THE COGNITIVE DATA SUPERCYCLE

Defining the Cognitive Data Stack (CDS)

The Infrastructure for the Next Trillion Dollars of Enterprise Value

by Hernan Asorey, Co-Founder, AVC Turing

March 2026

Investing in the Future of Intelligence

New York | Global

This paper is a research-driven market architecture and strategic framework. Company references and valuation snapshots are illustrative case studies, not endorsements or investment recommendations. See Disclaimers (Appendix C).

Published by AVC Turing Ltd.

The Cognitive Data Stack (CDS) framework was introduced by Hernan Asorey, Co-Founder of AVC Turing. Citation and reference to the framework is encouraged with attribution to the author.

1. Executive Summary: The Structural Shift

The "Modern Data Stack" (MDS) optimized the enterprise for retrospective analytics: decouple storage and compute, batch-transform data, and publish dashboards for humans. The **Cognitive Data Stack (CDS)** optimizes the enterprise for probabilistic reasoning and automated action: continuously refine truth, govern it, and safely write decisions back into operations.

CDS emerges from a structural decoupling in the AI market: model capability is expanding and diffusing rapidly (commercial competition + open-source pressure), while durable value increasingly accrues to the constraints: proprietary data, data locality, compliance, and the industrial systems that operationalize reasoning.

We map CDS into three layers, plus a cross-cutting Control Plane:

- **Physics Layer:** overcome data gravity and I/O limits so massive compute can “touch” the right data with minimal latency.
- **Genesis Layer:** manufacture truth via RLHF, expert feedback, evaluation, and the digitization of privileged vertical data assets.
- **Activation Layer:** close the loop, turn predictions into controlled actions inside systems of record (CRMs, ERPs, public-sector systems).
- **Control Plane (cross-cutting):** governance, identity, observability, evaluation, and safety rails that separate pilots from production autonomy.

2. The Thesis: From Deterministic Storage to Probabilistic Reasoning

Big Data (≈2010–2020) was defined by volume: store everything. MDS (≈2020–2024) was defined by organization: clean and model data for analytics. We are now in the **Cognitive Data era (≈2025–present)**: data is no longer a static asset for reporting; it is dynamic fuel for probabilistic reasoning systems that act.

In this paradigm, the primary constraints shift from "can we store it?" to: (i) can we move the right data to the right compute at the right time, and (ii) can we manufacture ground truth that models can reliably learn from.

The term "Cognitive Data Stack" synthesizes patterns that are independently observable across the industry. The IEA projects AI-related data center electricity consumption will more than double by 2030, reflecting the Physics Layer's growing infrastructure demands. Gartner forecasts that 40% of enterprise applications will embed task-specific AI agents by end of 2026, up from low single digits in 2024, a direct measure of Activation Layer adoption. The Linux Foundation's Agentic AI Foundation (co-founded December 2025 by OpenAI, Anthropic, Google, Microsoft, AWS, and Block) standardizes the protocol layer between agents and tools, institutionalizing the connective tissue this paper describes. The EU AI Act (obligations phasing in 2025–2027) is creating mandatory demand for exactly the governance primitives the Control Plane provides. CDS is a descriptive framework for an observable structural shift, not a proprietary thesis.

2.1 Core Definitions (Operational)

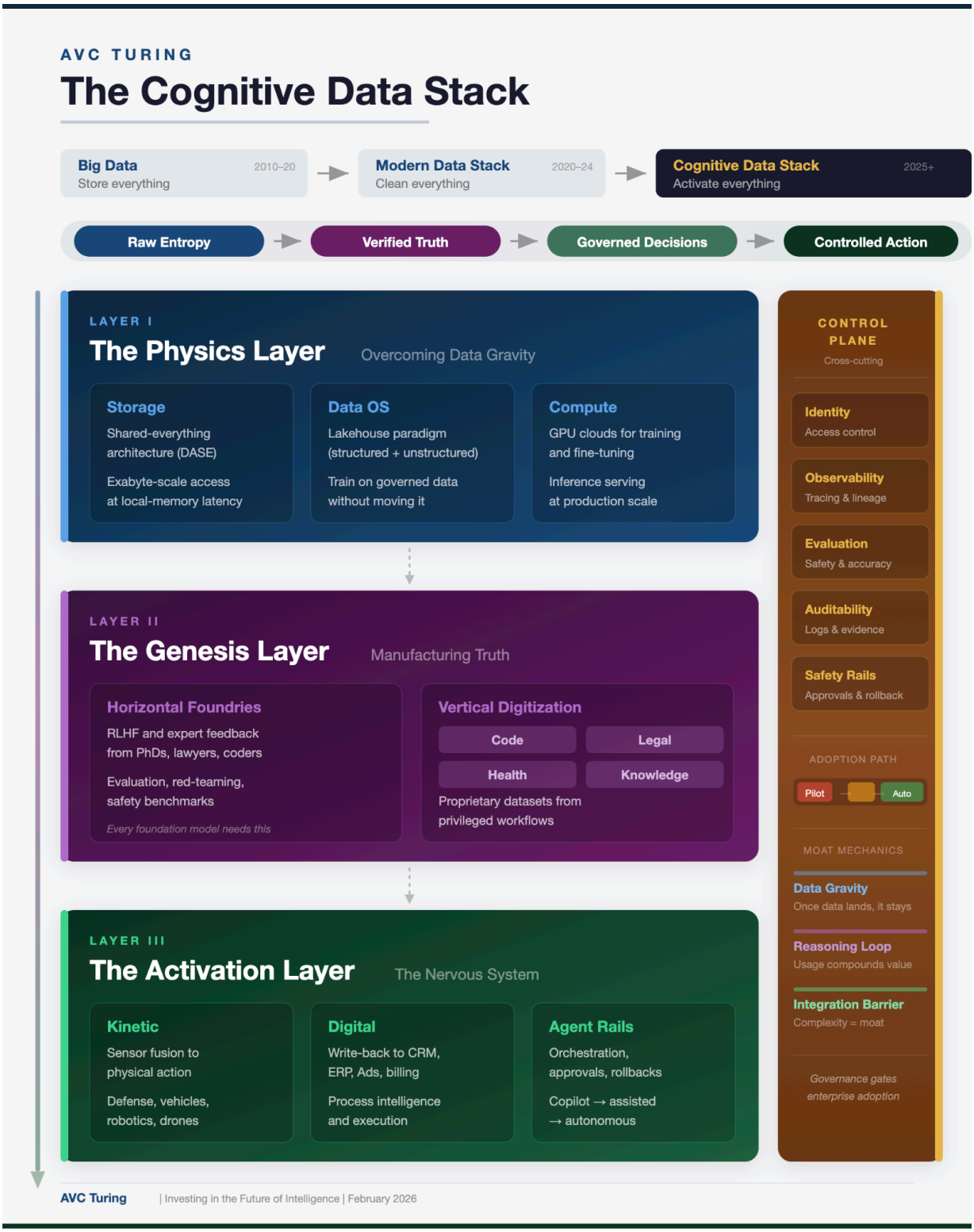
- **Data Gravity:** the practical physics and economics of moving large datasets; value concentrates where the data already lives.
- **Data Veracity:** the production of ground truth (labels, preferences, evaluations, corrections) that reduces model entropy and hallucination.
- **Reasoning Data:** any data explicitly shaped to teach, evaluate, or constrain model behavior (RLHF, expert traces, verified corpora, safety tests).
- **Activation:** the controlled capability to write back into operational systems (not just observe).

3. The CDS Framework: Physics, Genesis, Activation, Plus the Control Plane

CDS is best understood as an industrial pipeline: **raw entropy → verified truth → governed decisions → controlled action**. Each layer has distinct buyers, failure modes, and moat mechanics.

Derivation. The CDS layers are not arbitrary categories; they are derived from four persistent engineering constraints that remain regardless of which model architecture prevails. First, data must be physically moved to compute (or compute to data) at speeds that prevent hardware from idling: this is the Physics constraint. Second, models require ground truth to learn from, and high-quality public data is largely exhausted: this is the Genesis constraint. Third, predictions are economically valueless until they change a system of record: this is the Activation constraint. Fourth, autonomous action without governance is unacceptable to regulated enterprises: this is the Control Plane constraint. We considered and rejected alternative taxonomies (e.g., a two-layer model collapsing Genesis and Activation; a model-centric taxonomy organized by foundation model provider). Both alternatives failed to explain where value was actually concentrating in private markets. The 3+1 structure was selected because it maps to distinct buyer profiles, procurement budgets, and moat mechanics that we observe empirically.

The Control Plane is not “nice to have.” As autonomy rises, governance, evaluation, and auditability become the gating factor for deployment. In practice, the Control Plane often determines which platforms can cross the chasm from experimentation to enterprise-wide adoption.


AVC TURING
The Cognitive Data Stack
Big Data
2010–20
Store everything
Modern Data Stack
2020–24
Clean everything
Cognitive Data Stack
2025+
Activate everything
Raw Entropy
Verified Truth
Governed Decisions
Controlled Action
LAYER I
The Physics Layer
Overcoming Data Gravity
Storage
Shared-everything architecture (DASE)
Exabyte-scale access at local-memory latency
Data OS
Lakehouse paradigm (structured + unstructured)
Train on governed data without moving it
Compute
GPU clouds for training and fine-tuning
Inference serving at production scale
LAYER II
The Genesis Layer
Manufacturing Truth
Horizontal Foundries
RLHF and expert feedback from PhDs, lawyers, coders
Evaluation, red-teaming, safety benchmarks
Every foundation model needs this
Vertical Digitization
Code
Legal
Health
Knowledge
Proprietary datasets from privileged workflows
LAYER III
The Activation Layer
The Nervous System
Kinetic
Sensor fusion to physical action
Defense, vehicles, robotics, drones
Digital
Write-back to CRM, ERP, Ads, billing
Process intelligence and execution
Agent Rails
Orchestration, approvals, rollbacks
Copilot → assisted → autonomous
CONTROL PLANE
Cross-cutting
Identity
Access control
Observability
Tracing & lineage
Evaluation
Safety & accuracy
Auditability
Logs & evidence
Safety Rails
Approvals & rollback
ADOPTION PATH
Pilot
Auto
MOAT MECHANICS
Data Gravity
Once data lands, it stays
Reasoning Loop
Usage compounds value
Integration Barrier
Complexity = moat
Governance gates enterprise adoption
AVC Turing | Investing in the Future of Intelligence | February 2026

3.1 Scope, Selection Criteria, and Evidence Base

This paper profiles 19 private companies selected using three criteria: (i) primary value proposition maps to one or more CDS layers; (ii) valuation at or above approximately $1B based on publicly reported private-market transactions; (iii) company remains privately held as of Q1 2026, with public companies acknowledged for context where structurally relevant. Model providers (OpenAI, Anthropic, Google DeepMind, Mistral, Cohere) are excluded: CDS maps the infrastructure that feeds and activates models, not the model layer itself. Public infrastructure companies (CoreWeave, Snowflake, Palantir, Datadog) are referenced where they intersect CDS layers but are not profiled as exemplars. The 19-company set is illustrative, not exhaustive; additional private companies operate within each layer and are noted where relevant.

3.2 CDS Value Map: Enterprise Value by Layer

The following summary aggregates reported or estimated private-market valuations across the 19 profiled companies. All figures are approximate, derived from publicly available sources, and subject to the limitations described in Appendix C.

Physics Layer (5 companies: VAST Data, WEKA, Databricks, Lambda, Fireworks AI): ~$183B aggregate. Dominated by Databricks ($134B), reflecting the Data Lakehouse's gravitational pull as the enterprise AI operating system.

Genesis Layer (8 companies: Scale AI, Surge AI, Snorkel AI, Anysphere/Cursor, Glean, Harvey, Abridge, Hippocratic AI): ~$108B aggregate. Split roughly evenly between horizontal foundries (~$55B) and vertical digitizers (~$53B), validating both sub-markets.

Activation Layer (6 companies: Anduril, Applied Intuition, Celonis, Peregrine, Hightouch, LangChain): ~$63B aggregate. Anduril ($30.5B) anchors the kinetic segment; Celonis ($13B) and Applied Intuition ($15B) lead digital and simulation activation respectively.

Control Plane: No $1B+ private pure-play company identified. Largest structural white space in the CDS. Adjacent players (Collibra at $5.25B, Alation at $1.7B) are extending from data governance but have not yet demonstrated AI-specific traction at scale.

Service Layer (Adjacent): Turing ($300M+ ARR, profitable). Structural enabler across layers, not a CDS layer itself.

Aggregate: approximately $250–$300B in private-market enterprise value across the profiled companies. The range reflects valuation uncertainty in reported but unconfirmed rounds (Surge AI, Lambda). A detailed company-level summary is provided in Appendix A.

3.3 Addressable Markets (Directional)

Precise TAM estimates for nascent categories carry significant uncertainty; the following are directional references drawn from third-party sources. The AI infrastructure and platform market (spanning Physics and portions of the Activation Layer) is projected at $150–$200B by 2028 (IDC, Gartner). The data labeling and RLHF market (core Genesis Layer) is estimated at $15–$25B by 2028 (Grand View Research, Cognilytica), though this substantially underestimates the vertical digitization opportunity, which overlaps with the $500B+ professional services markets being restructured (legal, healthcare, financial advisory). AI governance, risk, and compliance (the Control Plane) is projected at $5–$10B by 2028 (MarketsandMarkets), growing at 30%+ CAGR as regulatory obligations crystallize. These figures suggest the CDS collectively addresses a

multi-hundred-billion-dollar infrastructure opportunity, of which the $250–$300B in current enterprise value represents early-stage concentration rather than market saturation.

3.4 Why the Framework Persists

The CDS is a structural framework, not a market forecast. Its durability rests on the same logic that sustained the Modern Data Stack as an organizing taxonomy for a decade: it maps to engineering constraints that are invariant to which model, vendor, or architecture prevails. As long as AI systems require data to be moved to compute at speed (Physics), ground truth to be manufactured from human expertise and privileged data (Genesis), predictions to be translated into controlled actions inside operational systems (Activation), and governance to make all of the above safe and auditable (Control Plane), these four layers describe the industrial reality. Companies within each layer will rise, consolidate, and in some cases be displaced. The layers themselves will not.

3.5 The Evidence: Capital, IP, and Talent Are Flowing Into the CDS

Sections 4 through 7 present the empirical evidence for the CDS thesis. We profile 19 private companies, representing approximately $250–$300 billion in aggregate enterprise value, that have reached $1B+ valuations by solving constraints within specific CDS layers. These are not predictions; they are observable concentrations of private-market capital, engineering talent, and proprietary intellectual property across the exact infrastructure categories the framework describes. Where the CDS identifies white space (most notably the Control Plane), we find no $1B+ pure-play company, precisely confirming the framework's diagnostic value. The exemplars validate the framework; the framework explains the exemplars.

4. Layer I. The Physics Layer: Overcoming Data Gravity

The Infrastructure of Intelligence.

Frontier training and real-time inference increasingly bottleneck on I/O throughput and data locality. When models are deployed as products, latency becomes a revenue and safety variable. The Physics Layer is the infrastructure that prevents compute from stalling behind storage and networking.

4.1 Architectural Primitives

- Disaggregated shared-everything architectures (high-throughput access to unstructured data).
- Lakehouse operating systems (unify data + AI workloads on open formats; reduce duplication and movement).
- Compute-to-data patterns (bring training/inference to the data boundary for sovereignty and speed).
- Inference serving infrastructure (optimize model deployment for latency and cost at production scale).

4.2 Valuation Drivers

- **Category standardization:** once a storage/AI OS becomes default, switching costs compound across teams and workloads.
- **Embedded distribution:** platform primitives become “invisible” but indispensable inside AI clouds and enterprises.
- **Gross margin durability:** performance moats tend to protect pricing longer than feature moats.

4.3 Exemplars (Illustrative Case Studies)

The following companies are selected to validate the Physics Layer’s structural logic across its key sub-markets (storage architecture, data operating system, compute cloud, performance acceleration, inference serving), not as an exhaustive market map. Public-market incumbents (e.g., Snowflake, CoreWeave post-IPO) are referenced for context but not profiled.

VAST Data — The Storage Architecture

Valuation: ~$30.0B (Late-Stage/Secondary) | Private

Company Overview. VAST Data, founded in 2016, re-architected the physics of storage from first principles. Their Disaggregated Shared Everything (DASE) architecture eliminates traditional storage tiering, allowing massive GPU clusters to access exabytes of unstructured data with the latency of local memory.

CDS Positioning. VAST is the functional monopoly on high-performance I/O for the AI Cloud. As compute clusters scale to 100,000+ GPUs, legacy storage becomes a catastrophic bottleneck. VAST acts as the central nervous system for AI infrastructure, providing the storage substrate upon which CoreWeave, Lambda, xAI, and others build their compute clouds.

Financial Profile. Estimated $200M ARR by early 2025, projecting $600M by 2026. Signed a $1.17B commercial agreement with CoreWeave (November 2025). Three customers represent over $100M in commitments each. In talks to raise at ~$30B from CapitalG and NVIDIA. Total funding: ~$380M. Reports free cash flow positive status.

Structural Moat. Architectural moat: DASE collapses the entire storage hierarchy into a single tier, eliminating data movement latency. Competitors (DDN, Pure Storage, NetApp) still rely on tiered approaches that fundamentally cannot match VAST’s throughput for random-access AI workloads.

Source: *Reuters (Aug & Nov 2025); VAST press release (Nov 2025); Sacra; Blocks & Files.*

WEKA — The Performance Accelerator

Valuation: ~$1.6B (Series E, May 2024) | Private

Company Overview. WEKA (WekaIO), founded in 2013, provides an AI-native data platform with a software-only architecture that accelerates GPU utilization. Their NeuralMesh architecture (2025) is purpose-built for NVIDIA Blackwell GPUs.

CDS Positioning. If VAST is the storage substrate, WEKA is the performance layer. Software-only deployments on commodity hardware deliver extreme throughput for on-premise AI clusters in finance and media. Over 300 of the world's largest GPU deployments run on WEKA, including 12 of the Fortune 50.

Financial Profile. Surpassed $100M ARR in 2024 with eight-figure enterprise deals. Series E ($140M, led by Valor Equity Partners) raised entirely with existing investors. Total funding: ~$415M. Investors: NVIDIA, Generation Investment Management, Qualcomm Ventures, Samsung, Cisco.

Structural Moat. Hardware-agnostic, software-only approach. WEKA can optimize existing GPU clusters without replacing the storage layer. Gartner Visionary for three consecutive years. At $1.6B versus VAST's $30B, represents a potential value opportunity.

Source: *WEKA press release (May 2024); Blocks & Files; Gartner; Tracxn.*

Databricks — The Operating System

Valuation: $134.0B (Series L) | Private

Company Overview. Databricks unified storage and compute into the Data Lakehouse paradigm. Through MosaicML integration, they pivoted from "storing data" to "training on data," becoming the de facto operating system for enterprise AI.

CDS Positioning. Enterprises prefer training and deploying AI where governed data already lives. By offering training-as-a-service directly on customer data (without data movement), Databricks solves the critical data sovereignty challenge for the Fortune 500.

Financial Profile. Crossed $5.4B annual revenue run-rate. Latest financing: $5B equity + $2B debt at $134B valuation. The most valuable private data company in the world.

Structural Moat. Data Gravity lock-in: once 50PB of training data sits in the Lakehouse, egress cost is prohibitive. Databricks owns the gravitational center of enterprise AI.

Source: *Databricks press release (Feb 2026); Reuters (Feb 2026).*

Lambda — The Developer Cloud

Valuation: ~$2.5–$15B (Series D/E) | Private, Pre-IPO

Company Overview. Lambda (formerly Lambda Labs), founded in 2012, is a developer-first GPU cloud provider. If CoreWeave is the industrial grid, Lambda is the developer's workbench: instant access to high-performance GPUs with a data-centric software stack.

CDS Positioning. Lambda solves the compute access bottleneck. Competitive pricing (H100 at ~$2.49/hr vs. CoreWeave's ~$4.25) and developer-centric tooling ("1-Click Clusters") make it the accessible on-ramp for AI at scale.

Financial Profile. Estimated $500M revenue run-rate by mid-2025. Series D: $480M (Feb 2025, $2.5B valuation). Series E: $1.5B (Nov 2025) anchored by multi-billion-dollar Microsoft partnership. Hired Morgan Stanley, JPMorgan, Citi for potential H1 2026 IPO.

Structural Moat. Developer loyalty + competitive pricing + Microsoft partnership. NVIDIA is both investor and hardware partner, ensuring preferential GPU access. Gross margins ~50–61%.

Source: *Sacra; TechCrunch; Bloomberg; The Information; StockAnalysis.com.*

Fireworks AI — The Inference Engine

Valuation: $4.0B (Series C, October 2025) | Private

Company Overview. Founded in 2022 by the PyTorch team (Meta), Fireworks provides a generative AI platform-as-a-service for building, tuning, and deploying AI using open-source models. Processes 10+ trillion tokens daily for 10,000+ customers.

CDS Positioning. Fireworks occupies inference serving infrastructure — how enterprises serve AI models at production scale. Provides access to hundreds of open-source models across text, image, audio, and multimodal domains. Customers: Uber, Shopify, Genspark.

Financial Profile. Estimated $130M ARR by mid-2025 (20x YoY growth). Series C: $254M (led by Lightspeed, Index, Evantic; Sequoia participation) at $4B. Total funding: $331M.

Structural Moat. PyTorch DNA = deep model optimization expertise. Application-tailored tuning democratizes production-grade model customization that frontier labs guard closely.

Source: *BusinessWire (Oct 2025); Sacra; Crunchbase; TechCrunch.*

5. Layer II. The Genesis Layer: The Manufacturing of Truth

From Data Mining to Data Manufacturing.

As high-quality public data saturates and model capacity rises, performance becomes increasingly gated by the availability of high-fidelity ground truth and privileged vertical corpora.

5.1 Genesis Sub-Markets

- **Reasoning Foundries:** RLHF, expert feedback, evaluation sets, red-teaming, and preference data that align models with human intent.
- **Vertical Digitization:** turning high-value, legally/physically inaccessible workflows into proprietary datasets (law, medicine, code, enterprise knowledge).

5.2 Valuation Drivers

- **Ground-truth supply chains scale operationally, not just technically;** the moat is process + QA + workforce orchestration.
- **Verified corpora creates data network effects**: more usage → more corrections → higher accuracy → more usage.
- **Regulated verticals reward trust and compliance as product features** (audit trails, provenance, governance).

5.3 Exemplars: Horizontal Foundries

Exemplars are selected to validate the Genesis Layer's sub-market structure: horizontal foundries (model-agnostic RLHF and data labeling) and vertical digitizers (industry-specific proprietary data assets). Numerous additional private companies operate in both segments.

Scale AI — The Reasoning Foundry

Valuation: ~$29.0B (Series G) | Private

Company Overview. Scale is a "Reasoning Foundry" employing subject matter experts (PhDs, lawyers, coders) to create the ground truth that aligns models with human intent. Functions as the universal supply chain for intelligence: regardless of which foundation model wins, Scale gets paid to train it.

CDS Positioning. Scale functions as the "Federal Reserve" of data. High-fidelity feedback and evaluation remain necessary across competing model ecosystems. Meta's 49% stake ($14.3B, June 2025) validates the strategic value but creates customer concentration risk.

Financial Profile. Estimated $1.5B+ ARR by 2025. Meta investment valued the company at $29B. Expanded into government contracts, autonomous vehicle data, and evaluation tools (Scale Seal, Scale Leaderboard).

Structural Moat. Network effects (more customers attract more annotators) + switching costs (RLHF pipelines deeply integrated). Strategic risk: Meta ownership concentration has triggered customer diversification by Google, OpenAI, xAI.

Risk Note: Meta's 49% stake creates potential customer conflicts. Synthetic data improvements could erode the human labeling moat, though frontier models still require human ground truth.

Source: *Reuters (Jun 2025); Gun.io (Dec 2025); Sacra.*

Surge AI — The Bootstrapped Juggernaut

Valuation: $15–$25B (reported fundraise) | Private, Bootstrapped

Company Overview. Founded in 2020 by Edwin Chen (ex-Google/Meta), Surge connects AI developers with ~1 million expert annotators. Reached $1B+ revenue without raising external capital, one of the most capital-efficient technology companies in history.

CDS Positioning. Direct structural parallel to Scale AI. Customers: OpenAI, Google, Microsoft, Meta, Anthropic, U.S. Air Force. Where Scale's moat is breadth, Surge's is quality and capital efficiency: premium annotator pay produces consistently higher-quality outputs.

Financial Profile. Exceeded $1.2–1.4B revenue by mid-2025 with ~121 FTEs (~$11.5M revenue per employee). Bloomberg (July 2025): in talks to raise ~$1B at $25B. Bloomberg (July 2025) reported the company was in talks with a16z, Warburg Pincus, and others to raise ~$1B at $25B.

Structural Moat. Bootstrapped profitability equals no stakeholder conflicts, no customer concentration from strategic investors, extraordinary operating leverage. Asset-light model scales as costs are variable and directly linked to task volume.

Risk Note: Class-action lawsuit (CA, May 2025) alleging worker misclassification of annotators. Multiple subsidiary platforms with unclear ownership relationships.

Source: *Bloomberg (Jul 2025); Reuters; Sacra; Wikipedia; Equidam.*

Snorkel AI — The In-House Data Factory

Valuation: $1.3B (Series D, May 2025) | Private

Company Overview. Born from Stanford AI Lab, Snorkel pioneered programmatic labeling: users write "labeling functions" that automatically annotate large datasets. Evolved into a full enterprise AI data development platform (Snorkel Flow / Snorkel Foundry).

CDS Positioning. Automated data curation for enterprises that cannot send data externally. While Scale and Surge are outsourced foundries, Snorkel enables in-house data manufacturing. Structurally critical for regulated industries: banking, insurance, healthcare, government.

Financial Profile. Series D: $100M (May 2025) at $1.3B. Investors: In-Q-Tel, BlackRock, Accenture. Five of top ten U.S. banks use the platform. Total funding: ~$285M.

Structural Moat. Data sovereignty guarantee: only enterprise-grade alternative keeping data on-premises. In-Q-Tel investment validates intelligence community importance. Limitation: programmatic labeling works best for classification; RLHF still requires humans.

Source: *Gun.io (Dec 2025); VentureBeat; CB Insights; Forbes AI 50.*

5.4 Exemplars: Vertical Digitization

Generic models are reaching asymptotic convergence. The next alpha is in Vertical Data: proprietary datasets legally or physically inaccessible to public crawlers.

Anysphere (Cursor) — Code Digitization

Valuation: $29.3B (Series D, November 2025) | Private

Company Overview. Founded 2022 by MIT classmates. Cursor is an AI-native code editor (VS Code fork) with multi-file reasoning, repository-wide search, and agentic workflows. 1M+ daily active users, 50,000 businesses.

CDS Positioning. Digitizing the creation of software. Every interaction generates training data on how engineers think, debug, and build, a proprietary data loop training next-generation coding models.

Financial Profile. Crossed $100M ARR in January 2025 (fastest SaaS ever). Surpassed $1B annualized revenue by November 2025, without marketing spend. Series D: $2.3B (Accel/Coatue, NVIDIA, Google) at $29.3B. Prior Series C: $900M at $9.9B (June 2025).

Structural Moat. Reasoning Feedback Loop: every interaction widens the gap. Graphite acquisition (code review, Dec 2025) signals end-to-end development lifecycle.

Risk Note: Pricing backlash (July 2025). Microsoft/GitHub Copilot, now operating as a multi-model platform (integrating Claude, OpenAI, and proprietary models), retains deeper enterprise distribution. Model providers are increasingly building native coding and agentic infrastructure, compressing the addressable space for independent developer tools..

Source: *Wikipedia; Contrary Research; Bloomberg; TechCrunch.*

Glean — Knowledge Digitization

Valuation: ~$7.2B (Series F) | Private

Company Overview. The “Work AI” platform. Connects to every enterprise SaaS app (Slack, Jira, Salesforce, Confluence) to build a unified Enterprise Knowledge Graph. 100M+ agent operations processed.

CDS Positioning. The Knowledge Graph is the “ground truth” for enterprise AI agents. Without a unified index of organizational knowledge, agentic AI cannot operate reliably inside enterprises.

Financial Profile. Raised $150M (June 2025) at $7.2B valuation.

Structural Moat. Breadth and depth of enterprise knowledge index. Each deployment creates a proprietary graph that becomes more valuable with every connected application and query. Competing requires integration with hundreds of tools.

Source: *36Kr/AiCoin (Apr 2025); Crunchbase; Forbes AI 50.*

Harvey — Legal Digitization

Valuation: ~$8–$11B (Series F) | Private

Company Overview. Productized "billable hours" as software by securing exclusive access to case law and contracts from top firms (PwC, Allen & Overy). Owns the "Reasoning Engine" for law.

CDS Positioning. Vertical Data > Generic Models. Generic models cannot replicate this performance because they lack access to privileged, private legal data assets.

Financial Profile. Valuation: $8B announced (Dec 2025); reporting on potential $11B round (Feb 2026).

Structural Moat. Legal data is uniquely defensible: privileged, confidential, not publicly crawlable. Each firm partnership deepens Harvey's corpus in ways no generic model can replicate.

Source: *Harvey blog (Dec 2025); TechCrunch (Dec 2025, Feb 2026); Reuters (May 2025).*

Abridge — Health Digitization (Clinical Documentation)

Valuation: $5.3B (Series E) | Private

Company Overview. Captures the "Dark Data" of medicine: ambient doctor-patient audio structured into clinical notes. Building the largest medical reasoning dataset in history.

CDS Positioning. Ambient clinical conversations are high-value, high-frequency data with strong governance requirements. The proprietary dataset constitutes a deeper moat than any open-source model.

Financial Profile. $5.3B valuation in a $300M raise (June 2025).

Structural Moat. Reasoning Feedback Loop: every note makes Abridge smarter at clinical context, terminology, and physician reasoning. Governance requirements (HIPAA, institutional trust) further protect the position.

Source: *Reuters (Jun 2025); Forbes AI 50; Digital Journal.*

Hippocratic AI — Health Digitization (Clinical Engagement)

Valuation: $3.5B (Series C, November 2025) | Private

Company Overview. Founded 2022. Safety-focused generative AI agents for healthcare: chronic care, post-discharge, pre-op, clinical trials. Explicitly non-diagnostic, non-prescribing.

CDS Positioning. Complementary to Abridge: where Abridge digitizes the doctor-patient conversation, Hippocratic digitizes the patient engagement workflow. Together they represent both sides of healthcare digitization.

Financial Profile. Series C: $126M (Avenir Growth) at $3.5B. Total: $404M. 50+ health systems/payors/pharma across 6 countries. 1,000+ use cases. 115M+ patient interactions, zero safety issues. Investors: a16z, General Catalyst, Kleiner Perkins, NVIDIA, Google CapitalG.

Structural Moat. Polaris Safety Constellation Architecture (multi-model cross-checking). Safety focus earns trust from risk-averse health systems (Cleveland Clinic, Northwestern, Guy's & St Thomas' NHS). 115M+ interactions = compounding proprietary dataset.

Source: *BusinessWire (Nov 2025); Fierce Healthcare; Axios Pro.*

6. Layer III. The Activation Layer: The Nervous System

Closing the Loop from Insight to Action.

Analytics created insight. Activation creates outcomes. As AI becomes agentic, the winning systems are those that can safely translate predictions into actions within systems of record, while enforcing permissions, approvals, and auditability.

The complexity of activation is compounding. Enterprise deployments are moving from single-agent copilots to multi-agent systems in which specialized agents coordinate across functions: one agent retrieves, another reasons, a third executes, and a fourth audits. In industry surveys, 46% of enterprise teams now cite integration with existing systems, not model capability, as their primary constraint on agent deployment. This shift reframes the Activation Layer: the scarce resource is not intelligence but the governed connective tissue that lets intelligence act safely across heterogeneous environments.

6.1 Activation Primitives

- **Reverse ETL and decision activation:** push computed truth back into operational tools (CRM, ads, support, ERP).
- **Integration OS:** unify messy, multi-source operational environments (public sector, industrial, field ops).
- **Agent rails:** policies, approvals, and rollback paths for autonomous actions.
- **Kinetic activation:** AI that executes in the physical world (defense, autonomous systems, robotics).

6.2 Valuation Drivers

- **Measurable ROI:** activation is where savings/revenue lift become visible and budget-justifiable.
- **Deep workflow embedding:** once agents touch the system of record, switching costs become operational, not just technical.
- **Risk-managed autonomy:** the platform that can prove safety and traceability wins regulated procurement.

6.3 Exemplars: Kinetic Activation

Exemplars are selected across two Activation sub-markets: kinetic (AI acting in the physical world) and digital (AI acting in enterprise systems of record). The distinction matters because moat mechanics, buyer profiles, and regulatory requirements differ substantially between them.

Anduril Industries — Defense Activation

Valuation: ~$30.5B (Series G, June 2025) | Private

Company Overview. Founded 2017 by Palmer Luckey (Oculus VR). Lattice OS fuses sensor data (radar, thermal, visual) to autonomously task drones, interceptors, and unmanned systems.

CDS Positioning. The Activation Layer for national defense. Proves that AI value is not in image recognition but in kinetic response. Lattice collapses the entire CDS into a single system: Physics (sensors), Genesis (real-time fusion), Activation (autonomous response).

Financial Profile. Raised $2.5B (June 2025) at $30.5B. Benefits from $1T+ U.S. defense budget and bipartisan support for software-first procurement.

Structural Moat. Integration Barrier applied to defense: connecting to classified sensor networks and weapons systems creates a barrier measured in years and billions. Government procurement cycles (12–24 months) further protect incumbents.

Source: *Crunchbase; Defense News; TFN (Jan 2026).*

Applied Intuition — Simulation & Validation

Valuation: ~$15.0B (Series F) | Private

Company Overview. Simulation and validation layer for autonomous systems. Enables automotive and defense companies to “activate” AI in the real world by first proving it works in physics-based simulation.

CDS Positioning. Safety layer between AI training and physical deployment. Before an autonomous vehicle or defense drone can act, its behavior must be validated across millions of scenarios.

Financial Profile. Reached $15B valuation in Series F. Serves major automotive OEMs and defense contractors.

Structural Moat. Scenario library deepens with every deployment. Each edge case, each validated behavior adds to a proprietary corpus of “how AI should behave in the physical world.”

Source: *Crunchbase; Forbes AI 50; PitchBook.*

6.4 Exemplars: Digital Activation

Celonis — Process Intelligence

Valuation: ~$13.0B+ | Private

Company Overview. Owns Process Intelligence. Execution Management System (EMS) automatically triggers actions to fix supply chain bottlenecks, billing errors, and operational inefficiencies.

CDS Positioning. The “Cortex” translating data into enterprise efficiency. Bridges the Genesis Layer (where insights are generated) and actual operational improvement. Natural orchestration layer for enterprise AI agents.

Financial Profile. Valued at $13B+. Significant Fortune 500 penetration. Based in Munich, the most valuable CDS company outside the United States.

Structural Moat. Process mining data creates an organizational "digital twin." Each deployment maps actual execution of business processes, creating the foundation for AI-driven optimization.

Source: *Crunchbase; PitchBook; Forbes.*

Peregrine Technologies — Radical Integration

Valuation: $2.5B (Series C, March 2025) | Private

Company Overview. Operating system for public safety. Fuses disparate sensors (CCTV, dispatch, police records, utility data) into a single operational picture.

CDS Positioning. Activation in complex environments is won through messy integration and trusted deployment. Connecting 5,000 legacy record systems is a brute-force problem that, once solved, creates a high-entropy barrier to entry.

Financial Profile. $2.5B valuation in Series C (March 2025).

Structural Moat. Integration Barrier: each connected system, each normalized format, each automated workflow adds to a proprietary integration layer competitors need years to replicate.

Source: *Peregrine press release (Mar 2025); Built In SF (Mar 2025).*

Hightouch — The Activation Rails

Valuation: $1.2B (Series C, February 2025) | Private

Company Overview. The "API Rails" for AI Agents. If an AI decides to "refund a customer," Hightouch executes that transaction in the ERP. Safety layer between the probabilistic brain and the deterministic database.

CDS Positioning. Decision activation creates measurable ROI by embedding outputs back into systems of record. As AI Agents take autonomous actions, rails for safe write-back become mission-critical.

Financial Profile. Series C: $80M at $1.2B. Investors: Sapphire Ventures, Bain Capital Ventures. Total funding: $171M.

Structural Moat. Position as "last mile" of the AI stack. Integrations with hundreds of downstream tools create network effect: more destinations = more value.

Source: *Hightouch blog (Feb 2025); PRNewswire (Feb 2025).*

LangChain — The Agent Engineering Framework

Valuation: $1.25B (Series B, October 2025) | Private

Company Overview. Founded 2022 by Harrison Chase. Open-source framework evolved into full platform: LangChain (agent builder), LangGraph (orchestration/memory), LangSmith (testing/observability). 118K GitHub stars.

CDS Positioning. Connective tissue of the Activation Layer. Strategic investors Databricks, Datadog, ServiceNow, Workday, and Cisco participated in Series B, signaling broad ecosystem support.

Financial Profile. Series B: $125M (IVP-led) at $1.25B. New investors: CapitalG, Sapphire Ventures. Enterprise adoption: Cisco, Replit, Cloudflare, Workday, ServiceNow. Total funding: ~$160M.

Structural Moat. Open-source ecosystem moat (118K stars). Monetization follows MongoDB/Elastic/Databricks playbook: free tools build community, LangSmith monetizes enterprise. Risk: OpenAI/Anthropic/Google building native agent infrastructure.

Source: *TechCrunch (Oct 2025); Fortune (Oct 2025); GitHub.*

6.5 The Agent Protocol Layer: MCP and A2A

As agents proliferate, ad hoc integrations do not scale. The industry has converged on two complementary open protocols that now form the standard wiring of the Activation Layer.

Model Context Protocol (MCP), introduced by Anthropic in November 2024, standardizes the vertical connection between an agent and external tools, databases, and services. It solves the context problem: rather than building custom connectors for every data source, MCP provides a universal interface through which agents discover and invoke capabilities. By February 2026, MCP had crossed 97 million monthly SDK downloads (Python and TypeScript combined) and had been adopted by every major AI provider. Agent-to-Agent Protocol (A2A), introduced by Google in April 2025, standardizes the horizontal connection between peer agents. It enables agents built on different frameworks, by different vendors, to discover each other's capabilities, delegate tasks, and coordinate execution across enterprise systems. Over 50 technology partners (Atlassian, Salesforce, SAP, ServiceNow, among others) support A2A. In December 2025, the Linux Foundation launched the Agentic AI Foundation (AAIF), co-founded by OpenAI, Anthropic, Google, Microsoft, AWS, and Block, as the permanent governance home for both MCP and A2A.

This protocol convergence has three structural implications for the CDS thesis. First, it validates the Activation Layer's core premise: the scarce resource is not intelligence but governed connectivity between intelligence and systems of record. MCP is, in effect, the standardization of what Hightouch, Peregrine, and Celonis solve through proprietary integration; its existence confirms the problem is structural, not niche. Second, protocol standardization lowers switching costs at the connectivity layer while raising switching costs at the workflow layer. An enterprise can swap the protocol client; it cannot easily replicate 5,000 integrated police records (Peregrine) or a process-mined digital twin of its supply chain (Celonis). Companies whose moats rest on depth of integration rather than

proprietary connectors are structurally advantaged. Third, the A2A requirement for agent identity, authentication, and capability discovery maps directly onto the Control Plane primitives described in Section 7. As multi-agent coordination scales, the governance infrastructure (who authorized this agent, what is its scope, what audit trail exists) becomes inseparable from the activation infrastructure. The Activation Layer and Control Plane converge at the protocol level.

Source: *Anthropic (Nov 2024); Google Cloud (Apr 2025); Linux Foundation AAIF announcement (Dec 2025); DEV Community MCP/A2A analysis (Mar 2026); Arcade State of AI Agents Report (Dec 2025).*

7. The Cross-Cutting Control Plane: Trust, Governance, and Evaluation

As systems move from assistance to autonomy, the enterprise's limiting reagent becomes **trust**. The Control Plane is the set of primitives that make AI deployable: identity, access, provenance, monitoring, evaluation, and incident response.

7.1 What a Production-Grade Control Plane Must Provide

1. **Identity & Authorization:** least-privilege access for humans and agents; separation of duties.
2. **Observability:** tracing, data lineage, cost/latency monitoring, and drift detection across pipelines and agents.
3. **Evaluation:** continuous tests for accuracy, safety, bias, and policy compliance; regression prevention.
4. **Auditability:** immutable logs, explainability artifacts, and evidence trails for regulators and customers.

Control Plane winners tend to emerge where trust requirements are highest (regulated verticals, public sector, finance, healthcare), and where failures are costly enough to demand formal governance rather than "best effort."

7.2 The Three-Way Race for Control Plane Ownership

Unlike every other CDS layer, the Control Plane has not yet produced a dominant private-market platform above $1B. This absence is not due to lack of demand, it reflects a structural contest among three categories of competitors, each with distinct advantages and limitations.

Contestant 1: Hyperscalers (AWS, Azure, GCP) — Bundling the Basics

Cloud providers are aggressively embedding governance into their AI platforms. AWS Bedrock Guardrails offers content filtering, PII redaction, hallucination detection (via Automated Reasoning checks), and prompt injection defense—at no additional charge beyond model usage. At re:Invent 2025, AWS launched AgentCore, explicitly positioned as "the governance layer that makes autonomous AI acceptable to enterprises." Azure AI Content Safety and Google Vertex AI Model Armor provide parallel capabilities. The structural advantage is clear: hyperscalers already own identity (IAM/Entra ID), logging (CloudWatch/Azure Monitor), and compliance certifications (SOC 2, FedRAMP, HIPAA). Adding governance to an existing cloud bill is frictionless for procurement. The risk for independent players is that basic guardrails become table stakes: bundled free, "good enough," and impossible to displace on price.

Contestant 2: Established Platform Companies — Lateral Extension

Established companies with enterprise distribution are extending into AI governance from adjacent positions. Among public companies (acknowledged but outside the private-market scope of this paper): Datadog (NASDAQ: DDOG, ~$40B market cap) launched LLM Observability in 2024 and AI Agent Monitoring in June 2025; CrowdStrike and Palo Alto Networks are absorbing AI security into the cybersecurity stack; DataRobot has built observability and guardrails directly into its ML platform.

These companies have distribution, enterprise trust, and revenue at scale, advantages no private startup can match today.

More relevant to private-market investors: two established private data governance companies are extending into AI governance. Collibra ($5.25B valuation, Series G, 2021; ~$596M total funding; 1,000+ employees) is a Gartner Magic Quadrant Leader for Data and Analytics Governance and recently acquired Deasy Labs (unstructured data governance) and Raito (data access management) to extend its platform toward AI model governance. Alation ($1.7B valuation, Series E, 2022; ~$109M revenue in 2024; Forrester Wave Leader in Data Governance, Q3 2025) acquired Numbers Station AI (May 2025) to build agentic workflows and launched an Agent Builder for enterprise-grade AI agents on structured data. Both companies govern data rather than models, but as data governance and AI governance converge, their enterprise footprint (Collibra: 500+ enterprises including 70% of the largest U.S. banks; Alation: 600+ enterprises including 40% of the Fortune 100) gives them distribution that purpose-built AI governance startups lack. The risk: both carry stale valuations from 2021–2022 and have yet to demonstrate that traditional data catalog customers will pay incremental budget for AI-specific governance.

Contestant 3: Private Startups — AI-Native but Pre-Scale

A cohort of AI-native startups is purpose-built for the Control Plane but none have crossed the $1B valuation threshold. Arize AI ($131M raised, Series C, February 2025) is the most prominent: an AI observability and evaluation platform backed by Adams Street, M12 (Microsoft's venture fund), and Datadog itself. Credo AI focuses on AI governance and compliance automation, named to CB Insights' AI 100. Prompt Security specializes in LLM red-teaming and prompt injection defense. These companies are technically strong and growing, but they face a "pincer" from hyperscalers bundling below and public incumbents extending above. The entire private Control Plane category remains pre-product-market-fit at scale.

7.3 The Regulatory Forcing Function: The Only Structural Moat?

If the Control Plane were purely a technical problem, hyperscalers would likely win by default, they bundle at zero marginal cost and own the procurement relationship. But regulatory dynamics may prevent this consolidation. The EU AI Act (obligations phasing in from August 2025 through 2027) requires auditable, traceable evidence of model behavior. Regulators in finance, healthcare, and public safety are unlikely to accept self-assessment from the same provider hosting the model and serving the inference. This creates a structural demand for **independent, cross-cloud evaluation and audit infrastructure**, the one capability that neither cloud providers nor security incumbents can credibly deliver for their own models. The AI governance market is projected to grow from $227M in 2024 to $4.83B by 2034. Whether that value accrues to independents or gets absorbed by incumbents will depend largely on whether regulatory frameworks mandate separation between model hosting and model evaluation.

7.4 Structural Assessment

The Control Plane is the least developed but potentially most decisive segment of the CDS. The likely outcome is a layered market: hyperscalers own the commodity layer (basic content filtering, PII redaction, prompt safety), public platform companies own the operational layer (observability, drift detection, cost monitoring), and independent specialists own the compliance layer (regulation-grade audit, cross-cloud evaluation, red-teaming certification). The investable thesis for private capital is

narrow but high-conviction: the winners will be companies that establish themselves as the trusted, vendor-neutral “auditor” of AI systems, an analogy closer to Moody’s or Deloitte than to Datadog. Without the regulatory forcing function, however, the white space closes and the Control Plane gets absorbed into existing infrastructure. This remains the highest-risk, highest-optionality segment of the Cognitive Data Stack.

8. Adjacent Frontier: The Service Layer

Infrastructure requires builders. The Service Layer supplies the specialized human capital and AI-native tools that turn CDS architecture into production systems.

The CDS cannot be built by software alone. Each layer demands a distinct talent profile: the Physics Layer requires systems engineers and storage architects; the Genesis Layer depends on domain experts (clinicians, lawyers, linguists) who manufacture ground truth; the Activation Layer needs integration specialists who understand both the AI stack and the legacy systems it must touch; the Control Plane demands governance, compliance, and security professionals. In aggregate, talent availability is the single most common constraint enterprises cite when scaling AI infrastructure.

Critically, the talent layer itself is bifurcating. Human expertise remains irreplaceable for tasks requiring judgment, domain knowledge, and regulatory accountability: RLHF annotation, clinical validation, legal review, and safety auditing. But AI-native building tools are increasingly capable of performing implementation tasks that previously required specialized engineers. Frontier coding agents (GitHub Copilot, Cursor, Claude Code, Gemini Code Assist), orchestration frameworks (LangChain, LangGraph, CrewAI), and agentic development platforms are compressing the time and headcount required to build, deploy, and maintain CDS components. The practical implication: enterprises building the CDS will increasingly staff with a combination of human domain experts and AI development tools, and the talent platforms that can orchestrate both will capture disproportionate value.

The Service Layer is structurally adjacent to the CDS rather than embedded within it: it enables all layers without being a layer itself. We profile one exemplar below; the broader landscape includes enterprise consultancies (Accenture, Deloitte, McKinsey's QuantumBlack), specialized AI implementation firms, and the growing ecosystem of AI-powered development tools referenced above.

Turing — The Talent Cloud

Disclosure: Turing (the company profiled below) is an independent, unaffiliated entity. It has no ownership, commercial, advisory, or other relationship with AVC Turing Ltd., the publisher of this report. The shared name is coincidental. AVC Turing Ltd. has no affiliation with any company referenced in this document (see Disclaimers).

Status: Profitable ($300M+ ARR)

Company Overview. Turing (turing.com), founded in 2018, operates as the talent engine for the AI era. Captures labor spend for data labeling, RLHF, and AI implementation, especially when internal hiring cannot keep pace.

CDS Positioning. As enterprises struggle to find talent to build out the CDS, Turing captures the labor spend of the AI budget. Critical enabler of the Genesis Layer.

Financial Profile. Revenue run-rate tripled to $300M in a profitable year (January 2025 reporting).

Structural Moat. Vetted talent network + matching algorithms connecting enterprise needs with specialized AI skills. As CDS expands, the talent bottleneck intensifies.

Source: *Reuters (Jan 2025); Business Wire (Jan 2025).*

9. Strategic Moats: Why This Architecture Wins

The durability of the CDS is rooted in three structural dynamics that protect these assets from commoditization.

1. The “Data Gravity” Lock-In

In the Physics Layer (VAST, Databricks, WEKA), value is sticky because data is heavy. Once an enterprise moves 50PB of training data into a specific architecture, the cost of egress is prohibitive. These companies do not just rent storage; they own the *state* of the enterprise. Migration cost alone can exceed annual contract value, creating a retention moat that strengthens with every additional petabyte.

2. The “Reasoning” Feedback Loop

In the Genesis Layer (Scale, Surge, Anysphere, Abridge, Hippocratic), the product improves with use. Every line of code written in Cursor and every medical note generated by Abridge makes their proprietary model smarter. This creates a widening gap that a generic competitor cannot bridge simply by buying GPUs: they need the *human usage data*. Vertical data assets are not merely defensible, they are compounding.

3. The “Integration” Barrier

In the Activation Layer (Peregrine, Anduril, Celonis, Hightouch), the moat is complexity. Connecting to 5,000 legacy police record systems or integrating with kinetic defense hardware is a brute-force problem. Once solved, it creates a high-entropy barrier to entry. Each new integration makes the platform more valuable to the next customer, an integration network effect.

10. Boundary Conditions and Design Constraints: Underwriting the CDS Thesis

A research-grade framework should not only explain why it works, it should specify the boundary conditions under which it holds. This section is an underwriting lens: it clarifies key frictions, failure modes, and signals that would force a CDS update.

10.1 Market Frictions and Failure Modes

- **Distribution re-centralizes at the model layer:** if one provider owns interface, identity, payments, and enterprise procurement, pricing power can shift upward.
- **Platform bundling compresses stand-alone margins:** hyperscalers can absorb primitives across storage, governance, and activation, forcing vendors to win on step-function performance or embedded distribution. This risk is most acute in the Control Plane, where AWS Bedrock Guardrails, Azure AI Content Safety, and GCP Model Armor are already bundling governance at zero marginal cost (see Section 7.2).
- **Synthetic data overreach creates fragility:** without continuous grounding in reality, systems can drift or optimize for benchmarks instead of the world. Provenance and rights become binding constraints.
- **Sovereignty and privacy fragment deployments:** cross-border and sector restrictions raise integration cost and slow standardization unless the architecture brings compute to data.
- **Activation liability stalls write-back:** enterprises will not permit autonomous actions without approvals, audit trails, and safety controls, especially in regulated workflows.
- **Open source accelerates feature parity:** durable moats must be rooted in proprietary data assets, workflow entrenchment, distribution, and measurable performance advantages.

10.2 Why CDS Remains Structurally Advantaged Under Stress

- Regardless of model quality, enterprises still need governed data plumbing, security, access control, lineage, observability, and integration are compliance and operational requirements.
- As inference costs fall, the economic bottleneck shifts to data readiness and workflow activation. Commoditized intelligence *increases* the value of the substrate connecting intelligence to systems of record.
- Bundling pressure tends to create a smaller set of default substrates. Defensible companies either deliver step-function performance gains or become the control plane every stack must adopt.
- Synthetic data scales volume, but verification becomes the constraint. The winners manufacture truth through expert feedback, evaluation harnesses, and continuous measurement against reality.
- Write-back adoption is staged: copilots → assisted workflows → approvals → bounded autonomy. Platforms providing rails, audit logs, and reversible actions capture value across the entire adoption path.

10.3 Disconfirming Signals: What Would Shift Value Within the Framework

An important distinction: the CDS framework itself is structural, not speculative. As long as AI systems require data to be moved, ground truth to be manufactured, actions to be executed in operational systems, and governance to make all of this safe, these four layers exist. Like the Modern Data Stack before it, the CDS describes the industrial reality of how data infrastructure serves intelligence. The MDS did not “fail” when specific vendors were displaced; the framework persisted because the underlying engineering constraints persisted. The same logic applies here.

What can change is which companies occupy each layer, which moat mechanisms hold, and where marginal value concentrates. The following signals would force a reallocation of emphasis within the framework, not its abandonment:

- Enterprises achieve reliable autonomy at scale with minimal incremental governance infrastructure beyond standard IT controls: value within the Control Plane shifts from independent platforms toward bundled hyperscaler offerings, compressing stand-alone opportunity.
- Model providers or hyperscalers deliver sovereign, end-to-end CDS capabilities at materially lower total cost with credible interoperability: independent substrates face margin compression, and value migrates toward integrated platforms. The layers persist, but fewer stand-alone companies occupy them.
- Vertical digitizers cannot retain data rights, or regulators prohibit durable derivation of proprietary datasets: Genesis Layer moats weaken for specific companies, though the structural need to manufacture ground truth persists. Value shifts toward in-house data curation or regulatory-compliant architectures.
- Hardware architectures make data movement and I/O constraints largely irrelevant for frontier training and enterprise activation: Physics Layer differentiation compresses for current incumbents, though the layer itself remains; the constraint simply moves to a different engineering frontier (e.g., energy, cooling, chip architecture).

11. A CDS Lens for Capital Allocation

Educational; not investment advice.

CDS is actionable as a diligence lens because each layer has distinct buyer psychology, procurement patterns, and unit economics. The goal is to evaluate where leverage compounds.

11.1 The 12-Question Rubric (Non-Exhaustive)

5. What is the true constraint removed (I/O, governance, veracity, integration)?
6. Does the product move compute to data (or force data movement)?
7. What is the "switching cost surface" (technical + operational + compliance)?
8. Is the moat architectural (physics), procedural (ops/QA), or proprietary (data asset)?
9. Who is the buyer and who is the blocker (security, compliance, procurement)?
10. What is the proof artifact (audit logs, eval suites, ROI dashboards)?
11. How does it handle unstructured data at scale (audio, video, documents, sensors)?
12. What failure mode is contained (hallucination, drift, unsafe actions)?
13. What is the path from pilot to production (time-to-value, integration burden)?
14. Does the business compound with usage (data network effects, workflow entrenchment)?
15. What gets commoditized next (and what stays scarce)?
16. What would disintermediate this company (platform bundling, standards, regulation)?

12. Conclusion: The Infrastructure Imperative

The market often mistakes infrastructure for commodity. However, the most durable value accrues to the layers that control the flow of the critical resource. In the AI era, that resource is **Reasoning Data**.

The capital expenditure cycle from hyperscalers is currently funding the build-out of the Cognitive Data Stack. As these companies mature from “Promising Unicorns” to “Critical Global Infrastructure,” they are establishing the standards that will define enterprise computing for the next decade.

The expanded CDS framework now maps 19 companies across three layers plus a cross-cutting Control Plane, representing approximately $250–$300 billion in aggregate enterprise value. This concentration in private markets underscores both the opportunity and the access challenge.

The mandate for the industry is clear: The alpha lies in the “Picks and Shovels” of the Cognitive Age; the builders of the Physics, Genesis, and Activation layers.

The Cognitive Data Stack is not a thesis about specific companies. It is a structural description of how data infrastructure must be organized to serve intelligence at scale. The companies profiled in this paper are the evidence that the framework is already operative; they are not the framework itself. As in every technology cycle, individual players will rise, consolidate, and in some cases be displaced. The constraints they solve will persist. The four layers will continue to be filled by capital, talent, and intellectual property for as long as the world demands governed, reliable, actionable AI. The question is not whether the CDS exists. It is who will build it.

About the Author

Hernán Asorey is Co-Founder and Managing Partner of AVC Turing, a New York-based venture and research firm investing at the intersection of applied AI, critical infrastructure, and global markets. His practitioner credentials span three of the most consequential technology organizations of the past two decades. At Salesforce, he served as the company's first Chief Data Officer, architecting a modern data platform, launching the Einstein AI platform, and embedding real-time intelligence into the enterprise product loop; that work earned recognition from the World Economic Forum at Davos and a dedicated feature in Marc Benioff's book *Trailblazer*. At Microsoft, he partnered with OpenAI to design and deliver the first business-ready generative AI experiences, including enterprise AI copilots and LLM-powered applications across industries, well before the category had a name. At Google, he led data science and engineering globally across the company's most iconic products, shaping growth strategy and overseeing the experimentation and measurement infrastructure that governs decisions at scale.

Asorey serves as Independent Board Member of Contentsquare, Head of the Board of Advisors at Klaviyo (NYSE: KVYO), and holds board and advisory roles at Alation, Syncro, Faros AI, and Trebellar. In 2021, he was named to the HITEC 100 as one of the most influential Hispanic professionals in technology. A Stanford GSB alumnus, he began his career as a university lecturer in Buenos Aires, a teaching instinct that continues to shape how he approaches research: every framework is, at its core, a tool for helping practitioners think more clearly about systems that are changing faster than conventional wisdom can track. The Cognitive Data Stack is that kind of tool.

Appendix A. CDS Portfolio Summary

Company	CDS Layer	Category	Valuation	Structural Moat
VAST Data	I. Physics	Storage (DASE)	~$30B	AI Cloud Storage Monopoly
WEKA	I. Physics	Storage Acceleration	$1.6B	AI-Native GPU Performance
Databricks	I. Physics	Data Operating System	$134B	Lakehouse / Training-as-a-Service
Lambda	I. Physics	GPU Cloud Compute	$2.5–$15B*	Developer Cloud / MSFT Partner
Fireworks AI	I. Physics	Inference Platform	$4B	PyTorch Team / Open-Source
Scale AI	II. Genesis	Horizontal Foundry	$29B	Universal RLHF Supply Chain
Surge AI	II. Genesis	Horizontal Foundry	$15–$25B*	Bootstrapped / $1.4B Revenue
Snorkel AI	II. Genesis	In-House Data Factory	$1.3B	Programmatic Labeling / On-Prem
Anysphere	II. Genesis	Code Digitization	$29.3B	Developer Data Flywheel
Glean	II. Genesis	Knowledge Digitization	$7.2B	Enterprise Knowledge Graph
Harvey	II. Genesis	Legal Digitization	$8–$11B	Privileged Legal Data Monopoly
Abridge	II. Genesis	Health Digitization	$5.3B	Medical Reasoning Dataset
Hippocratic AI	II. Genesis	Health Digitization	$3.5B	Clinical Engagement at Scale
Anduril	III. Activation	Defense / Kinetic	$30.5B	Sensor-to-Action Autonomy
Applied Intuition	III. Activation	Simulation	$15B	Physical AI Safety Layer
Celonis	III. Activation	Process Intelligence	$13B+	Enterprise Execution Engine
Peregrine	III. Activation	Integration OS	$2.5B	Government OS
Hightouch	III. Activation	Reverse ETL	$1.2B	AI Decision Execution
LangChain	III. Activation	Agent Framework	$1.25B	118K GitHub Stars Ecosystem
Turing	Service	Talent Cloud	Profitable	$300M+ ARR

** Valuations with asterisk denote reported fundraising discussions (Bloomberg, Reuters, Sacra) where a round may not have formally closed.*

Appendix B. Glossary (Selected)

- **MDS (Modern Data Stack):** cloud-native tools optimized for analytics/BI; batch pipelines; dashboards.
- **CDS (Cognitive Data Stack):** infrastructure to store, refine, evaluate, govern, and activate data for AI reasoning and action.
- **RLHF:** Reinforcement Learning from Human Feedback; alignment via human preferences/ratings and iterative optimization.
- **Reverse ETL:** operational activation of analytics/ML outputs by syncing them into business tools.
- **DASE:** Disaggregated Shared Everything; storage architecture eliminating tiering bottlenecks.
- **Data Gravity:** the physics/economics of moving large datasets; value concentrates where data already lives.
- **Reasoning Data:** data explicitly shaped to teach, evaluate, or constrain model behavior.

Appendix C. Disclaimers

The materials, content and opportunity described herein are for informational purposes only and for the exclusive use of investors outside the United States of America that are considered sophisticated and/or professional investors in the jurisdictions of their residence and/or organization; this opportunity is not suitable for retail investors.

***No U.S. Distribution:** This document is not directed at, nor intended for distribution to, any person or entity in the United States, nor to U.S. persons (as defined under Regulation S of the U.S. Securities Act of 1933), and may not be distributed in any jurisdiction where such distribution would be unlawful.*

***Confidentiality:** This document and its contents are confidential and for the exclusive use of the person(s) to whom they are delivered and should not be copied or distributed, in whole or in part, or disclosed by such person(s) to any other person (and by receiving this document the recipient agrees so). This document is an outline of matters for discussion only.*

***No Advice:** This document does not constitute an offer to sell or the solicitation of an offer to buy any securities and should not be interpreted as advice, including legal, tax or accounting advice. Potential investors are encouraged to conduct their own evaluation and seek independent advice before making an investment decision. References to specific companies are for illustrative and educational purposes only and do not constitute any recommendation, endorsement, or view on the suitability of any investment in such companies.*

This document has not been registered, reviewed or approved by any regulatory authority and does not constitute an offer of securities. Such an offer may only be made by means of confidential offering materials.

***Risk Warning:** Investment in private equity involves a high degree of risk, it is an illiquid investment, speculative in nature and is subject to a risk of loss, including a risk of loss of principal. Past performance is not indicative of future results, and no express or implied guarantees are or can be provided regarding returns or capital recovery.*

AVC Turing Ltd. does not provide investment advice, portfolio management services, or personalized recommendations; all information herein is of a general and educational nature. AVC Turing Ltd. has no affiliation with any of the companies mentioned in this report, and does not have access to non-public, confidential, or insider information relating to any of these companies.

***Data Sources & Accuracy:** All financial data, valuations, ratios, and projections presented in this analysis are derived from publicly available sources including research reports, news articles, and market data providers as of Q1 2026. These sources may contain errors, inconsistencies, or incomplete information. Private companies' information and official financial disclosures are limited, and much of the data represents estimates or third-party analyses that cannot be independently verified. No information in this document has been independently verified by AVC Turing Ltd. or its affiliates.*

***Private Market Limitations:** Private equity shares are not publicly traded and are subject to transfer restrictions. Valuations are based on limited private market transactions and may not reflect actual realizable value. Liquidity is extremely limited and investment recovery may be difficult or impossible.*

***No Reliance:** No person should rely on the information contained in this document for the purposes of making any investment decision. Any investment decision should be based solely on independent due diligence and, where applicable, formal offering documentation.*

AVC TURING

Investing in the Future of Intelligence

New York | Global