Data Catalogs: Governing & Provisioning Data in a Data Driven Enterprise

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

This 1-day course looks in detail at what a data catalog is and what the main reasons are for needing one. In particular, you will look at the challenge companies are dealing with respect to data complexity and the need to implement both an enterprise data governance program and provide technologies to help people find, engineer, provision and consume 'business ready' data for both analytical and operational use cases.

WHY ATTEND

You will learn:

- How data catalogs work and what their capabilities are?
- How to use data catalogs to discover, classify and catalog data in multiple data stores both on-premises, across multiple clouds, and at the edge?
- How to use data catalogs in an enterprise data governance program to systematically classify data and set policies to govern classified data and content across their distributed data estate from a single place. This includes automatic discovery and classification of sensitive data, governance of data access security, data privacy, data loss prevention, data sharing, data usage, data retention, and data quality?
- How to use data catalogs to discover data they can engineer in data integration pipelines to produce data products that can be published in a data marketplace?

WHO SHOULD ATTEND

This course is intended for business and IT professionals responsible for data engineering, data provisioning and enterprise data governance including data access security, data privacy, data sharing, data usage, data retention, data quality of both structured data and content. This includes Chief Data Officers, Citizen and professional IT Data Engineers, Data Architects, Data Scientists, Heads of Data Governance, Data Stewards, Solution Architects and Enterprise Architects.

COURSE DESCRIPTION

PREREQUISITES

This course assumes a basic understanding of data governance, data management, metadata, data warehousing, data cleansing, data integration etc.

INSTRUCTOR

MIKE FERGUSON

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 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 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 INTRODUCTION

This module looks at typical existing setups and challenges that companies are facing to explain why data catalogs are needed.

- The ever-increasing complex distributed data landscape – on-premises, multiple clouds, SaaS applications and the edge
- The growth in new data sources
- Disparate operational transaction systems – SaaS applications, on-premises and cloud
- Existing siloed analytical systems – data warehouse, data lakes, data lakehouses
  - Data engineering problems in a siloed environment
- The emergence of Data Mesh and its impact to data engineering and data architecture
- The impact of ungoverned data on
  - Business operations, decision making and risk
  - Business profitability and ability to respond to competitive pressure
  - Ability to comply with data privacy legislation in one or more jurisdictions
  - Data security, the possibility of data breaches and their business impact
- Major requirements facing companies with respect to data
  - The need for end-to-end data governance to:
    - Make data easy to find and understand no matter where it is
    - Know where sensitive data is so it can be governed
    - Comply with multiple data privacy regulations and legislation
    - Avoid data breaches in a complex distributed data estate
    - Understand where data quality is poor
    - Govern data access, privacy, sharing, usage, retention and quality
  - The need to accelerate data engineering and provision business-ready data to share and reuse across the enterprise
  - Key technologies needed – the data catalog and data fabric

02 WHAT IS A DATA CATALOG AND WHY HAVE ONE?

This module looks at what a data catalog is, why you need it, software vendors in the market and outlines data catalog capabilities.

- What is a data catalog?
- Why have them?
  - Data Catalog use case

COURSE OUTLINE

- The Data Catalog Marketplace
  - Alation, Ataccama, Atlan, AWS Glue Data Catalog, BigID, Cambridge Semantics Anzo Data Catalog, Collibra Data Catalog, data.world, Google Data Catalog, Hitachi Vantara Lumada, IBM Watson Knowledge Catalog, Informatica IDMC Data Governance and Catalog, Microsoft Purview, Oracle, SAP, Qlik (Talend) Data Catalog, TopQuadrant TopBraid, Truist Zaloni Data Catalog
- Core data catalog capabilities
  - Business glossary
  - Automated data discovery
  - Manual and automated mapping to business glossary
  - Manual and automated data governance classification
  - Automated data quality profiling
  - Data governance policy management
- What types of metadata and asset types are stored in a data catalog?
- Supporting different types of personas
- Skills needed to use one
- Data Catalog maintenance – manual, automated

03 DATA CATALOG CAPABILITIES: THE IMPORTANCE OF A BUSINESS GLOSSARY

This module looks at the need to understand your data landscape from a business perspective. The key to making this happen is to establish a common business vocabulary in the business glossary of a data catalog to create common data names and definitions for your data. This enables you to search for and govern data across your data estate from a business perspective.

- Data standardization using a shared business vocabulary
- Business glossary software – now a capability of a data catalog
- The purpose of a common vocabulary in data governance
  - Alation, Amazon Glue, Collibra, Informatica IDMC Business Glossary, IBM Watson Knowledge Catalog, Microsoft Azure Purview, Qlik (Talend) Business Glossary and Data Catalog, SAS Business Data Network, TopQuadrant TopBraid EDG Business Glossary
- Planning for a business glossary
- Glossary roles and responsibilities
- Glossary term submission, voting approval and dispute resolution processes
- Approaches to creating a common vocabulary
- Organizing data definitions in a business glossary
- The role of a data concept model
- Utilizing a common vocabulary in Data Modelling, ETL, BI, ESB, APIs, & MDM

COURSE OUTLINE

04 DATA CATALOG CAPABILITIES: UNDERSTANDING YOUR DATA LANDSCAPE USING AUTO DATA DISCOVERY, CATALOGUING AND MAPPING TO A BUSINESS GLOSSARY

Having defined your data, this session looks at discovering what data you have, where it is and how it maps to your business glossary to provide a business understanding of your data landscape.

- Understanding your data landscape - the critical role of data catalog software
- The data discovery process
- Registering data sources for discovery
- Automated data discovery, profiling, using a data catalog
- Mapping data assets to a business glossary

05 DATA CATALOG CAPABILITIES: CLASSIFYING DATA AND CONTENT TO KNOW HOW TO GOVERN IT

This module looks at manually and automatically labeling data using a data catalog to know how to govern it using predefined classifiers, user-defined classification schemes and trainable classifiers. It then looks at how classified data shows up in a data catalog and how policies can be assigned to label data to govern it across your data estate.

- What is data classification?
- Automated sensitive data type detection and classification using pre-defined trained classifiers
- Creating your own data classification schemes for data confidentiality and retention
- Manually classify content using your own classification scheme, e.g. Office Documents, SharePoint, Email, Chat, Microsoft Teams or Zoom Meetings
- Training classifiers to automatically label content
- Using trained classifiers to auto-label content in the cloud and on-premises
- Using your own classification schemes with a data catalog
- Automatically classifying sensitive structured data and objects using a data catalog
- Using classification insights to understand sensitive data proliferation and data redundancy across your estate
- Setting policies in a data catalog to govern data across your data estate

06 DATA CATALOG USE CASE: WHAT'S NEEDED TO GOVERN DATA ACROSS A DISTRIBUTED DATA LANDSCAPE

This module looks at what the requirements are to govern data in a modern enterprise and how the requirements can be met using a data catalog.

- Key requirements for governing data and content across a distributed data landscape

COURSE OUTLINE

- What do you need to know to govern data?
- Introducing a data governance framework to help meet the challenge
- People
  - Key roles and responsibilities
  - Getting the organization and operating model right
  - Data owners, data stewards, data governance control board and working groups
- Core processes needed to establish and govern commonly understood data
- Types of policies and rules needed to govern
  - Data quality
  - Data access security
  - Data privacy
  - Data retention
  - Data loss prevention
  - Data sharing
  - Data use and maintenance
- The role of the data catalog
  - Automated sensitive data discovery
  - Trainable classifiers
  - Data catalog policy creation
  - Integration with policy enforcement technologies
    - Dynamic data masking
    - Data loss prevention
- Core data governance capabilities needed
- Tasks involved in governing a distributed data landscape

07 DATA CATALOG USE CASE: FINDING, ENGINEERING AND PROVISIONING DATA

This module looks at what the requirements are to accelerate data engineering in a modern enterprise and how the requirements can be met using a data catalog and data fabric software.

- Defining data products in a business glossary
- Automatically discovering, mapping and classifying data in a data catalog
- Integration with Data Fabric
- Building pipelines to produce data products
- Creating a data marketplace as a data catalog application to share business-ready data
- Publishing and consuming data products using a data marketplace and data catalog metadata

PRICING

The fee for this course is EUR 725,00 (+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

16 MAY 2024	AMSTERDAM
27 NOVEMBER 2024	STOCKHOLM