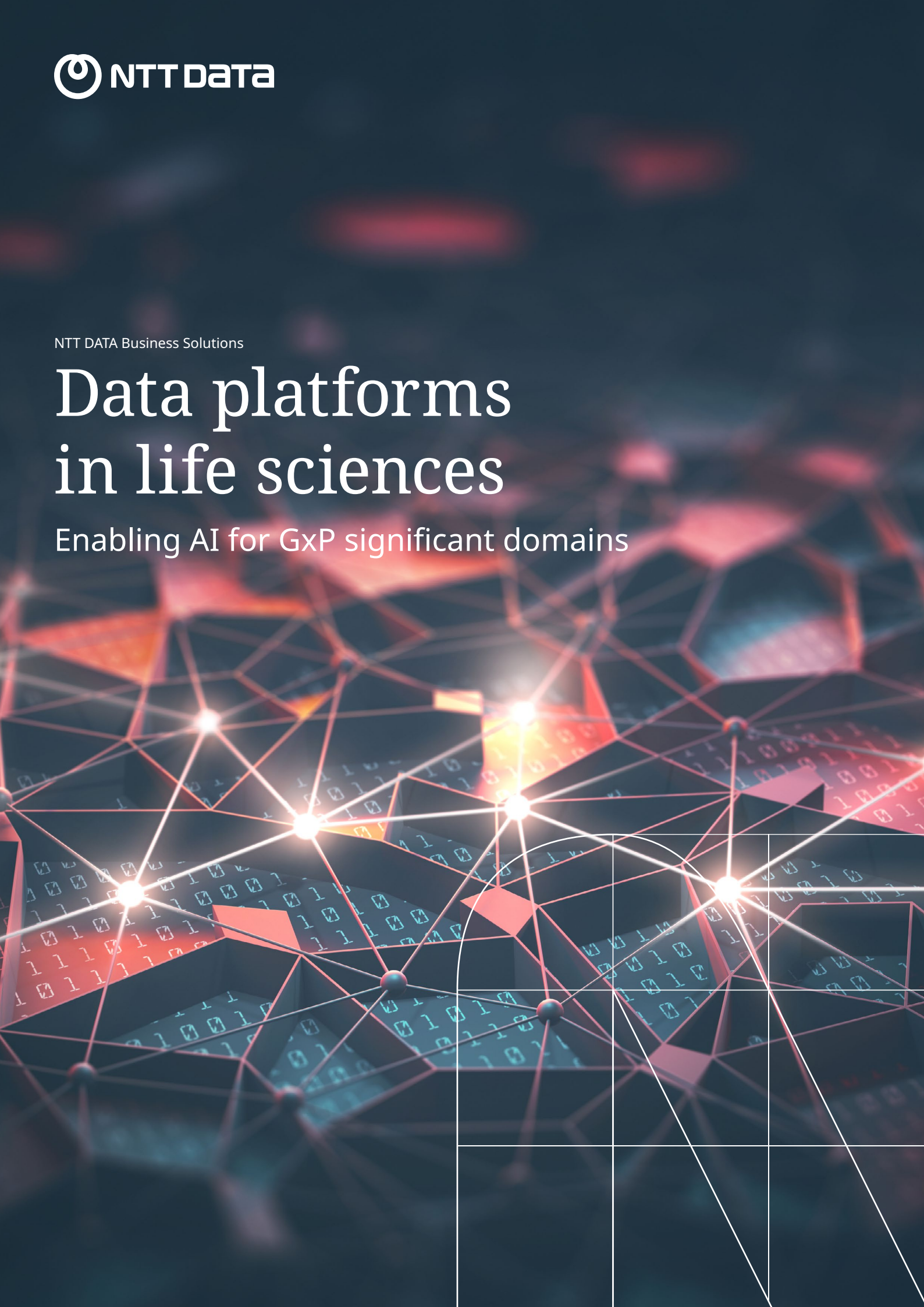
NTT DATA Business Solutions

Data platforms in life sciences

Enabling AI for GxP significant domains



Content

> Every company can embed generative AI with the flick of a switch, however getting value out of it requires a significant cultural change, where employees are data literate, data is treated as a product and where data quality is paramount.

1. Introduction

Big pharma has adapted to evolving technological changes by aligning IT strategy with overall company strategy. These strategies need to be dynamic as rapid advancements in technologies such IoT, AI, graph technologies and hyperautomation lead to larger competitive advantages. IT cannot be seen only as a back office function, but in the forefront of driving the business forward.

As businesses can now collect more data, different data, higher quality, outside of factory walls, we see an explosion in the variety of data we can work with. Combine this with big data analysis capabilities through mathematical methods and new insights can be generated that we have never thought possible. Companies integrating these insights into manufacturing business processes are outperforming pharmaceutical counterparts that do not adopt such technologies (Industry 4.0 for pharmaceutical manufacturing)[1].

Because of the pharmaceutical industry's focus on quality and documentation, this industry is very well suited to perform quantitative analyses for data science. The data is typically of high quality and all processes are well-documented. Currently a lot is captured in paper (data as a burden for auditing purposes), but when digitized the archives containing over 20+ years of data such as documented change procedures can be analyzed and modelled to improve decision making and automation. Examples include the application of graph database in risk assessment, speeding up change procedures by 70% across all sites around the globe.[2]

One of the biggest advancements of this decade: Generative AI. The IDC report on Generative AI in the Life Sciences Industry (Nov 2023) points out that from the 104 of the responding companies, 90% see GenAI as a top priority for their organization, where the average investment for over 50% the companies will exceed a 10% increase. Most of the focus is on applying GenAI for Quality, Risk and Compliance (89%). The biggest barrier seems to be security concerns according to 54% of the respondents. The most important expected benefit is on the employee side: 41% of respondents think they can cut employee workforce up to 10% within 12 months by applying GenAI throughout the company and even 20% in two years (39%)

But what made GenAI such a popular investment opportunity for pharmaceutical companies? What are the requirements to embed GenAI into your own business? What options do you have for integrating all your data? And what should I take into consideration before I embark on a journey implementing AI and GenAI throughout my business? This white paper will dive into these topics and provide a structured answer, supporting you in your preparations to digitally transform the business into the next generation of Pharma companies.

[1] https://www.mckinsey.com/capabilities/operations/our-insights/the-ai-revolution-will-be-virtualized

[2] https://aws.amazon.com/solutions/case-studies/merck/

2.
Drivers

Next-generation pharma companies have benefitted from various drivers which have led to the enablement of these new and promising technologies. Even though risk averseness is one of the key characteristics of the industry, life sciences has always been around experimenting and applying new methods and technologies to create new products and treatments, and to optimize processes such as improving product quality and patient safety of existing ones. To become next-gen pharma, these companies must invest in new technologies to innovate and apply research and experimentation throughout the company to thrive in the world of tomorrow.

Data collection advancements
One of the key developments that led to a cumulative growth of available data, is the utilization of sensor data at scale. Whilst sensors were already common technology in pharmaceutical production plants to keep production parameters in check during batch production, these were mainly utilized in closed loop systems allowing only direct feedback through Product Control Systems (PCS) and alerting when parameters went out of spec.

The internet of things gave an ability to start measuring environmental parameters live, through utilizing battery-powered chips with integrated sensors, utilized outside of production line domains; in warehouses, supply chains and patient wearables. All generating large volumes of data that can be used to derive insights and establish cause-effect relations through statistical modelling.

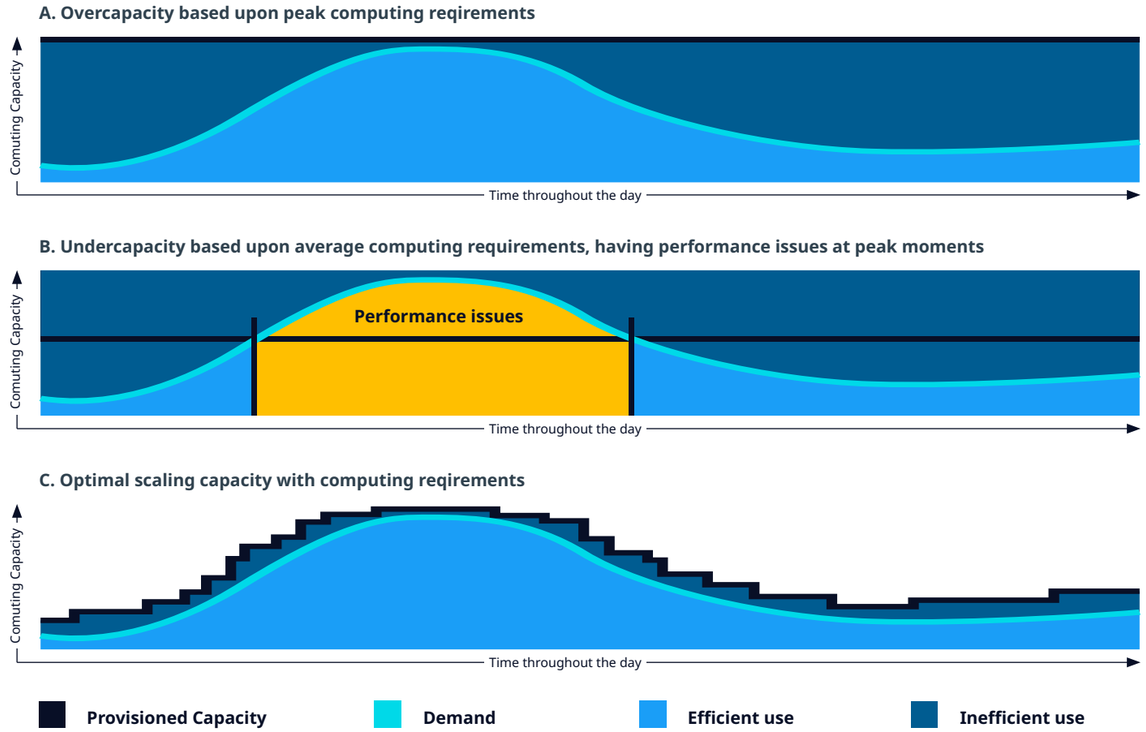
Data variety and volume development
Not only the volume of data has changed, but also its diversity. Where we used to work with structured data in table formats, and semi-structured file formats, we now see the possibility to also work with unstructured data, such as audio files, photo's, video's and genomics files containing the full human genome. Ingestion can be quite challenging as for example genomics files for a full genomics sequence are around 100gb big[3], and of course dedicated processing algorithms are required to make the data interpretable for humans.

Strategically this unstructured data provides a lot of value; for example analyzing production lines through video in order to detect anomalies and defective products, photo's to perform automated inspection of incoming goods and even audio files that can detect audio signals in machine rooms allowing to infer if machinery is at the brink of breaking down, enabling preventive maintenance scenarios.

Hyperscalers
The advancements in cloud computing has given the ability to gather larger volumes of data and the power required for modelling through the use of hyperscalers, who cater for these new use cases in hyperspeed. Amazon Web Services, Microsoft Azure and Google Cloud Platform all offer integrated Platforms as a service, allowing you to bring your IT landscape to the cloud and become increasingly agile, addressing the demand coming from the business; integrating new technologies such as IoT, AI or experimental domains like quantum computing, but also scale storage and computing capacity to the demand at that moment in time. Where on-prem systems were always equipped to cater for the peak load moments, these hyperscaler platforms can scale continuously to the business demand improving your cost efficiency.

[3] https://www.strand-ngs.com/support/ngs-data-storage-requirements

Performance efficiency
For an architecture to perform well and be scalable, it should properly match resource capacity to demand. Traditionally, cloud architectures accomplish this balance by scaling applications dynamically based on activity in the application. Demand for services changes, so it's important for your architecture to be able to adjust to demand. By designing your architecture with performance and scalability in mind, you'll provide a great experience for not only your customers but also customers and business partners while being cost-effective.



Hyperscaler capability to scale capacity with use, balancing performance and costs in an optimal way.

Artificial Intelligence (AI)
Artificial intelligence, or the capability of IT mimicking human capabilities, is a big IT breakthrough. Computers can be equipped with any of the senses that humans have and apply intelligence to what is sensed and transform that into outputs. Including the capability to learn and become more accurate over time. From simple cause and effect relationships purely derived from data, to complete robotic production lines that can fully manage themselves. These AI capabilities require both enormous amounts of data and large clusters of computing resources to build and train models on this data.

GenAI
Gen AI is the next advancement in the AI domain. Revolutionized by products like OpenAI's ChatGPT, computers are now capable of modelling any amount of information and look for the right data, users are looking for, based on prompts via an interface (mainly chat or speech). It can write complete blogs tailor-made to a specific audience, provide detailed instruction on difficult topics like computer programming and help you understand complex matters that previously required a complete academic study. AI tools will significantly enhance all parts of the pharamceutical value chain, such as clinical decision making, risk determination in change procedures, production site regulatory audits and pharmacovigilance..

As public Large Language Models (LLMs) like ChatGPT are using data publicly available on the internet, we now see a shift to applying GenAI to private datasets, such as the ones owned by a company. These kinds of GenAI capabilities are expected to take over most of the cumbersome administrative work and change professional environments significantly. The World Economic Forum had special interest groups focus on this topic, as its potential impact can no longer be denied.[4]

[4] https://initiatives.weforum.org/ai-governance-alliance/home

3.
Use Cases

To understand what these technologies can bring to the pharmaceutical industry, it is good to have a look at some of the use cases created by companies already integrating these technologies into their operation, bringing significant benefits in the GxP domains and across the value chain

Personalized Medicine
The enablement of personalized medicine across a network of various instances to create treatments tailor-made to the patient. Local clinics that can perform genomics analysis to map the DNA, which then move the data into the cloud to translate these sequences into disease variants. These pinpointed variants are then checked against an algorithm to determine the right treatment for the patient. This triggers a job towards a manufacturing facility to create the treatment and ship it to the right person as fast as possible, ensuring a full chain of custody and chain of identity across the chain. None of this was possible, without the immense scaling possibilities and technological offering that the hyperscalers have. An integrated cloud data platform allows to easily share data across actors, with taking privacy data of the patient into account and ensuring everyone is looking at the same data (single source of truth).

Drug Discovery and Production Optimization
Drug discovery is transformed in the same way, where huge amounts of data on compounds allow computers to suggest potentially promising combinations to treat certain diseases. On a process level, AI allows companies to proactively manipulate production line parameters like temperature, humidity, and pressure, ensuring optimal output in terms of yield and quality. This can be attained by implementing predictive processing capabilities or feed-forward production control.

Large Language Models for administrative tasks
LLMs have high potential in revolutionizing the way pharmaceutical companies perform the more administrative tasks. AI can perform 24/7, take away human errors along the way, outperform the workforce as much more parameters can be considered, and allow employees to focus on value-adding activities instead of the cumbersome administrative ones, the impact will be significant. Some very interesting examples contain fully automated pharmacovigilance, automated batch record and transportation documentation, automated risk assessments for change procedures and even apply an LLM as an interface for audits directly retrieve the right documentation from the right system the auditor is asking for.

An integrated cloud data platform allows to easily share data across actors, with taking privacy data of the patient into account and ensuring everyone is looking uniformly at data as a single source of truth.

4.
How to get started

To start working with these next-gen platforms to gain advanced insights, predict and hyper-automate business processes, there are specific requirements that should be met:

Centralize your data
The data required for the designated use cases should be available in a central place in the cloud, either in a structured format through a data warehouse or in semi- and unstructured formats that are stored in a data lake. Both technologies can also be combined to allow for all format types in a data lakehouse. The cloud allows for unlimited storage capacity, on-demand scaling of compute capacity when required and a central place where all parts of the business can access that data. Also not only batch ingestion is possible, but newer data integration techniques include data streaming capabilities such as Apache Kafka

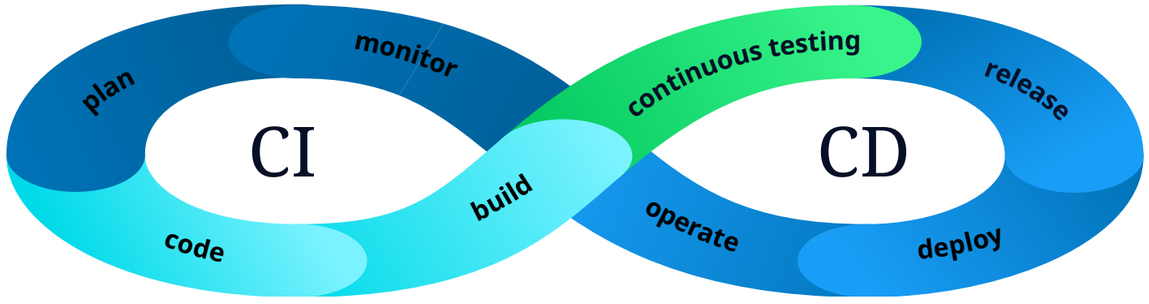
Ingest your data
To ingest, integrate and transform data, monitoring the complete flow, optimizing data quality and consuming data with BI and AI tooling, a technology stack of different components is required. Either single vendor or best-of-breed from various vendors, the set of chosen technologies is combined into a Data Fabric. According to Gartner the Data Fabric is a collection of technologies and different storage types integrated in a single platform

A data fabric is an emerging data management design for attaining flexible, reusable, and augmented data integration pipelines, services and semantics. A data fabric supports both operational and analytics use cases delivered across multiple platforms, processes, and business domains.

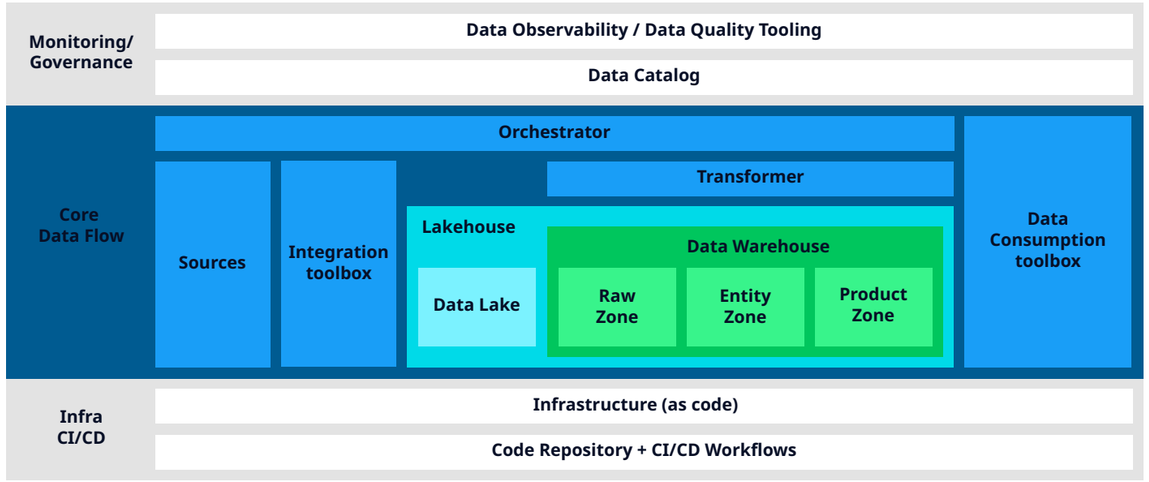
Modelling your data
Next to having the platform in place, methodologies to work with that data should be embedded on the platform. Think of the modelling approach the business will follow, such as the Inmon, Kimball model or a Data Vault model, to ensure flexibility, evolution, and time-travelling in the data fabric. Furthermore, a governance model needs to be put in place, ensuring ownership and responsibilities.[5]

Leveraging the Modern Data Stack
New trends are decentralized approaches like a data mesh and utilizing concepts such as data products and data contracts to understand data requirements certain use cases have. Another important tool is the catalog, containing all metadata, which is key in understanding what data is available in the data platform (some catalogs even integrate data quality and other trustworthiness indicators). Finally change control, especially on data models use in GxP critical workloads, is paramount to allow a controlled way of bringing changes into the platform. For this CI/CD flows with integrated testing and a high degree of automation are required, such as Github Actions.

[5] https://www.gartner.com/en/documents/3901169



A typical data fabric architecture



Options
Simple BI or advanced analytics: the use case drives the architecture, not vice versa

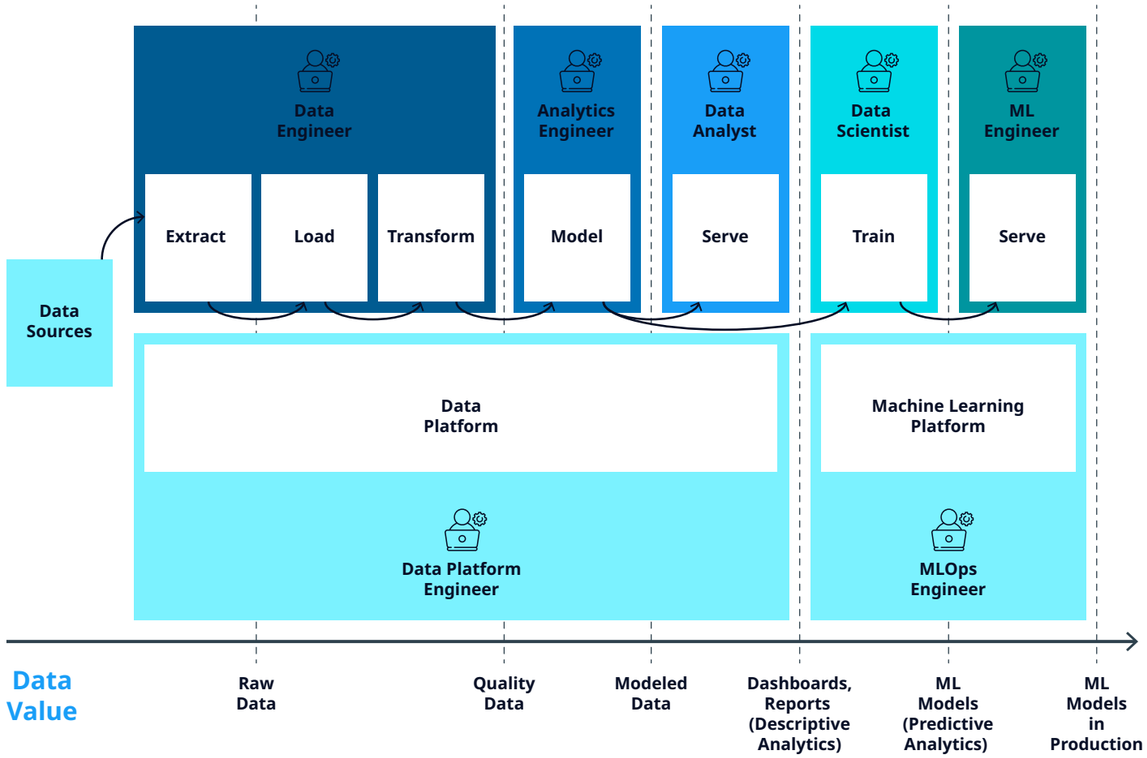
Determining a starting point for the technology stack to deploy always requires to work backwards from the use cases that will be deployed initially. If the initial idea is to start with simple BI dashboards, for example to replace the current BI tooling in a singular stack, you don't need all bells and whistles required to deploy Machine Learning algorithms. Functionalities and capabilities can be added gradually and can evolve over time, so ensure that you only deploy what you need but choose those components that allow architectural evolution in the direction of your vision

Make vs Buy
You can select your own set of technologies and vendor solutions and combine them into a custom data fabric (make) or it is possible to buy complete data fabric offerings in the market. Especially solutions like Microsoft Fabric and SAP datasphere are making waves in the market as one-stop-shop's for all analytical purposes, however you have no control over the functionalities of the stack itself and definitely not the roadmap. If you select your own set of tooling and integrate them into a fabric, you can also adjust components when necessary and reduce vendor lock-in, however these integrations need to be managed and can impose additional points of failure.

Single vendor vs Best of breed
Just like pre-integrated data fabrics in the market, the fabric can be composed of solutions coming from a single vendor. Currently only the hyperscalers cover all services required to compile a complete fabric, but technology vendors like Snowflake are quickly making up with promising roadmaps. But note, that not all solutions available for the modern data stack integrate well and each have their own capabilities. If you want to compose a platform that fits the business requirements, it can be better creating a best-of-breed stack. It is crucial to ensure that it does not transform in an unmanageable set of different tooling, also referred to as Frankenstack

Key considerations here are: costs, requirements coverage, maintainability and cross-component integration capabilities

Example roles required for running a modern data stack



Self-managed vs Fully managed (as-a-service)
Another key element that should be considered to determine the right stack is making the decision of managing the platform yourself within your own IT departments, or that you will leverage the capabilities from a service provider. In the case of self-management, you have full control yourself, over the stack, the roadmap and how it all integrates with IT, compliance, and governance Practices. It can also impose quite a heavy burden because the low-value activities such as database management, implementing patches and organizing in-depth testing in case of updates are all on the account of the pharmaceutical company. By leveraging an IT service provider/system integrator, you can leave these low-value activities up to that provider and only focus on those activities adding value such as generating insights and building solutions on top of the data. Because IT providers typically manage the landscape for multiple customers, they can do it for lower prices as they can create synergy by performing activities across clients and attain economies of scales. Typical examples are implementing updates at scale and providing all the documentation (regression tests, functional tests, training documentation, etc.) to integrate updates into the ecosystem, especially if the throughput times are long to integrate new updates when selfmanaging, the as-a-service option can significantly increase agility, reducing iteration times to implement new technologies for innovation.

Internal Roles vs Service Delivery by suppliers
It is good to understand what roles on the data platform will be covered by internal resources and for what roles you need a partner to cater for. As the modern data stack requires different capabilities in the IT landscape, it is good to understand which roles are required for your current stack. Depending on what roles you will be covering in-house, you can determine the right tooling that fit the job, skills, and preferences of those employees. It will be a key driver in the determining the technical requirements of the platform and impact the solution selection procedure.

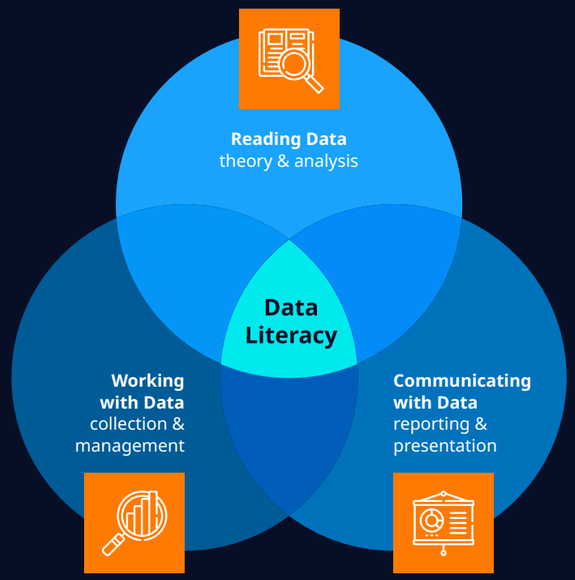
5. Considerations

Everything should be done in the right order, as you cannot start deploying LLMs and expect valuable insights from it tomorrow:

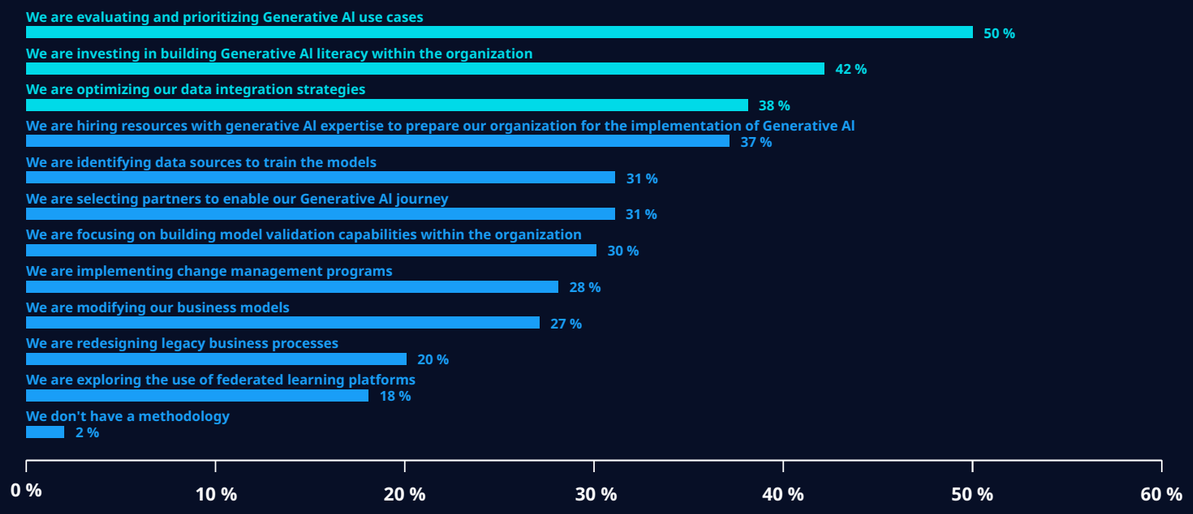
The first step in the journey is determining your data strategy. The strategy should be aligned with the business strategy and corresponding goals, ensuring that the data/IT strategy will support attaining those goals. A data strategy allows you to determine the direction you want to grow your data stack into, allowing you to create strategic roadmaps with initiatives that logically add up to attain that future stack.

The organization should be made ready to become data-driven, as it is not just a technological matter. Employees should understand how insights are derived, how algorithms are created and work in practice, how data is modelled and how dashboards are created, to ensure fruitful collaborations between technical IT experts that are able to create the insights, and the business domain experts requiring these insights. This is also called data literacy and is a key requirement to shape the organizational culture into a data-driven one.

What is Data Literacy?



How is your organization currently ensuring the maximization of the business value /ROI for its Generative AI initiatives?



Use case identification, followed by building literacy and optimizing data integration strategy were key to optimizing ROI for Generative AI initiatives currently

Infra

When the data strategy is in place, and data literacy programs are initiated, the infra selection procedure can start. Work from technical and user requirements as a starting point and cater for vendor selection procedures based on the organizational preferences. Important considerations to be made related to infra are:

Security-by-design
As for cloud platforms, security is one of the key aspects. You want to ensure the data in transit and at rest cannot be eavesdropped or stolen through techniques such as encryption and routing the data over networks isolated from the internet. The setup of the cloud virtual network into subnets, usage of network security groups and firewalls and integration of 2-Factor Authentication (2FA) for all tools are crucial examples. From an architectural point of view these security techniques can be embedded in the infrastructure; also known as security-by-design.

Infrastructure-as-a-code
The modern cloud infrastructures provide the possibility to declaratively manage the infrastructure as a code, allowing to ensure version control and CI/CD procedures for infra elements. This allows for standardization, rollbacks, and collaboration for an effective data stack. Typically done through tools such as Terraform or hyperscaler native tools like Azure Resource Manager (ARM/ BICEP templates). Allowing for self-service deployment of pre-approved infrastructural components and having pro-active compliance checks in place, such as allowed data center regions, storage types and having RBAC enabled.

Control your data
Invest in the right tooling to ensure you can keep track of your data from source to consumption. How is data ingested and transformed and have checks in place if this was performed successfully. For any reason, pipeline jobs can fail and require direct alerting on those use case consuming that data, to prevent decision are made on incorrect/outdated data. Observability tools can alert on data problems, notify the right people, and allow data specialists to easily find the root cause of the problem and restore the data in a trustworthy state as soon as possible. Especially in GxP contexts this is a functionality that should not be overlooked.

Data Quality and Integrity

Data Quality is paramount, period. Your outputs and insights are as good as the data they use to derive these, and therefore data quality is a key pillar for modern data platforms. Data quality obviously has many elements, such as completeness, uniformity, and referential integrity, and all should be covered to ensure trust in the data across the business. In the extension of data quality, we find data integrity (ALCOA+) which is one of the focus points of the regulatory bodies in life sciences (EMA, FDA, etc.). Data platforms should be able to profile the data in it, specify what the estimated quality is, indicate the data owner and show where the data is generated to fix data quality problems at the source. This can not only be solved by the platform itself, as fixing the data is typically a job that must be performed at scale through dedicated data remediation programs. But only if the data is of the right quality,it makes sense to attain insights from it. If decisions are made based on incorrect data, the effects in a life sciences context can be detrimental. Also, if the data quality is low, and data is not trusted, business users will not adopt the platform and its tooling and will refrain from becoming data driven. Furthermore don't overlook the challenge of incorporating privacy data from customers and/or patients. Regulations are very strict and sometimes require decentralizing storage of data to ensure data is not stored or processed in areas that fall outside of the jurisdiction of that regulation. Especially GDPR and HIPAA are interesting, but also the effect of the 'EU-US Data Privacy Framework' enforces a strategic approach to privacy-related data.

Leverage supplier expertise

Where possible make use of the expertise of your IT service provider. As these new landscapes will become increasingly complex, you cannot keep track of all the details and developments in the market if you are not a big pharma company with deep pockets, in-depth knowledge of building and qualify the platform, and complete departments that run them whilst maintaining the qualified state. Also, more and more new roles come into existence with advent of these new technologies, such as the prompt engineer role for GenAI, which can be hard to fulfill. Either because there is little supply or no full-time requirement for these kinds of roles. By making use of your suppliers to manage the qualified platform including the provisioning of documented proof, the employees of the pharmaceutical companies can focus on value adding activities such as insights generation. Utilize the knowledge, experience and expertise of your IT service provider to make the right decisions, and establish a strategic partnership to grow together.

The more custom the algorithm, the more competitive the advantage

Obviously, a lot of common algorithms are commercially available in the market. Deep vision models that can identify photo or video content, text algorithms allowing to extract structured data from documents, and predictive algorithms that can help you prevent machinery breakdown within production facilities. However, our experience is that the most value can be gained from custom created/tailor-made algorithms. For example, improving production line yield is highly depending on the product that is produced, raw materials utilized, environmental parameters in production facilities and much more. If you can model this in your own dedicated model, you might be able to significantly increase production yield, which can easily justify the investment in the algorithm. By having these dedicated algorithms in place, a true competitive advantage can be established.

Start collecting now, but leverage your history

Pharmaceutical companies have a high potential of leveraging data, as a lot of aspects are documented according to strict standards. Think of production line data, quality control tests, change forms, batch records, risk assessments, training records, maintenance forms, and much more. Start collecting all that data right away, as you can analyze it later, which prevents the growth of a gap with competitors that already do so. Unfortunately, the industry is infamous for executing a lot of procedures still on paper, which does not allow to analyze this data at scale as it is now available in a digital format. Therefore, companies should strive to digitize paperwork archives, sometimes containing over 20 years of data on specific processes and/or procedures. Change from having paper data as a burden purely for auditing into gathering value from this data and turning this into insights.

6. Importance of GxP focus from the start

Pharma companies should utilize data platforms to get insights on GxP-critical information to drive actions, incorporating this aspiration in their strategy determination phase. As this area requires the highest compliance standards such as platform qualification and use case validation, it should be considered from the start. GxP domains contain the most valuable use cases, such as procurement, manufacturing, quality and maintenance, with examples such optimizing yield (and thus revenue) and the reduction of compliance costs. If GxP is not considered, and the data platform is only established for non-critical areas, it is nearly impossible to incorporate GxP workloads at a later point in time.

As data integrity is a key focus point in the industry, it should already be considered at the design phase, which we call data integrity by design. Understanding the data lifecycle and ensuring checks to cover the ALCOA+ requirements including documented proof that the data is necessary to convince auditors that you are in control of your data estate. Without a clear practice to manage data integrity, and indirectly data quality, there is no foundation to perform GxP and business-critical critical actions on a cloud data platform. The ISPE GPG on enabling innovation states the following:

This also includes data security aspects, such as ensuring that all data in transit and at rest is encrypted, but also have procedures in place when problems with the data occur (like disaster recovery or in case of a data breach).

> “ Regulated companies must ensure that appropriate procedural and technical controls are in place to ensure data integrity is maintained throughout the data lifecycle.
>
> **(Paragraph 4.5.3)**

Obviously by integrating all data within, but also outside of the pharmaceutical company’s walls, there will be a lot of data and technical components present that will not be used in GxP critical actions, but in other non-critical use cases. As the management of GxP use cases requires tight control, a good understanding of the risks, effective change management, extensive documentation and strict monitoring on compliance, applying these practices to all elements of the data platform is overkill. Doing this all elements in the platform is overkill. Ensure that you have mechanisms in place that can indicate if certain data or technical components are leveraged in GxP-critical use cases, and only apply tight qualification and validation procedures to these elements. All else can take the shortcut in case of development and/or changes, and don’t have to follow strict regulatory guidelines. This will enlighten the burden of aspects like documentation significantly in areas where it does not make sense.

This whitepaper highlights the complexity of using critical data in a regulated environment. With a digital mindset, a step by step approach can be developed to become data driven organization and start to benefit from the advantages rapidly evolving technology can bring. NTT DATA Business Solutions would be happy to discuss the strategy you require to take you on this journey.

Links to existing blogs:

GxP Data lake overall:

Driving Pharma 4.0: Operationalizing Your GxP Data Lake →

Machine learning for Pharma blog series:

Machine Learning for Pharmaceuticals | Blog Series | Part 1 →

Machine Learning for Pharmaceuticals | Blog Series | Part 2 →

Machine Learning for Pharmaceuticals | Blog Series | Part 3 →

IoT:

How IoT Disrupts Manufacturing in Pharma →

How IoT Revolutionizes Traditional Medicine →

About NTT DATA Business Solutions

We Transform. SAP® Solutions into Value

We understand the business of our clients and know what it takes to transform it into the future. At NTT DATA Business Solutions, we drive innovation - from advisory and implementation, to managed services and beyond, we continuously improve SAP solutions and technology to make them work for companies – and for their people.

nttdata-solutions.com

NTT DATA

