# Designing, Operating and Managing an Enterprise Data Lake

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

**Tweedaags seminar met Mike Ferguson**

AdeptEvents

- Een data lake ontwerpen, bouwen, beheren en succesvol inzetten
- Het belang van een information catalog voor het leveren van Data-as-a-Service
- Voorkom overbodige complexiteit en chaos in een gedistribueerde data omgeving
- Een strategie voor het creëren van betrouwbare data-services in een gedistribueerde omgeving van verschillende data stores en databronnen
- Technologieën en implementatiemethoden om grip te krijgen op uw data

**LOCATIE**
Omgeving Utrecht/Hilversum

**TIJD**
van 9:30 tot 17:00 uur

**REGISTRATIE**
www.adeptevents.nl

# Designing, Operating and Managing an Enterprise Data Lake

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

Veel organisaties gebruiken tegenwoordig meerdere data stores voor hun informatie. Data wordt opgeslagen in de cloud en in on-premise gebaseerde transactieverwerkingssystemen, meerdere datawarehouses, data-marts, reference data management (RDM) systemen, master data management (MDM) systemen, content management (ECM) systemen en recentelijk ook Big Data NoSQL-platformen zoals Hadoop en andere NoSQL-databases. Tegelijkertijd stijgt het aantal databronnen dat gebruikt wordt enorm, voornamelijk van buiten de organisatie. Het gevolg hiervan is dat men verschillende tools moet gebruiken voor het beheren van de informatie binnen en tussen de verschillende systemen met verschillende governance niveaus. Daarnaast is de IT-afdeling tegenwoordig niet meer de enige die aan data-integratie doet. Zakelijke gebruikers maken ook steeds meer gebruik van self-service data wrangling tools. De vraag is nu, is dit de enige manier om data te beheren? Kunnen we het beheer en governance van data niet naar een ander niveau tillen in een steeds complexer data-landschap?

In dit tweedaagse seminar behandelt Mike Ferguson de uitdagingen van organisaties die te maken hebben met een explosief groeiende hoeveelheid databronnen, het opslaan van gegevens in verscheidene data stores (in de Cloud en on-premise) en gebruik van meerdere Analytics tools. Wat heeft u nodig om in staat te zijn betrouwbare en hoogwaardige informatie te definiëren, beheren en te delen in een gedistribueerde en hybride IT-omgeving? Mike behandelt een nieuwe aanpak om data-architecten, zakelijke gebruikers en ontwikkelaars te laten samenwerken om een enterprise data lake te bouwen en beheren zodat u weer grip krijgt op uw gegevens. Dit omvat mede het opzetten van een data refinery en een information catalog voor het produceren en publiceren van enterprise data services voor gebruik binnen uw organisatie alsmede het opzetten van gedistribueerde uitvoering en governance over meerdere data stores.

## Leerdoelen

In dit seminar leren deelnemers:

- Een strategie te definiëren voor het creëren van betrouwbare data services in een gedistribueerde omgeving van verschillende data stores en databronnen
- Data in een gedistribueerde data-omgeving dusdanig te organiseren om overbodige complexiteit en chaos te voorkomen
- Hoe u een data lake ontwerpt, bouwt, beheert en succesvol inzet
- Het belang van een information catalog voor het leveren van Data-as-a-Service
- Hoe datastandaardisatie en business glossaries kunnen helpen bij het definiëren van de data om zeker te zijn dat het begrepen wordt

- Een operationeel model voor effectieve gedistribueerde informatie governance
- Welke technologieën en implementatiemethodologieën nodig zijn om grip te krijgen op uw data
- Het toepassen van methoden om master en reference data, big data, datawarehouse data en ongestructureerde data makkelijk te beheren ongeacht of het in de cloud opgeslagen is of on-premise.

## Bestemd voor ú!

Dit seminar is ontworpen voor business data analisten die zich bezighouden met self-service data integratie, data architecten, chief data officers, master data management professionals, content management professionals, database administrators, big data professionals en data-integratie developers. Het is zeker ook relevant voor compliance managers die verantwoordelijk zijn voor datamanagement en zich bezighouden met bijvoorbeeld metadata management, data-integratie, datakwaliteit, master data management of enterprise content management. Het seminar is niet alleen geschikt voor 'Fortune 500'-bedrijven maar voor elke organisatie die te maken heeft met Big Data, meerdere

dataopslaglocaties en databronnen. Er wordt vanuit gegaan dat u de basisprincipes van datamanagement kent en kennis heeft van begrippen als datamigratie, datareplicatie, metadata, datawarehousing, datamodellering en data cleansing.

## MIKE FERGUSON

Mike Ferguson is oprichter van Intelligent Business Strategies Ltd. en als analist en consultant gespecialiseerd in business intelligence, big data, data management en enterprise business integration. Hij kan bogen op meer dan 30 jaar ervaring in IT, ondermeer op gebied van BI en Corporate Performance Management, Data Management en Big Data Analytics (Hadoop, MapReduce, Hive, Graph DBMSs). Mike opereert afwisselend op bestuursniveau, IT management niveau en ook gespecialiseerde technische IT niveau's voor de terreinen BI, corporate performance management strategie, technologie- en toolselectie, enterprise architectuur, MDM and data-integratie. Hij is een veelgevraagd spreker op internationale conferenties en heeft veelvuldig artikelen gepubliceerd in de vakbladen en via weblogs. Eerder was Mike partner en mede-oprichter van Codd and Date Europe Limited, Chief Architect bij NCR voor het Teradata DBMS en Europees directeur van Database Associates. Hij verzorgt regelmatig seminars en workshops met als onderwerpen Big Data, Operational BI, Enterprise Data Governance, Master Data Management, Data Integration and Enterprise Architecture.

# Programma-overzicht

## 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

# Informatie

## DATUM EN TIJD

Het seminar vindt periodiek plaats in het voorjaar en/of najaar. De exacte data en aanvangstijden vindt u op onze website. Het programma start om 9:30 uur en duurt tot 17:00 uur. Registratie is mogelijk vanaf 8:30 uur.

## PLAATS

Adept Events werkt met verscheidene accommodaties in de omgeving van Utrecht/Hilversum. Zodra de definitieve locatie bekend is, publiceren wij dit op de website. Controleer dit voor uw vertrek.

## AANMELDEN

Aanmelden kan via ons online inschrijfformulier op **www.adeptevents.nl**. Geeft u de voorkeur aan schriftelijk aanmelden? Stuur de PDF van uw aanmelding of inkoopopdracht naar **seminars@adeptevents.nl**.Vermeld altijd duidelijk het e-mailadres van de deelnemer(s) alsmede dat van de crediteurenafdeling. Na ontvangst van uw aanmelding krijgt u de bevestiging en factuur per e-mail toegestuurd.

## KOSTEN

Deelname aan dit seminar kost € 1.305,– per persoon indien u zich uiterlijk 30 dagen voor aanvang registreert en daarna € 1.450,– (exclusief BTW)*. Documentatie, maaltijd en koffie zijn inbegrepen. Leden van DAMA ontvangen 10% korting op de deelnemersprijs. Deze en andere lidmaatschapskortingen kunnen niet worden gecombineerd. Werkt u bij een gemeente of provincie? Dan kunt u BTW terugvorderen via het BTW compensatiefonds.

## AANTREKKELIJKE KORTINGEN

Meldt u tegelijkertijd meerdere personen van één bedrijf aan voor hetzelfde evenement, dan geldt al vanaf de tweede deelnemer een korting van 10% per deelnemer. Vanaf vier deelnemers ontvangen álle deelnemers 15% korting (de deelnemers dienen op dezelfde factuur te staan)*.

* Prijzen of kortingen in deze pdf brochure kunnen (tijdelijk) afwijken van de website. In dat geval prevaleert altijd de informatie op de website.

## ANNULEREN

Annuleren dient schriftelijk te geschieden. U kunt annuleren tot drie weken voor het evenement plaatsvindt. Er wordt echter wel € 75,– (excl. BTW) administratiekosten in rekening gebracht. Annuleren is niet meer mogelijk vanaf drie weken voordat het evenement plaatsvindt. Vervanging door een ander dan de aangemelde persoon is te allen tijde mogelijk.

## MEER INFORMATIE

+31(0)172 742680

**http://www.adeptevents.nl**

**seminars@adeptevents.nl**

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

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

**https://www.facebook.com/AdeptEventsNL**

**https://google.com/+AdeptEventsNL**

**Bezoek onze Business Intelligence en Datawarehousing website www.biplatform.nl en download de App**

**Bezoek ook onze website over Software Engineering in al zijn facetten: www.release.nl en download de App**

## IN-HOUSE SESSIES VOOR UW MEDEWERKERS

Wilt u dit seminar binnen uw organisatie aanbieden als in-house sessie voor een groep medewerkers? Bel of stuur een e-mail via ons contactformulier. Op de **In-house pagina** van onze site vindt u meer informatie over de mogelijkheden van In-house seminars en workshops.