Designing, Operating and Managing an Enterprise Data Lake

Governing your Information across Hadoop, Cloud Storage, Data Warehouses, MDM & NoSQL Data Stores

Two day seminar by Mike Ferguson

///AdeptEvents

- Design, build, manage and operate a distributed or centralised data lake
- Information catalog and Data-as-a-Service
- How to organise data in a distributed data environment to overcome complexity and chaos
- Defining a strategy for producing trusted data services in a distributed environment of multiple data stores and data sources
- Technologies and implementation methodologies to get your data under control

VENUE
Area Utrecht/Hilversum, The Netherlands

TIME
9:30 – 17:00 hours

REGISTRATION
www.adeptevents.nl

Designing, Operating and Managing an Enterprise Data Lake

Governing your Information across Hadoop, Cloud Storage, Data Warehouses, MDM & NoSQL Data Stores

Most organisations today are dealing with multiple silos of information. These include cloud and on-premises based transaction processing systems, multiple data warehouses, data marts, reference data management (RDM) systems, master data management (MDM) systems, content management (ECM) systems and more recently Big Data NoSQL platforms such as Hadoop and other NoSQL databases. In addition the number of data sources is increasing dramatically especially from outside the enterprise. Given this situation it is not surprising that many companies have ended up managing information in silos with different tools being used to prepare and manage data across these systems with varying degres of governance. In addition, it is not only IT that is now integrating data. Business users are also getting involved with new self-service data preparation tools. The question is, is this the only way to manage data? Is there another level that we can get reach to allow us to more easily manage and govern data across an increasingly complex data landscape consisting of multiple data stores?

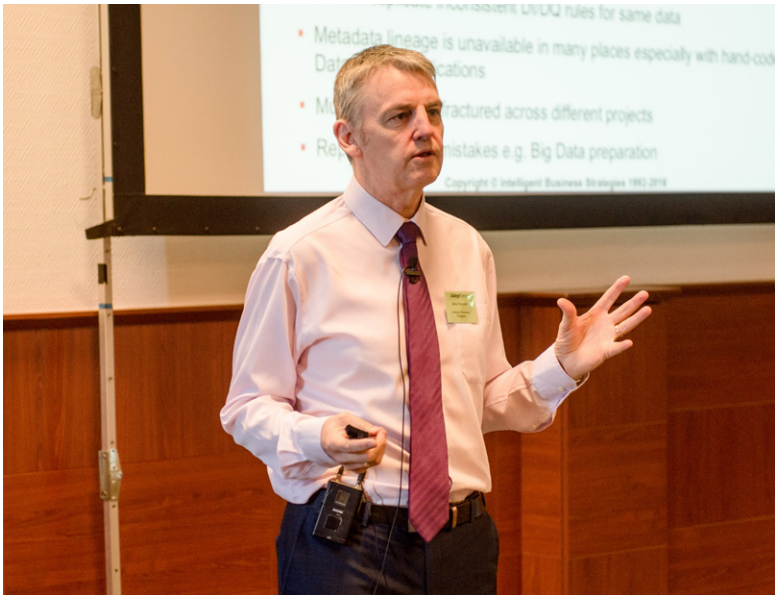
This 2-day seminar looks at the challenges faced by companies trying to deal with an exploding number of data sources, collecting data in multiple data stores (cloud and on-premises), multiple analytical systems and at the requirements to be able to define, govern, manage and share trusted high quality information in a distributed and hybrid computing environment. It also explores a new approach of how IT data architects, business users and IT developers can collaborate together in building and managing a logical data lake to get control of your data. This includes data ingestion, automated data discovery, data profiling and tagging and publishing data in an information catalog. It also involves refining raw data to produce enterprise data services that can be published in a catalog available for consumption across your company. We also introduce multiple data lake configurations including a centralised data lake and a 'logical' distributed data lake as well as execution of jobs and governance across multiple data stores. It emphasises the need for a common collaborative approach to governing and managing data of all types.

Learning objectives

Attendees 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 organise 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
- 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.



Target Audience

This seminar is intended for business data analysts doing self-service data integration, data architects, chief data officers, master data management professionals, content 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 seminar is not only for 'Fortune 500 scale companies' but for any organisation that has to deal with Big Data, multiple data stores and multiple data sources. It 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.

MIKE FERGUSON

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specialises in business intelligence / analytics, data management, big data and enterprise business integration. With over 35 years of IT experience, Mike 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 Big Data, Predictive and Advanced Analytics, Fast Data and Real-time Analytics, Enterprise Data Governance, Master Data Management, Data Virtualisation, Building an Enterprise Data Lake and Enterprise Architecture.

MODULE 1: STRATEGY & PLANNING

This session 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 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 inclusive approach to governing and managing data
- Introducing the data lake and data refinery
- Data lake configurations – what are the options?
  - Centralised, distributed or logical datalakes
- Information Supply Chain use cases – establishing a multi-purpose data lake
- The rising importance of an Information catalog
- Key technology components in a data lake
- 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

MODULE 2: METHODOLOGY & TECHNOLOGIES

Having understood strategy, this session looks at why information producers need to make use of multiple methodologies in a data lake information supply chain to product 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 structured data governance
- Why the methodology has to change for semi-structured and unstructured data
- Methodologies for structured Vs multi-structured data

MODULE 3: DATA STANDARDISATION, THE BUSINESS GLOSSARY AND THE INFORMATION CATALOG

This session 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, e.g Alteryx Connect Glossary, ASG, Collibra, Global IDs, Informatica, IBM Information Governance Catalog, Microsoft Azure Data Catalog Business Glossary, SAP Information Steward Metapedia, SAS Business Data Network, TIBCO Information Server
- 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?
- Information catalog technologies, e.g. Alation, Alteryx Connect, Amazon Glue, Apache Atlas, Collibra Catalog, IBM Information Governance Catalog & Watson Knowledge Catalog, Informatica EIC & Live Data Map, Microsoft Azure Data Catalog, Podium Data, Waterline Data, Zaloni Mica
- Information catalog capabilities

MODULE 4: ORGANISING AND OPERATING THE DATA LAKE

This session 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 distributed data lake
- Creating zones to manage data
- New requirements for managing data in centralised and distributed data lakes
- Creating collaborative data lake projects
- Hadoop 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 systematic disparate data and data relationship discovery using Information catalog software
- Using domains and machine learning to automate and speed up data discovery and tagging
- Alation, IBM Watson Knowledge Catalog, Informatica CLAIRE, Silwood, Waterline Data Smart Data Catalog
- Automated profiling and tagging and cataloguing of data
- Automated data mapping
- The data classification and policy definition processes
- Manual and automated data classification to enable governance
- Using tag based policies to govern data

MODULE 5: THE DATA REFINERY PROCESS

This session looks at the process of refining data to get produce trusted information

- What is a data refinery?
- Key requirements for refining data
- 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, Paxata, Tableau (Project Maestro), Tamr, Talend, Trifacta
- Automated data profiling using analytics in data preparation tools
- Executing data refinery jobs in a distributed data lake using Apache Beam to run anywhere
- Approaches to integrating IT ETL and self-service data preparation
- Apache Atlas Open Metadata & Governance
- Joined up analytical processing from ETL to analytical workflows
- Publishing data and data integration jobs to the information catalog
- Mapping produced data of value into your DW and business vocabulary
- Data provisioning – provisioning consistent information into data warehouses, MDM systems, NoSQL DBMSs and transaction systems
- Provisioning consistent refined data using data virtualisation, a logical data warehouse and on-demand information services
- Governing the provisioning process using rules-based metadata
- Consistent data management across cloud and on-premise systems

MODULE 6: REFINING BIG DATA & DATA FOR DATA WAREHOUSES

This session looks at how the data refining processes can be applied to managing, governing and provisioning data in a Big Data analytical ecosystem and in traditional data warehouses. 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

MODULE 7: INFORMATION AUDIT & PROTECTION – THE FORGOTTON SIDE OF DATA GOVERNANCE

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 session 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 you 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, Hortonworks Ranger, IBM (Watson Data Platform, Knowledge Catalog, Optim & Guardium), Imperva, Informatica Secure@Source, Micro Focus, Privitar
- Can these technologies help in GDPR?
- How do they integrate with Data Governance programs?
- How to get started in securing, auditing and protecting you data

Information

DATE AND TIME
The workshop will take place once or twice a year with the exact date and time available on our website. The programme starts at 9:30 am and ends at 5:15 pm on both days. Registration commences at 8.30 am and we recommend that you arrive early.

VENUE
Adept Events works with several accomodations in the area of Utrecht/Hilversum. Once the accomodation is confirmed, the information will be visible on the website. Please check the website prior to your departure.

HOW TO REGISTER
Please register online at **www.adeptevents.nl**. For registering by print, please scan the completed registration form and send this or your Purchase Order to **customerservice@adeptevents.nl**. We will confirm your registration and invoice your company by e-mail therefore please do not omit your e-mail address when registering.

REGISTRATION FEE
Taking part in this two-day workshop will only cost 1305 Euro when registering 30 days beforehand and 1450 Euro per person afterwards (excl. 21% Dutch VAT). This also covers documentation, lunch, tea/coffee.

Note: This seminar may also be offered 'Online' or as 'Face-to-face and live streaming'. In that situation, the prices for attending online differ from the prices listed here. On the **Registration Fee** page of our website you will always find the current rates for all available formats of this seminar.

Members of the DAMA are eligable for 10 percent discount on the registration fee.

In completing your registration form you declare that you agree with our Terms and Conditions.

TEAM DISCOUNTS
Discounts are available for group bookings of two or more delegates representing the same organization made at the same time. Ten percent off when registering 2 - 3 delegates and fifteen percent off for all delegates when registering four or more delegates (all delegates must be listed on the same invoice). This cannot be used in conjunction with other discounts. All prices are VAT excluded.

PAYMENT
Full payment is due prior to the workshop. An invoice will be sent to you containing our full bank details including BIC and IBAN. Your payment should always include the invoice number as well as the name of your company and the delegate name. For Credit Card payment please contact our office by e-mail mentioning your phone number so that we can obtain your credit card information.

CANCELLATION POLICY
Cancellations must be received in writing at least three weeks before the commencement of the workshop and will be subject to a € 75,– administration fee. It is regretted that cancellations received within three weeks of the workshop date will be liable for the full workshop fee. Substitutions can be made at any time and at no extra charge.

CANCELLATION LIABILITY
In the unlikely event of cancellation of the workshop for any reason, Adept Events' liability is limited to the return of the registration fee only. Adept Events will not reimburse delegates for any travel or hotel cancellation fees or penalties. It may be necessary, for reasons beyond the control of Adept Events, to change the content, timings, speakers, date and venue of the workshop.

MORE INFORMATION
+31(0)172 742680

http://www.adeptevents.nl/edl-en

seminars@adeptevents.nl

@AdeptEventsNL / **https://twitter.com/AdeptEventsNL**

http://www.linkedin.com/company/adept-events

https://www.facebook.com/AdeptEventsNL

Visit our Business Intelligence and Data Warehousing website www.biplatform.nl and download the App

Visit our website on Software Engineering, www.release.nl and download the App

IN-HOUSE TRAINING
Would you like to run this course in-company for a group of persons? We can provide a quote for running an in-house course, if you offer the following details. Estimated number of delegates, location (town, country), number of days required (if different from the public course) and the preferred date/period (month). Please find more info on the **In-house page on our website**.