






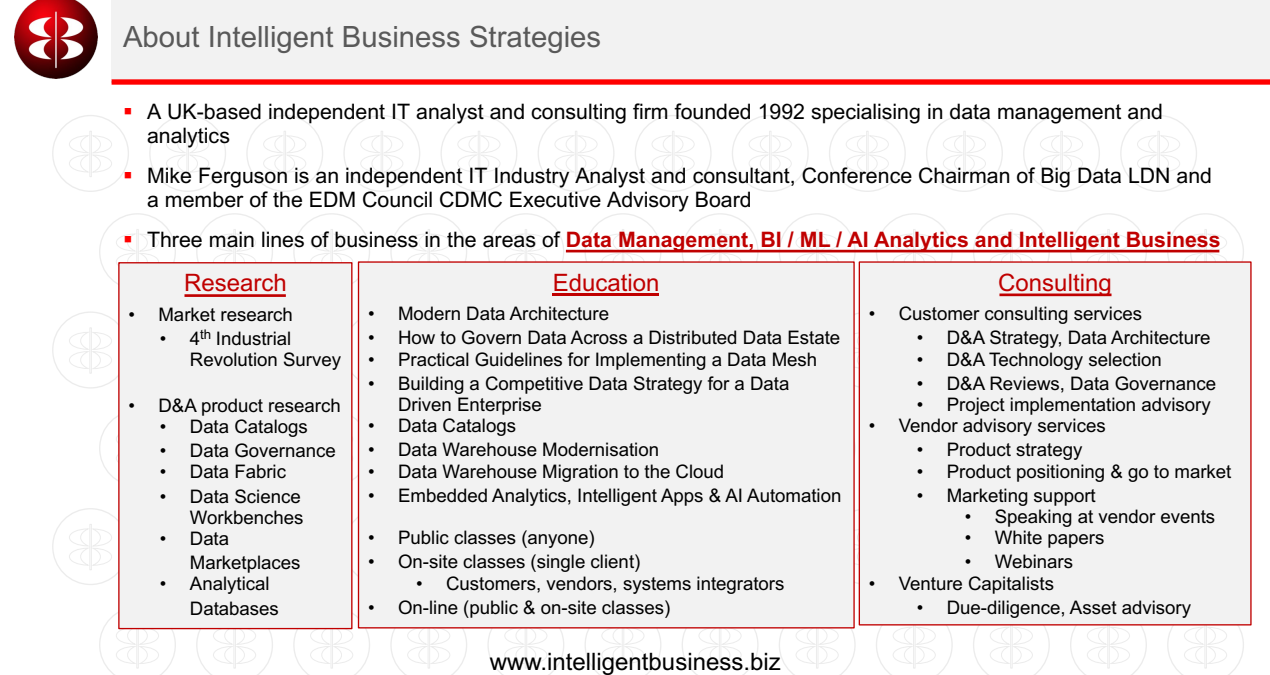
Data Architecture Evolution – The Impact on Analytics Data Platforms


Mike Ferguson
CEO, Intelligent Business Strategies

Adept Events Data Warehousing & BI Summit
Utrecht, March 2024



X @mikeferguson1





About Intelligent Business Strategies

- A UK-based independent IT analyst and consulting firm founded 1992 specialising in data management and analytics
- Mike Ferguson is an independent IT Industry Analyst and consultant, Conference Chairman of Big Data LDN and a member of the EDM Council CDMC Executive Advisory Board
- Three main lines of business in the areas of **Data Management, BI / ML / AI Analytics and Intelligent Business**

<u>Research</u>	<u>Education</u>	<u>Consulting</u>
<ul style="list-style-type: none"> • Market research <ul style="list-style-type: none"> • 4th Industrial Revolution Survey • D&A product research <ul style="list-style-type: none"> • Data Catalogs • Data Governance • Data Fabric • Data Science Workbenches • Data Marketplaces • Analytical Databases 	<ul style="list-style-type: none"> • Modern Data Architecture • How to Govern Data Across a Distributed Data Estate • Practical Guidelines for Implementing a Data Mesh • Building a Competitive Data Strategy for a Data Driven Enterprise • Data Catalogs • Data Warehouse Modernisation • Data Warehouse Migration to the Cloud • Embedded Analytics, Intelligent Apps & AI Automation • Public classes (anyone) • On-site classes (single client) <ul style="list-style-type: none"> • Customers, vendors, systems integrators • On-line (public & on-site classes) 	<ul style="list-style-type: none"> • Customer consulting services <ul style="list-style-type: none"> • D&A Strategy, Data Architecture • D&A Technology selection • D&A Reviews, Data Governance • Project implementation advisory • Vendor advisory services <ul style="list-style-type: none"> • Product strategy • Product positioning & go to market • Marketing support <ul style="list-style-type: none"> • Speaking at vendor events • White papers • Webinars • Venture Capitalists <ul style="list-style-type: none"> • Due-diligence, Asset advisory

www.intelligentbusiness.biz

Copyright © Intelligent Business Strategies 1992-2024



Copyright © 1992-2024 Intelligent Business Strategies Limited. All rights reserved.

All materials pertaining to this presentation are the intellectual property and copyright of Intelligent Business Strategies Ltd and cannot be reproduced or distributed in any form without prior written permission from Intelligent Business Strategies Ltd. In addition, they must not be used by any party for hire and/or to compete against Intelligent Business Strategies Ltd

Copyright © Intelligent Business Strategies 1992-2024

3



Topics

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
- The challenge of a distributed data estate
- Data fabric and how can they help build data products
- Data architecture options for building data products
- The impact of open table formats and query language extensions on architecture modernisation
- Conclusions

Copyright © Intelligent Business Strategies 1992-2024

4



Topics – Where Are We?

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
 - The challenge of a distributed data estate
 - Data fabric and how can they help build data products
 - Data architecture options for building data products
 - The impact of open table formats and query language extensions on architecture modernisation
- Conclusions

Copyright © Intelligent Business Strategies 1992-2024

5

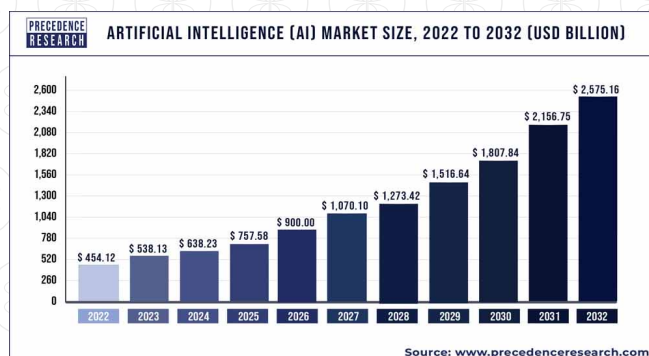


Demand Is Increasing For Data And AI

State of Data & Analytics Investment and Business Results

	2024
Investments in Data & Analytics are a Top Organizational Priority	87.9%
Our Organization is Increasing its Investment in Data & Analytics	82.2%
Delivering Measurable Business Value from Data & Analytics Investments	87.0%
Strong Business Leadership and Partnership In Place at Our Organization	84.3%

Source: Data And AI Leadership Executive Survey 2024 – New Vantage Partners (A Wavestone Company)



6

Executive Expectations Of Data And AI Are Huge But Maximum Return On Investment Will Not Happen Unless There Is A Solid Foundation Of High-Quality, Reusable Data

DATA & AI

The Data and AI Driven Enterprise

Disruption Transformation

Speed & Agility Competitive Advantage

High Quality & Governed Data Foundation

7

Many Companies Do Not Have A Data Foundation
 – They Have Multiple Siloed Analytical Systems With Data Redundancy

GraphDB

DW mart mart

DW mart mart

Hadoop cluster Hadoop cluster

DW mart mart

Data Lake (cloud storage)

Data Lake (cloud storage)

Cloud Lakehouse

Bronze tables (Raw data)

Silver tables

Gold tables

Feature store tables

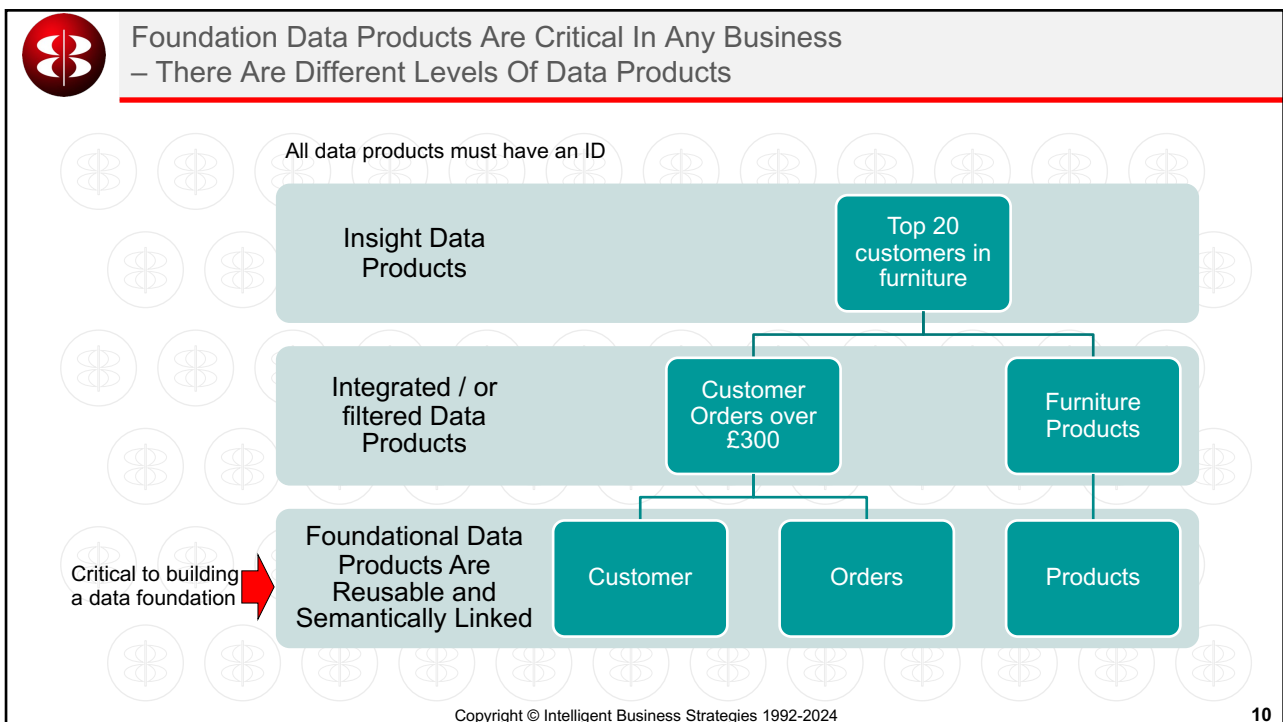
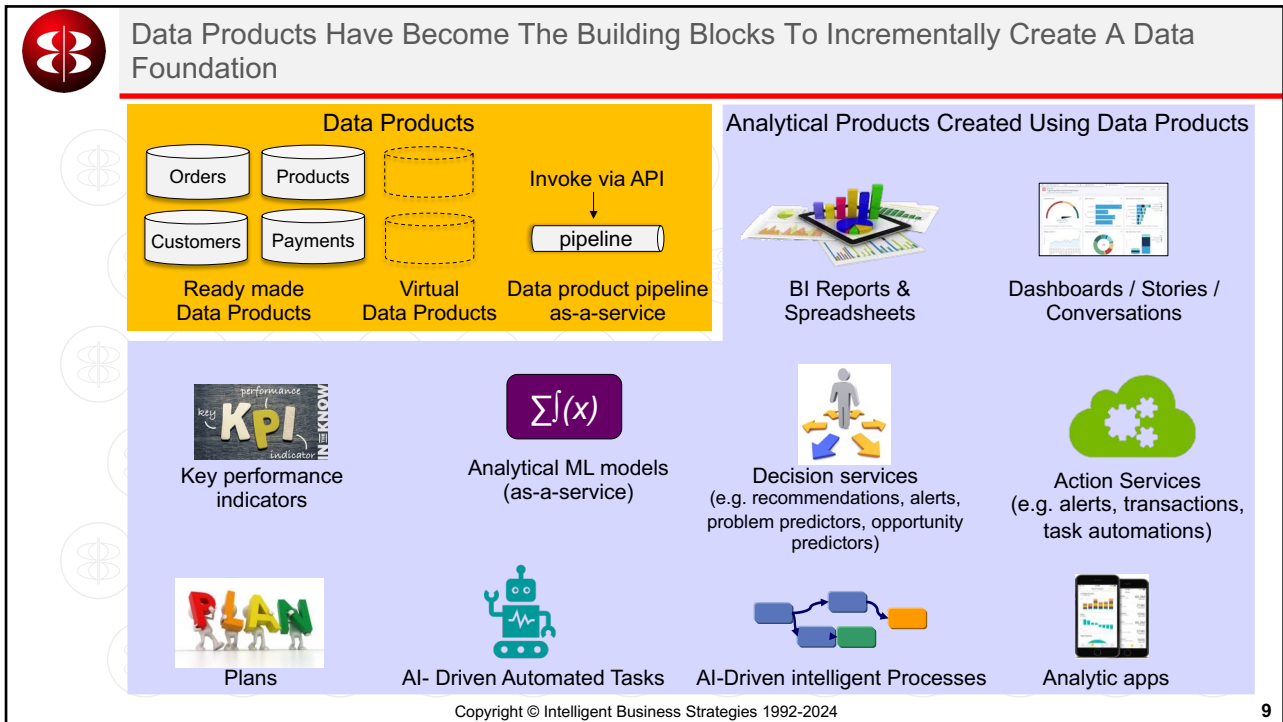
DW / data mart tables

Lakehouse

Data Lake (cloud storage)

Copyright © Intelligent Business Strategies 1992-2024

8

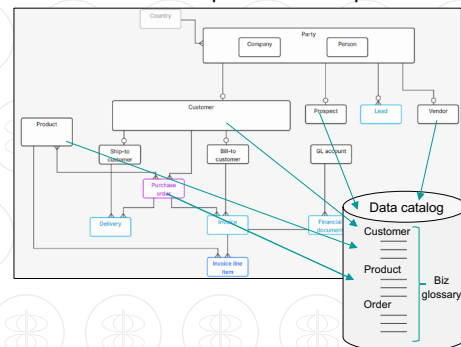




Steps To Creating For Data Products – From Data Concept Model To Data Marketplace

1. Identify the data concepts, properties and relationships and construct a data concept model
2. The data concepts become the 'skeleton data entities' in the common vocabulary
3. Each data concept and its attributes should be defined in the business glossary as a data entity with a data owner
4. Use the catalog to discover the data for each data entity in underlying data stores across the data landscape
5. Design DataOps component-based pipelines to create the data products with common vocabulary data names
6. Publish all data products in a data marketplace

Data Concept Model Example

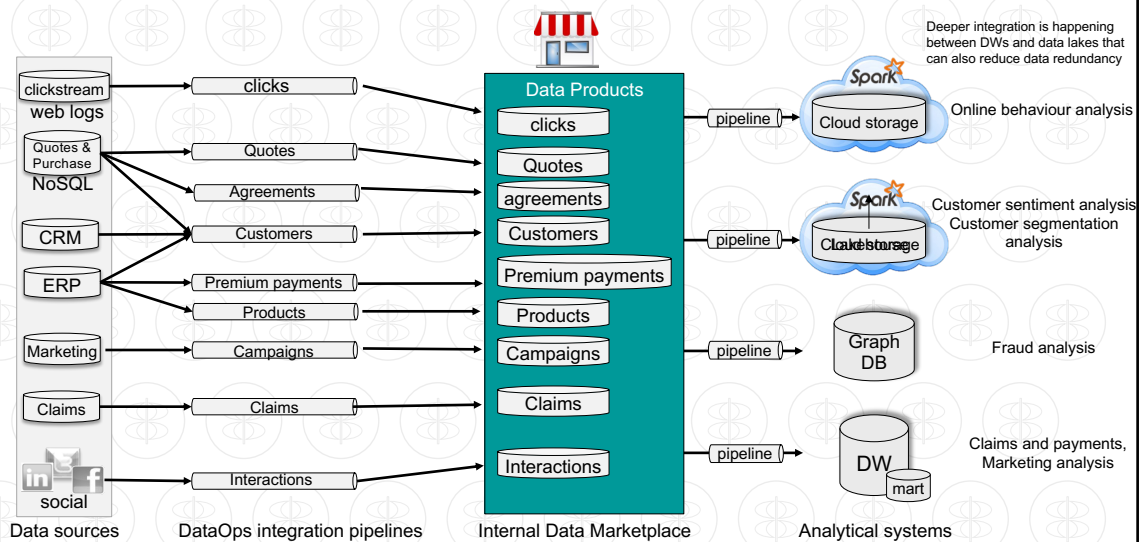


Copyright © Intelligent Business Strategies 1992-2024

11

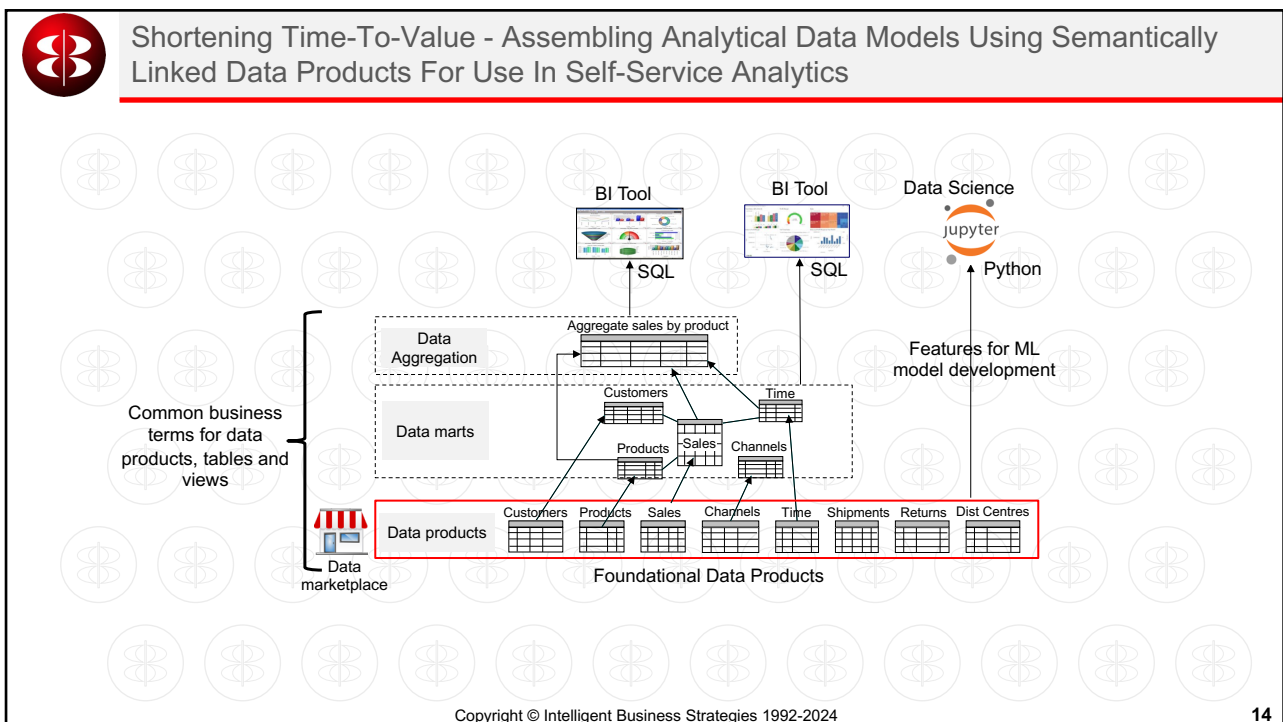
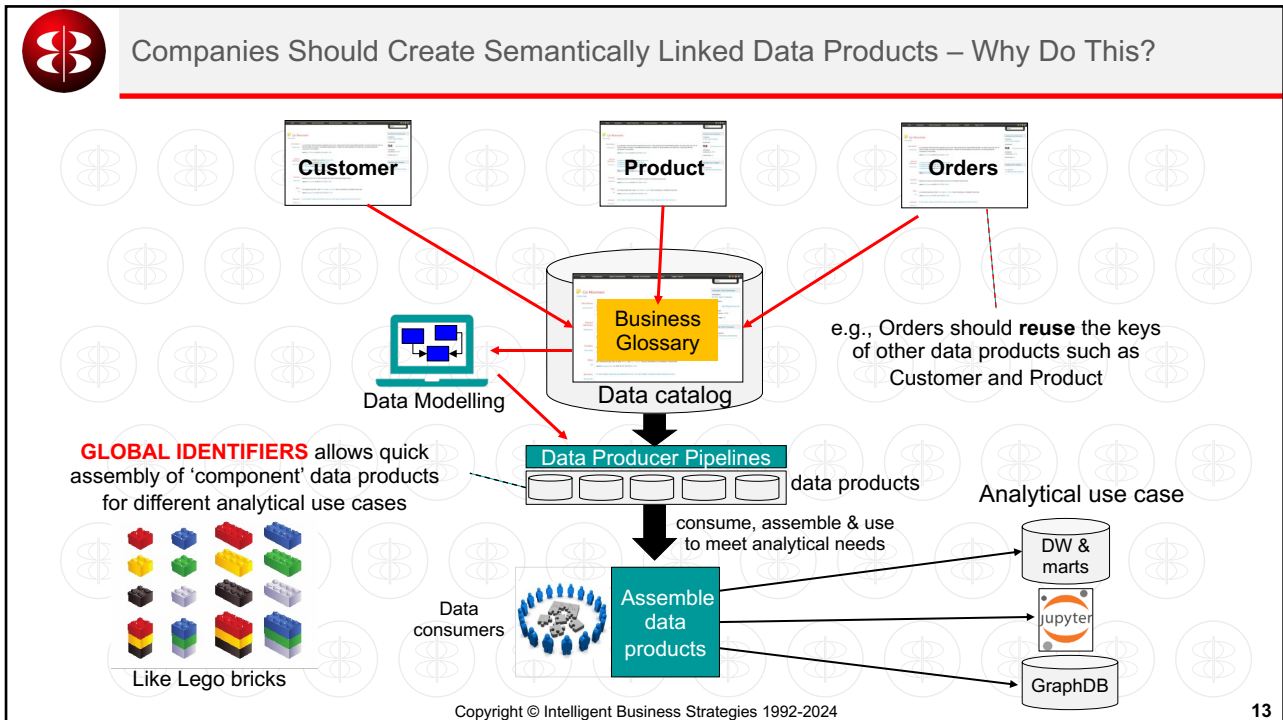


A Data Foundation Includes Reusable 'Business Ready' Data Products That Are Semantically Linked, Built Once And Reused Everywhere For Different Analytical Workloads



Copyright © Intelligent Business Strategies 1992-2024

12



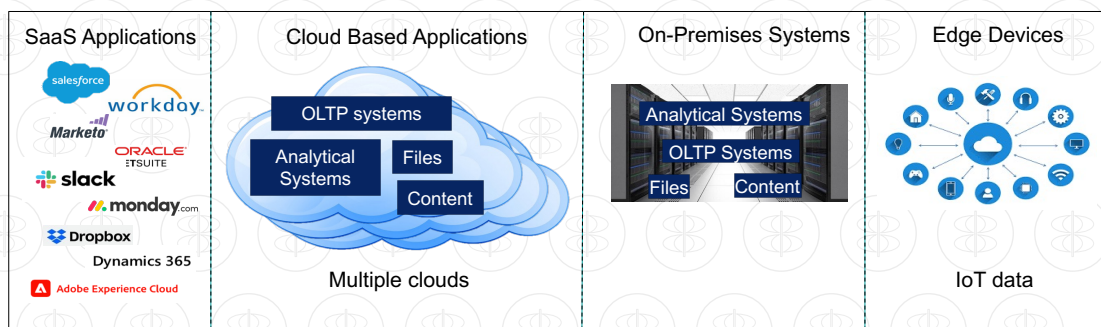


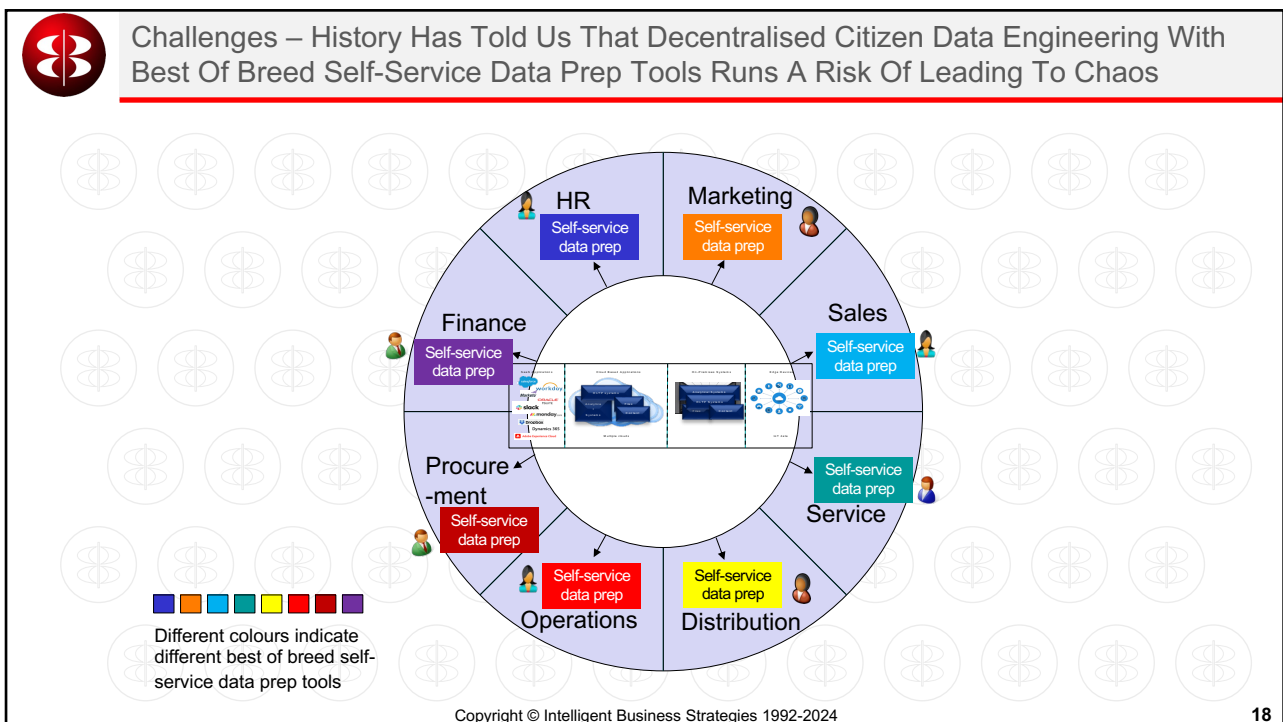
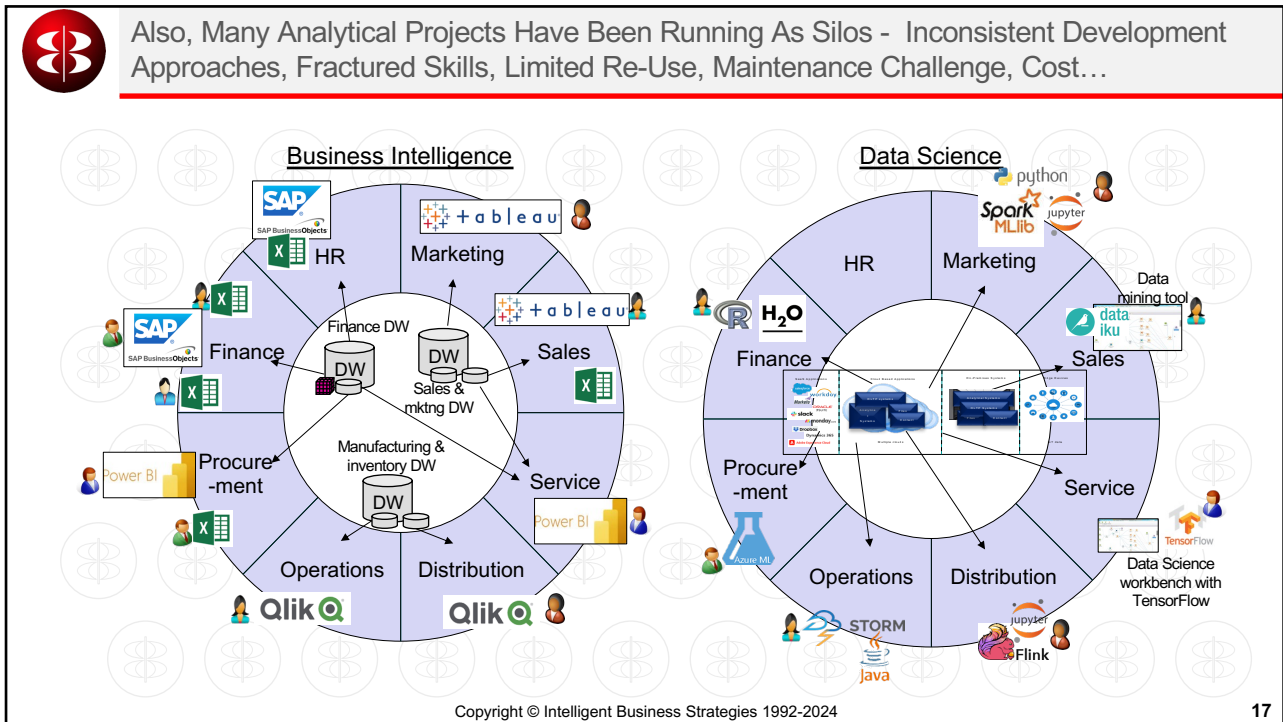
Topics – Where Are We?

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
- The challenge of a distributed data estate
- Data Fabric and how can they help build data products
- Data architecture options for building data products
- The impact of open table formats and query language extensions on architecture modernisation
- Conclusions



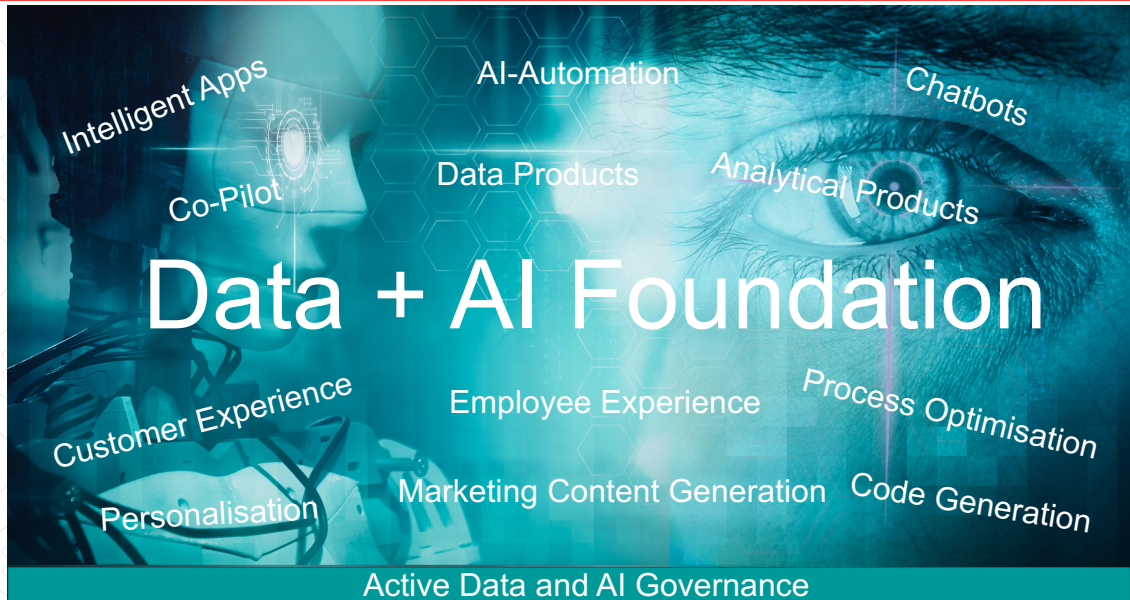
Challenges When Creating A Data Foundation - Many Companies Today Have Data Housed In Multiple Data Stores Across A Hybrid, Multi-Cloud Distributed Data Estate







We Are Now In A New Era - Companies Are Beginning To Put Data And AI First

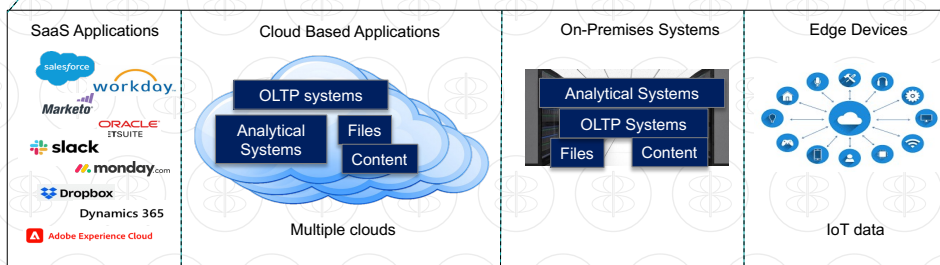
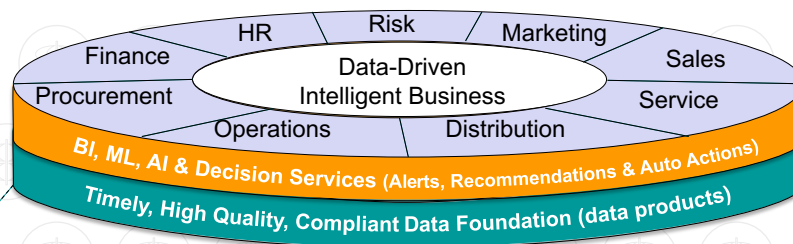


Copyright © Intelligent Business Strategies 1992-2024

19



The Challenge Of Data Complexity – How Do You Build A Data Foundation Providing Timely, High Quality, Compliant Data Products On A Highly Distributed Data Estate?



Data is spread across a hybrid distributed data landscape and stored in files, Relational DBMSs, NoSQL DBMSs, Cloud storage, Hadoop systems, SaaS applications and Content stores

Copyright © Intelligent Business Strategies 1992-2024

20

The Challenge Of Data Complexity – How Do You Build A Data And AI Foundation Providing Timely, High Quality, Compliant Data And Analytical Products

Data is spread across a hybrid distributed data landscape and stored in files, Relational DBMSs, NoSQL DBMSs, Cloud storage, Hadoop systems, SaaS applications and Content stores

Copyright © Intelligent Business Strategies 1992-2024

21

Trends – Companies Want A Common Data And Analytics Software Platform With Integrated Tools And Shared Metadata To Accelerate Development

Build data products, analytical products and AI-driven applications faster with end-to-end governance

Single Vendor Data Fabric Vs **The Modern Data Stack**

- One platform with integrated services and shared metadata
 - Manage, build and deploy data and analytical products
 - Govern data and AI models across the enterprise
 - Bundled data catalog or 3rd party catalog integration
- Some vendors also bundle integrated BI / ML and AI tooling
- Attractive because of no standard to share metadata between tools

- A D&A stack made up of complementary best-of-breed tools from multiple vendors who have partnered
- Pre-built integrations across the stack
- Formed to compete against single vendor data fabric
- Several variations of the modern data stack have appeared on the market, confusing prospective buyers who are unsure on what tools are integrated

Copyright © Intelligent Business Strategies 1992-2024

22

Creating Reusable Data Products Is An Approach Fast Gaining Momentum As A Way of Incrementally Building Up A Secure And Compliant Data Foundation

Copyright © Intelligent Business Strategies 1992-2024 23

Topics – Where Are We?

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
- The challenge of a distributed data estate
- Data Fabric and how can they help build data products
- Data architecture options for building data products
- The impact of open table formats and query language extensions on architecture modernisation
- Conclusions

Copyright © Intelligent Business Strategies 1992-2024 24

Multiple Architecture Options Have Emerged For Decentralised Creation Of Data Products?

Multiple approaches to creating data products

1. One storage account, multiple zones on cloud storage
2. Multiple zoned storage accounts on a single cloud
3. Multiple zoned storage accounts on multiple clouds
4. One or multiple zoned lakehouses (one per domain)
5. Create data products in a data warehouse staging tables
6. Leave data where it is and produce virtual data products using data virtualisation / federated queries
7. Create data products using Kafka topics

Data products need to be semantically linked (by using enterprise wide primary and foreign keys)

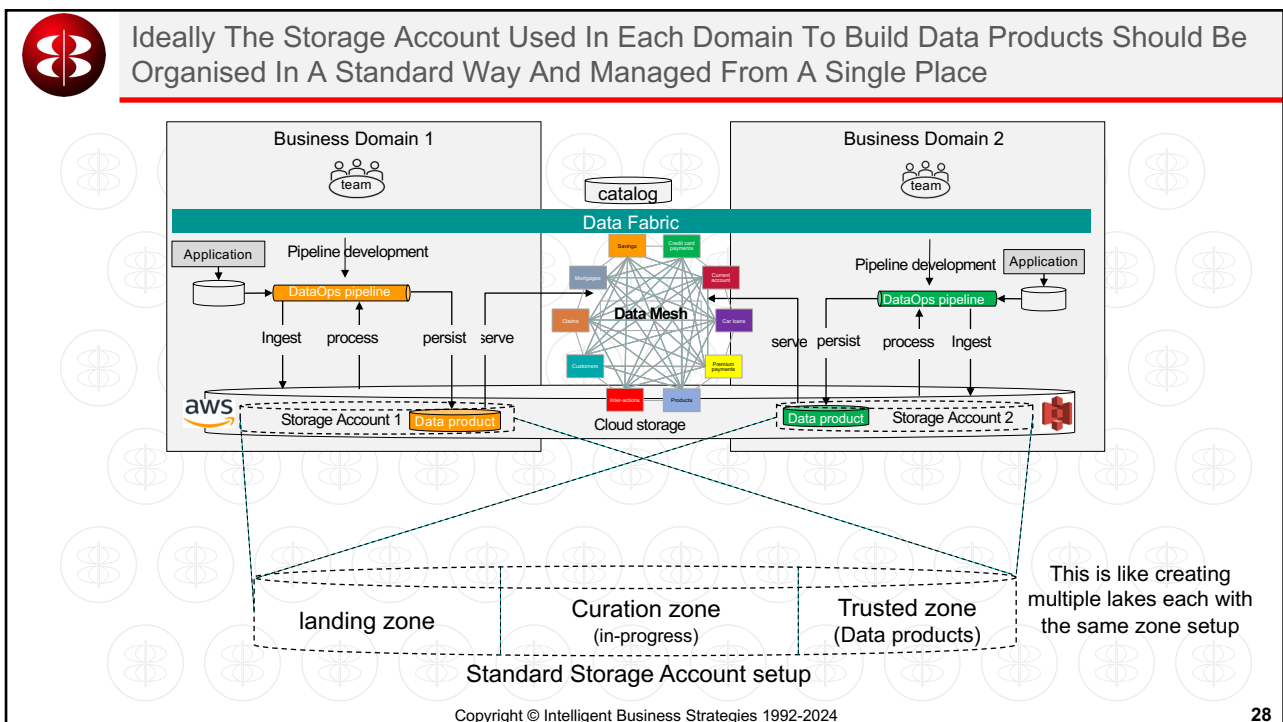
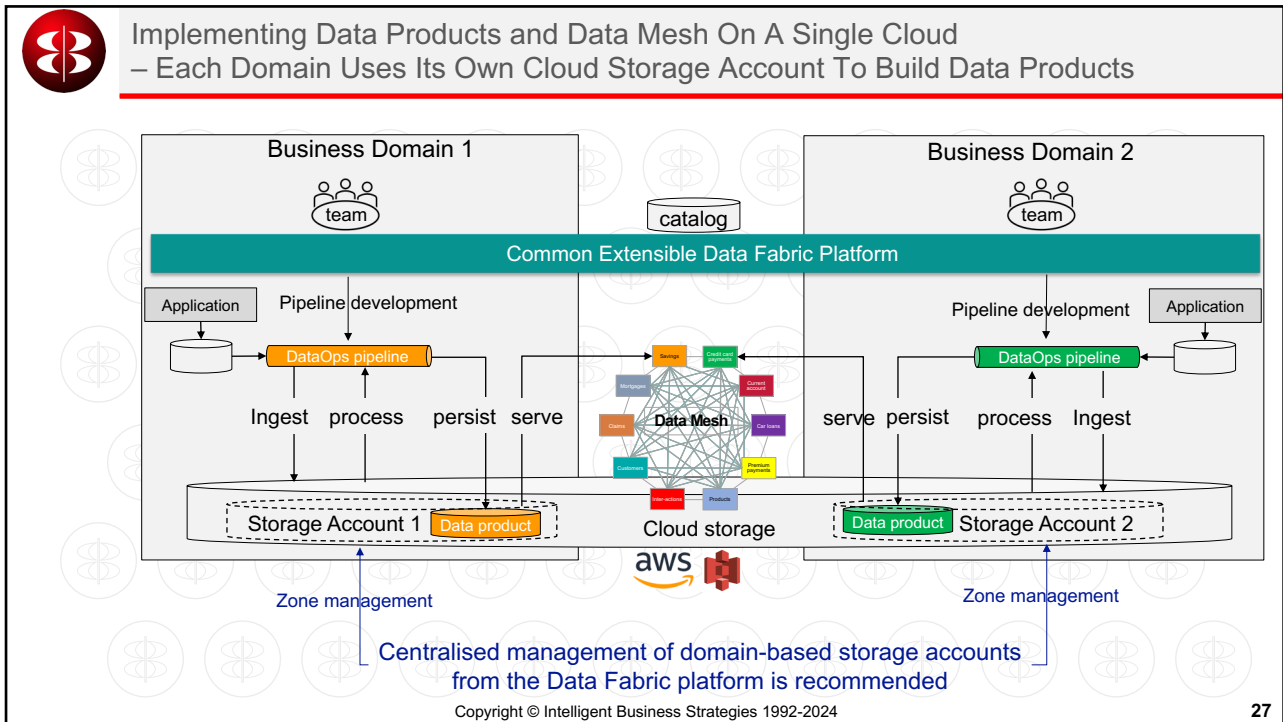
Copyright © Intelligent Business Strategies 1992-2024 25

Modern Data Architecture – In 2022 Multiple Competing Data Architecture Patterns Emerged As Options To Create A Data Foundation – Zoned Cloud Data Lake

- **Characteristics**
 - Custom architecture
 - Vendor agnostic
 - Requires organised storage and governance
 - using data fabric software and a data catalog
- **Issues**
 - No ACID properties on data lake files so data could be inconsistent if updated by multiple engines
 - No schema governance or change control
 - Data modelling often ignored by data scientists
 - Data redundancy can proliferate if unguarded
 - Can deteriorate if storage is unguarded and data products are not well defined

Custom built

Copyright © Intelligent Business Strategies 1992-2024 26





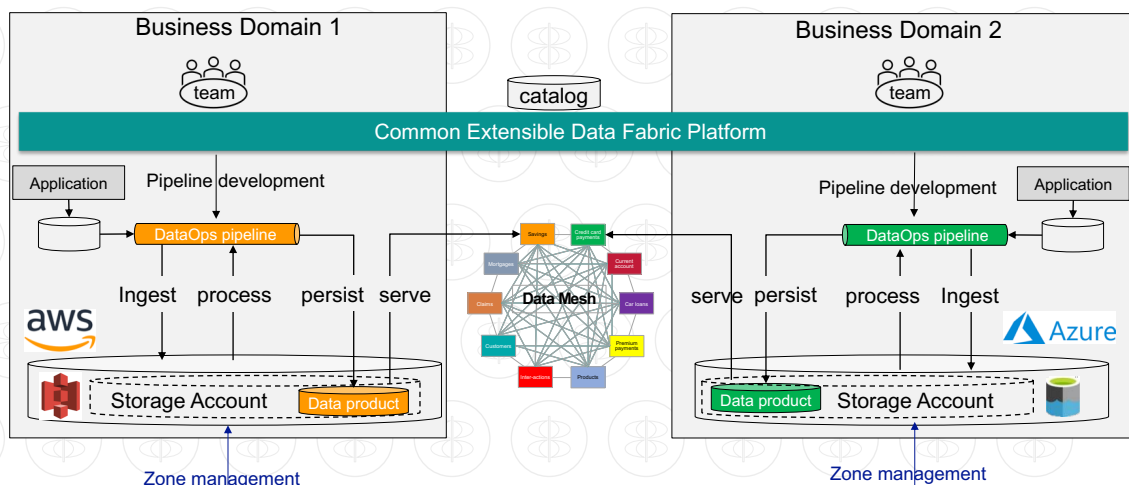
Storage Accounts Can Be Managed As Multiple Lakes With Zones In A Decentralised Environment Using Data Fabric Software – E.g. Google Dataplex

Display name	Type	Status	Assets	Last modified	Label
Landing Zone	Raw	Active	1342	May 26, 2021, 2:34:55 PM	ingestiontemplate: c85ad-secondary...
Raw Zone	Raw	Active	1	May 26, 2021, 10:44:12 AM	
Data Analytics Zone	Curated	Active	2	May 26, 2021, 10:44:12 AM	
Data Science Zone	Curated	Active	1	May 26, 2021, 10:44:12 AM	

You can create a lake per business domain within your organisation and create data zones that map to data readiness and usage (landing zone, raw data zone, data products zone, data science zone, etc.)



Implementing A Data Mesh Across Multiple Clouds Should Also Be Possible – Each Domain Uses Its Own Scalable Polyglot Storage Managed From Data Fabric

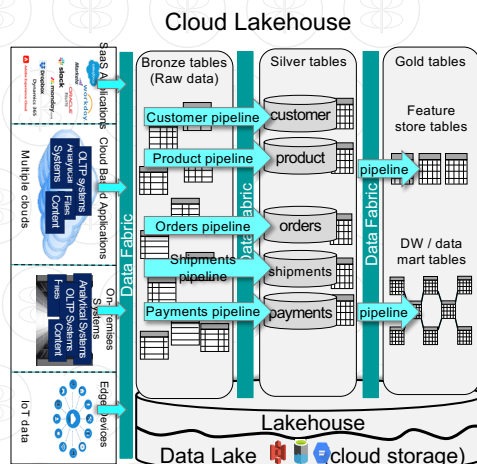


Centralised management of domain-based storage accounts from the Data Fabric platform is recommended



Modern Data Architecture – In 2022 Multiple Competing Data Architecture Patterns Emerged As Options To Create A Data Foundation – Lakehouse

- **Characteristics**
 - Centralised architecture governed by a data catalog
 - Can also be a distributed architecture with multiple lakehouses
 - Data is stored in open table format
 - ACID properties guarantees consistency
 - Data has to be ingested before processing
 - Data engineering using platform specific software or 3rd party data fabric
- **Issues**
 - Could end up building silos on a single platform
 - Exploratory storage Vs auditability
 - Is data engineering done before BI and data science?
 - Are BI tools accessing raw data or engineered data?
 - Which data modelling design technique is used?
- **Vendor example:**
 - Databrick Lakehouse platform with Unity Catalog

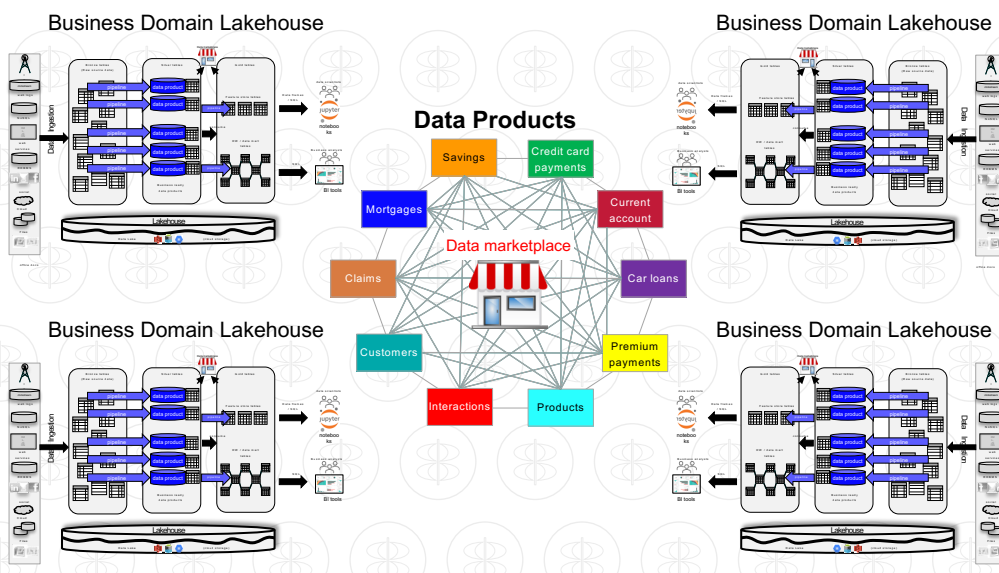


Copyright © Intelligent Business Strategies 1992-2024

31



Distributed Options For Decentralised Data Product Development On A Lakehouse – E.g., Multiple Domain-Oriented Lakehouses All On A Single Cloud



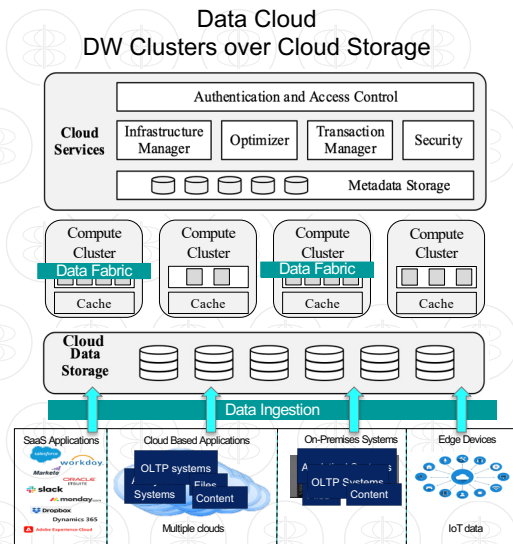
Copyright © Intelligent Business Strategies 1992-2024

32



Modern Data Architecture – In 2022 Multiple Competing Data Architecture Patterns Emerged As Options To Create A Data Foundation – Data Cloud

- **Characteristics**
 - Centralised architecture governed by the DBMS
 - Data typically ingested before processing
 - Data stored in proprietary table formats
 - Supports open table formats and federated query
 - Storage separated from compute
 - Workload separation via different compute clusters
 - Data engineering using 3rd party data fabric
- **Issues**
 - Primarily used for data warehouse workloads
 - Unstructured data mainly processed outside the DBMS
 - SQL on everything requires invoking Deep Learning/AI
- **Vendor example:**
 - Snowflake Data Cloud



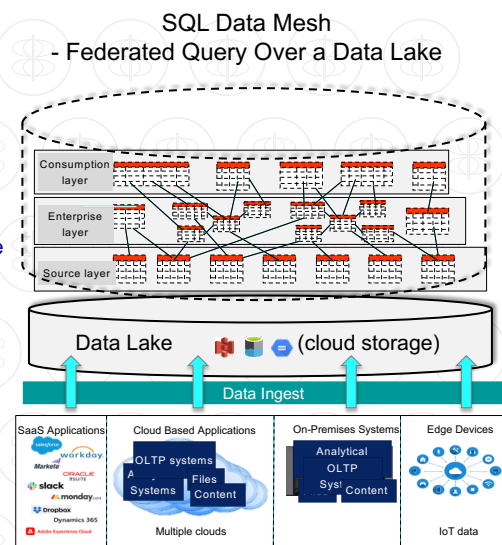
Copyright © Intelligent Business Strategies 1992-2024

33



Modern Data Architecture - In 2022 Multiple Competing Data Architecture Patterns Have Emerged As Options To Create A Data Foundation – SQL Data Mesh

- **Characteristics**
 - Centralised architecture
 - Data typically ingested before data engineering
 - Data stored in files e.g., Parquet, ORC
 - Supports federated query beyond cloud storage
 - Data engineering using data virtualisation
 - Virtual data products
 - Metadata specifications to process data are all in one place
- **Issues**
 - Ungoverned files in cloud storage could create a swamp
 - SQL on everything - unstructured data is a challenge
 - Data modelling can be overlooked
 - What about data cleansing?
 - Performance of file access on cloud storage
 - Is there a parallel engine to compensate?
- **Vendor example:**
 - Starburst, Denodo (supports parallel engine)



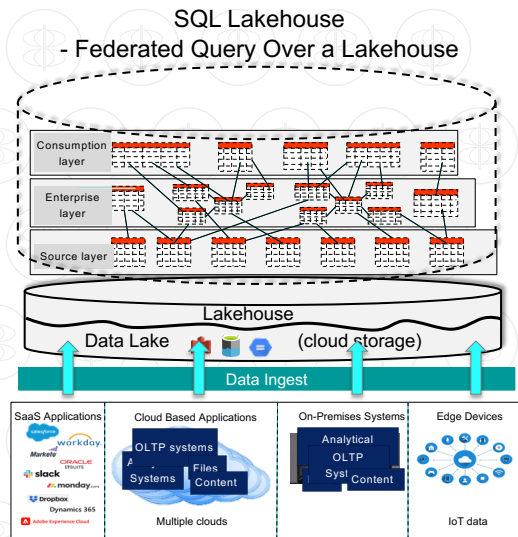
Copyright © Intelligent Business Strategies 1992-2024

34



Modern Data Architecture - In 2022 Multiple Competing Data Architecture Patterns Have Emerged As Options To Create A Data Foundation – SQL Lakehouse

- Characteristics
 - Centralised architecture
 - Data ingested before processing
 - Open table format e.g. Apache Iceberg, Delta Lake
 - ACID support - multiple engines can access and update
 - Schema governance and data partitioning
 - Federated query beyond the lakehouse
 - Data engineering on ingest and using data virtualisation
 - Virtual data products
 - Metadata specifications all in one place
 - Universal semantic layer
 - Universal data access control
- Issues
 - Unstructured data is difficult to deal with
 - Data modelling can be overlooked
 - SQL engine performance for concurrent BI users
- Vendor example:
 - Dremio (Read/write only on Iceberg, read on Delta)
 - Arctic catalog of Iceberg Tables



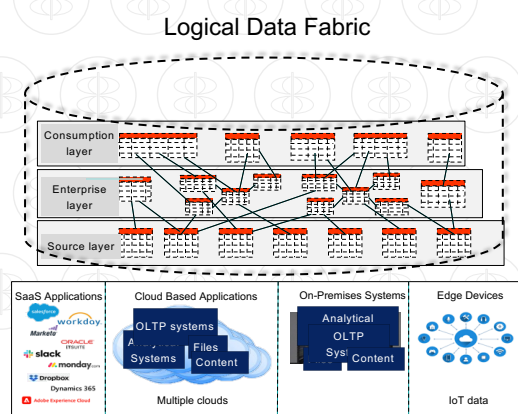
Copyright © Intelligent Business Strategies 1992-2024

35



Modern Data Architecture - In 2022 Multiple Competing Data Architecture Patterns Have Emerged As Options To Create A Data Foundation – Logical Data Fabric

- Characteristics
 - Leave data where it is
 - Federated query architecture to multiple data sources including open table formats
 - Data engineering using data virtualisation
 - Query with pushdown and ML driven caching
 - Parallel engine on cloud storage for performance
 - Metadata specifications all in one place
 - Virtual view materialisation and persistence to data warehouse or lakehouse
 - Virtual and physical data products
 - Universal semantic layer
 - Universal data access control
- Issues
 - SQL on everything - unstructured data is a challenge
 - Data cleansing
- Vendor example:
 - Denodo, Data Virtuality



Copyright © Intelligent Business Strategies 1992-2024

36

Modern Data Architecture – Multiple Competing Architectures Have Confused The Market

custom built zoned data lake feeding multiple analytical data stores
 Data science (Graph DB, Jupyter, MDM) feeds into cloud storage (raw, in-progress, ready to go). **Centralised**

Organised Lakehouse
 Bronze tables, Silver tables, Gold tables, Feature store tables, DW / data mart tables. **Centralised**

Data cloud
 Authentication and Access Control, Cloud Services (Infrastructure Manager, Optimizer, Security), Cloud Storage, Compute Clusters, Caches. **Centralised**

Virtual data products
 Data Mesh (virtual views on data in a data lake) **Centralised**
 SQL Lakehouse (virtual views on lakehouse open tables) **Centralised**
 Logical data fabric (virtual views on everything) **Federated**

Which architecture? Should it be centralised, distributed or federated?

Copyright © Intelligent Business Strategies 1992-2024

A Popular Data Architecture Option That Emerged In 2022 Was Data Warehouse / Data Lake Integration With External Tables Pointing To Files In A Cloud Storage Data Lake

Extended Data Warehouse DBMS
 External table, External table, External table, External table. **Centralised**

Data Lake (cloud storage)
 Parquet files, Trusted Data Products, Parquet files

Data Fabric

SaaS Applications (Salesforce, Workday, Marketo, Oracle TS Suite, Slack, Monday.com, Dropbox, Dynamics 365, Adobe Experience Cloud)

Cloud Based Applications (OLTP systems, Analytical Systems, Files, Content, Multiple clouds)

On-Premises Systems (Analytical Systems, OLTP Systems, Files, Content)

Edge Devices (IoT data)

Legend:
 Grid icon: DBMS proprietary table format
 Grid icon with red border: DBMS external table

Problems:
 • No ACID properties
 • No Schema governance
 • Scalable concurrent user BI query performance on files in cloud storage

Copyright © Intelligent Business Strategies 1992-2024



Some Vendors Went Further - Google BigQuery Architecture Allow Its Query Engine To Run On Other Clouds – BigQuery Omni

- BigQuery's architecture separates compute from storage and so processing happens where the data is stored
- Query results can be returned to Google Cloud over a secure connection

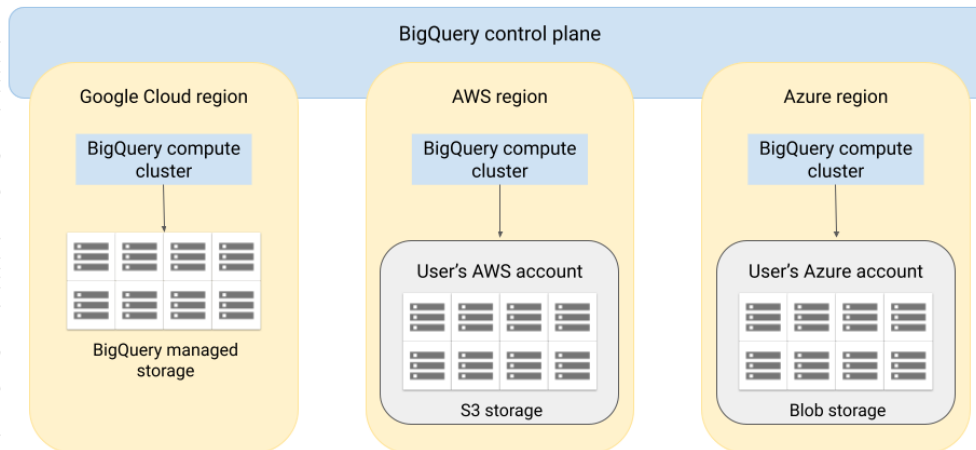


Image source: <https://cloud.google.com/bigquery/docs/omni-introduction>

Copyright © Intelligent Business Strategies 1992-2024

39



Topics – Where Are We?

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
- The challenge of a distributed data estate
- Data Fabric and how can they help build data products
- Data architecture options for building data products
- The impact of open table formats and query language extensions on architecture modernisation
- Conclusions

Copyright © Intelligent Business Strategies 1992-2024

40

Data Architecture Evolution – Impact Of Data Lakehouse And Open Table Formats

- Open table formats
 - Apache Iceberg
 - Apache Hudi
 - Delta Lake
- The lakehouse supports
 - Open tables accessible and updatable by multiple engines
 - Transaction integrity ACID properties for insert / update
 - Schema enforcement and governance
 - Schema evolution, version control and historical time travel
 - Multiple query engines with SQL access to data in the data lake
 - Streaming and batch unification on open tables
 - Data in compressed columnar file format, e.g., Parquet, ORC
 - Hidden data partitioning
 - Integration with Spark for access via other languages e.g. Python
 - Row level security
 - Organised data lake with ELT processing

ACID = Atomicity, Consistency, Durability and Isolation
Ensures that data is current and transactionally consistent

Copyright © Intelligent Business Strategies 1992-2024 41

Modern Data Architecture Continues To Change – Open Table Formats Means Multiple Query Engines Can Update And Access Open Tables, e.g., DW DBMS, Spark, Flink, etc.

Google BigQuery example

```
CREATE EXTERNAL TABLE myexternal-table
WITH CONNECTION `myproject.us.myconnection`
OPTIONS (format = 'ICEBERG',
uris = ["gs://mybucket/mydata/mytable/metadata/iceberg.metadata.json"])
```

Copyright © Intelligent Business Strategies 1992-2024 42



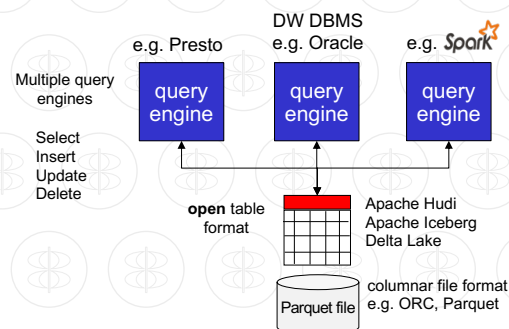
Impact Of Open Table Formats

- Enabling the integration of lakehouses and data warehouses as opposed to them being alternatives
- ACID transaction with a transaction log supported which is critical for data consistency
 - Offers access by multiple engines with full data integrity
- Better performance of SQL queries
 - Support for hidden data partitioning, data skipping etc.
- Schema change on immutable files
- Time travel for historical queries
 - Transaction log in Delta Lake, Snapshots in Apache Iceberg
- Enables data lakes to:
 - Be used for more than just exploratory analysis
 - Be governed, secured and managed
 - Provide consistent data with data integrity in support of data science and DW/BI workloads
 - Support training and retraining of ML models on consistent data



In 2023 Open Table Formats Have Gained Significant Momentum With Broad Adoption By DBMS Vendors

Open Table Format	Supported By
Apache Hudi	<ul style="list-style-type: none"> • AWS Athena • Onehouse • OpenText Vertica
Apache Iceberg	<ul style="list-style-type: none"> • AWS Athena • Cloudera • Dremio • Flink • Google • IBM • Oracle • Presto • Snowflake
Linux Foundation Delta Lake	<ul style="list-style-type: none"> • Databricks • Microsoft Fabric • SAP DataSphere • Teradata



- Data Lake 3.0 Universal Format
 - Read Delta tables in Delta, Hudi, & Iceberg formats
 - Auto generates metadata for Apache Iceberg or Apache Hudi

Apache Iceberg Table Support In Snowflake (Announced August 2022)

Image source: <https://www.snowflake.com/blog/iceberg-tables-powering-open-standards-with-snowflake-innovations/>

Iceberg tables work like Snowflake native tables with 3 key differences:

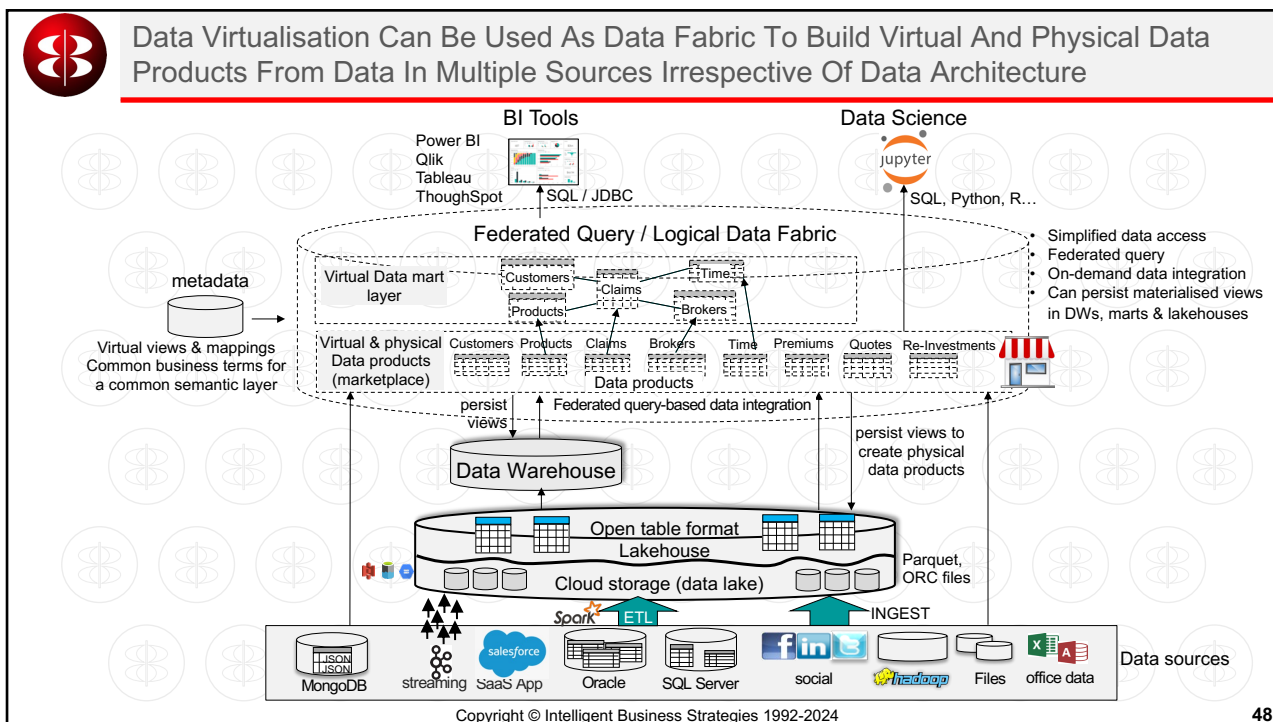
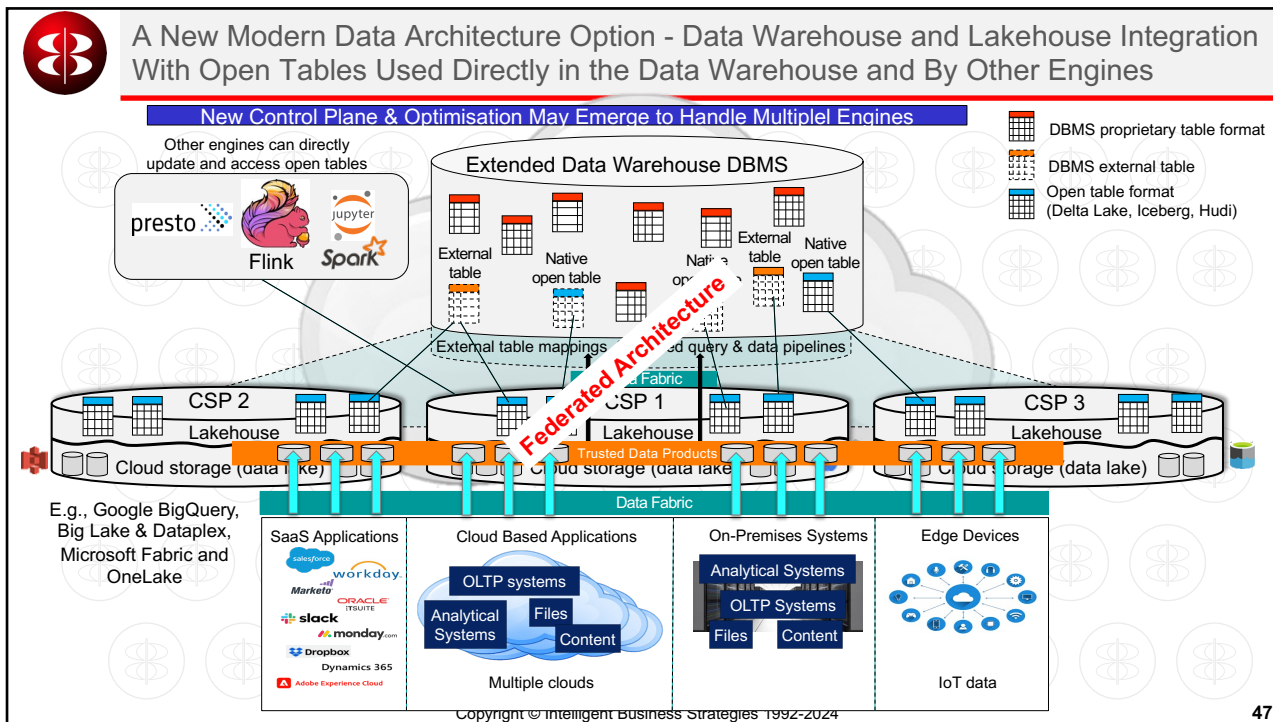
1. Table metadata is in Iceberg format
2. Data is stored in Parquet files
3. Both table metadata and data is stored in customer-supplied storage (customer managed)

Copyright © Intelligent Business Strategies 1992-2024 45

Using Iceberg Tables In Snowflake – Can Use External Tables Or Native Iceberg Tables

Table Type	Code Example
External Table	<pre>create external table <table> table_format = iceberg file_format = parquet refresh_on_create = false auto_refresh = false snapshot_location = @mystage/file.json</pre>
Native Iceberg table	<pre>-- Create an External Volume to hold Parquet and Iceberg data create or replace external volume my_ext_vol STORAGE_LOCATIONS = ((NAME = 'my-s3-us-east-1' STORAGE_PROVIDER = 'S3' STORAGE_BASE_URL = 's3://my-s3-bucket/data/snowflake_extvol/' STORAGE_AWS_ROLE_ARN = '****')); -- Create an Iceberg Table using my External Volume create or replace iceberg table my_iceberg_table with EXTERNAL_VOLUME = 'my_ext_vol' as select id, date, first_name, last_name, address, region, order_number, invoice_amount from sales;</pre>

Copyright © Intelligent Business Strategies 1992-2024 46





Open Table Formats - Critical Questions

Question	Answer
Can you swap query engines?	Not yet as not all are read/write plus some restrict write operations
Will SQL differences remain?	Yes
Will DBMS vendors try to prevent other engines accessing open table data by adding proprietary extensions?	<ul style="list-style-type: none"> I hope not! – so far so good Snowflake setting a precedent by not attempting to add their own proprietary micro-partitioning technique?
Will DBMS vendors keep their own metadata proprietary in addition to basic open table metadata?	Yes
Will metadata for open tables remain separate from metadata for proprietary tables	Yes, for now but pressure will emerge to create an industry standard to synchronise open table metadata into DBMS query engine catalogs

- The battle of the query engines will become a more visible in the next 12-18 months
- Open table formats will not be the differentiator because they will all be very similar by Q1 2025
- It is the ecosystem of vendor and partner tools around the platform that be the differential
- It is likely some vendors will roll out a single control plane to manage multiple engines

Copyright © Intelligent Business Strategies 1992-2024

49



In-Database Machine Learning Has Been Available For Years But Now On Open Tables And Multiple Clouds - E.g., Google BigQuery ML Integrates With Google BigLake And Dataplex

- A set of SQL extensions to build and deploy ML models using data stored in Google BigQuery
 - Abstracts development of ML models into a simple SQL syntax
 - Aimed at making ML more accessible to SQL developers
 - Limited to Linear and Binary Logistic regression only

Create a model

```
CREATE MODEL dataset.model_name
OPTIONS(model_type='linear_reg', input_label_cols=['input_label'])
AS SELECT * FROM input_table;
```

Train a model

```
SELECT * FROM ML.TRAINING_INFO(MODEL `my model`)
```

Evaluate a model

```
WITH eval_table AS ( SELECT *, label FROM `my dataset` )
SELECT * FROM ML.EVALUATE(MODEL `my model`, TABLE eval_table)
```

Execute model to get predictions

```
SELECT game_id, predicted_label
FROM ML.PREDICT(MODEL `my model`, table dataset_to_predict) AS predict
```

Source: <https://www.linkedin.com/pulse/machine-learning-your-database-case-against-bigquery-ml-rodriguez/>


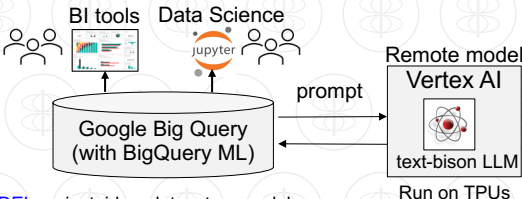
Copyright © Intelligent Business Strategies 1992-2024

50

Generative AI In The Database Can Access Open Tables Across Clouds – Google Duet AI In BigQuery, BigQuery ML Support For LLMs Used in SQL Queries Across BigLake

Duet AI in BigQuery enables you to:

- **Generate** a SQL query
- **Complete** a SQL query
- **Explain** a SQL query

Uses

- Classification
- Sentiment Analysis
- Entity extraction
- Extractive Question Answering
- Summarization
- Re-writing text in a different style
- Ad copy generation
- Concept ideation

```

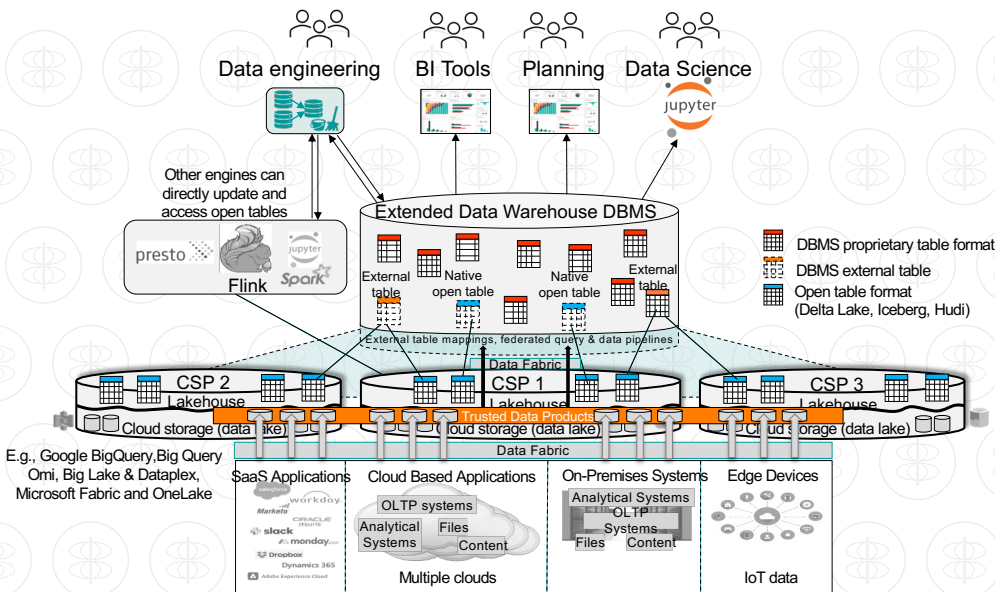
CREATE MODEL project_id.mydataset.mymodel
REMOTE WITH CONNECTION `myproject.us.test_connection`
OPTIONS(REMOTE_SERVICE_TYPE="CLOUD_AI_LARGE_LANGUAGE_MODEL_V1")

SELECT * FROM ML.GENERATE_TEXT( MODEL mydataset.llm_model TABLE mydataset.prompt_table,
STRUCT( 0.2 AS temperature, 75 AS max_output_tokens, 0.3 AS top_p, 15 AS top_k, TRUE AS flatten_json_output));
    
```

- Provides prompt data from a table column that's named prompt
- Returns a shorter generated text response
- Returns a more probable generated text response
- Flattens the JSON response into separate columns


51

Architecture Evolution – A New Opportunity To Combine Multiple Types of Analytics Across Multiple Clouds and On-Premises IF We Create and Share Data Products



Copyright © Intelligent Business Strategies 1992-2024

52



Data Architecture Evolution

– New Capabilities In SQL – ISO SQL/PGQ (June 2023)

Annual Meeting 2023
Applications OBP English

ISO Standards About us News Taking part Store
Search

← ICS ← 35 ← 35.060

ISO/IEC 9075-16:2023

Information technology — Database languages SQL — Part 16: Property Graph Queries (SQL/PGQ)

General information

Status : Published Publication date : 2023-06


Edition : 1 Number of pages : 269

Technical Committee : ISO/IEC JTC 1/SC 32 Data management and interchange

ICS : 35.060 Languages used in information technology

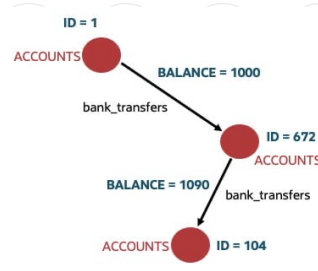
Preview

Copyright © Intelligent Business Strategies 1992-2024
53



SQL/PGQ Is Already Available In Some Database Management Systems

– Oracle 23c Release (September 2023)



```

graph TD
    V1((ID = 1  
ACCOUNTS)) -- "BALANCE = 1000  
bank_transfers" --> V2((ID = 672  
ACCOUNTS))
    V2 -- "BALANCE = 1090  
bank_transfers" --> V3((ID = 104  
ACCOUNTS))
            
```

CREATE PROPERTY GRAPH bank_graph
VERTEX TABLES (
 bank_accounts as ACCOUNTS
 PROPERTIES(ID, BALANCE)
)
EDGE TABLES (
 bank_transfers
 SOURCE KEY (from_acc) REFERENCES ACCOUNTS(ID)
 DESTINATION KEY (to_acc) REFERENCES ACCOUNTS(ID)
 PROPERTIES (amount)
)

Symbol	Name	Example
()	Vertex	(v1) and (v2) are bank accounts
[]	Edge	[e1] represents a cash transfer between them
{ }	<Path length	{1,3}
->	Directed edge	

```

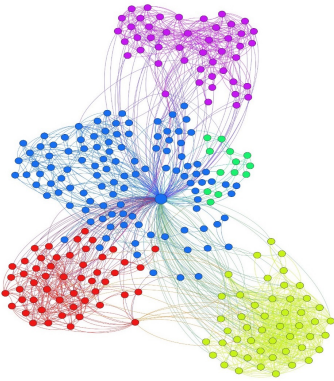
SELECT distinct(account_id)
FROM GRAPH_TABLE(bank_graph
MATCH (v1)-[is bank_transfers]->{3,5}(v1)
COLUMNS (v1.id as account_id)
);
            
```

Copyright © Intelligent Business Strategies 1992-2024
54



SQL/PGQ Opens Up Graph Analytics To Tools Generating SQL And Also To Developers – E.g., Community Detection (also known as Graph Clustering)

- Goal: Find group of vertices well-connected internally and poorly-connected externally



Applications

- Community detection in social networks
- Community detection in prospects, customer segments and across segments
- Document clustering
- Unsupervised learning
- And many more...

Image Source: <https://mattbagott.wordpress.com/2012/06/18/visualization-of-clusters-of-friends-on-facebook/>

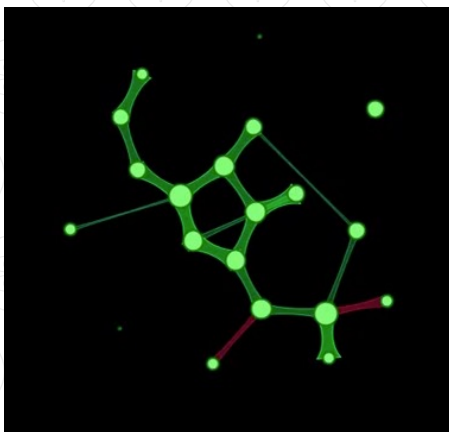
Copyright © Intelligent Business Strategies 1992-2024

55



SQL/PGQ On Historical Data Enables Understanding Of Community Evolution Monitoring The Velocity Of Market Changes, Changes In Operations, e.g. Supply Chain, Distribution, Web Traffic

This is the rate that the graph dynamically changes over time
Changes to a graph are indicated by new nodes, deleted nodes, new edges...



e.g. Communities can grow and shrink

See communities evolution over time

6 categories of evolving communities:

- Growth
- Contraction
- Merging
- Splitting
- Birth
- Death

Combine with ML and PREDICT community evolution

Could be hugely valuable in marketing, fraud pattern evolution, planning (resource usage)

What temporal interaction patterns are there?

How do these patterns help to identify communities?

Copyright © Intelligent Business Strategies 1992-2024

56

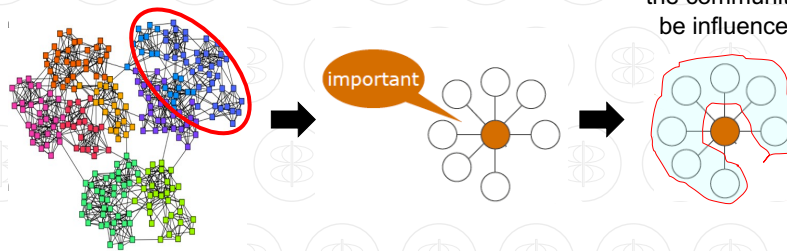


Sequencing Graph Analysis – Find The Marketing Opportunity

Community detection
“find the communities”

Centrality “find who are the
influencers in the community”

Calculate the
susceptibility of
each member of
the community to
be influenced?



Copyright © Intelligent Business Strategies 1992-2024

57



SQL / PGQ And GQL Implications

- Property graphs are much more widely adopted and easier to understand than RDF
 - RDF is a steep learning curve and is gradually becoming legacy
- SQL/PGQ is now part of the ISO/SQL standard as of June 2023
- GQL (not the same as GraphQL for REST APIs) is a proposed standard graph query language for property graphs that will become a standard in early 2024
 - A property schema for GQL will be delivered to ISO this year
 - Being done by Linked Data Benchmark Council (LDBC - came out of EU FP7)
- Graph Normal Form (GNF) or 6NF is likely to cause the merge of RDF with property graphs

Implications

- Data catalogue vendors likely to favour SQL/PGQ and GQL over RDF by 2025
- BI, Planning, Data Governance, Data Engineering tools can exploit graph analytics in data warehouse DBMSs, data mart DBMSs, and lakehouse SQL engines supporting SQL/PGQ
- Graph analytics combined with machine learning and AI could be a game changer

Copyright © Intelligent Business Strategies 1992-2024

58



Topics – Where Are We?

- The demand for data and AI
- The need for a data foundation to underpin data and AI initiatives
- The emergence of data mesh and data products
- The challenge of a distributed data estate
- Data Fabric and how can they help build data products
- Data architecture options for building data products
- The impact of open table formats and query language extensions on architecture modernisation
- **Conclusions**



Data Architecture Evolution - A New Opportunity to Combine Analytics

- Reach data in open tables across multiple clouds via SQL and federated query in the DBMS
 - Hybrid multi-cloud analytics
- Live streaming data in open table formats and invoke ML models via SQL
 - Stream directly into open tables and SQL PREDICT via SQL
 - Real-time event driven decision intelligence, automation and dynamic planning
- Analysis of structured, semi-structured and unstructured data in pipelines
 - Process unstructured data using LLMs
 - Generate content from structured data using LLMs
 - e.g. marketing content, personalised plans
- Invocation of ML models and LLMs from inside the database via SQL
- Graph analytics via SQL/PGQ
- Combine graph analytics and ML using SQL



Conclusions

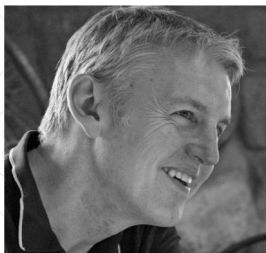
- The adoption of open table formats by analytical relational DBMS vendors is causing integration between data warehouse and lakehouse
- Multiple workloads are converging on fewer copies of data
 - Traditional BI
 - Data science ML model development
 - Streaming analytics
 - Graph analysis
- It is also possible to do this across data stored in multiple different cloud data stores
- Multiple engines can be exploited to suit different analytical workloads on the same data
- This opens up the opportunity to combine different types of analytics to create totally new insights for competitive advantage
 - CPM using LLMs, ML and graph queries on open tables
 - BI using LLMs, ML and graph queries on open tables

Copyright © Intelligent Business Strategies 1992-2024

61




About Mike Ferguson



 www.intelligentbusiness.biz

 mferguson@intelligentbusiness.biz

 @mikeferguson1

 (+44) 1625 520700

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, technology selection, enterprise architecture, and data management. Mike is also conference chairman of Big Data LDN, the largest data and analytics conference in Europe and a member of the EDM Council CDMC Executive Advisory Board. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date – the inventors of the Relational Model that caused the birth of relational databases and SQL, Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates. He teaches popular master classes in Data Strategy, Data Catalogs, Data Warehouse Modernisation, Practical Guidelines for Implementing a Data Mesh, Big Data Fundamentals, How to Govern Data Across a Distributed Data Landscape, Machine Learning and Advanced Analytics, and Embedded Analytics, Intelligent Apps and AI Automation



Thank You!

Copyright © Intelligent Business Strategies 1992-2024

62