MIKE FERGUSON

Building an Enterprise Data Lake

- The Route to Trusted Enterprise Data as a Service

info@intelligentbusiness.biz
www.intelligentbusiness.biz

OVERVIEW

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 degrees 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 wrangling 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?

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 an enterprise data lake to get control of your data. This includes data ingestion, 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 and governance across multiple data stores. It emphasises the need for a common collaborative process and common approach to governing and managing data of all types.

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, small 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.

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 distributed or centralised data lake within their organisation
- The critical importance of an information catalog for delivering data-as-a-service
- How data standardisation and business glossaries can help define the data to make sure it is understood
- An operating model for effective distributed information governance
- What technologies they need and implementation methodologies 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.

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 wrangling 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 M'gmt
- 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?
- The rising importance of an Information catalog
- Key roles and responsibilities - getting the operating model right
- Types of policy to govern data
- Formalising governance processes, e.g. the dispute resolution process
- Integrating a data lake into your enterprise analytical architecture

MODULE 2: METHODOLOGY & TECHNOLOGIES

Having understood strategy, this session looks at multiple methodologies and the technologies needed to help apply it to your structured and multi-structured data to bring it under control. It also looks at how platforms like Hadoop and common data services provide the foundation to manage information across the enterprise

- A best practice step-by-step methodology structured data governance
- Why the methodology has to change for semi-structured and unstructured data
- Technology components in the new world of distributed data
- Hadoop as a data staging area
- Why Hadoop is not enough
- Data management technology platforms, e.g. Actian, Global IDs, IBM, Informatica, Oracle, SAP, SAS, Talend
- Self-service data wrangling tools, e.g. Alteryx, Paxata, Trifacta, Tamr, ClearStory Data
- Self-service data integration in BI tools
- Implementation options
  - Centralised, distributed or federated data lakes
  - Self-service DI – the need for collaborative development, data management and data governance
  - Dealing with data on-premise and on the cloud
  - Common data services for service-oriented data management

MODULE 3: DATA STANDARDISATION & THE BUSINESS GLOSSARY

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.

- Semantic data standardisation using a shared business vocabulary within an information catalog
- SBV vs. taxonomy vs. ontology
- The role of a SBV in MDM, RDM, SOA, DW and data virtualisation
- How does an SBV apply to data in a Hadoop data lake?

- Approaches to creating an SBV
- Business glossary products
  - ASG, Cisco, Collibra, Global IDs, Informatica, IBM Information Governance Catalog, SAP Information Steward Metapedia, SAS Business Data Network
- Planning for a business glossary Organising data definitions in a business glossary
- Business involvement in SBV creation
- Using governance processes in data standardisation

MODULE 4: ORGANISING 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
- Zoning the data lake
  - Data ingestion zones, approved raw data zones, data exploration zones, data archive zones, trusted refined data zones, internal data marketplace
- 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
- Beyond structured data - from business glossary to information catalog
- Information catalog technologies e.g. Waterline Data, Alation, Amazon Glue, Apache Atlas, Collibra Catalog, IBM Information Governance Catalog, Informatica Live Data Map, Microsoft Azure Data Catalog, Podium Data, Semanta, Waterline Data, Zaloni Mica
- The power of a graph database for storing metadata – dynamic tracking of data and data relationships in real-time
- The data ingestion process
- Tools and techniques for data ingestion
- Using domains and machine learning to speed up auto tagging

MODULE 5: THE DATA REFINERY PROCESS

This session looks at the process of discovering where your data is and how to refine it to get it under control

- Implementing systematic disparate data and data relationship discovery
- Data discovery tools Global IDs, IBM Watson Data Platform, Informatica, Silwood, Waterline Data Smart Data Catalog
- Automated profiling and tagging of data
- Automated data classification and cataloguing to enable governance
- Automated data mapping
- Automated data profiling using analytics in data wrangling tools
- Generating data cleansing and integration jobs using common metadata
- Key approaches to scalable data integration using Apache Spark
- Self-service data Wrangling tools for Spark and Hadoop
- Executing data refinery jobs in a distributed data lake using Apache Beam to run anywhere
- Approaches to integrating IT ETL and self-service data wrangling
- Joined up analytical processing from ETL to analytical workflows
- Publishing data and data integration jobs to the information catalog
- Mapping discovered data of value into your DW and business vocabulary
- Data provisioning – provisioning consistent information into data warehouses, MDM systems, NoSQL DBMSs and transaction systems
- Achieving consistent data provisioning through re-usable data services
- Provisioning consistent refined data using data virtualisation 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?
- Securing and protecting Big data using tag based policies
- How can you use it for GDPR?
- What technologies are available to tackle this problem? – Apache Knox, Cloudera Sentry, Dataguise, Hortonworks Ranger, HP Enterprise, IBM Optim & Guardium, Imperva, Privitar
- How do they integrate with Data Governance programs?
- How to get started in securing, auditing and protecting you data

PRESENTER

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 architecture. 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 and Enterprise Data Lake and Enterprise Architecture.

INTELLIGENT BUSINESS STRATEGIES
Tel/Fax: (+44) 1625 520700
Email: info@intelligentbusiness.biz
http://www.intelligentbusiness.biz