Federated Data & AI Platform (FDAP)

Jay Bergeron

Pfizer Digital

Data and Intelligence

November 17, 2026

Acknowledgements

Key Individuals Digital

- Kate Ostbye
- Shiva Kumar
- Anjali Aggarwal
- Bassam Alkhamis
- Dora Karamanidou
- Afroditi Kordella
- Peter Woodhams

Key Individuals Oncology

- Tamara Mansfield
- Paul Rejto
- Elijah Johnston
- George Khan
- Aakash Sur
- Garret Cornelison
- Chris Newman

Key Consultants

- Talia Haller
- Diego Malaver
- Samrat Banerjee

FDAP as the central hub that unifies Oncology Research data

Focus on creating AI-ready research data products

FDAP Research Assistant, Databricks GenieOne

Produce scientific insights, with corresponding rationale, through natural language prompts. Ability to inspect LLM-generated SQL statements.

FDAP Marketplace

Discover data products, their ownership and context, as well as customer-created and approved dashboards



FDAP Data Products

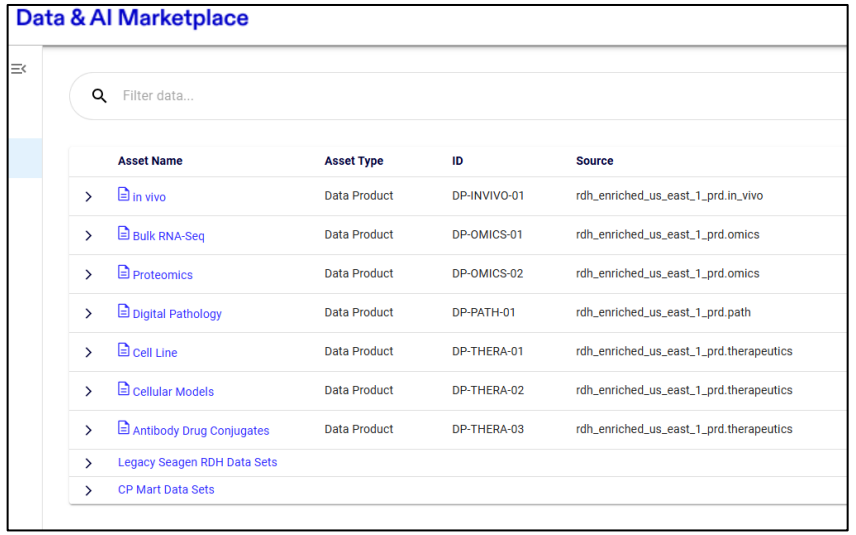
High quality domain-specialized data structures: standardized, governed, AI-ready

FDAP Platform

Robust ingestion, processing, quality, ontology, CI/CD and access solutions

Artisanal & Research Web Apps

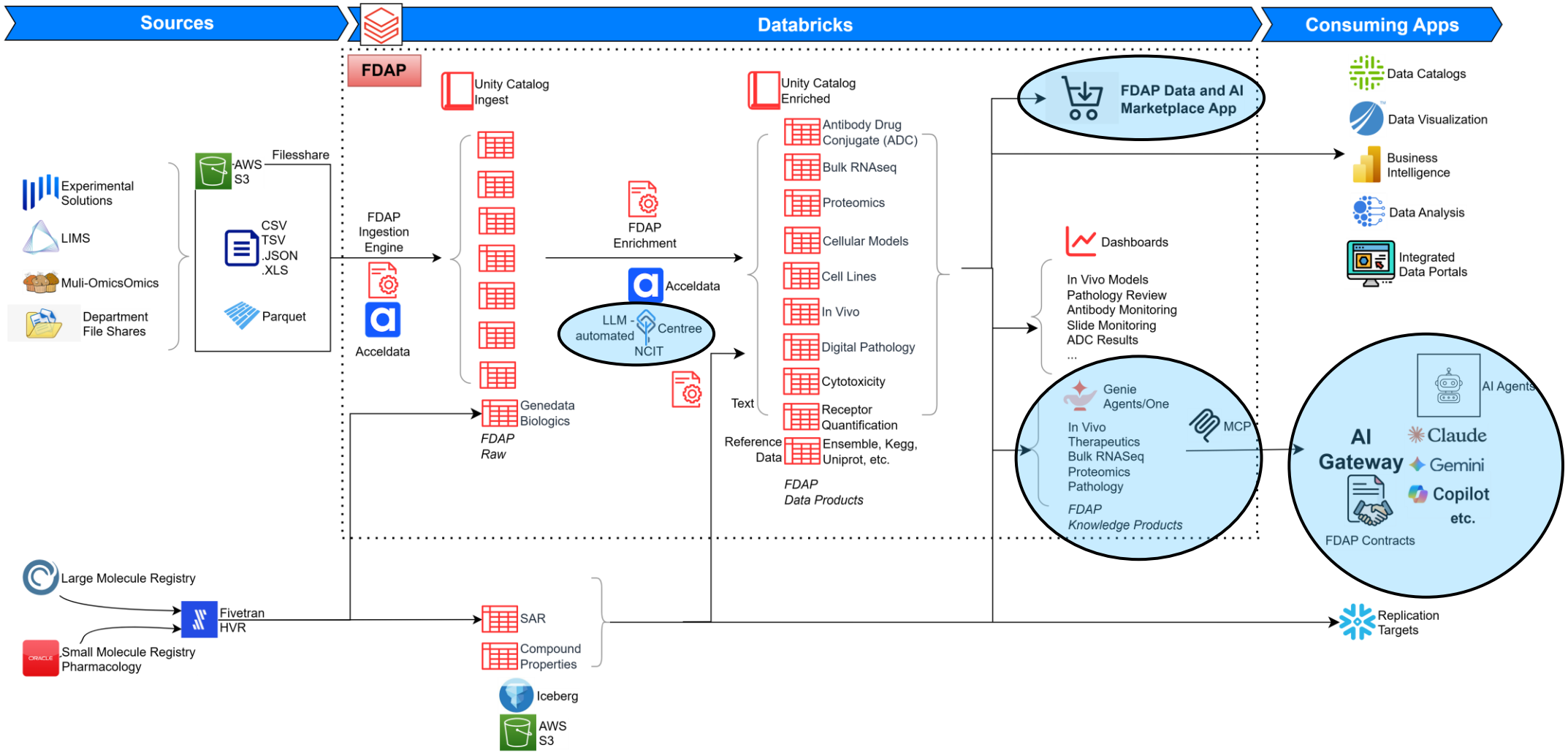
Research Tools developed by scientific staff (e.g. PFLOOKUP and In Vivo Analysis) powered by FDAP Data Products



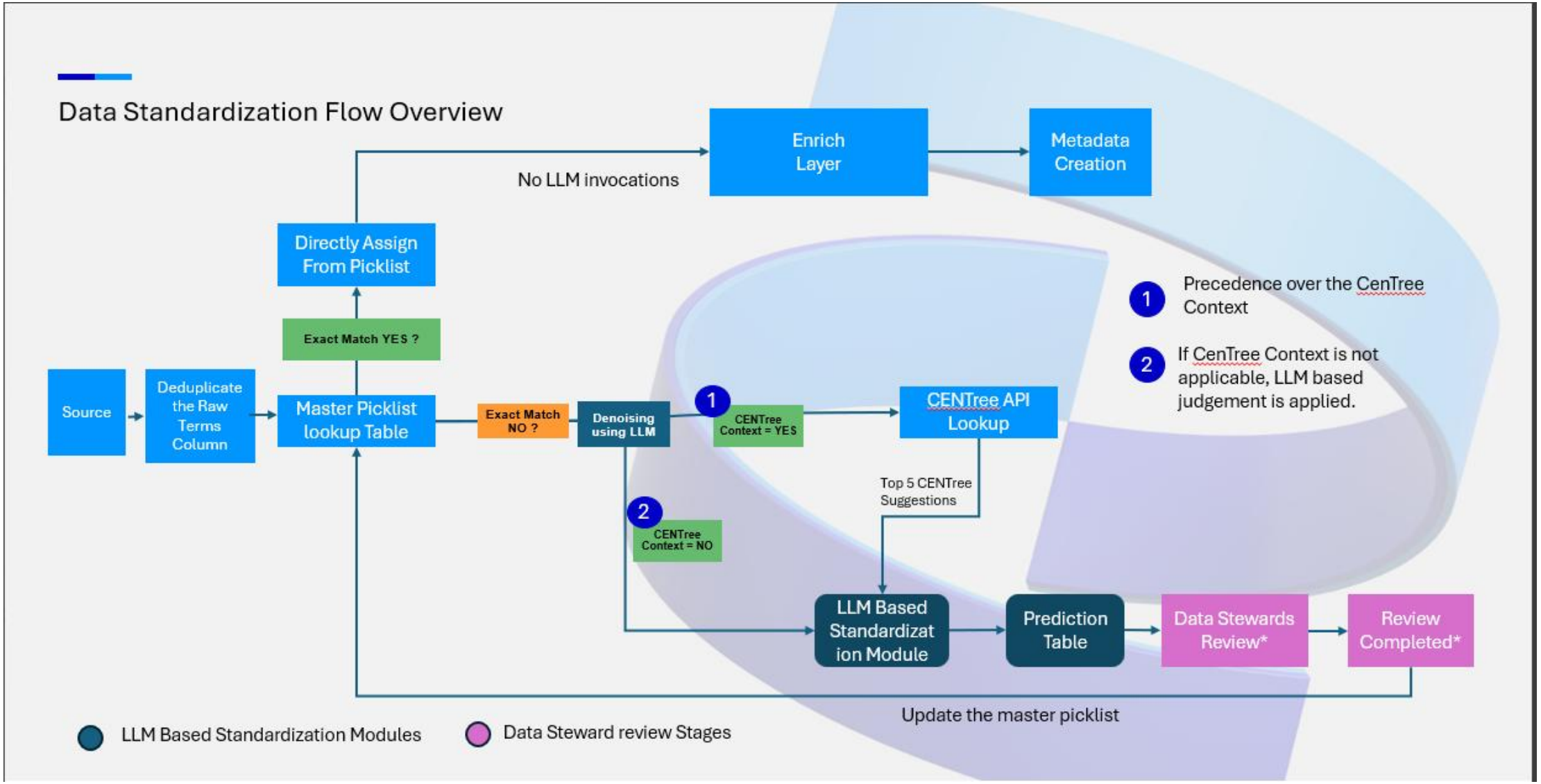
Contents

- Legacy Seagen Data Assets
- In Vivo studies
- RNA-Seq
- Proteomics
- Digital Pathology
- Cell Lines
- Cellular Models
- ADC
- Cytotoxicity
- Receptor Quantification
- Reference Data

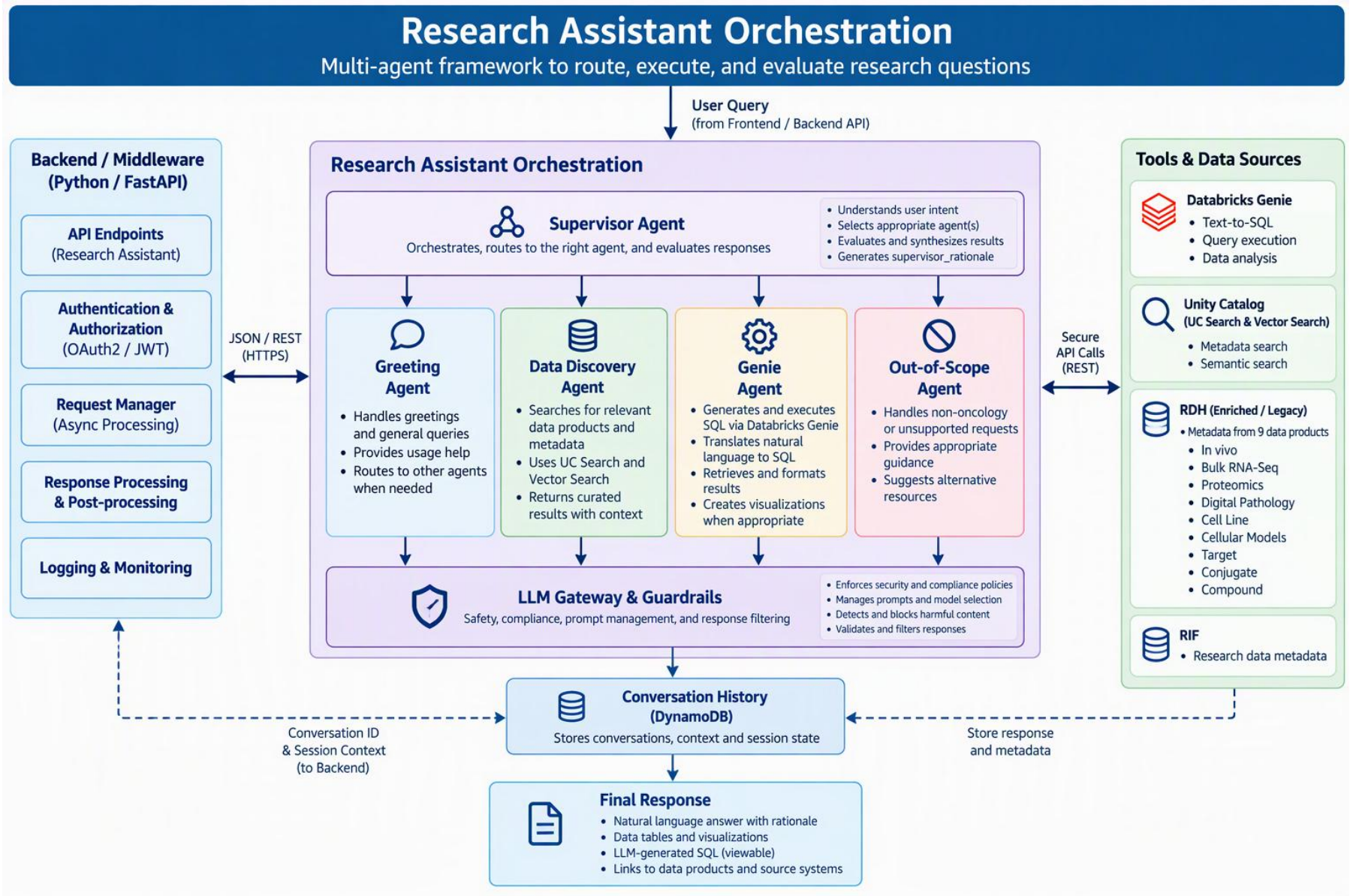
FDAP General Architecture



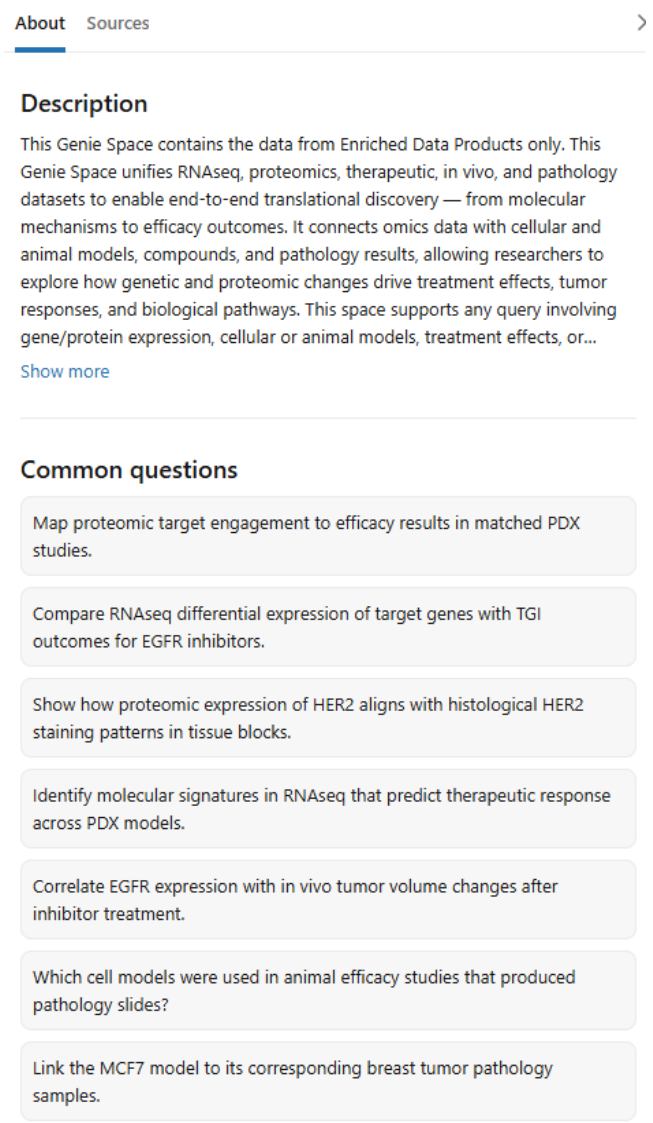
LLM-based data standardization



Research Assistant



Lakehouse Agents



The Conjugate table is the central backbone of the therapeutics schema and serves as the starting point for assembling all information related to an Antibody–Drug Conjugate (ADC). Each row represents a unique ADC identified by conjugate_id, which links to multiple other tables to provide full biological, chemical, and experimental context.
Through its key foreign key columns, the Conjugate table connects to:
• Production batches via the conjugate_batch table, enabling retrieval of all associated pf_batch identifiers.

Metadata/Descriptions/Relationships
Agent scope
Ground Truth and benchmarks

Proteomics	rdh_enriched_us_east_1_prd.omics	integrated_sample_metadata
		integrated_surface_proteomics_rnaseq
		integrated_total_proteomics_rnaseq_vw
		sgpa_surface_abundance_and_turnover_proteomics_sample
		sgpa_surface_abundance_proteomics_counts
		sgpa_surface_turnover_proteomics
		sgpa_total_proteomics_counts
		sgpa_total_proteomics_sample
	rdh_enriched_us_east_1_prd.metadata_master	tier1_metadata
		tier2_metadata
		tier3_metadata
	rdh_enriched_us_east_1_prd.therapeutics	cell_line
		cellular_models
		compound
		conjugate
		target
		reagent_antibody
		recombinant_antibody
		reagent_antibody_batch
		recombinant_antibody_batch
	rif_us_east_1_dev.cpmart	cpm_compound
		cpm_compound_batch
		cpm_compound_parent
		cpm_computed_property
		cpm_parent_name

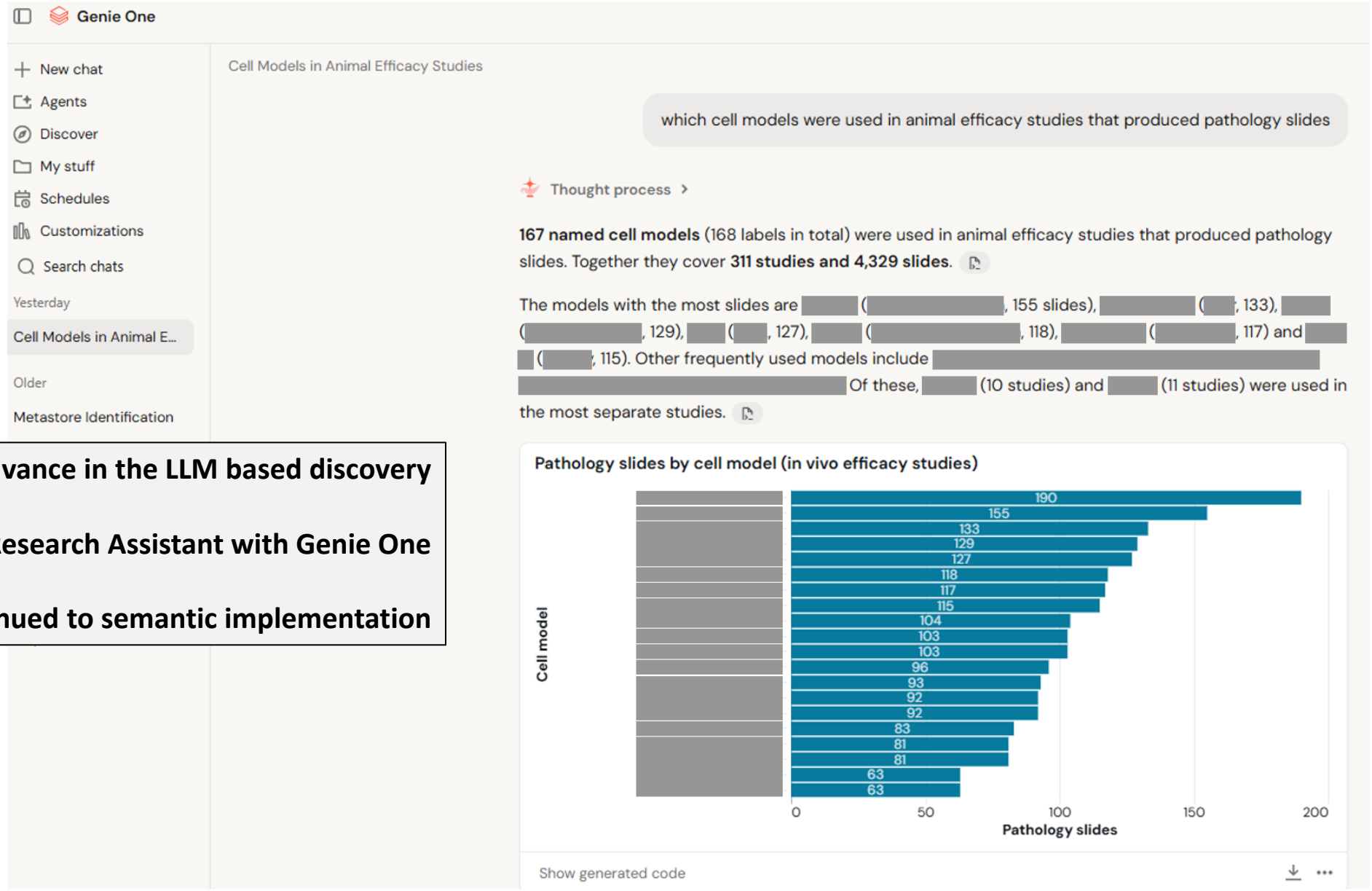
Status	Genie Space	Use Case	Query	Ideal Response (Textual	Ideas for Follow-up Questions	Ideal SQL
Ready For SME Review	Bulk RNA Seq	Gene-level expression lookup	What are the TPM counts for B2M and ACTB for Certis PDX project?	The ideal response should include: 1. Project &	•Identify and report the protein expression levels of the specified genes within the proteomics dataset. •Retrieve gene expression results in mouse for	SELECT c.global_sample_id, c.genesymbol,
Ready For SME Review	Bulk RNA Seq	Sample search by biospecimen type	List the global sample IDs of the bulk RNA studies where the sample is taken from brain	1.Purpose of Global Sample IDs Global sample IDs are used to trace samples within the rdh_enriched_us_east_1_prd.omi	Would you like to explore gene expression levels for specific genes within the brain-derived samples? Are you interested in reviewing any related proteomics studies conducted on brain samples?	WITH labeled AS (SELECT CASE WHEN sample_type_category = 'brain' THEN 'Exact match'
Ready For SME Review	Bulk RNA Seq	Cohort identification & demographics	How many PDX donors have sarcoma as the cancer type and what is the patient demographics: age, gender?	1.Ideal Response Should provide a Summary that reports:	1. Additional PDX Donor Information Would you like to view data from other PDX donors who have different cancer types?	SELECT project_name, sample_name,

April 2026

Extraordinary advance in the LLM based discovery

Replaced the custom Research Assistant with Genie One

Continued to semantic implementation



Expediting Data Onboarding

TIMELINE COMPARISON · ONBOARDING LIFECYCLE

Where the Time Goes, Step by Step

8–12 → ~4–5

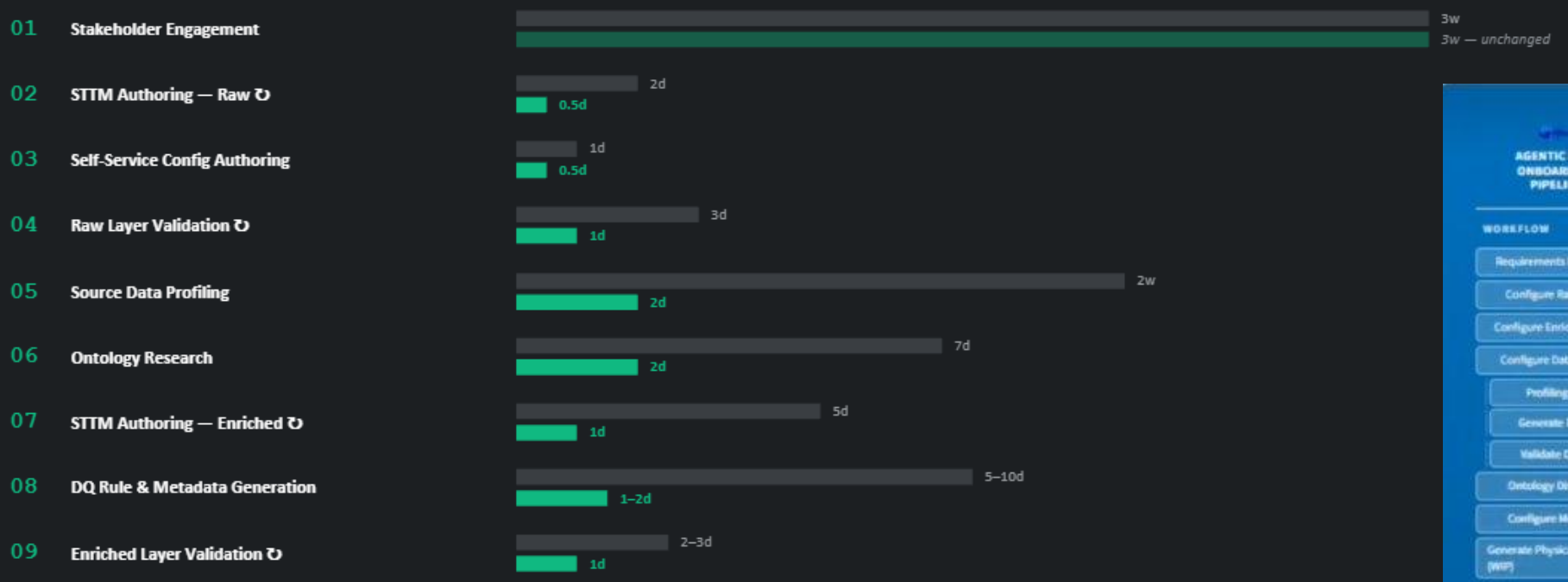
WEEKS PER DATA PRODUCT

excludes UAT & deployment

Per-stage time today vs. our planning estimate once agents are in place (Steps 2–9).

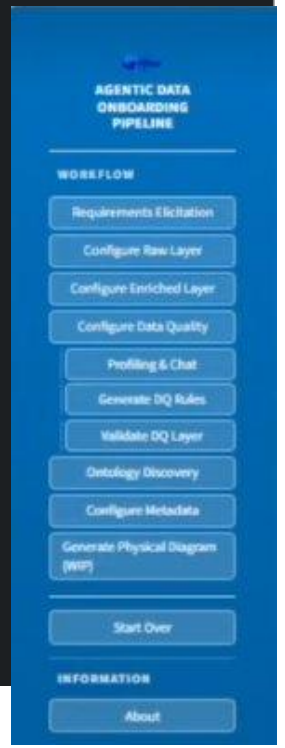
CURRENT STATE WITH AGENTS

↻ Steps 2–4 & 6–9 often loop 2–3× before sign-off



Steps 2–9: about 7–8 weeks today → about 2 weeks with agents (directional estimate, step 1 excluded).

Agent timelines are tentative planning estimates, not commitments — to be validated through the POC. UAT, environment promotion, and prod cutover run afterward and aren't reduced by agents in this view yet.



Conclusions

Reception and Use
- 20-30 accounts/month using Genie One (100-150 sessions)
- 200K-300K queries/month (traditional dashboards and applications)

Research Assistant
- Overcome by the astounding pace of foundation model development
- Semantic implementation
  - 50% of development consumption costs
    - 4-5 agents
    - ~20 benchmarks (not including scientist evaluations)
    - 30 sets of ground truth
    - ~4-5 months of time (1 primary developer and support staff)
- Use greatly lags traditional methods
  - And visualizations are exceeding easy to construct
    - Cost of token use alone is similar to that of general consumption
  - pragmatics must be monitored (access controls, comprehensive testing)

Data curation/processing
- A focus area of opportunity
- Remains a challenging/costly/time consuming task
  - even in a LLM world

FDAP is where data stop being files we move around and starts becoming knowledge product AI in use. It is a fixture in my browser as our standardized data products evolve into AI-optimized assets.

A fast, modern data lake with curated data products has enabled novel advances, powering notebooks that always reflect current data. FDAP tables have overcome long-time hurdles in aggregating molecule and target data.

FDAP is the **only solution** for standardizing and collating data across the department to make it AI ready. Hosting data in one place eliminates integration errors and significantly accelerates my work.

What's Next
Divisional and Enterprise Gateways and Orchestration