TECHNOLOGY TRANSFER PRESENTS

MIKE FERGUSON

Creating Re-Usable Data Products for Analytics
Data Lake vs. Lakehouse vs. Data Mesh

ONLINE LIVE STREAMING

MARCH 28-29, 2022

info@technologytransfer.it
www.technologytransfer.it

ABOUT THIS SEMINAR

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, new data architectures have emerged attempting to accelerate creation of data for use in multiple analytical workloads. Data Mesh is a decentralised data architecture with domain-oriented data ownership and decentralised self-service data engineering to create a mesh of data products serving multiple analytical systems. Also, Data Lakes can be used for the same thing and integrated with Data Warehouses or Lakehouses so lower latency data products can be created once and used in streaming analytics, Business Intelligence, Data Science and other analytical workloads.

This 2-day class examines the strengths, and weaknesses of Data Lakes, Data Mesh and Data Lakehouses and at how multiple domain-oriented teams can use common data infrastructure software to create trusted, compliant, reusable, data products in a Data Mesh or Data Lake for use in Data Warehouses, Data Lakehouses and Data Science to drive value. The objective is to shorten time to value while also ensuring that data is correctly governed in a decentralised environment.

It also looks at the organisational implications of these architectures and how to create sharable data products for Master Data Management and for use in multiple analytical workloads. Technologies discussed includes data Catalogs, self-service data integration, Data Fabric, DataOps, Data Warehouse automation, Data Marketplaces and Data Governance platforms.

WHO SHOULD ATTEND

- Business Data Analysts
- Data Architects
- Chief Data Officers
- Master Data Management Professionals
- Data Scientists
- IT ETL Developers
- 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.

ABOUT THIS SEMINAR

LEARNING OBJECTIVES

- Strengths and weaknesses of centralised data architectures used in analytics
- The problems caused in existing analytical systems by a hybrid, multi-Cloud data landscape
- What is a Data Mesh a Data Lake and a Data Lakehouse? What benefits do they offer?
- What are the principles, requirements, and challenges of implementing these approaches?
- 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 as a service
- How business glossaries can help ensure data products are understood and semantically linked
- An operating model for effective federated Data Governance
- What common data infrastructure software is required to operate and govern a Data Mesh, a Data Lake or a Data Lakehouse?
- 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 and automation software can be used to generate DataOps pipelines
- Managing data quality, privacy, access security, versioning, and the lifecycle of data products
- Publishing semantically linked data products in a data marketplace for others to consume and use
- Consuming data products in an MDM system
- Consuming and assembling data products in multiple analytical systems to shorten time to value

OUTLINE

1. What is Data Mesh, a Data Lake and a Lake house? Why to use them?

This session looks at the challenges facing companies trying to become data driven and at the strengths and weaknesses of current centralised data architectures used in analytics. It then introduces Data Lakes, Data Mesh and Data Lakehouse as potential ways to address current problems. It explores the pros and cons of each of these, explains how they work and how they enable creation of trusted, reusable data products for use in multiple analytical workloads. It also asks if combining these approaches is advantageous or not.

- Data complexity in a hybrid, multi-Cloud environment
- The growth in new data sources
- Centralised data architectures in use in existing analytical systems
- Strengths and weaknesses of the centralised approach
- What is a Data Mesh?
- Data Mesh principles
- How does decentralised Data Mesh work?
- What is a data product?
- What types of data product can you build?
- Decentralised development of data products
- Pros and cons of Data Mesh
- What are the challenges with this decentralised approach?
- Is Data Management software ready for Data Mesh?
- How will Data Mesh impact your current IT or ganisation and data culture?
- Is federated Data Governance possible?
- Pros and cons of Data Lakes
- The merging of Data Warehouses and Data Lakes
- The move from just Data Science to multi-purpose-Data Lakes
- What is a Data Lakehouse?
- How does a Data Lakehouse work?
- Pros and cons of a Data Lakehouse
- Can you combine Data Lakes, Lakehouses and Data Mesh and why would you do this
- Implementation requirements to create data products
    - o Federated operating model
    - o Common business vocabulary
    - o Data producers and data consumers
    - o Architecture independence
    - o A unified data platform for building any pipeline to process any data
    - o DataOps - component-based CI/CD pipeline developmen
    - o Distributed pipeline execution
    - o Reusable, semantically linked data products
    - o Governance of a distributed data landscape
- Key technologies: Data Fabric, Data Catalogs, data classifiers, Data Marketplace, Data Warehouse Automation tools
- Vendor's offerings in the market - Alation, AWS, BigID, Cambridge Semantics, Collibra, Global IDs, Google, IBM, Informatica, Microsoft, Oracle, Qlik, Talend, SAP, SAS, WhereScape, Zaloni

2. Methodologies for creating Data Products

This session looks at 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. It also looks how master data products can also be produced for use in Master Data Management.

- Creating a program office
- Decentralised development of data products in a Data Mesh, Data Lake or Lakehous
- The special and critical case of Master Data
- A best practice step-by-step methodology for building reusable data products
- How does structured, semi-structured and unstructured data impact the methodology?
- Applying DataOps development practices to data product development?

3. Using a Business Glossary to define Data Products

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 a Data

Mesh or a Data Lake. 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, Amazon Glue, Cambridge Semantics ANZO Data Catalog, Collibra Catalog, Data.world, Denodo Data Catalog, Google Data Catalog, Hitachi Vantara Lumada, IBM Watson Knowledge Catalog, Informatica Axon and EDC, Microsoft Azure Purview Data Catalog, Qlik Catalog, Zaloni Data Platform
- 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?

4. Standardising development and operations in a Data Mesh, Data Lake or Lakehouse

This session looks at how to standardise the setup in each business domain to optimise development of data products in a Data Mesh, a Data Lake or Lakehouse.

- The importance of a program office
- Implementing Data Mesh on a single cloud Versus a hybrid multi-cloud environment
- Implementing a Data Lake or Lakehouse
- Standardising the domain implementation process - ingest, process, persist, serve
- Creating zones in a Data Mesh domain, a Data Lake or Lakehouse to produce and persist data products
- Selecting Data Fabric software for building data product
- Steps-by-step data product development
    - o Data source registration
    - o Automated data discovery, profiling, sensitive data detection, governance classification, lineage extraction and cataloguing
    - o Data ingestion
    - o Global and domain policy creation for federated governance of classified data
    - o Data product pipeline development
    - o Data product publishing for consumption

5. Building DataOps Pipelines to create multi-Purpose 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
- Processing unstructured data in a pipeline using ML
- Generating data pipelines using Data Warehouse Automation tools
- Making data products available for consumption in a Data Mesh or Data Lake using a Data Marketplace
- The Enterprise Data Marketplace - enabling information consumers to shop for data
- Serving up trusted data products for use in multiple analytical systems and in MDM
- Consuming data products in other pipelines for use in Data Warehouses, Lakehouses, Data Science sandboxes, graph analysis and MDM

6. 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 in a Data Lakehouse
and 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 a Data Mesh, a Data Lake or Data Lakehouse
- 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
- 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 and data fabric technologies to help govern data across a distributed data landscape

    - Types of data governance offerings
    - Alation, Ataccama, Collibra, Dataguise
    - Google Cloud IAM, Data Catalog, BigQuery, Dataplex and DLP
    - IBM Cloud Pak for Data, Watson Knowledge Catalog, Optim & Guardium
    - Hitachi Vitara, Immuta, Imperva
    - Informatica EDC and Axon
    - Microsoft 365 Compliance Centre and Azure Purview
    - Okera, OneTrust Data Governance Suite
    - Oracle Enterprise Data Management Cloud, Privitar, SAP Data Intelligence
    - Talend, TopQuadrant

Speaker

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specialises in Business Intelligence and Enterprise Business Integration. With over 39 years of IT experience, he has consulted for dozens of companies on Business Intelligence Strategy, technology selection, enterprise architecture, and data management. 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 and European Managing Director of Database Associates. He teaches popular master classes in Operational Business Intelligence, New Technologies in DW and BI for the Agile Enterprise, Big Data Multi-Platform Analytics, Master Data Management and Enterprise Data Governance.

INFORMATION

PARTICIPATION FEE

€ 1100

The fee includes all seminar documentation.

SEMINAR TIMETABLE

9.30 am - 1.00 pm
2.00 pm - 5.00 pm

HOW TO REGISTER

You must send the registration form with the receipt of the payment to:
TECHNOLOGY TRANSFER S.r.l.
Piazza Cavour, 3 - 00193 Rome (Italy)
Fax +39-06-6871102

PAYMENT

Wire transfer to:
Technology Transfer S.r.l.
Banca: Credit Agricole
Agenzia 1 di Roma
IBAN Code:
IT 03 W 06230 03202 000057031348
BIC/SWIFT: CRPPIT2P546

GENERAL CONDITIONS

DISCOUNT

The participants who will register 30 days before the seminar are entitled to a 5% discount.

If a company registers 5 participants to the same seminar, it will pay only for 4.

Those who benefit of this discount are not entitled to other discounts for the same seminar.

CANCELLATION POLICY

A full refund is given for any cancellation received more than 15 days before the seminar starts. Cancellations less than 15 days prior the event are liable for 50% of the fee. Cancellations less than one week prior to the event date will be liable for the full fee.

CANCELLATION LIABILITY

In the case of cancellation of an event for any reason, Technology Transfer's liability is limited to the return of the registration fee only.

MIKE FERGUSON
CREATING RE-USABLE DATA PRODUCTS FOR ANALYTICS DATA LAKE VS. LAKEHOUSE VS. DATA MESH

March 28-29, 2022

Registration fee: 1100€

If anyone registered is unable to attend, or in case of cancellation of the seminar, the general conditions mentioned before are applicable.

first name .....

surname .....

job title .....

organisation .....

address .....

postcode .....

city .....

country .....

telephone .....

fax .....

e-mail .....

Stamp and signature

Send your registration form with the receipt of the payment to:
Technology Transfer S.r.l.
Piazza Cavour, 3 - 00193 Rome (Italy)
Tel. +39-06-6832227 - Fax +39-06-6871102
info@technologytransfer.it
www.technologytransfer.it