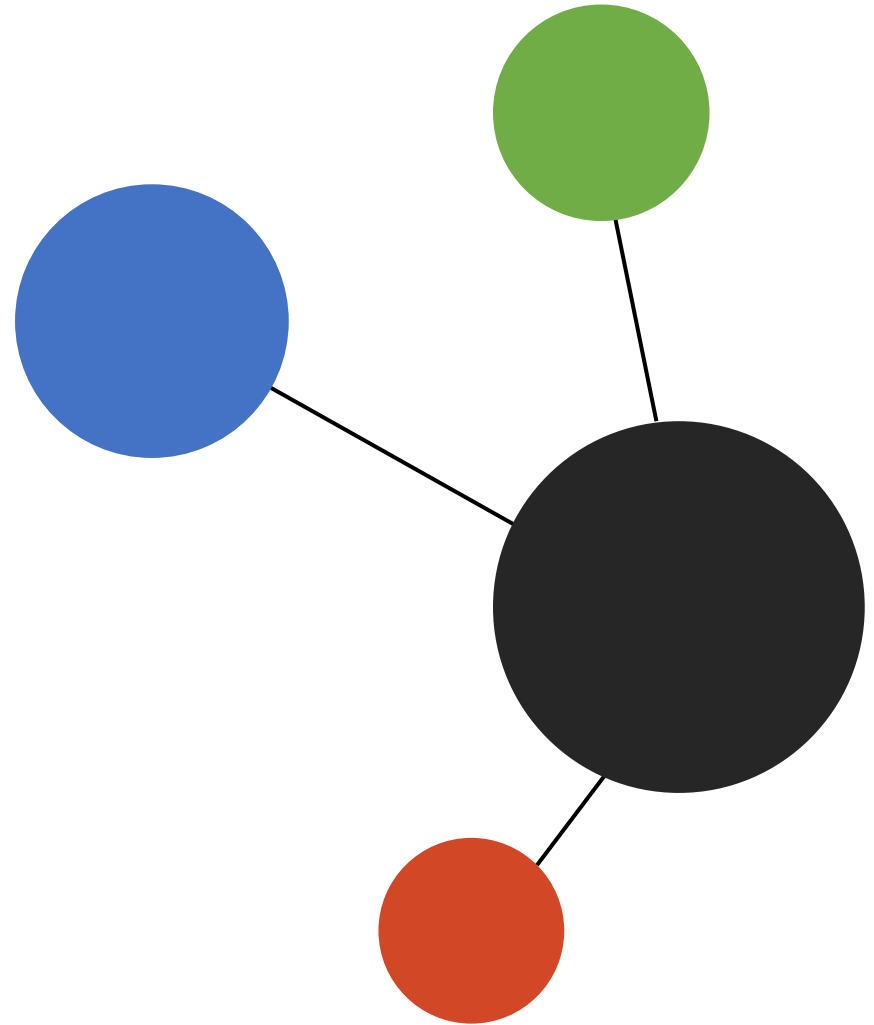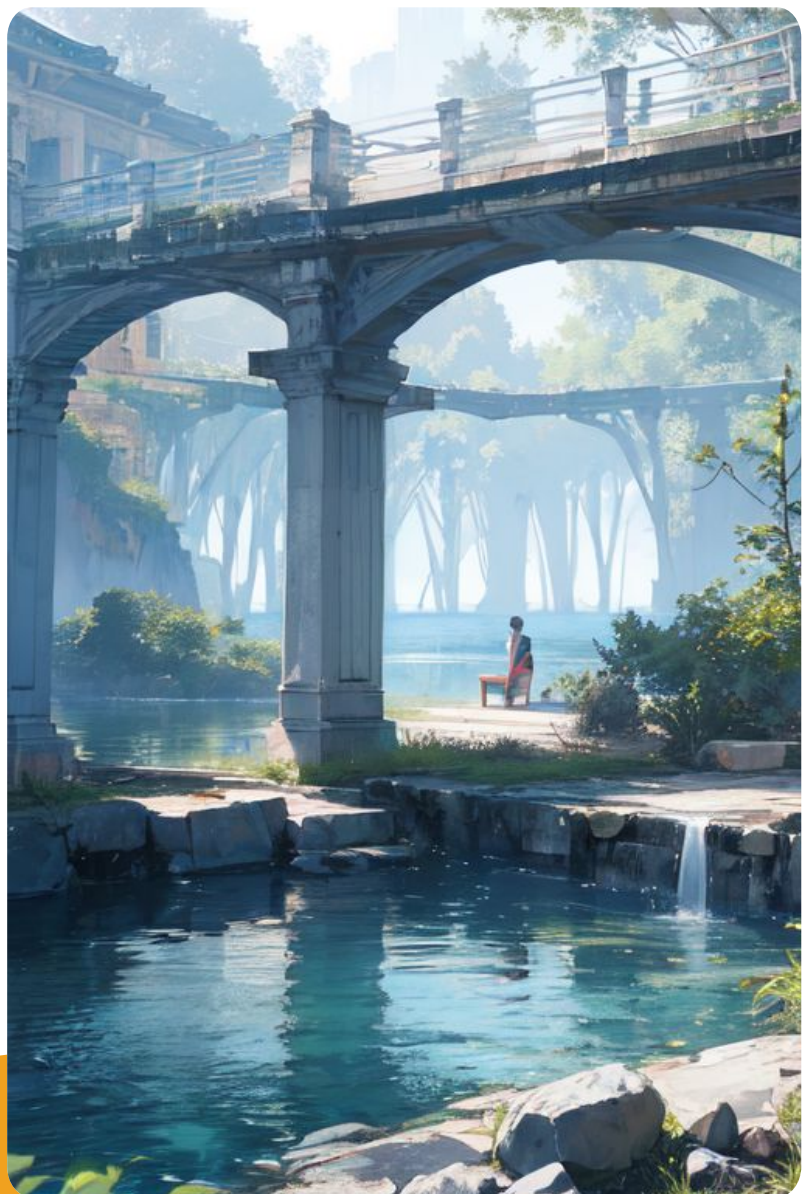# Twinblue Knowledge Lake

AI Data Requirements and AI Knowledge Lakes – A New Era

Introduction

AI Knowledge Lakes vs. Traditional Data systems

- **Natural Language Querying**: AI systems use natural language, making interactions more intuitive, unlike traditional SQL.

- **Cross-Related Documents**: Multiple documents contribute to single topics, enhancing depth and context. AI knowledge lakes manage these cross links.

- **Dynamic Correlations**: Document correlations are query-dependent; AI knowledge lakes correlate data at runtime.

- **Flexible Output Formats**: Query answers can be dynamically formatted to meet various needs.

- **Massive Data Volumes**: AI systems manage vast amounts of data, with even small organizations handling terabytes of knowledge data.

AI Knowledge lake data formats and challenges

Knowledge lakes must support all data types, including Rich Data

- Knowledge lakes support structured, semi-structured and unstructured data

- Must store data in its native format for accuracy and reference.

- Includes rich data – for example documents, images, audio, video, and more.

- AI applications require the Knowledge Lake to transform and utilize this diverse data.

- KTL is a new concept. It implies a requirement to knowledge transform the data for AI applications
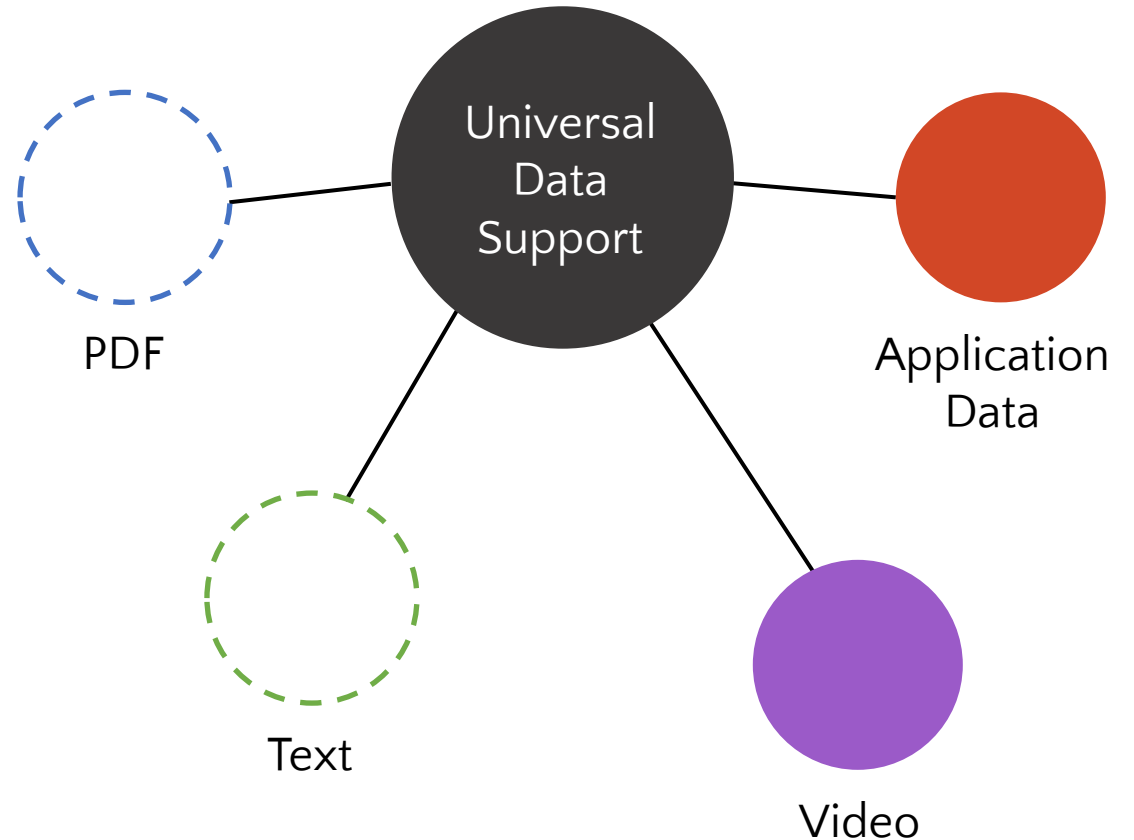
Knowledge Transformation & Load (KTL)

- **Essential for Knowledge Lakes**: Transforms any data into AI–compatible formats (semantic vectors, frequency tables, knowledge content tables).

- **Specialized Storage and Processing**: AI data structures require special processing during the search phase; traditional databases or data lakes are insufficient.

- **Knowledge Enrichment**
  - Lookup and add to the knowledge by querying external and internal systems during ingestion of documents and knowledge
  - Using AI LLMs to generate distill additional knowledge from the incoming document and add it to the knowledge lake

- **Native Data Preservation**
  - Retains data in its original format
  - Includes reference pointers to KTL data for AI applications

PDF / Word

Universal Data Support

Application Data

Text

Video
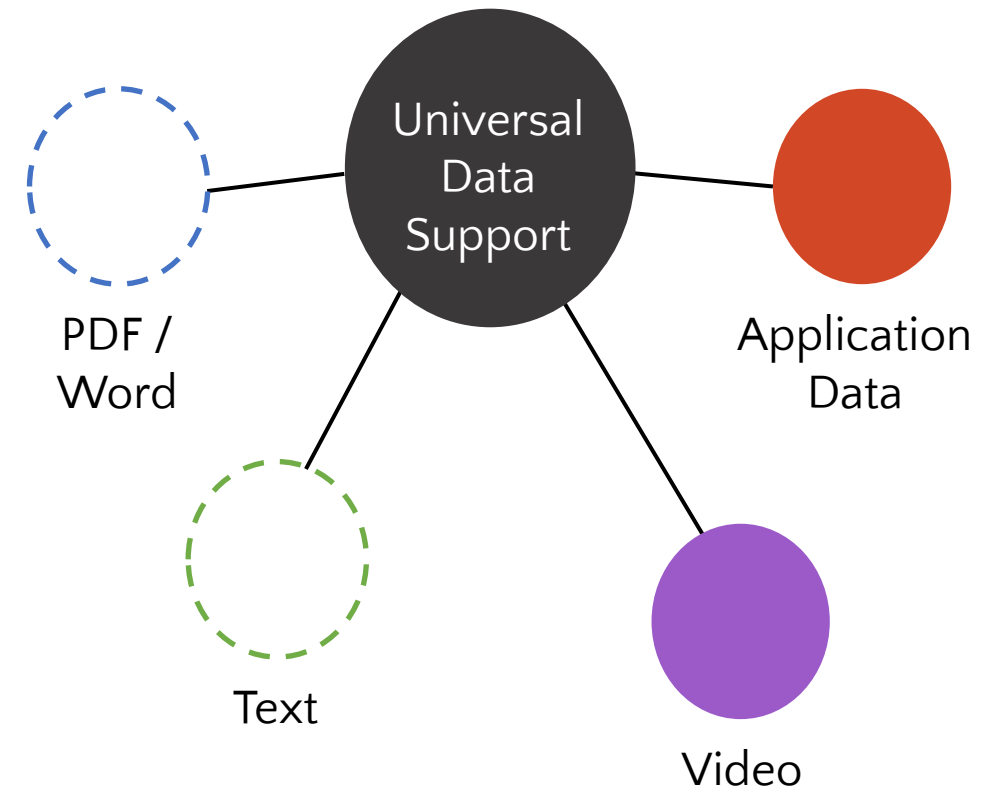
Scalability

Distributed and Redundant Storage

- Ensures high availability and fault tolerance.

- Data is replicated across multiple nodes to prevent loss and improve access speed.

- Supports scaling as data volume grows, accommodating vast amounts of data seamlessly.

Distributed Query Execution

- Enables handling of large datasets efficiently.

- Breaks down complex queries into smaller tasks that are processed in parallel across multiple nodes.

- Reduces query execution time, providing faster insights and responses.

- Can use GPUs or multi-core systems for efficient AI modeling and calculations, helping in richer query response

By integrating these distributed systems, AI knowledge lakes can manage and analyze massive data models effectively, ensuring robust performance and scalability.

Enhancing AI Lakes with Metadata

Rich Metadata Support for Lookups and Search

Query targeting

- Enable searches within specific subsets of documents.

- Improve efficiency and accuracy by focusing on relevant data.

Document Tagging

- Facilitate easier lookups with comprehensive tagging systems.

- Support conditional searches based on tags.

Elastic Metadata Formats

- Adapt to various formats and conditions.

- Ensure flexibility during querying to meet diverse needs.

The knowledge lakes have a necessity of rich metadata support for effective lookups and search functionalities including conditional searches based on metadata (eg dates) or document tags.

Security and Partitioning

Security Partitioning into Knowledge Islands

- Knowledge Islands for Role-Based or Department based access.

- Restrict access to sensitive data based on user roles.

- Prevent unauthorized access to secure information, e.g., financial data.

Application based partitioning into Sub-Lakes

- Segment datasets for specific use cases and AI applications.

- Optimize performance and management by isolating data relevant to distinct processes within sub-lakes

Bridges

- Join segmented partitions for deep lake searches or correlations across larger swathes or the entire lake

- Allow knowledge deep dives for an across-the-org query.

AI Querying in Knowledge Lakes

AI knowledge lakes primarily support querying with Natural Language

Input as Plain Text Query

- Users can input queries in natural language, making it accessible and intuitive.

- No need for complex query languages, reducing the barrier to entry.

Incorporation of Metadata Conditions

- Enhance searches by combining plain text queries with metadata conditions.

- Enables more precise and context-aware results.

- Increases knowledge lake performance by dynamically segmenting the datasets using metadata conditioning lambda functions

Defining Output Format

- Users can specify the desired format for the output results.

- Customizable outputs ensure the data is presented in the most useful and relevant way for the user's needs.

- Output formats can range from document sets, to context nuggets, to tabular format or other user defined formats

AI Knowledge lakes are the backbone of AI applications

AI Knowledge Lakes should allow for easy integration of AI application frameworks

- AI application frameworks utilize the knowledge lake as the LLM backend in various AI applications.

- Framework Example: Neuranet:

- Twinblue Knowledge Lake seamlessly integrate frameworks like Neuranet.

This allows converting the knowledge lake data and queries into actionable results using AI LLMs.

AI application use cases

Answers to Questions:

- Generate precise, contextually relevant answers from natural language queries.

- Utilize AI to understand and respond to complex questions effectively.

Analytics and Reports:

- Produce detailed reports based on the data within the Knowledge Lake. Visualize data through AI-generated charts and graphs.

- Facilitate easier interpretation and analysis of large datasets.

Agent Flows:

- Develop and manage agent workflows using AI insights.

- Optimize processes and automate decision-making for enhanced efficiency.

There is an elegance and simplicity in the correct solution

Most AI challenges are solved with Knowledge Lakes.

Expensive GPU problem

- Knowledge lakes can power completely disconnected AI that does not need expensive CPUs.
- Knowledge lake + Micro LLM (that can run on consumer grade laptops) > GPT 4. Eg https://arxiv.org/abs/2404.14219

Security

- Knowledge lake islands solve most AI security challenges elegantly – role-based access, data isolation, data privacy concerns.

Data Integration to Knowledge transformation

- Knowledge lakes KTL integrates and transforms all business data into useable AI knowledge.

Real time AI

- Knowledge lakes are real-time and help building analytics, and real-time response AI applications and agents.

Skill challenges

- Knowledge lakes remove complexity – any developer can shift to AI paradigm as the concepts are similar to what they know – Query knowledge lake, pass response to AI LLM to get the final answers.

Large data and knowledge storage

- Knowledge lakes accumulate and organize terabytes of data and provide knowledge access to AI applications, without requiring any other databases or software.

Twinblue
Knowledge
Lake

Twinblue – Industry's first Knowledge Lake

Tekmonks' Twinblue knowledge lake product provides every one of these features, for an industry first and complete AI application backbone data and knowledge management system.

•Contact our sales for a briefing and to plan your knowledge lake and AI projects.

THANK YOU!