Welcome!

Module 3

- 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

© Copyright IBM Corporation 2006

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

© Copyright IBM Corporation 2006

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	

© Copyright IBM Corporation 2006

IBM Global Business Services

Course Title:

Module 3 : Data Warehouse Architecture - Part 1 - Generic

Disclaimer
(Optional location for any required disclaimer copy.
To set disclaimer, or delete, go to View | Master | Slide Master)

© Copyright IBM Corporation 2006

Module Objectives

- At the completion of this chapter you should be able to understand:
  - What is Data Warehouse Generic Architecture
  - Components of a Generic Data Warehouse Architecture
  - Data Warehouse Components - Terminology

© Copyright IBM Corporation 2006

Module 3: DW Architecture - Part 1- Generic : Agenda

- Topic 1. Data Warehouse Generic Architecture
- Topic 2. DW Components - Terminology

© Copyright IBM Corporation 2006

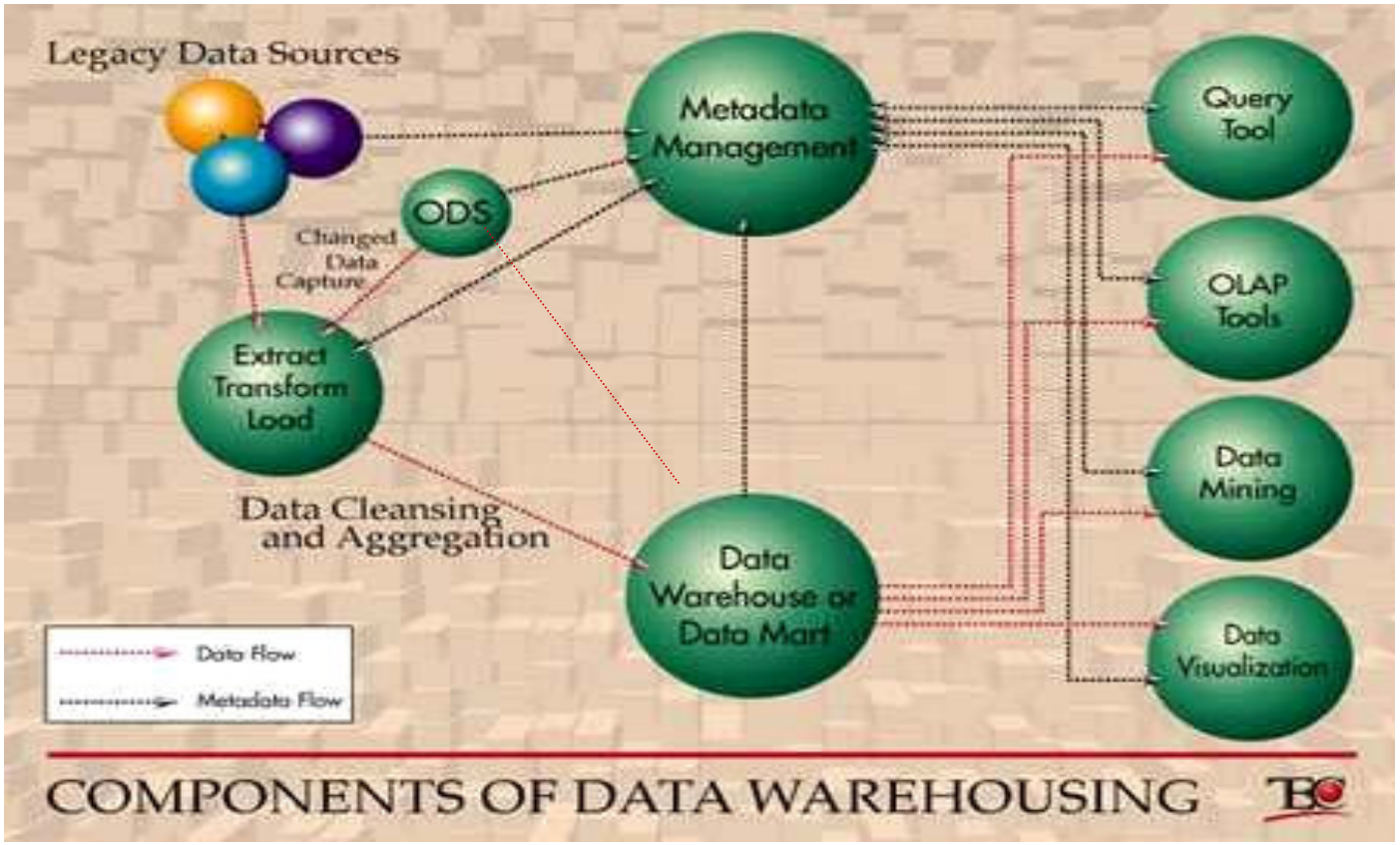
Module 3: > Topic 1: Data Warehouse Generic Architecture

- DW Data Flow Diagram
- DW Generic Architecture Diagram

© Copyright IBM Corporation 2006

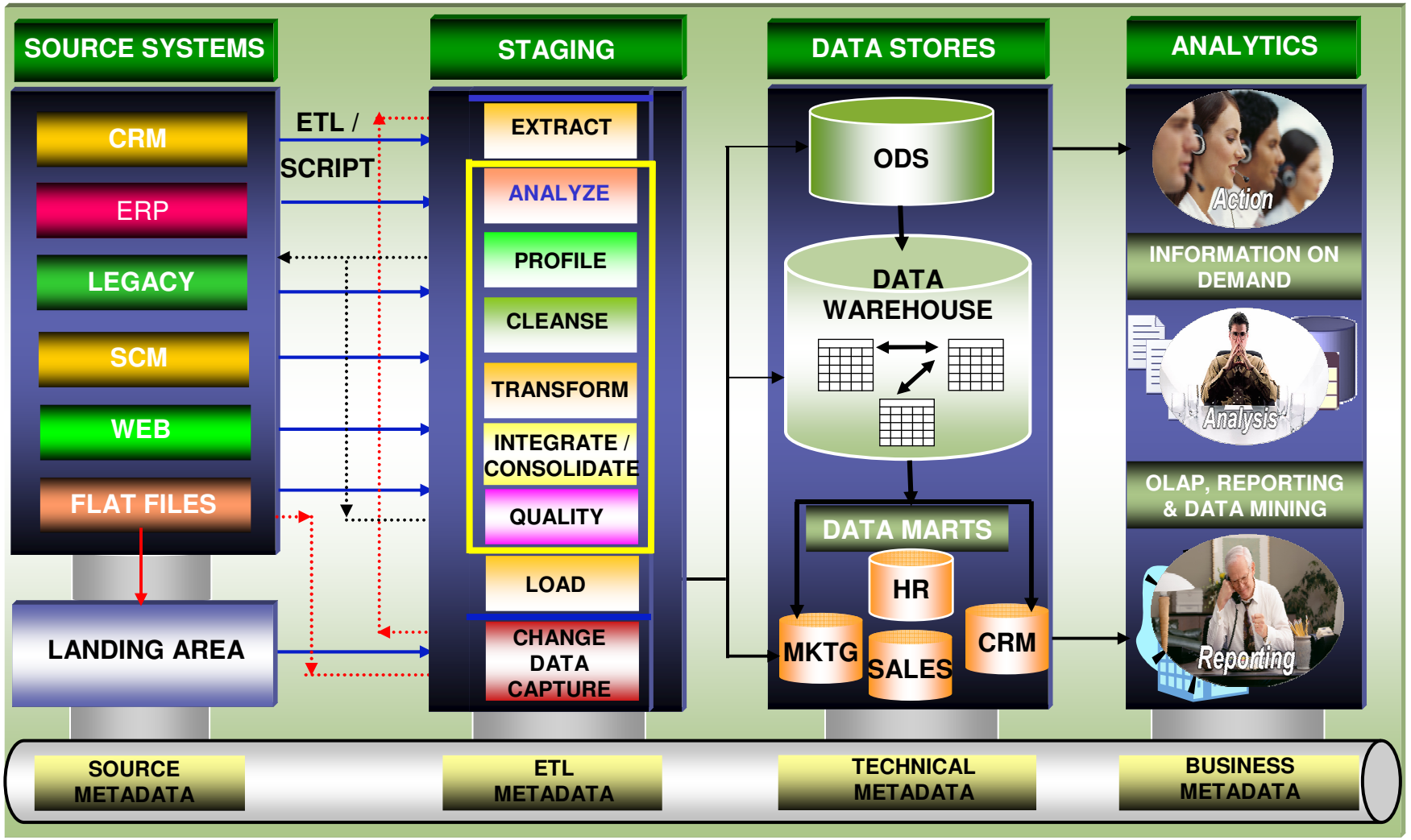
Module 3: > Topic 1: Data Warehouse Generic Architecture

DW Data Flow Diagram



© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture



© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Source Systems -

The Source Systems of a Data Warehouse can be Legacy data sources, ERP's, simple flat files to complex SAP sources, or COBOL sources or other data sources like RDBMS, AS400, Web application data, etc.

Commonly, these sources, or operational data (OLTP Data Sources) are known as Transactional Data Sources

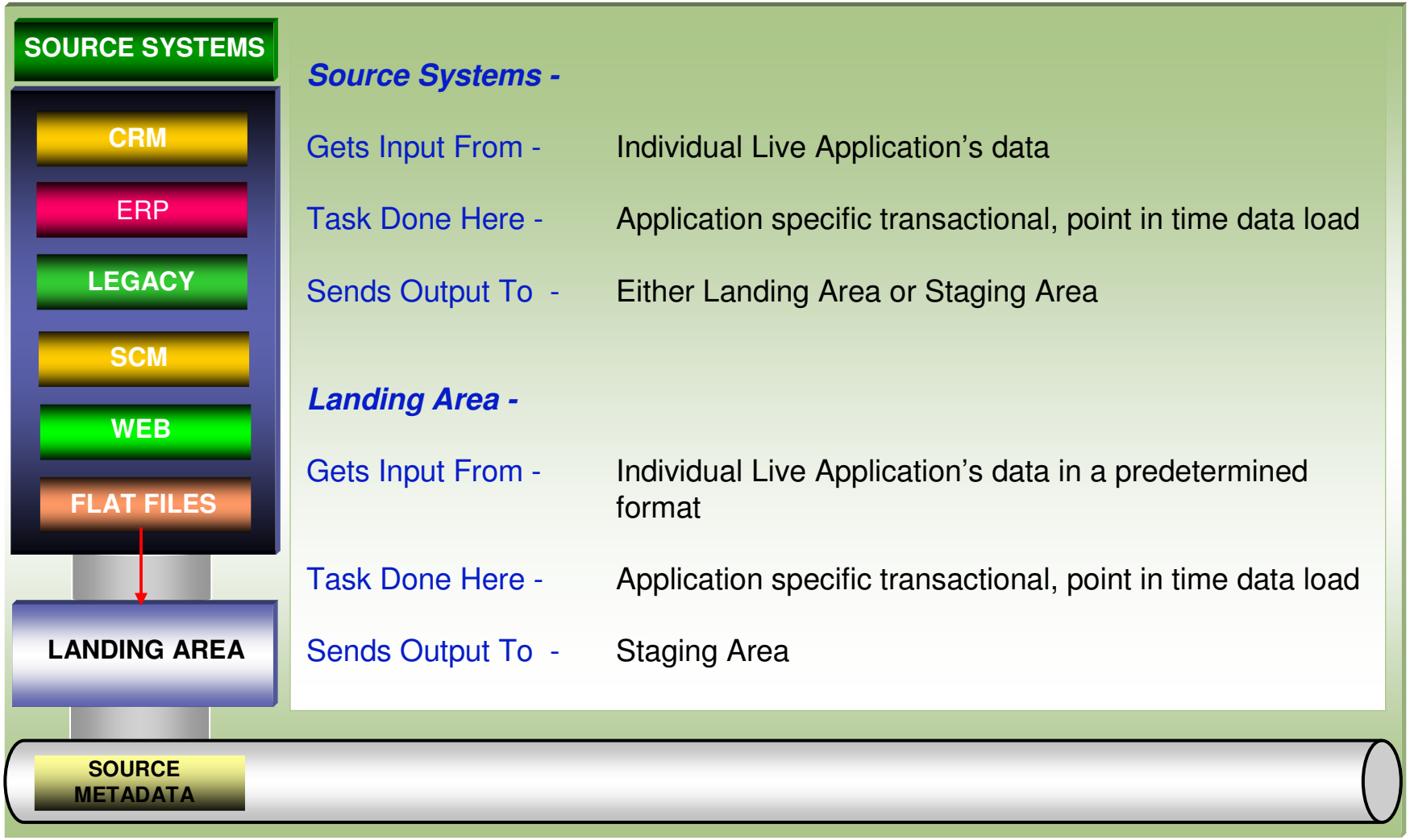
Landing Area -

A volatile intermediate area for operational data before transformation takes place. This is implemented to insulate the OLTP systems from the developer, avoid access on load on 'ONLINE' Source system applications, & to abide by federal laws in some cases. Source system data is either pushed or pulled into the Landing area in a pre-determined format from respective Source Systems. This data is then loaded into the staging area.

In some scenarios its quite possible that data can directly be sourced from the Source System area to the Staging area instead of routing it through the Landing Area

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture



Source Systems -

Gets Input From - Individual Live Application's data

Task Done Here - Application specific transactional, point in time data load

Sends Output To - Either Landing Area or Staging Area

Landing Area -

Gets Input From - Individual Live Application's data in a predetermined format

Task Done Here - Application specific transactional, point in time data load

Sends Output To - Staging Area

© Copyright IBM Corporation 2006

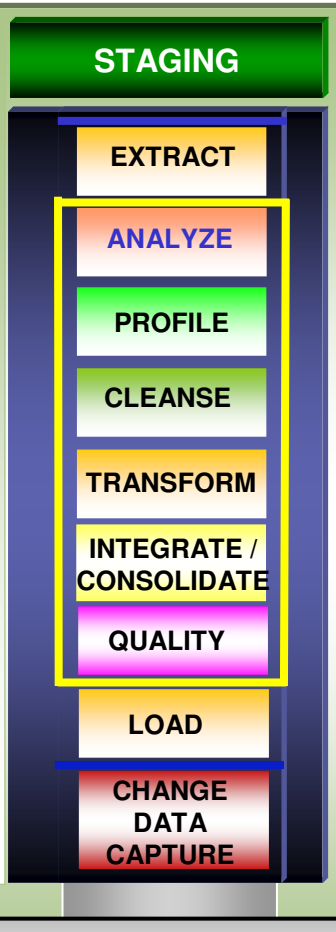
Module 3: > Topic 1: Data Warehouse Generic Architecture

Staging Area -

Gets Input From – Landing Area or Individual Source Systems

Task Done Here – Extraction, Cleansing, Transformation, Integration, Standardization of disparate source systems data to generate a 'COMPLETE' & 'CONFORMED' record.

Sends Output To - Volatile, Integrated, Point in Time data moved to either 'ODS' or 'DW' or 'DM'



Staging Area -

Is a place where you hold temporary tables on data warehouse server. We basically need staging area to hold the data, and perform data cleansing and merging, before loading the data into warehouse. Sometimes Staging Area is also required to hold a Subset of Source for Data Profiling activities

Data quality (information quality) is defined as standardizing and consolidating customer and/or business data

By cleansing / enhancing / merging / scrubbing the data and combining/aggregation related records to avoid duplicate entries you are able to create a single record view. Staging area can also hold reference & standardization tables

© Copyright IBM Corporation 2006

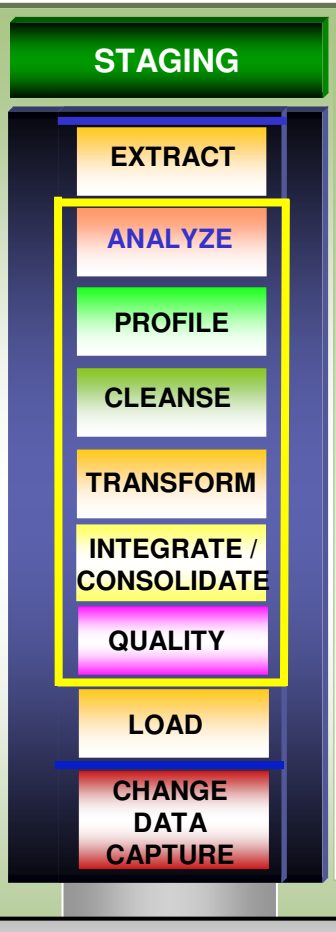
Module 3: > Topic 1: Data Warehouse Generic Architecture

CDC -
Change Data Capture

Is a set of software design patterns used to determine the data that has changed in a database so that action can be taken on the changed data.

CDC solutions occur mostly in DW environments since capturing and preserving the state of data across time is one of the core functions of a DW

It can be in source, in landing or in staging area



ETL -
Extract Transform Load

EXTRACT - Extract Data from either landing area, or directly from source systems using ETL tools preferably, or using custom scripts

TRANSFORM - Transformation would involve the following:

1. *Analyze the Data*
2. *Profile the Data (optional) – required for DQ*
3. *Cleanse the Data*
4. *Integrate the Data*
5. *Standardize the Data*
6. *Data Quality*

LOAD - Load integrated, complete, & conformed system of record into either 'ODS', 'DW' or 'DM'

© Copyright IBM Corporation 2006

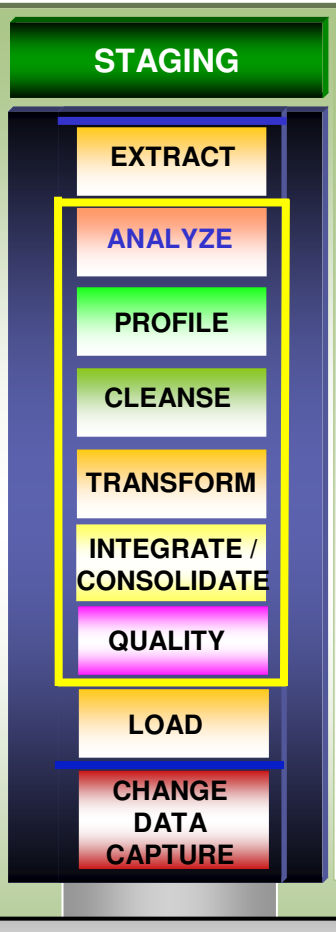
Module 3: > Topic 1: Data Warehouse Generic Architecture

CDC -
Change Data Capture

Gets Input From – Source System's Databases, or ETL Metadata Repository, or Log or Journal entries in databases either on Source, Staging, or Landing

Task Done Here – Verification of records that have been either inserted (new), Updated (modified), or deleted (removed)

Sends Output To - Perform necessary updated & deletes on Target systems



ETL -
Extract Transform Load

Gets Input From – Landing Area or Individual Source Systems or Staging Area

Task Done Here – Extraction, Cleansing, Transformation, Integration, Standardization of disparate source systems data to generate a 'COMPLETE' & 'CONFORMED' record. Generates ETL Metadata

Sends Output To - Integrated, complete & conformed record moved to either 'ODS' or 'DW' or 'DM'

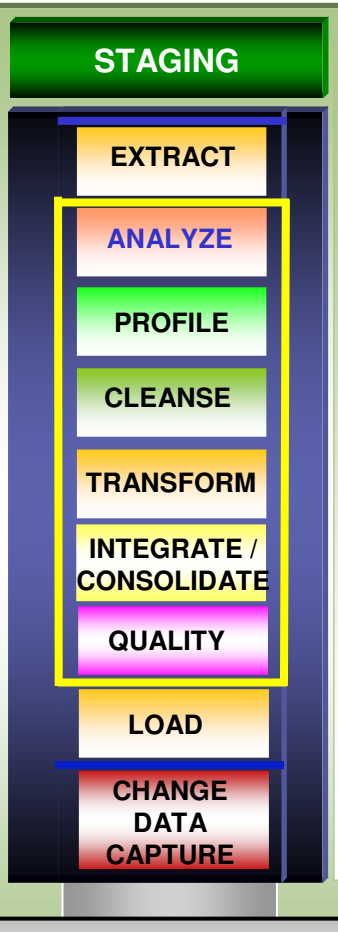
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. ***Analyze the Data***
2. *Profile the Data*
3. *Cleanse the Data*
4. *Integrate the Data*
5. *Standardize the Data*
6. *Data Quality*



ANALYZE THE DATA

This involves Analysis of metadata and data values; detection of differences between defined and inferred properties

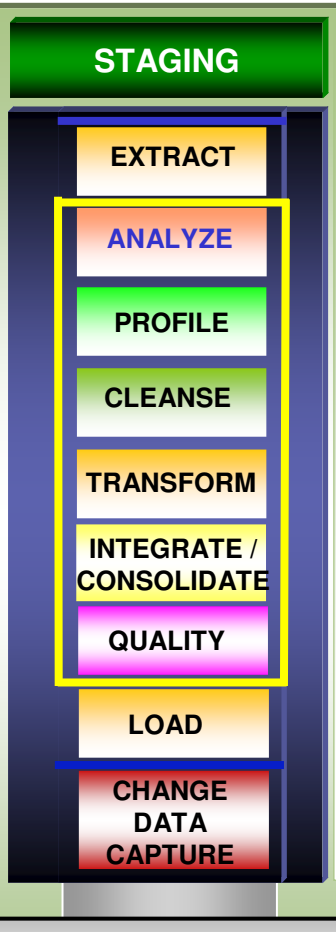
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. *Analyze the Data*
2. ***Profile the Data***
3. *Cleanse the Data*
4. *Integrate the Data*
5. *Standardize the Data*
6. *Data Quality*



PROFILE THE DATA

Data profiling is a process to assess current Data conditions, or to monitor data quality over time

It begins with collecting measurements about your data, and then looking at the results individually and in various combinations to see where anomalies exits

Data quality by understanding the metadata of your data sources (structure and the relationships within and among them) supported through efficient tooling

It helps to identify data entry errors.
It Avoids Data Migration re-work
It quantifies Data Corruption

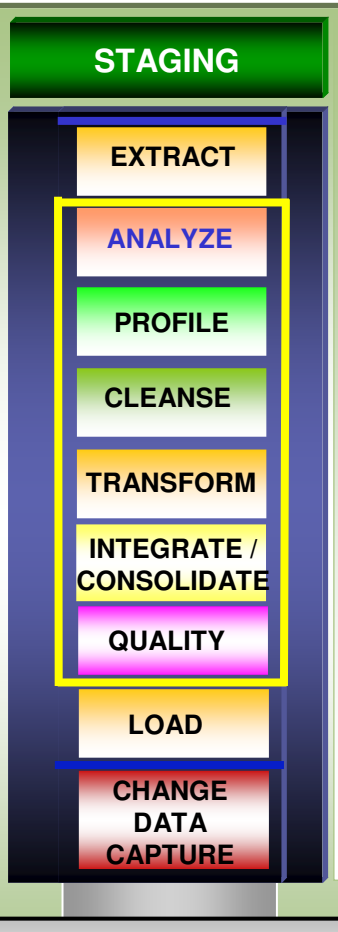
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. *Analyze the Data*
2. *Profile the Data*
3. ***Cleanse the Data***
4. *Integrate the Data*
5. *Standardize the Data*
6. *Data Quality*



CLEANSE THE DATA

Data cleansing also referred to as data Scrubbing is the act of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

The term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this dirty data.

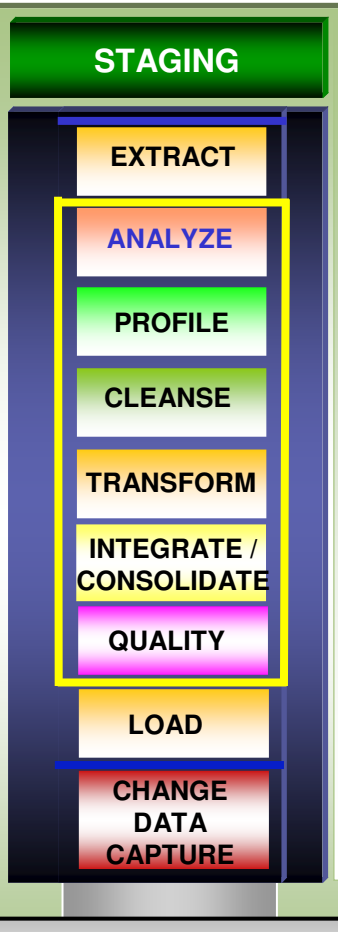
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. *Analyze the Data*
2. *Profile the Data*
3. *Cleanse the Data*
4. ***Integrated the Data***
5. *Standardize the Data*
6. *Data Quality*



INTEGRATE THE DATA

This involves integration & consolidation of data from various source systems to form a single system of record

Example – In Supply Chain, a product's Life cycle history make involve various stages like manufacturing, processing, inventory, shipping, fault & maintenance, etc. Each one of these process is possibly tracked by various OLTP applications (source systems)

Essentially to understand the complete lifecycle of this product means to integrate these Different records for these different processes Into a single system of record

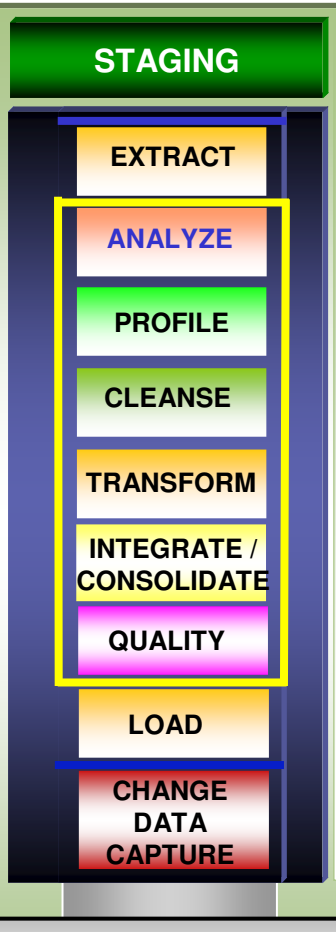
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. *Analyze the Data*
2. *Profile the Data*
3. *Cleanse the Data*
4. *Integrated the Data*
5. ***Standardize the Data***
6. *Data Quality*



STANDARDIZE THE DATA

Data standardization transforms different input formats into a consolidated output format. It helps in –

1. *Creating single domain fields*
2. *Incorporating business &*
3. *Industry standards*

Example – Application A processes Country name as 'INDIA', Application B processes Country name as 'IND'. Adhering to business rules we can then have either 'INDIA' as a standard name to refer to the country India.

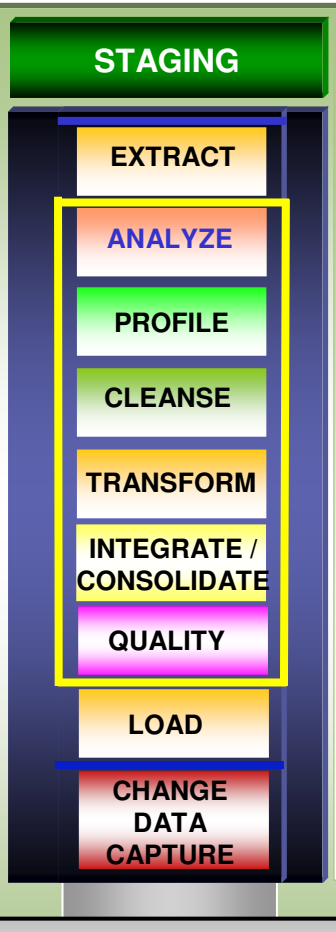
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ETL -

TRANSFORM -

1. *Analyze the Data*
2. *Profile the Data*
3. *Cleanse the Data*
4. *Integrated the Data*
5. *Standardize the Data*
6. ***Data Quality***



DATA QUALITY

Without accurate data users loose confidence In the data and make improper decisions

Data Quality addresses issues like -

1. *Business Rules violations like*
   1. *Missing data*
   2. *Use of Default (1, or 0, or 9999, or ?)*
   3. *Data with Logic embedded (Item code starts with1, product code starts with 9)*

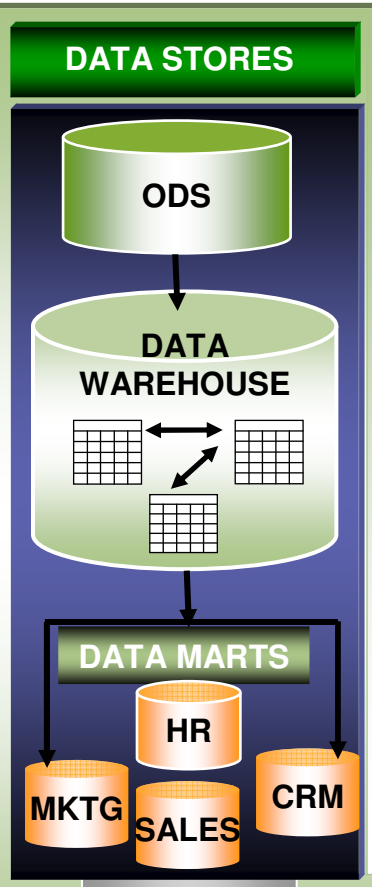
- *Data Integrity violations like*
  1. *Duplicate Primary Key*
  2. *One Entity have different key identifiers*
  3. *No reference data*
  4. *Multiple variation of same value*

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

ODS – Operational Data Store

Is a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-the-second, operational, integrated, collective information



Audience - *Data Analysts*

Data Model - *ER (Normalized), detailed & lightly summarized*

Database size - *Moderate*

Data Update - *Field by Field*

Philosophy - *Support Day to Day decisions & operational activities*

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

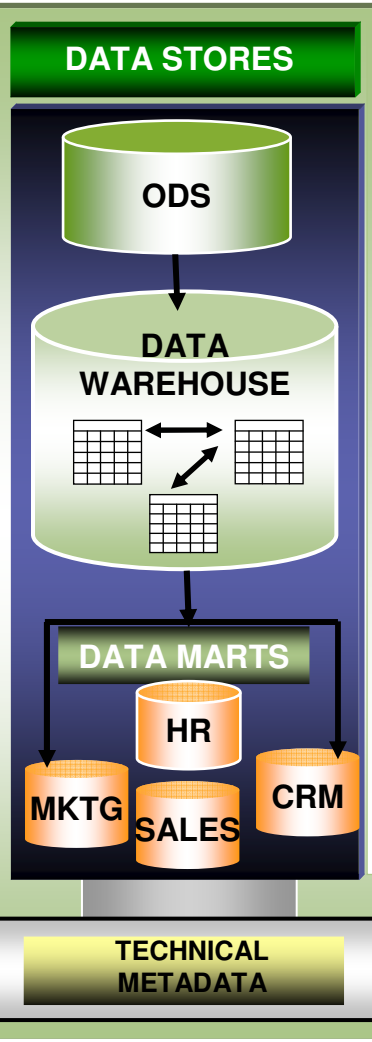
ODS – Operational Data Store

Is it OK to use a DW system to store the real-time updates instead of a separate ODS ?

Most Clients find that conflicting Service Level Agreements (SLA's) make it more desirable to have the ODS and Data warehouse (DW) operate on distinct schemas, or even totally separate environments.

Examples of conflicting SLA's include the need to have data in the ODS faster than it can be transformed into the standardized DW format, as well as irresolvable business priority issues between the ODS user and the demands placed on the DW by ad-hoc queries

Its principal key differentiators are the update frequency & the direct update paths from applications, compared to the controlled update path of data warehouse or data mart



ODS -

Gets Input From –
Staging Area

Task Done Here –
1. Data Storage for current period
2. Alters key structures
3. Reformats data
4. Lightly summarizes data
5. Recalculates data
6. Queried by analysts.

Sends Output To -
DW and/or DM.

© Copyright IBM Corporation 2006

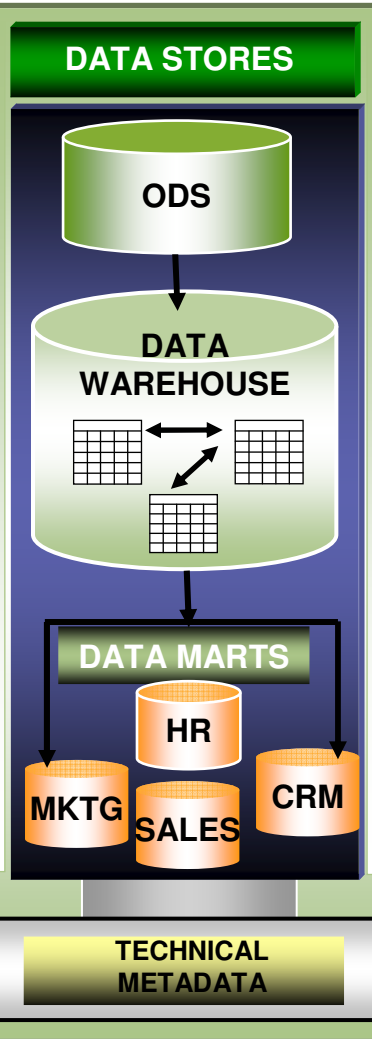
Module 3: > Topic 1: Data Warehouse Generic Architecture

DW - Data Warehouse

Is a "A subject oriented, integrated, non-volatile, time-variant, collection of data organized to support management needs."
(W. H. Inmon, Database Newsletter, July/August 1992)

Also referred to as CENTRAL DATAWAREHOUSE (HUB)

Data warehouse serves as a single-source hub of integrated data upon which all downstream data stores are dependent. The Data Warehouse has roles of intake, integration, and distribution



DW -

Audience - *Manager & Analysts*

Data Model - *ER (Normalized), detailed & historical*

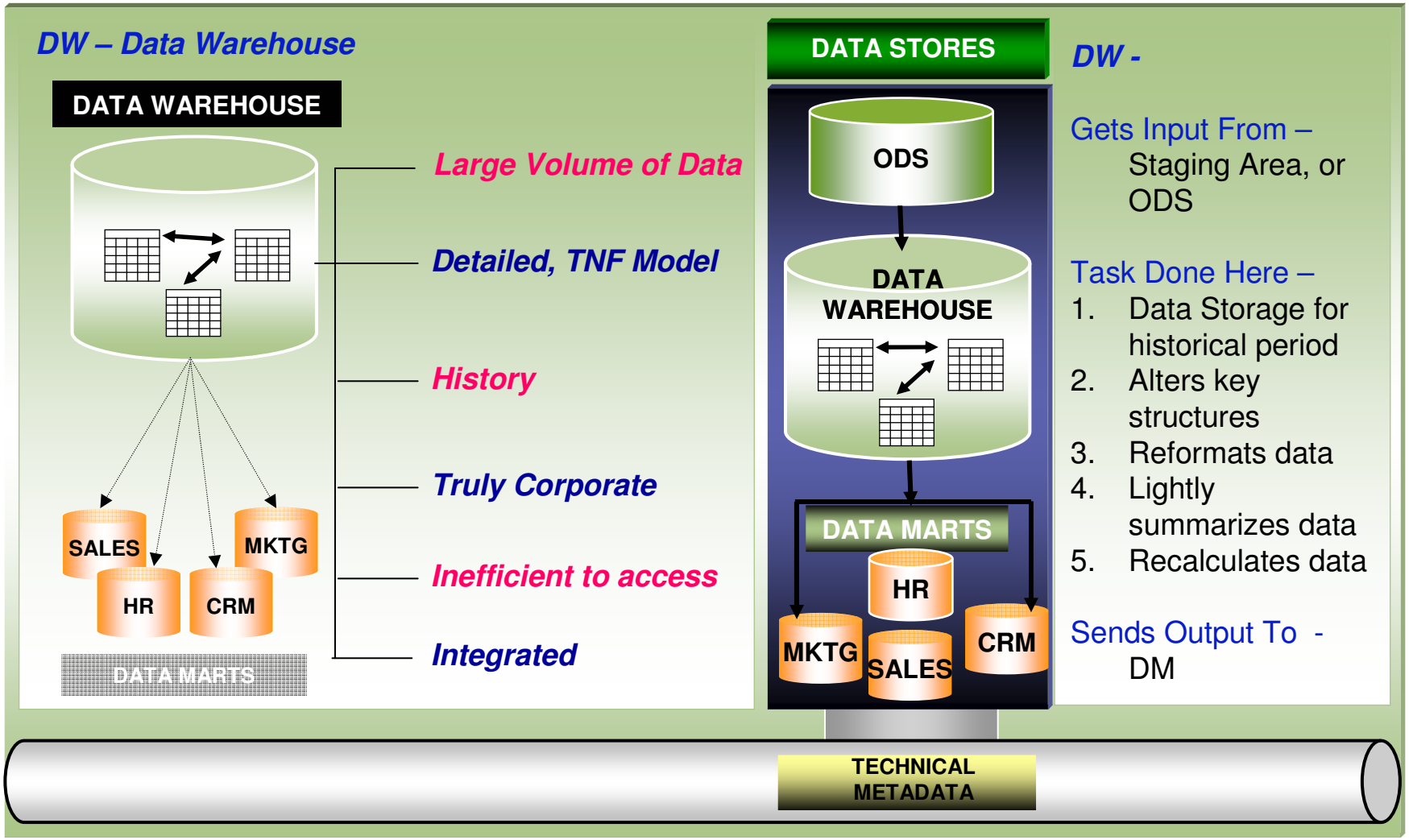
Database size - *Large to Very Large*

Data Update - *Batch, Controlled*

Philosophy - *Support managing the enterprise*

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture



© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Data Mart -

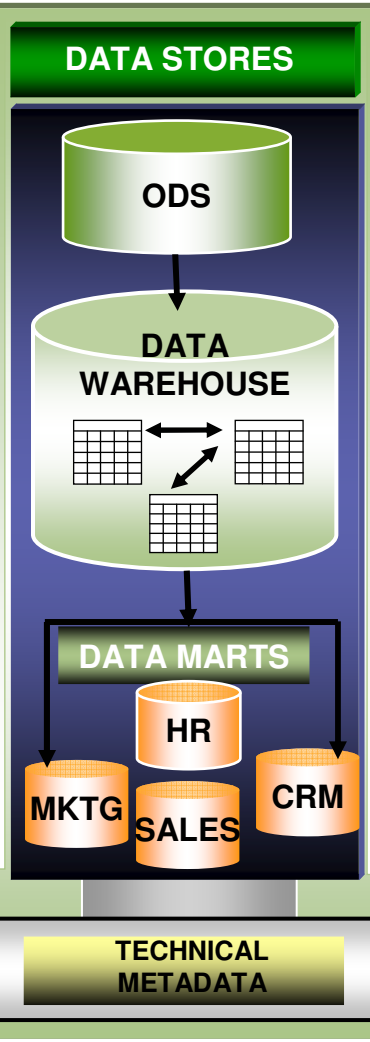
Data flows from the Data Warehouse to various Departments for their customized DSS usage. These departmental DSS data bases are called Data Marts

A Data Mart is a body of DSS data for a department that has an architectural foundation of a data warehouse

A Data Mart can also represent a Business Process that can proliferate across many departments

There are several names for data marts. Those names include:

1. *Departmental DSS data bases*
2. *OLAP data bases*
3. *Multi Dimensional database (md DBMS)*



Data Mart -

Audience - Executives. Manager & Analysts

Data Model - Dimensional, Summarized

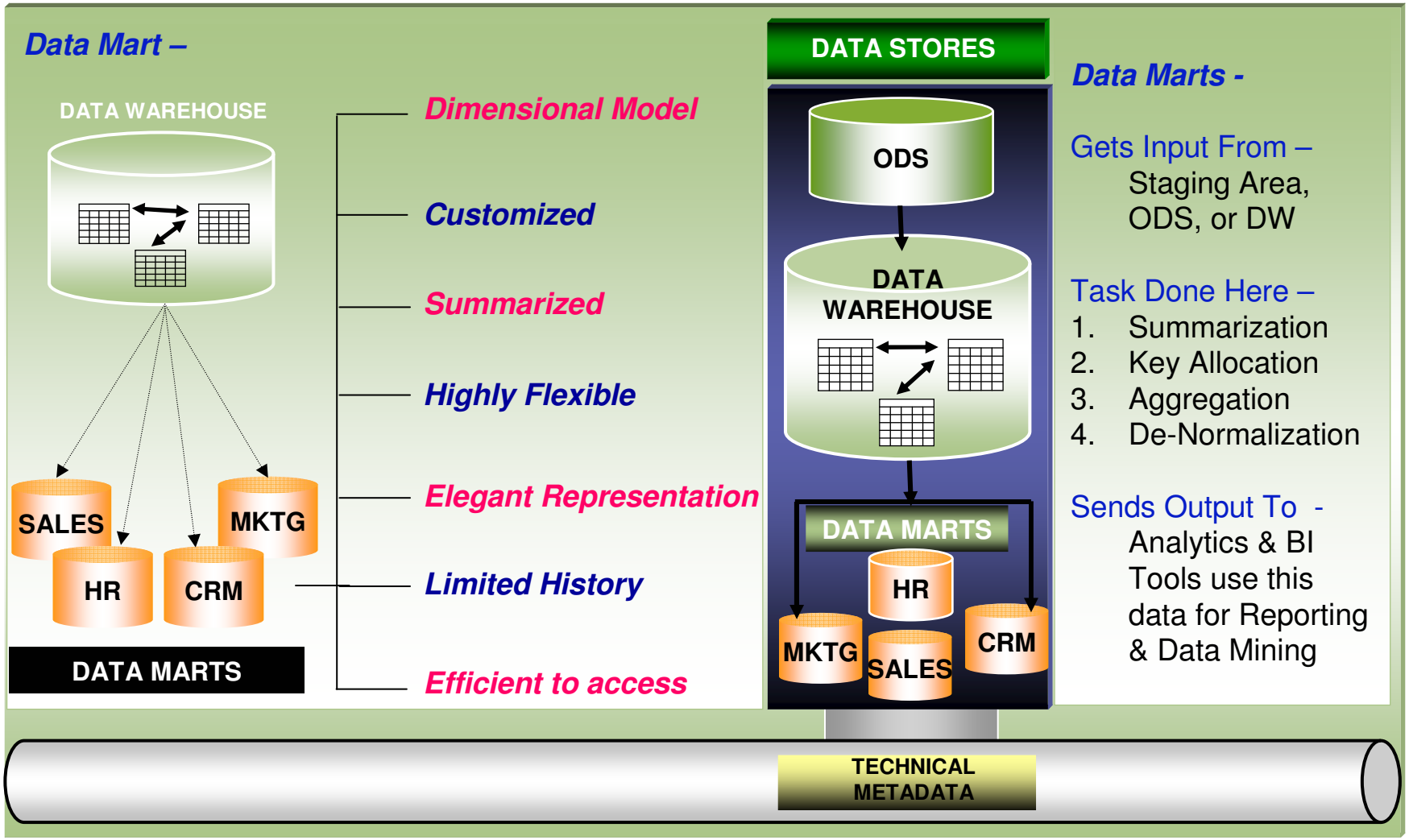
Database size - Moderate to Large

Data Update - Batch, Controlled

Philosophy - Operational Efficiency

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture



© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Characteristics	ODS	Data Warehouse	Data Mart
Audience	Data Analysts	Managers & Analysts	Executive Mangers & Analyts
Data Model	ER, Detailed & Lightly Summarized	ER, Normalized, Detailed & History	Dimensional, Summarized
Database Size	Moderate	Large to Very Large	Moderate to Large
Data Update	Field by Filed	Batch, Controlled	Batch, Controlled
Philosophy	Support Day to day decisions & operational Activities	Support Managing the Enterprise	Operational Efficiency
Type of Data	Detailed, Point in time, Integrated	Detailed, Historical, Integrated	Business Process Specific &/or Departmental, Summarized

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Analytics –

Analytics is "The SCIENCE of Analysis"

It defines how a Business or an Entity arrives at an optimal or realistic decision based on existing data

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

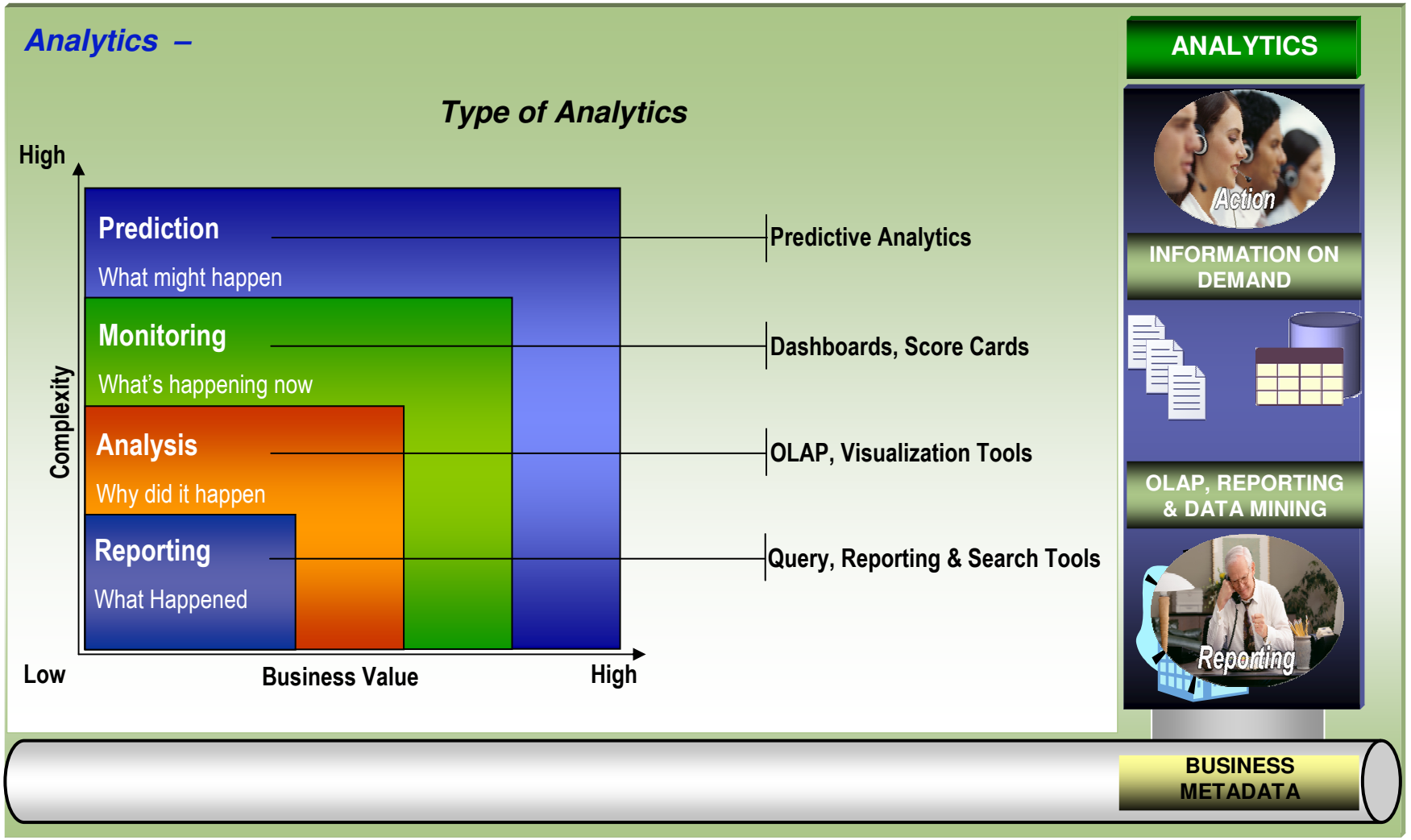
In other words ***Applied Business Analytics*** *is "****Business Intelligence****"*

BI Analytics consists of -

1. ***Query, Reporting & Search Tools***
2. ***OLAP, Visualization & Data Mining Tools***
3. ***Executive Dashboards & Scorecards***
4. ***Predictive Analysis Tools***

© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture



© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Metadata –

Two contractors are assigned a task of building a bridge. One is to start building from East end and the other is to start building from the West end. Both have to meet in the center and then merge. When they arrived at the center point one end of the bridge was higher than the other by a few inches. This was because one group of contractors & their engineers used kilograms and meters, while another used pounds and feet. It caused the parent company losses in billions!.

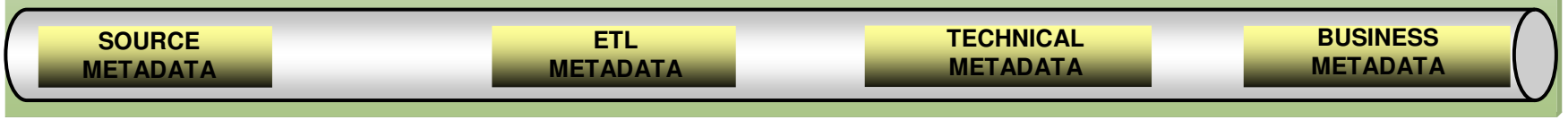
Reason - ***It wasn't the data that was faulty; it was the metadata!***

*Metadata is "**Data about Data**"*

*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 :

1. *What kind of Data ?*
2. *Who is the owner of this data ?*
3. *How was the data created ?*
4. *What are the attributes and significance of the data created or collected ?*

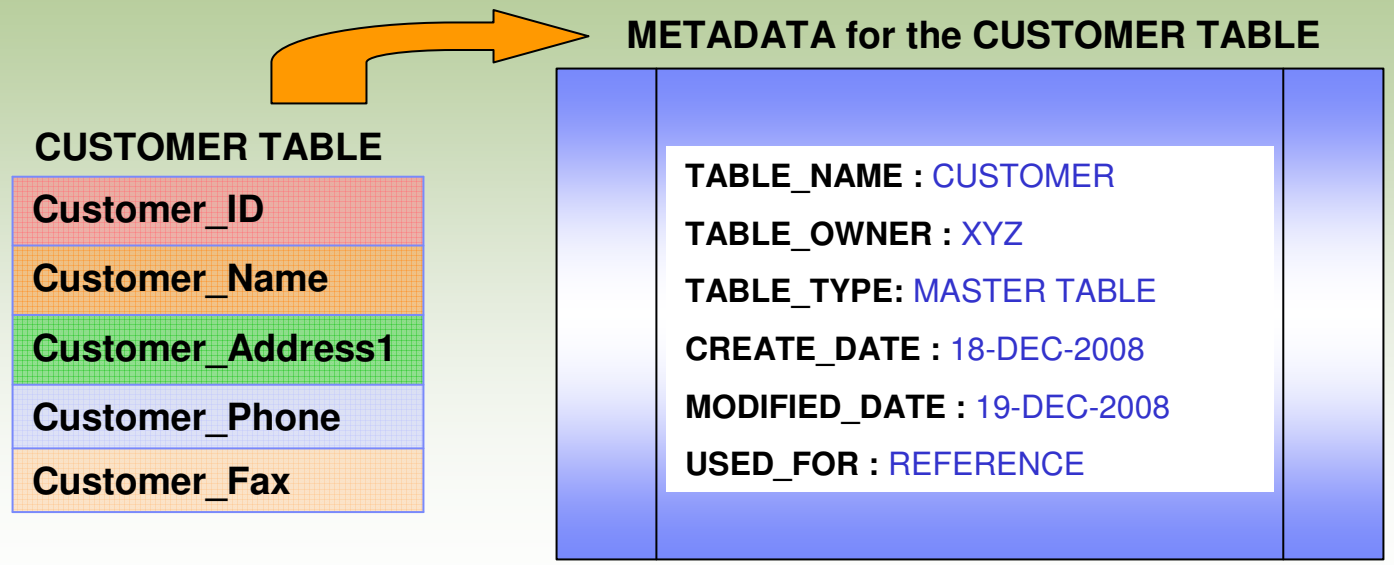


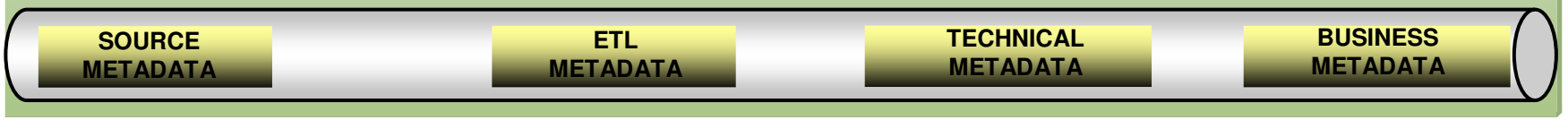
© Copyright IBM Corporation 2006

Module 3: > Topic 1: Data Warehouse Generic Architecture

Metadata –

Metadata for a Customer Database Table -





© Copyright IBM Corporation 2006

Module 3: > Topic 1: DW Generic Architecture Summary

- Having completed this topic, you should be able to:
  - Generic Data Warehouse Architecture
  - Components of Data Warehouse Architecture like –
    - Source Systems
    - Landing Area
    - Staging Area & its sub-components
    - ETL & brief about it
    - ODS – Operational Data Store
    - DW – Data Warehouse
    - DM – Data Marts
    - Analytics & its types

© Copyright IBM Corporation 2006

Module 3: > Topic 1: DW Generic Architecture Review

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***Data Warehouse****: A Data Warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process*

- ***Data Warehousing****: Intrinsic process of Design, Build & Maintenance of a Data Warehouse*

- ***ETL****: Extract, Transform & Load. Is a set of process by which operational data is prepared for Data Warehouse*

- ***Metadata****: Data about Data. There are broadly two types of metadata, Technical Metadata and Business Metadata*

- ***Operational Data Store****: Subject-oriented, integrated, volatile, current valued data store, containing* only corporate detailed data. Stores Operational/Current or Near Real time data

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***Data Mart***: *Subset of a data derived from Data Warehouse which caters to a specific business process or department*

- ***Data Mining***: *A class of undirected queries against the most atomic data set to that seek to find unexpected trends in data*

- ***Dimensional Model***: *A data model suited for representation of data in a Data Warehouse environment. It is business friendly and consists of two types of tables namely Dimension and Facts (Measures). It allows for high performance access to data*

- ***Dimension***: *It is the entity that one tries to model in a Data Warehouse environment. Customer, Product, Date, Time, Location are examples of Dimensions. The data represented in a Dimension Table is descriptive*

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***Fact**: It consists of business metrics or measurements for data that is represented in a dimensional model. Profit Margin, Sales, Revenue are examples of a Facts. Most often the data represented in a Fact Table is numeric*

- ***STAR Schema**: It's a form of representation of data in a Dimensional Model. It consists of Single Data (FACT) Table surrounded by multiple descriptive (DIMENSION) tables resulting in a STAR like structure. Provides high speed access to data.*

- ***SNOWFLAKE Schema**: It's a form of representation of data in a Dimensional Model. Low Cardinality Redundant attributes of a dimension are move to a sub-dimension tables. Useful for maintenance of rapidly changing large dimensions.*

- ***Hierarchy**: It defines the navigational path for Drill Up & Drill Down. Dimension tables can represent multiple hierarchical roll-ups*

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***OLAP**: Online Analytical Processing. It provides the end users an optimal way of browsing through the DW data*

- ***MOLAP**: Multidimensional OLAP. Data stored in multidimensional Cubes. Base data & pre-calculated results stored in multi-dimensional arrays. Each Side of the Cube represents an entity or DIMENSION and the cell within the cube represents a measure or FACT*

- ***ROLAP**: Relational OLAP. Data Storage in Relational form. Access data stored in Relational Data Warehouse for OLAP analysis*

- ***HOLAP**: Hybrid OLAP. Combination of MOLAP & ROLAP. RDBMS used for Data Storage and MDB used for high speed data OLAP data access, analysis & calculations*

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***Data CUBES****: Structure for storing & representing data in a multi-dimensional form. Each Side of the Cube represents an entity or DIMENSION and the cell within the cube represents a measure or FACT*
- ***Drill Across****: Data analysis across Dimensions*
- ***Drill Down****: Data analysis to a child attribute*
- ***Drill Through****: Data analysis that proliferates from an OLAP cube into a relational database*
- ***Drill Up****: Data analysis to a parent attribute*
- ***Slice and Dice****: Operations for browsing data through visualized cubes*

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Components - Terminology

- ***Aggregation****: Process for summarization of data for faster retrieval. Facts or measures are summed up across required Dimensions. The resulting aggregate table has fewer rows, thus making retrieval of data faster*

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Comp. - Terminology Summary

- Having completed this topic, you should be able to understand:

  DW Components Terminology Like -

  - Data Warehouse
  - Data Warehousing
  - ETL
  - Metadata
  - ODS
  - Data Mart
  - Dimensional Modeling
  - Dimension
  - Fact
  - Star Schema
  - Snow Flake Schema
  - Hierarchy
  - OLAP
  - ROLAP
  - MOLAP
  - HOLAP
  - Drill Through
  - Drill Across
  - Drill Down
  - Drill Up
  - Data Cubes
  - Slice and Dice
  - Aggregation
  - Data Mining

© Copyright IBM Corporation 2006

Module 3: > Topic 2: DW Comp. - Terminology Review

© Copyright IBM Corporation 2006

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

© Copyright IBM Corporation 2006