Welcome!

Module 5

- Data Warehouse – Business Intelligence Concepts

"A collection of integrated, subject-oriented databases designed to support the DSS function where each unit of data is relevant to some moment in time..."

Inmon, Imhoff and Sousa, The Corporate Information Factory

"A copy of transaction data specifically structured for query and analysis."

Ralph Kimball, The Data Warehouse Toolkit

Introduction

About Me

- **Parwaz Dalvi**
  Sr. Architect / Consultant DW-BI & Database

  TOGAF 8 Certified (The Open Group Architecture Framework)

My Session For you

- **Data Warehouse Concepts**

Session‘s Objective

- Understand what Data Warehousing means
- Realize the Need, Advantages & Challenges in implementation of a DW Solution
- Understand Data Warehouse Architecture and its components
- Understand IBM Reference DW-BI Architecture
- Understand IBM’s IOD initiative and realize how DW-BI helps in achieving this objective
- Know the DW-BI Tools and Products, the trends in DW-BI
- Know your Growth Prospects in the DW-BI Arena within IBM

Course Content

Module	Content	Duration
1	Data Warehouse Evolution	
2	Data Warehouse Concepts	
3	Data Warehouse Architecture – Part 1 – GENERIC	
4	Data Modeling in DW-BI	
5	***Data Warehouse Architecture – Part 2 – SPECIFIC***	
6	DW-BI - IBM Reference Architecture & IOD	
7	DW-BI Tools and Products	
8	Trends in DW-BI	
9	Growth Path of DW-BI Professionals	

IBM Global Business Services

Course Title:

Module 5 : Data Warehouse Architecture - Part 2 - Specific

Module Objectives

- At the completion of this chapter you should be able to understand:
  - Data Warehouse Architecture - Types
    - Data Stores – Responsibilities in Data Warehouse
    - The Inmon Data Warehouse - Integration Hub
    - The Kimball Data Warehouse – Bus Integration
    - Architecture Diagrams - Type 1, Type 2, Type 3
  - ETL – Insights
  - Analytics & BI - Insight
  - Metadata – Insight
  - Master Data – An Introduction
  - Data Mining – An Introduction
  - Data Governance – An Introduction

Module 5: DW Architecture - Part 1- Specific : Agenda

- Topic 1. Data Stores – Responsibilities in Data Warehouse
- Topic 2. The Inmon Data Warehouse
- Topic 3. The Kimball Data Warehouse
- Topic 4. The Inmon versus Kimball Data Warehouse
- Topic 5. Data Warehouse – Architecture Types

Module 5: > Topic 1: Data Warehouse Architecture - Types

- Data Store - Responsibilities in DW
  - Intake
  - Integration
  - Distribution
  - Delivery
  - Access
- Inmon Model
  - Top Down Approach
  - Hub Integration
  - Corporate Information Factory
- Kimball Model –
  - Bottom Up Approach
  - Data Warehouse Bus Architecture
  - Bus Integration & Conformed Dimensions

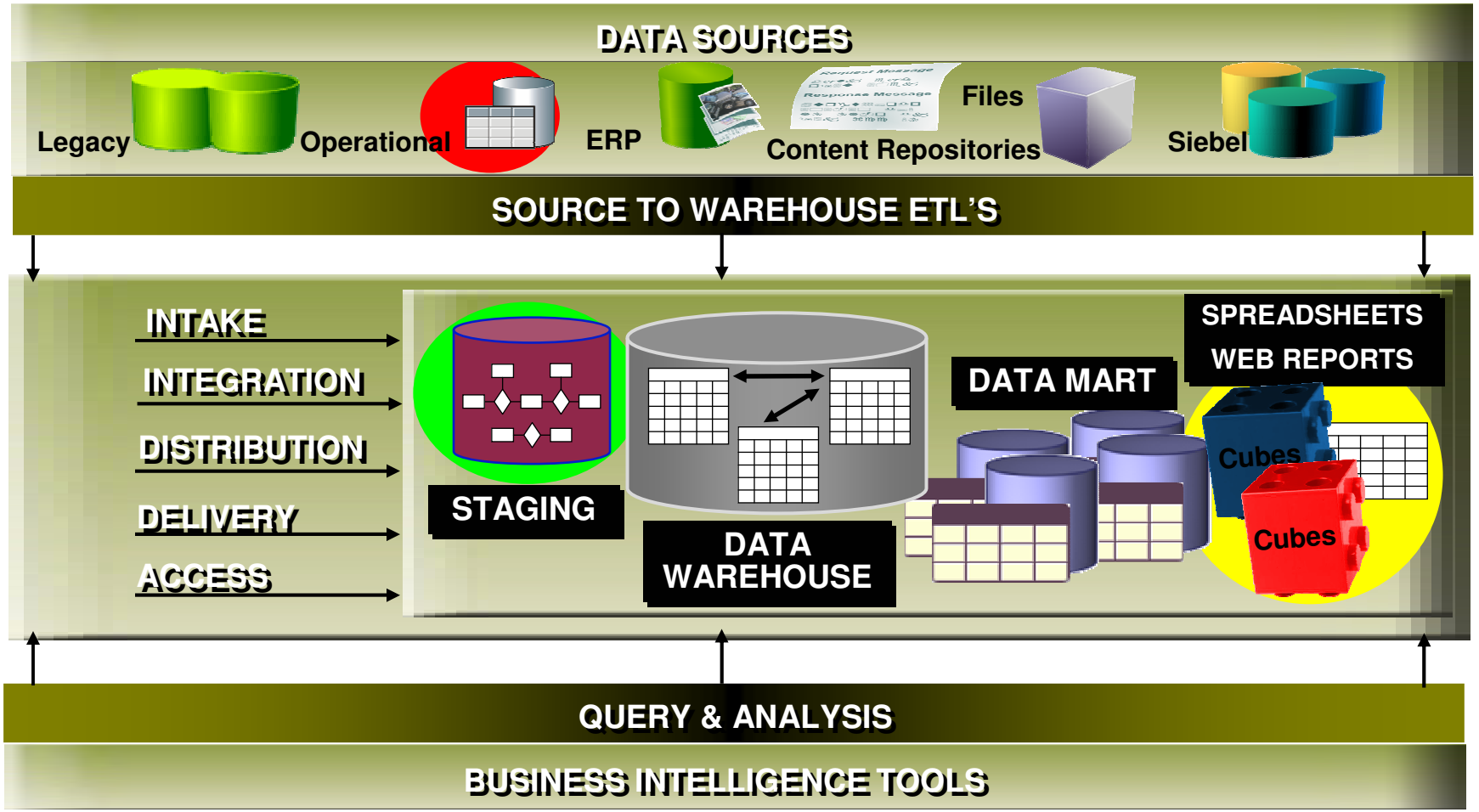
Module 5: > Topic 1: Data Warehouse Architecture - Types

Data Store – Responsibilities in DW

- *Five different Roles in DW Environment*
  - *Intake*
  - *Integration*
  - *Distribution*
  - *Delivery*
  - *Access*

Module 5: > Topic 1: Data Warehouse Architecture - Types

Data Store – Responsibilities in DW



Module 5: > Topic 1: Data Warehouse Architecture - Types

Data Store – Responsibilities in DW

- *Intake -*
  - *Data stores with intake responsibility receive data into warehousing environment. Data is acquired from multiple source systems, of varying technologies, at different frequencies, and into numerous warehousing files and/or tables. Further, the data typically requires many and diverse transformations. Most data is extracted from operational systems whose data is most certainly not all clean, error-free and complete. Data cleansing is commonly performed as part of the intake process to ensure completeness and correctness of data*
- *Integration -*
  - *Integration describes how the data fits together. The challenge for warehousing architect is to design and implement consistent and interconnected data that provides readily accessible, meaningful business information. Integration occurs at many levels – "the key level, the attribute level, the definition level, the structural level, and so forth …" (Data Warehouse Types, www.billinmon.com) Additional data cleansing processes, beyond those performed at intake, may be required to achieve desired levels of data integration*

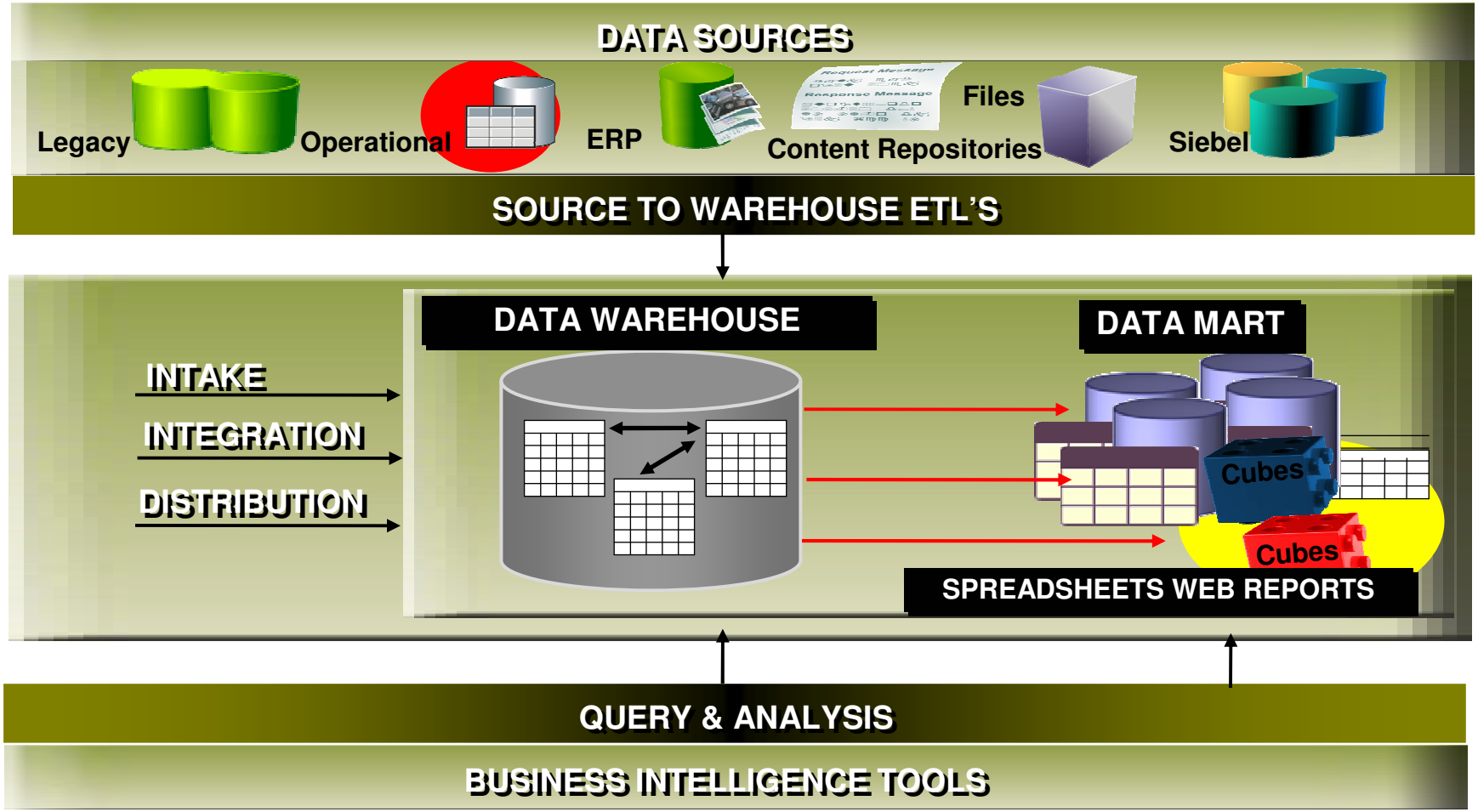
Module 5: > Topic 1: Data Warehouse Architecture - Types

Data Store – Responsibilities in DW

- *Distribution -*
  - *Data stores with distribution responsibility serve as long-term information assets with broad scope. Distribution is the progression of consistent data from such a data store to those data stores designed to address specific business needs for decision support and analysis*
- *Delivery -*
  - *Data stores with delivery responsibility combine data as "in business context" information structures to present to business units who need it. Delivery is facilitated by a host of technologies and related tools – data marts, data views, multidimensional cubes, web reports, spreadsheets, queries, etc*
- *Access –*
  - *Data stores with access responsibility are those that provide business retrieval of integrated data – typically the targets of a distribution process. Access-optimized data stores are biased toward easy of understanding and navigation by business users*

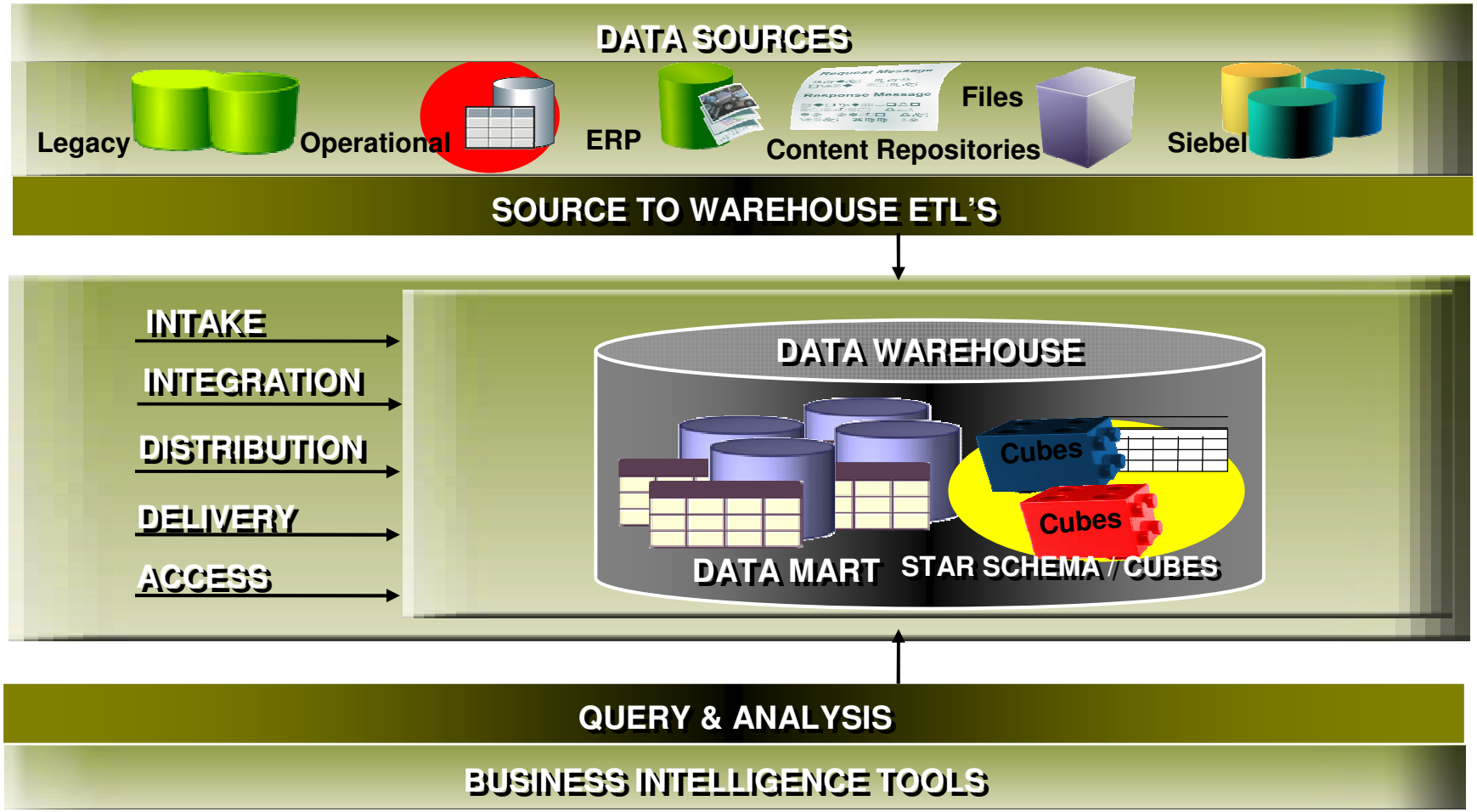
Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Inmon" Data Warehouse -



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Kimball" Data Warehouse -



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Inmon" vs. "Kimball" Data Warehouse – Data Store Responsibilities

	Inmon Warehouse	Kimball Warehouse
Intake	*Fills the Intake role but may be downstream for the staging area*	*Fills the intake role – downstream form backroom staging*
Integration	*Primary Integrated Data Store. Data stored at atomic level*	*Integration through standards & conformity of Data Marts*
Distribution	*Designed & optimized for distribution to Data Marts*	*Distribution is insignificant as Data Marts are considered a sub-set of Data Warehouse*
Access	*May provide limited data access to some "POWER" Users*	*Specially designed for Busienss Access & Ananlysis*
Delivery	*Not deisgned or intended for delivery*	*Supports delivery of Information to the Business*

Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Inmon" vs. "Kimball" Data Warehouse

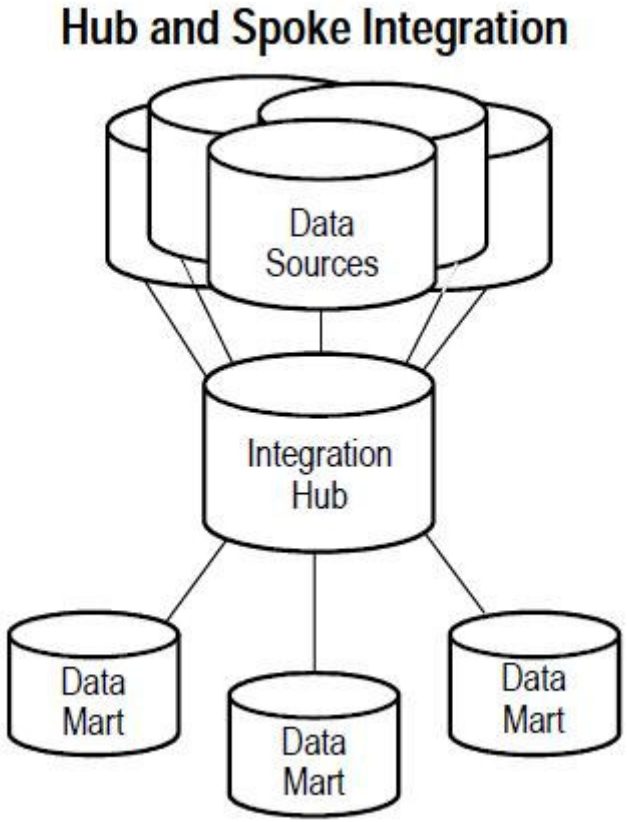
- *Inmon's Central Data Warehouse - Hub & Spoke Architecture*
  - *Inmon defines a data warehouse "A subject oriented, integrated, non-volatile, time-variant, collection of data organized to support management needs." (W. H. Inmon, Database Newsletter, July/August 1992) The intent of this definition is that the data warehouse serves as a single-source hub of integrated data upon which all downstream data stores are dependent. The Inmon data warehouse has roles of intake, integration, and distribution*

- *Kimball's Definition – Bus Architecture*
  - *Kimball defines the warehouse as "nothing more than the union of all the constituent data marts." (Ralph Kimball, et. al, The Data Warehouse Life Cycle Toolkit, Wiley Computer Publishing, 1998) This definition contradicts the concept of the data warehouse as a single-source hub. The Kimball data warehouse assumes all data store roles -- intake, integration, distribution, access, and delivery*

Module 5: > Topic 1: Data Warehouse Architecture - Types

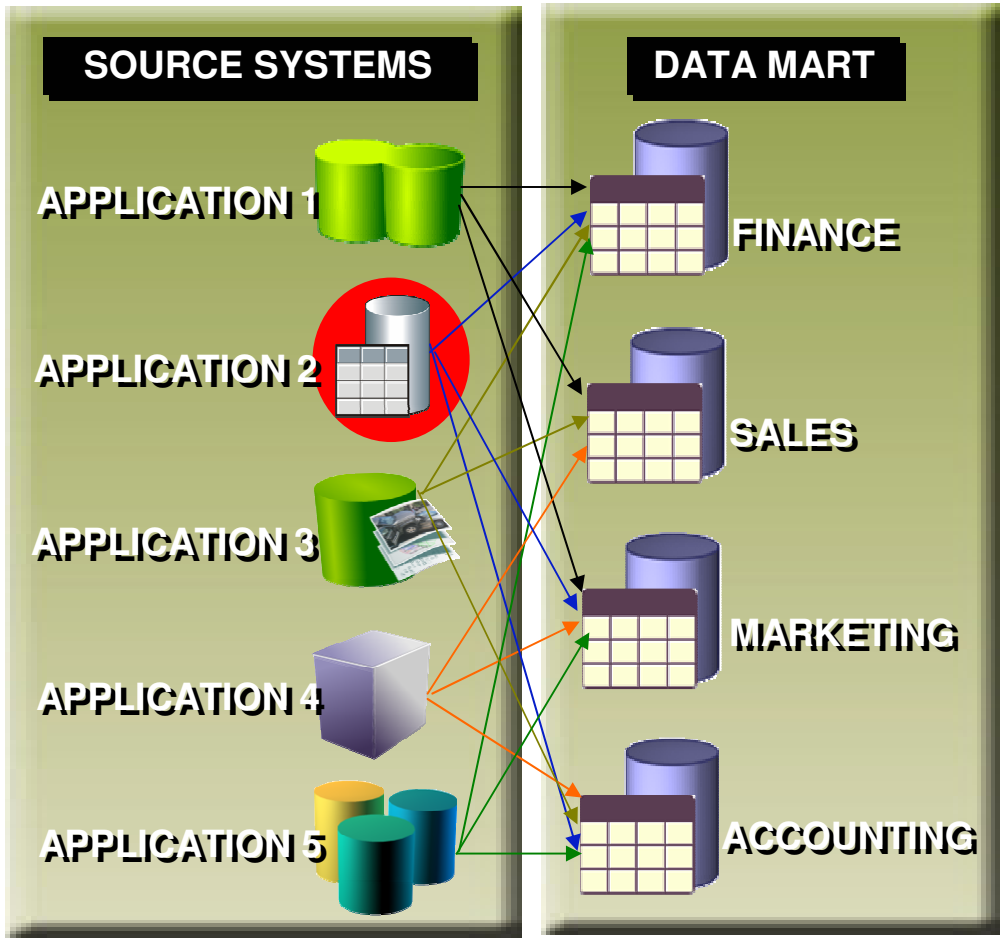
The "Inmon" Data Warehouse -

- *Hub & Spoke Architecture*
  - *The hub-and-spoke architecture provides a single integrated and consistent source of data from which data marts are populated. The warehouse structure is defined through enterprise modeling (top down methodology). The ETL processes acquire the data from the sources, transform the data in accordance with established enterprise-wide business rules, and load the hub data store (central data warehouse or persistent staging area). The strength of this architecture is enforced integration of data*



Module 5: > Topic 1: Data Warehouse Architecture - Types

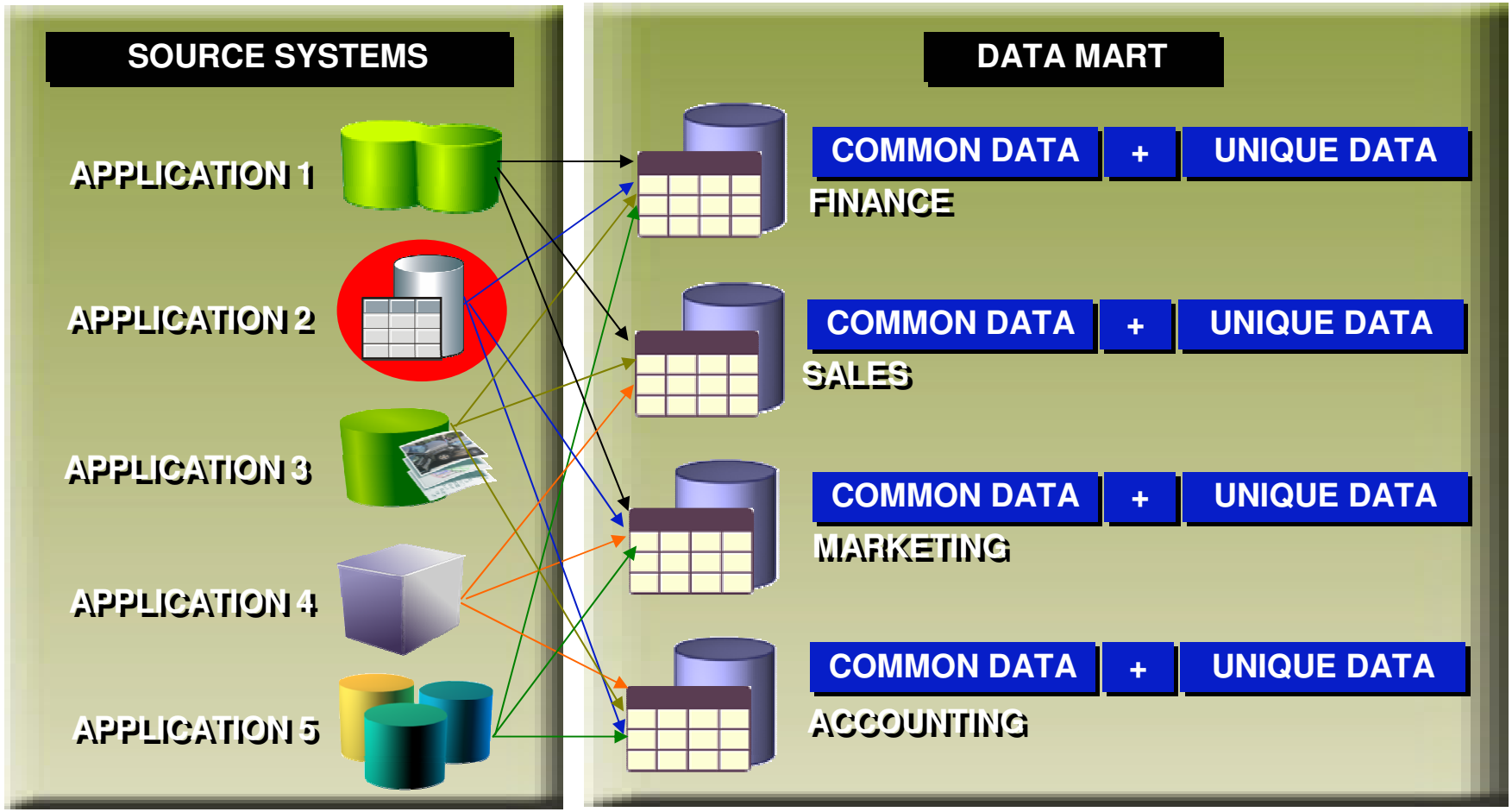
The "Inmon" Data Warehouse – Why the need for Integration Hub



- *Interface between applications & data marts is a complex one*
- *There are many programs that interface the two environments that must be build & maintained*
- *The amount of hardware required to move the data along all the interfaces is considerable*
- *If there are 'm' Applications & 'n' Data Marts, then 'm x n' Interfaces will have to be built , executed & maintained*
- *Data redundancy is possible. This may lead to data synchronization issues which in turn may lead to data quality issues. Overall impact business may loose faith in the data as there is no single, consistent, reliable & accurate source of data*

Module 5: > Topic 1: Data Warehouse Architecture - Types

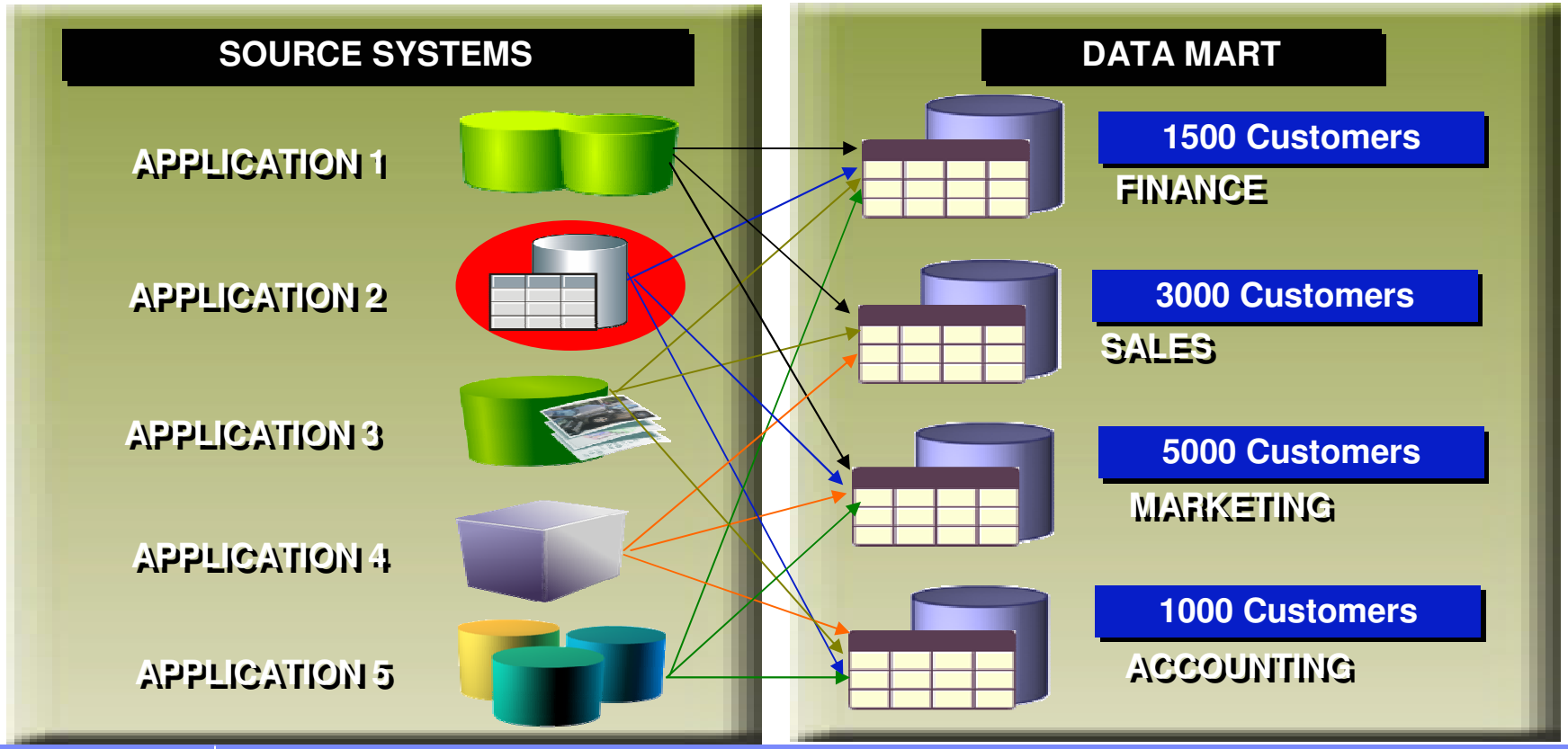
The "Inmon" Data Warehouse – Why the need for Integration Hub



Module 5: > Topic 1: Data Warehouse Architecture - Types

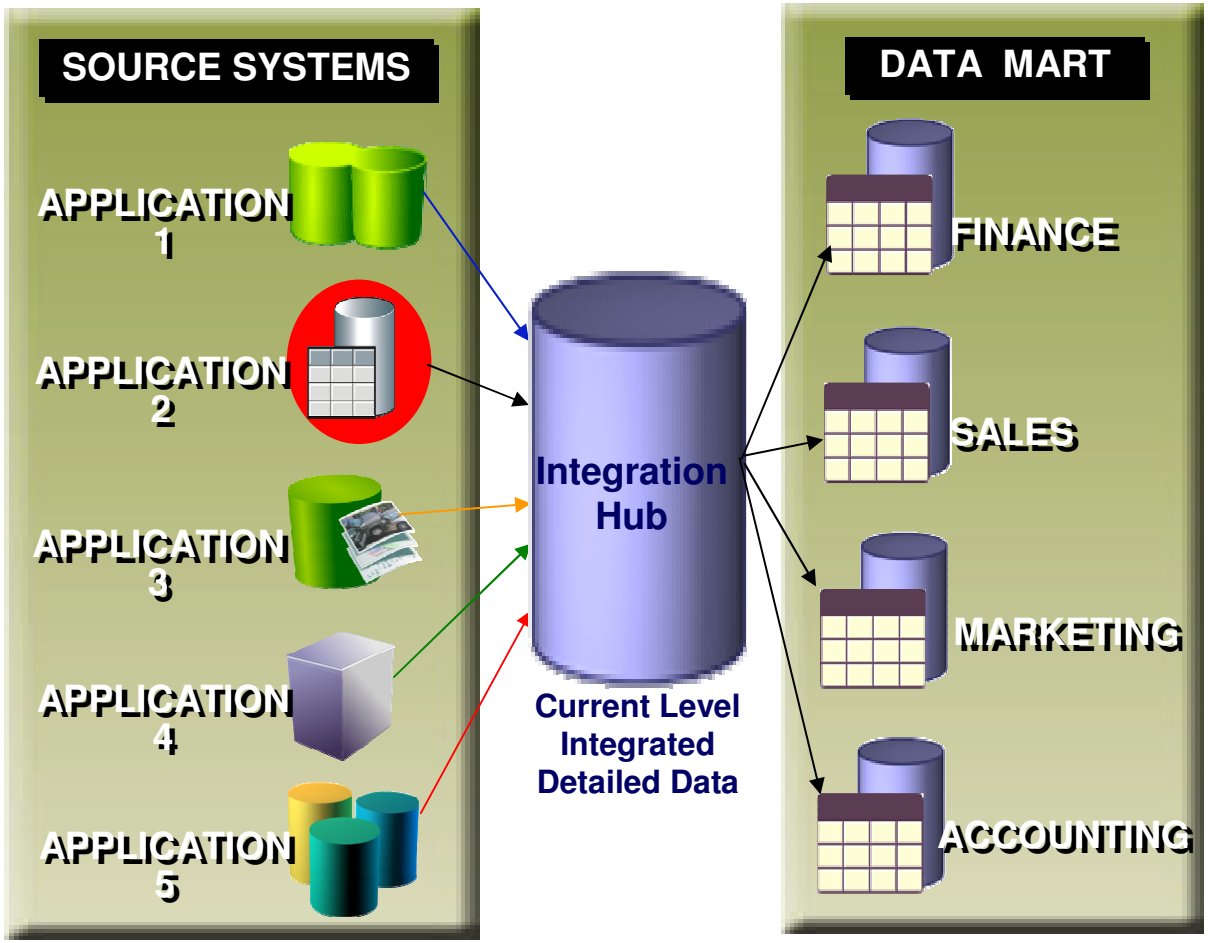
The "Inmon" Data Warehouse – Why the need for Integration Hub

How many Customers are there ?



Module 5: > Topic 1: Data Warehouse Architecture - Types

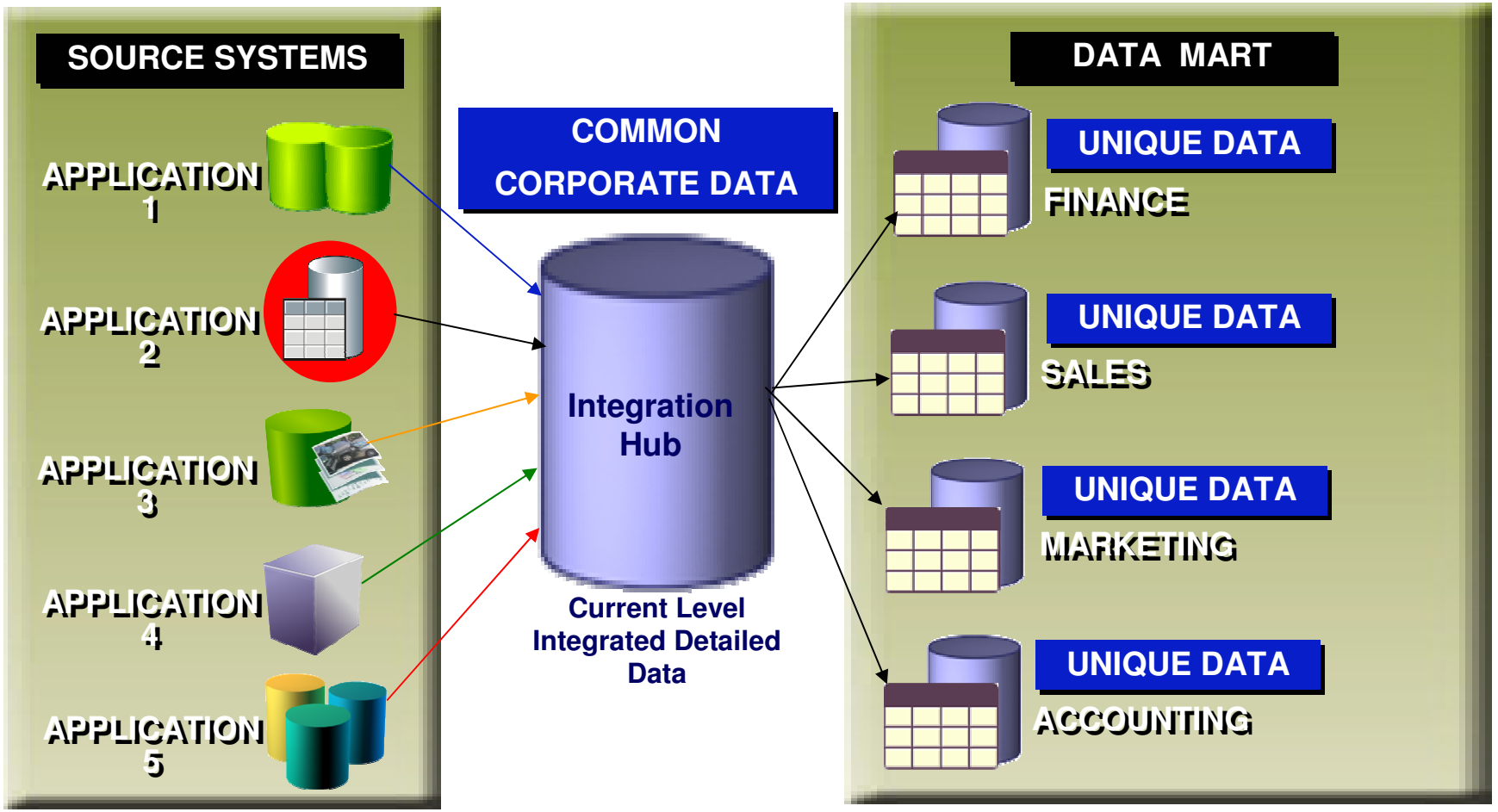
The "Inmon" Data Warehouse – Why the need for Integration Hub



- *There is orderly approach to building of interfaces*
- *If there are 'm' Applications & 'n' Data Marts, then only 'm + n' Interfaces will have to be built , executed & maintained*
- *Data consistency, reliability & accuracy is increased*
- *Common Corporate Data & Business process specific or departmental data is clearly demarcated, common data resides in the Integration Hub & Departmental Data resides in the Data Marts*

Module 5: > Topic 1: Data Warehouse Architecture - Types

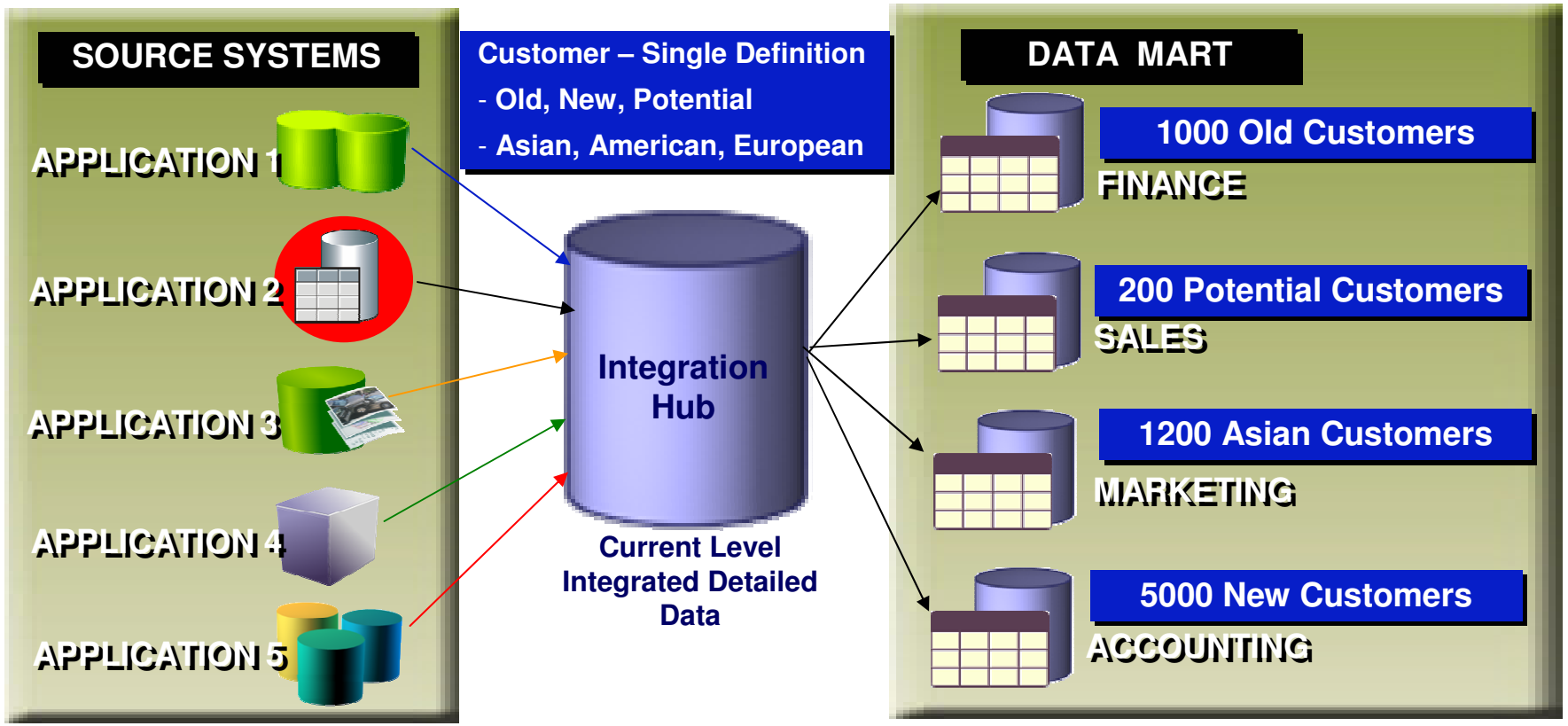
The "Inmon" Data Warehouse – Why the need for Integration Hub



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Inmon" Data Warehouse – Why the need for Integration Hub

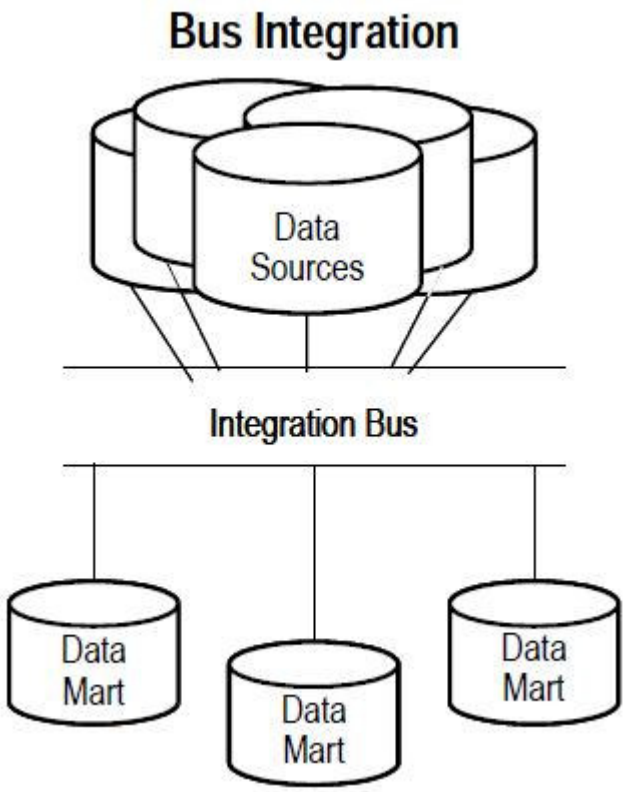
How many Customers are there ?



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Kimball" Data Warehouse - Bus Integration

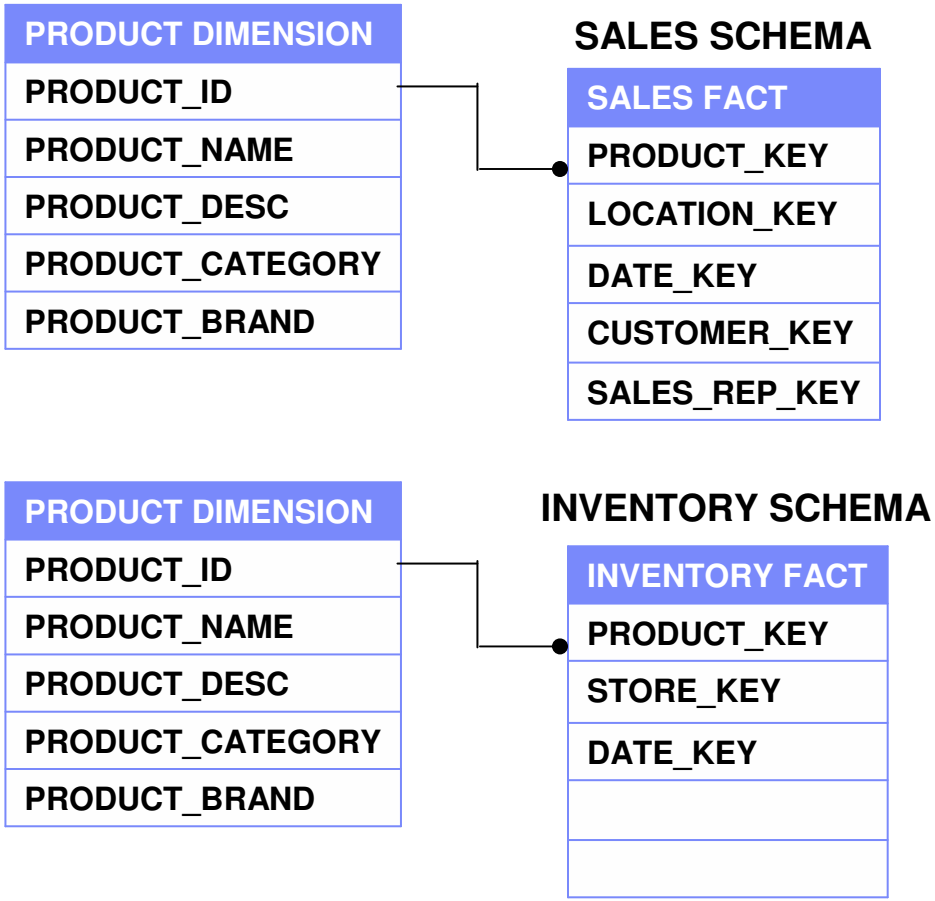
- *Bus Integration Architecture*
  - *The Bus Architecture relies on the development of conformed data marts populated directly from the operational sources or through a transient staging area. Data consistency from source-to-mart and mart-to-mart are achieved through applying conventions and standards (conformed facts and dimensions) as the data marts are populated. The strength of this architecture is consistency without the overhead of the central data warehouse*



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Kimball" Data Warehouse - Bus Integration

- *Conformed Dimension –*
  - *Dimension which retains the same business & technical nomenclature even if shared across Business processes.*
  - *Shared dimensions should conform.*
  - *Identical dimensions should have the same definitions, keys, labels & values*



Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Kimball" Data Warehouse - Bus Integration

BUSINESS PROCESS

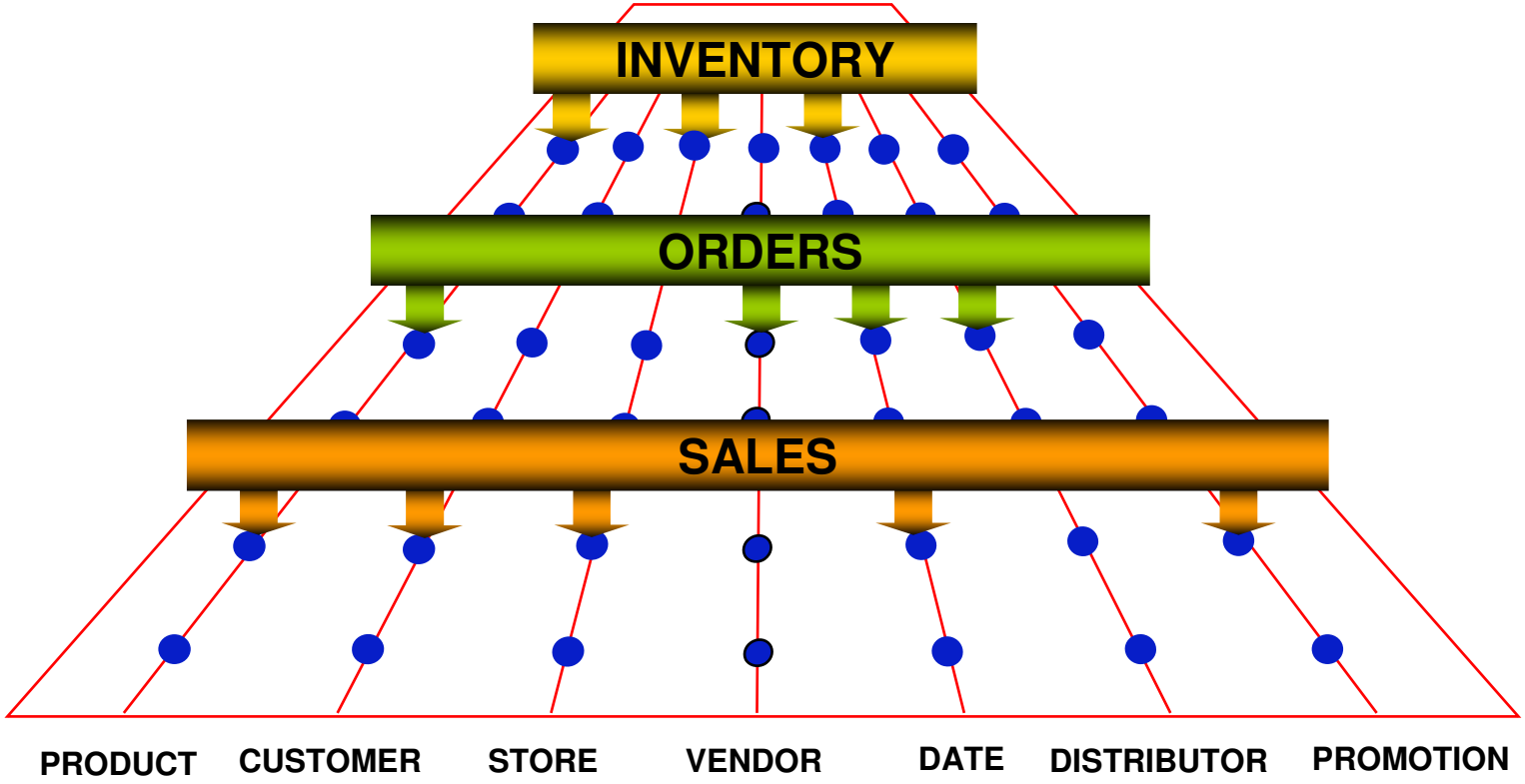
	PRODUCT	CUSTOMER	STORE	VENDOR	DATE	DISTRIBUTOR	PROMOTION
SALES	X	X	X		X		X
ORDERS	X			X	X	X	
INVENTORY	X		X		X		

CONFORMED DIMENSIONS

- *Business Analysts & Architects from different business streams arrive at single description of a dimension & its attributes. This results in a conformed dimension*
- *Conformed Dimensions are listed on the 'X' Axis, Business Process are listed on the 'Y' Axis*
- *The Matrix is completed by filling in an 'X' at intersection of a Business Process & Dimension, implies 'This dimension required for this business process'*
- *Once finalized, parrallel development of Data Marts can begin, each business process corresponding to a Data Mart*

Module 5: > Topic 1: Data Warehouse Architecture - Types

The "Kimball" Data Warehouse - Bus Integration



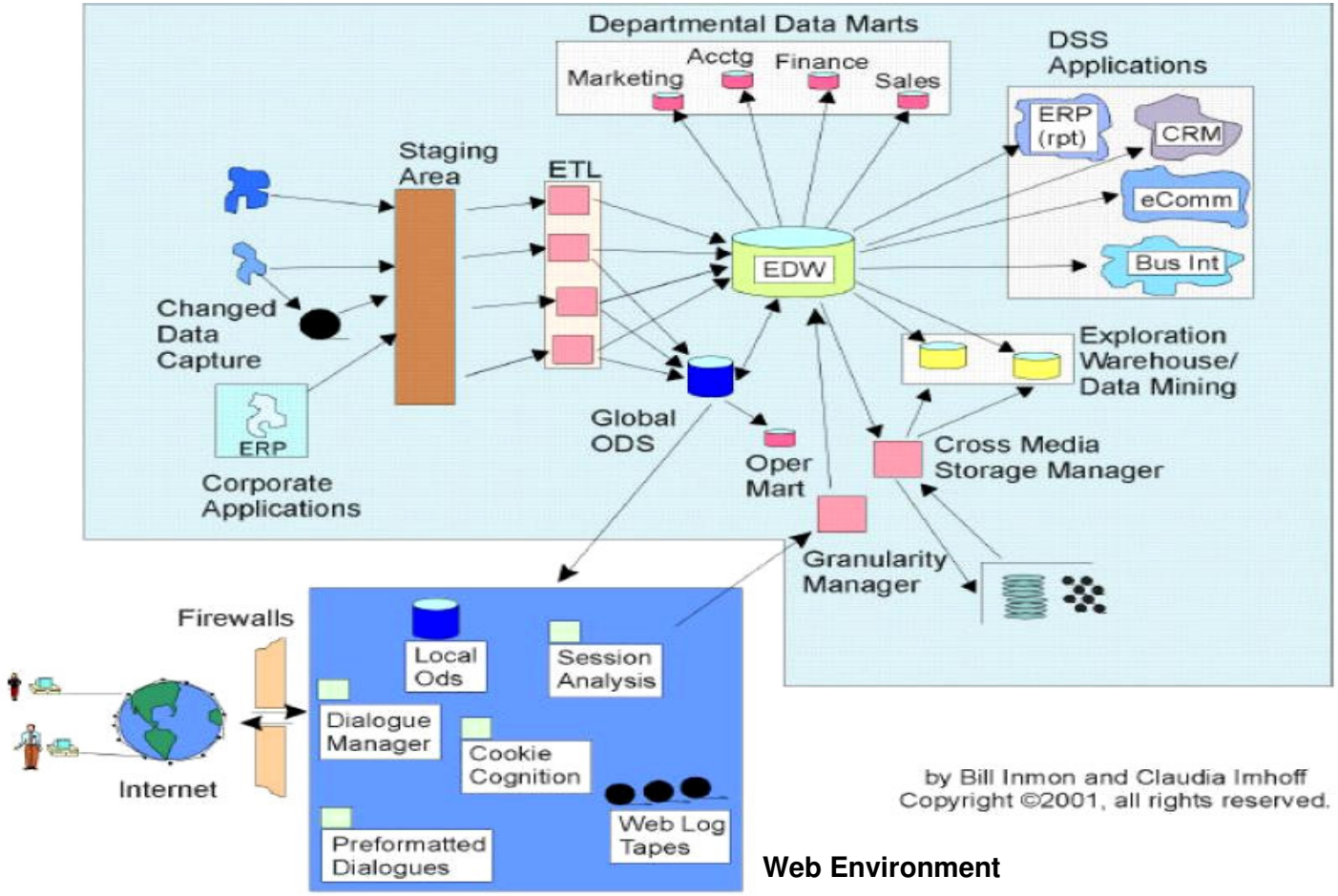
Module 5: > Topic 1: Data Warehouse Architecture - Types

Data Warehouse - Architecture Types -

- *There are 3 schools of thought related to DW Architecture*
  - *Type 1 : ER Model for EDW & Star Schema for Data Marts – Inmon*
  - *Dimensional Model all through - Kimball*
  - *ER model for DW - Teradata*

Module 5: > Topic 1: Data Warehouse Architecture - Types

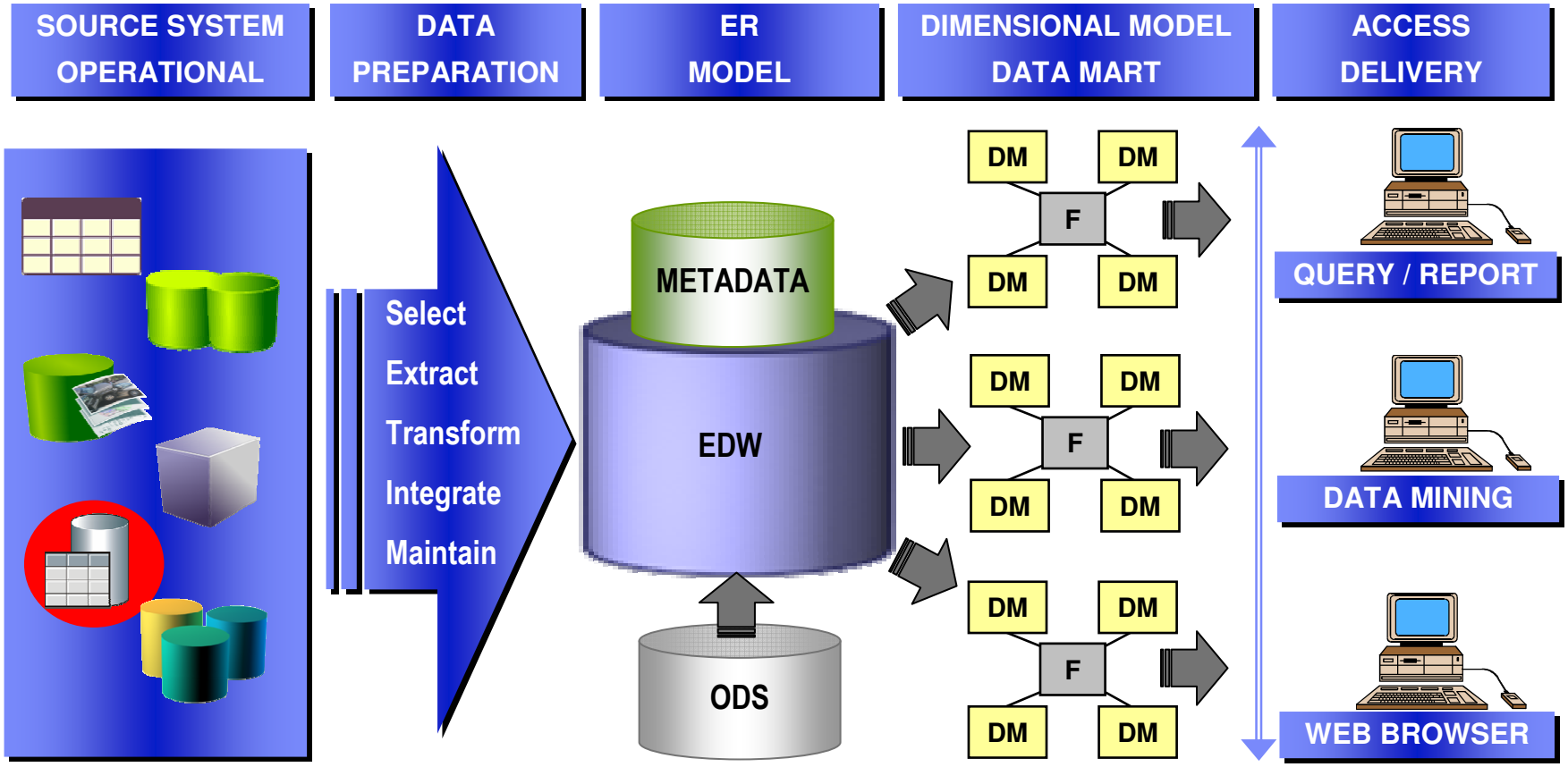
Architecture Diagram – Type 1 : Inmon - Corporate Information Factory (CIF)



Web Environment

Module 5: > Topic 1: Data Warehouse Architecture - Types

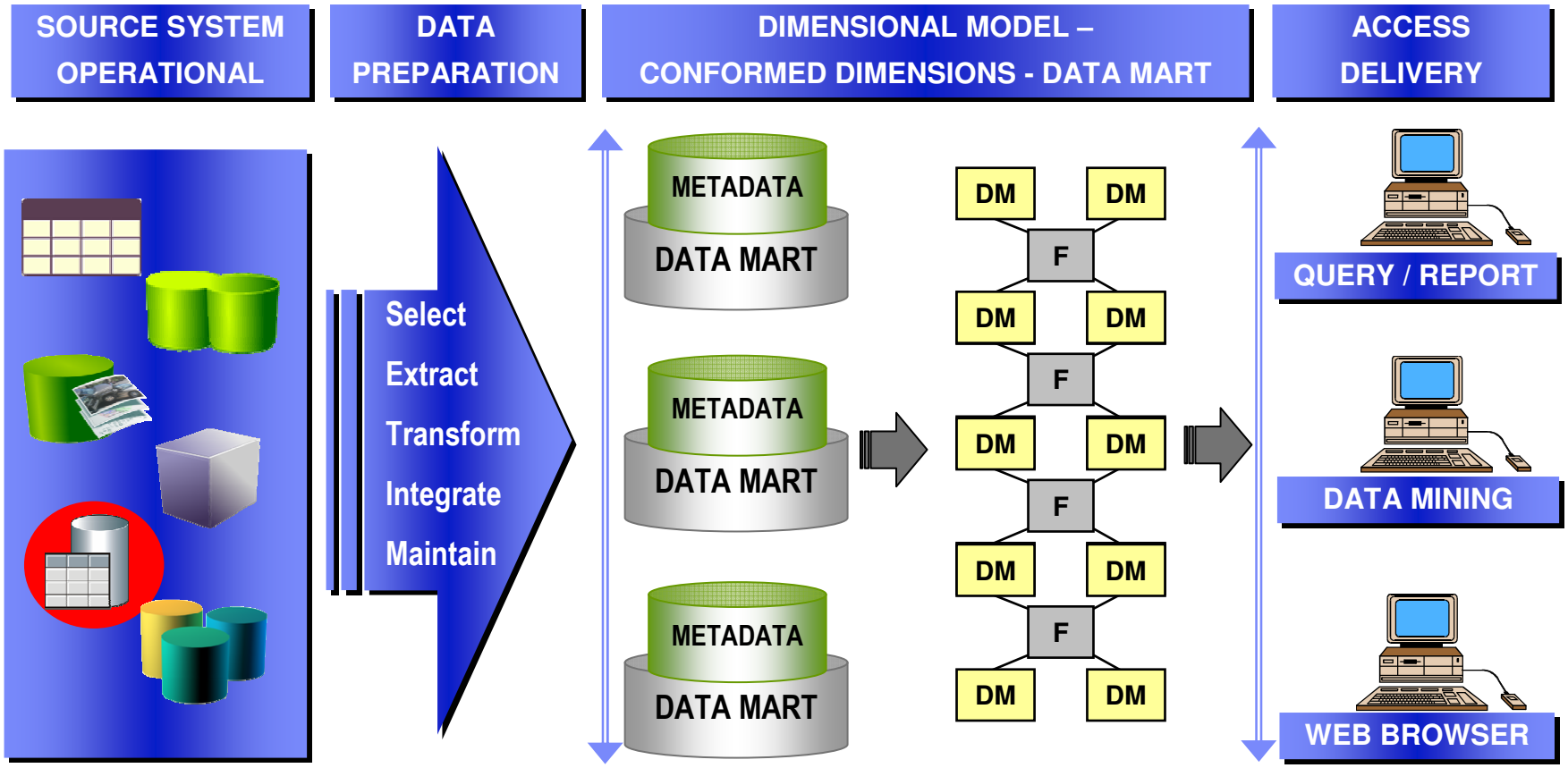
Architecture Diagram – Type 1 : Inmon Model



ER Model for EDW & Dimensional Model for Data Marts

Module 5: > Topic 1: Data Warehouse Architecture - Types

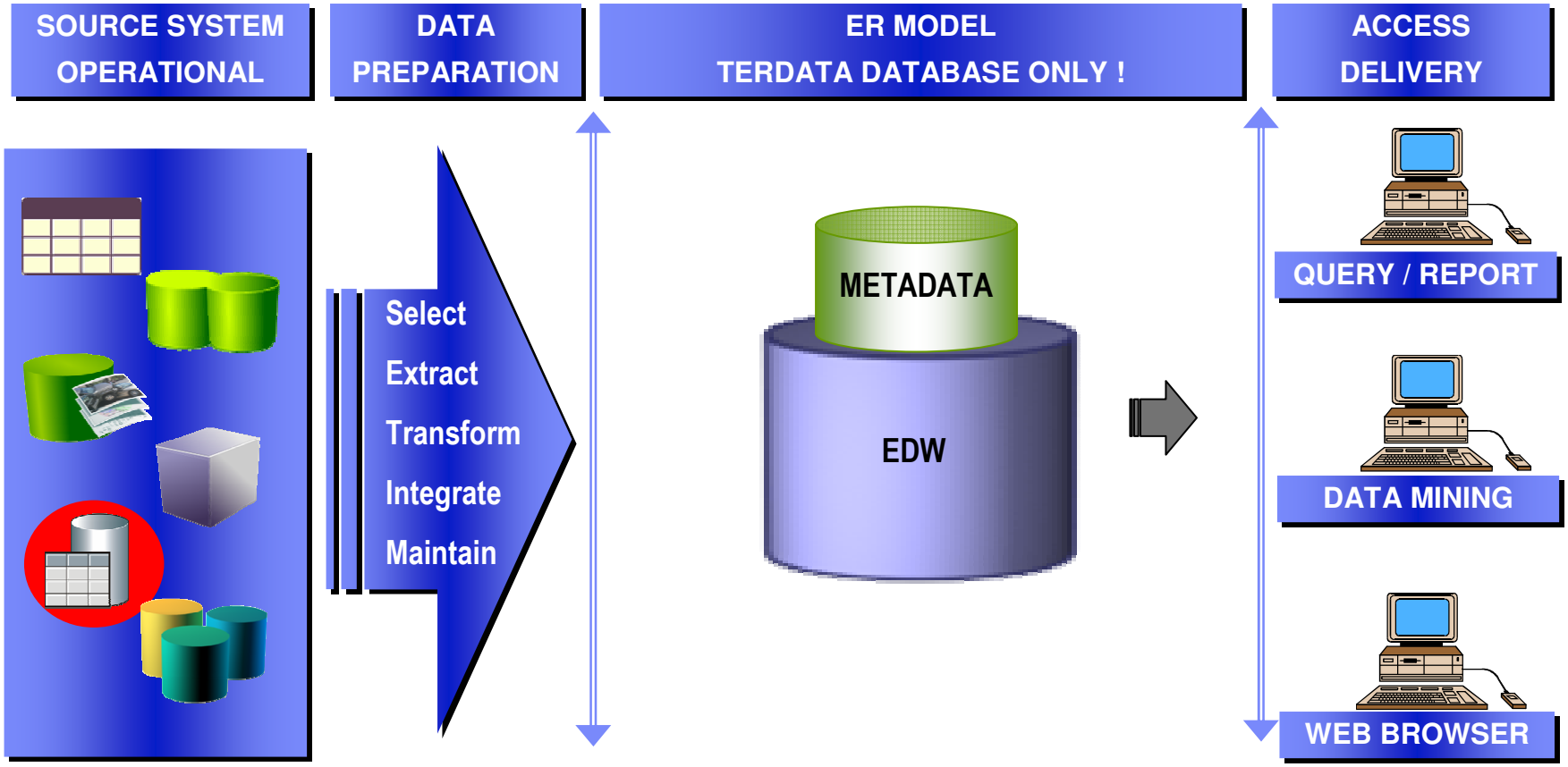
Architecture Diagram – Type 3 : Kimball Model



Multiple Data Marts With Conformed Dimensions

Module 5: > Topic 1: Data Warehouse Architecture - Types

Architecture Diagram – Type 2 : Teradata Model



ER Model for Data Warehouse – For Teradata Database Only !

Module 5: > Topic 1: DW Architecture - Types Summary

- Having completed this topic, you should be able to:
  - Data Stores – Five Roles in Data Warehouse
    - Intake
    - Integration
    - Distribution
    - Delivery
    - Access
  - Inmon's Data Warehouse Approach
    - Hub & Spoke Architecture
    - Hub integration
  - Kimball's DW Approach
    - DW Bus Architecture
    - Bus Integration & Conformed Dimensions

Module 5: > Topic 1: DW Architecture - Types Review

Module 5: > Topic 2: ETL - Insight

- ETL Classification
- ETL & Related Technologies

Module 5: > Topic 2: ETL - Insight

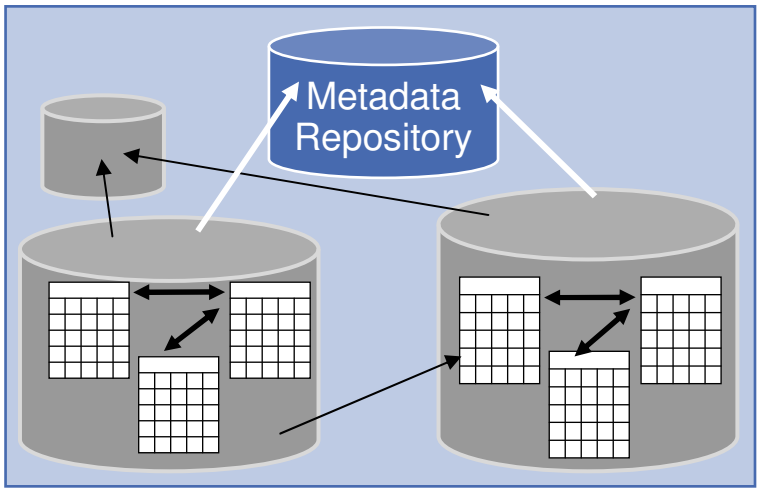
ETL - Classification

- Data Profiling
- Data Cleansing
- Data Integration, Consolidation & Population
- Data Replication
- Data Federation

Module 5: > Topic 2: ETL - Insight

ETL – Classification – Data Profiling

- Features
  - Analysis of metadata and data values; detection of differences between defined and inferred properties
  - Discovery of dependencies within source tables (functional dependencies, primary key) and across (detect common domain: redundancy, foreign-key relationship)
  - Recommendation for target data model (e.g., primary key, foreign key, normalized design)
- Benefits
  - Data quality by understanding the metadata of your data sources (structure and the relationships within and among them) supported through efficient tooling

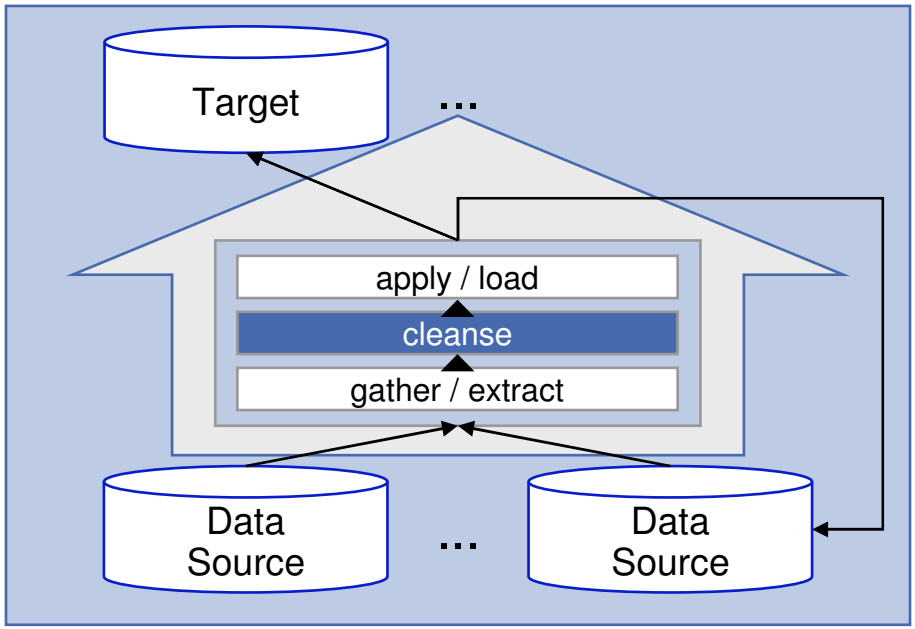


Module 5: > Topic 2: ETL - Insight

ETL – Classification – Data Cleansing

- Features
  - Data standardization transforms different input formats into a consolidated output format
  - Creating single domain fields
  - Incorporating business & industry standards
  - Data matching
  - Data enrichment
  - Data survivorship

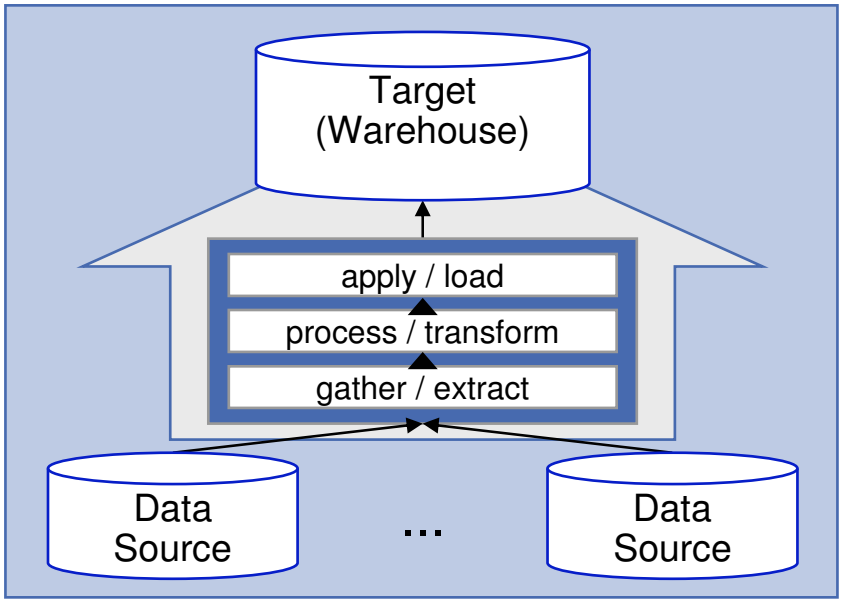
- Benefits
  - Reduce costs by improving data quality and consistency



Module 5: > Topic 2: ETL - Insight

ETL – Classification – Data Integration, Consolidation & Population

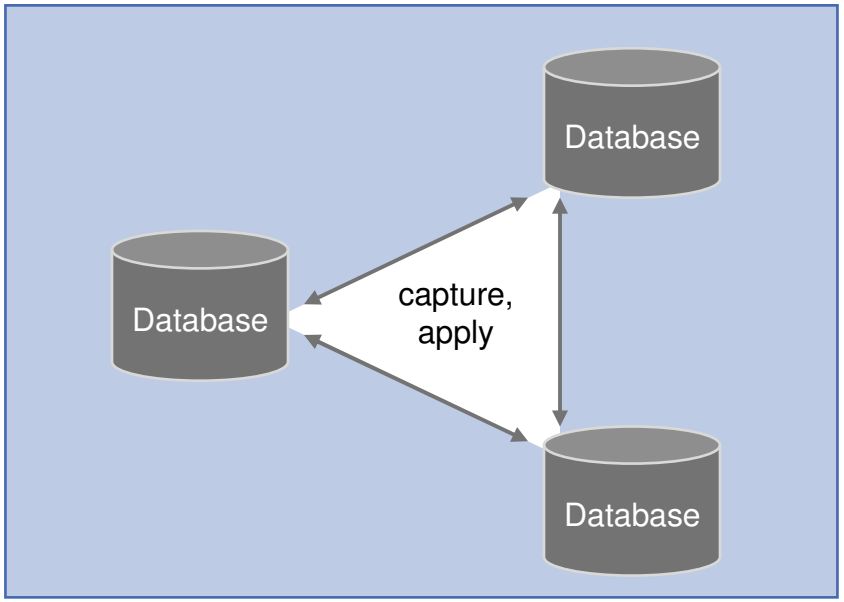
- Features
  - Complex transformations
  - High data volume (billions of records)
  - Performance and scalability of target access more important than data currency in target
  - De-coupled model: minimal impact on source systems due to target access
  - Target may collect historical snapshots of integrated information
- Benefits
  - Gain insight through single version of truth in distributed, heterogeneous and possibly low-quality data environment



Module 5: > Topic 2: ETL - Insight

ETL – Classification – Replication

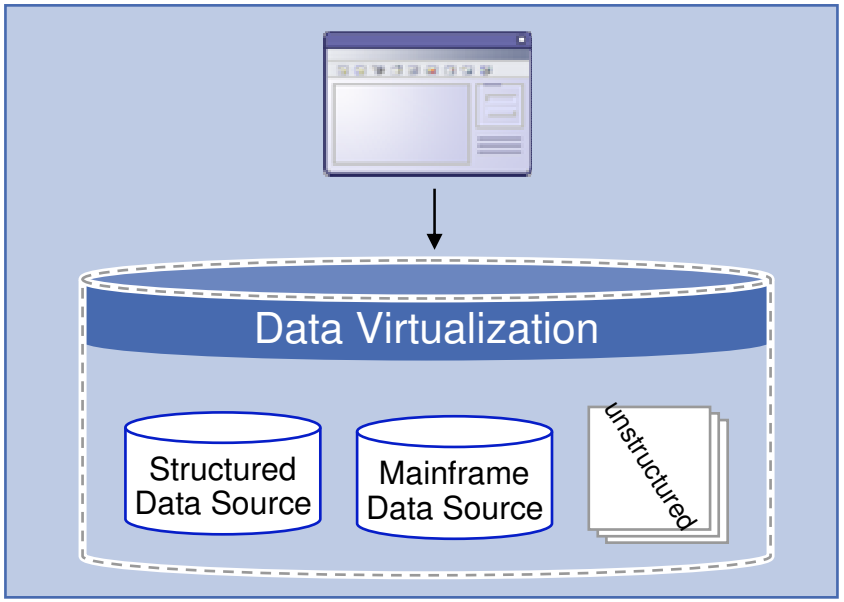
- Features
  - Unidirectional data distribution or Bidirectional data synchronization
  - Increased performance & scalability through distribution of load to multiple specialized copies of information
  - Increased availability and reliability for failover scenarios
  - Low-latency, high throughput data movement with queue-base replication
  - Automated and system-supported conflict resolution for bi-directional replication
- Benefits
  - Improved performance, scalability, reliability, and availability while guaranteeing consistency



Module 5: > Topic 2: ETL - Insight

ETL – Classification – Federation

- Features
  - On demand integration instead of copy management and data redundancy
  - Real-time access to distributed information as if from a single source
  - Flexible and extensible integration approach for dynamically changing environment
  - Query optimization
  - Integration of structured and unstructured information
- Benefits
  - Time to market and control costs when joining distributed (rather homogeneous) information



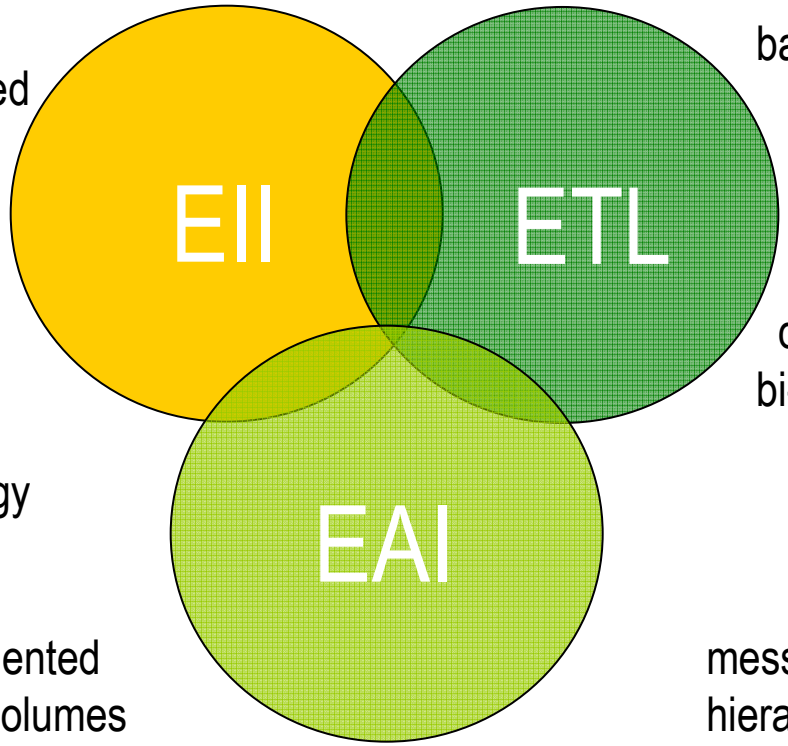
Module 5: > Topic 2: ETL - Insight

ETL - Technologies

- ETL - Extract Transform Load
- EII - Enterprise Information Integration
- EAI - Enterprise Application Integration
- ELT – Extract Load Transform

Module 5: > Topic 2: ETL - Insight

ETL - Technologies



query oriented
dynamic, real-time access
transparency across
different sources
metadata support

less a tool than a SQL
access layer and methodology

batch oriented
latency tolerant
complex transformations
massive volumes
code-less specification
deep metadata support
bi-tool integration

transaction-oriented
transactional volumes
developer audiences

C, Java, etc.

message-based
hierarchical structures
EDI, SWIFT, HPPA

lower emphasis on metadata integration

ensured delivery 24 by 7 security encryption restart/recovery

Module 5: > Topic 2: ETL - Insight

ETL - Technologies

- ETL - Extract Transform Load
  - Set-oriented, point-in-time transformation for migration, consolidation, and data warehousing.
  - Supports the large scale loading of a data warehouse or migration of vast quantities of data between systems
    - **Example : Informatica, Data Stage, Ab-Initio, OWB, BODI**

- EII - Enterprise Information Integration
  - Optimized & transparent data access and transformation layer providing a single relational interface across *all* enterprise data
  - Allows users to easily combine data warehousing reports with newly acquired real time analytic information with transparent queries – without caring where the data lives
    - **Examples : Ipedo, Data Mirror**

Module 5: > Topic 2: ETL - Insight

ETL - Technologies

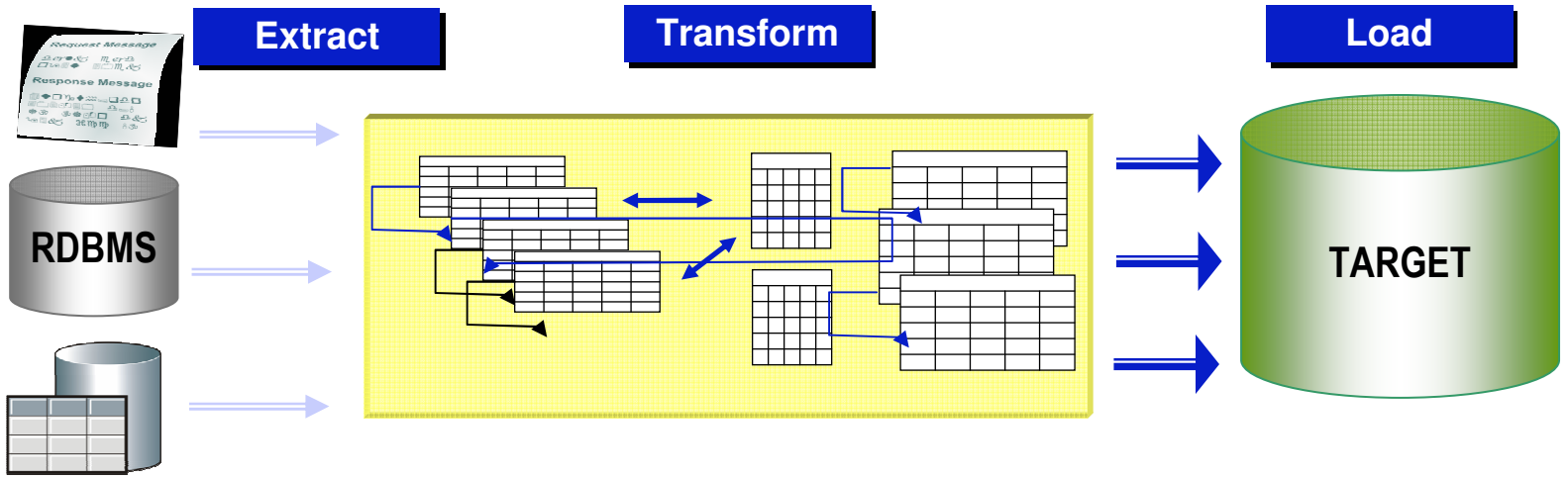
- EAI – Enterprise Application Integration
  - Message-based, transaction-oriented, point-to-point (or point-to-hub) brokering and transformation for application-to-application integration
  - Enables data sharing among partners in a supply chain, or brings transactional applications together after acquisition
    - **Example : BizTalk, Tibco**

- ELT - Extract Load Transform
  - Set-oriented, point-in-time loading for migration & transformation for data warehousing.
  - Supports the large scale loading of a data warehouse or migration of vast quantities of data between systems
    - **Example : Sunopsis**

Module 5: > Topic 2: ETL - Insight

ETL versus ELT

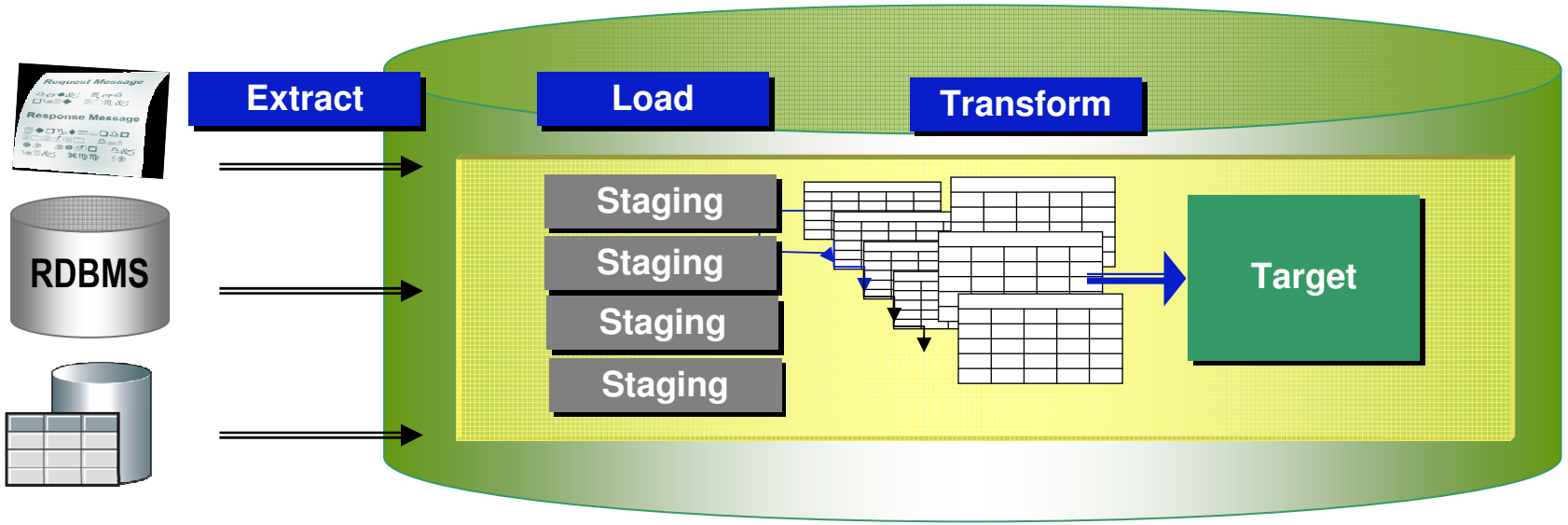
- ETL – Extract Transform Load
  - The data is manipulated outside the database to cleanse and sort, & then the result is loaded into the database



Module 5: > Topic 2: ETL - Insight

ETL versus ELT

- ELT – Extract Load Transform
  - The raw data from the source system loaded in staging tables in database, then it is cleansed and loaded



Module 5: > Topic 2: ETL - Insight Summary

- Having completed this topic, you should be able to:
  - Sub-Classification of ETL into
    - Data Profiling
    - Data Cleansing
    - Data Integration, Consolidation & Population
    - Data Replication
    - Data Federation
  - ETL & Related Technologies like
    - EII
    - EAI
    - ELT

Module 5: > Topic 2: ETL - Insight Review

Module 5: > Topic 4: Metadata - Insight

- Metadata - Definition
- Need for Metadata
- Types of Metadata
- Components of Metadata in DW Environmnet
- Characteristics of Metadata
- Metadata Architecture for DW Environment

Module 5: > Topic 4: Metadata - Insight

- *Metadata is "**Data about Data**". It provides a basis for Trust in information, providing visibility into lineage, relationships to other systems, & business definitions*
- *It refers to data that tries to describe a data set in terms of its **Value**, **Content**, **Quality**, **Significance***
- *It provides insight into data for information like :*
  - *What kind of Data ?*
  - *Who is the owner of the data ?*
  - *How was the data created ?*
  - *What are the attributes and significance of the data created or collected ?*

Module 5: > Topic 4: Metadata - Insight

Need for Metadata -

- *Faster Development, Faster Maintenance*
  - *Helps accelerate development by actively sharing knowledge through the analysis, design, and build process, even with external technologies*
  - *Serves as a automatic form of documentation to make maintenance easier, and provides the ability to assess the impact of changes prior to making them*
- *Better Business & IT Collaboration*
  - *Aligns business and IT understanding by linking business terms, rules, and taxonomies to technical artifacts*
  - *Allows business and IT resources to collaborate while using tools tailored to their roles*
- *Trust*
  - *Supports a higher degree of trust in information by keeping a record of collaboration, and the ability to see where information comes from*

Module 5: > Topic 4: Metadata - Insight

Need for Metadata -

- *Improve Consistency, Accuracy & Speed of data for DW by providing business specific & technical specific information of available DW data*
- *Reduce development time by integrating & merging data from disparate sources into the DW*
- *Increase data reliability by having consistent definitions & nomenclatures of data within the DW*
- *Helps in integration of data across enterprise, especially when Acquisitions & Mergers are the order of the day*

Module 5: > Topic 4: Metadata - Insight

Types of Metadata -

Developer / Technical Analysts

DBA / System Administrator

Business Analysts / Managers / Executives

Technical Metadata
Attribute Name
Entity Relationship
Domain Values
Transformation Rules

Business Metadata
Entity Business Definition
Attribute Business Definition
Report Business Description
Business Transformation Rules

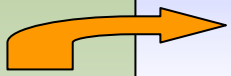
Module 5: > Topic 4: Metadata - Insight

Metadata – Example

Metadata for a Customer Database Table -

CUSTOMER TABLE

CUST_ID
CUST_NAME
CUST_ADDR1
CUST_ADDR2
CUST_TYPE
CUST_PHONE_HOME
CUST_PHONE_OFF
CUST_PHONE_CELL
CUST_IS_ACTIVE



METADATA for the CUSTOMER TABLE

TECHNICAL METADATA

TABLE_NAME : CUSTOMER
TABLE_OWNER : XYZ
TABLE_TYPE: MASTER TABLE
CREATE_DATE : 18-DEC-2008
MODIFIED_DATE : 19-DEC-2008
USED_FOR : MASTER TABLE

CUSTOMER_ID : SHOULD BE ALPHANUMERIC
CUST_TYPE : DETERMINED BY BUSINESS RULES
CUST_IS_ACTIVE : BASED ON CREDIT HISTORY

BUSINESS METADATA

Module 5: > Topic 4: Metadata - Insight

Metadata Components in DW Environment -

- *Metadata Components can be found in DW at the following areas -*
  - ***Operational / Source Systems***
  - ***Transformation / ETL***
  - ***ODS***
  - ***EDW***
  - ***Data Mart***
  - ***Development***
  - ***Versioning***
  - ***Enterprise Modelers***
    - ***ETL Tools***
    - ***Reporting Tools***
    - ***Universes***
    - ***Data Modeling Tools***

Metadata	Component
Owner	Data Owner Application Owner
General Characteristic	Technical Name Business Name Data Type Data Lenght Comments
Rules	Relationship Security Business Rules & Policies
Physical Characteristics	Data Source Physical Location – DB Transformation

Module 5: > Topic 4: Metadata - Insight

Characteristics of DW Metadata -

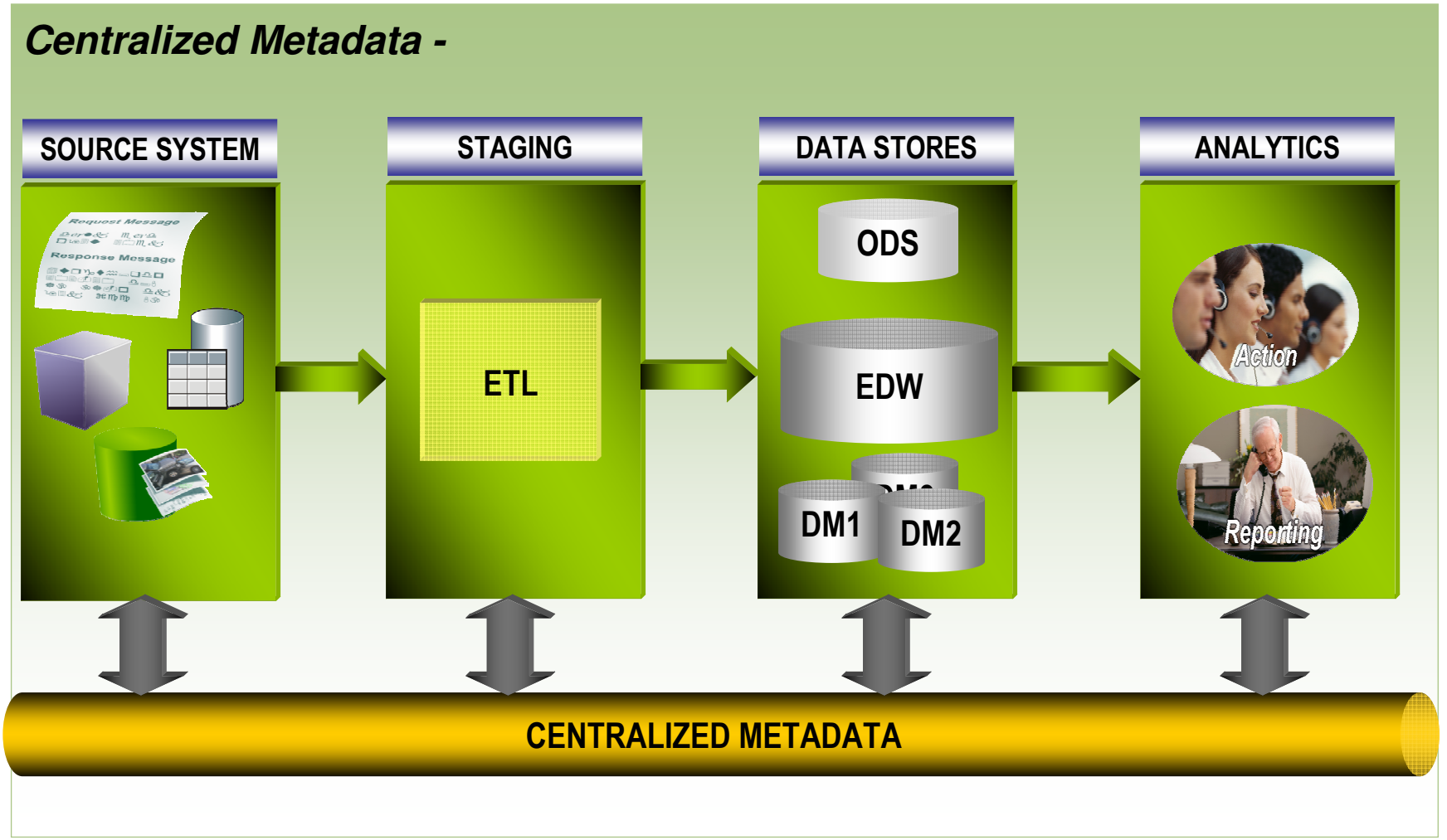
- *DW Metadata typical helps in tracking the following -*
  - ***Extract Information***
    - *Last Refresh / Load - Date / Time*
  - ***Historical Information about data & meta data***
    - *Versioning*
    - *Data Access Patterns over a period of time*
  - ***Data Mapping Information***
    - *Source to Target*
    - *Transformation Rules*
  - ***Summarization***
    - *Aggregation Algorithm*
  - ***Archiving***
    - *Period of Data Purging*
  - ***Reference & Standardization***
    - *Aliases*
    - *Lookups*

Module 5: > Topic 4: Metadata - Insight

Metadata Architecture for DW Environment -

- *Centralized Metadata Architecture*
  - *Metadata is stored & managed centrally. Rights of creation, maintenance, distribution & decimation of metadata reside with a central authority*
- *Distributed Metadata Architecture*
  - *Metadata is stored in individual repositories of most of the components of the Data Warehouse like – Source Systems, ETL Tools, Jobs, Versioning, EDW, DM & ODS*
  - *It means that metadata resides & is managed locally. This implies that an EDW will create, update & delete its own metadata*

Module 5: > Topic 4: Metadata - Insight

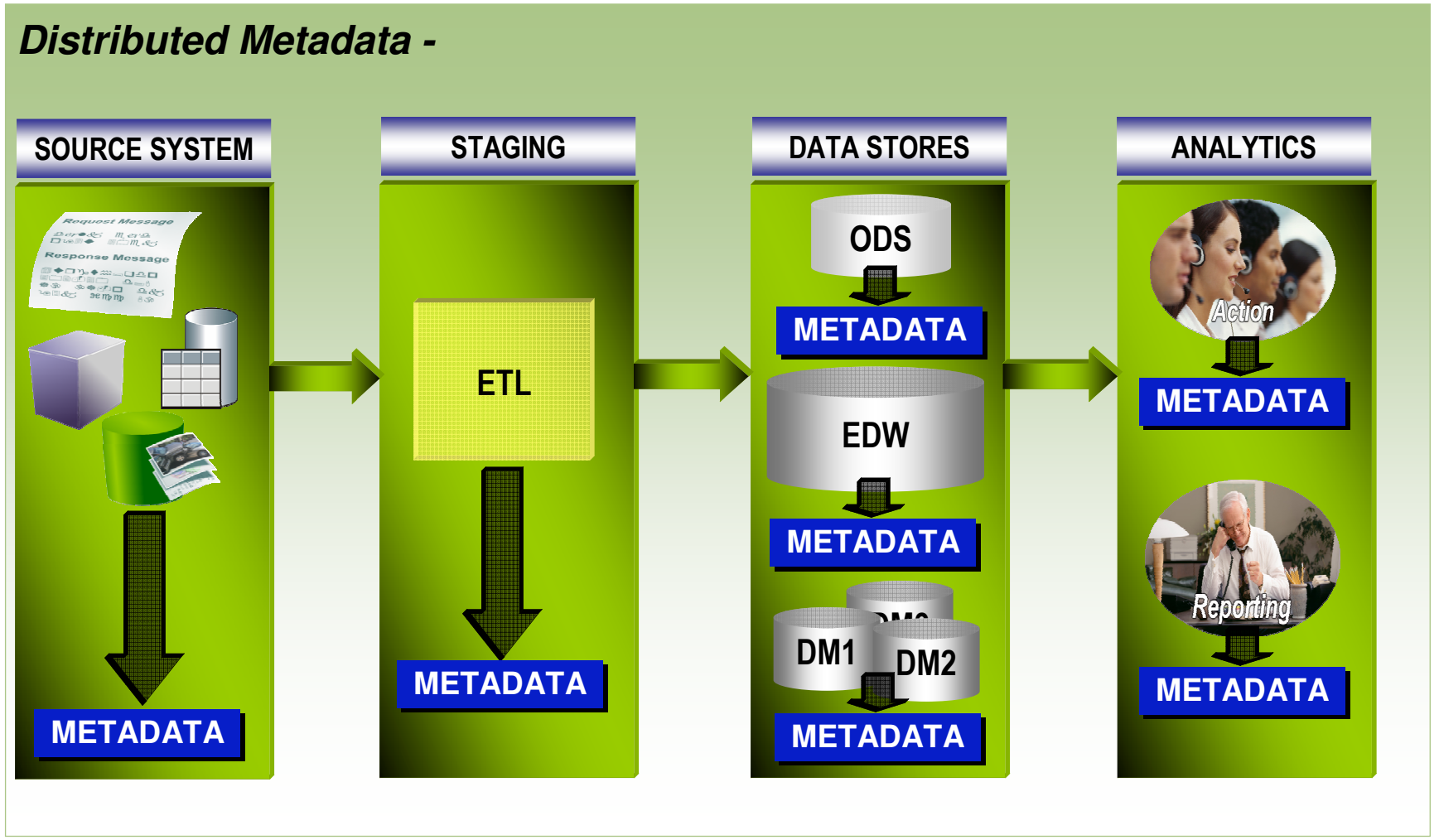


Module 5: > Topic 4: Metadata - Insight

Centralized Metadata : Advantages -

- *Centralized control*
- *Minimal data redundancy*
- *Better data sharing and integration*
- *Improves data consistency*
- *Ease in enforcing data standards*
- *Ease of Maintenance*

Module 5: > Topic 4: Metadata - Insight



Module 5: > Topic 4: Metadata - Insight

Centralized versus Distributed Metadata :

- *Goal -*
  - *Enable Metadata to be seamlessly shared amongst the different components of the DW Environment*
  - *Metadata consistency across repositories without compromising the flexibility*
  - *In cases where metadata is distributed, there should be technology that is capable, compatible & available to connect & integrate this metadata across the whole DW environment*
- *As such an hybrid solution is the preferred approach while implementing Metadata solutions in a Data Warehouse Environment*
  - *Centralized Metadata Repository for Corporate Data Structures, Business definitions & Rules*
  - *Distributed Metadata Repository for specific data models within the DW Framework that require the corporate metadata to be overwritten in some cases and / or specific metadata that is not appropriate for the corporate repository*
    - *Like departmental, or region specific data*

Module 5: > Topic 4: Metadata - Insight Summary

- Having completed this topic, you should be able to:
  - What is Metadata
  - Need for Metadata
  - Types of Metadata
  - Components of Metadata in DW Environment
  - Characteristics of Metadata
  - Metadata Architecture for DW Environment
  - Centralized & Distributed Metadata

Module 5: > Topic 4: Metadata - Insight Review

Module 5: > Topic 5: MDM - Introduction

- MDM - Definition
- Need for Master Data Management
- Challenges in Implementing MDM
- MDM Domains
- Approach to MDM

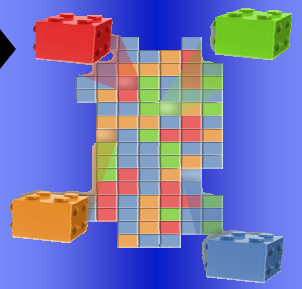
Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Business Drivers for MDM

"Unless enterprises figure out how to synchronize [master] data among departments, divisions and enterprises, the value promised from business process fusion will be much less than expected."

Gartner, October 2003

ORAGNIZATION SCENARIO	WHY NEEDED	WHAT IS REQUIRED	PROBLEM FACED
MERGER & ACQUISITION	BUSINESS COMPULSIONS	DATA CONSOLIDATION	1. DUPLICATE 2. INCONSISTENT 3. INACCURATE 4. FRAGMENTED DATA
INFORMATION EXCHANGE	SHARE INFO. WITH BUSINESS		
INFORMATION MANGEMENT	DATA MGMT. GOVERNANCE		



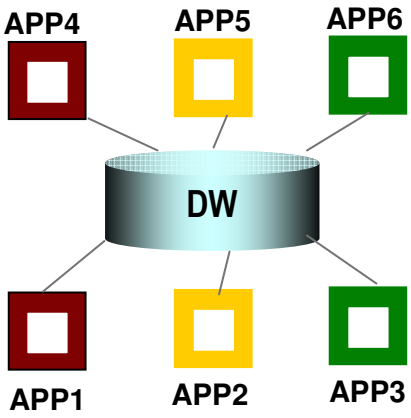
Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Challenges Faced with Master Data -

- *Duplicates:*
  - *Distinct Supplier Records for 'IBM', 'I.B.M.' & 'International Business Machines'*
- *Multiple conflicting views:*
  - *Material Master data is out of sync between ERP systems*
- *Data Quality Issues:*
  - *Customer movement … clean data erodes quickly*
- *Fragmentation:*
  - *Product Cost & specification managed by discrete out of sync ERP systems*

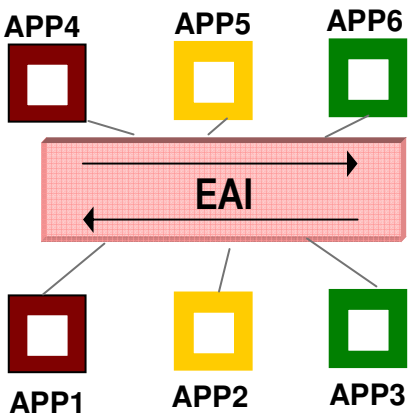
Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Challenges Faced with Master Data -



Data Warehouse

1. Different Versions of truth
2. Unidirectional
3. Batch updates .. Synchronization difficult



EAI

1. No History
2. Event driven
3. Investment for data synchronization is huge

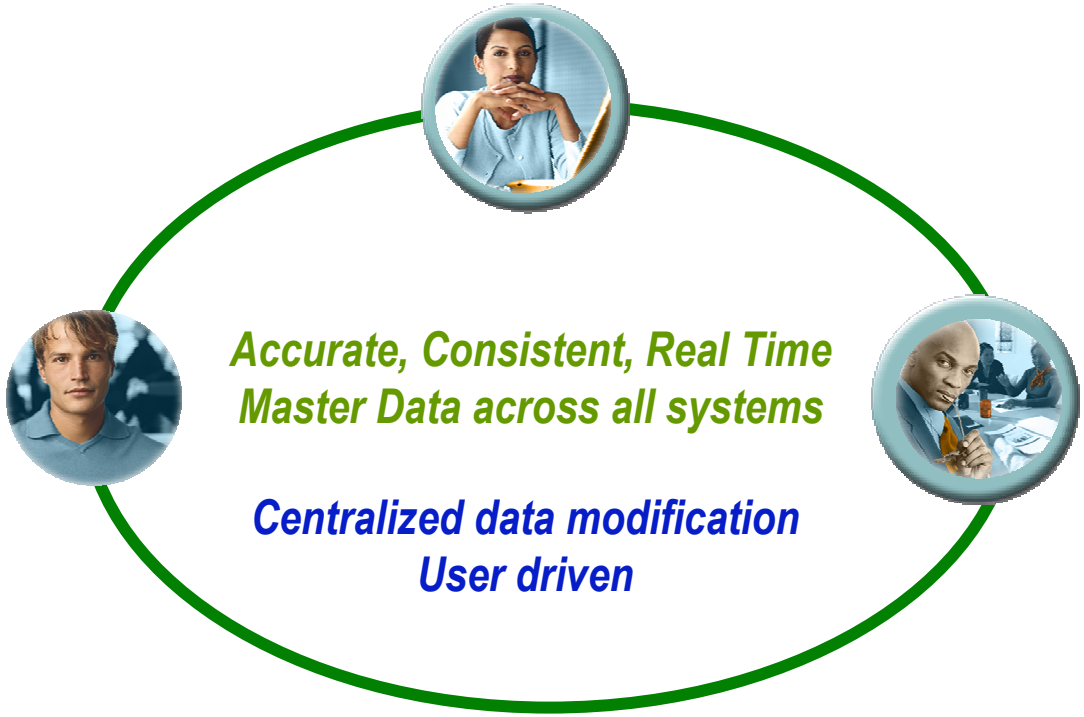
ERP

1. Proliferation of data
2. No sync between different ERP's
3. High Investment for consolidation

Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

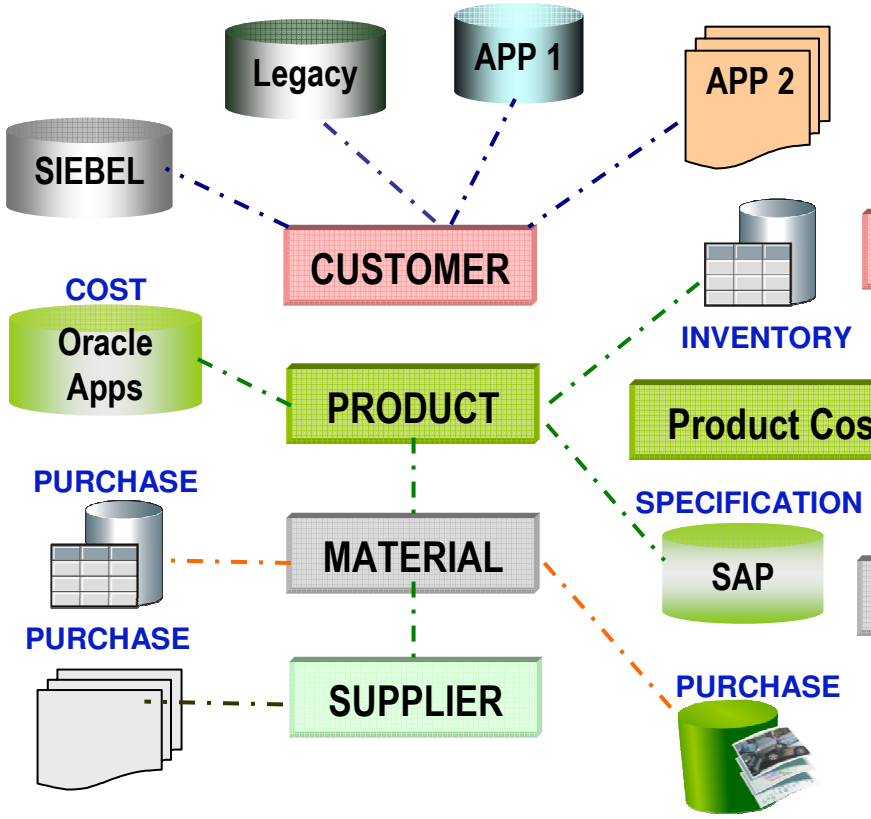
Role of Master Data

Single version of Truth (Information & Knowledge) across the enterprise



Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Role of Master Data



Business Scenario	Master Data Domain
Who are our best customers, In what Age Group?	Customer Name , Customer Age, Customer Adddress, Customer Credit
No Single Consistent Source of Customer Data	
What are our profit margins?	Product Price, Product Cost, Product Sales
Product Cost & Specification managed by discrete ERP systems	
Are our supplliers giving us the best price?	Supplier, Material
Material Master data is out of sync.	
How much should we produce this quarter? Which product requires more marketing?	Product Name, Product Category, Product Brand

Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Role of Master Data

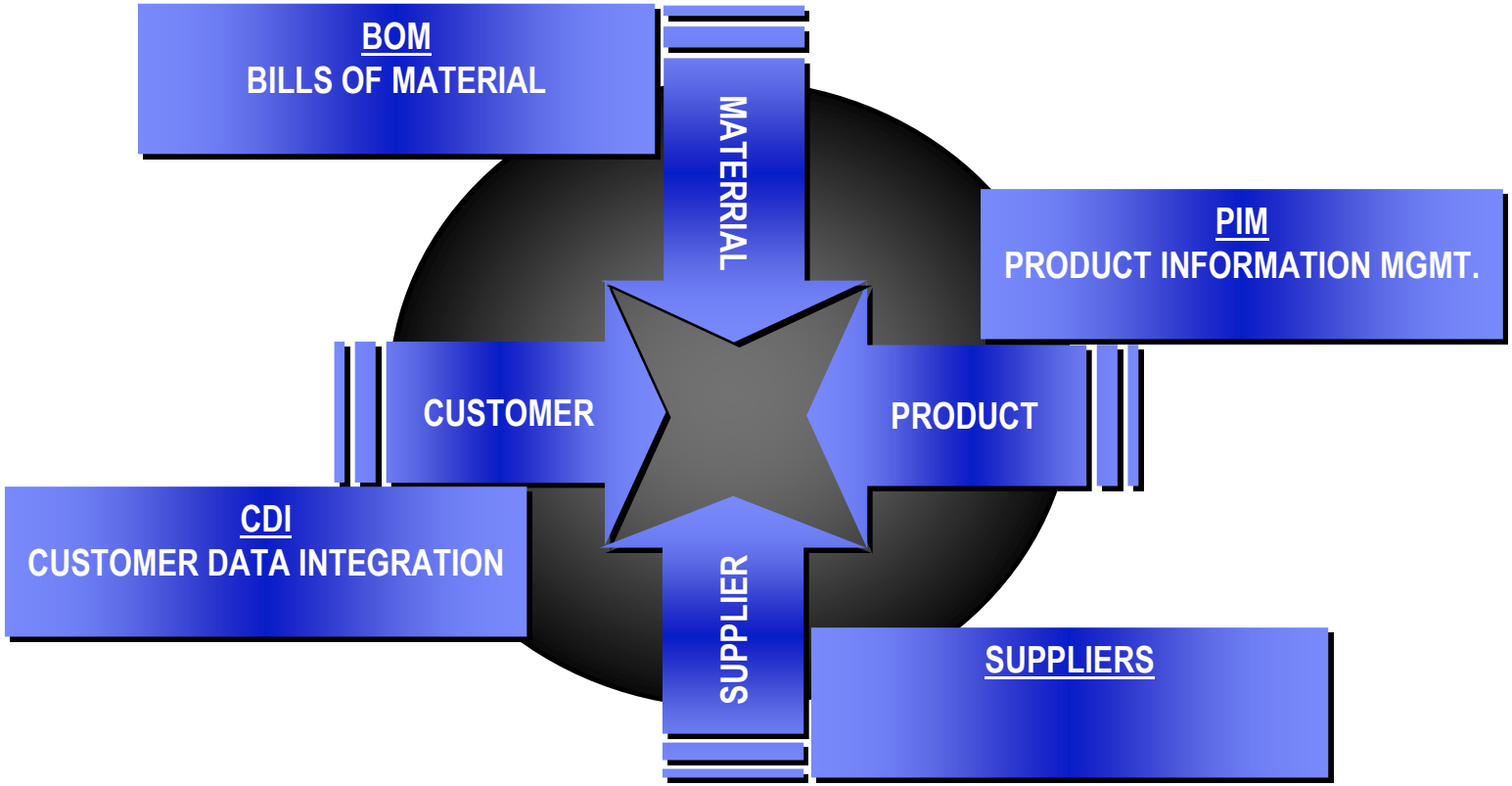
- ***Main purpose***
  - *Decouple master data from individual applications and provide single version of truth for master data (analytical, operational, reference, etc.)*

- ***What is Master Data?***
  - *Describe Core business entities : customers, suppliers, partners, products, materials, chart of accounts, location & employees*
    - *High value Information used repeatedly across business processes*
    - *Generally used across multiple LOB (Line of Business)*
  - *Gives business context by providing concrete data models processes for a particular domain*

- ***What are the benefits of Master Data ?***
  - *Common authoritative source of accurate, consistent and comprehensive master information for business services to access critical business information*
  - *Common business services supporting consistent information-centric procedures across all applications within the enterprise and extended enterprise*
  - *Business process support to integrate with or drive business processes across heterogeneous applications by making data actionable*

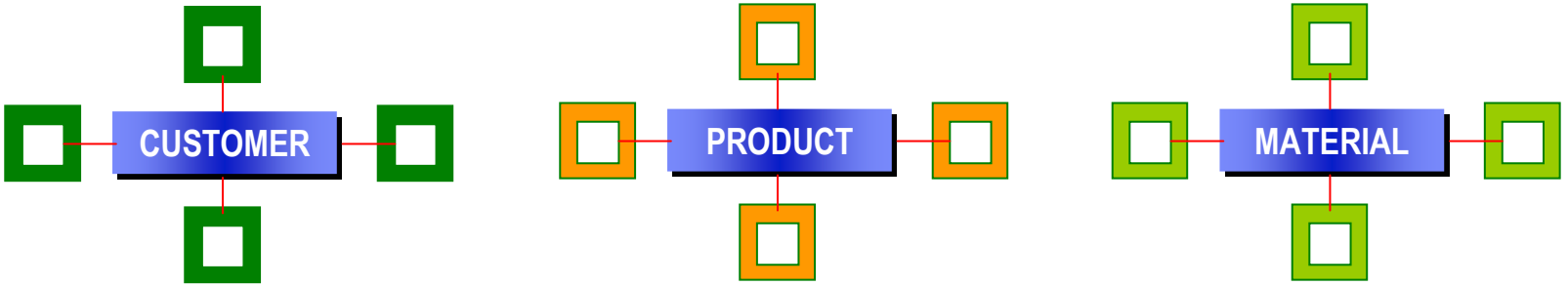
Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Domains & Flavors of MDM



Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

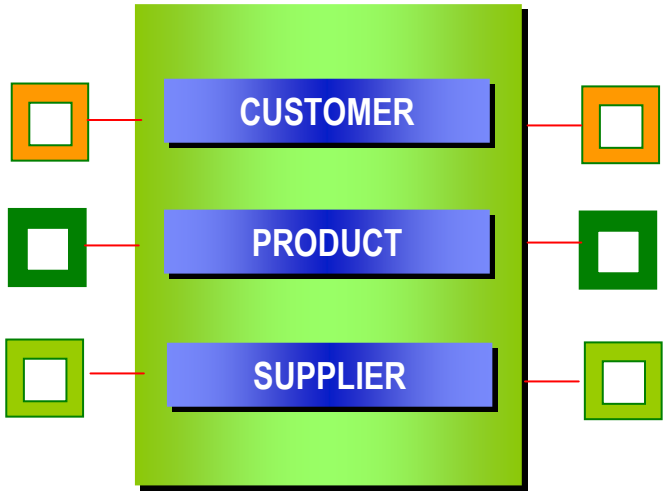
Approaches to MDM – Distinct MDM per Master



- ***Strengths***
  - ***Easy to build & maintain***
  - ***Cost Effective***
- ***Weakness***
  - ***Poor Enterprise Scalability***
  - ***Possibility of Master Data Fragmentation***

Module 5: > Topic 5: Master Data Mgmt. (MDM) - Introduction

Approaches to MDM – Platform Centric Approach



- ***Strengths***
  - ***Enterprise Scalability***
  - ***Open Extensible Architecture***
- ***Weakness***
  - ***Complex***
  - ***Roadmap & Plan Required***

Module 5: > Topic 5: MDM - Introduction Summary

- Having completed this topic, you should be able to:
  - What is Master Data
  - Need for Master Data
  - Challenges in implementing Master Data Management
  - Domains in MDM
  - Approaches to MDM

Module 5: > Topic 5: MDM - Introduction Review

Module 5: > Topic 6: Data Mining – An Introduction

- Data Mining – Definition
- Need for Data Mining
- Advantages of Data Mining
- Data Mining Examples
- Data Mining - Process
- Data Mining – Applications & Tools

Module 5: > Topic 6: Data Mining – An Introduction

Data Mining -

- *Data Mining is the detection of unknown, valuable & non trivial information in large volumes of data using automated statistical analysis*
- *It helps in trying to predict future trends & discover patterns and behavior that have previously been un-noticed or un-earthed*
- *It leads to simplification & automation of statistical process of deriving information from huge volumes of data*

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions

Module 5: > Topic 6: Data Mining – An Introduction

Need for Data Mining

Who are our best Customers?

How can we detect fraud?

What do we need to know in order to predict & prevent losses?

- ***In competitive market answers to the business questions above can make all the difference in profitability & loss of market share, Data mining provides IT with tools to answer these questions thus producing & discovering new information & knowledge that decision makers can act upon***

- *It does this by using sophisticated techniques such as artificial intelligence to build a model of the real world based on data collected from a variety of sources including corporate transactions, customer histories and demographics, and from external sources such as credit bureaus*

- ***This model produces patterns in the information that can support decision making and predict new business opportunities***

Module 5: > Topic 6: Data Mining – An Introduction

Need for Data Mining

Can we replace skilled business analysts with data mining?

How is Data Mining related to DW-BI?

Can Data mining replace OLAP & reporting applications?

- ***Data Mining does not replace business analysts & managers, It compliments these users to confirm their empirical observations, find new patterns, that yield steady incremental improvement & breakthrough insight***

- ***Data mining follows Data Warehousing, Data to be mined is extracted from EDW, the need for data cleansing, integration, & consolidation is thus eliminated. With EDW as foundation, 70% of the data mining effort is eliminated thereby saving time, money & increasing reliability & delivering faster results***

- ***Data mining cannot replace OLAP & reporting tools, they compliment each other. The outcome of the patterns discovered using data mining need to be analyzed before being put into action, in order to know the implications of such patterns. OLAP tool can allow the analysts to get answers to these queries***

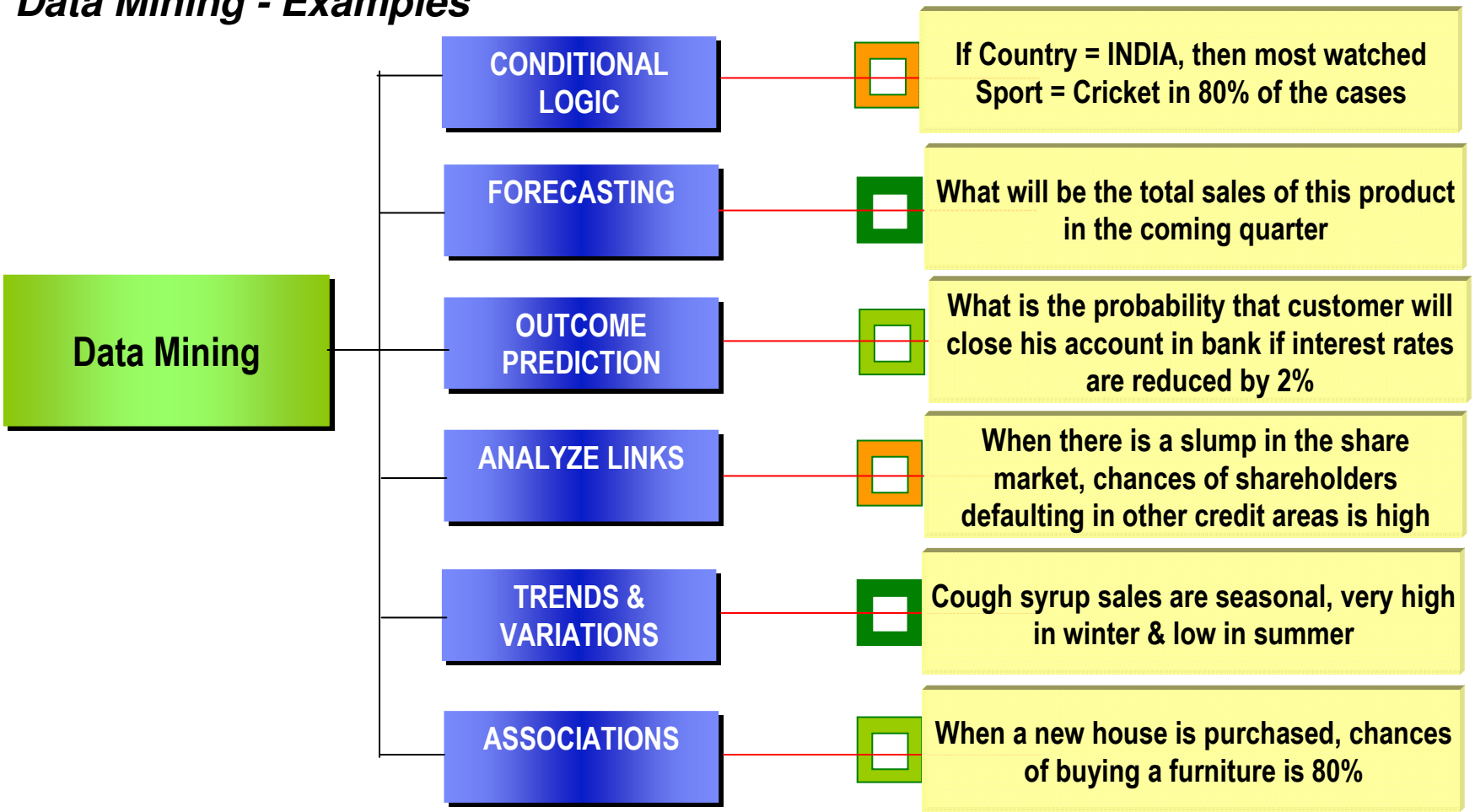
Module 5: > Topic 6: Data Mining – An Introduction

Advantages of Data Mining

- *Add value to data holding – data collected as part of DW-BI initiatives*
- *Competitive Advantage*
- *More efficient & effective decision making*
- *Data volumes are ever increasing, business feels there is value in historical data, Automated knowledge discovery is the only way to explore this data*
- *Supports high level & long term decision making*
- *Allows business to be proactive & prospective*

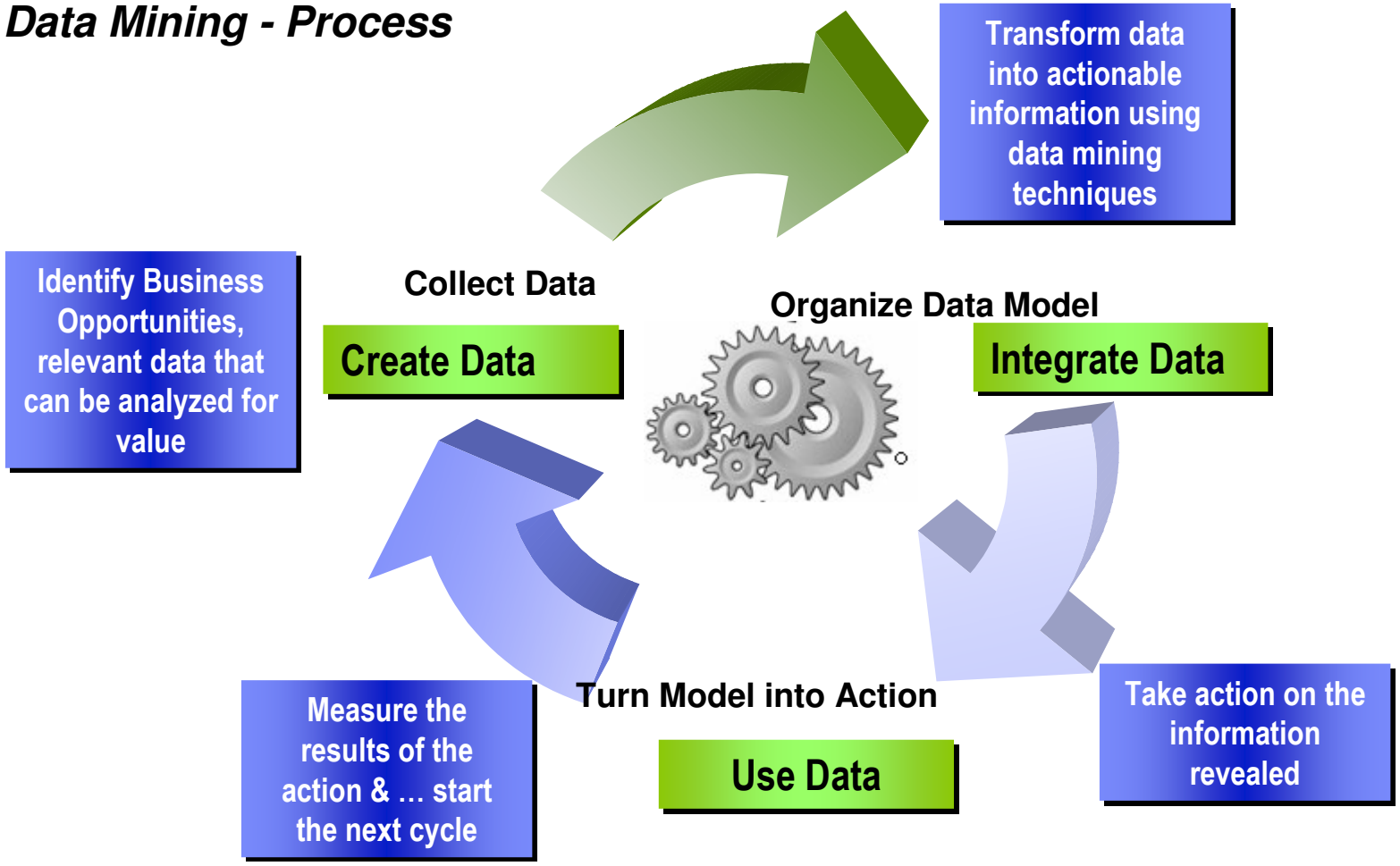
Module 5: > Topic 6: Data Mining – An Introduction

Data Mining - Examples



Module 5: > Topic 6: Data Mining – An Introduction

Data Mining - Process



Module 5: > Topic 6: Data Mining – An Introduction

Data Mining – Process – Data Preparation

- Data Preparation
  - Collection
  - Assessment
  - Consolidation & Cleaning
  - Data Selection
  - Cross Validation
    - Break up data into groups of small size
    - Use one group for testing and one group for building the mining model

Module 5: > Topic 6: Data Mining – An Introduction

Data Mining – Process – Type of DM Models

Trying to predict the future or trying to predict the state of the world?

- Descriptive Models -
  - Clustering
  - Associations
  - Sequence discovery
- Predictive Models -
  - Classification
  - Regression
  - Time Series

Module 5: > Topic 6: Data Mining – An Introduction

Data Mining – Applications & Tools

- Applications -
  - Target Marketing
  - Churn Analysis
  - Customer Profiling
  - Bioinformatics
  - Fraud Detection
  - Medical Diagnostics

- Tools & Vendors -
  - IBM – Intelligent Miner
  - SAS – Enterprise Miner
  - SGI – Mine Set

Module 5: > Topic 6: Data Mining – An Introduction Summary

- Having completed this topic, you should be able to:
  - What is Data Mining
  - Need for Data Mining
  - Advantages of Data Mining
  - Data Mining Examples
  - Data Mining Process & Models
  - Data Mining Applications & Tools

Module 5: > Topic 6: Data Mining – An Introduction Review

Module 5: > Topic 7: Data Governance – An Introduction

- Data Governance – Definition
- Need for Data Governance
- Advantages of Data Governance
- Implementation Approach
- Data Stewardship
- Characteristics of a governed organization

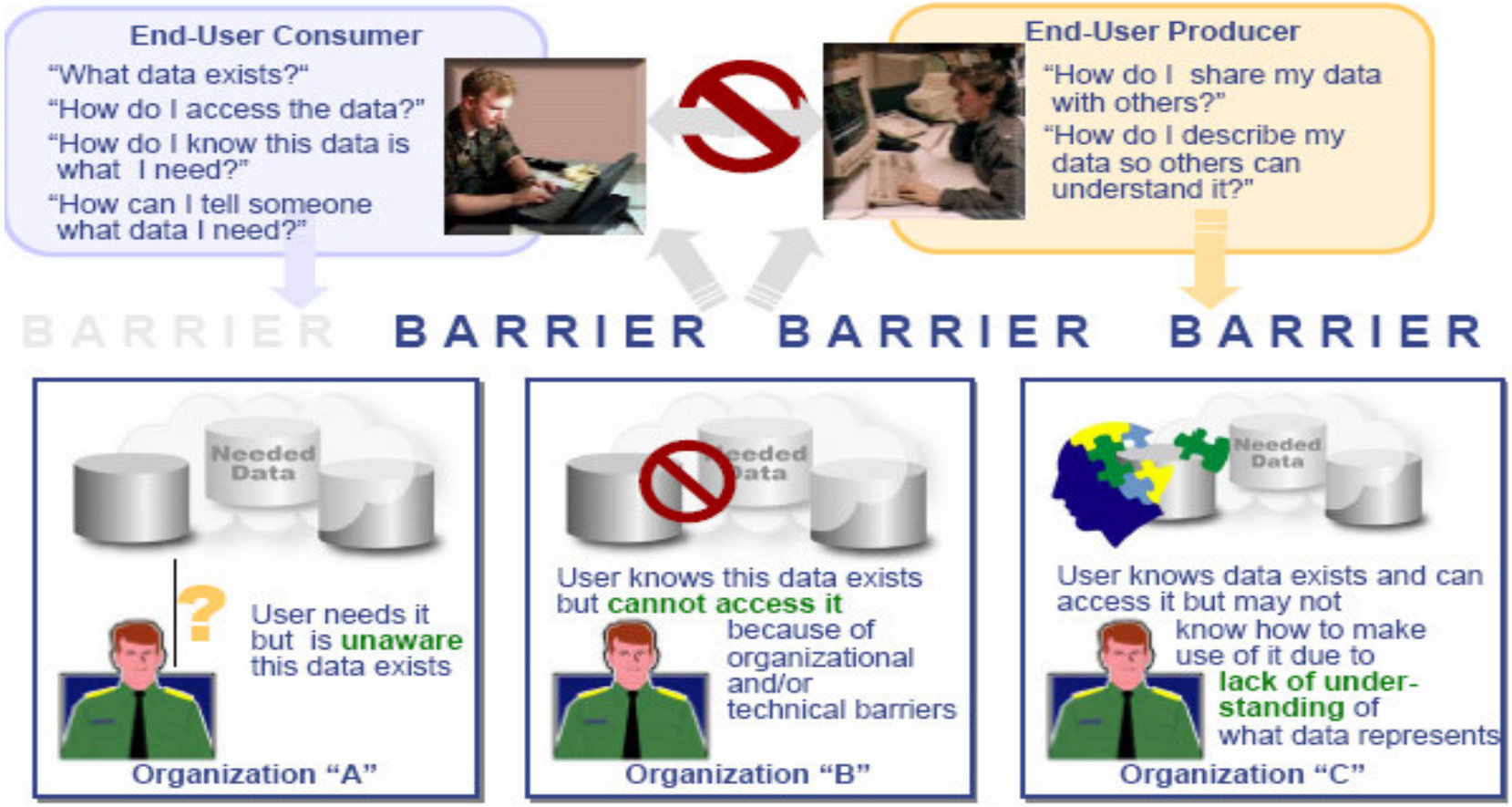
Module 5: > Topic 7: Data Governance – An Introduction

Data Governance - Definition

- *Data Governance refers to the overall **management** of **Availability**, **Usability** and **Security** of data employed in an enterprise*
- *It includes a Governing body, a Defined set of procedures & a Plan to execute these procedures*
- *It involves Stewardship and Data Security*
- *It also helps in adhering to Regulatory Compliances.*

Module 5: > Topic 7: Data Governance – An Introduction

Need for Data Governance -



DAMA International
Wilshire Conferences, Inc.

Module 5: > Topic 7: Data Governance – An Introduction

Need for Data Governance -

- ***"Dirty" Data is a Business Problem, Not an IT Problem***
  - *Gartner March 2007*

- ***Over the next two years, more than 25 percent of critical data in Fortune 1000 companies will continue to be flawed, that is, the information will be inaccurate, incomplete or duplicated…***
  - *Gartner*

Businesses are discovering that their success is increasingly tied to the quality of their information. Organizations rely on this data to make significant decisions that can affect customer retention, supply chain efficiency and regulatory compliance. As companies collect more and more information about their customers, products, suppliers, inventory and finances, it becomes more difficult to accurately maintain that information in a usable, logical framework

Module 5: > Topic 7: Data Governance – An Introduction

Need for Data Governance -

- *The amount of data is increasing every year, IDC estimates that the world will reach a zettabyte of data (1,000 exabytes or 1 million pedabytes) in 2010.*
- *A significant portion of all corporate data is flawed*
- *Process failure and information scrap and rework caused by defective information costs the United States alone $1.5 trillion or more*
- *The amount of data - & the prevalence of bad data - is growing steadily*

Module 5: > Topic 7: Data Governance – An Introduction

Need for Data Governance -

- *Enterprise data is frequently held in disparate applications across multiple departments and geographies*

- *The confusion caused by this disjointed network of applications leads to poor customer service, redundant marketing campaigns, inaccurate product shipments and, ultimately, a higher cost of doing business*

- *To address the spread of data – and eliminate silos of corporate information – many corporate implement enterprise wide data governance programs, which attempt to codify and enforce best practices for data management across the organization*

Module 5: > Topic 7: Data Governance – An Introduction

Advantages of Data Governance -

- *Data Governance employs a '**Holistic**' approach to the management of **People**, **Policies** & **Technology** that manage enterprise data, thereby providing the following benefits -*
  - ***Better data drives more effective decisions across every level of the organization***
  - ***With more unified view of the enterprise, managers & executives are able to devise strategies that make the company more profitable***
  - ***Consistent enterprise view of organizations data, leads to increase in consistency & confidence in decision making***
  - ***Decrease the risk of regulatory fines by adhering to rules, processes & standards for creation, acquisition, usage, decimation, security, maintenance, & availability of data***
  - ***Consistent Information Quality***
  - ***Accountability for Information creation, usage & decimation***

Module 5: > Topic 7: Data Governance – An Introduction

Implementation Approach -

- ***Data Governance is an evolutionary process***
- ***Set up of Data Resource Management team and supervision by business data stewards***
- ***Define and maintain data strategy and policies, manage data issues, estimate data value and data management costs, and justify the budget for data management programs***
- ***Enforce Data management policies and programs promote them***
- ***Make users aware of these policies and programs***
- ***Have Data Stewardship, Strategy & Governance in place***

Module 5: > Topic 7: Data Governance – An Introduction

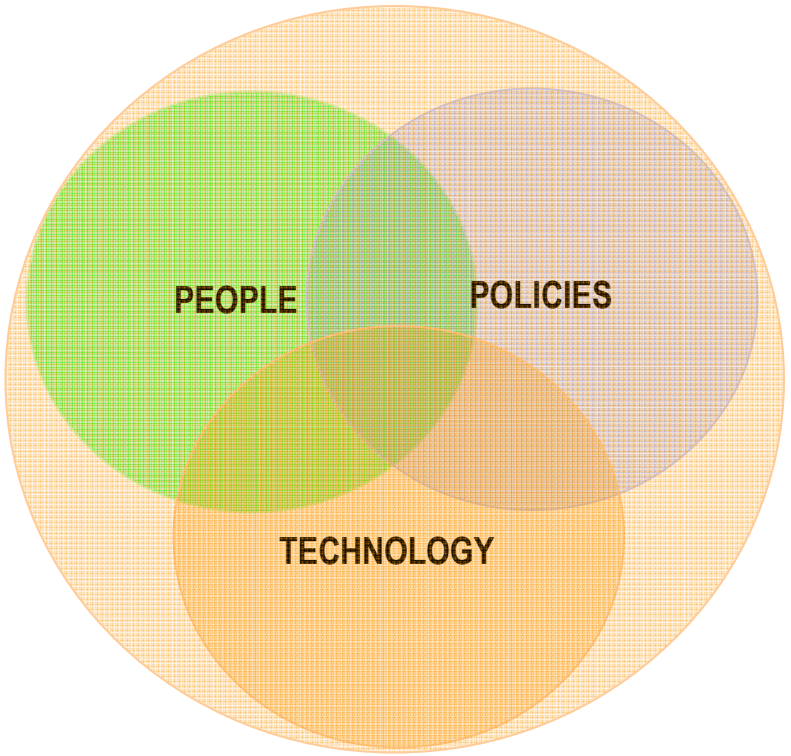
Data Stewardship -

- *It is a role assigned to a person responsible for maintaining data element in a metadata registry. Its main objective is to manage an organization's data assets in order to improve its integrity, usability, accessibility and quality. A Data Steward ensures that each data element –*
  - ***Has clear and unambiguous data definition***
  - ***Does not conflict with other data elements in the metadata registry***
  - ***Documents the origin & source of each metadata element***
  - ***Has adequate documentation on appropriate usage***
  - ***Has data security specification & retention criteria***

Module 5: > Topic 7: Data Governance – An Introduction

Characteristics of a Governed Organization -

At the **Governed** stage, an organization has a unified data governance strategy throughout the enterprise.
Data quality, data integration and data synchronization are integral parts of all business processes, & the organization achieves impressive results from a single, unified view of the enterprise



Module 5: > Topic 7: Data Governance – An Introduction

Characteristics of a Governed Organization -

PEOPLE

- ***Data governance has executive-level sponsorship with direct CEO support***
- ***Business users take an active role in data strategy and delivery***
- ***A data quality or data governance group works directly with data stewards, application developers and database administrators***
- ***Organization has "zero defect" policies for data collection and management***

Module 5: > Topic 7: Data Governance – An Introduction

Characteristics of a Governed Organization -

POLICIES

- ***New initiatives are only approved after careful consideration of how the initiatives will impact the existing data infrastructure***
- ***Automated policies are in place to ensure that data remains consistent, accurate and reliable throughout the enterprise***
- ***A service oriented architecture (SOA) encapsulates business rules for data quality and identity management***

Module 5: > Topic 7: Data Governance – An Introduction

Characteristics of a Governed Organization -

TECHNOLOGY

- ***Data quality and data integration tools are standardized across the organization***
- ***All aspects of the organization use standard business rules created and maintained by designated data stewards***
- ***Data is continuously inspected – and any deviations from standards are resolved immediately***
- ***Data models capture the business meaning and technical details of all corporate data elements***

Module 5: > Topic 7: Data Governance – An Introduction

Characteristics of a Governed Organization -

RISKS & REWARDS

- ***Risk: Low. Master data tightly controlled across the enterprise, allowing the organization to maintain high-quality information about its customers, prospects, inventory and products***
- ***Rewards: High. Corporate data practices can lead to a better understanding about an organization's current business landscape – allowing management to have full confidence in all data-based decisions***

Module 5: > Topic 7: Data Governance – Summary

- Having completed this topic, you should be able to:
  - What is Data Governance
  - Need for Data Governance
  - Advantages of Data Governance
  - Approach to implementing data governance
  - What is Data Stewardship
  - Characteristics of a Governed Organization

Module 5: > Topic 7: Data Governance – Review

Module 5: > Topic 3: Analytics & BI - Insight

- Analytics & BI
- Types of Analytics
- Query, Reporting & Search Tools
- OLAP, Visualization Tools
- Dashboards & Scorecards
- Predictive Analytics

Module 5: > Topic 3: Analytics & BI – Insight

Analytics & BI

- ***Analytics is the SCIENCE of ANALYSIS***

- *Analytics leverage data in a particular functional process (or application) to enable context-specific insight that is actionable." It can be used in many industries in real-time data-processing situations to allow for faster business decisions - Gartner*

- *Analytics is different from BI, although BI products play a role in analytics*

- *Application of Analytics include the study of business data using statistical analysis in order to discover and understand historical patterns with an eye to predicting and improving business performance*

- ***Applied Business Analytics is "Business Intelligence"***

Module 5: > Topic 3: Analytics & BI - Insight

Analytics & BI

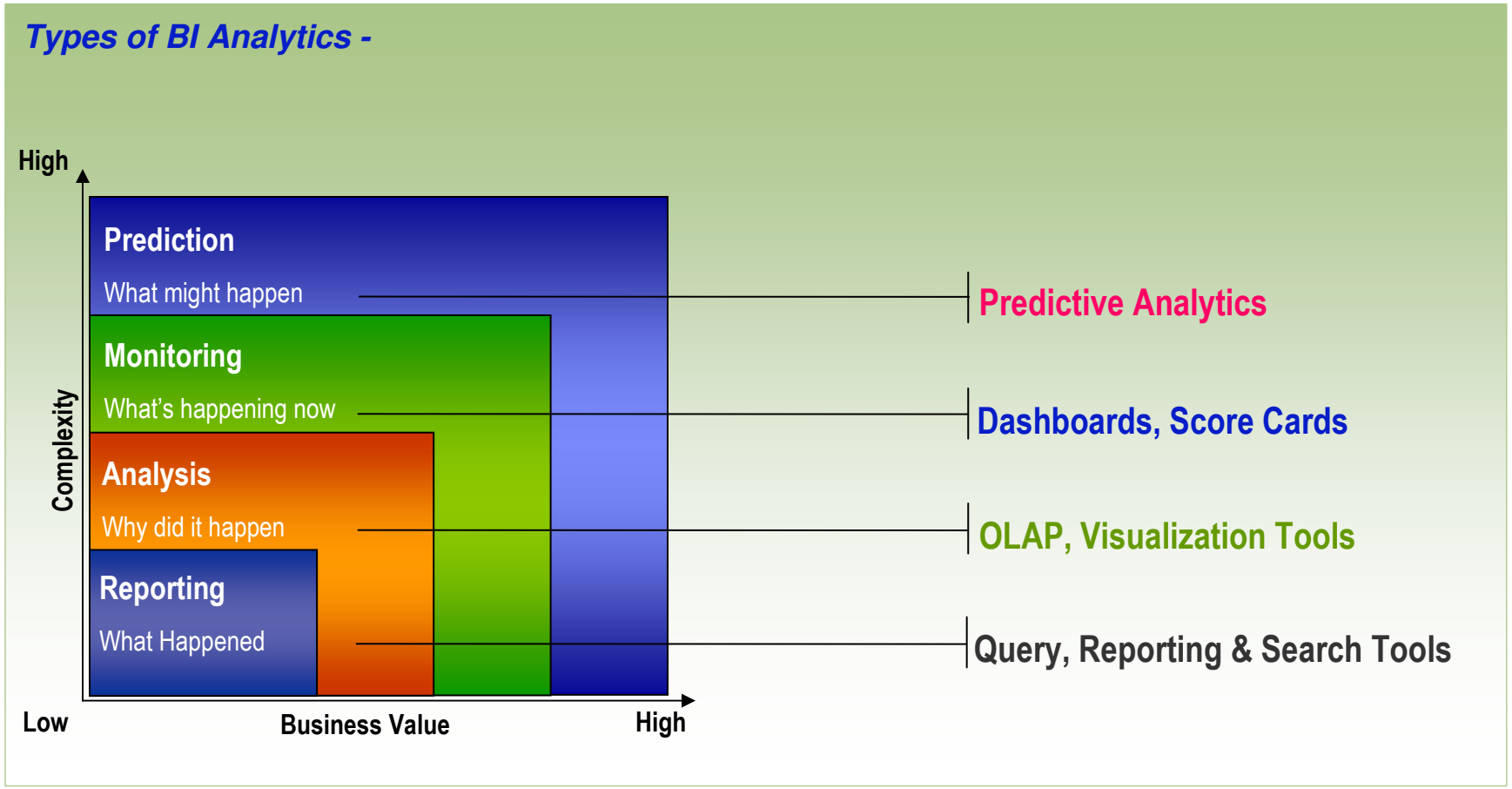
- *BI is neither a product nor a system.*
  "It is an architecture and a collection of integrated operational as well as decision-support applications and databases that provide the business community easy access to business data."

- *BI is all about how to **capture**, **access**, **understand**, **analyze** and turn one of the most valuable assets of an enterprise – **raw data** – into **actionable information** in order to improve business performance*

- ***Business Analytics** is the **analytical process** of **Reasoning**, **Forecasting**, & **Measuring** **Business Actions & Processes** based on extracted patterns in collected business data & business plans*

Module 5: > Topic 3: Analytics & BI - Insight

Types of Analytics



Module 5: > Topic 3: Analytics & BI - Insight

Query, Reporting & Search Tools

- *A Category of data access solution in which information is represented in the form of reports. Reporting Tools are also referred to as Query, Search & Reporting tools*

- *It presents state of data / information at that point in time in a report format*

- ***Types of Query & Reporting Tool –***
  - ***Ad Hoc Query & Reporting Tool***
  - ***Managed Query Tool (Canned Reports)***

Module 5: > Topic 3: Analytics & BI - Insight

OLAP, Visualization Tools

- ***Features***
  - ***Selection***
  - ***Drill Down***
  - ***Exception reporting***
  - ***Calculations***
  - ***Data Entry Options***
  - ***Web based Reporting***
  - ***Broadcasting***
  - ***Graphics***
  - ***Customization***
  - ***Printing***

Module 5: > Topic 3: Analytics & BI - Insight

OLAP, Visualization Tools

- ***Selection** – Is the criteria for filtering the number of records viewed based on some criteria which can be either fixed or user defined*

- ***Drill Down** – It is an action that opens the children of a selected parent. Drilling path could be based on the hierarchical structures defined in the database*

- ***Exception reporting** – It is a feature to spot the exceptional items. Implemented mostly using color coding. Percentile Analysis report is an example of this*

- ***Calculations** – Allows users to create simple calculations in the report including the four basic arithmetic calculations*

- ***Data Entry Options** – In some cases data entry from user end is allowed, gives more flexibility, however care should be taken database level security & other constraints are not violated*

Module 5: > Topic 3: Analytics & BI - Insight

OLAP, Visualization Tools

- ***Web based reporting*** *– A better way to provide basic OLAP reports to users anywhere, anytime*

- ***Broadcasting*** *– Feature which allows distribution of static reports to large no of static users. It also provides alerts in case of some data driven important events. Options available for broadcasting are Email, FAX, Mobiles & PDA's*

- ***Graphics*** *– Better representation of data. Helps quickly visualizing & analysis of data. Includes ability to switch between different graphical forms & representation of data*

- ***Customization*** *– Ability to customize report which includes creation of a template which when recalled displays the latest structures & members*

- ***Printing*** *– Feature to format a report for hard copy output. Trend is to integrate OLAP reports with third party reporting tools which have better printing capability*

Module 5: > Topic 3: Analytics & BI - Insight

Dashboards & Scorecards

- ***Dashboards & Scorecards*** *- They are multi layered performance management systems, build on Business Intelligence & Data Integration Infrastructure, that enable organizations to measure, monitor, & manage business activity using both financial & non-financial measures*

- ***Dashboards*** *– They tend to monitor the performance of operational process, they are able to display charts & tables with conditional formatting*

- ***Scorecards*** *– They tend to chart the progress of tactical & strategic goals, they use graphical symbols & icons to represent the status & trends of key metrics*

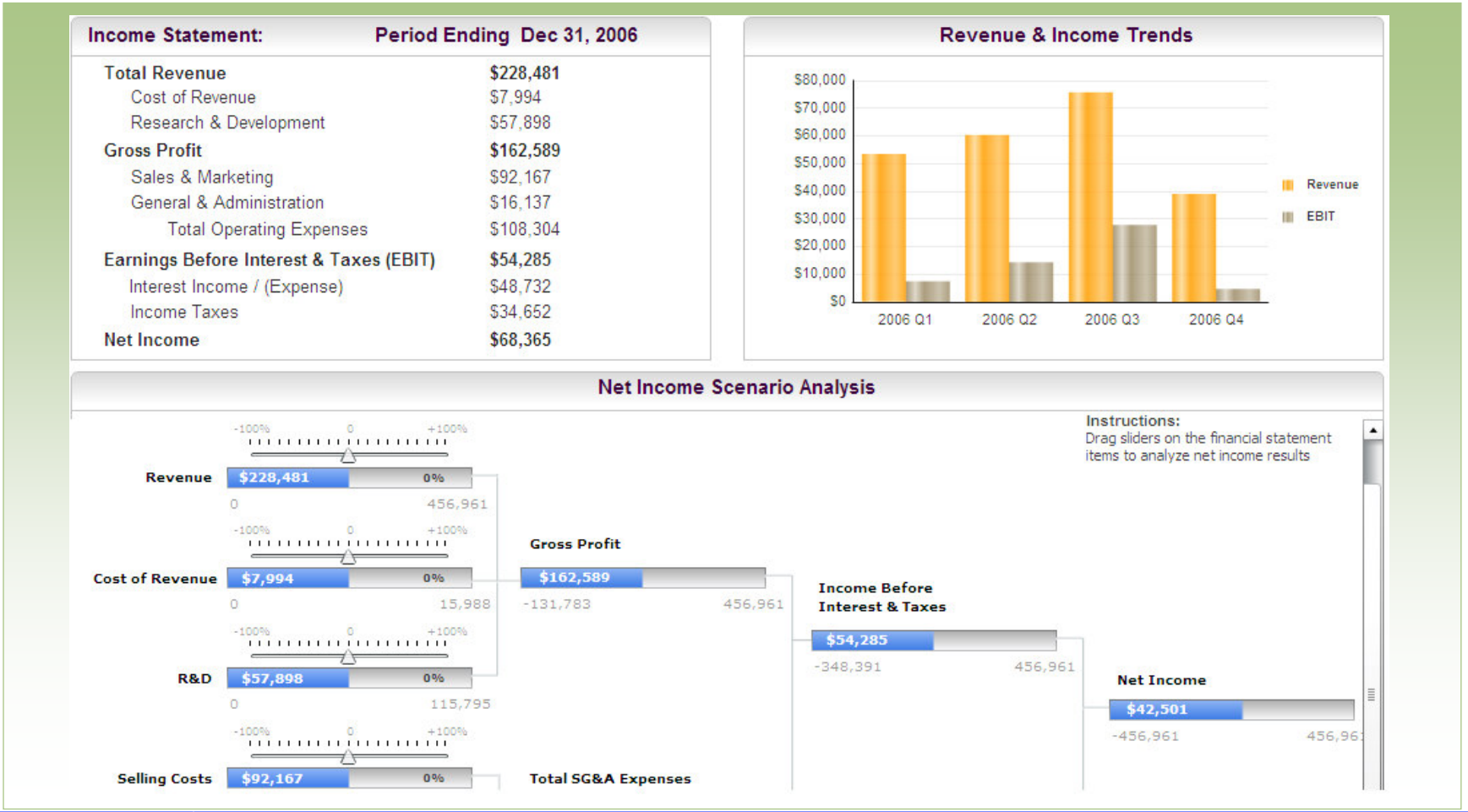
Module 5: > Topic 3: Analytics & BI - Insight

Dashboards & Scorecards - Comparison

	DASHBOARD	SCORECARD
Purpose	Measure Performance	Charts Progress
User	Managers, Staff	Executives, Managers, Staff
Updates	Real-Time to Right Time	Periodic Snapshots
Data	Events	Summaries
Top-Level Display	Charts & Tables	Symbols & Icons

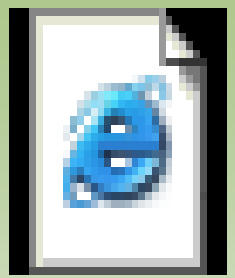
Module 5: > Topic 3: Analytics & BI - Insight

Dashboards & Scorecards - Dashboards - Example



Module 5: > Topic 3: Analytics & BI - Insight

Dashboards & Scorecards - Dashboards - Example



ManagementDashboardVFWebSite.swf

Module 5: > Topic 3: Analytics & BI - Insight

Predictive Analytics -

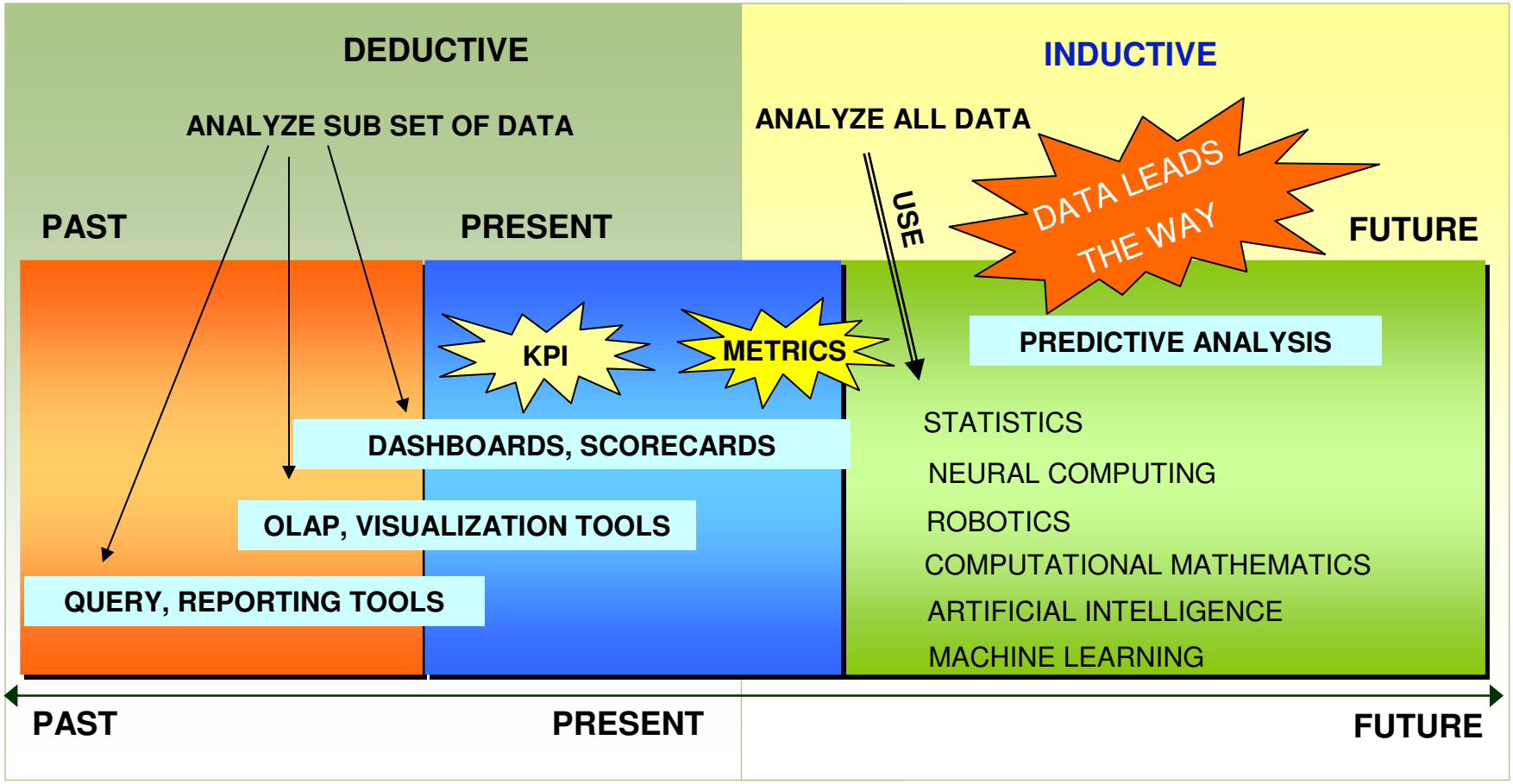
- *Predictive Analytics is a set of Business Intelligence technologies that uncovers relationships & patterns within large volume of data that can be used to predict behavior & events*

- *Unlike other BI technologies, predictive analytics is forward-looking, using past events to anticipate the future*

- *Predictive Analytics can help*
  - *companies optimize existing processes*
  - *Better understand customer behavior*
  - *Identify unexpected opportunities*
  - *Anticipate problem before they occur*
  - *And in doing so achieve breakthrough business results*

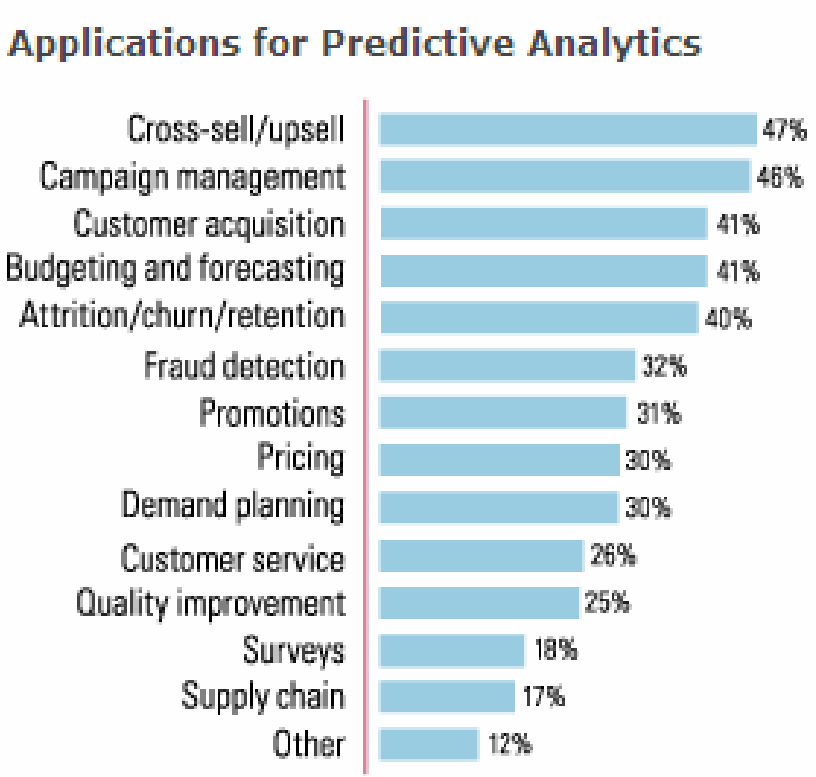
Module 5: > Topic 3: Analytics & BI - Insight

Predictive Analytics -



Module 5: > Topic 3: Analytics & BI - Insight

Predictive Analytics - Applications



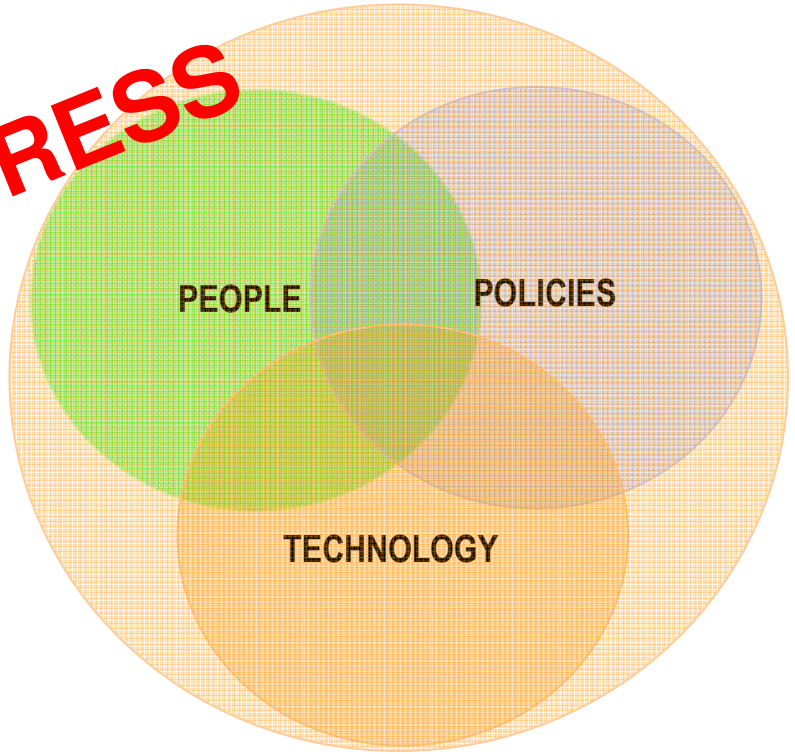
- ***It can identify the customers most likely to churn next month or to respond to next month's direct mail piece***
- ***It can anticipate when factory floor machines are likely to breakdown***
- ***Figure out which customers are likely to default on bank loan***

Module 5: > Topic 3: Analytics - Insight

BI Analytics -

At the **Governed** stage, an organization has a unified data governance strategy throughout the enterprise. Data quality, data integration and data synchronization are integral parts of all business processes, & the organization achieves impressive results from a single, unified view of the enterprise

WORK IN PROGRESS



Module 5: > Topic 3: Analytics - Insight Summary

- Having completed this topic, you should be able to:
  - What is Data Governance
  - Need for Data Governance
  - Advantages of Data Governance
  - Approach to implementing data governance
  - What is Data Stewardship
  - Characteristics of a Governed Organization

Module 5: > Topic 3: Analytics - Insight Review

WORK IN PROGRESS

References

- DM Review … www.dmreview.com
- Wikipedia … http://en.wikipedia.org
- Bill Inmon … www.billinmon.com
- Ralph Kimball … www.ralphkimball.com
- TDWI – The Data Warehousing Institute … www.tdwi.org