Practical Guidelines for Implementing a Data Mesh

quest for knowledge

W. www.q4k.com
E. info@q4k.com
P. +31 76 57 21 99
P. +32 2 808 99 46
P. +46 8 525 07 005

COURSE DESCRIPTION

OVERVIEW

Most companies today are storing data and running applications in a hybrid multi-cloud environment. Analytical systems tend to be centralized 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 centralized 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 centralized, siloed approach including multiple tools to prepare and integrate data, reinvention of data integration pipelines in each silo and centralized data engineering with a 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 the creation of data for use in multiple analytical workloads. That approach is Data Mesh. Data Mesh is a decentralized business domain-oriented approach to data ownership and data engineering to create a mesh of reusable data products that can be created once and shared across multiple analytical systems and workloads. A Data Mesh can be implemented in a number of ways. These include using one or more cloud storage accounts on cloud storage, on an organized data lake, on a Lakehouse, on a data cloud, using Kafka or using data virtualization. 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 course looks at:

- Data Mesh in detail and examines its strengths, and weaknesses
- Data Mesh implementation options
- Which architecture is best to implement a Data Mesh
- How to co-ordinate multiple domain-oriented teams
- The use of common data infrastructure software like Data Fabric to create high-quality, compliant, reusable, data products in a Data Mesh.
- How to use a data marketplace to share data products
- The objective is to shorten the time to value while also ensuring that data is correctly governed and engineered in a decentralized environment.
- Organizational implications of a Data Mesh
- 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 like 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

COURSE DESCRIPTION

WHY ATTEND

You will learn about:

- Strengths and weaknesses of centralized data architectures used in analytics
- The problems caused in existing analytical systems by a hybrid, multi-cloud data landscape
- What is a Data Mesh and how does it differ from a Data Lake and a Data Lakehouse?
- What benefits does a Data Mesh offer and what are the implementation options?
- What are the principles, requirements, and challenges of implementing these approaches?
- How to organize to create data products in a decentralized 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
- An operating model for effective federated data governance
- What software is required to build, operate and govern a Data Mesh of data products for use in a Data Lake, a Data Lakehouse or Data Warehouse?
- 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 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 like data warehouses, lakehouses and graph databases to shorten time to value

WHO SHOULD ATTEND

This course is intended for business data analysts, data architects, chief data officers, master data management professionals, data scientists, ETL developers and data governance professionals.

COURSE DESCRIPTION

PREREQUISITES

This course assumes that you have an understanding of basic data management principles and data architecture plus a reasonable understanding of data cleansing, data integration, data catalogs, data lakes and data governance.

INSTRUCTOR

Mike Ferguson is the Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specializes 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. 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 Data Warehouse Modernization, Big Data Architecture & Technology, Centralised Data Governance of a Distributed Data Landscape, Practical Guidelines for Implementing a Data Mesh (Data Catalog, Data Fabric, Data Products, Data Marketplace), Real-Time Analytics, Embedded Analytics, Intelligent Apps & AI Automation, Migrating your Data Warehouse to the Cloud, Modern Data Architecture and Data Virtualisation & the Logical Data Warehouse.

COURSE OUTLINE

01 WHAT IS A DATA MESH, A DATA LAKE AND A LAKEHOUSE? WHY USE THEM?

This module looks at the challenges facing companies trying to become data-driven and at the strengths and weaknesses of current centralized data architectures used in analytics. It then introduces Data Lakes, Data Lakehouse and Data Mesh as potential ways to address current problems. It explores the pros and cons of each of these and explains how you can enable the creation of trusted, reusable data products in a Data Mesh using different architecture options such as cloud storage accounts, data lakes, cloud data platforms etc., for use in multiple analytical workloads. It also asks if combining multiple architectural approaches is advantageous or not.

- Data complexity in a hybrid, multi-cloud environment
- The growth in new data sources
- Centralized data architectures in use in existing analytical systems - data warehouses, data lakes
- 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
- What is a Data Mesh?
- Data Mesh principles
- How does decentralized Data Mesh work?
- What is a data product?
- What types of data products can you build?
- Decentralized development of data products
- Pros and cons of Data Mesh
- What are the challenges with this decentralized approach?
- Is data management software ready for Data Mesh?
- How will Data Mesh impact your current IT organization and data culture?
- Is federated data governance possible?
- What are the architectural options for implementing Data Mesh?
- Implementing Data Mesh on Data Cloud vs Kafka vs Cloud Storage vs Lakehouse vs Data Virtualization
- 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, Data Marketplace, Data Automation tools
- Vendor's offerings in the market – Alation, AWS, BigID, Cambridge Semantics, Collibra, Dremio, Global IDs, Google, IBM, Informatica, Microsoft, Oracle, Qlik, Talend, SAP, SAS, Software AG, Starburst Data, StreamSets, IDERA WhereScape, Truist Zaloni

COURSE OUTLINE

02 METHODOLOGIES FOR CREATING DATA PRODUCTS

This module looks at how to produce business-ready, reusable data products for use by data consumers in multiple analytical use cases who need them to drive business value. It also looks at how master data products can also be produced for use in master data management.

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

03 USING A BUSINESS GLOSSARY TO DEFINE DATA PRODUCTS

This module 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?

04 STANDARDIZING DEVELOPMENT AND OPERATIONS IN A DATA MESH, DATA LAKE OR LAKEHOUSE

This module looks at how to standardize the setup in each business domain to optimize the development of data products in a Data Mesh.

- The importance of a program office
- Implementing Data Mesh on a single cloud vs a hybrid multi-cloud environment

COURSE OUTLINE

- Implementing Data Mesh on a Data Lake or Lakehouse
- Standardizing 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 in a Data Mesh
- Using Kafka as an option to persist data products in a Data Mesh
- Selecting Data Fabric software as a platform for domain-oriented teams to build data products
- Step-by-step data product development
  - Data source registration
  - Automated data discovery, data quality profiling, sensitive data detection, governance classification, lineage extraction and cataloging
  - Data ingestion
  - Global and domain policy creation for federated governance of classified data
  - Data product pipeline development
  - Standardizing on best practices and taking the complexity away from citizen data engineers
  - Data product publishing for consumption

05 BUILDING DATAOPS PIPELINES TO CREATE MULTI-PURPOSE DATA PRODUCTS

This module 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 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

COURSE OUTLINE

06 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 principle, data product development is decentralized, and data products are shared and consumed across the organization. It is also paramount across the whole hybrid multi-cloud data landscape. This module 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, Confluent Schema Registry and Catalog, Dataguise
  - Google Cloud IAM, Data Catalog, BigQuery, Dataplex and DLP
  - IBM Cloud Pak for Data, Watson Knowledge Catalog, Optim & Guardium
  - Hitachi Vantara
  - Immuta, Imperva
  - Informatica EDC and Axon
  - Microsoft Purview
  - Okera, OneTrust Data Governance Suite
  - Oracle Enterprise Data Management Cloud
  - Privitar
  - SAP Data Intelligence
  - Software AG StreamSets DataOps Platform
  - Talend, TopQuadrant

PRICING

The fee for this 2-day course is EUR 1.450 (+VAT) per person.

We offer the following discounts:

- 10% discount for groups of 2 or more students from the same company registering at the same time.
- 20% discount for groups of 4 or more students from the same company registering at the same time.

Note: Groups that register at a discounted rate must retain the minimum group size or the discount will be revoked. Discounts cannot be combined.

COURSE DATES

14-15 MAY 2024	STOCKHOLM
25-26 NOVEMBER 2024	AMSTERDAM