quest for knowledge

q4k.com

Beyond the Data Lake: Building a Trusted Data Platform

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

In today’s digital economy, companies are creating and capturing more and more data from an increasing number of internal and external sources that they want to analyze to make more effective decisions. Data is being captured in scalable messaging systems, edge databases, SaaS applications, cloud storage, file systems, NoSQL databases, Hadoop systems, data warehouse staging areas and more. The result is that data is spreading out across an increasingly complex distributed data landscape in many different types of data store on-premises, in multiple clouds and at the edge. It is not surprising that many companies are struggling with knowing what data is available, whether they can trust it and how they go about integrating it. Many have ended up managing information in silos like data warehouses, MDM, data science sandboxes and graph databases with different tools being used to prepare and integrate data. In addition, both IT and business users are now integrating data and the danger is total chaos. The question is, do we let this continue or is there another way to govern and unify data across an increasingly complex data landscape that can shorten time to value?

This 2-day course looks at how companies can solve this problem by using a Data Catalog, data fabric software and component based development of DataOps pipelines to rapidly ingest, organise and process data to produce trusted, reusable data products that can published in a data marketplace and consumed and used in multiple analytical workloads to drive value. It looks at different data architectures to support this approach including a centralised data lake, a lakehouse, a logical data lake and data mesh architectures discussing the pros on cons of each. It also looks at how teams of IT and business users can work together in a DataOps environment to build ready-made trusted, reusable data products once and make them available to others to consume and use to drive value. The objective is to reduce time to value, avoid reinvention while governing and unifying data quickly in a multi-cloud, multiple data store, hybrid computing environment.

✓ WHO SHOULD ATTEND

This course is intended for business data analysts doing self-service data integration, data architects, chief data officers, master data management professionals, database administrators, big data professionals, data integration developers, and compliance managers who are responsible for data management. This includes metadata management, data integration, data quality, master data management and enterprise content management. The course is not only for ‘Fortune 500 scale companies’ but for any organisation that has to deal with big data, small data, multiple data stores and multiple data sources.

COURSE DESCRIPTION

OVERVIEW

Learn how to discover, and ingest any kind of data into a data lake, organize it, make it findable and process it to produce trusted, reusable data assets in an enterprise data marketplace to make it accessible and to shorten time to value. This 2-day course shows why a data lake is now the starting point in a data architecture, why you need a data lake to lower the cost of data integration and why you can't do DataOps without one. It includes best practices in setting up data lake zones, streaming and file based ingestion of any kind of data, using a data catalog to automate data discovery and map what it discovers to your business glossary. It shows why a data catalog is critical, and provides best practices on collaborative pipeline development to enable rapid unified data delivery of trusted reusable data assets into an enterprise data marketplace.

WHY ATTEND

You will learn:

- How to define a strategy for producing trusted data as-a-service in a distributed environment of multiple data stores and data sources
- How to organize data in a centralised or distributed data environment to overcome complexity and chaos
- How to design, build, manage and operate a logical or centralised data lake within their organisation
- The critical importance of an information catalog in understanding what data is available as a service
- How data standardisation and business glossaries can help make sure data is understood
- An operating model for effective distributed information governance
- What technologies and implementation methodologies they need to get their data under control and produce ready-made trusted data products
- Collaborative curation of trusted, ready-made data products and publishing them in a data marketplace for people to shop for data
- How to apply methodologies to get master and reference data, big data, data warehouse data and unstructured data under control irrespective of whether it be on-premises or in the cloud
- Fuelling rapid 'last mile' analytical development to reduce time to values

COURSE DESCRIPTION

✓ PREREQUISITES

This course assumes that you have an understanding of basic data management principles as well as a high level of understanding of the concepts of data migration, data replication, metadata, data warehousing, data modelling, data cleansing, etc.

✓ WHY ATTEND

You will learn:

- How to define a strategy for producing trusted data as-a-service in a distributed environment of multiple data stores and data sources
- How to organize data in a centralised or distributed data environment to overcome complexity and chaos
- How to design, build, manage and operate a logical or centralised data lake within their organisation
- The critical importance of an information catalog in understanding what data is available as a service
- How data standardisation and business glossaries can help make sure data is understood
- An operating model for effective distributed information governance
- What technologies and implementation methodologies they need to get their data under control and produce ready-made trusted data products
- Collaborative curation of trusted, ready-made data products and publishing them in a data marketplace for people to shop for data
- How to apply methodologies to get master and reference data, big data, data warehouse data and unstructured data under control irrespective of whether it be on-premises or in the cloud
- Fuelling rapid ‘last mile’ analytical development to reduce time to values

COURSE OUTLINE

01 ESTABLISHING A DATA STRATEGY FOR RAPID UNIFICATION OF TRUSTED DATA ASSETS

This module introduces the data lake together with the need for a data strategy and looks at the reasons why companies need it. It looks at what should be in your data strategy, the operating model needed to implement, the types of data you have to manage and the scope of implementation. It also looks at the policies and processes needed to bring your data under control.

- The ever-increasing distributed data landscape
- The siloed approach to managing and governing data
- IT data integration, self-service data preparation or both? Data governance or data chaos?
- Key requirements for data management
    - Structured data – master, reference and transaction data
    - Semi-structured data – JSON, BSON, XML
    - Unstructured data - text, video
    - Re-usable services to manage data
- Dealing with new data sources - cloud data, sensor data, social media data, smart products (the internet of things)
- Understanding the scope of your data lake
    - OLTP system sources
    - Data Warehouses
    - Big Data systems, e.g. Hadoop
    - MDM and RDM systems
    - Data virtualisation
    - Streaming data
    - Enterprise Content Management
- Building a business case for data management
- Defining an enterprise data strategy
- A new collaborative approach to governing, managing and curating data data
- Introducing the data lake and data refinery
- Data lake configurations – what are the options?
    - Centralised, distributed or logical data lakes
- Establishing a multi-purpose data lake and Information Supply Chain to produce data products for the enterprise
- DataOps – a component-based approach to curating trusted data products
- The rising importance of an Information catalog and its role as a data marketplace
- Key technology components in a data lake and information supply chain – including data fabric software
- Using Cloud storage or Hadoop as a data staging area and why it is not enough
- Implementation run-time options – the need to execute in multiple environments
- Integrating a data lake into your enterprise analytical architecture

COURSE OUTLINE

02 INFORMATION PRODUCTION METHODOLOGIES

Having understood strategy, this module looks at why information producers need to make use of multiple methodologies in a data lake information supply chain to produce trusted structured and multi-structured data for information consumers to make use of to drive business value.

- Information production and information consumption
- A best practice step-by-step methodology to structured data governance
- Why the methodology has to change for semi-structured and unstructured data
- Methodologies for structured Vs multi-structured data

03 DATA STANDARDISATION, THE BUSINESS GLOSSARY AND THE INFORMATION CATALOG

This module looks at the need for data standardisation of structured data and of new insights from processing unstructured data. The key to making this happen is to create common data names and definitions for your data to establish a shared business vocabulary (SBV). The SBV should be defined and stored in a business glossary and is important for information consumers to understand published data in a data lake. It also looks at the emergence of more powerful information catalog software and how business glossaries have become part of what a catalog offers.

- Semantic data standardisation using a shared business vocabulary within an information catalog
- The role of a common vocabulary in MDM, RDM, SOA, DW and data virtualisation
- Why is a common vocabulary relevant in a data lake and a Logical Data Warehouse?
- Approaches to creating a common vocabulary
- Business glossary products storing common business data names
- Alteryx Connect Glossary, ASG, Collibra, Informatica Axon, IBM Information Governance Catalog, Microsoft Azure Data Catalog Business Glossary, SAS Business Data Network and more
- Planning for a business glossary
- Organising data definitions in a business glossary
- Key roles and responsibilities - getting the operating model right to create and manage an SBV
- Formalising governance of business data names, e.g. the dispute resolution process
- Business involvement in SBV creation
- Beyond structured data - from business glossary to information catalog
- What is an Information Catalog?
- Why are information catalogs becoming critical to data management?

COURSE OUTLINE

- Information catalog technologies, e.g. Alation, Alteryx Connect, Amazon Glue, Apache Atlas, Collibra Catalog, Cambridge Semantics ANZO Data Catalog, Denodo Data Catalog, Google Data Catalog, IBM Information Governance Catalog & Watson Knowledge Catalog, Informatica EDC & Live Data Map, Microsoft Azure Data Catalog, Qlik Data Catalyst, Waterline Data, Zaloni Data Platform
- Information catalog capabilities

04 ORGANISING AND OPERATING THE DATA LAKE

This module looks at how to organise data to still be able to manage it in a complex data landscape. It looks at zoning, versioning, the need for collaboration between business and IT and the use of an information catalog in managing the data.

- Organising data in a centralised or logical data lake
- Creating zones to manage data
- New requirements for managing data in centralised and logical data lakes
- Creating collaborative data lake projects
- Hadoop or cloud storage as a staging area for enterprise data cleansing and integration
- Core processes in data lake operations
- The data ingestion process
- Tools and techniques for data ingestion
- Implementing automated disparate data and data relationship discovery using Information catalog software
- Using domains and machine learning to automate and speed up data discovery and tagging
- AI in the catalog - Alation, IBM Watson Knowledge Catalog, Informatica CLAIRE, Silwood, Waterline Data Smart Data Catalog
- Automated profiling, PII detection, tagging and cataloguing of data
- Automated data mapping and lineage discovery
- The data governance classification and policy definition processes
- Manual and automated data governance classification to enable governance
- Using tag-based policies to govern data

05 THE DATA REFINERY PROCESS

This module looks at the process of refining data in an information supply chain to produce trusted data products.

- What is Data Warehouse Automation?
- What is a data refinery?
- Key requirements for refining data

COURSE OUTLINE

- The need for multiple execution engines to run in multiple environments
- Options for refining data – ETL versus self-service data preparation
- Key approaches to scalable ETL data integration using Apache Spark
- Self-service data preparation tools for Spark and Hadoop, e.g. Alteryx Designer, Informatica Intelligent Data Lake, IBM Data Refinery, Azure Data Factory Wrangling FLows, Paxata, Tableau Prep Builder, Tamr, Talend, Trifacta
- Automated data profiling using analytics in data preparation tools
- Executing data refinery jobs in a logical data lake using Apache Beam to run anywhere
- Approaches to integrating IT ETL and self-service data preparation
- ODPi Egeria for metadata sharing
- Joined up analytical processing from ETL to analytical workflows
- Publishing data and data integration jobs to the information catalog
- Mapping produced data products into your business vocabulary
- Data provisioning – publishing trusted, ready-made data products into an Enterprise Data Marketplace
- The Enterprise Data Marketplace – enabling information consumers to shop for data
- Provisioning trusted data using data virtualisation, a logical data warehouse and on-demand information services
- Consistent data management across cloud and on-premise systems

06 UNIFYING BIG DATA, MASTER DATA AND DATA WAREHOUSE DATA TO DRIVE BUSINESS VALUE

This module looks at how the data refining processes can be applied to governing, unifying and provisioning data across a big data, MDM and traditional data warehouses to drive new business value. How do you deal with very large data volumes and different varieties of data? How do you load and process data in Hadoop? How should low-latency data be handled? Topics that will be covered include:

- A walk through of end-to-end data lake operation to create a Single Customer View
- Types of big data & small data needed for single customer view and the challenge of bringing it together
- Connecting to big data sources, e.g. web logs, clickstream, sensor data, unstructured and semi-structured content
- Ingesting and analysing clickstream data
- The challenge of capturing external customer data from social networks
- Dealing with unstructured data quality in a big data environment
- Using graph analysis to identify new relationships
- The need to combine big data, master data and data in your data warehouse
- Matching big data with customer master data at scale
- Governing data in a data science environment

COURSE OUTLINE

07 INFORMATION AUDIT & PROTECTION – GOVERNING DATA ACROSS A DISTRIBUTED DATA LANDSCAPE

Over recent years we have seen many major brands suffer embarrassing publicity due to data security breaches that have damaged their brand and reduced customer confidence. With data now highly distributed and so many technologies in place that offer audit and security, many organisations end up with a piecemeal approach to information audit and protection. Policies are everywhere with no single view of the policies associated with securing data across the enterprise. The number of administrators involved is often difficult to determine and regulatory compliance is now demanding that data is protected and that organisations can prove this to their auditors. So how are organisations dealing with this problem? Are the same data privacy policies enforced everywhere? How is data access security co-ordinated across portals, processes, applications and data? Is anyone auditing privileged user activity? This module defines this problem, looks at the requirements needed for Enterprise Data Audit and Protection and then looks at what technologies are available to help you integrate this into your data strategy.

- What is Data Audit and Security and what is involved in managing it?
- Status check - Where are we in data audit, access security and protection today?
- What are the requirements for enterprise data audit, access security and protection?
- What needs to be considered when dealing with the data audit and security challenge?
- Automatic data discovery and the information catalog – a huge help in identifying sensitive data
- What about privileged users?
- Using a data management platform and information catalog to govern data across multiple data stores
- Securing and protecting data using tag-based policies in an information catalog
- What technologies are available to protect data and govern it? – Apache Knox, Cloudera Sentry, Dataguise, IBM (Watson Knowledge Catalog, Optim & Guardium), Immuta, Informatica Secure@Source, Imperva, Micro Focus, Okera, Privitar
- Can these technologies help in GDPR?How do they integrate with Data Governance programs?
- How to get started in securing, auditing and protecting your data

INSTRUCTOR

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 35 years of IT experience, Mike has consulted for dozens of companies. He has spoken at events all over the world and written numerous articles. Mike is Chairman of Big Data LDN – the fastest growing Big Data conference in Europe, and chairman of the CDO Exchange. 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 Analytics, Big Data, Data Governance & MDM, Data Warehouse Modernisation and Data Lake operations.

COURSE DATES

14-15 OCTOBER 2021 VIRTUAL LIVE CET TIME

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.