LEARNING MADE EASY

Cloudera® Special Edition

Accelerating the Pharmaceutical Value Chain with AI

for dummies®
A Wiley Brand

Build a pharma knowledge graph

Secure and govern your data

Operationalize AI across the value chain

Brought to you by

CLOUDERA

Jeremiah Morrow
Rameez Chatni, PhD

About Cloudera

Cloudera is the only hybrid data and AI platform company that large organizations trust to bring AI to their data anywhere it lives. Unlike other providers, Cloudera delivers a consistent cloud experience that converges public clouds, on-prem data centers, and the edge, leveraging a proven open-source foundation. As the pioneer in big data, Cloudera empowers businesses to apply AI and assert control over 100% of their data, in all forms, improving security, governance, and real-time and predictive insights. The world's largest brands across all industries rely on Cloudera to transform decision-making and ultimately boost bottom lines, safeguard against threats, and even save lives. Learn more at cloudera.com.

Accelerating the Pharmaceutical Value Chain with AI

Cloudera® Special Edition

by Jeremiah Morrow and Rameez Chatni, PhD

for dummies®
A Wiley Brand

Accelerating the Pharmaceutical Value Chain with AI For Dummies®, Cloudera® Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-394-47619-0 (pbk); ISBN 978-1-394-47620-6 (ebk); ISBN 978-1-394-47621-3 (ebk)

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor/Project Manager: Rebecca Senninger

Acquisitions Editor: Traci Martin

Editorial Manager: Rev Mengle

Business Development Representative: Jeremith Coward

Production Editor: Tamilmani Varadharaj

Table of Contents

Introduction

Pharmaceutical and life sciences companies are under increasing pressure to bring new therapies to market faster while reducing costs and wasted resources, improving the probability of success, and maintaining rigorous standards for quality, safety, and compliance. AI has the potential to fundamentally transform the way pharma companies operate across the entire value chain, from research and clinical development to manufacturing and commercialization. But realizing that potential is difficult when crucial insights are hidden in fragmented data and disparate systems. To capitalize on the AI opportunity, pharma companies need to unify their data and provide the context AI needs to deliver accuracy, consistency, and value.

About This Book

In this book, I explore how pharmaceutical companies can build the foundation required to accelerate the value chain with AI. I discuss how to connect and contextualize fragmented data, how to establish security and governance in highly regulated environments, and how to apply machine learning, generative AI, and agentic AI to the right use cases. By the end of this book, you should have a practical roadmap for connecting your data and delivering AI outputs that improve productivity, accelerate decisions, and ultimately bring therapies to patients faster.

Icons Used in This Book

Throughout the book, I use icons to indicate special information. Here's a guide to what those icons mean.


TIP

The Tip icon indicates information that you can apply to your own projects to make them work better. Tips can save you time and help you avoid frustration.

REMEMBER

The Remember icon highlights information that's worth retaining after you put down this book.

WARNING

The Warning icon is for serious situations, where if you don't take care, you can cause harm to your work.

Beyond the Book

This book introduces you to important AI concepts for pharma and life sciences. The following resources complement and build on that knowledge.

- Podcast: Discover the life sciences blueprint for scalable AI: `https://www.healthcareitnews.com/podcast/discover-life-sciences-blueprint-scalable-ai`.
- Blog: Navigating GxP Compliance in the Age of Precision Medicine and AI: `https://www.cloudera.com/blog/business/navigating-gxp-compliance-in-the-age-of-precision-medicine-and-ai.html`.
- Webinar: How to Implement AI in Healthcare Responsibly and at Scale: `https://www.cloudera.com/content/dam/www/marketing/resources/webinars/how-to-implement-ai-in-healthcare-responsibly-and-at-scale.landing.html`.

IN THIS CHAPTER

- Accelerating drug development
- Use cases for AI in pharma

Chapter 1
The AI Transformation in Pharmaceuticals

Pharmaceutical and life sciences companies are facing a critical inflection point. Bringing new therapies to market has never been more complex, expensive, or data intensive. Research and development (R&D) costs continue to rise, clinical trials require increasingly specialized patient populations, and precision medicine depends on understanding relationships across enormous volumes of scientific and clinical information.

At the same time, much of the pharmaceutical value chain still operates across fragmented research systems, clinical applications, manufacturing environments, regulatory repositories, and commercial platforms. Teams often spend significant time finding, reconciling, and interpreting information before they can act on it.

Artificial intelligence (AI) has the potential to improve processes and accelerate workflows across the value chain. By helping scientists, clinicians, regulatory teams, and business leaders analyze more information and make better decisions earlier, AI can accelerate work and produce better outcomes. But AI can only deliver those benefits when it is built on a foundation of trusted, connected, secure, and governed data.

Reducing the Drug Development Timeline

Bringing a new therapy to market can take 10 to 15 years. During that process, a promising scientific discovery must progress through target identification, preclinical development, multiple clinical trial phases, regulatory review, manufacturing, and commercialization.

To accelerate the route to market, pharmaceutical companies must:

- **Fail faster.** Failure is an unavoidable outcome of R&D, but the cost of failure increases dramatically as a candidate moves through development. Identifying an ineffective or unsafe compound during discovery is far less costly than discovering the same problem after years of clinical development. Better use of data can help researchers identify promising candidates earlier and discontinue weak candidates before committing additional time and resources.
- **Optimize clinical trials.** Clinical trials are one of the most costly aspects of drug development, and they are getting increasingly difficult to manage. Recruiting eligible participants can take months or years. Sponsors must identify trial sites, manage enrollment, monitor safety, respond to protocol changes, and maintain accurate regulatory documentation. Connecting clinical, operational, and real-world data can help teams reduce costs and manage the complexities associated with clinical trials.
- **Enable precision medicine.** Instead of developing therapies for broad populations, researchers increasingly need to understand how genetic, environmental, clinical, and lifestyle factors affect smaller groups of patients. Finding these relationships may require analyzing millions of scientific and clinical records and connecting information that spans multiple systems and datasets.

AI can help pharmaceutical companies manage these complexities at a scale that would be impossible with manual analysis alone. Models can analyze large datasets and help experts make decisions sooner.

But to do that, AI needs access to data that is accurate, connected, secure, and governed. Pharmaceutical companies must create a foundation of trusted data that enables AI to securely access the context it needs to help organizations shorten development timelines, reduce the cost of failure, improve clinical trial execution, and bring increasingly targeted therapies to patients faster and more safely.

Applications for AI across the Pharma Value Chain

AI can improve work across the entire value chain. The greatest value often comes from combining AI's ability to process information with the expertise of the people already performing the work. Here are a few areas where AI can help:

- **Drug discovery: accelerating research.** Drug discovery teams analyze enormous volumes of scientific literature, experimental results, genomic data, compound information, and disease research. Generative and agentic AI help researchers spot opportunities that would take months to find manually. An agent can federate queries across targets, pathways, and disease tables through the knowledge graph and deliver a ranked, evidence-linked shortlist to researchers.
- **Medical affairs: preparing Medical Science Liaisons (MSLs) faster.** MSLs must walk into a conversation with a physician fully briefed. They must understand trial data, competitive landscape, and prior interactions. Generative AI can assemble this briefing automatically, pulling only entities that are relevant to that physician and compound, cutting preparation time from hours to minutes.
- **Pharmacovigilance: safety reporting, attribution, and case analysis.** Pharmacovigilance is one area where grounding matters most because errors can have serious clinical and regulatory consequences. AI can triage incoming adverse event reports, checking each against the resolved data for attribution, and draft the case analysis and regulatory submission. Every step should be checked against

resolved, entity-linked data, rather than general training data, and human oversight is the key to building confident automation with this high-risk workflow.

- **Manufacturing: batch records and yield enhancement.** Manufacturing generates enormous volumes of data, including batch records, deviation reports, and sensor logs. AI can review batch records for anomalies, cross-reference deviations against historical patterns, and surface likely causes of yield variance. Yield calculations and process control can and should stay in deterministic systems, while AI does pattern spotting.
- **Clinical trials: clinical trial optimization.** Clinical trials represent the largest R&D cost for pharmaceutical companies, which makes AI-driven optimization an especially attractive opportunity. Graph-grounded AI can identify optimal trial sites from historical enrollment performance, flag patient populations likely to meet inclusion criteria, and detect protocol deviations earlier. Automation can reduce timelines without compromising rigor.
- **Real-world evidence.** Real-world evidence, which includes claims, electronic health records, and patient registries, is messy, unstructured, and enormous. Extracting entities and grounding them in canonical vocabulary turns what are initially scattered signals into a queryable resource, enabling teams to see how a drug performs outside of a trial's controlled conditions.

IN THIS CHAPTER

- Understanding what a knowledge graph is
- Giving AI accurate and consistent context
- Building a knowledge graph

Chapter 2
Unifying the Pharma Value Chain with Knowledge Graphs

Every pharmaceutical company owns a mountain of data. That data lives in hundreds of disconnected systems that speak different languages. A compound may have one name in a research system, another identifier in a clinical trial platform, and a third name in a regulatory submission. A scientist may understand that these records all refer to the same therapy, but a traditional data system or an AI model might not.

A knowledge graph connects those records and captures the relationships among them. It gives scientists and business teams a consistent map of the pharmaceutical value chain, and it gives AI the context it needs to interpret organizational data accurately.

What Is a Knowledge Graph?

A knowledge graph organizes data around relationships instead of rows and columns. Instead of rigid tables, it connects individual pieces of data, called *entities*, through labeled relationships describing how they relate. It's more of a map than a spreadsheet.

In pharma, entities might include a compound, a trial, a patient population, a gene, or an adverse event, and the relationships between them carry real meaning: A compound targets a gene, and a trial tests a compound. This web of connections (Figure 2-1 shows an example) makes a knowledge graph a powerful source of context for AI, which can capture the meaning behind a connection, and analyze relationships rapidly at scale.

REMEMBER

A knowledge graph isn't just a database. Databases answer questions the user anticipated when designing them, while a knowledge graph answers questions the user hasn't thought of yet, because relationships carry the meaning, not the schema. It also shows relationships across traditionally siloed organizations, spanning research, clinical development, manufacturing, regulatory affairs, safety, and commercialization; something a traditional database often can't do.

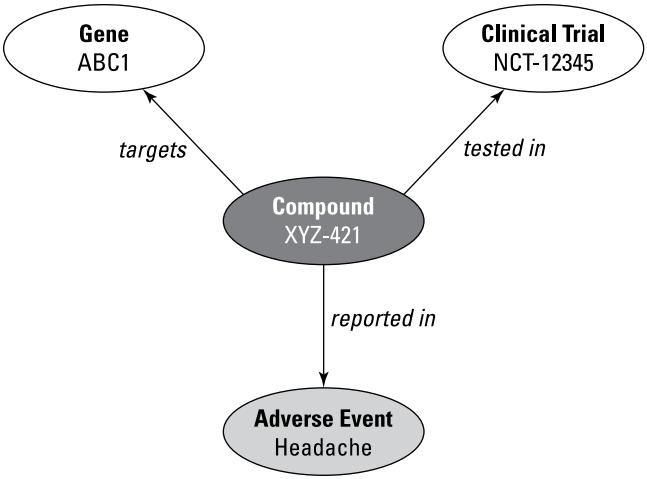

FIGURE 2-1: A simple pharma knowledge graph showing a drug compound connected to a target gene, a clinical trial, and a reported adverse event.

How a Knowledge Graph Supports Accurate and Consistent AI

AI models are only as trustworthy as the data and context they receive. LLMs generate fluent, confident text, but fluency isn't the same as accuracy. Models can hallucinate (they can make statements that sound plausible but aren't true). In pharma, the consequences of a hallucination can be much greater than a minor inconvenience.

Grounding AI in facts

Retrieval-augmented generation (RAG), a common method of providing context to an AI model, pulls chunks of relevant text and feeds it to the model as a reference. But unstructured text is often messy: A paragraph might say a compound was not associated with an adverse event, and a model skimming fragments can lose track of that negation, flipping "did not cause" into "caused."

A knowledge graph solves this challenge by storing facts as explicit relationships, not sentences that can be left up to interpretation. The graph simply doesn't create a relationship that isn't true. RAG built on a knowledge graph, or graph-grounded RAG, is far more precise than RAG built on text chunks alone. Chapter 4 compares graph-grounded RAG to other AI approaches in more depth.

Knowledge graphs deliver accurate and consistent AI with the following capabilities:

- **Verified connections replace guesses.** The AI checks against established relationships instead of predicting what sounds right.
- **Negation is handled explicitly.** A relationship either exists in the graph or it doesn't.
- **Traceability comes built in.** Every fact links back to a source system, so users can always trace an answer to where it came from.
- **Consistency across the organization.** Every team queries the same underlying map, so everyone gets the same answer.

WARNING

Without a knowledge graph or similar grounding, an AI model has no way of knowing whether, say, the compound it just mentioned even exists. Never deploy genAI in a regulated pharma workflow without grounding in verified data.

Finding new connections

Although accuracy and consistency are critical for AI success in pharma, a knowledge graph's real super power is inference. It can surface relationships that were never directly stated by reasoning across the ones that were.

Consider drug discovery: If the graph knows Compound A inhibits Gene X, and Gene X is overexpressed in a rare pediatric cancer, it can infer Compound A may be worth investigating for that cancer even though no document directly states that.

The same power applies elsewhere. In manufacturing, a graph might infer that a quality deviation at one facility and a supplier's material substitution are related, despite being logged in separate systems. In rare diseases, where evidence is thin, a graph can infer off-label activity worth investigating, such as noticing that physicians treating a rare condition have independently been prescribing a drug approved for something else.

One map as the source of truth

Pharma value chains cross dozens of systems: Laboratory information management systems (LIMS) contain experimental and laboratory results. Clinical trial management systems (CTMS) track trial operations. Enterprise resource planning (ERP) platforms manage manufacturing and supply chain processes. Other applications support safety, regulatory affairs, quality, medical affairs, and commercialization. Each system often has its own name for the same drug, trial, or patient population.

Unifying these different names into one consistent identity is the first step towards building a working knowledge graph. When CTMS calls a compound "XYZ-421" and regulatory calls it "Compound B, Study 4471," entity resolution recognizes them as the same entity. That step happens at the data layer, before a graph is built. Entity resolution enables an AI model to reason across the whole value chain instead of being trapped inside one silo.

Speaking a Common Language

Entity resolution establishes that different records refer to the same thing. An ontology establishes what kinds of things and relationships exist in a particular domain.

An *ontology* is the vocabulary and structure behind a knowledge graph. It defines concepts such as compounds, genes, diseases, trials, batches, sites, and adverse events, along with the relationships that connect them.

The pharma value chain contains several domains with distinct vocabularies. Research and development teams think in terms of targets, compounds, assays, and pathways. Clinical development teams work with protocols, cohorts, sites, and enrollment. Manufacturing teams focus on batches, yields, deviations, and facilities.

Forcing every function to agree on one complete vocabulary from the beginning can derail a knowledge graph before it's ever built. A more practical approach is to start locally and connect globally:

- **Start with the vocabulary the use case needs.** A drug discovery use case may need to define compounds, targets, genes, diseases, pathways and assays. It doesn't need to model every concept used in commercial operations or supply chain management.
- **Use established standards where they fit.** Healthcare and life sciences already have controlled vocabularies and identifiers for many common entities. Organizations can align relevant concepts to established standards when doing so improves consistency and interoperability.
- **Not every internal term needs an external equivalent.** A manufacturing concept may have no meaningful counterpart in a clinical vocabulary. Don't force a mapping simply to create the appearance of one universal language.
- **Connect domains through shared entities.** Different domains don't need identical vocabularies to work together.

  Research and manufacturing teams may describe their work differently, but both may refer to the same compound. Once a cross-functional team has assigned a consistent identity to a compound, users can follow it from target discovery through clinical development and manufacturing, even when the surrounding terms are different.

TIP

Don't build every vertical's ontology at once. Select one valuable business question, define the entities and relationships needed to answer it, and expand after proving the approach.

Building a Knowledge Graph

The goal of building a knowledge graph shouldn't be to connect every pharmaceutical system immediately. The more practical path forward is to start with a use case where better access to connected information can create measurable value.

The following steps provide a practical path.

1. **Start with a high-value question.**

   A knowledge graph should begin with a business or scientific problem. Potential starting questions include:

   - Which compounds targeting a particular pathway have shown promising efficacy without significant safety signals?
   - Which trial sites are best suited to recruit patients with a specific biomarker?
   - Which manufacturing deviations may be connected to changes in raw materials or suppliers?
   - Which previously investigated therapies may be candidates for a rare disease?

   The question determines which data, entities, and relationships matter.

   A narrow first use case also makes success easier to measure. Teams can evaluate whether the graph improves research time, increases the quality of results, reduces manual reconciliation, or accelerates a specific decision.

2. **Define the entities and relationships.**

   For a compound-to-disease discovery use case, entities might include compounds, targets, genes pathways, diseases, assays, and studies. Relationships might include targets, inhibits, participates in, associated with, and evaluated in.

TIP

Keep the first model focused. You can add more entities and relationships as the use case develops.

3. **Connect the relevant data.**

   Identify the structured and unstructured sources needed to answer the question.

   Structured sources may include laboratory, clinical, regulatory, safety, or manufacturing databases. Unstructured sources may include study reports, scientific publications, protocols, regulatory documents, and laboratory notes.

   For documents, extraction tools can identify relevant entities and relationships and convert them into structured information that can be connected to the graph. All this data should ideally be secure and governed, a topic I cover in greater depth in the next chapter.

4. **Standardize and resolve entities.**

   Different sources will use inconsistent names and identifiers. Those records must be mapped to consistent entities before users or AI models can reason across them reliably.

   Entity resolution may combine established identifiers, controlled vocabularies, matching rules, and human review.

   Don't force uncertain matches. If the system can't confidently determine whether two records refer to the same entity, route the result to a subject matter expert rather than guessing. A false match can contaminate relationships throughout the graph and create misleading results.

REMEMBER

   A knowledge graph is only as trustworthy as the identities and relationships beneath it.

5. **Preserve provenance.**

   Every important entity and relationship should retain information about where it came from.

   Provenance enables a scientist, regulator, quality professional, or AI user to trace an answer back to its supporting data. This traceability is especially important when knowledge graphs support regulated workflows or provide context to AI. The graph shouldn't merely say that a compound is associated with an adverse event. It should help users find the clinical record, safety report, or study from which that relationship was derived.

6. **Choose how the graph connects to data.**

   Organizations have more than one way to implement a knowledge graph.

A virtual knowledge graph maps entities and relationships to data that remains in existing systems. The graph acts like an intelligent map, showing what the data means and where it can be found without requiring every record to be copied into a new repository. This approach can help organizations connect distributed data, preserve source-level governance, and reduce the operational burden of maintaining multiple copies.

Other use cases may benefit from storing entities and relationships directly in a dedicated graph database. A materialized graph can provide faster performance for frequent or highly complex relationship queries. However, it also requires pipelines to populate the graph, processes to keep it synchronized, and infrastructure to operate it.

Neither option is automatically right for every workload. Start with the architecture that solves the business problem with the least unnecessary complexity. Materialize relationships when performance or repeated use justifies the additional operational work, not just because a dedicated graph sounds more sophisticated.

7. **Expand incrementally.**

Once the first use case demonstrates value, add adjacent sources, entities, relationships, and domains.

A discovery graph may expand to include clinical outcomes. A clinical graph may connect to safety and regulatory information. A manufacturing graph may incorporate supplier and quality data.

Over time, these domain-specific graphs can connect through shared entities. The result is a connected view of the pharma value chain built through useful projects and organizational value.

IN THIS CHAPTER

- Protecting sensitive data across the pharma value chain
- Maintaining GxP compliance
- Deploying explainable AI

Chapter 3
Securing and Governing Your AI

Pharmaceutical companies manage large volumes of sensitive and valuable data. Patient records, genomic information, clinical trial results, manufacturing processes, regulatory submissions, and intellectual property (IP) all require high degrees of protection. Researchers and business teams also need access to information to discover new therapies, conduct trials, maintain product quality, and monitor patient safety.

AI makes balancing safety with innovation even more difficult. Models perform best when they have access to large volumes of current, relevant data from which to draw context. But giving models broad access to data without sufficient controls can expose protected information, compromise IP, introduce unreliable results, and create regulatory risk.

Security and governance can't be bolted on as an afterthought. It must be embedded into the data architecture, and protect the entire data and AI lifecycle, from ingestion at the edge to model training and development at the core to consumption.

Understanding the Challenge

Security and governance are challenging in almost every large organization, but the pharmaceutical industry adds several additional layers of complexity.

The pharma value chain spans laboratory research, preclinical development, clinical trials, regulatory affairs, manufacturing, supply chains, medical affairs, and commercial teams. Each function uses point solutions that generate data in different formats. That information may reside in public and private clouds, data centers, edge and Internet of Things (IoT) devices, and systems operated by external partners.

That data may also contain personally identifiable information (PII), protected health information (PHI), intellectual property (IP), product quality and patient safety information, and other forms of sensitive data that must be protected.

Security and governance policies that are based only on what system the data resides in is not sufficient. Pharmaceutical companies may know that a solution is protected inside its corporate environment, but as data moves and as is consumed by AI, firms must be able to answer several questions:

- Who (and which model) is allowed to access each dataset?
- Has sensitive information been masked or anonymized?
- Where did the data originate?
- How has the data changed?
- Which systems and AI models are consuming the data?
- Can that data legally move between regions?
- Which decisions have been made with the data?

Meeting GxP Requirements

GxP is an umbrella term for the good practice requirements that help pharmaceutical companies ensure the quality, safety, integrity, and reliability of their work. The *x* changes depending on the activity. Good laboratory practice applies to certain laboratory work, good clinical practice applies to clinical trials, and good manufacturing practice applies to pharmaceutical production.

Although requirements vary by process and jurisdiction, GxP environments share an important principle: work that affects product quality, patient safety, or regulatory compliance must be performed consistently and documented reliably.

For electronic systems, records must remain trustworthy throughout their lifecycle. Organizations need controls governing access, changes, approvals, retention, and electronic signatures. They must also maintain audit trails that show what happened, when it happened, and who was responsible.

Data integrity is critical for GxP compliance, and pharmaceutical companies must ensure the completeness, accuracy, and consistency of data across the value chain.

AI use places a greater burden on organizations to explain every output. For models that influence critical functions like clinical eligibility, product quality, or regulatory submissions, pharma companies need to document the model version, prompts, the data that provided the context, and any human approvals associated with those decisions.

REMEMBER

GxP compliance is about demonstrating that a pharmaceutical company's processes, systems, and data are reliable, and AI governance is an important part of compliance.

Building a Unified Data Fabric

Pharmaceutical data rarely lives in a single repository. Scientific workloads may run in one cloud, clinical systems might be in another cloud, and manufacturing systems may live on premises. Privacy and sovereignty requirements may also prevent certain datasets from leaving a country or region.

Instead of forcing all the data into a single location, pharmaceutical companies need a consistent layer of security, governance, and metadata that works across distributed environments. That layer is called a *unified data fabric.* A unified data fabric connects data across different infrastructures while providing common controls for how that data is discovered, accessed, protected, and used. It enables pharmaceutical companies to apply security and governance consistently without eliminating the systems that individual functions rely on. It consists of several important capabilities:

- **Connecting and unlocking data everywhere.** Pharmaceutical data spans clouds, data centers, edge environments, specialized applications, and external partners and third-party databases. A unified data fabric

connects those distributed sources without requiring every workload to run in the same place. It also supports data in motion, including real-time and streaming data, so information can be captured and made available as events occur instead of only after it lands in a repository. This enables pharmaceutical companies to combine historical data with fresh operational data for analytics and AI.

- **Adding business context.** Connected data is only useful if people and models understand what it means. A unified data fabric uses metadata, catalogs, and knowledge graphs to identify what data represents, who owns it, how it relates to other information, whether it is current, and how it should be used. This helps users discover relevant data, and helps AI retrieve and interpret it accurately.
- **Applying consistent security and governance.** Security and governance policies should apply consistently wherever data resides and moves. Identity and access controls determine which users, applications, models, and agents can access specific information, while encryption, masking, tokenization, and anonymization protect sensitive data. Classifications and permissions should remain associated with data as it moves across environments instead of being rewritten. This consistency reduces complexity and risk.
- **Tracking end-to-end lineage.** End-to-end data lineage shows where data originated, how it's ingested, how it moves and transforms, and how it's consumed for analytics and AI. For AI workloads, lineage can help teams trace the output back to the data that contributed to it, providing the evidence pharma companies need for audits, regulatory reviews, and model explainability.

By implementing a unified data fabric, data remains understandable, protected, and governed as it moves across environments and as it is consumed for analytics and AI.

Making AI Explainable

Generative AI is probabilistic. Given the same prompt multiple times, users can expect different results because models identify patterns and generate outputs based on likelihood. While probabilistic models add a lot of value for many use cases, their outputs can be difficult to understand.

Explainability is the ability to understand and communicate how an AI system reached a result. The level of explanation required depends on how the output is used. For low-risk productivity tools, showing the documents used to generate an answer may be sufficient. But for an AI system that influences a clinical, safety, manufacturing, or regulatory decision, teams may need to understand and document much more.

A unified data fabric contributes to AI explainability by providing:

- **Source attribution:** Identifying the records, documents, and datasets used by the model.
- **Data lineage:** Showing how information changed before reaching the model.
- **Model lineage:** Tracking model versions, training data, configurations, and deployments.
- **Auditability:** Recording prompts, outputs, approvals, overrides, and actions.
- **Policy context:** Demonstrating that the model accessed only authorized information.
- **Reproducibility:** Preserving enough information to investigate or repeat an outcome.

Explainability gives stakeholders evidence to evaluate whether models and their outputs are appropriate for a decision.

Bringing AI to the Data

Leveraging AI in a public cloud environment carries risk: risk related to IP and data leakage, and risk of losing control over the models. One way to reduce that risk is to bring AI to the data instead of bringing data to the AI.

In a private AI architecture, pharmaceutical companies maintain control over the models, data, infrastructure, and access policies used for an AI workload. The model may run in a private cloud environment, a private data center, or on premises near a laboratory or manufacturing facility.

This approach enables organizations to use proprietary and regulated information without exposing it to a public cloud or model.

It can also support data residency requirements by enabling companies to keep data within a specific country, region, or system.


TIP

There is often a cost advantage to running AI on infrastructure the organization owns. Costs can quickly add up in the typical pay-as-you-go model, whereas you have already paid for on-premises servers. Consider running different components of the AI workflow on the infrastructure where it makes the most technical and financial sense.

Applying Guardrails to AI

While security and governance controls determine what models can access, guardrails determine what it can do. They are technical and procedural boundaries that constrain AI behavior. They may prevent a model from revealing sensitive information, require it to use approved sources, prohibit certain actions, route high-risk outputs to a human reviewer, and reduce resource utilization.

Common guardrails include:

- Restricting retrieval to trusted and approved data
- Filtering sensitive information from prompts and outputs
- Requiring source citations for generated answers
- Setting confidence thresholds for recommendations
- Preventing agents from using unauthorized tools
- Requiring human approval before certain actions
- Monitoring models for changes in performance or behavior

Guardrails should match risk and impact. More consequential decisions require stronger controls and oversight.

IN THIS CHAPTER

- Telling generative and agentic AI apart
- Matching the right AI to the right job
- Taking practical first steps with AI

Chapter 4
Generative and Agentic AI in Pharma

AI has become shorthand for a range of technologies that behave very differently. A chatbot drafting a clinical summary and a system rerouting a trial supply shipment are both labeled AI, but they perform different jobs, they carry different risks, and they require different levels of oversight.

Understanding those differences is critical for AI success. Pharmaceutical companies don't need the largest model or the most autonomous system for every workload. They need the right form of AI for the problem, and it must be grounded in trusted data and operating within the proper security, governance, and guardrails framework.

Generative AI vs. Agentic AI

At the simplest level, generative AI creates and agentic AI acts. Figure 4-1 illustrates the differences.

Generative AI responds to a prompt by producing an output, such as summarizing a clinical study report, drafting a medical

affairs briefing, explaining a regulatory requirement, or creating a first draft of a trial protocol. The user typically decides what happens next.

Agentic AI goes further by pursuing a goal through a sequence of steps. It can plan work, use approved tools and data, observe results, adjust its approach, and continue until the task is complete or human input is required. For example, an agent might gather information from multiple systems, compare trial sites against enrollment requirements, identify missing documentation, or escalate unresolved issues for review.

Think back to the map metaphor from Chapter 2. Generative AI can describe a route across the pharma value chain. Agentic AI can follow that route, respond to obstacles, and keep moving toward the destination.

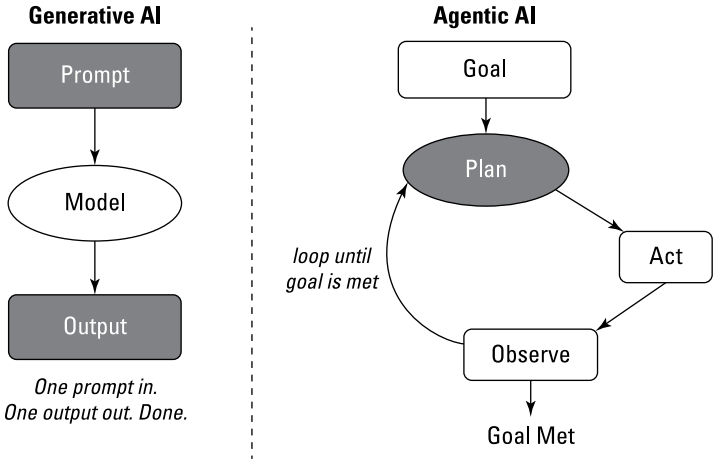

FIGURE 4-1: Generative AI produces a single output from a prompt; agentic AI plans, acts, observes results, and continues toward a goal.

Matching the AI to the Job

Not every pharmaceutical use case needs the same model or level of autonomy. Using more AI than the task requires wastes resources and may introduce risk without creating additional value.

Teams have two decisions to make as they leverage AI: They must select the right model and choose the right operating mode.

Choosing the right model size

It's tempting to assume bigger is always better. The largest general-purpose models can perform sophisticated reasoning and synthesize information across many topics, but they may be expensive and slow for narrow, repetitive work.

A large model may make sense when a regulatory specialist needs to compare complex requirements across several markets. It may be unnecessary for extracting dates, trial identifiers, or laboratory values from thousands of standardized documents.

Pharmaceutical companies can generally choose among:

- **Large general-purpose models:** Useful for open-ended reasoning, complex synthesis, and tasks where nuance matters more than speed or cost.
- **Smaller specialized models:** Useful for narrow, high-volume tasks such as classification, document extraction, and standardized summarization.
- **Traditional machine learning:** A single-purpose model that is often the best choice for pattern recognition, prediction, anomaly detection, and other tasks that don't require natural-language generation.

TIP

Cost and latency matter more at scale. To control costs, start with the smallest model that reliably performs the task, moving to a larger one only when added capability produces measurable value.

Matching autonomy to risk

The second decision is how much freedom AI should have to act.

A generative assistant that drafts an internal summary carries less risk than an agent that changes a manufacturing schedule or communicates with a clinical trial site. The greater the potential impact, the stronger the validation, governance, and human oversight should be.

Three broad modes are useful:

- **Generative assistance:** AI produces an answer or draft for a human to review. This mode works well for research, summarization, and content creation.

- **Grounded AI:** AI retrieves context from trusted organizational data before answering. This should be the standard for work involving company-specific compounds, trials, manufacturing processes, or regulated information.
- **Agentic workflows:** AI completes a multi-step task using approved tools and systems.

The mode should be driven by the cost of the output being wrong, rather than the desire to achieve full autonomy.

Giving AI the Right Context

AI outputs depend on the context models receive, but more context doesn't automatically improve accuracy or consistency. Irrelevant or outdated information can distract a model, bury important qualifications, create conflicts, and expose sensitive data unnecessarily.

Instead, AI should retrieve the smallest amount of trusted, relevant, and current information needed for the task, and the knowledge graph helps identify that information. It connects entities and relationships across the value chain and helps AI locate supporting evidence without forcing it to search every source. The unified security and governance layer determines whether the user or model is authorized to access that information.

Together, these capabilities give AI context that is relevant to the question or task, drawn from trusted sources, authorized for that user and purpose, and traceable based on the original evidence.

The right context matters even more for agentic AI. A mistake in a single generated answer may remain contained to that output. An agent can carry a mistake into later steps, where it can influence additional decisions and actions. Agentic workflows should therefore check trusted information throughout the process rather than retrieving data only once at the beginning.

REMEMBER

The longer the workflow, the more opportunities an early error has to compound. Grounding, guardrails, and monitoring should apply at every important decision point.

Three Rules for Pharmaceutical AI

Although the use cases differ, pharma companies should apply three principles to AI deployed across the pharma value chain.

Keep humans involved where judgment matters. AI can reduce manual effort, but human experts remain responsible for scientific, clinical, safety, quality, and regulatory judgment.

Provenance makes that oversight scalable. When AI output links directly to its supporting sources, reviewers can verify the evidence instead of reconstructing the analysis from scratch. Ultimately, the role of AI is to enable humans to focus their attention where expertise and accountability matter the most.

Prepare data before selecting the model. AI adoption starts with the data, not the model. Models need access to information that is connected, consistent, secure, governed, and available in the right context. Without that foundation, even the most capable model knows very little about the organization's compounds, trials, manufacturing processes, or regulatory history. Focusing on the data foundation creates a repeatable AI capability across the value chain.

Keep deterministic work deterministic. Generative and agentic AI are probabilistic. They excel at interpreting language, finding patterns, and synthesizing evidence under uncertainty, but exact tasks such as dosage calculations, statistical computations, and process-control limits should be performed by validated software.

WARNING

When exactness matters, AI should call a validated deterministic tool rather than perform the calculation itself.

Where and How to Get Started with AI

Nearly every pharmaceutical company is under pressure to make progress with AI. The fastest visible action is often purchasing chatbot licenses for employees. That may provide some productivity value, but access to a chatbot is not an AI strategy.

A general-purpose model knows nothing about an organization's proprietary data unless the company gives it trusted access to that context. Employees who receive confident but irrelevant or unreliable answers quickly stop using the tool. The initiative is often described as an adoption failure when the real problem was that the model was never given the context it needed to deliver value.

Follow these steps to ensure successful deployments of AI:

1. **Start with a valuable workflow.**

   Select a workflow that requires employees to spend significant time manually working with information. Good candidates have a clearly defined business or scientific objective, accessible and trustworthy data, measurable outcomes, and subject matter experts who are willing to help redesign the work.

2. **Build on trusted data.**

   Connect the relevant sources, resolve conflicting identities, preserve provenance, and apply consistent security and governance. The goal is to ensure information required for the selected workflow is context-ready.

3. **Design oversight and observability from the beginning.**

   Organizations need visibility into what context AI receives, what it produces with that context, which tools it uses, and what actions it takes. This visibility — often called *observability* — helps teams identify performance changes, investigate errors, evaluate adoption, and prove compliance.

REMEMBER

   It is easier to build this into the initial workflow than it is to bolt it on after AI is already running.

4. **Rethink the workflow.**

   Rather than simply inserting AI into an existing, unchanged process, the companies that achieve meaningful returns from AI work cross-functionally with subject matter experts across the organization to transform the workflow.

REMEMBER

Adoption and return on investment with AI come from embedding the right model into a redesigned workflow built on a foundation of trusted and governed data.

Chapter 5
Ten Steps Toward AI Transformation in Life Sciences

We've talked throughout this book how to leverage AI to accelerate the pharma value chain and bring new therapies to market faster. Here, we give some takeaways:

- **Get your data estate in order.** Successful AI transformation starts with building a foundation of trusted, secured, and governed data.
- **Leverage a unified data fabric.** A consistent security and governance framework across environments and systems is crucial for trusted data and AI.
- **Build a knowledge graph.** Knowledge graphs unify data across the value chain and provide context for accurate and consistent AI outputs.
- **Focus on one use case to start.** Instead of trying to build a shared ontology across the entire organization, target the relevant entities and relationships for a specific use case and grow from there.

- **Consider a Private AI architecture.** Private AI enables pharmaceutical companies to build AI using more of their data in a safe, secure, and controlled environment.
- **Implement guardrails.** Guardrails provide the constraints necessary to trust AI with business and operational decisions.
- **Ensure every AI output is explainable.** AI governance in life sciences requires teams to understand and document AI decisions and the conditions, including the context, that led to those outputs.
- **Choose the right AI for each job.** Generative AI, agentic AI, and machine learning are all best suited for different types of workloads. Choose the right mode for the job.
- **Match autonomy to risk.** Start with humans reviewing and taking actions, then increase AI autonomy only as workflows demonstrate reliable performance. The higher the cost of being wrong, the stronger the oversight should be.
- **Rethink the workflow.** Work with the people performing the job to determine what AI should automate, what humans should own, and how the workflow itself can be redesigned to create measurable value.

CLOUDERA

The Great AI Re-Architecture

Dive into the next stage of AI's fast-paced evolution

Workloads are becoming more distributed, data is moving between environments more frequently, and leaders are under heightened pressure to balance business impact with spend.

Turn pharma data into insights with AI

Pharmaceutical companies are facing rising research and development costs, a complex and dynamic regulatory landscape, and greater pressure to produce more targeted therapies. AI has the potential to accelerate the pharma value chain, reducing the time to market for life-saving therapies, but only if pharma companies can unify fragmented data and deliver the context AI needs to deliver accurate and consistent results. This book explores how pharma companies can build a foundation of trusted, governed data and operationalize AI to deliver value and automate workflows across the value chain.

Inside...

- Bring therapies to market faster
- Build a pharma knowledge graph
- Secure and govern your data
- Apply AI across the value chain
- Match models to the right workflows
- Prioritize AI projects based on value
- Redesign workflows around AI

CLOUDERA

Jeremiah Morrow is has been writing about technology, data and AI, and digital transformation for more than a decade. **Rameez Chatni, PhD** is an engineer-scientist-turned-AI leader, driving AI-impact across the pharma value chain.

ISBN: 978-1-394-47619-0
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.