SEI

Data Strategy
The Foundation for GenAI Success

Data and a well-defined Data Strategy are *crucial* to successful GenAI Adoption

Is your Data an Enabler or a Deterrent?

We are at an exciting crossroads with AI and GenAI a top priority for organizations across all industries.
Here are some key fundamental reasons that make maturing their Data Capabilities crucial.

GenAI is Only as Good as the Data It Consumes

Data Strategy Aligns AI with Business Goals

Governance and Compliance Are Built on Data Foundations

Metadata, Context, and Semantics Matter

Operationalization Depends on Data Infrastructure

Feedback Loops Require Data to Improve

A deep dive…

Why is a deliberate Data Strategy an imperative for GenAI success?

A Data Strategy should be a pre cursor to your Gen AI solutions before they are deployed in Production.

Failing to do that, may cause challenges that erode trust, cost more and run the risk of getting defunded.

GenAI is Only as Good as the Data It Consumes

GenAI models rely heavily on high-quality, relevant, and structured data to generate accurate, valuable, and context-aware outputs. If the input data is fragmented, biased, outdated, or lacks depth, GenAI outputs will reflect those flaws, resulting in poor decisions, hallucinations, or reputational risk.

Data Strategy Aligns AI with Business Goals

A clear data strategy, with the right Data Governance Framework ensures that GenAI efforts are targeted at high-impact use cases, aligned with organizational priorities. It defines what data matters, who owns it, and how it will be governed, enabling scalable and responsible AI use.

Governance and Compliance Are Built on Data Foundations

GenAI introduces new risks related to data privacy, security, copyright, and explainability. A mature data strategy embeds governance frameworks to ensure regulatory compliance, ethical AI use, and trustworthy outputs, particularly critical in healthcare, finance, and regulated sectors.

Metadata, Context, and Semantics Matter

GenAI needs metadata, taxonomies, and knowledge graphs to understand the business context and produce domain-specific results. A strong data strategy helps define and manage this semantic layer, enabling more precise and useful generation. This is critical to ensure trust.
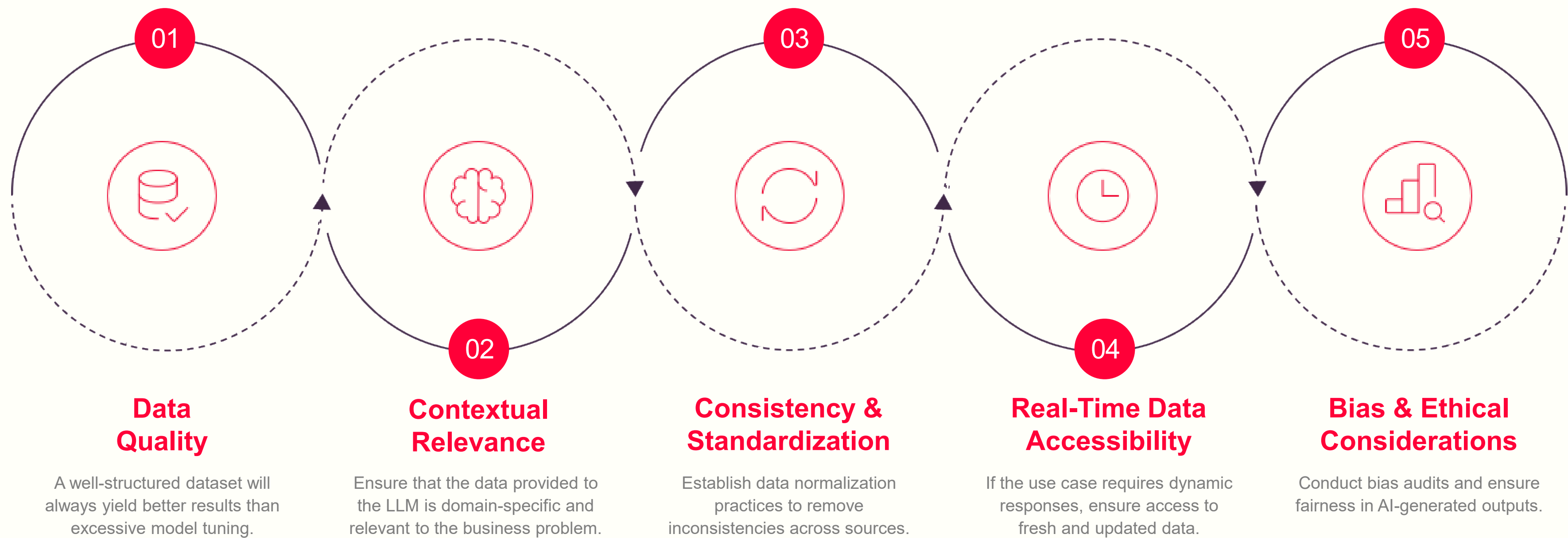
Operationalization Depends on Data Infrastructure

Deploying GenAI into production requires clean pipelines, data catalogs, feature stores, and APIs. A modern data architecture, enabled by a well thought out the data strategy, ensures that GenAI is not just a prototype, but a repeatable, secure, and governed solution.
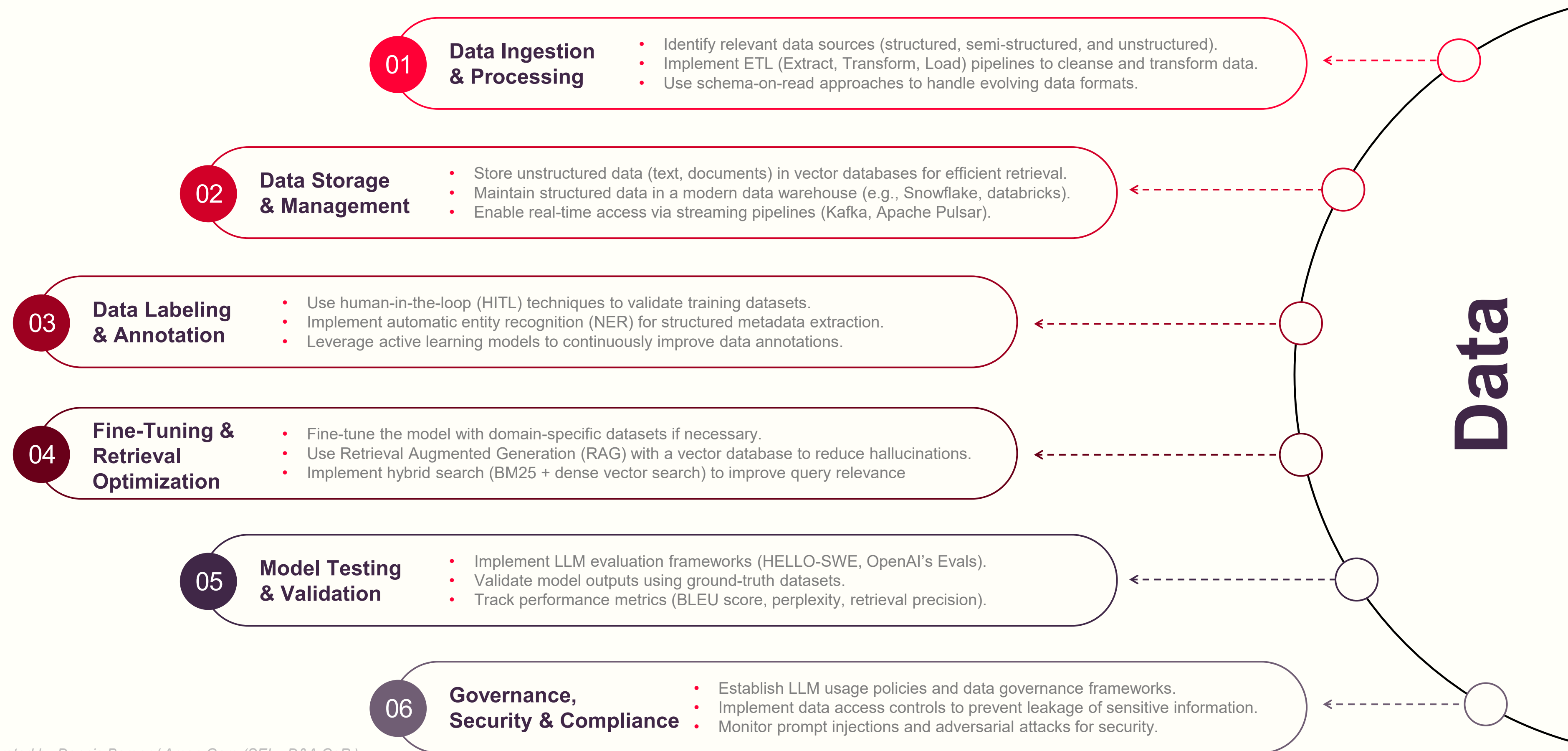
Feedback Loops Require Data to Improve

Continuous learning, fine-tuning, and reinforcement mechanisms need labeled data and user feedback. A data strategy ensures the organization has the systems to capture this feedback, close the loop, and refine the GenAI models over time.

Core Data Principles for LLM Performance Optimization

01

Data Quality

A well-structured dataset will always yield better results than excessive model tuning.

02

Contextual Relevance

Ensure that the data provided to the LLM is domain-specific and relevant to the business problem.

03

Consistency & Standardization

Establish data normalization practices to remove inconsistencies across sources.

04

Real-Time Data Accessibility

If the use case requires dynamic responses, ensure access to fresh and updated data.

05

Bias & Ethical Considerations

Conduct bias audits and ensure fairness in AI-generated outputs.

Making Data Usable, Valuable, and Error-Free

01 Data Ingestion & Processing
- Identify relevant data sources (structured, semi-structured, and unstructured).
- Implement ETL (Extract, Transform, Load) pipelines to cleanse and transform data.
- Use schema-on-read approaches to handle evolving data formats.

02 Data Storage & Management
- Store unstructured data (text, documents) in vector databases for efficient retrieval.
- Maintain structured data in a modern data warehouse (e.g., Snowflake, databricks).
- Enable real-time access via streaming pipelines (Kafka, Apache Pulsar).

03 Data Labeling & Annotation
- Use human-in-the-loop (HITL) techniques to validate training datasets.
- Implement automatic entity recognition (NER) for structured metadata extraction.
- Leverage active learning models to continuously improve data annotations.

04 Fine-Tuning & Retrieval Optimization
- Fine-tune the model with domain-specific datasets if necessary.
- Use Retrieval Augmented Generation (RAG) with a vector database to reduce hallucinations.
- Implement hybrid search (BM25 + dense vector search) to improve query relevance

05 Model Testing & Validation
- Implement LLM evaluation frameworks (HELLO-SWE, OpenAI's Evals).
- Validate model outputs using ground-truth datasets.
- Track performance metrics (BLEU score, perplexity, retrieval precision).

06 Governance, Security & Compliance
- Establish LLM usage policies and data governance frameworks.
- Implement data access controls to prevent leakage of sensitive information.
- Monitor prompt injections and adversarial attacks for security.

Data

*Curated by Dennis Barger / Aman Garg (SEI – D&A CoP)*

Challenges facing D&A Leaders

Today's leaders are faced with the challenge of delivering AI innovation without clear direction, skilled employees and in-depth understanding of the resources needed to make AI successful.

Data is an enabler for AI solutions. Enablement requires:

- Data strategies to increase data maturity across the organization

- Data platforms that support scalability, flexibility and acceleration of new solutions

- Organizational governance and literacy of data supporting business initiatives
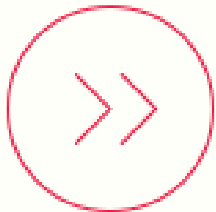
Pressure to Accelerate

D&A leaders are under pressure to deliver results faster, even if the company doesn't have a clear plan in place.
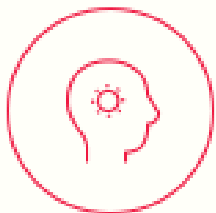
Upskilling Employees

Training employees to work with data is difficult due to partial support and data maturity across the organization.
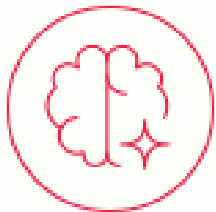
AI Knowledge

Leaders want to use AI, but there is a gap in understanding what is needed to make it work, including skills, budgets and resources.
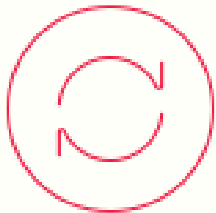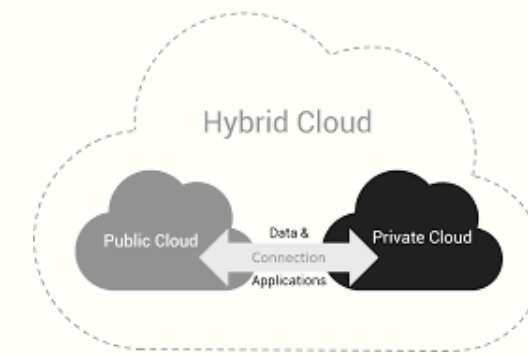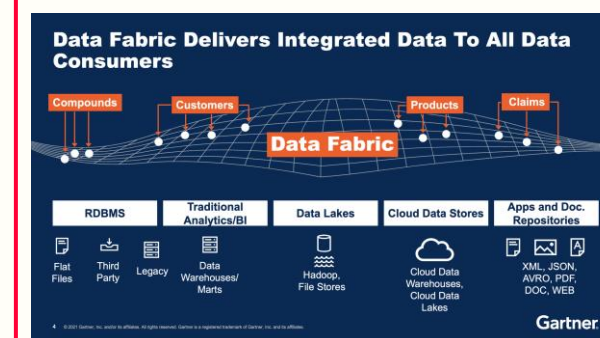
Evolving Role

AI is changing what D&A leaders do. They need to adjust their strategies and ways of working to keep up with growing demands.
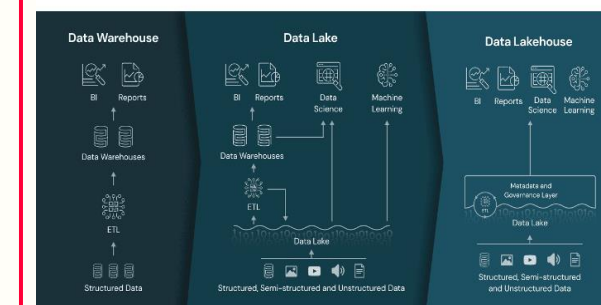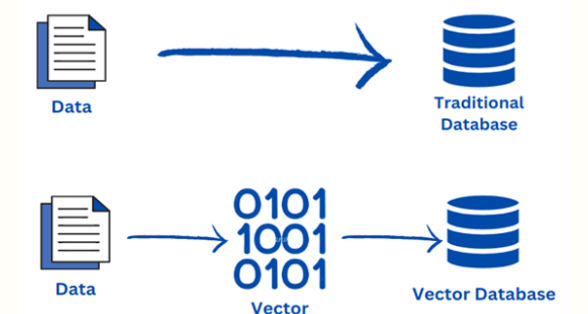
Data Technologies, Platforms & Frameworks

Microsoft Azure · aws · Google Cloud

Hybrid Cloud · databricks · snowflake®

Data Fabric	Lakehouse	Vector Databases

SEI

National D&A Practice Leads

Dennis Barger
dbarger@sei.com
(404) 431-1432

Aman Garg
agarg@sei.com
(513) 484-4151