Centralised Data Governance of a Distributed Data Landscape

- Data Catalog, Data Discovery, Data Classification, Data Security, Data Privacy, Data Retention, Data Sharing, Data Quality, Master Data Management

Instructor: Mike Ferguson
Duration: 2 Days

OVERVIEW

Many businesses today are operating in a distributed computing environment with data and processes running across on-premises systems, multiple clouds, on SaaS applications and the edge. It this environment, with so much going on, data is much harder to find and govern. Also, the number of data sources continues to grow and master data, the most widely used data in any business, is becoming harder to find, manage and keep synchronised across so many systems in a hybrid computing environment. This two-day in-depth class looks at this problem and shows how to successfully implement a data governance program to centrally govern data across a distributed data landscape. This includes governing data access security, data privacy, data sharing, data retention and data quality with data quality encompassing master data management to create a 360-degree view of customers, products, suppliers, and other core entities. It is intended for chief data officers, heads of data governance, enterprise architects, data architects, MDM professionals, business professionals, database administrators, data engineers, and compliance managers responsible for data governance.

The class takes a detailed look at the business problems caused by poorly governed data and how it can seriously impact business operations, cause unplanned operational costs, and destroy confidence in accuracy of business intelligence, machine learning model predictions and recommendations. It also defines the requirements that need to be met for a company to confidently define, manage, and govern data as well as create and share consistent reference and master data across operational applications and analytical systems both on-premises and in one or more clouds.

Having understood the requirements, you will learn what should be in a governance programme. This includes a data governance framework that includes data governance roles and responsibilities, processes, policies, technologies, and a core set of data governance capabilities to govern data across a distributed data landscape. It also includes a master data management strategy and what you need to do to bring your master data under control. We will look at how to make use of a business glossary, a data catalog with automated data discovery, data quality profiling, sensitive data classification, centralised data governance policy definition by data owners around the business and policy enforcement across a distributed data landscape. We look at data cleaning and data integration, to provision master data and reference data products and how Customer Master Data can be combined with data warehouse and big data to create a Customer Data Platform (CDP) for a customer intelligent omni-channel front-office.

During the seminar we take an in-depth look at the technologies and best practice methodologies and processes for governing data across on-premises systems, SaaS applications, multiple clouds, and the edge.

AUDIENCE

This seminar is intended for business and IT professionals responsible for enterprise data governance including data access security, data privacy, data sharing, data usage, data retention, data quality (includes master data management) of both structured data and content. It assumes a basic understanding of data governance, data management, metadata, data warehousing, data cleansing, data integration, etc.

LEARNING OBJECTIVES

Attendees will learn how to set up an enterprise data governance program to systematically govern data and content across their distributed data landscape from a single place. Using a data governance framework and key technologies like data catalogs, data classifiers, data fabric and MDM they will learn what is needed to discover, classify and govern data and content. This includes data access security, data privacy, data loss prevention, data sharing, data retention, and data quality.

MODULE 1: WHAT IS DATA GOVERNANCE AND WHY DO WE NEED IT?

This session looks at what data governance is and what the main reasons are for needing to implement a data governance program. It looks at the need to comply with multiple data privacy regulations and legislation in a global business, the need to avoid data breaches and the challenges posed by a growing number of data sources and an increasingly complex distributed data landscape. It looks at the problems ungoverned data can bring and how they impact business operations, decision making and increase risk.

- The ever-increasing distributed data landscape – on-premises, multiple clouds, SaaS applications and the edge
- The impact of ungoverned data on business profitability and ability to respond to competitive pressure
- Protecting personal data in a global business – the impact of data privacy legislation in multiple jurisdictions
- Data breaches and how they impact business

MODULE 2: WHAT ARE THE REQUIREMENTS AND WHAT'S NEEDED TO GOVERN DATA ACROSS DISTRIBUTED DATA LANDSCAPE?

This session looks at what the requirements are to govern data in a modern enterprise and what is needed to make it happen.

- Key requirements for governing data and content across a distributed data landscape
- 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 organisation 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
- Technology
  - Data catalog
  - Trainable classifiers
  - Data Fabric
  - Dynamic data masking
  - Data loss prevention
  - Master Data Management
- Core data governance capabilities needed
- Tasks involved in governing a distributed data landscape

MODULE 3: THE IMPORTANCE OF A BUSINESS GLOSSARY

This session 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 standardisation using a shared business vocabulary
- The purpose of a common vocabulary in data governance
- Business glossary software – now a capability of a data catalog
  - Alation, Amazon Glue, Collibra, Informatica Axon Business Glossary, IBM Watson Knowledge Catalog, Microsoft Azure Purview, 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

- Organising data definitions in a business glossary
- The role of a data concept model
- Utilising a common vocabulary in Data Modelling, ETL, BI, ESB, APIs, & MDM

MODULE 4: UNDERSTANDING YOUR DATA LANDSCAPE – 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 Catalog Marketplace
    - Alation, Ataccama, 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 Enterprise Data Catalog, Microsoft Azure Purview, Oracle, SAP, Talend Data Catalog, TopQuadrant TopBraid, Truist Zaloni Data Catalog
- The data discovery process
- Registering data sources for discovery
- Automated data discovery, profiling, using a Data Catalog
- Mapping data assets to a business glossary

MODULE 5: CLASSIFYING DATA AND CONTENT TO KNOW HOW TO GOVERN IT

This session looks at manually and automatically labelling data 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 labelled 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 classifying 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

MODULE 6: GOVERNING DATA SECURITY ACROSS YOUR DISTRIBUTED DATA LANDSCAPE

Having classified the data and content in your data estate, this module looks at protecting data and content in your data estate with focus on that which is classified as sensitive or confidential. It looks at setting and enforcing policies to govern data access and usage security as well as governing data loss prevention.

- Data security objectives
- Key technologies in governing data security
    - Policy establishment
    - Policy enforcement
- Steps to implement data security
- Setting policies centrally in your data catalog to govern data access across your data estate
    - Attribute based access control
- Unifying data access control across multiple data stores
- Universal authorisation fabric software (e.g. IBM, Immuta, Okera) and how they integrate with data catalogs
- Setting policies centrally in your data catalog to govern data usage across your data estate
- What are cloud application security brokers?
    - Auto discovery of cloud application usage

- o Setting policies to govern access to and use of sensitive data and content from applications
- o Monitoring cloud application activity

MODULE 7: GOVERNING DATA PRIVACY ACROSS YOUR DISTRIBUTED DATA LANDSCAPE

This session looks at governing access to personal data across your data estate to remain compliant with legislation in multiple jurisdictions that your company operates

- Data privacy objectives
- Data privacy legislation – GDPR, CCPA, HIPAA and more
- Steps involved in a central data privacy governance process
- Automatically identifying where unprotected personal data is located
- Data privacy policy enforcement across a distributed data landscape
  - o Linking your data catalog to other technologies
  - o Encrypting and de-identifying personal data
  - o Using data loss prevention (DLP) to avoid loss of personal data
  - o Protecting personal data in email, chat, documents, file shares, cloud storage, and endpoints

MODULE 8: GOVERNING DATA RETENTION ACROSS YOUR DISTRIBUTED DATA LANDSCAPE

This session looks at governing the lifecycle of data across your data estate and how you can set policies to control how long data is kept for and what happens to it on expiry. It also looks a special purpose condition such as "legal holds" placed on data by legal departments.

- Creating a data retention classification scheme
- Complying with country and region-specific legislation
- Classifying data and content using retention labels
- Setting policies centrally to retain data
- Setting actions to destroy or archive it on expiration

MODULE 9: GOVERNING DATA SHARING ACROSS YOUR DISTRIBUTED DATA LANDSCAPE

This session looks at producing trusted, compliant data to be shared across the enterprise and beyond and how data sharing can be governed

- Data sharing objectives
- Key technologies to help produce trusted data products for sharing
- Steps to creating data products
- A unified approach to producing trusted data products using Data Fabric and DataOps pipelines
- Publishing trusted, compliant data products in a data marketplace
- Governing data sharing and consumption using data sharing agreements and a data marketplace
- Creating a standard data sharing approval process for consumers
- Monitoring and tracking shared data consumption and usage

MODULE 10: GOVERNING DATA QUALITY ACROSS YOUR DISTRIBUTED DATA LANDSCAPE

This session looks at consistently governing data quality across your data estate.

- The business impact of bad quality data
- Common data quality metrics
- Setting up data validation and matching rules in your data catalog
- Using your data catalog to automatically profile and validate your data
- Monitoring data quality and policies across your distributed data estate
- Integrating data catalogs with self-service data cleansing software
- AI-assisted data cleansing

MODULE 11: GOVERNING DATA QUALITY USING MASTER DATA MANAGEMENT

This session looks at master data management (MDM) and reference data management (RDM) and how to implement them to improve data quality.

- What is master data management and why is it needed?
- What is reference data management?
- Components of an MDM solution

- MDM implementation styles and options
  - Real-time master data synchronisation
  - Virtual MDM (Index / Registry)
  - Single Entity Hub
  - Enterprise Multi-Domain MDM
- Identifying master data entities
- Defining a common vocabulary for master data entities, e.g. Customer, Supplier, Product
- Master data modelling
- Master Data Hierarchy Management
- Master Data discovery – identifying where your disparate master data is located using a data catalog
- Mapping your disparate master data to your business glossary
- Profiling disparate master data to understand data quality
- Using Data Fabric to clean, match and integrate master data to create trusted master data entities
- Master data matching – fuzzy matching and survivorship rules
- Implementing outbound master data synchronisation
- Standardising business processes to create and maintain master data
- Governing maintenance of master data
- The MDM solution marketplace
  - Ataccama, IBM, Informatica, Oracle, Profisee, Reltio, Syndigo Riversand, SAP, SAS, Semarchy, Stibo, Talend, Tamr, TIBCO & more
- Evaluating MDM products
- Integration of MDM solutions with data catalogs and data fabric
- MDM in the Cloud – what's the advantage?
- Accessing and maintaining master data via shared master data services
- Integrating MDM with operational applications and process workflows
- Using master data to tag unstructured content, e.g. Supplier contracts

MODULE 12: TRANSITIONING TO CENTRALISED MASTER DATA MAINTENANCE – THE CHANGE MANAGEMENT PROCESS

This session looks at the most difficult job of all – the change management process needed to get to centralised common approach to master data maintenance. It looks at the difficulties involved, what really needs to happen and how to make it happen.

- The impact of centralised update of master data on existing processes
- Transitioning from multiple data entry systems to one data entry system
- Planning for incremental change management
- Creating an MDM change management program
- Changing application logic to use shared MDM services
- Changing user interfaces
- Leveraging a REST APIs, GraphQL and a SOA to access MDM shared services
- Changing existing business processes to take advantage of MDM
- Changing ETL jobs to leverage master data as a data source
- Hierarchy change management in MDM

MODULE 13: COMBINING MDM WITH BIG DATA AND YOUR DATA WAREHOUSE TO CREATE A CUSTOMER DATA PLATFORM

This last session looks at combining Customer master data, Big Data, and your Data Warehouse to create a Customer Data Platform to support Marketing, Sales and Customer Service in the digital enterprise.

- Integrating Master Data with Big Data and Data Warehouses
- New data sources related to customers
- Creating new customer insights using analytics
- Creating a CDP in your enterprise
- Integrating CDPs with digital and traditional marketing, sales, and service applications

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. 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 Modernisation, 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.

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