



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Pharma Data Engineering Workshops: Databricks & AI Skills

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-03-03 · pharma data engineering · databricks training · healthcare ai · data lakehouse · gxp compliance · workforce upskilling · clinical data pipelines · ai foundations

Original article: <https://intuitionlabs.ai/articles/pharma-data-engineering-workshops-databricks-ai>



In this article

1. Executive Summary
 2. Introduction and Industry Background
 3. Section 1: The Pharma Data Landscape and Its Challenges
 4. Section 2: Data Engineering Fundamentals in Pharma
 5. Section 3: The Databricks Lakehouse Platform for Pharma
 6. Section 4: AI Foundations for Pharma Teams
 7. Section 5: Design and Delivery of Pharma Data Engineering Workshops
 8. Section 6: Case Studies and Real-World Examples
 9. Section 7: Data Analysis, Metrics, and Evidence
 10. Section 8: Challenges and Lessons Learned
 11. Section 9: Implications and Future Directions
 12. Conclusion
 13. References
-

Executive Summary

In recent years, the pharmaceutical industry has faced an unprecedented surge in data volume and complexity, from genomics and clinical trials to real-world evidence and manufacturing logs. To remain competitive and compliant, pharmaceutical companies are driving a **digital transformation** that relies on robust data engineering and advanced AI capabilities. However, a critical bottleneck persists: **workforce readiness**. Training plans should begin by assessing *skills shortages and talent gaps* before choosing a curriculum (www.fiercepharma.com) (www.techradar.com). In response, specialized training programs – notably **Pharma Data Engineering Workshops** centered on platforms like Databricks – have emerged to upskill teams in data pipeline development and foundational AI.

This report provides an in-depth analysis of the rationale, design, and impact of Pharma Data Engineering Workshops in 2026. We review the historical evolution of pharmaceutical data management, outline the technical foundations of data engineering and AI in life sciences, and examine the pitfalls posed by regulatory compliance (e.g. **GxP**, HIPAA, AI Act) and legacy silos (www.prnewswire.com) (www.covasant.com). We detail the **Databricks Lakehouse** platform – a unified data + AI environment – and its healthcare-specific extensions (from genomics libraries to integrated NLP tools) (www.prnewswire.com) (pages.databricks.com). Workflows on Databricks (Apache Spark, Delta Lake, MLflow, Databricks SQL, and GenAI features) are highlighted.

Central to our discussion are multiple training strategies and case studies. We examine industry-wide initiatives (e.g. Databricks' commitment to train 100,000 people in AI data skills (www.itpro.com)), corporate programs (including Novartis's dedicated Learn portal (pages.databricks.com)), and vendor-led workshops (e.g. joint Databricks–John Snow Labs pharmacovigilance labs (pages.databricks.com)). An example table outlines typical workshop curricula covering data engineering, ML/AI, analytics, and compliance (including tools like MLflow and Agent Bricks) (www.linkedin.com) (www.databricks.com). We also present industry

perspectives on training evaluation: vendor-reported changes in follow-up test performance (www.wave-access.com), stakeholder testimonials (www.edlitera.com) (www.edlitera.com), and governance of safe AI use (www.databricks.com).

Finally, we discuss future directions: emerging AI/ML trends (foundation models, federated data architectures), regulatory developments (AI Act, FDA's AI pilots), and the evolving data engineering roles required. By assembling cross-disciplinary viewpoints — from CIOs stressing data fragmentation (www.prnewswire.com) to regulators and training proponents emphasizing workforce development (www.techradar.com) (www.fiercepharma.com) — we underscore that **investing in pharma data engineering workshops on Databricks & AI** is not only timely but essential for the industry's strategic goals.

Introduction and Industry Background

The pharmaceutical industry is in the midst of a digital metamorphosis. Historically rooted in chemistry and biology, pharma now increasingly depends on *data* spanning drug discovery, clinical operations, supply chain, real-world outcomes, and commercial analytics. These data can be fragmented across silos and disparate formats. As one industry analysis notes, “traditional data silos are giving way to data lakehouses” precisely because *fragmented data hinder innovation in drug discovery, trial optimization, personalized medicine and predictive care* (www.covasant.com).

Compliance and governance further complicate the picture. As of 15 October 2026, assess obligations for each pipeline's intended use, records, jurisdiction, entity role, and AI risk classification. **FDA guidance on 21 CFR Part 11** ties applicability to qualifying electronic records under predicate rules; **HIPAA protections** generally apply to information held by covered entities and business associates. Assess GCP and GDPR applicability for the specific workflow rather than treating every pipeline as subject to the same checklist. The **EU AI Act** already has applicable obligations, with exceptions and later milestones for certain high-risk systems. Meeting these requirements demands traceability, **auditability**, and explainability. Meanwhile, new data types (real-world evidence from claims or remote sensors, unstructured text from physician notes, multimedia like medical imaging) require modern architectures. The industry has responded by transitioning away from legacy warehouses to a unified “**Data Lakehouse**” paradigm where raw and structured data coexist with governance (www.covasant.com).

The Growing AI Imperative

AI and machine learning have been hot topics in pharma. Over the past half-decade, the industry has witnessed an explosion of AI applications: from protein folding and small-molecule design to automated pathology, supply chain optimization, and personalized medicine recommendation engines (www.pharmtech.com) (www.techradar.com). PharmTech reports that **2025 was considered “The Year of AI”** for bio/pharma, marked by high-profile partnerships (e.g. Pfizer with AI platform firms (www.pharmtech.com)) and leadership poll results indicating strong belief in AI's power to accelerate R&D (www.pharmtech.com).

However, such promise brings challenges. NTT DATA's July 2025 research surveyed 425 healthcare decision-makers and influencers in 33 countries; 75% acknowledged a lack of necessary GenAI skills. These are organizational respondents' perceptions, not an employee competence rate (**NTT DATA**). The World Economic Forum's surveyed employers expected 39% of existing skill sets to be transformed or become outdated over

2025–2030 ([Future of Jobs Report 2025](#)). GlobalData surveyed 109 pharmaceutical industry professionals from 15 August to 11 October 2024; 49% identified a lack of specific skills and talents as the top hindrance to digital transformation ([GlobalData](#)).

These conditions set the stage for **Pharma Data Engineering Workshops**. The core idea is to rapidly elevate internal teams' capabilities by hands-on training in modern data platforms (notably Databricks) and AI fundamentals. The workshops typically cover end-to-end workflows – from data ingestion and processing to analytics and ML – tailored to pharma use-cases. By focusing on **practical skills and tools** (e.g. Apache Spark/Delta, MLflow, SQL/BI dashboards, and AI assistants), companies aim to accelerate time-to-insight and cultivate a data-driven culture. In effect, these workshops address the industry need identified by Sundar Srinivasan (NTT Data): “*develop comprehensive workforce training*” as a cornerstone of a successful AI strategy (www.techradar.com).

Section 1: The Pharma Data Landscape and Its Challenges

1.1 Data Sources and Types in Pharma

Pharma organizations generate and consume myriad data types:

- **Research & Development (R&D Data):** High-throughput screening outputs, genomics/proteomics assays, chemical libraries, computational biology models. Modern drug discovery alone can produce petabytes of specialized scientific data.
- **Clinical Trial Data:** Clinical Research Forms (CRFs), Electronic Data Capture (EDC) systems, lab results, imaging scans (MRI/CT), electronic health record (EHR) extracts, wearable device logs. Trials increasingly incorporate real-world evidence (RWE) from registries and devices.
- **Manufacturing and Quality Data:** Batch records, sensor telemetry from production lines (e.g. temperature, pH), batch release analytics. The emergence of “pharma 4.0” means advanced process-monitoring data is continuously streamed.
- **Regulatory and Quality Compliance Data:** Documentation for submission (e.g. lab notebooks, audit trails), QA/QC logs, pharmacovigilance databases (adverse event reports).
- **Market Access and Commercial Data:** Sales and marketing CRM data, insurance claims, payer outcomes databases, pharmacy dispensing records.
- **Patient and IoT Data:** Beyond trials, data from patient support apps, remote monitoring devices, and social media (e.g. public reports of symptoms).

These variables emphasize *volume, velocity, and variety* of pharma data. A survey of healthcare data projects notes organizations face “massive multi-format data volumes” (claims, genomics, trial data, device telemetry, etc.) (www.covasant.com). For example, wearable devices can produce thousands of data points per patient per day, and whole-genome sequencing adds enormous binary datasets. In terms of velocity, edge devices and continuous manufacturing produce near-real-time streams that require scalable processing and storage.

1.2 Data Silos and Fragmentation

A longstanding challenge is **fragmented data silos** across organizational domains. Different functional units (R&D, clinical, manufacturing, sales) often deploy disparate systems that were not designed to interoperate. These include legacy relational warehouses for manufacturing, specialized lab databases for R&D, isolated EHR systems for clinic data, and various third-party cloud analytics for commercial insights. The result: **silos** that inhibit end-to-end analysis. For instance, integrating a patient's genomic profile (R&D) with real-world clinic outcomes (trial/EHR) and manufacturing traceability often requires laborious data wrangling.

Databricks and industry experts explicitly highlight this issue. In a Databricks 2022 announcement, executives noted that healthcare and life sciences organizations historically **inhibited innovation** by “creating data silos and making advanced analytics difficult” (www.prnewswire.com). For example, GE Healthcare's CTO Joji George explains that their field needed “a comprehensive view of the patient” across care pathways, which was enabled by unifying data in a single lakehouse platform (www.prnewswire.com). Without unification, teams lack timely insights and duplicate costly data work.

Other analysis adds to this perspective. A 2024 industry blog asserts: “The current data fragmentation hinders innovation in most areas, be it drug discovery, clinical trial optimization, personalized medicine or predictive care” (www.covasant.com). This fragmentation arises from *incompatible data standards and isolated processes*. For instance, clinical data may reside in OMOP-formatted databases, while lab assay data uses custom schema. Such heterogeneity burdens data engineers who must reconcile and transform data for any enterprise analytics or AI model.

1.3 Regulatory and Compliance Constraints

Assess regulatory and governance requirements for the intended use and jurisdiction of each workflow. Workshop exercises should document the applicability decision rather than assume that every pharma pipeline has the same legal obligations.

- **Electronic records and signatures:** [FDA's Part 11 guidance](#) recommends determining and documenting which records fall under Part 11 based on predicate rules.
- **Good practices (GxP):** Identify the regulated activity and its applicable requirements before defining validation and evidence needs.
- **Patient privacy:** [HHS guidance](#) describes HIPAA protections for information held by covered entities and their business associates. Assess GDPR and other privacy requirements separately for the workflow and jurisdiction.
- **Interoperability:** Select and teach the schemas and standards required by the participating systems; do not treat every interoperability framework as a universal legal requirement.
- **AI regulation — status as of 15 October 2026:** The [European Commission's AI Act timeline](#) states that the Act became applicable on 2 August 2026 with exceptions. Prohibited-practice and AI-literacy obligations began applying on 2 February 2025, and governance and general-purpose AI obligations on 2 August 2025. It lists 2 December 2027 for certain Annex III high-risk uses and 2 August 2028 for high-risk systems embedded in Annex I regulated products. Assess the system's use, risk classification, and the organization's role before assigning obligations.

These factors affect workshop content too: data engineers must learn to incorporate privacy-preserving transformations and governance (e.g. unity catalogs, role-based access control). They also must understand healthcare schemas and terminology (ICD-10 codes, LOINC lab codes) because data lakes must correctly interpret clinical fields.

1.4 Opportunities and Use-Cases

Despite challenges, unified data infrastructures enable powerful use-cases:

- **Drug Discovery and Biomarker Identification:** Integrating assays (genomics, high-content screens) with literature mining via NLP can surface novel targets (www.covasant.com). A unified pipeline allows ML models to learn from diverse molecular data simultaneously.
- **Clinical Development & Real-World Evidence (RWE):** Cohort building using medical records, RWD (e.g. insurance claims) and trial data supports external control arms or patient stratification. Databricks' "Real World Evidence Suite" accelerates this by ingesting EHR, claims, and generating cohorts (www.prnewswire.com).
- **Pharmacovigilance:** With <5% of adverse events formally reported (pages.databricks.com), analyzing unstructured text (doctor notes, social media) via NLP can discover drug safety signals earlier. Databricks-hosted workshops have trained teams on using healthcare NLP to detect adverse drug events (pages.databricks.com) (pages.databricks.com).
- **Precision Medicine:** Unified patient data supports predictive modeling. For example, Mayo Clinic-type initiatives could predict disease risk from both genomic profiles and longitudinal health records; Databricks customers are working on such "disease prediction" models (www.prnewswire.com).
- **Manufacturing & Quality Analytics:** Real-time analytics on production sensors can predict equipment failure or ensure batch consistency. Databricks has case studies (e.g., Eli Lilly) showing how unified analytics shortens cycle times and prevents anomalies (www.databricks.com).
- **Supply Chain Optimization:** The COVID-19 era highlighted pharma supply chain fragility. Streaming analytics of raw material shipments (IoT trackers) can ensure continuity of drug supply (events.databricks.com).
- **Business Intelligence:** Consolidating sales, marketing, and expenditure data for forecasting and market access decisions. Shared data catalogs provide common reference for global teams.

In summary, the data engineering foundation unlocks cross-functional insights from R&D to commercialization, while AI adds predictive power. However, realizing these benefits requires teams skilled in modern "big data" architectures and AI tools – hence the push for specialized training.

Section 2: Data Engineering Fundamentals in Pharma

2.1 What is Data Engineering?

Data engineering is the discipline of **designing and building systems** that collect, ingest, store, and transform raw data into formats amenable for analytics and AI. It encompasses:

- **Data Ingestion:** Capturing data from diverse sources (databases, APIs, files, devices) using ETL/ELT (Extract-Transform-Load) processes or streaming ingestion.
- **Data Storage:** Organizing data in data lakes (distributed object stores) and warehouses (structured databases). Use of formats like Parquet or Delta that support large-scale analytics.
- **Data Transformation:** Cleaning, normalizing, aggregating, and integrating datasets. For example, joining trial data with lab results, or converting free-text into structured fields.
- **Data Quality & Governance:** Validating accuracy and enforcing schema/data quality rules. Maintaining lineage so any analytic result is audit-worthy.
- **Workflow and Pipeline Orchestration:** Automating data flows with tools like Airflow or Lakeflow pipelines, ensuring timely updates.
- **Scalability and Performance:** Architecting for cloud-scale compute (horizontal scaling) so pipelines can handle high-volume streaming or batch jobs without manual intervention.

In pharma, *data engineers* often build pipelines for regulated “science production”. Unlike consumer data (clickstreams), pharmaceutical data pipelines must frequently adhere to compliance checks at each stage. They also often reach global distribution: a secure data catalog (e.g. Databricks Unity Catalog) might be used to tag each dataset with sensitivity labels (public research vs protected PHI).

2.2 Role of Data Engineering in Analytics and AI

Data engineering is **the foundation** on which data science and AI are built. No machine learning model can be trained or deployed reliably without well-curated and accessible data. Key points:

- **Reproducibility:** Data engineers use version control for data (e.g. Delta Lake automatically tracks changes). This aligns with regulatory audit demands.
- **Timeliness:** They ensure data pipelines are updated frequently. For example, if a model predicts manufacturing throughput, the pipeline must ingest sensor data in near-real-time.
- **Bridging Domains:** Data engineers integrate domain-specific data. A pharmacovigilance text-mining pipeline requires matching drug codes in free-text to a reference database – a non-trivial engineering task.
- **Efficiency:** Automated pipelines (like Databricks *Lakeflow pipelines*) prevent ad-hoc scripts that break easily. As one workshop participant noted, these abstractions allow *automating workflows without endless scripting* (www.linkedin.com).
- **Enabling Self-Service:** By structuring and cataloging data, data engineers empower scientists and analysts (or even non-technical users) to query data without deep code. A 2026 Databricks summit example describes giving “scientific insights” to non-technical pharma users via dashboards and chatbots built on the lakehouse (www.databricks.com).

In essence, data engineers make it possible to “turn messy data into clear insights” (www.linkedin.com). Without them, analysts spend most time wrangling data; with them, the focus shifts to deriving value.

2.3 Key Technologies and Patterns

Distributed Processing (Apache Spark): A cornerstone of modern data engineering is distributed computing. Apache Spark (the engine behind Databricks) enables parallel processing of big datasets. Pharma engineers use Spark to run ETL and ML preprocessing at scale. For example, joining a billion-row genomics variant table with patient outcomes can be done via Spark.

Delta Lake (and ACID storage): Delta Lake is a storage layer providing ACID transactions on top of cloud storage (S3, ADLS, GCS). It prevents data corruption in concurrent pipelines and supports **schema evolution**. As Databricks notes, the lakehouse “eliminates the need for legacy data architectures” by providing an open, cloud-native approach (www.prnewswire.com). Delta's time-travel and logging also serve compliance (you can restore previous data states).

Data Catalogs and Governance (Unity Catalog): Managing who can see what in sensitive pharma data is crucial. Databricks Unity Catalog lets organizations register tables, documents, ML models, and define fine-grained access policies. Training includes catalog management to ensure, for instance, that only authorized teams can access identified patient records.

Workflow Orchestration (Lakeflow pipelines, Airflow): An emerging best practice is “pipeline as code.” Tools like Lakeflow pipelines allow declaratively defining data transformations. In workshops, trainers emphasize using Lakeflow pipelines to automate ingestion rather than hand-writing cron jobs. This ensures pipelines are monitored and easily recoverable.

Machine Learning Infrastructure (MLflow): MLflow originated at Databricks to manage ML experiments. Data engineering workshops include MLflow as a “one-stop shop” to track models from training to deployment (www.linkedin.com). This bridges data pipelines to AI – e.g. a feature store pipeline can feed directly to an MLflow experiment.

Querying & Analytics (SQL/BI): Not all pharma staff code. Databricks SQL provides a drag-and-drop interface and BI dashboards for reporting (www.linkedin.com). Training covers how SQL constraints and dashboards are built atop the lakehouse, so even clinical operations teams can slice data without heavy engineering.

Generative AI Assistants: The latest wave in training touches on *agentic AI*. Databricks introduced tools like “Genie/Agent Bricks” (GenAI utilities) that let users ask questions in natural language. For instance, a scientist might chat with a model that queries the lakehouse. Data engineering workshops show how to integrate such tools while maintaining data governance (www.databricks.com).

2.4 Building Scalable Pipelines: Best Practices

Workshops emphasize hands-on building of pipelines with best practices:

- **Modular Design:** Break pipelines into standardized stages (ingest, transform, validate). Use notebooks and job clusters efficiently.
- **Idempotency:** Ensure jobs produce the same result if re-run (important under uncertainty of streaming data).

- **Automated Testing:** Incorporate data quality checks (e.g., verify schema, null counts) in pipelines. Databricks Academy often suggests “shift-left” tactics like writing SQL sanity queries.
- **Documentation:** Catalog pipeline logic and dependencies. The Unity Catalog metadata and notebooks serve as living docs.
- **Performance Tuning:** Training often includes optimizing Spark (e.g. partitioning strategies, broadcast joins) to handle the high data volume in pharma.

An illustrative case is **Independent Health** (a US insurer, analogous to a healthcare org) migrating to the cloud. Their data migration team received intensive training on Python and PySpark. Use this vendor case as an example of a tailored curriculum; assess participants’ practical abilities separately.

2.5 Data Engineering in Regulatory Context

In pharma, data pipelines themselves often fall under **computer system validation (CSV)** frameworks. Training covers how to document evidence that a data pipeline meets FDA expectations: version-controlled code, change management, and audit logs. For example, if a data engineer creates a new ingestion job for adverse event data, the development process parallels software validation (requirements mapping, testing, sign-offs). Workshops introduce concepts like using **Git integration** on Databricks for tracking pipeline changes, and ensuring data snapshots can be produced for regulatory audits.

Data security is also paramount. Workshops should cover encryption, access controls, and the distinction between pseudonymization (such as replacing identifiers with hashes) and HIPAA de-identification. **HHS guidance** describes Safe Harbor and Expert Determination; hashing identifiers alone does not establish that a release meets either method. Teach removal of residual identifiers, assessment of linkage risks, and documentation of an expert’s methods and results where that method is used. PHI-derived hash codes can be included under an appropriate Expert Determination only with the associated keys protected from disclosure, including to data recipients.

For the proposed curriculum, use **Lakeflow pipelines** for declarative orchestration; historical accounts may retain the former product name.

Section 3: The Databricks Lakehouse Platform for Pharma

3.1 Overview of Databricks Lakehouse

Databricks, founded by the creators of Apache Spark and Delta Lake, pioneered the **Lakehouse** concept – a unified architecture merging data lakes and warehouses. The Lakehouse provides ACID transactions, schema management, and BI support on data lake storage, enabling both data science and analytics in one platform (www.prnewswire.com) (www.covasant.com). In practice, Databricks runs on cloud infrastructure (AWS, Azure, GCP) and offers:

- **Delta Engine:** High-performance Spark processing with optimizations for machine learning and SQL.
- **Managed Metastore (Unity Catalog):** A central registry of tables, files, streams with access control.
- **MLflow Integration:** One-click model registry, experiment tracking.
- **Databricks SQL:** A rich SQL editor and dashboard builder connected directly to Lakehouse tables.

- **Collaboration Notebooks:** Multi-languages (Python, R, Scala, SQL) in shared repositories.
- **GenAI Features:** AI-assisted query builders (Genie), built-in adapters for large language models.

It also integrates with partner tools (e.g. data ingestion from Apache Kafka, streaming FHIR ingest, etc.). Importantly, Databricks has **industry-specific accelerators**. For healthcare, it collaborates with domain libraries: for example, Glow (an open-source genomics toolkit) is optimized for Databricks. (www.prnewswire.com).

The platform's relevance to pharma is also reflected in dedicated offerings. In March 2022, Databricks announced the **Lakehouse for Healthcare and Life Sciences** package to target common use-cases like disease risk prediction and medical image classification (www.prnewswire.com). This initiative signals vendor recognition of pharma's needs. Notably, early adopters of the Lakehouse included **Regeneron** and **Thermo Fisher Scientific** – evidence that even conservative life science enterprises see value. Thermo Fisher's IT director said Databricks “has enabled us to eliminate costly data silos, unlock new opportunities to innovate, and become a more data-driven organization” (www.prnewswire.com).

Databricks emphasizes that unified data greatly aids precision medicine: Michael Sanky (Databricks' HLS lead) remarks that with the Lakehouse, companies can go “from measuring disease to predicting it” by accelerating biomarker discovery workflows (www.prnewswire.com). The platform supports end-to-end pharma analytics:

- **Data Ingestion:** Streaming FHIR APIs, connectors to EMRs, lab instruments, etc.
- **Data Storage:** Delta tables for trials, batches, genomics.
- **Data Modeling:** Medical or OMOP schemas applied centrally.
- **Analytics & BI:** SQL dashboards for dashboards & KPI tracking.
- **Machine Learning:** Scalable training of models for tasks like pathology image analysis.
- **AI/GenAI:** Query data via chatbots (Agent Bricks, Genie (www.databricks.com)) and integrate LLMs for summarizing scientific text.

3.2 Key Features and Innovations

3.2.1 Lakeflow Pipelines and Apache Spark

Use **Lakeflow Spark Declarative Pipelines**, now described in the documentation as Lakeflow pipelines, for the proposed curriculum. Databricks identifies this as the successor name to Delta Live Tables ([Databricks technical blog](#)). Participants can practice SQL or Python transformations, inspect the execution graph, and test data-quality checks using synthetic lab-result data.

For supported operations, **Photon** executes queries using a native C++ runtime; it is not a version of Spark ([Databricks Photon documentation](#)). Keep Spark planning/APIs and Photon execution distinct when teaching the runtime. Workshops can cover compute sizing and partitioning without promising a universal speedup.

Apache Spark began as a research project at UC Berkeley's AMPLab ([Apache Spark research history](#)).

3.2.2 Machine Learning Flow (MLflow)

MLflow is another cornerstone. Databricks training ensures teams can track experiments, package models, and deploy at scale using MLflow (www.linkedin.com). Participants use MLflow Tracking UI to log parameters (like model hyperparameters) and metrics. They practice deploying a model as a REST endpoint via MLflow. For pharma, examples include deploying a risk prediction model for clinical outcomes or a dosing recommendation model.

One outcome from Novartis training captured this: “MLflow’s one-stop shop for tracking experiments, packaging models, and deploying with confidence” (www.linkedin.com). In practice, this means when a scientist iterates on a trial success predictor, MLflow records each version continuously. In a regulated environment, this also provides an auditable trail of model versions – crucial if the model informs FDA submissions or patient care.

3.2.3 SQL Queries and BI Dashboards

Databricks SQL offers a **drag-and-drop dashboarding** interface, lowering the bar for business teams. For example, Novartis trainees noted how “Databricks SQL’s drag-and-drop dashboarding” empowers anyone to visualize data (www.linkedin.com). In workshops, users create dashboards of patient enrollment over time, or vaccine adverse events by region. Teach dashboard caching and scheduled refresh explicitly: **Databricks recommends configuring a dashboard schedule to reliably refresh its cache**. For the lab, measure query latency and data freshness on representative workloads, recording compute, query design, cache behavior, ingestion timing, and refresh settings. Avoid promising sub-second queries or continuously current dashboards.

3.2.4 GenAI and Agent Bricks

Generative AI integration is a new and rapidly evolving chapter. At the Data + AI 2026 Summit, sessions highlighted using *Genie* and *Agent Bricks* in pharma (www.databricks.com). These allow non-technical users to ask natural language questions about clinical study data. For example, an analyst preparing a presentation could ask, “What is the trend of adverse events in the last quarter for drug X?” and the AI agent queries the lakehouse and produces an answer. For a proposed lab, build **retrieval-augmented question answering** over vetted trial protocol documents. Treat a separate fine-tuning exercise as a different task; do not describe document retrieval as model fine-tuning.

A critical teaching point is safety and governance with GenAI (“GenAI best practices” (www.linkedin.com)). Engineers learn to mask PHI, to limit generations to pre-approved knowledge bases, and to use retrieval augmentation only on vetted corpora. For instance, a pharmaco-toxicology summary generator can be fine-tuned on curated literature; staff learn to vet difficult cases rather than blindly trusting the LLM – aligning with the regulatory expectation to keep a human in the loop (www.techradar.com).

3.3 Databricks Ecosystem and Community

Beyond the platform itself, Databricks fosters a rich ecosystem that benefits pharmaceutical users:

- **Healthcare Partnerships:** Databricks collaborates with domain specialists like John Snow Labs (healthcare NLP libraries) and ZS Associates (genomics pipelines) (www.prnewswire.com). Workshops often leverage these partnerships. For example, a sponsored workshop taught by John Snow Labs experts focused on pharmacovigilance (pages.databricks.com).
- **Industry Conferences:** The **Data + AI Summit** is a key venue. It features many pharma-focused sessions (e.g. Eli Lilly's manufacturing case (www.databricks.com), Novo Nordisk using chatbots (www.databricks.com)). Attendance is often part of training plans.
- **Training Programs:** Databricks runs global academies and free courses (as news articles noted, a \$10M initiative in UK/Ireland to train 100,000 people (www.itpro.com)). There are specific certifications (e.g., Databricks Data Engineer Associate).
- **User Community:** The online Databricks Community forum has sections on healthcare analytics. Professionals share notebooks and best practices. Indeed, even the creation of the Novartis self-service portal (see below) suggests that pharma tech leaders actively collaborate online.

All these amplify the impact of in-house workshops. After a workshop, teams can take advantage of recorded sessions, forums, and further training. This ecosystem approach ensures that the workshop is the beginning of a continuous learning journey.

Section 4: AI Foundations for Pharma Teams

While data engineering provides the pipelines, **AI foundations** impart the modeling and reasoning capabilities that unlock insights. Training in AI within pharmaceutical workshops typically covers:

- **Machine Learning Basics:** Core concepts (supervised learning, unsupervised clustering, model evaluation metrics relevant to healthcare like ROC-AUC for diagnostic tests). Courses may use healthcare datasets (e.g. predicting disease outcomes) for hands-on practice.
- **Deep Learning:** Neural network architectures (CNNs for medical imaging, RNNs for sequential data). Learners train models on example tasks, such as pathology slide classification or genomics variant effect prediction. Emphasis is placed on using MLflow to track these experiments.
- **Natural Language Processing (NLP):** Handling text (doctor notes, literature). Proposed exercises can use pretrained clinical embeddings and task-specific fine-tuning. Include **BioBERT** here as a biomedical encoder for text mining, rather than as a GPT-style generative assistant. As noted, a specific workshop focused on “*Healthcare NLP at Scale*” (pages.databricks.com) demonstrated this in a drug safety context.
- **Generative AI / LLM Skills:** Introduction to generative language models. Teams learn prompt engineering, fine-tuning LLMs on domain data (e.g. internal protocols, regulatory guidelines), and how to integrate them via APIs. Crucially, training emphasizes ethical/secure use – for example, not revealing proprietary formulas to a public LLM.
- **AI Governance and Ethics:** A more recent pillar. Ensures teams understand bias, fairness, and documentation for models used in medicine. For example, discussing the EU AI Act's requirements (IntuitionLabs research highlights pharma-specific compliance needs) underscores the need for model validation and interpretability.

- **Statistical Foundations:** Many programmers come without formal stats background. Workshops often include refreshers on hypothesis testing, confidence intervals, p-values. This is key since pharma R&D culture relies heavily on statistical evidence.
- **Data Ethics and Privacy:** How to handle healthcare data responsibly. Topics include differential privacy, de-identification, and secure multi-party computation (especially if sharing data across pharma companies or with regulators).

These AI topics complement data engineering. A typical workshop might first secure the data pipeline, then apply an AI model in the second phase. For example, after building a structured adverse-event dataset, the team might train an NLP classifier to flag high-priority safety issues. The notion is *Data engineering pipelines feed AI, and AI outputs require new data feedback for improvement.*

4.1 Bridging to Pharma-Specific AI

Generic AI skills need adaptation for pharma. Workshops address domain-specific challenges:

- **Small Data Regimes:** Unlike tech, pharma often has limited labeled data (especially early in drug development). Thus, training may cover techniques like transfer learning from other biomedical datasets, data augmentation, or synthetic data generation (e.g. generative models for chemical libraries).
- **Time-Series and Survival Models:** Clinical trial outcomes often involve censored data (patients drop out or haven't had events yet). Teams learn survival analysis techniques (Kaplan-Meier, Cox models) and how to implement them in Spark/MLflow.
- **Domain Ontologies:** Understanding SNOMED, RxNorm, UCUM (units) and linking these to data processing. For example, a generative AI model might need to reference a med code; engineers learn to map code-ids to human labels.
- **Integration with Clinical Workflows:** AI outputs must often be certified or reviewed by clinicians. Training includes best practices for creating sharable visualizations or summaries (for example, generating patient profile packets automatically).

Presentations from industry leaders reinforce AI's central role: Jennifer Cannon of Thermo Fisher and others at PharmSci360 emphasized that AI is permeating drug development and quality control (www.pharmtech.com). Pharma conferences now routinely feature sessions on "data-driven AI in QA/QC" and "executive strategies for AI adoption," making it clear that AI fundamentals are no longer optional.

4.2 Training Modalities for AI

Different learning methods are used to build AI foundations:

- **Project-based learning:** Trainees work on a mini-project, e.g. building an NLP pipeline to extract symptoms from trial notes. This cements theory with a tangible outcome.
- **Guest Lectures:** Experts from academia or specialized companies (like John Snow Labs) may lecture on advanced topics (e.g. "Deep Learning for Histopathology").
- **AI Simulators/Challenges:** Ethical hacking-style challenges where teams find vulnerabilities in an AI model, to appreciate robustness.

- **Certifications:** Use a practical assessment as a milestone and verify any optional credential's availability before recommending it. The **TensorFlow Certificate exam is closed**.

Ultimately, AI foundations training goals mirror data engineering: **practical readiness**. After training, participants should feel comfortable selecting and customizing AI algorithms on Databricks, monitoring model drift, and working in cross-functional teams (with clinicians or statisticians).

Section 5: Design and Delivery of Pharma Data Engineering Workshops

5.1 Workshop Objectives and Audience

Pharma Data Engineering Workshops target **cross-functional teams**: data engineers, data scientists, bioinformaticians, and even advanced business analysts. Executive champions sponsor these initiatives, sometimes coordinated by the CIO or head of R&D informatics. Common objectives include:

- **Skill Development:** Impart hands-on experience on Databricks tools (Delta, Spark, MLflow) and AI frameworks (TensorFlow, PyTorch).
- **Project Kickstart:** Often, workshops are aligned with a strategic project (e.g. a hospital onboarding data, or a trial analytics pilot). By training teams in context, the content directly contributes to actual business objectives.
- **Collaboration and Evangelism:** Workshops create internal champions who can spread best practices and evangelize the tech stack. Novartis's in-house 'Databricks Academy' portal (see below) exemplifies fostering a learning culture (pages.databricks.com).
- **Governance and Standards:** Standardizing approaches (naming conventions, coding standards, security protocols) across teams.

The audience may have mixed levels. As in the Independent Health case, novices may have SQL/ETL background, while experienced engineers dive into Spark. Thus, workshops often use a **blended learning** approach: foundational lectures (for everyone), followed by breakout lab exercises at varying difficulty.

The following proposed agenda spans up to five days. It is an illustrative curriculum, not a record of a particular company's training or its outcomes:

Day/Session	Topic	Description
Day 1: Databricks Fundamentals	Introduction to Spark and Databricks	<i>Covers Databricks platform basics, Spark architecture, and how to create your first notebook. (A historical Novartis portal lists a Jan 20 session covering: Data Problem, DB Platform, Introduction to Spark (pages.databricks.com).)</i>
Day 1 (pm): Data Ingestion & Delta	Building Data Pipelines	<i>Hands-on with Delta Lake tables, ingesting sample clinical data from CSV/JSON, using Auto Loader or Azure Event Hubs.</i>
Day 2: Data Engineering Practices	Lakeflow pipelines, SQL, ETL pipelines	<i>Participants build an automated pipeline using Lakeflow pipelines. Best practices like idempotent jobs and testing are taught.</i>

Day/Session	Topic	Description
Day 3: Analytics & Visualization	Databricks SQL and Dashboards	Create tables for analysis, write SQL queries, build interactive dashboards. (Emphasize drag-drop ease (www.linkedin.com).) Use BI examples: e.g., a dashboard of study enrollment by site.
Day 4: Machine Learning & AI	MLflow, Model Training, GenAI	Train a simple ML model on pharma data; track experiments with MLflow. Introduction to Databricks GenAI tools (Agent Bricks). Discuss AI governance.
Day 5: Domain-specific Workshop	NLP for Drug Safety or Clinical Analytics	Work with unstructured text (e.g., doctors' notes) to identify adverse events (pages.databricks.com). Alternatively, build a small cohort analysis (simulating RWE). Wrap-up and Q&A.

Use current [Lakeflow pipeline documentation](#) for the pipeline exercises. This is an illustrative breakdown. Actual content is often co-developed with the company's subject matter experts. For example, if the focus is pharmacovigilance, the agenda may lean heavily on NLP (as in the drug-safety workshop linked in the Day 5 row). If the audience is clinical operations, more time might be spent on ML models for patient stratification.

5.2 Teaching Methodology

Key pedagogical practices include:

- **Hands-on Labs:** Every concept is paired with a lab. For data engineering, labs might include ingesting a skeletal dataset (e.g. drug response data) and performing transformations. For AI, labs might include training a Jupyter notebook model end-to-end.
- **Realistic Repositories:** Use sanitized real-world datasets when possible. E.g., surrogate health or trial data that reflect pharma complexities – missing values, mixed units – so that learning is authentic.
- **Collaborative Problem-Solving:** Teams of 2–4 people work together; they mimic a project team. This collaborative setting improves retention and replicates actual work patterns.
- **Incremental Complexity:** Start with smaller data and simpler tasks, then progress to scaled-up versions. For instance, begin by loading a small CSV, then upgrade to streaming messages in a cluster.
- **Expert Guidance:** Instructors (often Databricks experts or consultants with pharma background) circulate to assist and ensure best practices. They point out documentation, help debug, and ensure no one falls behind.

Feedback loops are built in: daily Q&A sessions, code reviews, and final project presentations. Success of a workshop is measured not by a written test, but by the team's ability to build a prototype by day's end (e.g. deploying a simple model or dashboard).

5.3 Organizational Support and Reinforcement

Workshops are most effective when supported by follow-on initiatives:

- **Executive Sponsorship:** Leadership buy-in (CIO, Chief Data Officer, or heads of clinical) signals importance. This can align training outcomes with KPIs (e.g. improvement in time-to-insight).
- **Self-paced Follow-up:** Companies provide access to the Databricks Academy and other courses (as Novartis did (pages.databricks.com)). Embedding this into internal LMS incents continuous learning.

- **Data Science Communities of Practice:** Internal meetups for alumni to share use-cases, challenges, and tips.
- **On-the-job Projects:** Right after training, participants are often staffed into a priority project (e.g. migrating an existing pipeline to Databricks). This “live project” cements skills.

For example, Novartis created a **Self-Service Learning Portal** listing Databricks webinars and resources for employees (pages.databricks.com). It lists webinars on “Databricks Fundamentals” and “Introduction to Data Science.” The portal described the sessions as repeating quarterly.

5.4 Scaling Training (Train-the-Trainer)

Larger firms employ a **train-the-trainer** model. A select group of engineers gets intensive training (often with the vendor). These internal experts then tailor workshops for other departments. This addresses the resource constraints of always relying on external instructors (expensive for large scale).

In 2025, Databricks announced plans to train 100,000 people across the UK and Ireland in generative AI and data engineering (www.itpro.com). While not pharma-specific, it illustrates the model: using the free edition and standardized curricula, in partnership with universities, to rapidly scale knowledge. Pharma companies can similarly partner with service providers to train batches of employees or external partners (like CROs).

5.5 Measuring Workshop Impact

Evaluation is critical. Metrics include:

- **Knowledge Assessments:** Pre/post quizzes on Databricks concepts; though practical skills are often validated via project completion.
- **Project Outcomes:** Compare a completed pipeline or model with its documented baseline. Record runtime, cost, data quality, and maintenance effort without assuming a percentage improvement in advance.
- **Usage Statistics:** Monitor platform usage (logins, notebooks run) before and after.
- **Surveys:** Gather attendee feedback on confidence and relevance. The Independent Health quote shows self-reported confidence gains (www.edlitera.com).

Separate the effects of the workshop from those of a later implementation. For a training-content system, distinguish correct-answer rates on follow-up tests from retention or first-attempt exam pass rates; Section 6.6 reports the vendor’s specific measure.

Section 6: Case Studies and Real-World Examples

6.1 Novartis: Comprehensive Databricks Upskilling

The following is an individual participant’s account, not a measured evaluation of workshop effectiveness. A LinkedIn account by Prabhakar Pandey (Novartis) describes a **3-day on-site Databricks training** covering Data Engineering, ML, SQL/BI, and Generative AI (www.linkedin.com). Key points from that testimonial:

- Multi-format delivery: The training blended ILT, hands-on exercises, assignments, and continued-learning webinars (www.linkedin.com).
- Tools covered: Pandey mentions *Delta Live Tables* (automating data workflows), *MLflow*, *Databricks SQL* dashboards, and *GenAI best practices* (www.linkedin.com).
- Outcome: Pandey described learning to “turn messy data into clear insights” and build reliable data pipelines in the Lakehouse (www.linkedin.com).

The Databricks-hosted self-service learning portal for Novartis (2022) was dedicated to employees, offering quarterly webinars and access to Databricks Academy courses (pages.databricks.com). For example, scheduled sessions included “Databricks Fundamentals” and “Introduction to Data Science with Databricks” (pages.databricks.com) – an example of continued learning resources.

6.2 Thermo Fisher and GE Healthcare: Lakehouse Adoption

While not a formal “training workshop,” the Databricks launch PR (www.prnewswire.com) highlights real impacts at Thermo Fisher and GE Healthcare. Thermo Fisher’s Sr. IT Director Feng Liang noted the Lakehouse “enabled us to eliminate costly data silos” and “ [unlock] new opportunities to innovate” (www.prnewswire.com).

Similarly, GE Healthcare’s CTO praised using Databricks to unify patient views across care pathways (www.prnewswire.com). Treat these as platform-adoption examples, not evidence of training activities or measured workshop effectiveness. A proposed workshop could use a similar data-integration scenario, but that would be a hypothetical curriculum application.

6.3 Ardigen: Clinical Data Lakehouse for Trials

Ardigen, an AI-driven Contract Research Organization in drug discovery, partnered with Databricks to create a **Clinical Data Lakehouse** solution (ardigen.com).

6.4 Novartis Self-Service Portal (2022)

Returning to Novartis, the company’s **self-service learning portal** is a blueprint for internal education (pages.databricks.com). It illustrates how corporate training extends beyond one-off workshops:

- **Quarterly webinars:** Dedicated Databricks sessions repeating every quarter, on topics like platform intro and Spark.
- **Databricks Academy access:** Employees were guided to complete official Databricks Academy courses (e.g. Data Engineering Certification).
- **Team of Experts:** The portal invited engagement with a “dedicated Databricks Team” behind it.

This institutional support likely amplified workshop outcomes; continuous learning ensures retention. Companies initiating similar training programs should note that an *ongoing curriculum* is needed to reach the desired “100,000 skilled people” footprint touted by Databricks (www.itpro.com).

6.5 Independent Health Case Study (Healthcare Perspective)

Independent Health (a not-for-profit insurer) worked with Edlitera for a custom Python data engineering training (www.edlitera.com). Their use case – migrating data to the cloud – parallels many pharma IT modernization projects. Key takeaways from this case:

- **Blended Curriculum:** The training combined theory and hands-on practice, emphasizing Python and PySpark for scalable pipelines (www.edlitera.com).
- **Participant Testimonial:** Edlitera quotes Venkat K., a data architect, who reported confidence writing Python and PySpark in the team's Cloud Build Lab after a couple of weeks ([Edlitera case study](#)). This is one person's self-reported experience, not a team-wide production-proficiency measure.
- **Real Outcome:** Post-training, the team “hit the ground running” with cloud data integration, applying the new skills to their migration project.

Use the case to inform curriculum design, and measure productivity and production readiness separately for each organization.

6.6 Data-driven Training Improvement (Wave-Access Case)

WaveAccess is a consulting company that devised an AI system to analyze a pharma company's training program (www.wave-access.com) (www.wave-access.com). Before building their solution, they stressed the importance of an **AI Readiness Workshop** to plan the project. While not Databricks-related, the case offers relevant lessons:

- They performed an *AI Readiness Workshop* to **evaluate data and identify gaps** (www.wave-access.com), highlighting that workshops can serve as scoping sessions to formulate AI use-cases.
- **Vendor-reported system outcome:** WaveAccess reports that the percentage of correct answers on follow-up tests increased by 20–25% after the AI-assisted training-content system was implemented ([WaveAccess case study](#)). The report does not specify this as percentage points. Do not relabel this measure as knowledge retention or first-time exam pass rates, or attribute the system-level result to the readiness workshop alone.

For pharma training programs, the insight is twofold: first, a preliminary workshop can align stakeholders and ensure data quality before building AI features. Second, define the assessment measure before evaluating a later implementation; test-answer accuracy alone does not establish financial ROI. This reinforces the viewpoint that **workshops should not be one-off events** but integrated into a feedback loop of continuous improvement.

Section 7: Data Analysis, Metrics, and Evidence

Throughout these programs, organizations collect data to validate impact. Example metrics:

- **Skill Acquisition:** Pre/post assessment scores or certification pass rates. Record the assessment definition, cohort size, dates, and denominator.

- **Pipeline Performance:** Number of jobs automated, reduction in runtime or cost. Record a baseline and the observation period for maintenance effort.
- **Business Metrics:** Speed of insight (from weeks to days), improvement in decision time. For a training-content implementation, track follow-up-test answer accuracy and content-update time (www.wave-access.com).
- **Employee Feedback:** Qualitative evidence, such as the architect's self-reported confidence in a Cloud Build Lab (www.edlitera.com).
- **Adoption Rates:** Percentage of data projects migrated to target platform post-training. For example, at Novartis, the uptake of Databricks notebooks in new projects could be tracked.

Measure post-workshop platform use and project outcomes directly rather than infer adoption or retention from the existence of a learning portal. Treat the categories above as proposed measurements, not published benchmark values.

Section 8: Challenges and Lessons Learned

Implementing training faces obstacles:

- **Time Constraints:** Staff have day jobs; carving out several days for workshops requires executive permission and project flexibility.
- **Varying Skill Levels:** Ensuring content is neither too basic nor overwhelming is tough. The blended approach (modular labs, online follow-up) mitigates this.
- **Data Privacy for Training:** Use synthetic or properly de-identified datasets as the workshop default. Any proposed use of identifiable records should undergo review for a lawful purpose, appropriate permissions and agreements, minimum-necessary access, and security controls. [HHS permits certain supervised healthcare training uses under applicable policies](#); this does not automatically authorize a vendor data-engineering workshop.
- **Change Management:** Introducing a new platform like Databricks may meet resistance. Workshops often include change management components (e.g. success stories, executive endorsements).
- **Sustainability:** After initial training, continuous support (office hours, refreshers) is needed or skills erode.
- **Vendor Dependence:** Relying on vendor-run workshops raises concerns about biased content. Organizations balance this by involving internal SMEs and external consulting.

Pharma leaders emphasize that training should align with strategy. For example, if a company's priority is speeding up drug safety surveillance, the workshop labs should simulate that use-case. Design training around the work participants need to build, with theory supporting the practical exercises.

Section 9: Implications and Future Directions

9.1 Building a Data-Driven Culture

Training is a means to a cultural end: fostering **data-driven decision-making**. By empowering non-technical staff (through dashboards and AI agents (www.databricks.com)), companies aim for organization-wide literacy. In 2026 and beyond, we expect more cross-functional involvement (e.g. regulatory affairs analysts querying data sets, medical directors exploring insights via notebooks). Companies that successfully scale training (akin to Novartis's 2022 portal) may see *data fluency* become a standard competency alongside biology and chemistry in pharma talent profiles.

9.2 Technological Evolution

The platforms and tools will rapidly evolve:

- **Cloud-Native Data Mesh / Lakehouse 2.0:** There's a trend toward federated data mesh and multiple lakehouses (clinical, research, IoT) that interoperate. Pillars of this (like Databricks' Unity Catalog being extendable across multiple AWS accounts or Azure subscriptions) will become important. Training must adapt to mesh paradigms: e.g., how to orchestrate queries across multiple catalogs.
- **Agentic AI in Operations:** Agent Bricks and similar frameworks will mature. In training, this means a deeper dive into orchestrating multi-step AI workflows (e.g. an agent that can fetch patient data, run an analysis, and draft a report autonomously).
- **Augmented Data Management:** Tools that auto-suggest schema or cleaning steps (using AI) may reduce engineering drudgery. Future workshops may teach "AI-assisted pipeline design".
- **Regulatory Sandboxes:** As agencies adopt AI (USP AI Programs, FDA's IStand becoming permanent (www.pharmtech.com)), new "AI validation" requirements will emerge. Training must cover documenting model quality in GxP context.

9.3 Skills Pipeline and Education

The workforce pipeline is key. In parallel with corporate training, universities and bootcamps are offering life sciences data programs. The collaboration seen in the UK/Ireland training initiative (www.itpro.com) may expand globally. Pharma companies might partner with academic institutions to shape curricula (as Databricks did with LSE for broader AI skills (www.itpro.com)).

Moreover, the concept of "**AI readiness**" is migrating down to the individual: pharma companies may start requiring data certifications for certain roles. Peer networks and consortiums (like Pistoia Alliance) may create shared training resources. Indeed, the high-level consensus (PharmTech KOLs meetings) suggests that *talent development is now a strategic concern* (www.pharmtech.com).

Conclusion

Design pharmaceutical data engineering workshops on Databricks & AI around documented skills needs, practical data-platform exercises, and responsible AI use. Set explicit objectives for data integration, governance, analytics, and project delivery. The examples here include vendor reports and **individual**

testimonials; use them to inform curriculum design, and measure outcomes in the organization's own setting rather than assume a workshop will deliver the same results.

However, training is not a one-time fix. The landscape will keep shifting (new AI regulations, technology updates, data growth). Sustained investment in learning, tied to real projects and measured outcomes, is essential. Organizations that succeed will be those that treat workshops not as a checkbox but as part of an ongoing **culture of learning** – one where data engineering and AI competencies permeate every level of the company.

As Databricks themselves frame it, the Lakehouse for healthcare is a “*modern, open and collaborative platform*” that equips teams with “timely and accurate insights” (www.prnewswire.com). Realizing this promise requires ensuring that those teams **are trained and confident**. A practical next step is to define a pilot curriculum, document baseline capabilities, and evaluate completed work before expanding the program.

References

- ITPro: Databricks wants to train 100,000 people in AI across the UK and Ireland; plans to train 100k across UK/IR in AI and data skills (www.itpro.com) (www.itpro.com).
- TechRadar coverage of NTT DATA's healthcare GenAI survey; the primary study surveyed decision-makers and influencers, not a representative employee population (www.techradar.com) (www.techradar.com).
- FiercePharma (GlobalData survey): 49% of pharma cite digital skills shortage as top challenge (www.fiercepharma.com) (www.fiercepharma.com).
- PharmTech (Dec 2025): 94% agree AI boosts R&D; interviews confirming appetite for AI (www.pharmtech.com) (www.pharmtech.com).
- Databricks PR (Mar 2022): Lakehouse for Healthcare drives unification of data and AI; customer quotes (GE, Thermo Fisher) (www.prnewswire.com) (www.prnewswire.com).
- Databricks & John Snow Labs workshop promo (Dec 2021): <5% of ADEs officially reported; workshop on Lakehouse + NLP for pharmacovigilance (pages.databricks.com) (pages.databricks.com).
- Ardigen press release: Clinical Data Lakehouse on Databricks for trials; unified data management, AI insights (ardigen.com) (ardigen.com).
- Prabhakar Pandey's LinkedIn account: 3-day Databricks training (ILT) on Data Eng, ML, SQL/BI, GenAI; highlights of Delta Live Tables, MLflow, SQL dashboards (www.linkedin.com).
- Covasant blog: unified data foundations and an overview of lakehouse architecture (www.covasant.com) (www.covasant.com).
- Novartis training portal (2022): quarterly Databricks fundamentals & ML webinars for employees (pages.databricks.com).
- Independent Health/Edlittera case (Mar 2023): Python/PySpark data engineering bootcamp; “hit the ground running” with cloud integration (www.edlittera.com) (www.edlittera.com).
- Wave-Access case (Oct 2025): AI Readiness Workshop and an AI-assisted training-content system; see Section 6.6 for the vendor-reported follow-up-test metric (www.wave-access.com) (www.wave-access.com).

- Databricks Data + AI Summit session (2026): dashboards + chatbots for pharma, acting on Unity Catalog and SQL (www.databricks.com).

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://intuitionlabs.ai/articles/pharma-ai-training-upskilling-strategies>
2. <https://www.fiercepharma.com/marketing/skills-shortage-still-holding-back-pharmas-digital-transformation-survey>
3. <https://www.techradar.com/pro/healthcare-providers-really-want-to-try-out-ai-but-dont-really-have-the-skills>
4. <https://intuitionlabs.ai/articles/gxp-ai-compliance-inspection-readiness-guardrails>
5. <https://www.prnewswire.com/in/news-releases/databricks-introduces-lakehouse-for-the-healthcare-and-life-sciences-industries-to-drive-transformation-across-healthcare-ecosystem-889577449.html>
6. <https://www.covasant.com/blogs/data-lakehouse-pharma-healthcare-unified-analytics>
7. https://pages.databricks.com/202201-AMER-VE-Healthcare-and-Life-Sciences-Workshop-Improve-Drug-Safety-w_Registration-Page.html
8. <https://www.itpro.com/business/careers-and-training/databricks-wants-to-train-100-000-people-in-ai-across-the-uk-and-ireland-heres-how-to-get-involved>
9. <https://pages.databricks.com/Novartis-Onboarding-Sessions.html>
10. https://www.linkedin.com/posts/prabhakarpandey_databricks-academy-certificates-activity-7337080983940210688-ALkz
11. <https://www.databricks.com/dataaisummit/session/empowering-non-technical-users-pharma-dashboards-and-chatbots-working>
12. https://www.wave-access.com/public_en/blog/2025/october/22/how-a-data-driven-approach-improved-training-for-a-pharma-company/
13. <https://www.edlitera.com/blog/posts/case-study-data-engineering-training>
14. <https://intuitionlabs.ai/articles/pharma-ai-governance-gxp-compliance>
15. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application>
16. <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html>
17. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
18. <https://intuitionlabs.ai/articles/audit-trail-requirements-ai-gxp-compliance>
19. <https://www.pharmtech.com/view/the-year-of-ai-2025-bio-pharma-upskilling-revolution>
20. <https://www.nttdata.com/global/en/news/press-release/2025/july/072500>
21. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/digest/>
22. <https://www.globaldata.com/media/pharma/lack-skills-talent-still-top-hindrance-digital-transformation-pharma-says-globaldata/>
23. <https://www.databricks.com/dataaisummit/session/transforming-bio-pharma-manufacturing-eli-lillys-data-driven-journey>
24. <https://events.databricks.com/building-resilient-supply-chains-in-healthcare-and-life-sciences>
25. <https://docs.databricks.com/aws/en/data-engineering>
26. <https://community.databricks.com/t5/technical-blog/what-s-new-in-the-lakeflow-pipelines-editor/ba-p/156555>

27. <https://docs.databricks.com/aws/en/compute/photon>
28. <https://spark.apache.org/research.html>
29. <https://docs.databricks.com/aws/en/dashboards/caching>
30. <https://docs.databricks.com/aws/en/agents/retrieval-augmented-generation>
31. <https://arxiv.org/abs/1901.08746>
32. <https://www.tensorflow.org/certificate>
33. <https://docs.databricks.com/aws/en/ldp/>
34. <https://ardigen.com/news-ardigen-partners-with-databricks-to-transform-clinical-data-management-with-lakehouse-solutions/>
35. <https://www.hhs.gov/hipaa/for-professionals/faq/does-minimum-necessary-allow-students-to-access-patient-information/index.html>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.