

# Corporate Data, AI and Data Governance

## The dichotomy of needs

*October 2025*

# Agenda

---

1. Are you ready for Enterprise AI?
  - a) Chasing the benefits of using AI – The Reality!
  - b) Corporate data & AI agents or engines
  - c) Data Governance – A pre-requisite for AI enablement
2. Current methods, tools and pitfalls:
  - a) Unstructured Data
  - b) Data Security, Classification and Leakage
  - c) Semantic Layers
3. Suggestions for success:
  - a) Compare Layers – Medallion Architectures
  - b) Compare Cloud Data Warehouse Toolsets
  - c) Compare Methodologies

# Are you ready for Enterprise AI?

# Chasing the benefits of using AI – The Reality!

1. **AI relies on accurate data** – Inaccurate data will pollute your perspective and deliver incorrect results. Few are addressing data quality ahead of AI.
2. **AI training** – 99% of AI engines constantly require re-training every time new data is introduced.
3. **Public Domain Data vs Corporate** – Few if any corporates are exposing their corporate data to AI engines, most rely on some form of semantic view of underlying data to enable a level of control.
4. **Current value** – Building code, test cases, automation, documentation and research type analysis with immense time benefit but little tactical, operational or strategic benefit. Predictive analytics is still in its infancy.
5. **Low trust of outputs** – Without human intervention (secondary contributions) and review, there are very few automatically delivered corporate data AI initiatives.

1. **Sensitive corporate data** – Role Based Access Control (RBAC) type access is difficult to enforce on free format data (PDF, Doc, Jpeg, Img & mp3/4).
2. **Operational systems integrated data** – Standard relationships and embedded data within operational systems make some data isolation and restrictions difficult.
3. **Semantic layer definitions** – Often, the operational nature of the raw data is focussed on operational function and not logically separate 'enough' to define role-based access (RBAC) at a field level.
4. **Repositories & abstraction from the source of truth** – Any ETL/ELT type processing before AI usage adds a potential point of failure & inconsistency to AI results.
5. **Avoid at all costs unauthenticated AI agents** – Any unauthenticated user access permissions can lead to undesirable access or data leakage, ensure EntraID (or similar) linked users are used to manage user lifecycles and access control.

# Data Governance – A prerequisite for AI enablement

1. **Data Governance** – Attempting to retro-fit data governance will leave gaps.
2. **Data Ownership and Stewardship** – Must be explicitly controlled and owned for every source of AI, even if only at a system level. Responsibility changes human behaviour & improves control.
3. **Be sure to add KPI's** – Including measures and visibility makes errors easier to identify and correct. Earlier identification makes remediation easier.
4. **Data Lifecycle** – Early definition of data lifecycles avoids data bloat and, at the very least, hive off obsolete data to keep your AI performing.
5. **Semantic layer access** - Provides less opportunity to control what's visible to the AI agent since it doesn't know (without explicit training) what to ignore in a collection of data where some items may be wrong.
6. **Process impacts** - Often requires business process reviews and/or change to address gaps.

# Current Methods, Tools and Pitfalls

1. **Leave unstructured data where it lives** – Duplicating content means duplicating security, access permissions and data lifecycle rules.
2. **Data Lakes** – Cause data bloat of up to 80% which, due to increasing volumes of data becomes costly and unsustainable over time.
3. **Business Self-Service** - Data Lakes ensure the inability of business users to self serve and build a reliance on Data Science discovery activities.
4. **Data Leakage** - Risk of data leakage through absence of role-based restrictions on content means you will require additional filters.

**By its very nature, governance is an afterthought for unstructured data hence point 1 above.**



1. **Field Level Security** - Security access controls are most often at record level, field level is few and far between.
2. **User Access** - Complete system access restriction reduces the ability for self-service, builds reliance on data science for true cross system data usage.
3. **Common Nomenclature** - Organisational context is often represented differently across different systems, this makes for difficulty in consistent data cataloguing and single organisational nomenclature with inherent security complexity.
4. **Support** - Most tools modify the query to suit the access controls making dynamic execution varied, this causes problems debugging as the security controls dictate what data is being read and how.

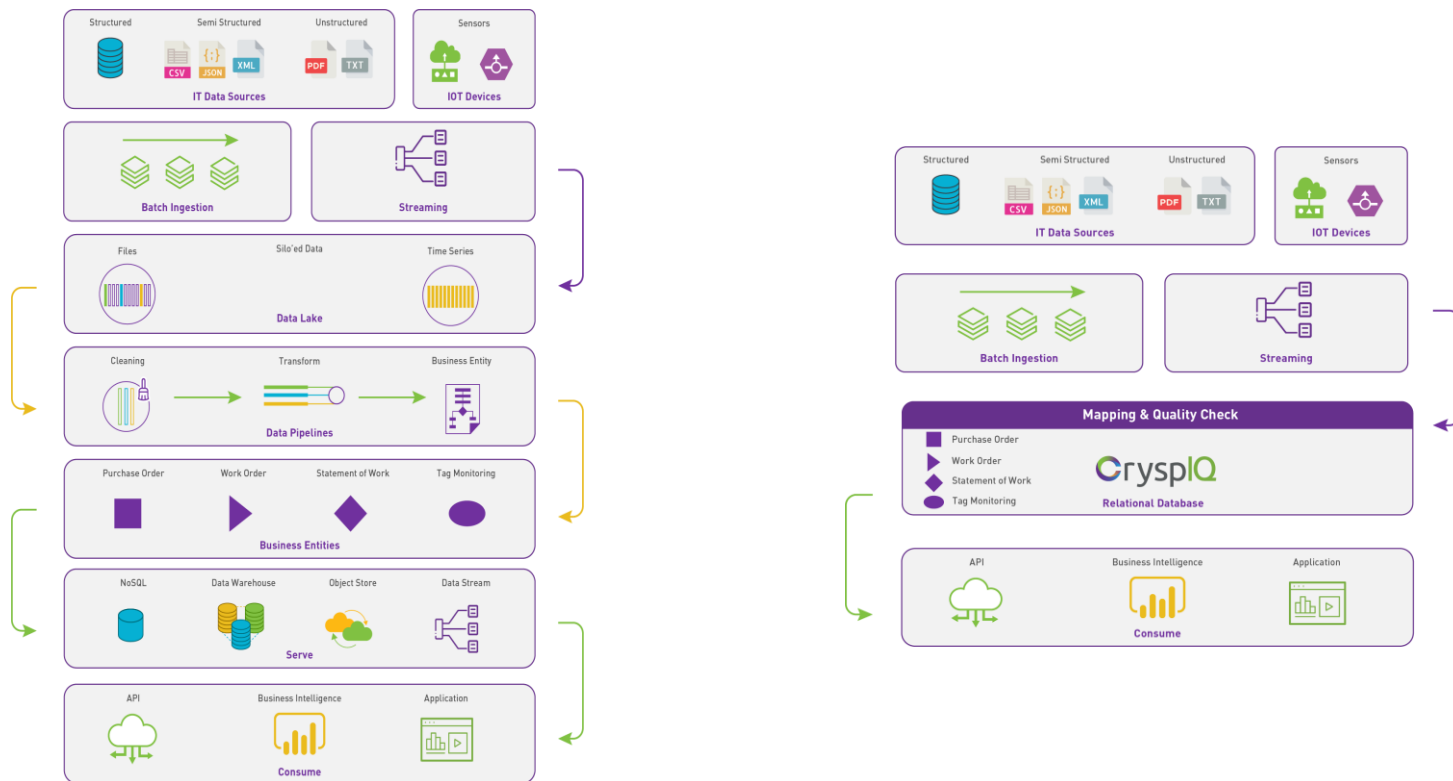
**Remember, if the data never leaves the server, it cannot be leaked inadvertently.**

1. **Semantic Layer Access** – BI/AI query engines using semantic layers are not throttled and will likely put additional load on operational systems at some stage.
2. **Change Management & Operational Configuration** - Modern system dynamic configuration capabilities means there is seldom a single semantic layer definition that suits all organisations without some form of tweaking.
3. **Multi-System Access** - These remain largely siloed and singular in their coverage and additional developments are required to access more than one system.
4. **Common Context** - Often the nomenclature of a semantic layer does not align to your organisations terminology and may lead to inconsistent understanding.

**Governance is not provided in any semantic layer so needs additional product(s) or development, especially for AI generated data**

## Suggestions for success


# Compare Layers – Typical Configuration vs CryspIQ



# Compare Cloud Data Warehouse Toolsets

Functional	CryspIQ	Snowflake	Databricks
Data Collection Model	Static	Subjective	Subjective
Factual Data (Business Critical) or All Data	Factual	All Data	All Data
Data Footprint	Small	Large	Large
Multi-Cloud	Yes	Yes	Yes
Separate Storage and Compute	Yes	Yes	Yes
Query Language	SQL	Snowflake SQL	SQL
Massively Parallel Processing (MPP)	Yes	Yes	Yes
Columnar	Yes	Yes	Yes
Foreign Keys	Yes	Yes	Yes
Structured Data	Yes	Yes	Yes
Unstructured Data	Yes	Yes	Yes
Concurrency	Yes	Yes	Yes
Automation	Yes	No	Yes

# Compare Methodologies

	Bill Inmon	Ralph Kimball	Data Lake	Data Vault 2.0	CryspIQ®
Definition	Method is the top-down or data-driven strategy, in which we start with the data warehouse and break it down into data marts.	Method is the bottom-up approach where data marts are first created to provide reporting and analytical capabilities for a function.	Method is storing data within a system or repository, in its natural format, that facilitates the collation of data in object blobs or files.	Method is designed to provide long-term historical storage of data coming in from multiple operational systems.	Method is the decomposition of source records to allow one to store the incoming data at the granular level clustered with data of like type.
Advantages	<ul style="list-style-type: none"> <li>Ensures data quality and consistency</li> <li>Facilitates data integration and maintenance</li> <li>Normalised source of truth across functions</li> <li>Supports the scope and depth of data analysis</li> </ul>	<ul style="list-style-type: none"> <li>Low upfront effort</li> <li>User-friendly and understandability</li> <li>Allows for flexibility and data model scalability</li> <li>Faster analysis and reporting</li> </ul>	<ul style="list-style-type: none"> <li>Source flexibility</li> <li>Low Upfront effort</li> <li>Speed to data availability</li> <li>No technical dependency</li> <li>Low cost</li> </ul>	<ul style="list-style-type: none"> <li>Scalable platform</li> <li>Source flexibility</li> <li>Lineage and traceability</li> <li>Enables automation</li> </ul>	<ul style="list-style-type: none"> <li>Scalable platform</li> <li>Source flexibility</li> <li>Lineage &amp; traceability</li> <li>Enables automation</li> <li>Low effort upfront</li> <li>Single source of truth</li> <li>User self service</li> <li>Data quality checking</li> </ul>
Disadvantages	<ul style="list-style-type: none"> <li>High upfront effort</li> <li>Reduces efficiency of data analysis and reporting.</li> <li>Less user-friendly</li> <li>Query complexity increases over time</li> <li>Limited flexibility and scalability</li> </ul>	<ul style="list-style-type: none"> <li>Limited scalability</li> <li>Compromises the quality and consistency</li> <li>Increases complexity and cost for ETL</li> <li>Introduces data redundancy, inconsistency and fragmentation</li> </ul>	<ul style="list-style-type: none"> <li>Limited scalability</li> <li>Moves complexity to customer</li> <li>Source system knowledge required</li> <li>Long term maintenance costs are high</li> <li>Silo-ed Data</li> <li>Data cataloguing required</li> </ul>	<ul style="list-style-type: none"> <li>Training required</li> <li>High upfront effort</li> <li>Performance issues</li> <li>Specialist modelling skills required</li> <li>Technical knowledge dependency</li> <li>No Data Quality checking</li> <li>Data Cataloguing required</li> </ul>	<ul style="list-style-type: none"> <li>Training required</li> </ul> 

# A proven solution that works

CryspiQ<sup>®</sup> is currently available as SAAS product that:

1. Prepares and organises your data,
2. Captures only valuable and useful data,
3. Enforces business context to collected data,
4. Automates data quality management,
5. Uses natural language to query the data,
6. Simplifies data protection and security,
7. Provides static foundation for your Enterprise AI.



# Any Questions?