AI Inference Distributed Data Challenge – NDNCOMM2025

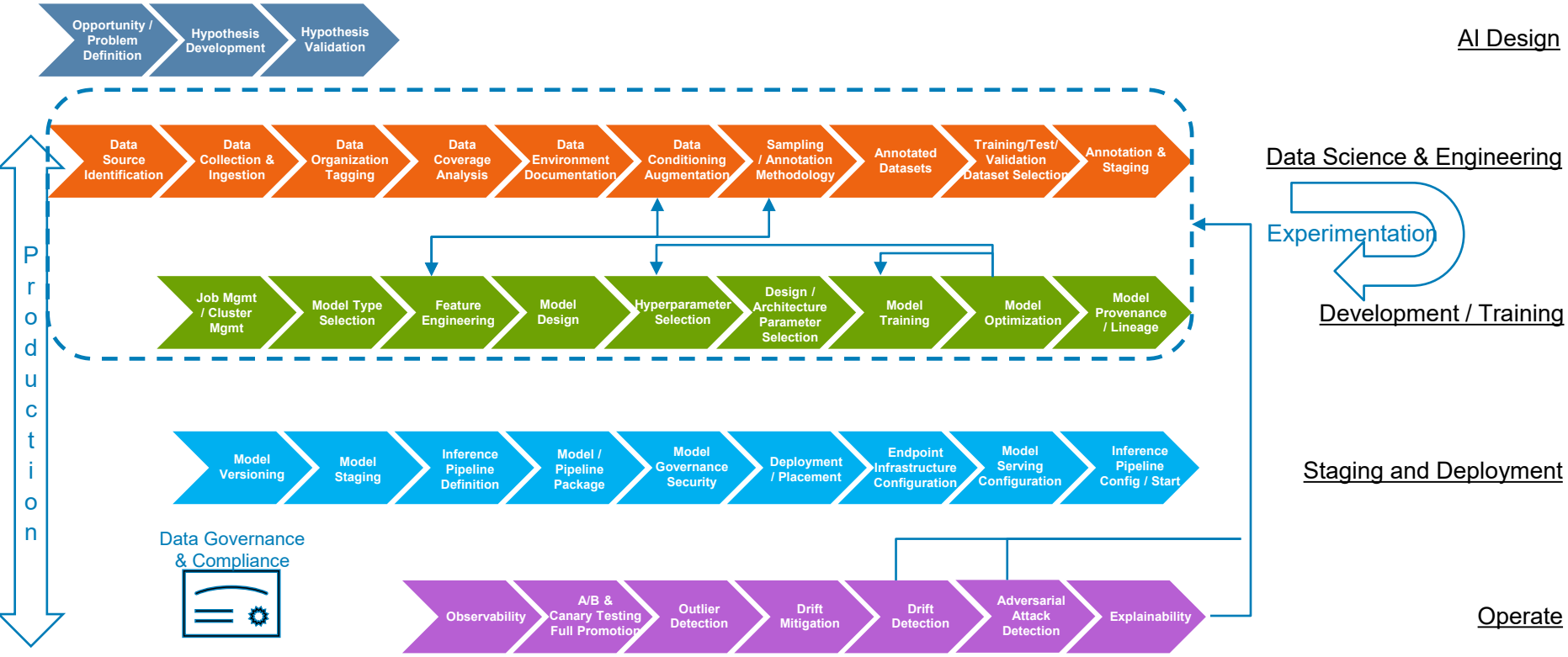
"We know the past but cannot control it. We control the future but cannot know it." Claude Shannon

Jeff White

April 17, 2025

DELL Technologies

AI DevOps CI/CD/CO Process



Cost Per Token for Foundation Models

Token
< 10 bytes

Embedded Representation
10's of Kb's

Embedding process expands tokens to a high dimensional dense vector representation.

RAG
Documents

Large Libraries of Embedded Content are Stored

RAG amasses large chunked embedded content.

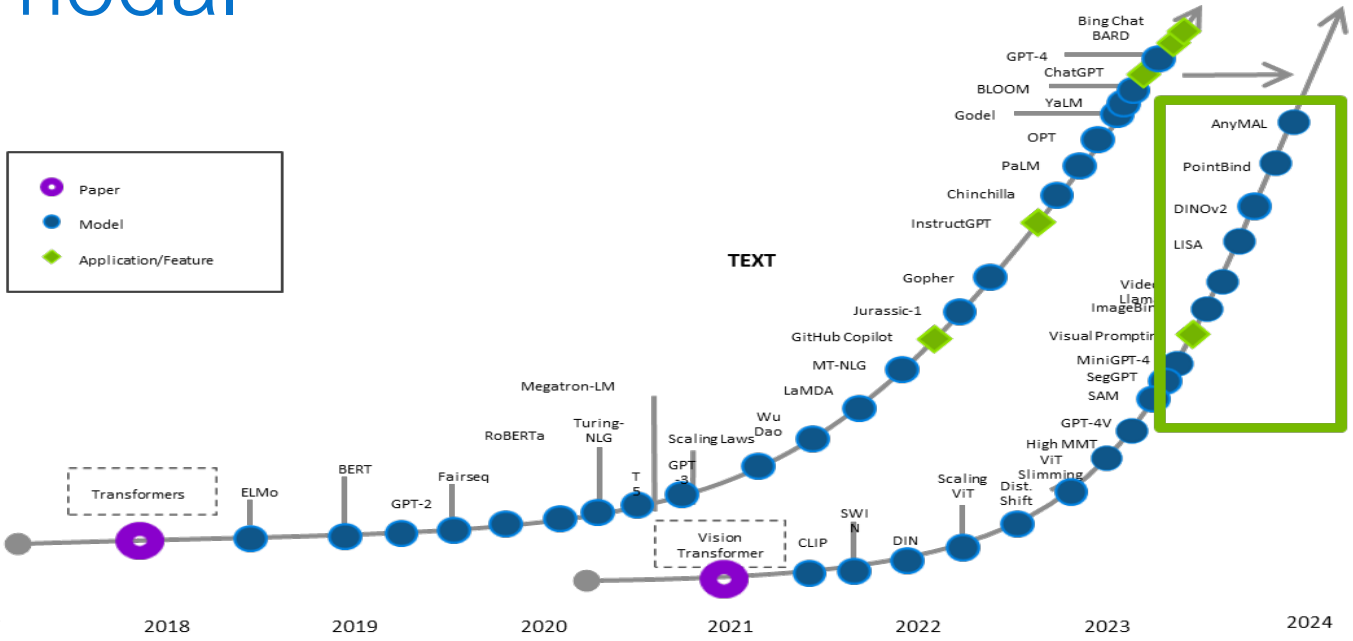
I/O Mgmt
Documents

Input and output sequence length Is increasing

With test time and other accuracy improvement techniques increase data requirements.

All these functions require computation, storage and networking.
Multimodal operation is greatly increasing the challenge of inferencing

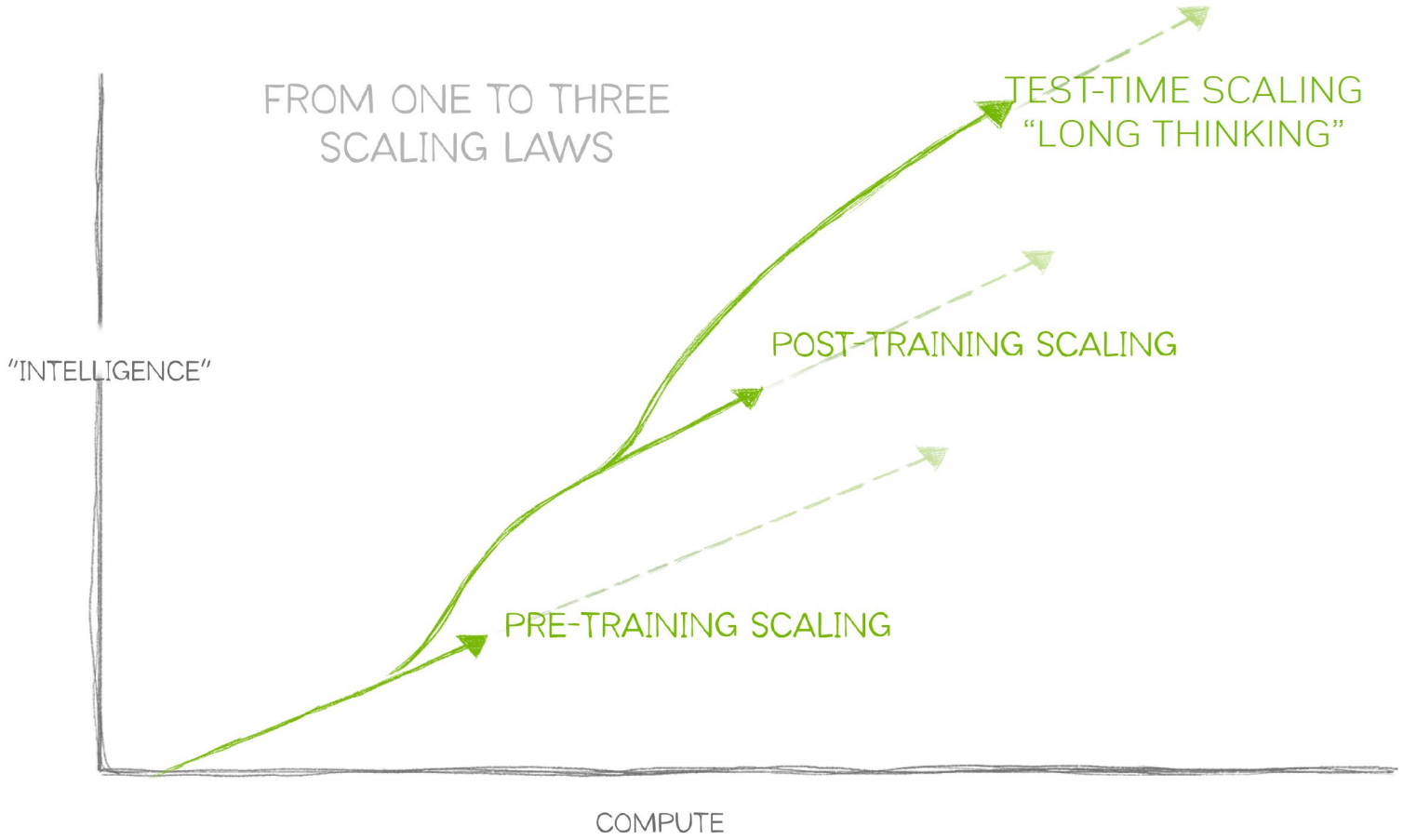
Generative AI is Accelerating – Led by Multimodal



Credit: Andrew Ng

AI Scaling Laws Drive Exponential Demand for Compute

New “long thinking” supercharges inference scaling



Trends in AI

- Prompt Length Extensions:
  - Enable better context – Hard due to Quadratic Complexity.
- Improvements to Foundation Model Execution:
  - Speculative decode, higher dimensionality (LoRA), KV caching, Input/Output controls, model distillation/compression.
- Modality Expansion:
  - Increase the number of use cases and ability to utilize foundation technology.
- Small Foundation Model Synthesis:
  - Improve methods to synthesize a Small Foundation Model that has same embedded representation of tokens as parent LFM.
- Model Architecture Evolution:
  - MoE and Multimodal improvements of transformer architecture, state space machine and new activation methods.
- Hybrid Discriminative and Generative Approaches:
  - Leveraging the strength of both AI to improve accuracy prediction, throughput and greatly increase addressable use cases.
- Objective Model Driven AI with Digital Twin:
  - Digital twins can model complex behavior and with generative AI can create control over large scale complex systems.
- Synthetic Data Content Evolution:
  - To create better datasets for training and to improve data collection methods
- Machine Concept Reasoning models with Function Calling and Inference Search:
  - Use of search techniques to optimize to complex reasoning with chain and graph mechanisms.
- Distributed and Disaggregated Inference:
  - Separate model functions and distribute them across multinodes and then multisite (tokenization, embedding, prefill. Decode

Future of Hybrid (Generative/Discriminative) Distributed Inference

- Generative AI based attention transformer foundation models have many components/functions:
  - Input processing (Tokenization/Embedding), Positional Encoding, Attention (Self/Cross/Multi-Head), Feed Fwd NN (MLP), Layer Normalization, Residual Connections, Output Processing (Decode, Activation/Softmax), Caching (KV, Decode), Inference Specific Optimization, Accelerator Mapping/Parallelism.
- Generative AI Inference also requires functions for Prompt/Input Management:
  - RAG (Augmented current search information), Guardrails (protection limits for inappropriate content and bias), Accuracy Evaluation, Prompt Logging/Context Tracking.
- Certain Model functions and Prompt/Input are latency, memory or computationally sensitive and embedded token size can grow quickly.
- Edge Inference can require multimodal/multisensor and use hybrid mixture discriminative+generative Multimodal inference is data intensive.
- SFM (<10B) can also be utilized where computational and management capabilities allow.
  - However, Input/Output functions must be present at the edge and MoE or Multimodal could increase model count and resource requirements.
- Some LFM/LLM functions can be architecturally distributed for inference to satisfy latency and data intensity.
  - Input/Output Management, Tokenization, Embedding, Position Encode, Caching, Accuracy Evaluation etc...

Symbolic

$\forall x \forall y (Rx \odot Cy < Qxy)$

Machine Learning

$f(x) = \sum_{i=1}^{n} \alpha_i * K(x_i, x) + b$

Reinforcement Learning

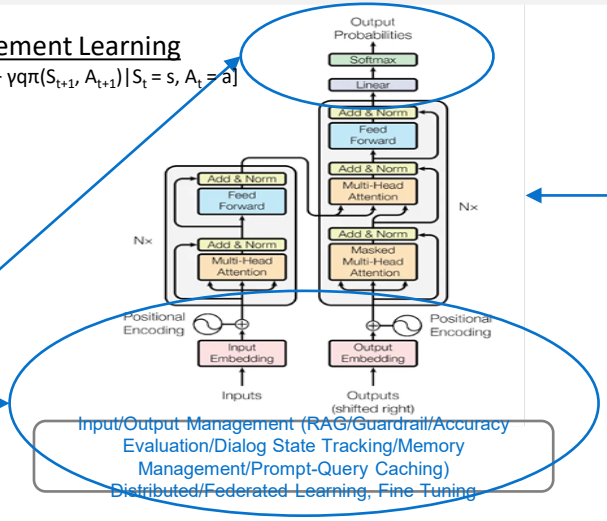
$q\pi(s, a) = E\pi[R_t+1 + \gamma q\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$

Discriminative

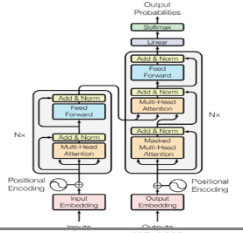
$P(Y|X)$

Generative

$P(X,Y)$



Small Local Foundation Model



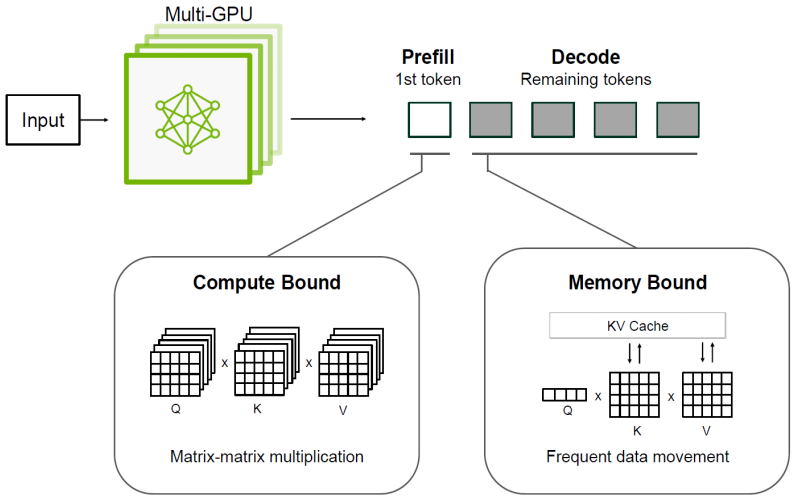
Input/Output Management (RAG/Guardrail/Accuracy Evaluation/Dialog State Tracking/Memory Management)
Distributed/Federated Learning, Fine Tuning

Potential Edge Distributed Functions

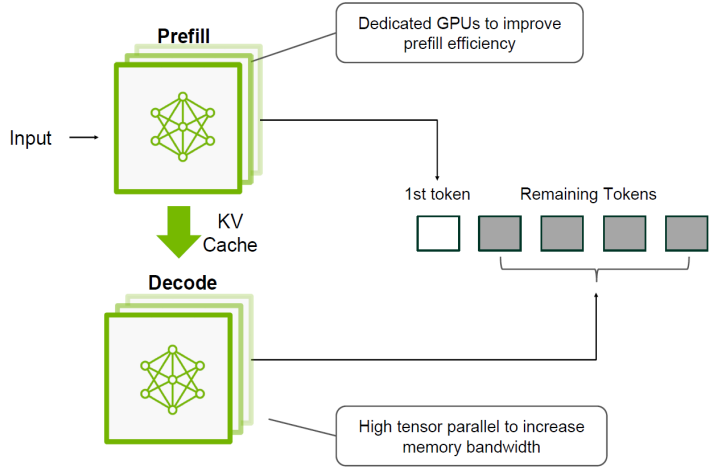
New Inference Optimization Techniques to Boost Inference

Disaggregated serving separates prefill and decode allowing each to be optimized independently

Traditional Serving

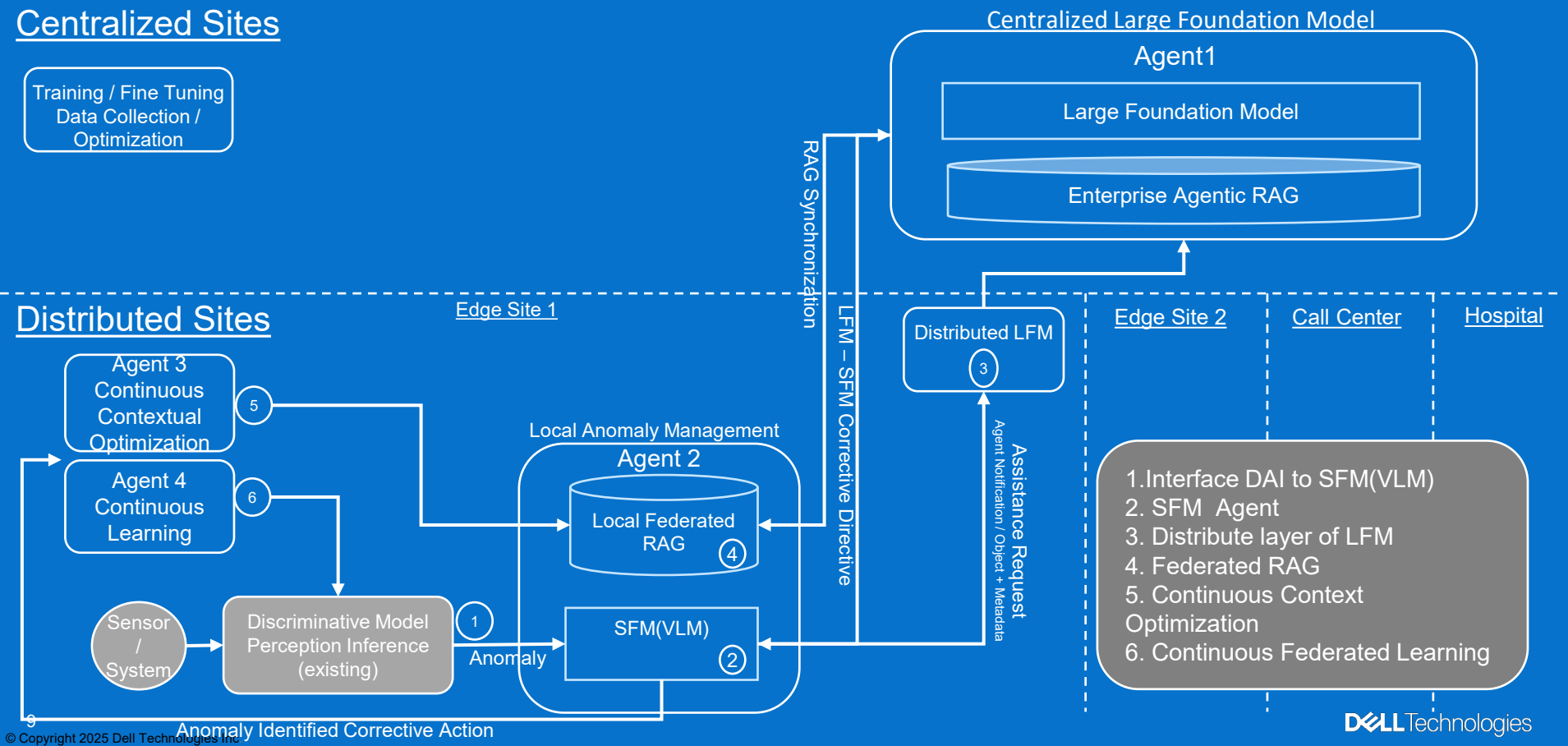


Disaggregated Serving



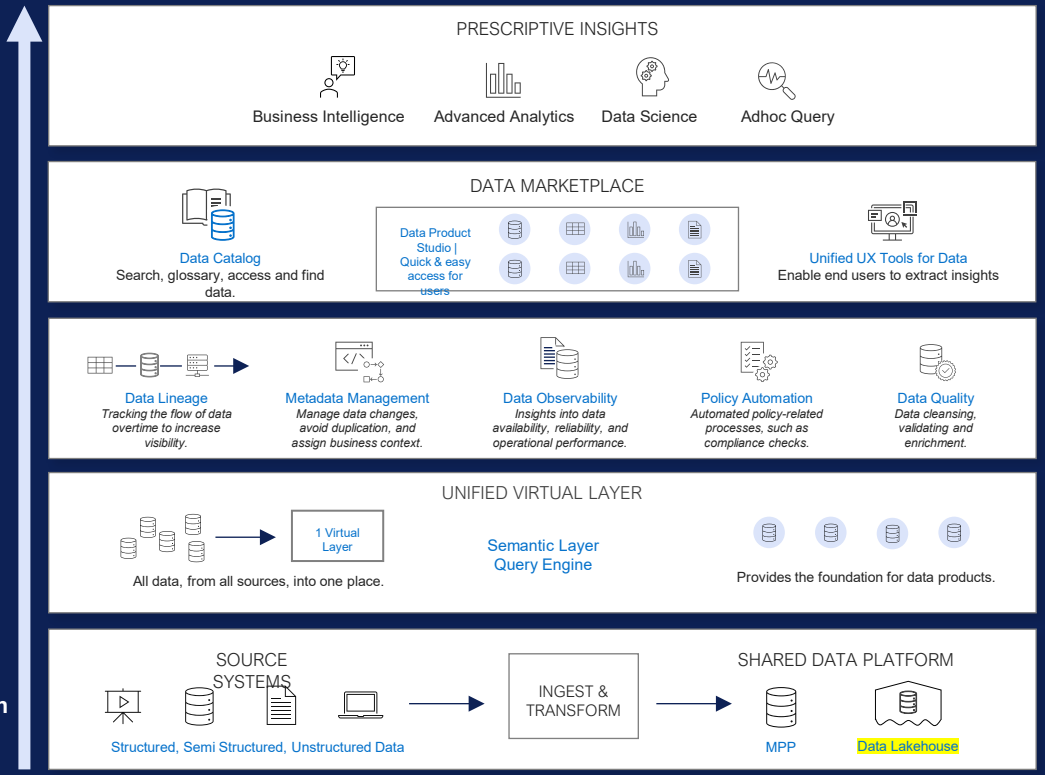
More flexibility to optimize cost and user experience

Hybrid Distributed Agent Based Inference



Modern Data Management Ecosystem

Capabilities to enable self service marketplace for data



- Cost-effective solution to enable multi-petabyte scale data ecosystem
- Data Mesh and Data Fabric methodologies will aid in this transformation
- Mindset shift from data as an asset, to data as a product – making data “FAIR” (Findable, Accessible, Interoperable, Reusable)
- Creating a seamless customer experience with pervasive self-service capabilities

AI Information Management Summary

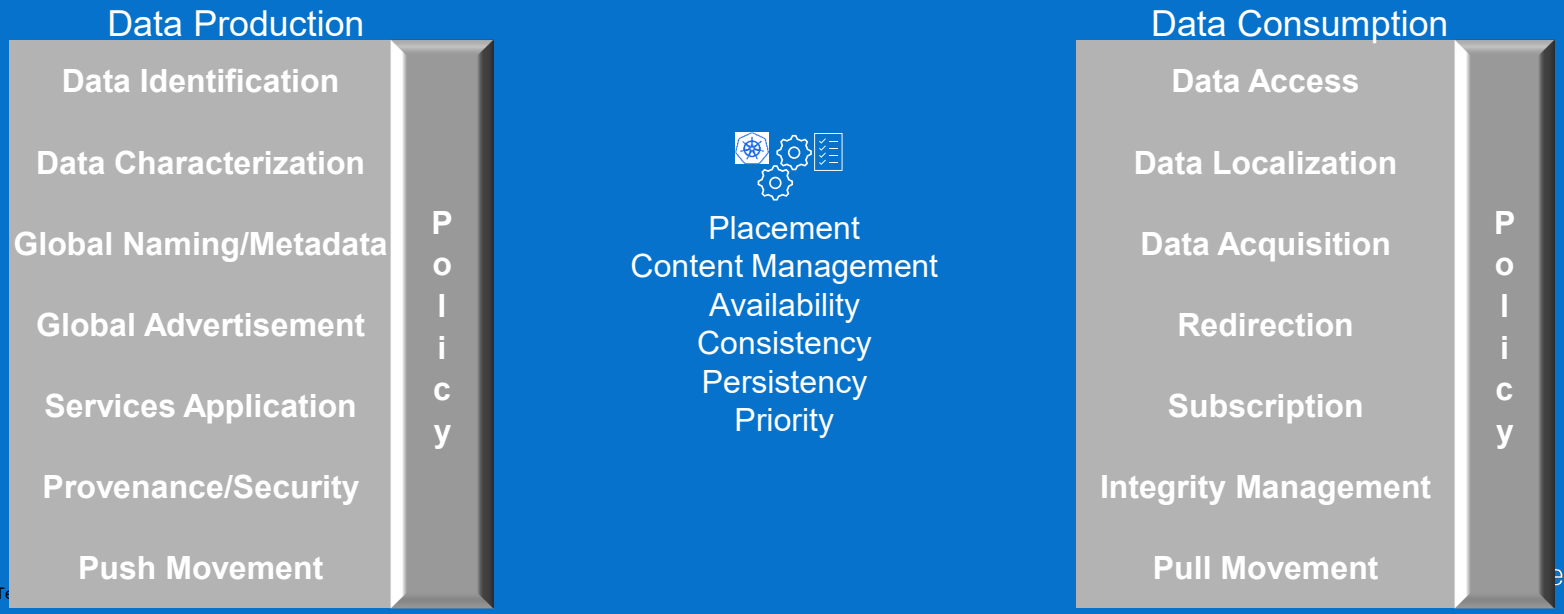
Integrated data management framework for dataverse management.

Data visibility, access, movement, replication and services applied autonomically.

Support for fixed and mobile data production/consumption.

Semantic platform known name space versus descriptive metadata (NDN!) with integration to platform control.

Policy control of data protection/security/naming and data services.



Thank You

“It is easier to build the future than to predict it” - Alan Kay

or

“All models are wrong, some are useful” - George Box

DELL Technologies