDF follows a **metadata-driven approach.** It encourages/enforces the collection and storage of all the metadata providing domain-specific context descriptions and technical storage details to make data reusable and promote interoperability. This allows for data transformation specification which is independent of the platform that will execute them to gain flexibility.

- A CDF should **facilitate or automate** the construction and execution of complex data transformation pipelines whenever possible, by taking advantage of new tools and the use of advanced technologies like Natural Language Processing (NLP) and knowledge graphs.

- A CDF is **composable by design.** It is made up of components that can be selected and assembled in various combinations to ensure it meets present and future needs.

- A CDF is **future-proof and software solutions agnostic** in order to guarantee its persistence over decades, capable of adapting to technological and regulation evolutions.

The components embedded into a CDF depend on the set of specific data needs, like the ones derived in the previous section, as described in the figure below (CDF component selection). This emphasizes the flexibility of the CDF architecture, and positions it as a key technological block to meet the life data processing challenges. Of course, this flexibility comes at a price. Selecting the relevant set of technologies and providers is not an easy task, and it requires the collaboration of different teams.


Needs
CDF COMPONENT SELECTION
CDF components
Centralized view on all the data available
Storage capabilities of a high volume of heterogeneous data
Efficiently store and manage metadata to provide context information
Data transformation capabilities
Create aggregated data views combining different data sources
Make the data available from one single point and manage access rights
Connectivity via APIs to enable data exchange with applications and existing databases
Make sure the data remains up-to-date / synchronized
Store data independently of the applications
Make the data searchable based on keywords
Data pipeline description across applications to maintain traceability
Define who can get access to what and when
Maintain data traceability and security when it is consumed by applications or analysts
A secure environment to train and test machine learning models connected to relevant data
Exposing prepared data to reporting tools and external data science environments
Automation capabilities: trigger data processing tasks based on events or a schedule
Ingestion technologies to extract data and metadata from free text environments
Bridge with external data warehouses to enrich generated or existing data
Data catalog
Data lineage
Access control
Virtualization
Secure Storage
Data ingestion
Data preparation
Orchestration
A.I. & M.L.(NLP)
Knowledge graph
Data consumption
Fully comply with regulation rules to collect, process, store (location), and consume sensitive patient data

THE CDF COMPONENTS AND WHAT THEY ARE FOR.

Data catalog

The data catalog is one of the key bricks in a metadata-driven approach. It is a collection/inventory of metadata extracted from all the connected source systems, along with some advanced search capabilities.

A data catalog can store many metadata. For a field/column in a data table, it can be: the field name, data type (number, text, date etc.) and its description. But it may also store information about the source of the data, the contract number for external data or some quality flags.

It provides an overview of all the data available in a company from one single point. This avoids having to perform time-consuming explorations across departments asking for data descriptions, which when they exist are usually not searchable. It is thus a huge time saver.

Data lineage

Data lineage is a data flow description from its source to its end point, containing all or the main transformations applied to the data. It is useful to evaluate the trust one can have on a data set, by checking if it has been combined with other sources or what are the cleaning steps and quality checks that have been performed.

A data lineage module can automatically extract data transformations from compatible systems and provide some visualization tools to see the data flows at different levels of detail (on a business level for data governance as well as on detailed data transformations for data engineers).

Access control

Global access control to all resources is always important to meet security requirements, as well as access monitoring to know who (person or application) has done what and when.

To implement the data governance rules, a data management team will have to rely on data catalog, data lineage and access control components.

Data virtualization

Data virtualization is a technology capable of connecting to data sources to extract the metadata and expose combined/transformed versions of the data contained in these data sources into virtual views containing metadata only. When a consumer (e.g. a data scientist or an application) is requesting the data in these virtual views, queries are launched in the source systems and only the relevant data are extracted, combined and transmitted to the consumer. So, from the consumer point of view, everything happens as if all the data were located in the virtual view.

This technology is particularly useful when it is necessary to combine different data sources, and you cannot copy the data due to governance restrictions or the data is changing constantly, which makes it difficult to maintain an up-to-date copy. For instance, a pharma company willing to compute statistics on current trials to improve its operations needs to connect to several CTMS containing the data on current studies.

Data storage

A myriad of storing systems can be embedded depending on the needs. From the classical relational database system storing the results of a clinical trial in CDISC-SDTM data format, to more specific storing systems for images, free texts, signals and DNA/RNA sequences.

Data ingestion

This component provides a set of connectors to existing data warehouse and data lake technologies, as well as tools to interact via APIs with operational systems like CTMS and LIMS systems. Hopefully, nowadays, most operational software platforms provide data interaction capabilities via APIs. In addition, this component provides a set of tools to ingest raw data like text files, Excel files, Word documents and image data.

Data preparation

Data preparation encompasses a variety of different actions: replacing missing or aberrant values, selecting a subset of the data, joining/combining two data sets together, etc.

Orchestration

Orchestration is needed to operate long data pipelines embedding different tasks, using different technological modules that must be interconnected. Orchestration systems help define data pipelines and make sure they are properly executed. They often provide error monitoring (trigger an alert) and advanced scheduling features.

Suppose a data manager needs to combine data coming from different labs. Each time a lab submits a data sample, they need to: update a dashboard to monitor the data compliance; share an updated aggregated version of all the data collected so far with bio-statisticians, and send a report to an operational tool used by clinical project managers. These simple tasks can be fully automated using an orchestration module, enabling faster delivery and freeing up time to perform other tasks.

A.I. & machine learning technologies

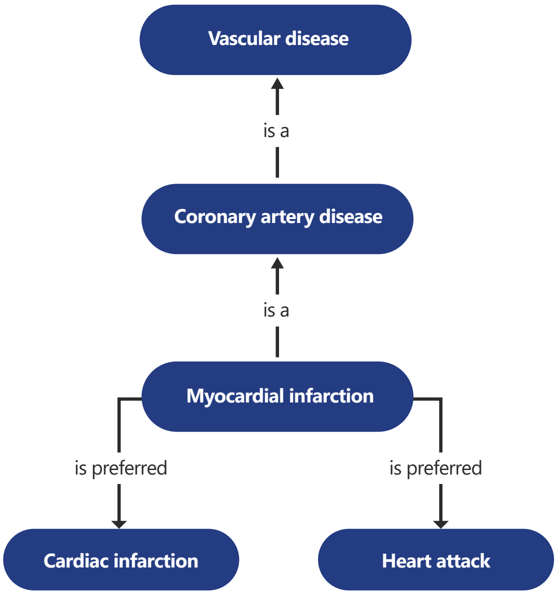
A data fabric supports data science initiatives by providing a single access point to a variety of data sources, as mentioned before. It thus acts first as a data provider layer for A.I. development environments or applications.

However, a CDF can also integrate an A.I. development environment for data scientists and data analysts. It can consist of an environment where R or Python notebooks are installed with a set of certified relevant packages. Being integrated into the CDF, this environment benefits from all the data governance features and data accessibility of the CDF. It enables data scientists to focus on creating insights and not on installing programs. It will reduce the efforts to push new initiatives further, in a domain where 'fail fast' approaches are the norm, and thereby allow for more use cases to be considered, increasing the opportunities to make new discoveries.

A.I. technologies and more specifically natural language processing (NLP) can also be used to automate and facilitate the creation of data pipelines. For instance, NLP can be combined with knowledge graphs to help the clinical data manager map data from one data format to another (like the mapping from the output of the EDC system to SDTM), suggest data transformations, or propose fields of information that could be combined to generate new insights. These technologies are under active development, and usually spread across the other data fabric components like the data catalog, the data preparation module or the consumption layer, enabling faster pipeline definitions and making data discovery easier no matter the data literacy of the consumer.

Knowledge graph

A knowledge graph is a mathematical structure, not to be confused with the graph of a function, composed of nodes/vertices connected with edges. It is thus extremely useful to represent objects/concepts and the connection between them. In particular, graphs can be used to represent metadata and knowledge we have about a specific topic. This includes terminologies like SNOMED linking equivalent terms together, providing concept IDs, defining preferred/official terms and synonyms. For instance, 'heart attack', 'cardiac infarction' and 'myocardial infarction' are synonyms and the latter is the preferred term. Besides, an ontology provides relationships between concepts. It links the 'myocardial infarction' concept to a 'coronary artery disease' which is linked itself to a 'vascular disease', see figure (Myocardial infarction knowledge graph). Terminologies and ontologies are essential to leverage the data in text files like clinical notes and scientific papers. They can be also used for data mapping, recognizing the synonyms in field names or field descriptions.



MYOCARDIAL INFARCTION KNOWLEDGE GRAPH

Ontologies can also be combined, creating more links and more knowledge. For instance, a specific disease can be linked to treatments, drugs, drug chemical properties, adverse effects, etc.

Knowledge graphs are also used to create a unified data model of different data sources. This graph data model is constructed from metadata (e.g. a radiography has an image ID, a link to a file, device used, date, person responsible) and data schemas (e.g. an entity relationship diagram for tabular data describing all the fields/columns and how to join them together) from the sources we need to combine. Then, it can be queried using specific query languages like SPARQL, to extract required data.

There are different types of graphs and technologies to store and query them. Though graph technologies are not widely used today in the data management field but more reserved for specific R&D applications, they are considered as one of the most promising components of a modern data fabric. They are the subject of intense technological developments.

The consumption layer

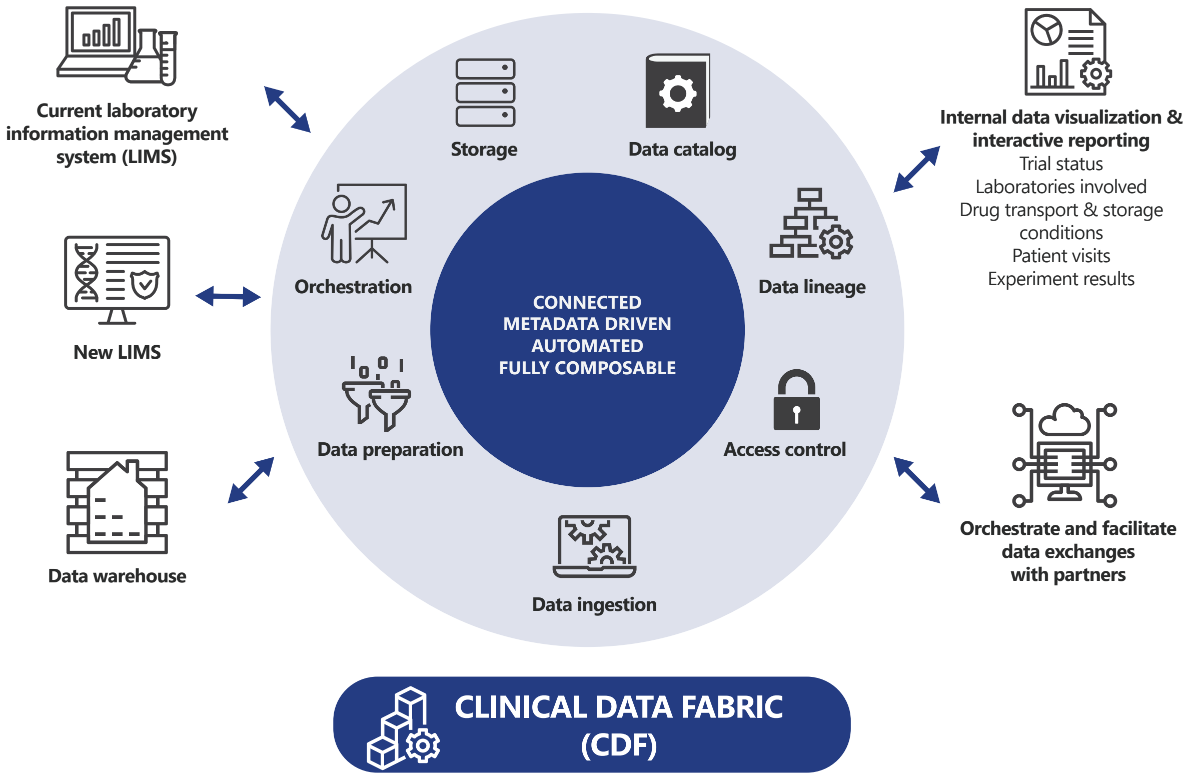
The data fabric enables any data consumers to get access to the data. It can be an application (chatbot, reporting tool, mobile-app) or a person (data scientist, clinical data manager, clinical project manager, business) no matter their data literacy to get access to the data they need and for which they are allowed access to. The consumption layer provides a set of tools to design APIs and create data extracts in a variety of formats to meet consumer requirements.

PART III

Examples of use cases

In this section, we illustrate how a well composed CDF can meet key data management objectives on two concrete use cases.

USE CASE 1 – EXCHANGE DATA INTERNALLY AND WITH PARTNERS



Context & objectives

A clinical laboratory is making a major change to its operational system to support clinical trials and lab testing. In order to support this change and allow for a smooth transition, they want to implement a future-proof platform to fulfil the current and future needs of the company regarding LIMS reporting, management reporting and operational reporting. This platform should be an integrated solution that increases the efficiency of the organization, by delivering usable and trustworthy information to the people (internal & partners) who need it quickly and efficiently, to allow fact-based decision-making in order to achieve pre-selected goals.

Why a data fabric?

- **Obtain a single and accurate source of integrated information.**

  The data consumption layer provides easy access to data for all the users (clinical data managers, data scientists, business & operational teams, applications and all the external partners) leveraging different data exchanges & export capabilities.

- **Make complete information available faster.**

  The data preparation and orchestration components support the development and automation of complex data pipelines (data aggregation, joining of data sources etc.) which reduces the time required to build, update and maintain the data flows feeding reporting tools and applications.

- **Ensure "best-fit" tool selection and scalability towards future needs.**

  The CDF is fully composable by design, meaning you can select the right set of functions/components to meet your requirements at a given point in time. Then, you can still adapt (change/update) this set based on your needs, e.g. by changing the storage technologies under the hood to obtain higher data throughput.

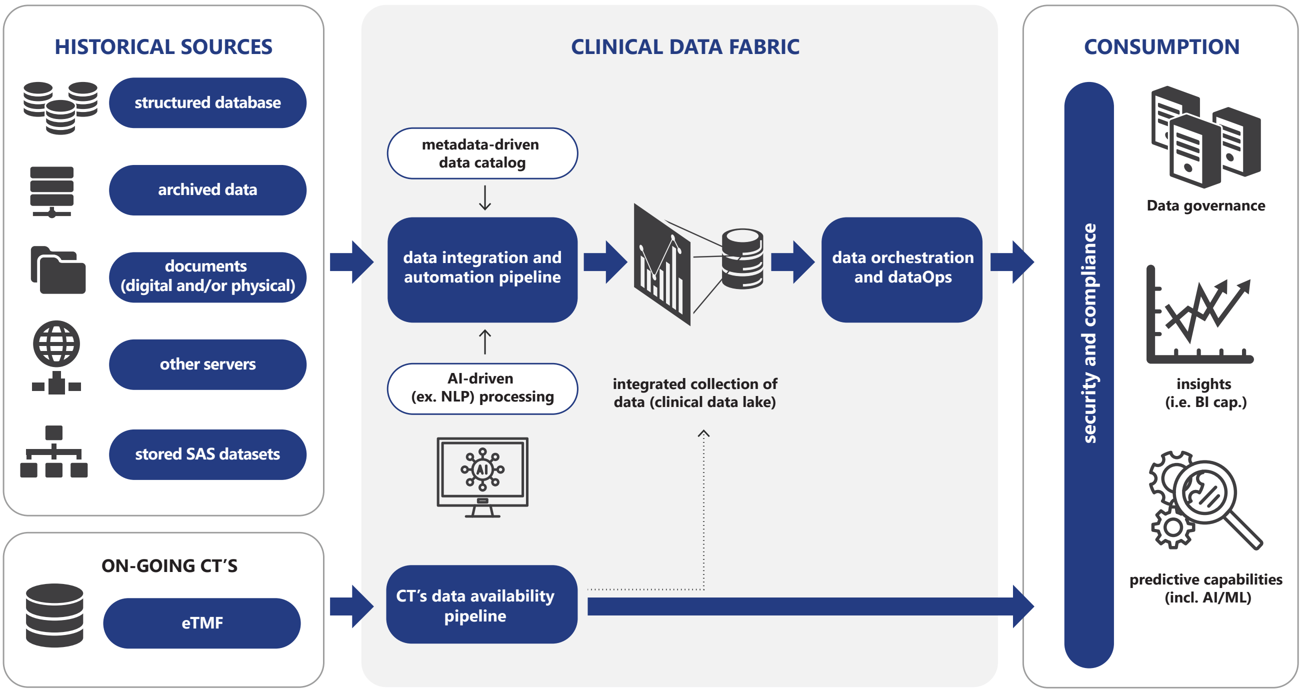
- **Connect and consolidate all the sources.**

  With its advanced data ingestion capabilities, the data fabric can connect to all kinds of sources: operational systems like CTMS/LIMS, existing data warehouses & data lakes and even lab measurement devices.

- **Compliance**

  The data fabric can be composed to meet any kind of regulatory requirements pertaining to data storage (security, location etc.) and data processing thanks to its data catalog, data lineage and access control components.

USE CASE 2 – CAPITALIZE ON PAST CLINICAL TRIAL DATA


HISTORICAL SOURCES
structured database
archived data
documents (digital and/or physical)
other servers
stored SAS datasets
ON-GOING CT'S
eTMF
CLINICAL DATA FABRIC
metadata-driven data catalog
data integration and automation pipeline
AI-driven (ex. NLP) processing
integrated collection of data (clinical data lake)
data orchestration and dataOps
CT's data availability pipeline
CONSUMPTION
security and compliance
Data governance
insights (i.e. BI cap.)
predictive capabilities (incl. AI/ML)

Context & objectives

A pharma company would like to further digitalize its clinical and R&D processes. In this purpose, they plan to build:

- a tailor-made data platform to collect, enhance, and consolidate all digitalized data, from multiple sources;
- BI and reporting capabilities exposing those data to all relevant stakeholders (internal and external), in order to accelerate integrated data availability, improve decision-making and unlock predictive capabilities.

Why a data fabric?

- **Gather current and past clinical trial data.**

  Thanks to the ingestion and storage components, the data fabric is capable of integrating and storing the numerous files in the TMF, the study results in SDTM and ADaM format as well as all the data stored in other operational systems.

- **Make this data consumable by reporting and prediction algorithms.**

  Leveraging NLP and the data preparation capabilities of the data fabric, enables information extraction from documents and its transformation to feed reporting systems.

- **Enable fast reporting.**

  The data catalog enables the BI team to have quick access to all the data available from raw data to already prepared data and their descriptions, enabling the reuse of previous works and a much easier identification of the data required to create insightful dashboards. In addition, the data orchestration component enables data processing automation.

- **Centralized metadata & governance.**

  The data catalog centralizes and gives access to all the metadata. When combined with data lineage and access control, it gives monitoring and supervision capabilities on all the data flows, enabling the deployment of governance rules.

- **Regulatory compliance (ALCOA+ framework).**

  The central position of the data fabric in all the data pipelines of an organization makes monitoring of any data transformations and movements across all sources and end-points (people and applications) possible.

PART IV

How to set up your own CDF

HOW TO START A CDF?

A CDF is a really versatile architectural concept that can be applied to a large variety of cases and is evolutive by nature. However, its activity touches several company departments. So, moving to such architecture requires having the right team in place. Collaboration between IT, data architects, clinical data managers and business holders is crucial to develop a successful CDF.

Due to the full flexibility in the components and technologies that can be embedded into a CDF, there is a risk of creating an overly complex platform that could be difficult to maintain or to forget some key components that could limit the scalability and reusability of the platform. It is therefore recommended to start with a detailed scoping study. Here are examples of topics that should be addressed during this as well as some advice.

Identification of current project objectives

A need for a data fabric is often triggered by looking at meeting the new requirements of one or two new data initiatives. This is a good starting point. However, you should also take this opportunity to analyze the friction points you have with your current data pipelines, and to elaborate a vision of your future needs as well as expanding its use to all data use cases a company is facing.

Overview of the current data architecture

The CDF is not meant to replace your current data infrastructure, but to interact with it and to complement it with the relevant functionalities. It is thus crucial to perform an exhaustive audit of your current data architecture and technological stack.

Definition of the new CDF architecture

Depending on your needs and objectives, you will need to identify the key components to embed in your CDF, to use and select the relevant solution providers. Particular attention should be placed here, on specific accreditations to process sensitive clinical data for storage location and all software components.

Deployment and maintenance of the CDF

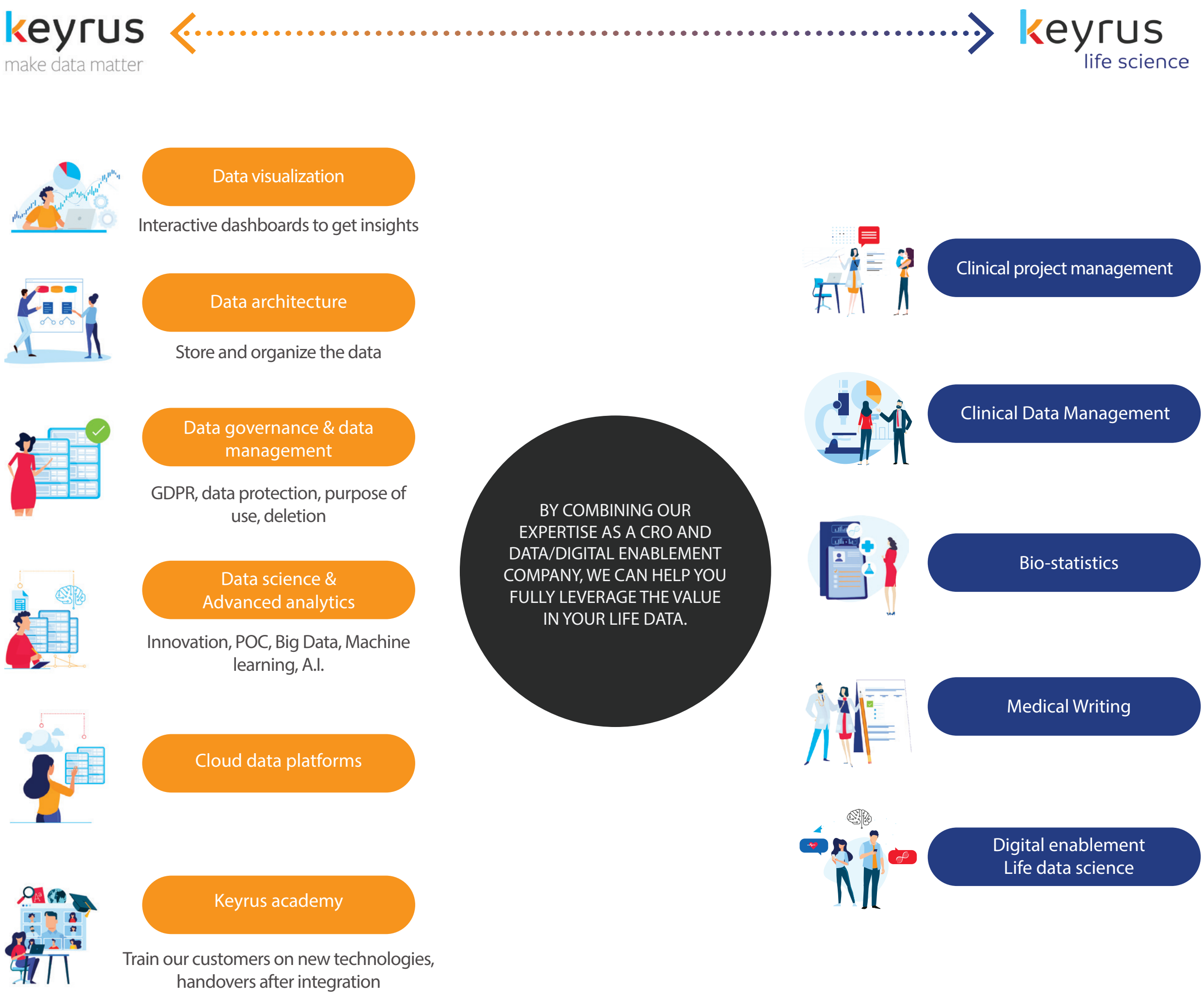
Deployments are usually made by an internal team with the help of a 3rd party company specialising in CDF architecture. Note that a healthy data fabric is constantly evolving and should be scalable by design. To keep up with new resource requests and technological upgrades, a devoted team composed of IT experts and data engineers is required to ensure maintenance and elaborate new user requirement specifications.

It is also important to decommission unused modules or data pipelines that may prevent the CDF from evolving.

Since implementation is never straightforward and requires analyzing and understanding many interdependent domains, an iterative methodology (the implementation project is split into multiple releases or iterations) is often the best approach compared to big bang waterfall delivery methods, for the following reasons:

- The delivery of quick results for the business and obtaining an early ROI on investment.
- Reducing the risk of failure of the total initiative due to having a manageable scope and size, and at the same time avoiding loss of credibility and sponsoring buy-in strategies caused by long delays in delivery.
- Ability to work with small teams that are focused on/specialized in the specific problem area covered by the iteration.
- Limitation of the amount of rework, thanks to early and regular feedback obtained from the stakeholders during each of the iterations.
- Ability for the business and IT organization to build up experience and skills in a gradual way.
- Capability to keep the solution scalable and reactive to any new needs that will emerge over time or evolutions from external factors.

HOW KEYRUS CAN HELP YOU?



keyrus make data matter

keyrus life science

Data visualization

Interactive dashboards to get insights

Data architecture

Store and organize the data

Data governance & data management

GDPR, data protection, purpose of use, deletion

Data science & Advanced analytics

Innovation, POC, Big Data, Machine learning, A.I.

Cloud data platforms

Keyrus academy

Train our customers on new technologies, handovers after integration

BY COMBINING OUR EXPERTISE AS A CRO AND DATA/DIGITAL ENABLEMENT COMPANY, WE CAN HELP YOU FULLY LEVERAGE THE VALUE IN YOUR LIFE DATA.

Clinical project management

Clinical Data Management

Bio-statistics

Medical Writing

Digital enablement Life data science

Keyrus can mobilize a diverse workforce of experts in the clinical domain (clinical data managers, bio-statisticians, interventional and non-interventional study medical leaders) as well as data experts (data architects, data scientists etc.) to help identify and confirm the added value of a clinical data fabric for your organization. Following this, we can design, implement, and deploy a tailor-made clinical data fabric. Depending on your needs, we can configure the required components, or identify external partners of cloud, BI, CTMS or ETL so to make sure you have the right and most exhaustive portfolio of technologies at your disposal. The management of your clinical data fabric can then be handed over to your team thanks to our provided training sessions, or Keyrus can maintain it on your behalf. If you are facing some of the challenges described above or want to know more about the clinical data fabric, do not hesitate to contact us for a data ideation design thinking workshop.

Contact us: https://keyruslifescience.com/

keyrus

An international player in the consulting and technology sectors and a specialist in data and digital technology, Keyrus is dedicated to helping enterprises take advantage of the data and digital paradigm to enhance their performance, facilitate and accelerate their transformation, and generate new drivers of growth and competitiveness.

Placing innovation at the heart of its strategy, **Keyrus** develops a value proposition that is unique in the market and centred around five major service groups, each comprised of multiple solutions:

- Automation and Artificial Intelligence
- Human-Centric Digital Experience
- Data and Analytics enablement
- Cloud and Security
- Business transformation and Innovation

Building on the combined expertise of more than 3,000 employees active across 22 countries and 4 continents, Keyrus is one of the leading international experts in data, consulting and technology.

For more information, visit **https://keyrus.com/**

keyrus
life science

With more than 25 years of experience, **Keyrus Life Science** makes data matter to address the biggest clinical challenges for clients around the globe to enable long-term success. As part of the **Keyrus Group**, we connect **CRO expertise**, life **data science and digital enablement** to improve clinical trial **efficiency** and **agility**, to the utmost benefit of society and human health. As an innovation-centric company by design, we anticipate the future health needs in clinical research and uncover new solutions and technologies to develop faster the cures of tomorrow. With the help of the most advanced partners in the clinical research domain, we provide full services to **optimize patient recruitment and engagement**, to **leverage insights from data**, and to unlock new horizons for **personalized therapeutic approaches**.

For more information, visit **https://keyruslifescience.com/**