LanceDB

Universal Multimodal Data Lake with Lance and Trino

Lei Xu
Co-Founder/CTO, LanceDB

Trino Fest
June 13, 2024

Multimodal Gen AI's Challenges

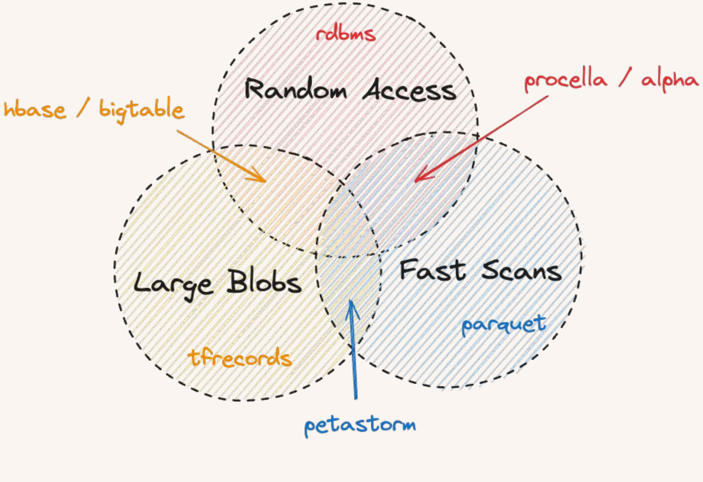
Spoil alert: it is not only about Nvidia

- Gen AI needs a different kind of "Big" data
  - **Modality**: Image, Video, Text, Embeddings
  - **Access pattern**: ingestion, training, random shuffling, and EDA
  - **Toolings**: image processing, video understanding, totalization, chunking, similarity search
- No standard practice
  - No two companies have the same stack
- Hard to build high performance AI pipelines
  - GPUs are scarce resources
  - Industry move at lightning speed, iteration speed

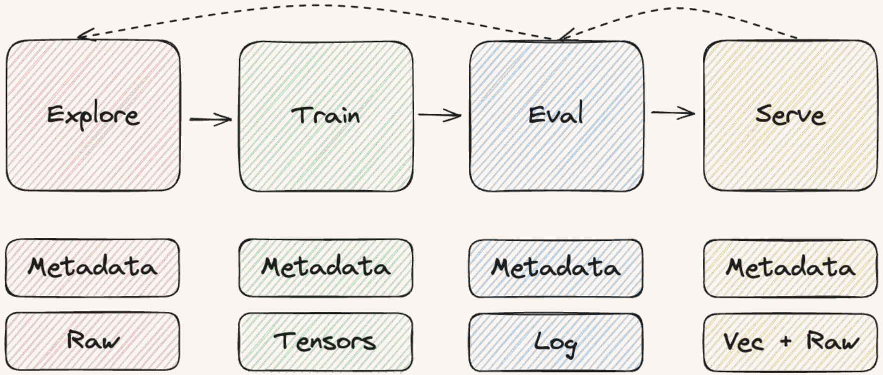
LanceDB

Not-your-father's Data Lake

- Multimodal AI - Data curation, explore, train, eval and serving
  - Parquet-based Data Lakes optimize for large-scan of primitive values
  - Shuffling and serving of large blobs are hard
- But people still love how easy the data toolings were in the old days
  - DataFrame, SQL, Schema Evolution



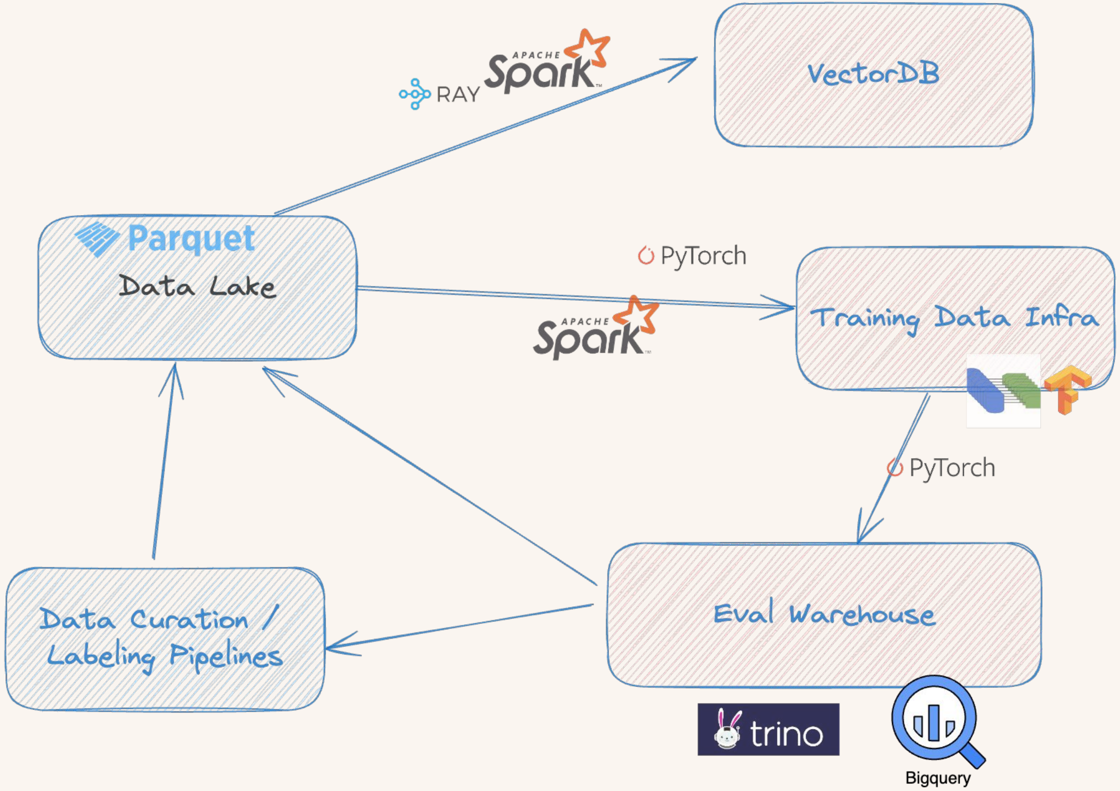
CAP Theorem for AI Data



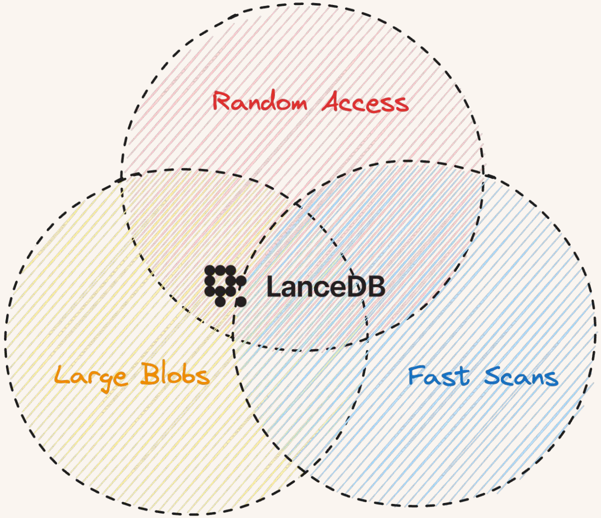
LanceDB

Multiple Single Purpose Systems

For different workloads



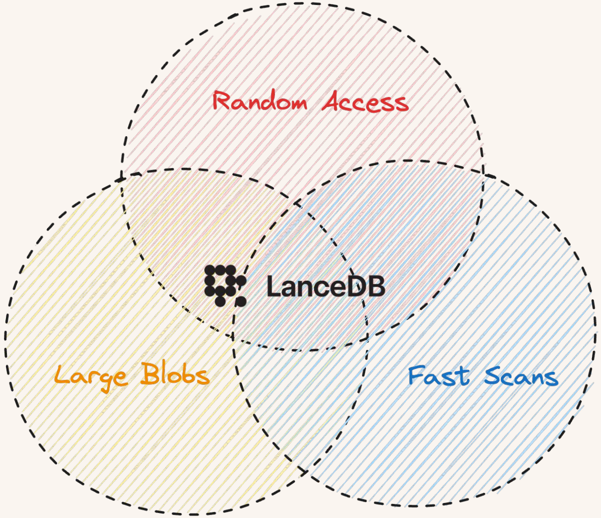
What if we can have it all?



~~CAP Theorem for AI Data~~

Have your cake and eat it too

What if we can have it all?



~~CAP Theorem for AI Data~~

Have your cake and eat it too

"..was evaluating LanceDB as our VectorDB, until our ML team told me that it can serve ML training...and easy feature engineering (via Zero-cost Schema Evolution)...My team is going to use Lance(DB) to replace two separated systems. I wish this technology existed 5 years ago"

LanceDB

A developer-friendly, open source multimodal database for AI.

Developer-friendly

- Apache Arrow and Dataframe based APIs
- Python, Typescript, Rust, and Java SDKs
- SQL-filter, Vector Search, Fts, Hybrid-search, reranking, embedding functions

Open Source

- Start with an embedded database
  - Similar to SQLite, but for vector search

Scale to Hyper-scale (Billions of Records)

- Low latency and high QPS
- Compute-Storage separation, easy to deploy
- Indices on S3, GCS, Azure or filesystems.
- Multi-tenancy.
- One system to serve vector, image, video and All.

LanceDB

Built on Lance Columnar Format

Solving the multimodal data problem from the ground up

File format

- Columnar format for fast scan
- Near O(1) random access
- Advanced encodings
- Store large blobs inline
- No row groups and prefetch-friendly layout
- Optimize for NVME SSD and Object Store

Table format

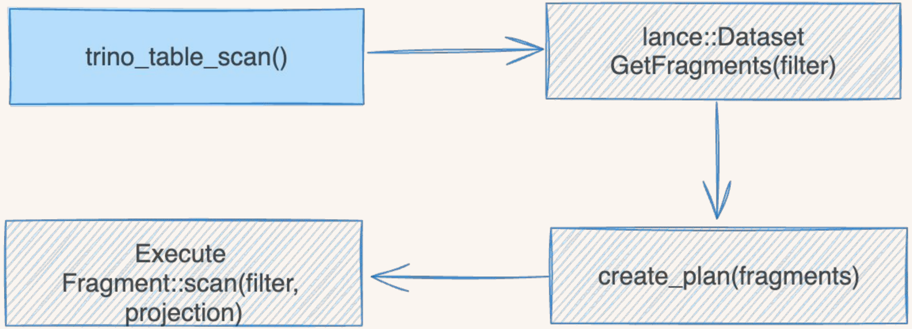
- Global Stable Row ID
- Versioning
- Zero-copy Schema Evolution
- WAL and LSM (*)
- Apache-arrow based Dataset API
- I/O execution plan optimized for randomly access large blobs in the dataset

Indexing

Built on its fast random access capability, Lance has an extensible sub-system to support various of disk-based indices.

- Scalar Index (B-Tree)
- Vector Index (IVF_PQ, IVF_HNSW, HNSW, …)
- Full text search (*)

Trino LanceDB Connector - In Development


trino_table_scan()
lance::Dataset GetFragments(filter)
create_plan(fragments)
Execute Fragment::scan(filter, projection)

Trino-LanceDB plugin (in-development)

- Thanks Rong Rong @ Character.ai
- Communicate Lance Rust core + JNI via Apache Arrow
- Distributed Execution Plan
  - Using Native Lance Fragment
  - Predicates and Project Pushdown (2)
- Directory based Catalog
- Run complex analysis on Trino
- Flexible adaptation of Dataset/DataLoading scheme

1. https://github.com/trinodb/trino/pull/21880
2. https://github.com/walterddr/trino/pull/2

Trino LanceDB Connector – Roadmap

- Support writes
  - Need more typing system support for multi-modal data
- Support vector search via remote_table_scan
- Create index via trino functions
- DDL to support schema evolutions via Lance
- Trigger distributed compactions jobs
- Multi-OS/arch release

https://github.com/trinodb/trino/pull/21880

LanceDB

We’re Hiring

Join Our Discord

LanceDB is a developer-friendly, open source database for multimodal AI. From hyper scalable vector search and advanced retrieval for RAG, to streaming training data and interactive exploration of large scale AI datasets, LanceDB is the best foundation for your AI application:

- https://lancedb.com/
- https://github.com/lancedb/lance
- LanceDB Cloud is in private beta!
- HIRING: Database Engineers

Trino + LanceDB @Character.AI

Accelerating LLM Research

Who am I?

Noah Shpak 👋

+ the team:

Nat Roth
Tian Xie
Ryan Vilim
Rong Rong
Sam Bean
Katherine Yoshida
Haowen Ge
Peng Xu
Adam Zhang

character.ai

character.ai

Personal AI Platform
Full-Stack AI Company

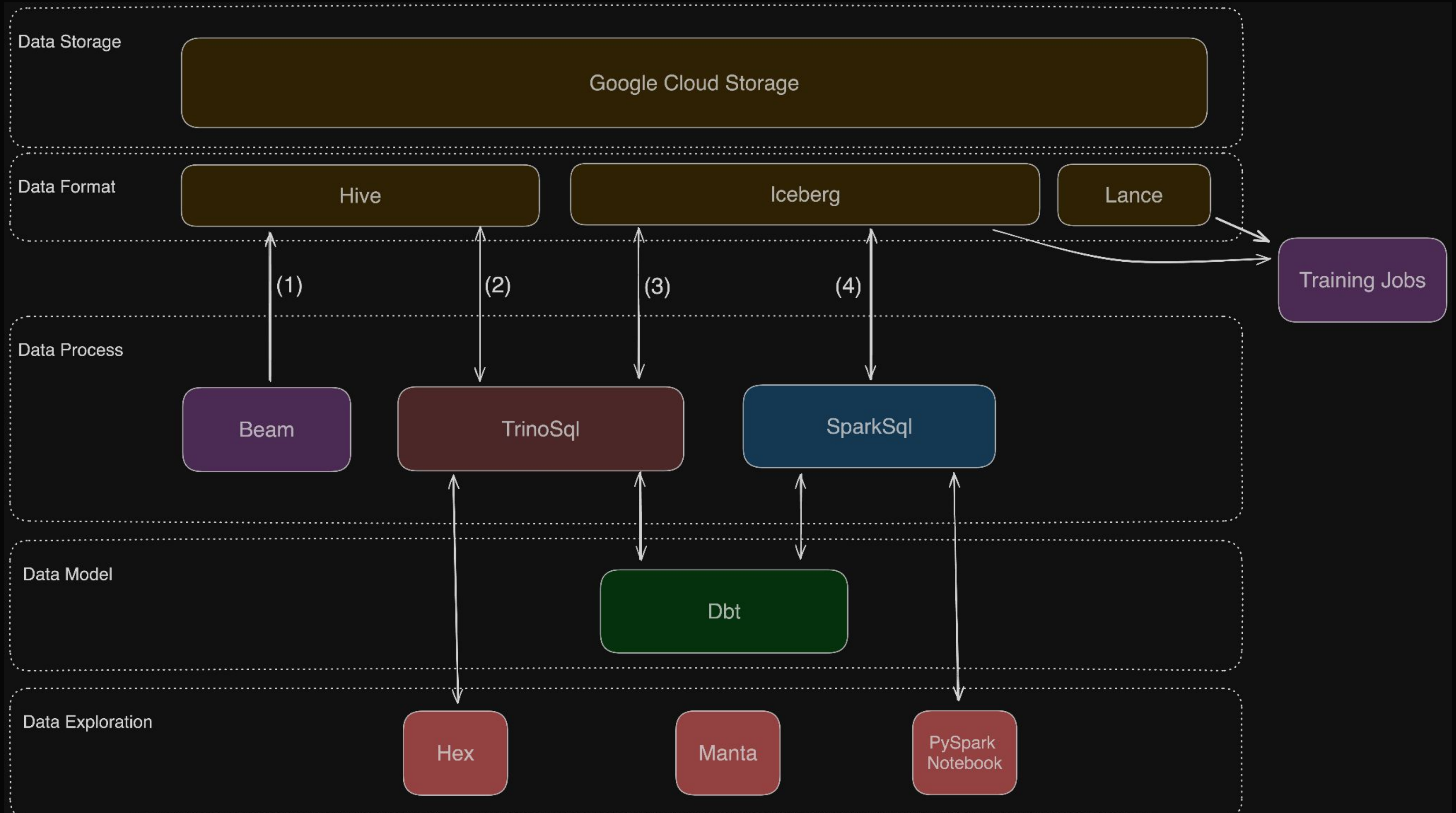
What is Platform @Character.AI?

LLM Research on Data Mixtures

GPU Services, Data Catalog, Infra & Tooling

The Data Lake

Making data go vroom while GPUs go brrr



Trino @Character

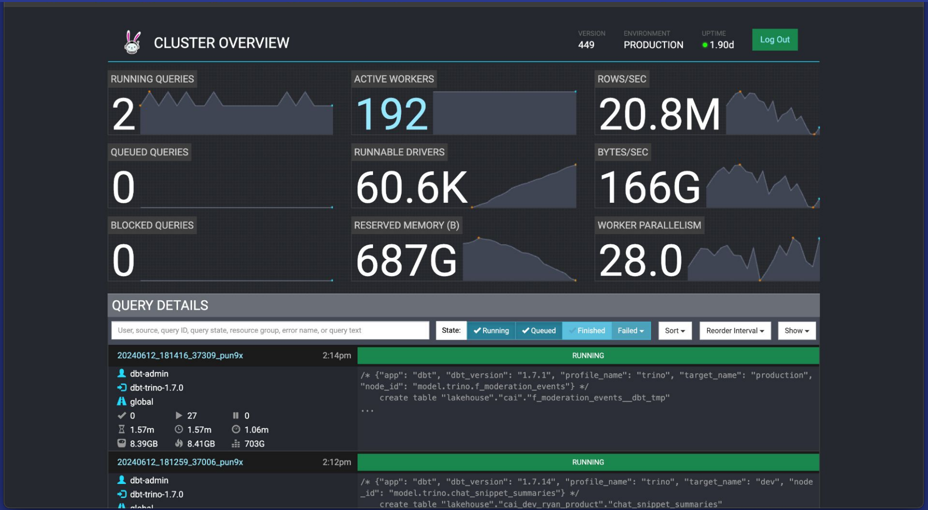
2 K8s Deployment

> 100s of queries / day

> thousands of cores

> 100TB of RAM

> 0.5TB of NVME fs cache



CLUSTER OVERVIEW

VERSION 449 · ENVIRONMENT PRODUCTION · UPTIME ● 1.90d · Log Out

RUNNING QUERIES 2

ACTIVE WORKERS 192

ROWS/SEC 20.8M

QUEUED QUERIES 0

RUNNABLE DRIVERS 60.6K

BYTES/SEC 166G

BLOCKED QUERIES 0

RESERVED MEMORY (B) 687G

WORKER PARALLELISM 28.0

QUERY DETAILS

User, source, query ID, query state, resource group, error name, or query text

State: ✓ Running ✓ Queued ✓ Finished Failed ▾ Sort ▾ Reorder Interval ▾ Show ▾

20240612_181416_37309_pun9x 2:14pm RUNNING

dbt-admin
dbt-trino-1.7.0
global
✓ 0 ▶ 27 ⏸ 0
1.57m 1.57m 1.06m
8.39GB 8.41GB 703G

```
/* {"app": "dbt", "dbt_version": "1.7.1", "profile_name": "trino", "target_name": "production",
"node_id": "model.trino.f_moderation_events"} */
    create table "lakehouse"."cai"."f_moderation_events__dbt_tmp"
...
```

20240612_181259_37006_pun9x 2:12pm RUNNING

dbt-admin
dbt-trino-1.7.0
global

```
/* {"app": "dbt", "dbt_version": "1.7.14", "profile_name": "trino", "target_name": "dev", "node
_id": "model.trino.chat_snippet_summaries"} */
    create table "lakehouse"."cai_dev_ryan_product"."chat_snippet_summaries"
```

Spark + Trino @Character

Self-hosted on Kubernetes

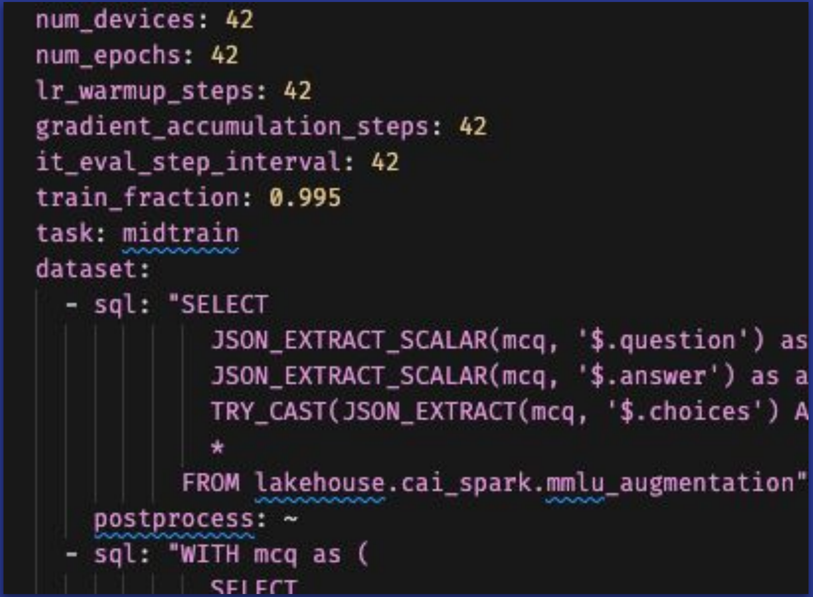
Data Processing & Analytics

Custom UDFs for AI Capabilities

- Prompting, Embedding, etc

Trino + Lance for Training

Rapid Prototyping



```
num_devices: 42
num_epochs: 42
lr_warmup_steps: 42
gradient_accumulation_steps: 42
it_eval_step_interval: 42
train_fraction: 0.995
task: midtrain
dataset:
 - sql: "SELECT
      JSON_EXTRACT_SCALAR(mcq, '$.question') as
      JSON_EXTRACT_SCALAR(mcq, '$.answer') as a
      TRY_CAST(JSON_EXTRACT(mcq, '$.choices') A
      *
     FROM lakehouse.cai_spark.mmlu_augmentation"
  postprocess: ~
 - sql: "WITH mcq as (
      SELECT
```

More Trino @Character

Iceberg Optimization

Query entire catalog

Export for Training

Analytics for Ingestion

Lance @Character

Search & Retrieval

Training Format

Analytics

Unification of Data Lake

Researchers

Data People

Data Recipes

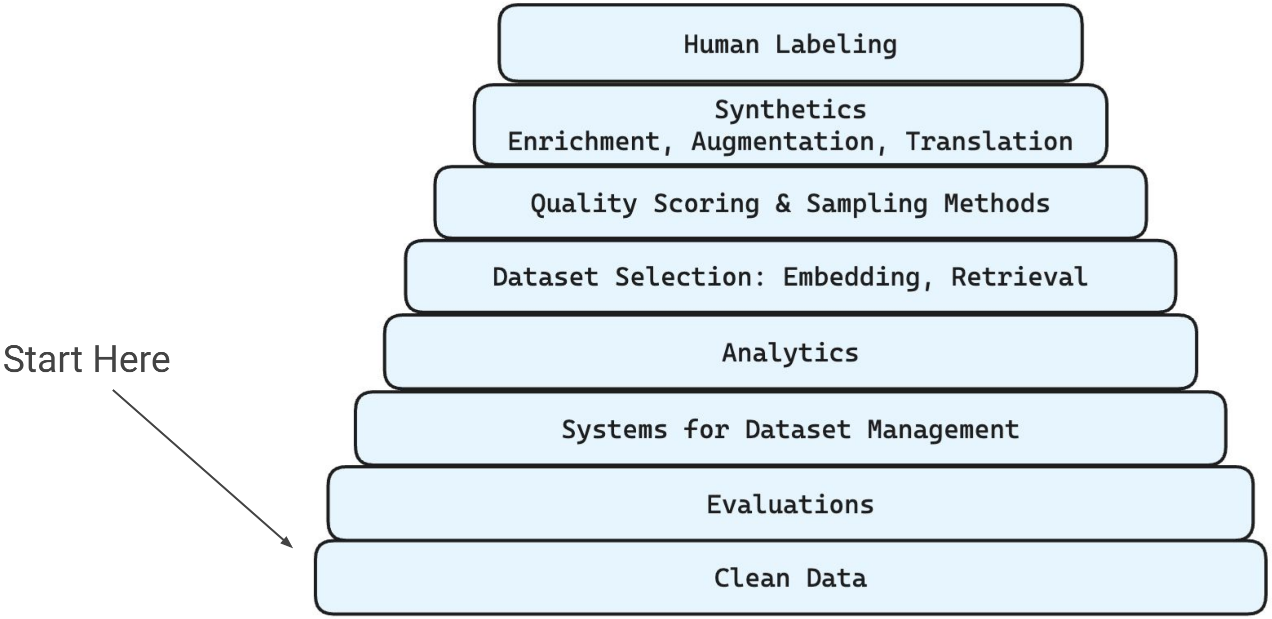
Token Budgets
Mixture Weights
Sampling Methods
Pre-processing
Prompting Mechanisms
…

Data Research

```
⛏️ + 📖
   →
   evaluate(
       train(data + 🧪))
         → SOTA
```

Hierarchy of Needs

for dataset development



Human Labeling

Synthetics
Enrichment, Augmentation, Translation

Quality Scoring & Sampling Methods

Dataset Selection: Embedding, Retrieval

Analytics

Systems for Dataset Management

Evaluations

Clean Data

Start Here

Lance for Multimodal

Strictly more difficult engineering problem

Storage becomes a bigger problem

Data Loaders have to become more involved

Excited to solve this with Lance

Thank You!