



INTELLIGENT BUSINESS STRATEGIES PRESENTS



Mike Ferguson

**Practical Guidelines for
Implementing Data Products**

(Data Catalog, Data Fabric, Data
Product Development, Data Marketplace)

www.intelligentbusiness.biz
training@intelligentbusiness.biz

OVERVIEW

Most companies today are storing data and running applications in a hybrid multi-cloud environment. Analytical systems tend to be centralised and siloed like data warehouses and data marts for BI, Hadoop or cloud storage data lakes for data science and stand-alone streaming analytical systems for real-time analysis. These centralised systems rely on data engineers and data scientists working within each silo to ingest data from many different sources, clean and integrate it for use in a specific analytical system or machine learning models. There are many issues with this centralised, siloed approach including multiple tools to prepare and integrate data, reinvention of data integration pipelines in each silo and centralised data engineering with poor understanding of source data unable to keep pace with business demands for new data. Also, master data is not well managed.

To address these issues, a new approach has emerged attempting to accelerate creation of data for use in multiple analytical workloads. That approach is to create Data Products. This is a decentralised business domain-oriented approach to data ownership and data engineering to create reusable data products that can be created once and shared across multiple analytical systems and workloads. Multiple data architecture options are available to create data products can be . These include using one or more cloud storage accounts on an organised cloud storage data lake, on a Lakehouse, on a cloud data warehouse, using Kafka or using data virtualisation. Data products can then be consumed in other pipelines for use in streaming analytics, Data Warehouses or Lakehouse Gold Tables, for use in business intelligence, data science and other analytical workloads.

This 2-day class looks at data products in detail and examines its strengths, and weaknesses. It also looks at the strengths and weaknesses of data product implementation options. Which architecture is best to implement this? How do you co-ordinate multiple domain-oriented teams and use common data infrastructure software like Data Fabric to create high-quality, compliant, reusable, data products?. Also, how can you use a data marketplace to govern and share data products? The objective is to shorten time to value while also ensuring that data is correctly governed and engineered in a decentralised environment. It also looks at the organisational implications of democratised data product development and how to create sharable data products for master data management AND for use in multi-dimensional analysis on a data warehouse, data science, graph analysis and real-time streaming analytics to drive business value? Technologies discussed includes data catalogs, data fabric for collaborative development of data integration pipelines to create data products, DataOps to speed up the process, data orchestration automation, data marketplaces and data governance platforms.

AUDIENCE

This seminar is intended for business data analysts, data architects, chief data officers, master data management professionals, data scientists, IT ETL developers, and data governance professionals. It assumes you understand basic data management principles and data architecture plus a reasonable understanding of data cleansing, data integration, data catalogs, data lakes and data governance.

LEARNING OBJECTIVES

Attendees will learn about:

- The problems caused in existing analytical systems by a hybrid, multi-cloud data landscape
- Strengths and weaknesses of centralised data architectures used in analytics
- What are data products and how do they differ from other approaches?
- What benefits do data products offer and what are the implementation options?
- What are the principles, requirements, and challenges of implementing Data Products?

- How to organise to create data products in a decentralised environment so you avoid chaos
- The critical importance of a data catalog in understanding what data is available
- How business glossaries can help ensure data products are understood and semantically linked
- A best practice organisational model for coordinating development of data products across different domains to succeed in implementation
- What software is required to build, operate and govern data products for use in data science, a data warehouse, graph analysis and other analytical workloads?
- What is data fabric software, how does it integrate with data catalogs and connect to data in your data estate
- An Implementation methodology to produce ready-made, trusted, reusable data products
- Collaborative domain-oriented development of modular and distributed DataOps pipelines to create data products
- How a data catalog, Generative AI and data automation software can be used to generate DataOps pipelines to create data products
- Managing data quality, privacy, access security, versioning, and the lifecycle of data products
- Pros and cons of different data architecture options for implementing data products
- Publishing semantically linked data products in a data marketplace for others to consume and use
- Federated data architecture and data products - the emergence of lakehouses open tables as a way for multiple analytical workloads to access shared data products
- Persisting master data products in an MDM system
- Consuming and assembling data products in multiple analytical workloads like data warehouses, data science and graph analytics to shorten time to value
- How to implement federated data governance

MODULE 1: WHAT ARE DATA PRODUCTS AND WHY ARE THEY NEEDED?

This session looks at the challenges facing companies trying to become data driven and at the problems of siloed analytical systems and data engineering. It then looks at the emergence of Data Mesh and its introduction of Data Products as a potential way to address current problems. It explains how you can enable the creation of trusted, reusable data products using different data architecture options such as data lakes, data lakehouses, data virtualisation, message oriented middleware topics etc., for use in multiple analytical workloads. It also asks if combining multiple architecture approaches is advantageous or not.

- Data complexity in a hybrid, multi-cloud environment
- The growth in new data sources
- Siloed analytical systems and the IT data engineering bottleneck
- The need to industrialise data engineering to shorten time to value
- The emergence of Data Mesh and Data Products
- What is a data product?
- What types of data product can you build?
- Decentralised development of data products
- What are the challenges with this decentralised approach?
- Is data management software ready for Data Products?
- How will decentralised development of Data Products impact your current IT organisation and data culture?
- Is federated data governance possible?
- What are the architectural options for implementing Data Product development and what are their strengths and weaknesses?
- Implementing Data Products on Cloud Storage Vs Lakehouse Vs Cloud Data Warehouse Vs Data Virtualisation Vs Kafka
- The promise of open table formats and a federated hybrid data architecture for building data products

- Implementation requirements to create data products
 - Federated operating model
 - Common business vocabulary
 - Data producers and data consumers
 - Architecture independence
 - A unified data platform for building any pipeline to process any data
 - DataOps – component-based CI/CD pipeline development
 - Distributed pipeline execution
 - Reusable, semantically linked data products
 - Governance of a distributed data landscape
- Key technologies: Data Fabric, Data Catalogs, data classifiers, Generative AI in data management, Data Marketplace, Data Automation tools
- Vendor's offerings in the market – Alation, AWS, Ataccama, Boomi, Cambridge Semantics, Collibra, Denodo, Dremio, Global IDs, Google, IBM, Informatica, Microsoft, Oracle, One Data, Qlik (Talend), SAP, SAS, SnapLogic, Stratio, Starburst Data

MODULE 2: ORGANISING AND STANDARDISING YOUR ENVIRONMENT TO SUPPORT DEMOCRATISED DATA PRODUCT DEVELOPMENT

This session looks at how to standardise the setup in each business domain to optimise development of data products

- The importance of a program office
- Federated organisational structure
- Implementing Data Products on a single cloud Versus a hybrid multi-cloud environment
- Implementing Data Products on a Data Lake or Lakehouse
- Standardising the domain implementation process – ingest, process, persist, serve
- Creating zones in a domain cloud storage account, a Data Lake or Lakehouse to produce and persist data products
- Using Kafka as an option to persist data products

- Selecting Data fabric software as a platform for domain-oriented teams to build data products
- Applying DataOps development practices to help standardise and version control data product development?

MODULE 3: METHODOLOGIES FOR CREATING DATA PRODUCTS

This session looks at methodologies on how to produce business ready, reusable data products for use by data consumers in multiple analytical use cases who need it to drive business value.

- A best practice step-by-step methodology for building reusable data products
- How does structured, semi-structured and unstructured data impact the methodology?
- Steps-by-step data product development
 - Data concept model
 - Business glossary
 - Data source registration
 - Automated data discovery, data quality profiling, sensitive data detection, governance classification, lineage extraction and cataloguing
 - Data ingestion
 - Data product pipeline development
 - Improving data pipeline development productivity using Generative AI
 - Standardising on best practice and taking complexity away from citizen data engineers
 - Data product publishing for consumption
 - Global and domain policy creation for federated governance of classified data

MODULE 4: DEFINING AND DESIGNING DATA PRODUCTS USING A CATALOG BUSINESS GLOSSARY AND DATA MODELLING

This session looks at how you can create common data names and definitions for your data products in a business glossary so data consumers can understand the meaning of the data produced and available in Data Products. It also looks at how business

glossaries have become part of a data catalog

- Why is a common vocabulary relevant?
- Data catalogs and the business glossary
- The Data Catalog market, e.g., Alation, Atlan, Amazon Glue, Ataccama ONE, BigID, Cambridge Semantics ANZO Data Catalog, Collibra Catalog, data.world, Denodo Data Catalog, Google Data Catalog, Hitachi Vantara Lumada, IBM Watson Knowledge Catalog, Informatica IDMC Data Governance & Catalog, Microsoft Purview Data Catalog, Oracle, Qlik (Talend) Catalog, SAP DataSphere, Top Quadrant TopBraid
- Roles, responsibilities, and processes needed to manage a business glossary
- Jumpstarting a business glossary with a data concept model
- Defining data products using glossary terms
- Using a catalog and glossary to ensure data products are semantically linked
- Design options for data product data models
- Incrementally building an Enterprise Data Model while designing and building data products
- Assembling data product data model components to create a data warehouse

MODULE 5: SOURCING, MAPPING AND DATA QUALITY PROFILING DATA FOR YOUR DATA PRODUCTS

This session looks at how you can use the capabilities of a data catalog to source data for your data products using automated data discovery. It also looks at how a data catalog can help you automate the mapping of automatically discovered raw data to the business terms of your data products defined in a business glossary. Finally, it looks at how data catalogs can also automatically profile the data quality of your data sources to quickly determine what data needs to be cleaned when building your data products.

- Sourcing data for data products using a data catalog for automated data discovery
- Mapping discovered physical data to data product business terms in your business glossary

- Using a data data catalog to automatically profile the quality of source data for your data products

MODULE 6: BUILDING DATAOPS PIPELINES TO CREATE REUSABLE DATA PRODUCTS

This session looks at designing and developing modular DataOps pipelines to produce trusted data products using Data Fabric software

- Collaborative pipeline development & orchestration to produce data products
- Designing component based DataOps pipelines to produce data products
- Using CI/CD to accelerate development, testing and deployment
- Designing in sensitive data protection in pipelines
- Processing streaming data in a pipeline
- Handling schema drift in a pipeline
- Processing unstructured data in a pipeline using ML and Generative AI
- Generating data pipelines using Generative AI and Data Automation tools
- Using data observability to monitor and improve pipelines
- Making data products available for consumption using a data marketplace
- The Enterprise Data Marketplace – enabling information consumers to shop for data
- Creating Data Contracts to govern the sharing of data products in a data marketplace
 - What should be in a data contract?
- Serving up trusted data products for use in multiple analytical workloads
- Consuming data products in other pipelines for use in data warehouses, lakehouses, data science sandboxes, graph analysis, BI tools and MDM

MODULE 7: IMPLEMENTING FEDERATED DATA GOVERNANCE TO PRODUCE AND USE COMPLIANT DATA PRODUCTS

With data highly distributed across so many data stores and applications, on-premises, in multiple clouds and the edge, many companies are struggling to govern data throughout its lifecycle. This is critically important in a Data Mesh where federated computational data governance is a

fundamental principal, data product development is decentralised, and data products are shared and consumed across the organisation. It is also paramount across the whole hybrid multi-cloud data landscape. This session looks at how this can be achieved.

- What is involved in federated data governance?
- How do you implement this across a hybrid, multi-cloud distributed data landscape?
- Understanding compliance obligations
- Types of data governance policies
- Understanding Global Vs local policies when creating Data Products
- Defining sensitive data types
- Using the data catalog for automated data profiling, quality scoring and sensitive data type classification
- Defining and attaching policies to classified data in a data catalog
- Creating sharable master data products and reference data products for MDM and RDM
- Ensuring data quality in data product development
- Protecting sensitive data in data product development for data privacy compliance
- Governing data product version management
- Standardising the process for publishing data products in a data marketplace
- Governing consumer access to data products containing sensitive data
- Prevent accidental oversharing of sensitive data products using DLP
- Governing data retention of data products in-line with compliance and legal holds
- Monitoring and data stewarding to ensure policy enforcement
- Data catalog, data fabric and data marketplace technologies to help govern data across a distributed data landscape
 - Types of data governance offerings
 - AWS Glue and Datazone
 - Alation Data Marketplace,
 - Ataccama, Collibra, Confluent Schema Registry and Catalog
 - Databricks Unity Catalog and Okera



- Google Cloud IAM, Data Catalog, BigQuery, Dataplex and DLP
- IBM Cloud Pak for Data, IBM Knowledge Catalog, & IBM Data Product Hub
- Immuta, Imperva
- Informatica IDMC Governance and Catalog, Data Marketplace and Cloud Data Access Management
- Microsoft Purview
- OneTrust Data Governance Suite
- Oracle Enterprise Data Management Cloud
- SAP DataSphere
- Qlik (Talend)

PRESENTER



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, data architecture, and data management. Mike is also conference chairman of Big Data LDN, the fastest growing 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 Europe Limited – the inventors of the

Relational Model, a Chief Architect at Teradata on the Teradata DBMS. He teaches popular master classes in:

- Building a Data and AI Strategy for A Data-Driven Enterprise
- Modern Data Architecture
- AI-Assisted Active Data Governance
- Practical Guidelines for Implementing Data Products (Data Catalog, Data Fabric, Data Product Development, Data Marketplace)
- Embedded Analytics, Intelligent Apps, AI Agents & AI Automation
- Data Warehouse Modernisation
- Migrating your Data Warehouse to the Cloud
- Data Catalogs – Governing and Provisioning Data in a Data Driven Enterprise
- Real-Time Analytics

ONSITE TRAINING

All training can be given as onsite education, tailored to your company's requirements. For further details please contact us at training@intelligentbusiness.biz

