

The Olympics of Data Ingestions

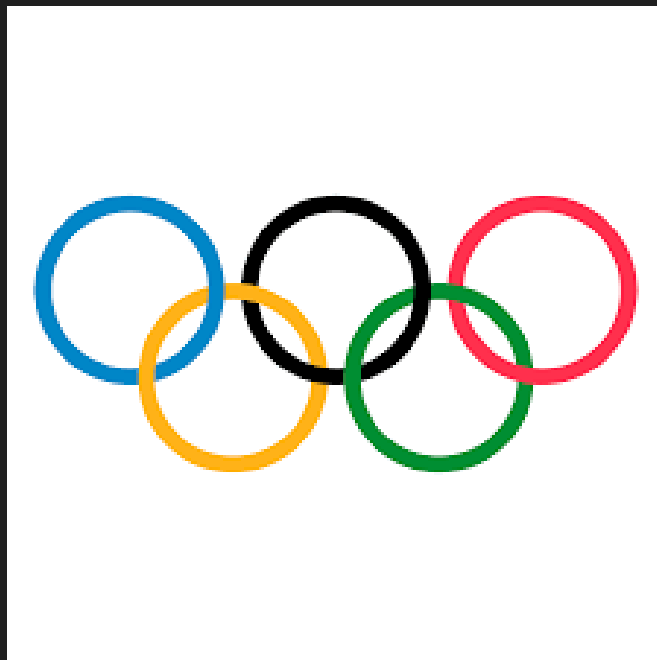
PHUSE USA 2025

DH09

Cathryn Herbst

Welcome to the Olympics of Data Ingestions

To win the gold in drug development, you need a data platform that's built for the speed, precision, and endurance of the data decathlon.



How we operate

01 Acquire drug assets

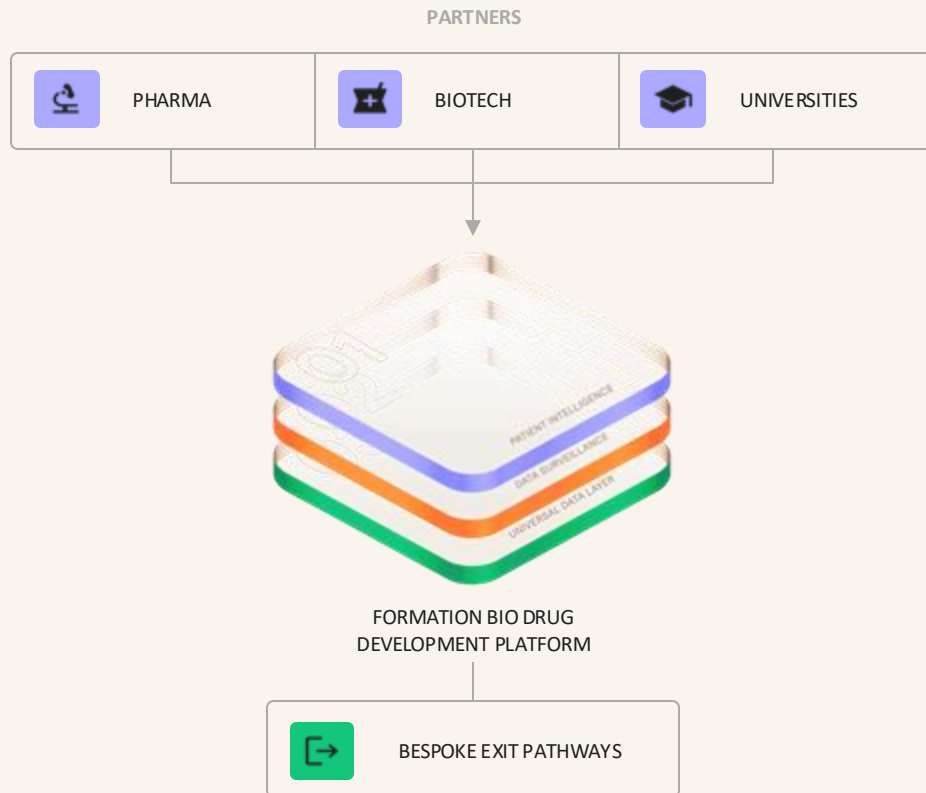
Acquire clinical-stage drugs in a variety of therapeutic areas

02 Develop drugs in-house

Leverage developmental engine to uniquely develop drugs faster and maximize value

03 Realize value inflection points

Realize increased value through bespoke exit and development pathways



Agenda

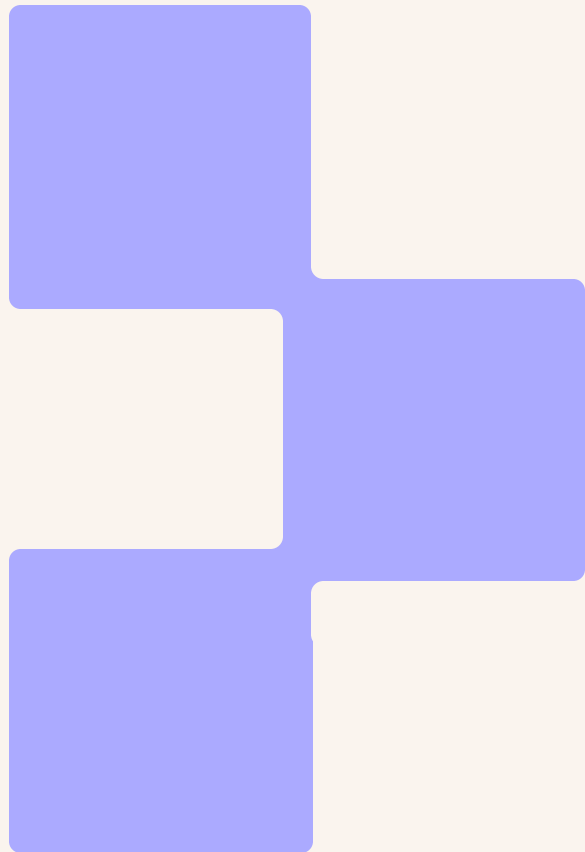
Current Data Hurdles

How to Win Gold with Your Data

- Data as a Product
- Medallion Framework
- Modern Infrastructure

Training Plan for Implementation





Current Data Hurdles

The preseason struggles:

Why teams struggle to make the most of their data

Data is fragmented across silos

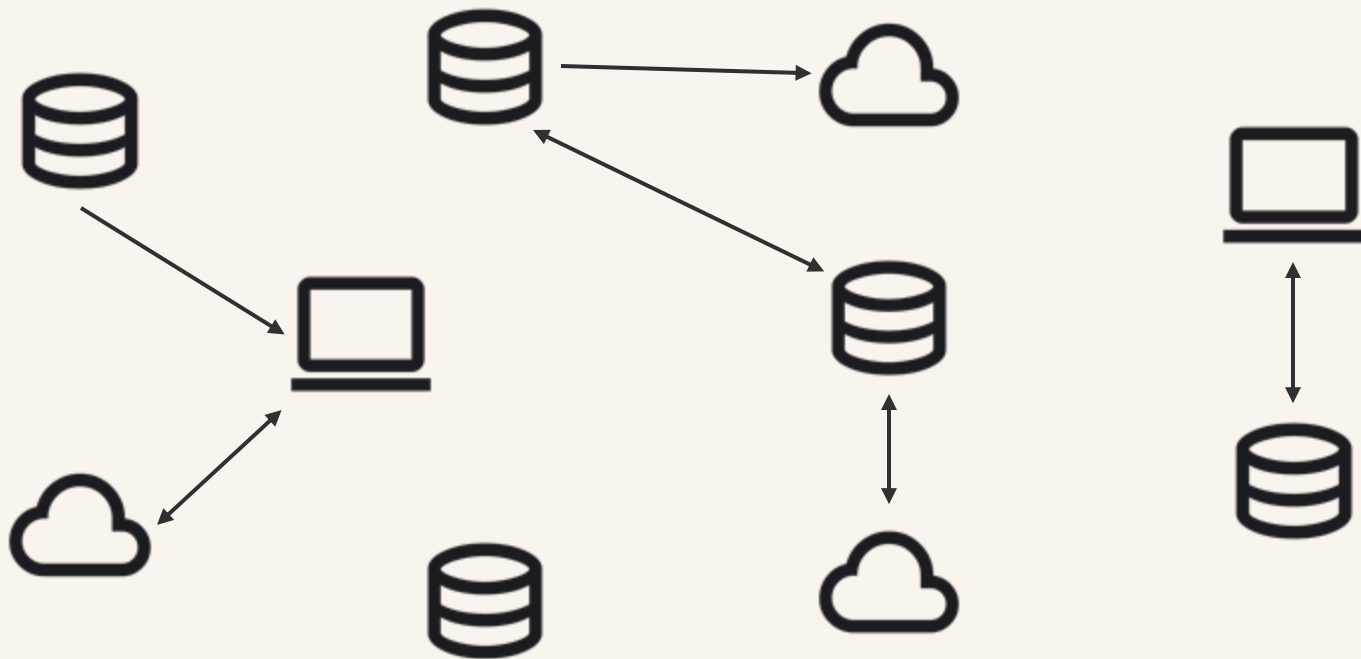
Processing is manual and slow

Teams reinvent the wheel

Data governance gymnastics



**Too many events, not enough coordination:
Data is fragmented across silos**



No real-time performance tracking: Data processing is manual and slow



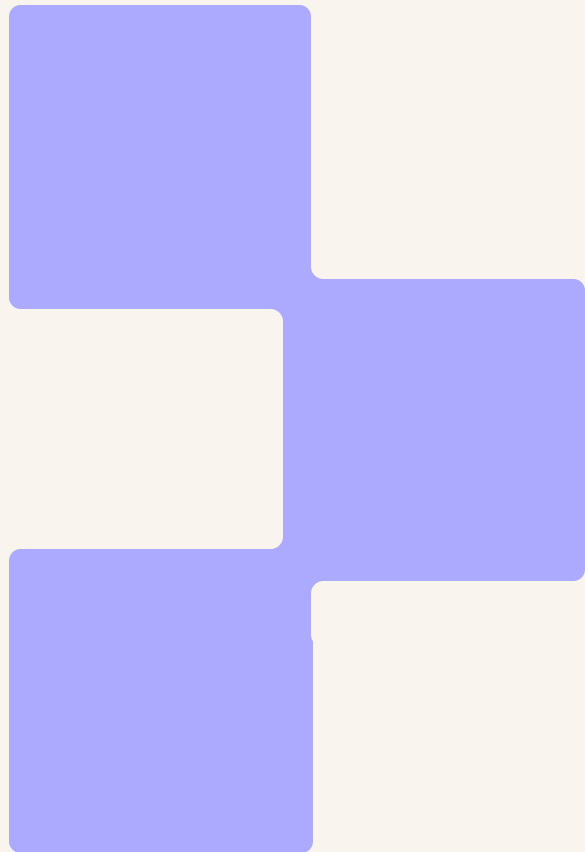
**Lack of historical context:
Reinventing the wheel for every use case**



Data governance?

More like data gymnastics





How to win Gold with Your Data

Training for gold: Modernizing data infrastructure for drug development

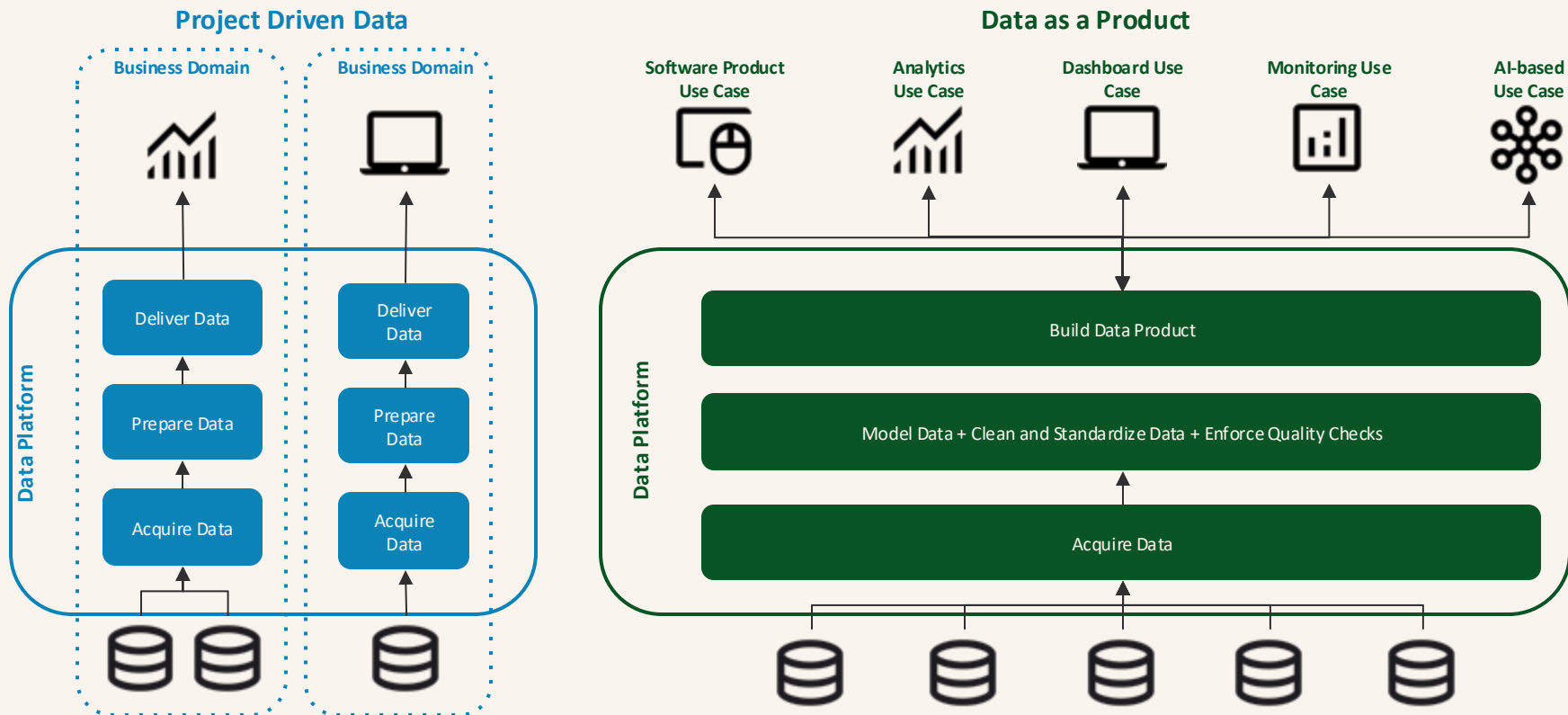
Data as a
product

Medallion framework

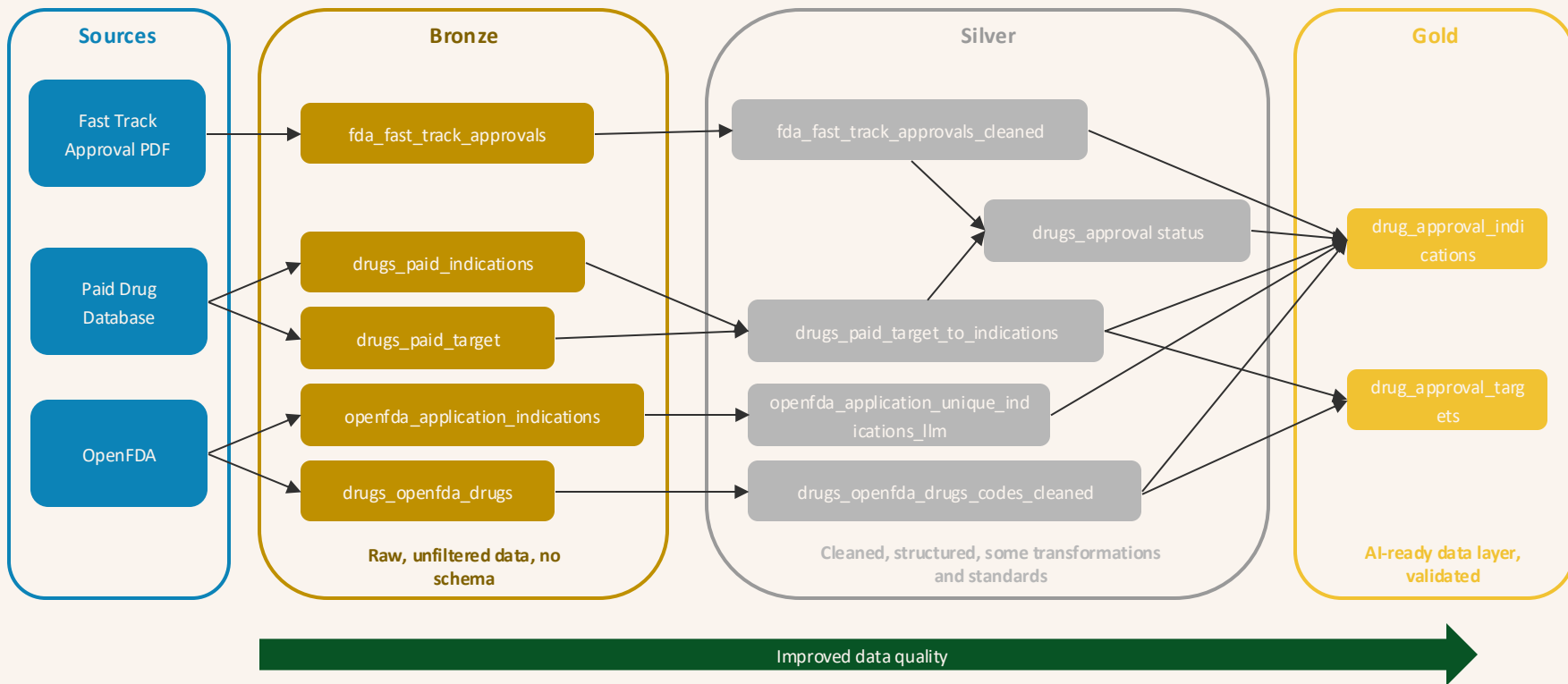
Modern data
infrastructure



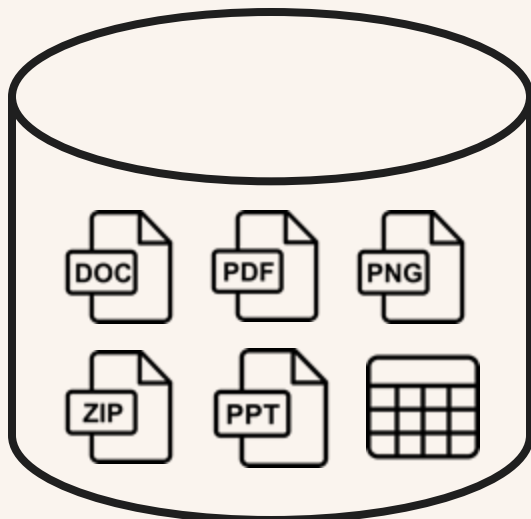
Thinking of data as a product: The key to a winning strategy



The medallion framework: Structuring data for success



Bronze medal: Taming the chaos of raw data



```
drug-drugsfda-0001-of-0001.json
{
  "submission_type": "SUPPL",
  "submission_number": "14",
  "submission_status": "AP",
  "submission_status_date": "20210428",
  "review_priority": "STANDARD",
  "submission_class_code": "LABELING",
  "submission_class_code_description": "Labeling",
},
{
  "submission_type": "SUPPL",
  "submission_number": "20",
  "submission_status": "AP",
  "submission_status_date": "20241121",
  "review_priority": "STANDARD",
  "submission_class_code": "LABELING",
  "submission_class_code_description": "Labeling",
},
{
  "application_number": "ANDA076174",
  "sponsor_name": "TARO",
  "openfda": {
    "application_number": [
      "ANDA076174"
    ],
    "brand_name": [
      "ETODOLAC"
    ],
    "generic_name": [
      "ETODOLAC"
    ],
    "manufacturer_name": [
      "Taro Pharmaceuticals U.S.A., Inc.",
      "NORTHSTAR RX LLC"
    ]
  }
}
```

FDA U.S. FOOD & DRUG ADMINISTRATION		
Home	Food	Drugs
Medical Devices	Radiation-Emitting Products	Vaccines, Blood & Biologics
Animal & Veterinary		
Search Orphan Drug Designations and Approvals		
FDA Home Developing Products for Rare Diseases & Conditions		
Results for All Designations		
Return to Orphan Products Designation Search Page		
1 - 25 of 7240		Designations: 72
#	Generic Name	Orphan Designation
1	2-((hydroxy((R)-2-(((5Z,8Z,11Z,14Z)-icosa-5,8,11,14-tetraenoyloxy)-3-(((E)-octadec-2-en-1-yloxy)propoxy)phosphoryloxy)ethan-1-ammonium	treatment of Leigh syndrome
2	Glycine, L-prolyl-L-prolyl-L-leucyl-L-seryl-L-glutamyl-L-(i-circumflex plus/minus)-glutamyl-L-threonyl-L-phenylalanyl-L-seryl-L-(i-circumflex plus/minus)-aspartyl-L-leucyl-L-tryptophyl-L-tyrosyl-L-leucyl-L-leucyl-L-tyrosyl-L-tyrosyl-L-tryptophyl-L-tyrosyl-L-methionyl-L-arginyl-L-arginyl-L-asparaginy-L-glutamyl-L-phenylalanyl-L-tryptophyl-L-valyl-L-tyrosyl-L-valyl-L-glutamyl-L-arginyl-	Treatment of Multiple Myeloma
3	N-[3-fluoro-4-[[1-methyl-6-[[1H-pyrazol-4-yl]-1H-indazol-5-yl]oxy]phenyl]-1-{4-fluorophenyl}-1,2-dihydro-6-[[1-methylethoxy]-2-oxo-3-pyridinecarboxamide	Treatment of acute myeloid leukemia
4	[[1c, 4c]-N1-(2-benzyl-7-(2-methyl-2H-tetrazol-5-yl)-9H-pyrimido[4,5-b]indol-4-yl)cyclohexan-1,4-diamine dihydrobromide dihydrate]-Expanded cord blood	Prevention of Graft-versus-Host-Disease
5	[(2S,5S)-5-(4-amino-5-(4-(2,3-difluorophenoxy)phenyl)imidazo[5,1-f][1,2,4]triazin-7-yl)pyrrolidin-2-yl)methanol	Treatment of Central Nervous System Lymphoma
6	[[4-(3-benzyl-4-hydroxybenzyl)-3,5-dimethylphenoxy)methyl]phosphonic acid	Treatment of X-linked adrenoleukodystrophy

Silver medal: Standardizing & structuring data for analysis



ID (TEXT)	INDICATION (TEXT)
87105facd2619ca72b1f644e82be5d56	effects of; anesthesia, in pregnancy
b91f3eb85cb4a387bbacd6845laab803	to measure systolic blood pressure postintubation

ID (TEXT)	OPENFDA_DRUG_APPLICATION_ID (TEXT)	UNIQUE_INDICATION_ID (TEXT)
00002151017d32af1e07e330edd20bef	ANDA207146	e96a4b57578b948c6b822256af92ae93
00049044af20elacc48687b9fbb01949	ANDA073664	9c0c6bbd73aeb349d59a85b5a0eb4bc7

0004fdfab887fe51e498b4a53
000509e8d4ce62229117b5d6c

ID (TEXT)	AACT_TRIAL_ID (TEXT)	UNIQUE_ID (TEXT)
168ba31fed5d7decc35e043e53f6415a	NCT03892330	f837711
18636872b5ae6d7cfc34b4168bf55219	NCT01955668	80a78
0feacff5c8acc225a0c7a5cca0b0bc59	NCT00744263	2c9d3

Gold medal: The trusted AI-Ready data layer



NAME	TYPE	DESCRIPTION
id	text	Unique identifier for the record
drug_id	integer	Identifier for the drug
indication_id	integer	Identifier for the medical indication

is_approved_for_indication	boolean	Flag indi	IS_APPROVED_FOR_INDICATION (BOOLEAN)	APPROVAL_DATE (DATE)	LATEST_TRIAL_COMPLETION_DATE	IS_UNDER_DEVELOPMENT_FOR_INDICATION (BOOLEAN)	IS_FAST_TRACK_APPROVED_FOR_INDICATION (BOOLEAN)	IS_ORPHAN_DRUG (BOOLEAN)
			false	---	---	false	false	false
			false	---	---	false	false	false
approval_date	date	Dot	false	---	---	false	false	false
			false	---	---	false	false	false
is_under_development_for_indication	boolean	Flag the	false	---	---	false	false	false
			false	---	---	false	false	false
			false	---	---	false	false	false
is_fast_track_approved_for_indication	boolean	Indi indi	false	---	---	false	false	false
			false	---	---	false	false	false
			false	---	---	false	false	false
is_orphan_drug	boolean	Indi the	false	---	---	false	false	false
			false	---	---	false	false	false
			false	---	---	false	false	false

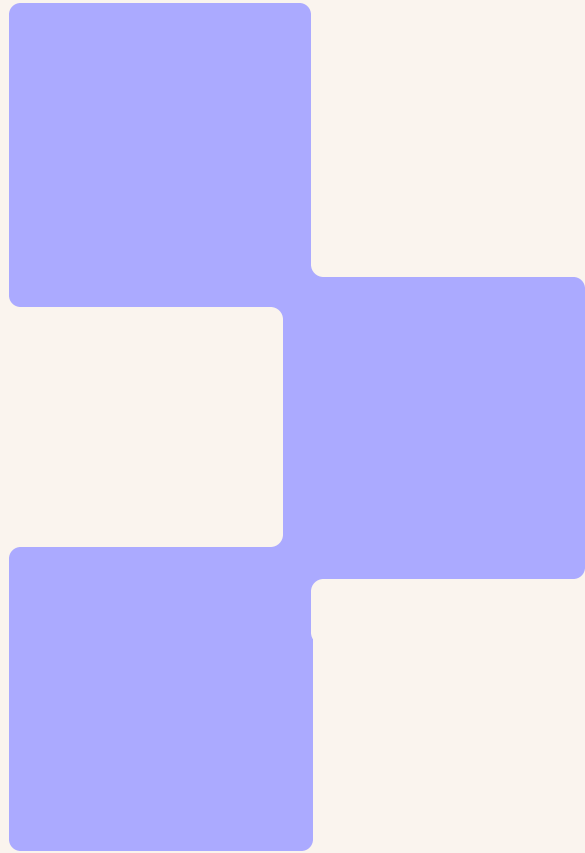
Adopting modern data infrastructure: The foundation for a winning data decathlon

Unified data platform
for cross-functional
collaboration

Scalability and
automation for faster
data processing

Evaluate data quality
and governance with
AI

Empower self-service
analytics for faster
insights



Training Plan for Implementation

Winning the drug development data decathlon

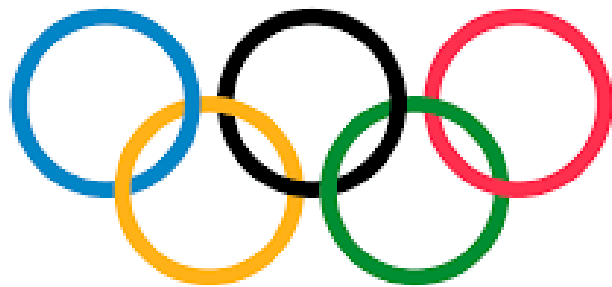
Step	What to do
1. Define and Prioritize Core Data Models for Clinical, Regulatory, and Business Needs	Define the essential data models needed for clinical, regulatory, and commercial functions to ensure interoperability and consistency.
2. Centralize Raw Data in a Data Lake (Bronze Layer)	Establish a single location (e.g., S3, Snowflake, or GCP Storage) to collect raw clinical, preclinical, regulatory, commercial, and diligence-related data. Ensure metadata tagging and lineage tracking.
3. Automate All Data Ingestions with Standardized Pipelines	Move from manual data pulls across all sources (EDCs, safety, real-world evidence, commercial, diligence) to automated extractions via APIs or scheduled batch processing.
4. Implement and Enforce Standardized Data Contracts & Schema Validation (Silver Layer)	Define and enforce common schemas for clinical, regulatory, and diligence data (e.g., SDTM/CDISC) to ensure standardization.
5. Implement Comprehensive Data Quality Checks (Silver Layer)	Add automated anomaly detection (e.g., missing values, protocol deviations, duplicate records) to all ingestion pipelines.

Winning the drug development data decathlon

Step	What to do
6. Develop and Deploy Real-Time Dashboards (Gold Layer)	Enable real-time views of enrollment, adverse events, diligence analysis, and protocol deviations with self-service analytics.
7. Integrate AI-Driven Data Harmonization (Gold Layer)	Automate mapping of clinical, regulatory, diligence, and real-world data sources into a unified model for longitudinal analysis.
8. Establish Automated Governance, Data Lineage, and Audit Trails (Gold Layer)	Enable automated lineage tracking, access controls, and data versioning to support regulatory audits and M&A diligence.
9. Enable Secure Data Sharing & Self-Service Analytics Across the Organization	Provide controlled access for biometrics, regulatory, commercial, and M&A teams to explore structured data securely.
10. Expand Advanced Predictive Analytics and AI for Business & Clinical Decision Support	Leverage AI for trend detection in clinical, regulatory, and commercial data to improve forecasting and decision-making.

Are you ready to win gold?

Where does your data stand in the Olympics? Run a data ingestion fitness test and assess whether you are operating at Bronze, Silver, or Gold.





Thank You

Cathryn Herbst (cathryn@formation.bio)