Lecture 6

Data Integration and Metadata

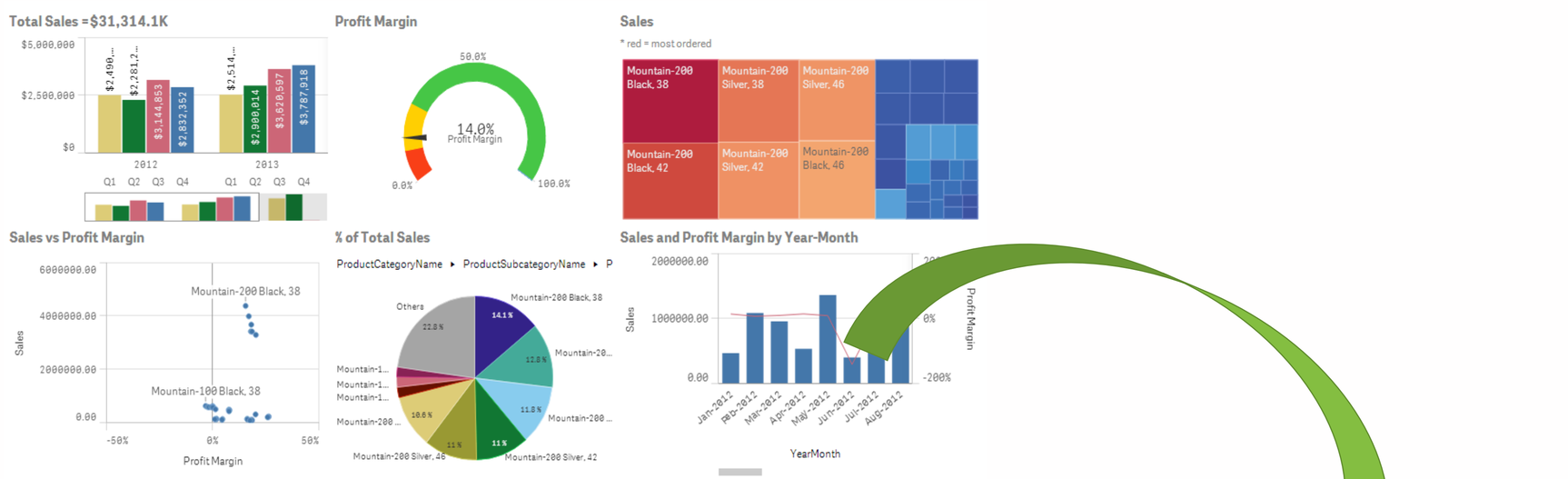
Dr Bikesh Raj Upreti
E-mail: b.upreti@uq.edu.au
Room: Joyce Ackroyd R530

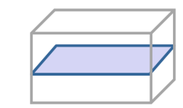
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

UQ BUSINESS SCHOOL

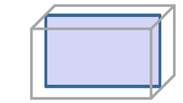
Recap: Dashboard Navigation - Data Cube Operators

- Sub-setting
  - Slicing
  - Dicing
- Navigation
  - Drill-down
  - Roll-up
  - Drill-across
- Pivoting

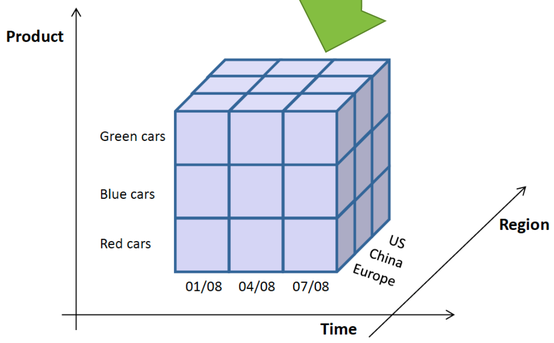


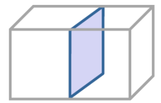


Product Manager's view:
Blue cars in all regions in all periods.

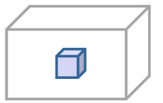


Regional Director's view:
All cars in China in all periods.



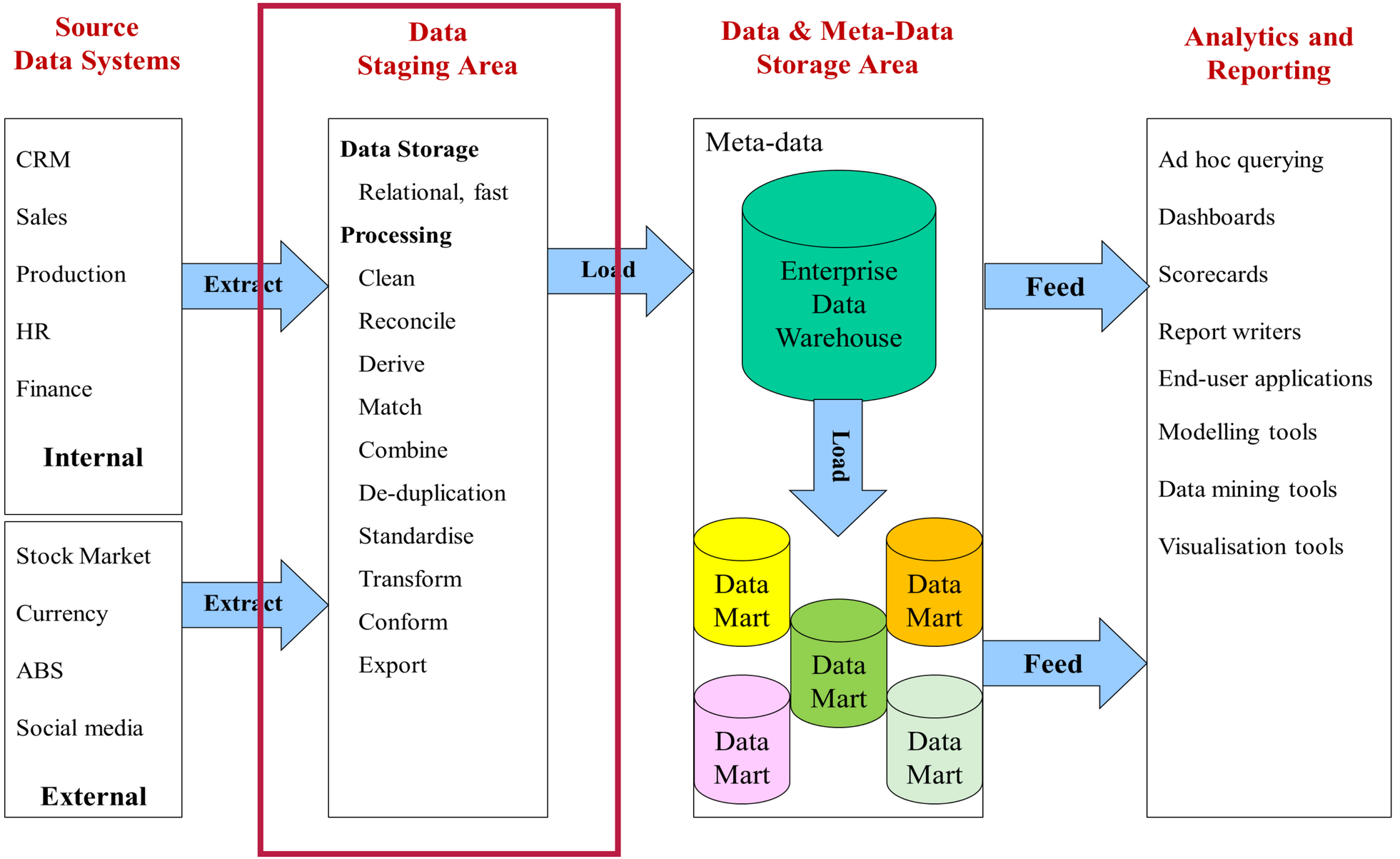


Controller's view:
All cars in all regions in a specific period.



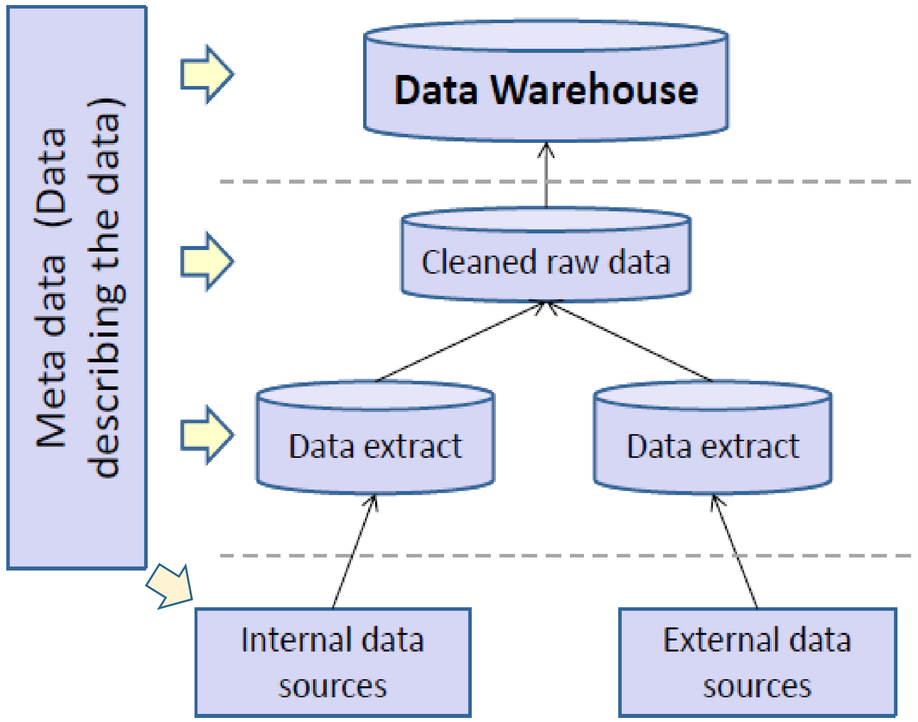
Ad-hoc view:
Specific car in a specific region in a specific period.

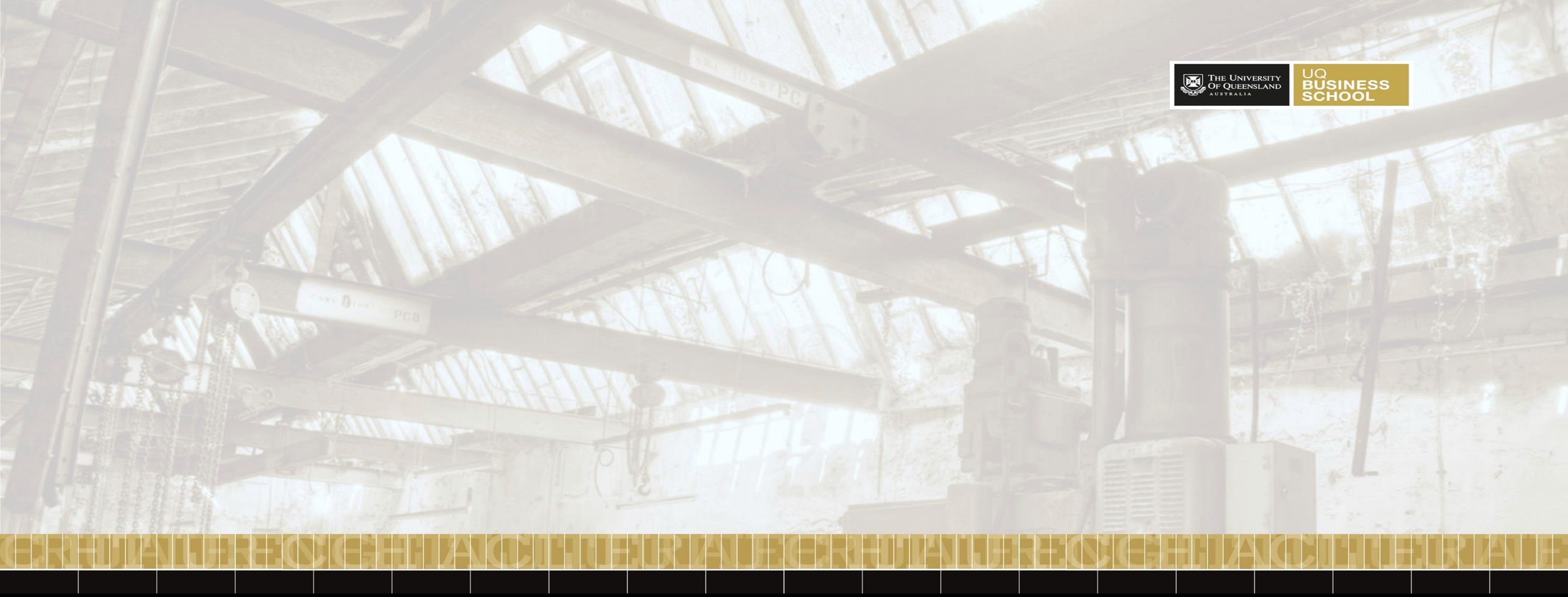
Recap: Business Analytics Framework



Agenda for today

- Data Extraction, Transformation and Loading
  - Overview
  - Requirements and Steps
  - Data Extraction
  - Data Loading
  - ETL
- Meta Data
  - Importance
  - Types
  - Business
  - Technical
  - Providing Metadata

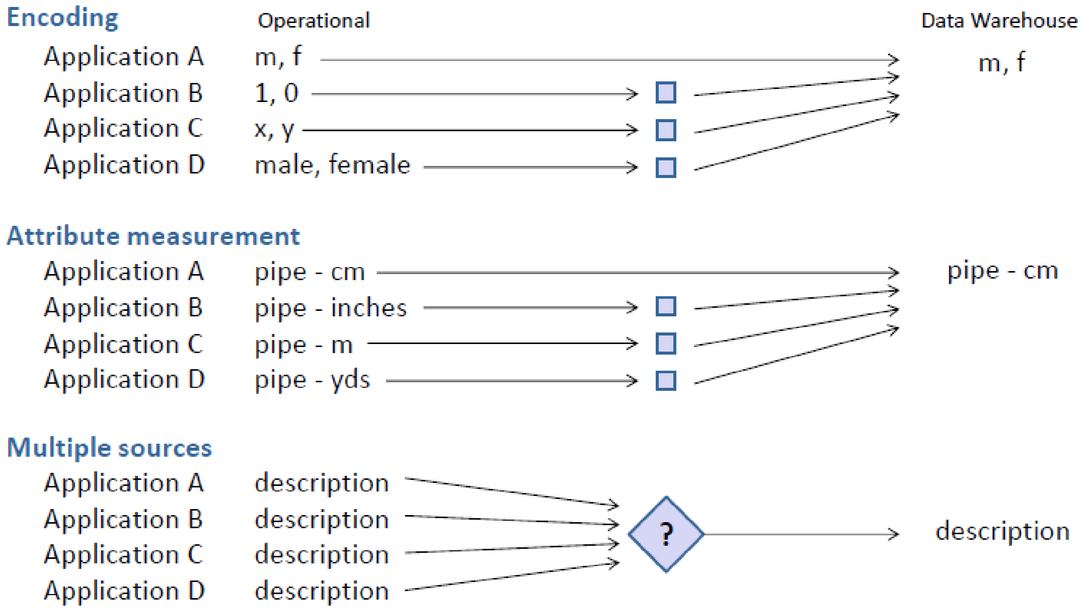


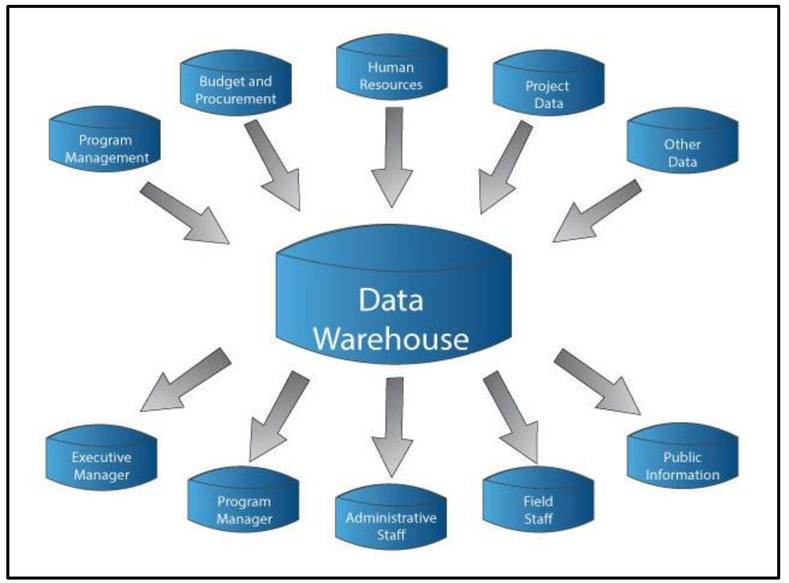


Data Extraction, Transformation and Loading (ETL)

WHY ETL?

- Move data from transactional environment to analytical environment
- Create a single version of truth
- Integrated high quality data!





- ✓ Subject-oriented
- ✓ Integrated
- ✓ Time variant
- ✓ Non-volatility

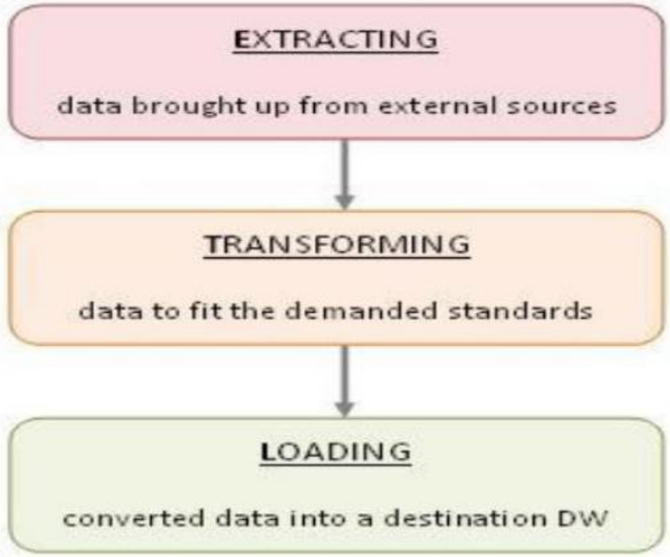
ETL process

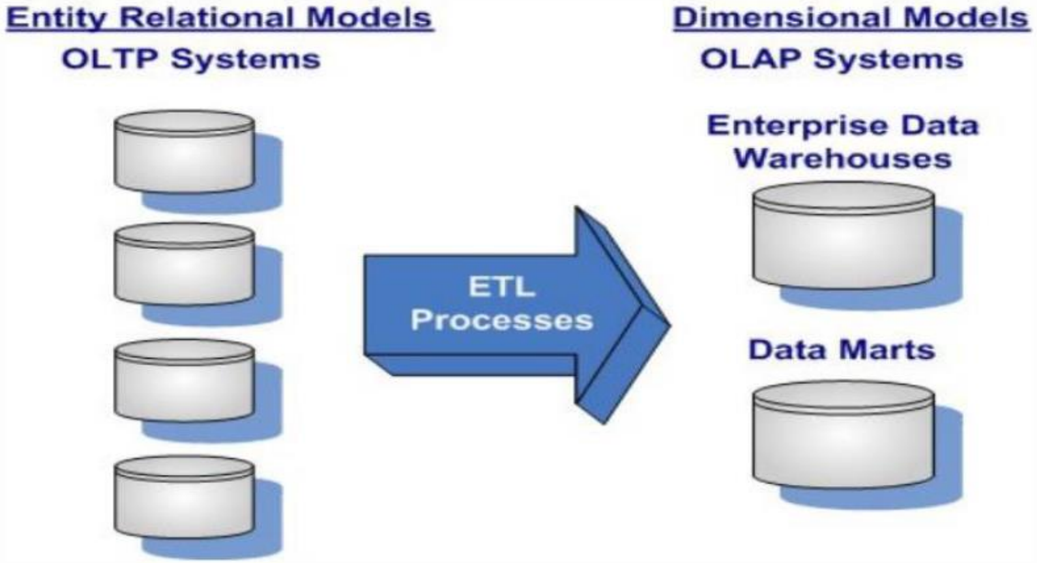
- Extract, Transform, Load
- We are essentially talking about the integration of enterprise data
- Ensuring we have high quality data for decision-making

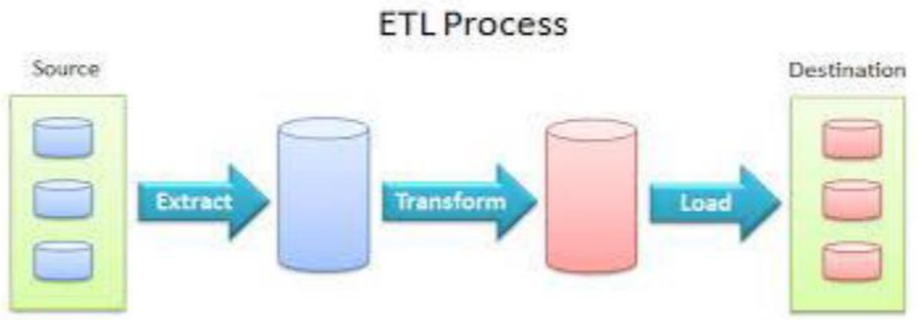
- Overview of ETL
  - Purpose is to load DW with integrated and cleansed data
  - Most important and most challenging activity for DW
  - Time consuming and arduous

Video – What is an ETL Tool?

ETL Process


EXTRACTING
data brought up from external sources
TRANSFORMING
data to fit the demanded standards
LOADING
converted data into a destination DW


Entity Relational Models
OLTP Systems
ETL Processes
Dimensional Models
OLAP Systems
Enterprise Data Warehouses
Data Marts


ETL Process
Source
Extract
Transform
Load
Destination

ETL Process: Major Tasks

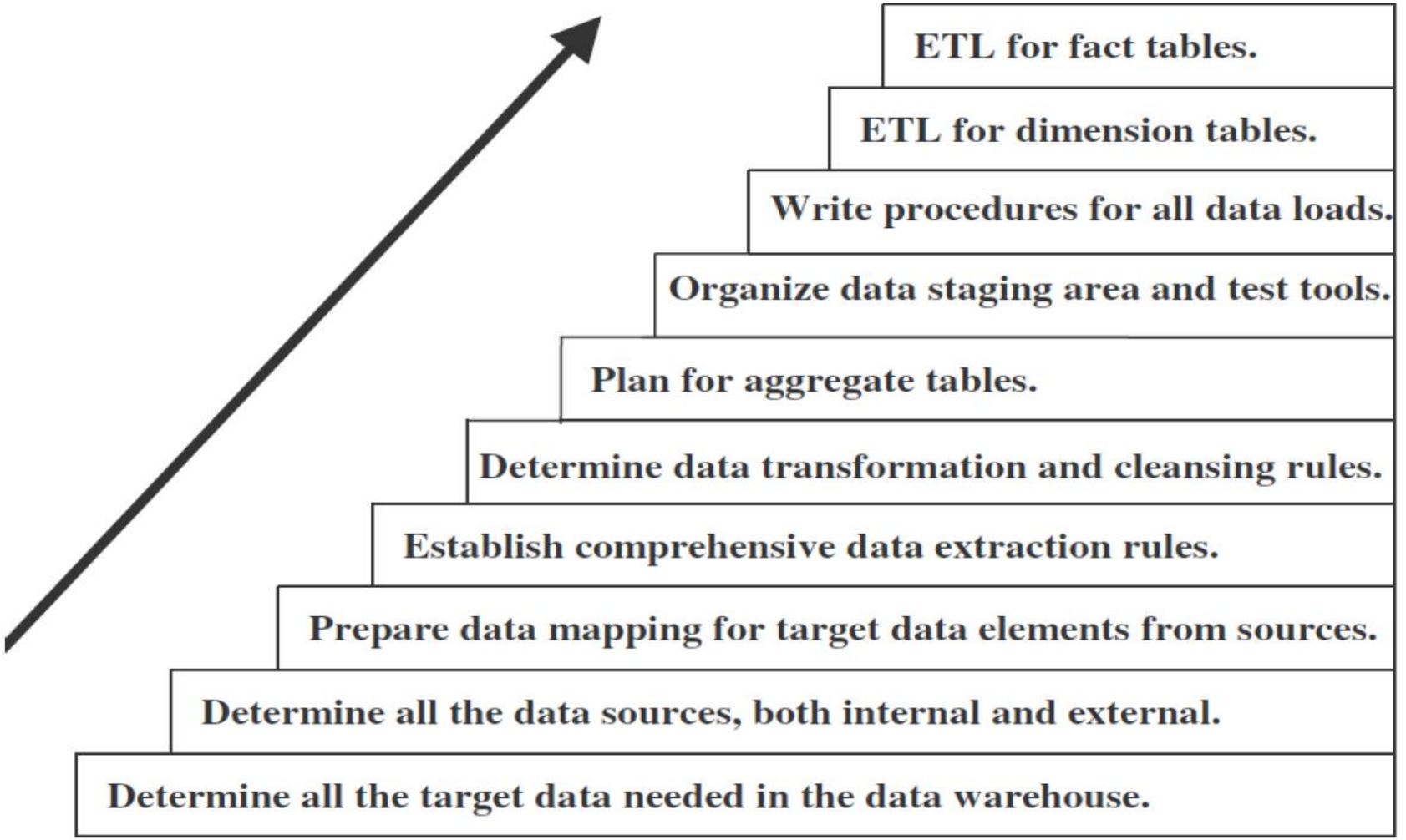
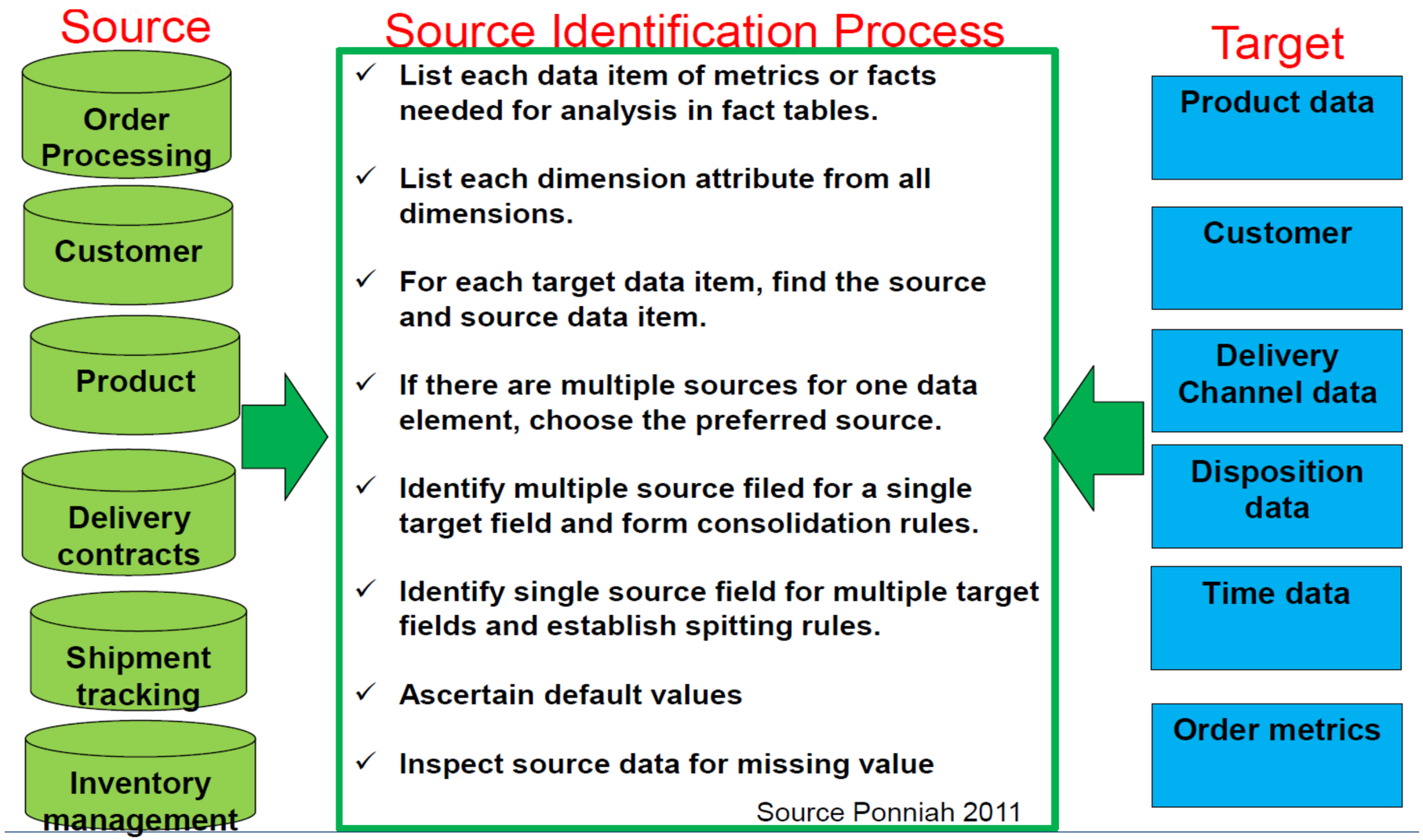


Figure 12-1 Major steps in the ETL process.

Source Ponniah 2011

Mapping source and target data



Data Extraction

Data Extraction

- We have done the formal requirements analysis, we know what data is required, where it is now, what needs to be done to it etc…

- Should know the information you need – WHY?
  - Because you have done a REQIREMENTS ANALYSIS
  - You have gathered all data, analytics, decision and business requirements

Source identification

- What are the PROPER data sources
  - Examine and verify - Can you get the necessary data for the DW
- Steps
  - List each data item needed for analysis in fact tables
  - List each dimension attribute
  - For each target data item find the source system and source data item
    - If there are multiple sources which one is the definitive one?
  - Identify multiple source fields for a single target field
    - Form consolidation rules for these
  - Identify single source fields for multiple target fields
    - Form splitting rules for these
  - Ascertain correct values
  - Inspect source data for missing values

Data extraction techniques

- Must have intimate knowledge of data sources
  - Time dependent data!
  - When do you update the DW?
- Also important to know how extracted data is used
  - When do we HAVE to update the data?
  - What is real time for the type of data?
- How do we handle historical data...
  - Customers over 3 years having 4 different addresses
  - Suppliers moving offices
    - Each of these may indicate the need for slowly changing dimensions
  - Lots of issues around this

Data in operational systems

- Current value – most common data type
  - Transient values – at a particular snapshot this is the value
  - Value can change at any time
  - If you need to preserve history of these values it gets very involved (especially if there is no aggregation taking place)

- Periodic status (not common)
  - Every time value is changed the old value is stored historically along with a timestamp
  - History is preserved in operational system
    - Thus easy to get history into DW

Data Extraction

- Two major types of Data Extraction
  - The static data (“as-is” at that point in time)
  - The revision data (what’s changed – also includes periodic data)

- This will depend on how we are loading the DW with data
  - Initial Load
  - Subsequent Load
  - Re-Load (same as running initial load and starting again)
  - This can happen with any table in the DW, individually or as a group

Data Extraction Types

- Immediate data extraction – REAL TIME!
- Multiple Methods
  1. Capture via transaction logs
     - A transaction log records all transactions and the database modifications made by each transaction
     - Reads transaction logs and selects all committed transactions
     - Must ensure you capture ALL logs
     - Great if data comes from a database

  2. Capture via database triggers
     - A piece of code that gets initiated when data changes in the database
     - Use triggers to generate data in separate file for all changes to data you want to track
     - Additional burden on development effort – also changing source databases by adding triggers
       - Additional overhead...

Data Extraction Types

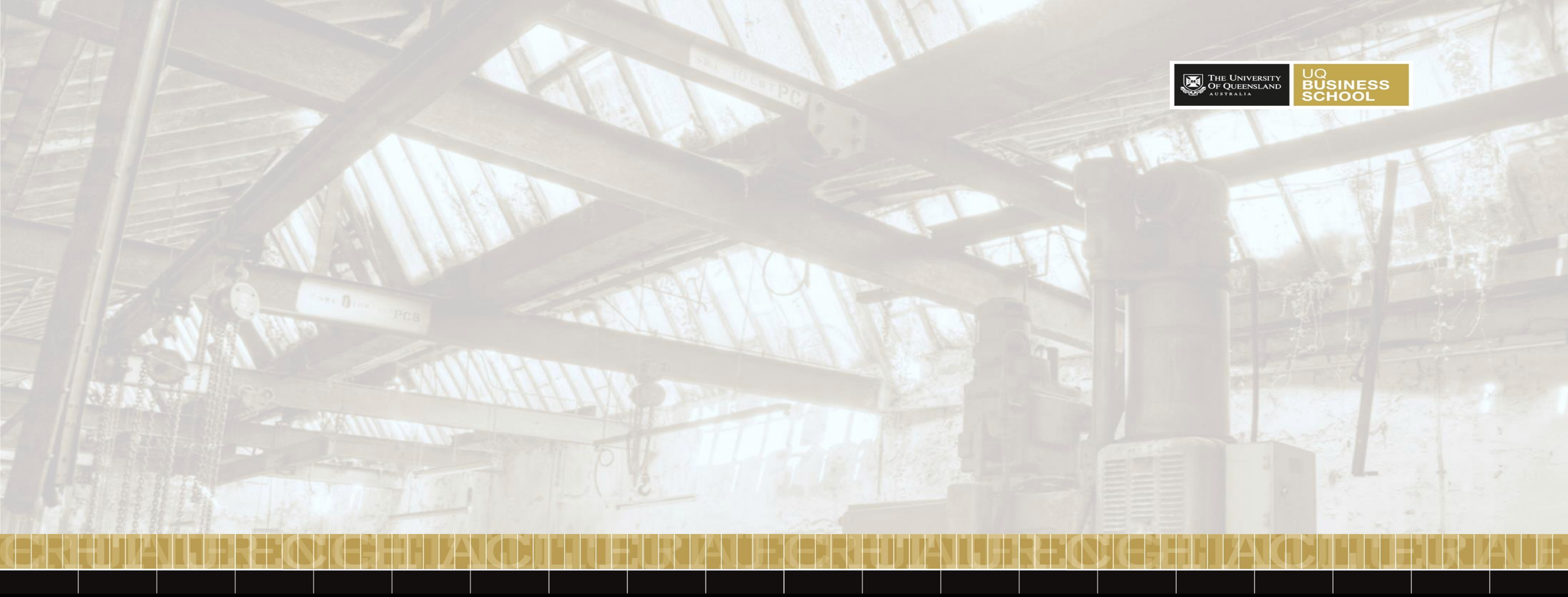
- Immediate data extraction – REAL TIME!
  3. Capture in source applications
     - Source applications are modified to ALSO capture data warehouse data
       - Don’t forget these will need to also be maintained
     - All relevant changes to data are written to separate files for the ETL process to use
     - Can be used for all types of data sources
       - Not just databases
     - May downgrade application performance

Data Extraction Types

- Deferred data extraction – NOT REAL-TIME
  1. Capture based on date and time stamp
     - All relevant items need to be time stamped
     - Use timestamp to identify changed data since last time and only extract these records.
     - Works well if small number of records
     - Deletions
       - Need to be marked initially and then after ETL runs they get deleted

Data Extraction Types

- Deferred data extraction – NOT REAL-TIME
  2. Capture by comparing files
     - Last resort
       - Especially for legacy systems with no timestamps or logs
     - Compare the data now with the data last time
       - Determine what’s changed and update it
       - Look at keys to identify deletions and insertions
     - On a large scale is inefficient
       - Especially in large tables



Data Transformation

Data Transformation

- Extracted data not good enough for the DW
  - Quality
  - Format

- We have the RAW data...

Basic Tasks

- Selection
  - Get whole or part records from source systems
  - May be carried out in extraction
    - not always
      - Source structure might not be amenable
      - So extract whole record and select as part of transformation
- Splitting/Joining
  - Manipulation of data
    - Splitting up records is uncommon
    - Joining info is very common (eg customer data)
- Conversion
  - Converting single fields to
    - Standardise
    - Make fields understandable to users

Basic Tasks

- Summarisation
  - Depending on level of detail required some data can be summarised
    - Egs
      - Balance per second, vs Balance at end of day
      - Each individual sale vs Sales per product per store per day

- Enrichment
  - Rearrangement and simplification of individual fields to make them more useful in the DW
  - Several fields from different source systems about an entity are combined
    - Eg, customer data

Major Transformation Tasks

- Format Revisions
  - Changes to data types and field length
    - Common

- Decoding of Fields
  - Which name is correct for each field
  - If many sources, probably different field names and definitions
    - Common
  - Field values changed to non cryptic
    - AC, IN, RE for instance should be Active, Inactive, Regular
    - In a gender field storing 1, 2 or M, F – need to fix

Major Transformation Tasks

- Calculated and derived values
  - May need to calculate data points
    - Eg average sales, profit margin
    - Common

- Splitting of single fields
  - Essentially normalising a single field
    - Address stored as 1 field instead of Street #, Name, etc
    - Customer name breakdowns also
  - Important
    - Can index things like postcode
    - Allows for analysis on components

Major Transformation Tasks

- Merging of information
  - Getting data about a particular thing all together in the DW
    - Merging info about a product from different sources
      - Eg code, description, package types, cost
- Character set conversion
  - Different systems use different character sets (may not be compatible)
    - Must convert to DW character set
      - Eg EBCDIC to ASCII
- Conversion of units of measurements
  - What is the standard of measurement for the organisation
    - May need to convert from imperial to metric
    - Currency is an example

Major Transformation Tasks

- Date/time conversion
  - Different systems may use different formats
  - Need to be clear
    - 11/12/2011
      - 11 Dec 2011 or 12 Nov 2011
      - Store it in a standard format
        - 11 DEC 2011
- De-duplication
  - Get rid of the duplicate records that you find

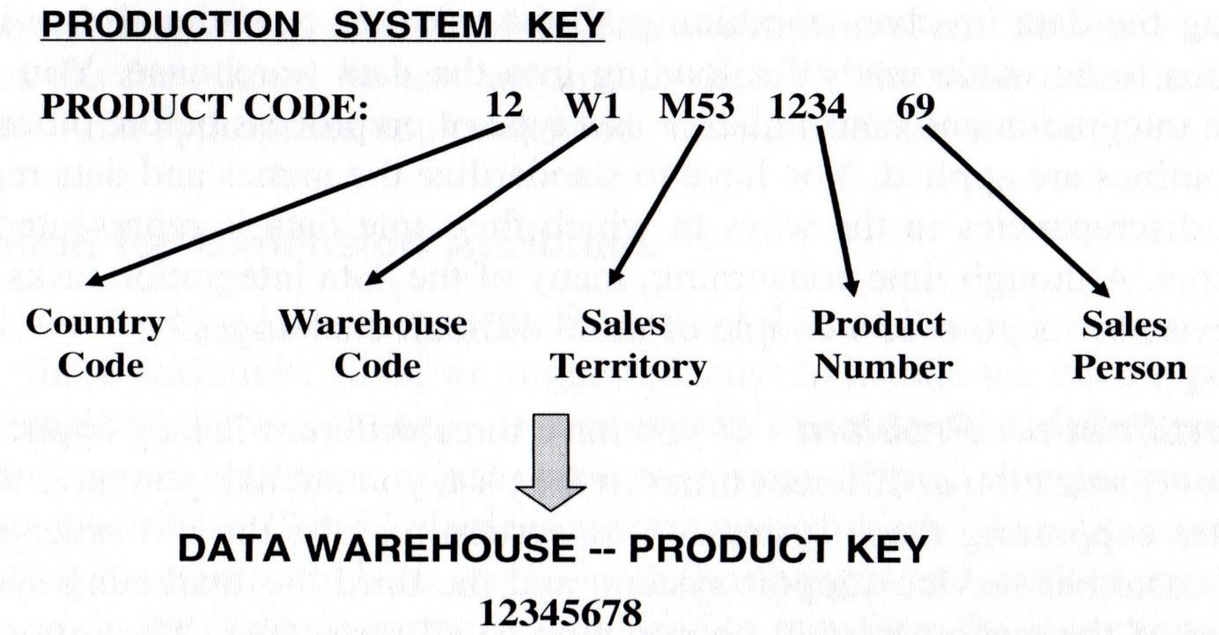
Major Transformation Tasks

Surrogate key

Former primary key

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

- Key restructuring
  - Need to give new keys in the DW
  - Avoid keys with built in meaning
    - In the below example if the product is stored in a different warehouse it gets a different key... So you lose it in the DW



Source: Ponniah (2010) p299

Data Integration and Consolidation

- Biggest Challenge
  - Lots of disparate data sources
    - Business rules changed over time
    - Different
      - Naming conventions
      - Standards for data representation
  - And your job, should you choose to accept it, is to consolidate it all into a DW

Data Integration and Consolidation

- Entity identification problem
  - The Customer Entity
    - Data from 3 systems
    - All with different identifier formats
    - How do / can you identify the same customer in all 3 systems to integrate the data?
    - Same for suppliers, employees etc...
  - Algorithms group alike “customers” together
    - Manual process then to decide if they are the same customer...
  - A common, complex and perplexing problem

Data Integration and Consolidation

- Multiple Sources Problem
  - What do you do if you have the same data point from multiple source systems
    - Eg “cost of product” has 2 values from 2 different systems
    - Which system is correct?
  - Have to decide where to go for the definitive data

Data Loading

Data Loading

- Types
  - Initial Load
    - Populating the DW for the 1st time
  - Incremental Load
    - Applying ongoing updates to the DW in a periodic manner
  - Full Refresh
    - Erase the DW data, and run Initial Load again!
- When to load?
  - Full Loads take a long time to run
  - DW offline during loads
    - Partially or fully
  - Need to find a time where they can be accomplished
  - Test load times – so you know how long the system will be down.

Applying the data to the DW

- Four ways to copy data to DW tables
  - Load
    - Apply data directly to table, overwrites anything there
  - Append
    - Adds data to the table, preserving what is already there
  - Destructive Merge
    - Adds data to the table, if the key exists overwrite that record
  - Constructive Merge
    - Adds data to the table, if the key exists mark that row as old and add the new row
      - Allows history to be stored
      - One way of doing slowly changing dimensions

Summary of Data Application

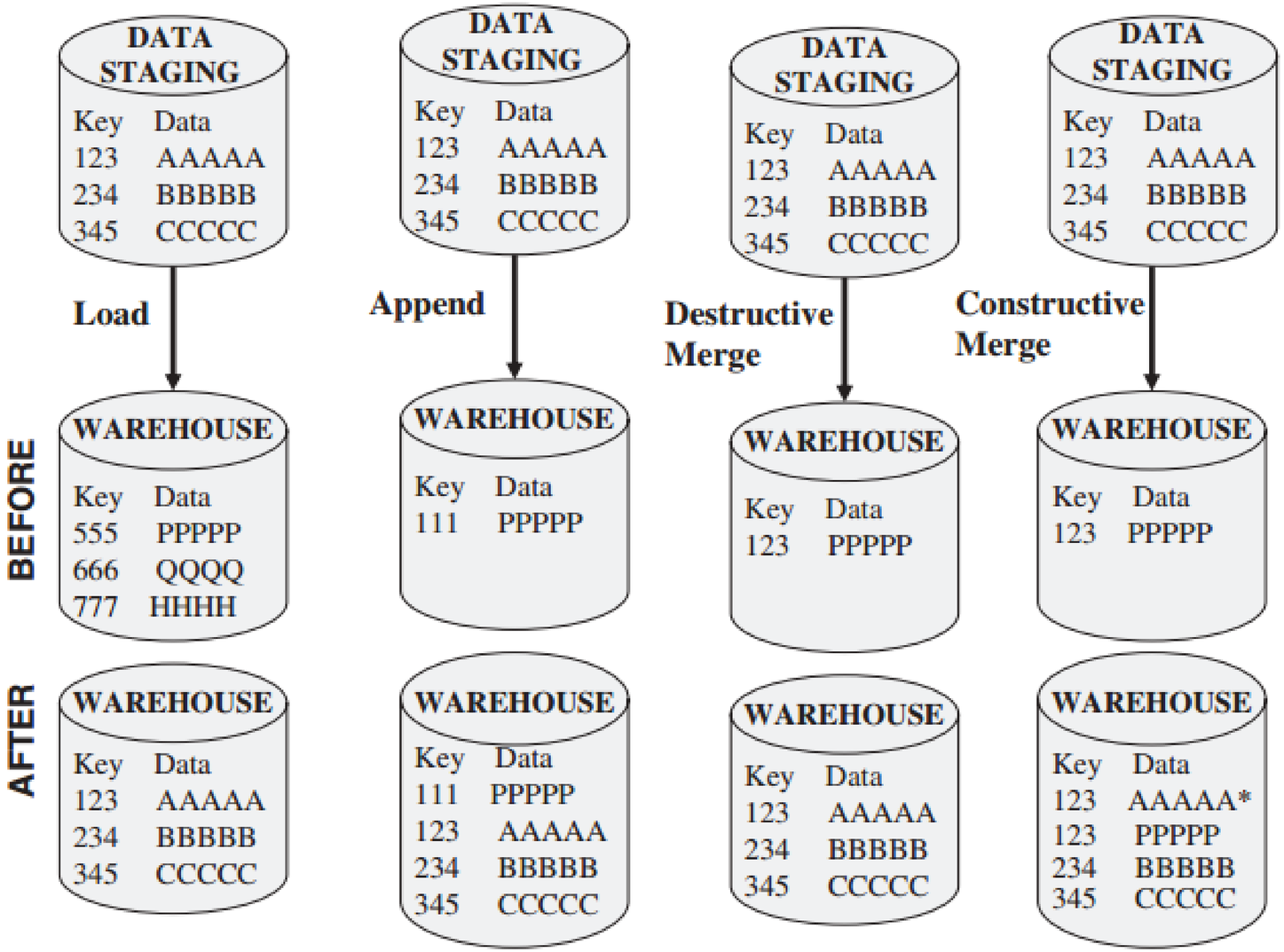


Figure 12-11 Modes of applying data.

Ponniah (2010) p304

ETL Summary

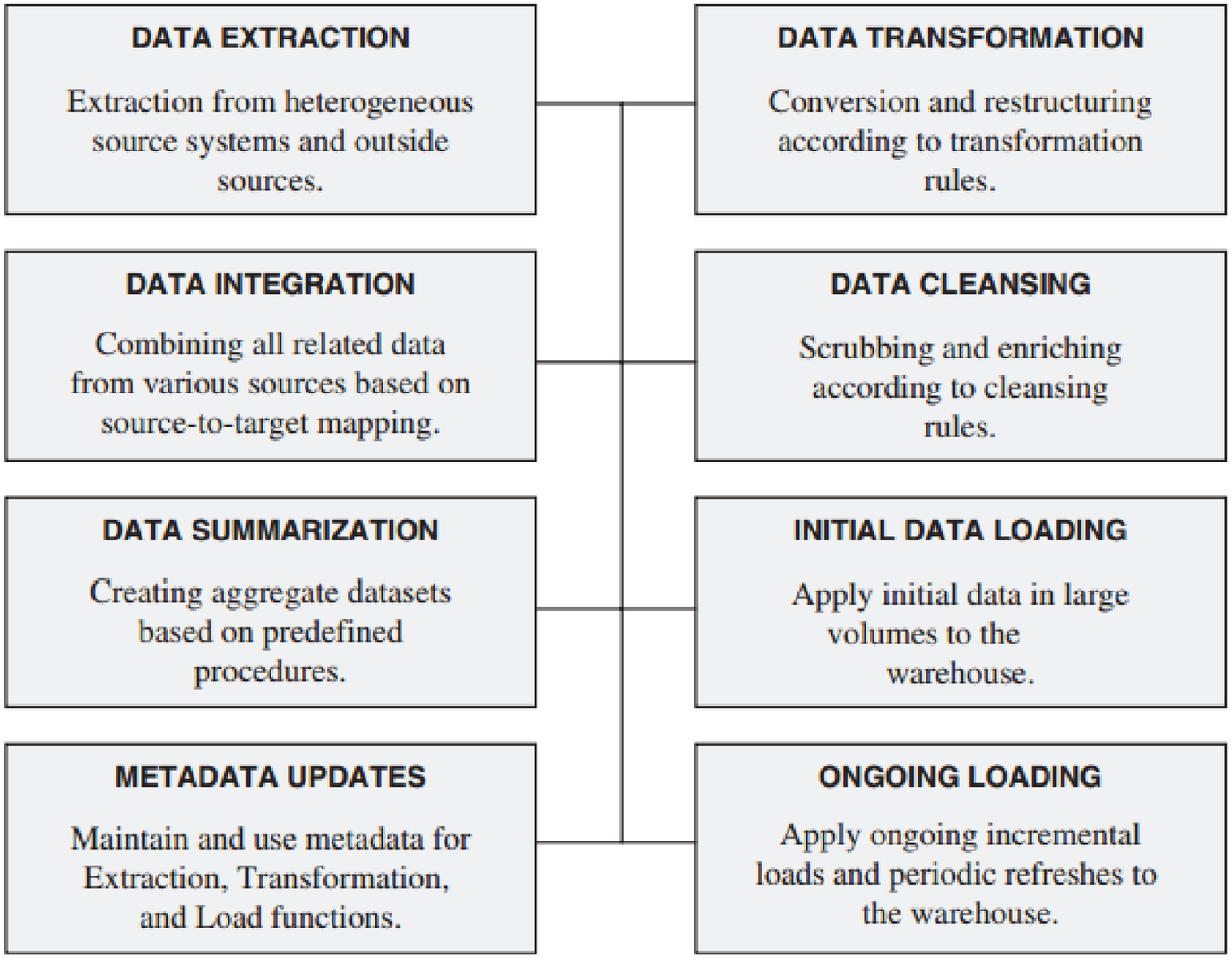


Figure 12-14 ETL summary.

Ponniah (2010) p310

ETL Tools

- Its not all manual labour after all...

ETL Tools

- 3 types of tools
  - Data Transformation Engines
    - Dynamic and sophisticated data manipulation algorithms
    - Capture data from set of sources, transforms data and sends results to target environment
    - Functionality covers whole ETL process
  - Data Capture via replication
    - Tools use DBMS recovery logs
    - Data replicated to staging area in near real-time
    - Translation and loading can then take place (outside tool)
  - Code Generators
    - Specifically for ETL
    - You define where the data is and the rules and then code is generated to do it
    - Can further enhance code with own code if required

ETL Tools

- The good news is that there are commercial and in-house products to do these tasks...
- Many DBMS vendors sell inbuilt tools also (a fairly inexpensive option)
- Examples
  - Anatella
  - Oracle Data Integrator
  - Pentaho
  - Safe Software
  - Benetl
  - Syncsort DMExpress
  - Informatica
  - Pervasive Software
  - SAS Data Integration Server
  - SAP BusinessObjects Data Integrator
  - SQL Server Integration Services
  - Talend Open Studio

Microsoft
SQL Server Integration Services

- SSIS
  - A platform for data integration and workflow applications
  - Features a data warehousing tool used for data extraction, transformation, and loading
- The Basics – A flash demo (lots of them)
- Shows the basics of using the software
  - https://blog.pragmaticworks.com/topic/ssis
  - https://docs.microsoft.com/en-us/sql/integration-services/lesson-1-create-a-project-and-basic-package-with-ssis?view=sql-server-ver15
- We will do lots more in the 2 labs on it and then you will be asked to do some ETL for the Assignment

Metadata

A Customer Example

- What metadata might there be for a customer dimension in a DW?
  - Aliases (client, account)
  - Definition (A person or Organisation that purchases goods or services from the company)
  - Remarks (Customer dimension includes regular, current and past customers)
  - Source Systems (FinGoodsOrders, Maintenance Contracts, OnlineSales)
  - Create Date (Feb 3, 2015)
  - Last update date (Jan 20, 2017)
  - Update cycle (Weekly)
  - Last full refresh (Dec 31, 2016)
  - Full refresh cycle (every 6 months)
  - Data Quality Reviewed (Dec 5, 2016)
  - Last De-duplication (Dec 18, 2016)
  - Planned Archival (every 6 months)
  - Responsible user (Slim Dusty)

Why – Metadata for…

- Using the DW
  - DW is different that operational systems as users create own reports
  - Users need meta data to work out definitions, what is available in the system to query

- Building the DW
  - ETL specialists need to know data sources, and their transformations
  - DBA’s need to know the logical database structure, load cycles etc.

- Administering the DW
  - Need lots of info around the management of the DW – including new data etc.

User Metadata Needs

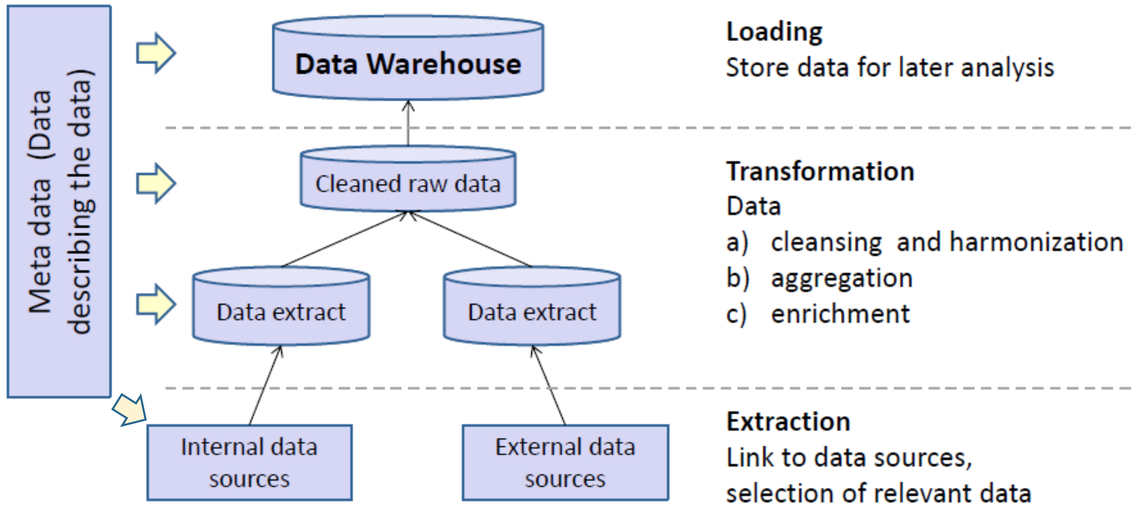
	IT Professionals	Power Users	Casual Users
Information Discovery	Databases, tables, columns		List of predefined queries and reports, business views
	Server platforms		
Meaning of Data	Data structures		Filters, data sources, conversion, data owners
		Business terms	
	Data definitions		
	Data mapping, cleansing functions, transformation rules		
Information Access	Program code in SQL, 3GL, 4GL, front end applications, security	Query toolsets, database access for complex analysis	Authorisation requests, information retrieval to desktop applications like spreadsheets

Providing Metadata

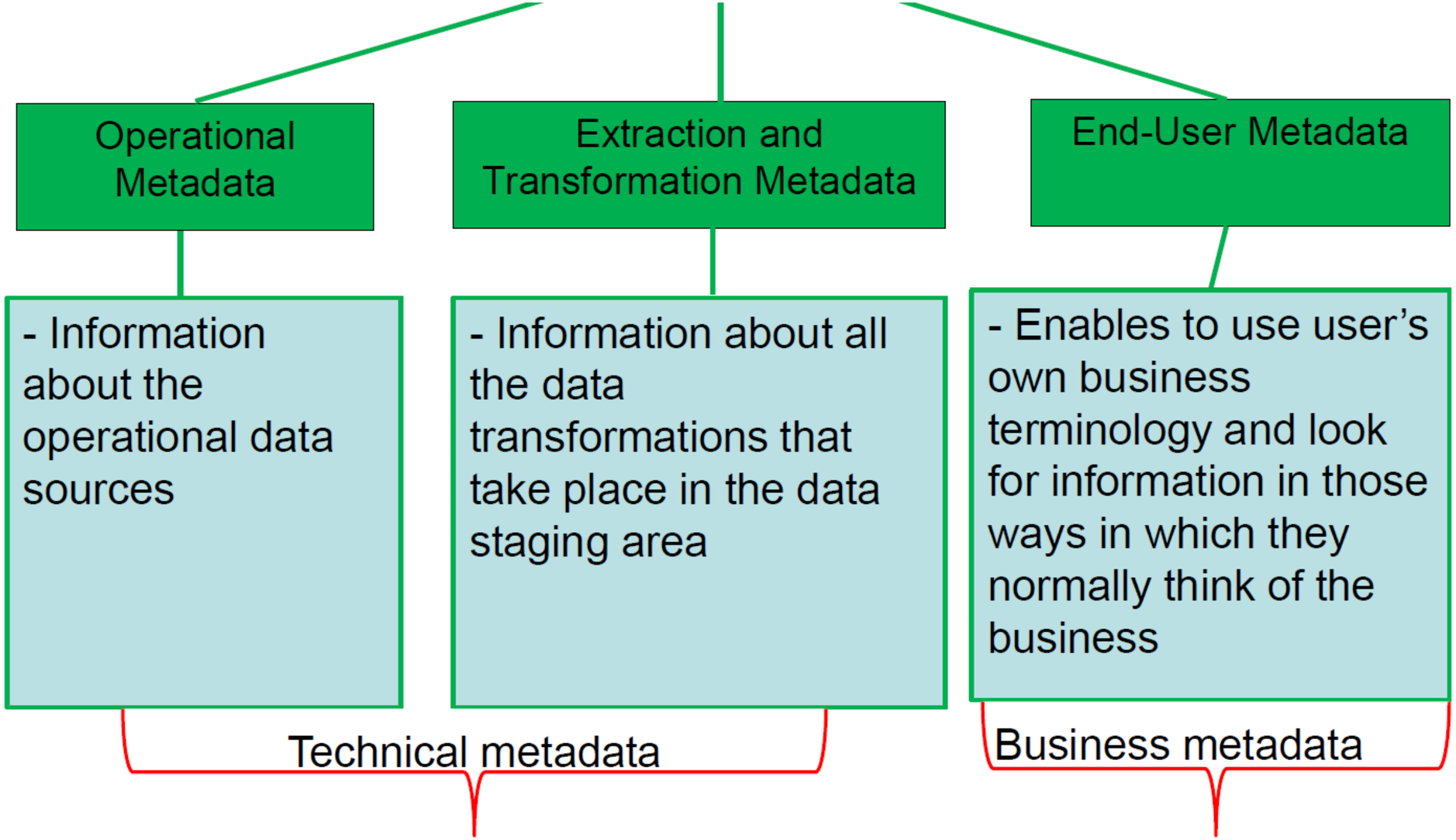
- Metadata must serve as a roadmap to the DW
  - For users, & developers and administrators

- Some meta data comes from source system metadata and is then added to in the ETL process

- Must have processes in place to:
  - Standardise metadata across systems
  - Revise metadata across systems if it is changed
  - Exchange metadata across systems
  - Allow querying of metadata



Metadata Types

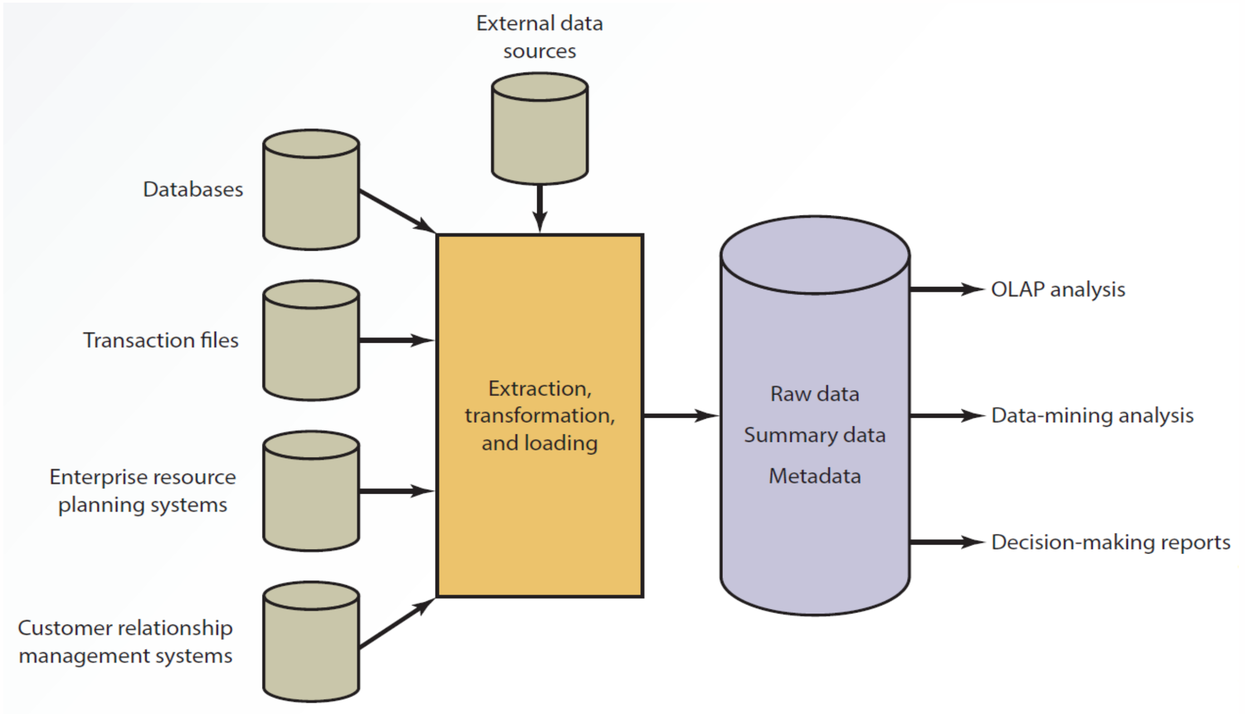


Business Metadata for End-Users

- Data Content
- Summary data
- Business dimensions
- Business metrics
- Navigation paths
- Source systems
- External data
- Data transformation rules
- Last updates dates
- Data load update cycles
- Query templates
- Report formats
- Predefined queries
- OLAP data

Sources of Metadata

- Many of the tools and techniques used at stages of the DW development and implementation process generate metadata – some examples below.
- Source Systems
  - Data Models
  - Data Definitions
    - Documentation
    - Data dictionary
  - File layouts
  - Program specifications



Sources of Metadata

- Data Extraction
  - Data on source platforms
  - Layouts and definitions of selected data sources
  - Field definitions
  - Rules for standardising data types and lengths
  - Data extraction schedules
  - Extraction methods for incremental changes
  - Criteria for merging into initial extract files

Sources of Metadata

- Data Transformation and Cleansing
  - Specifications for mapping extracted files to data staging area
  - Conversion rules
  - Default values for fields with missing values
  - Business rules for validity checking
  - Sorting and resequencing arrangements
  - Audit trail details

Sources of Metadata

- Data Loading
  - Specifications for mapping data staging files to load images
  - Rules for keys
  - Audit trail for data staging to loading
  - Schedules for full, and incremental data loads

Sources of Metadata

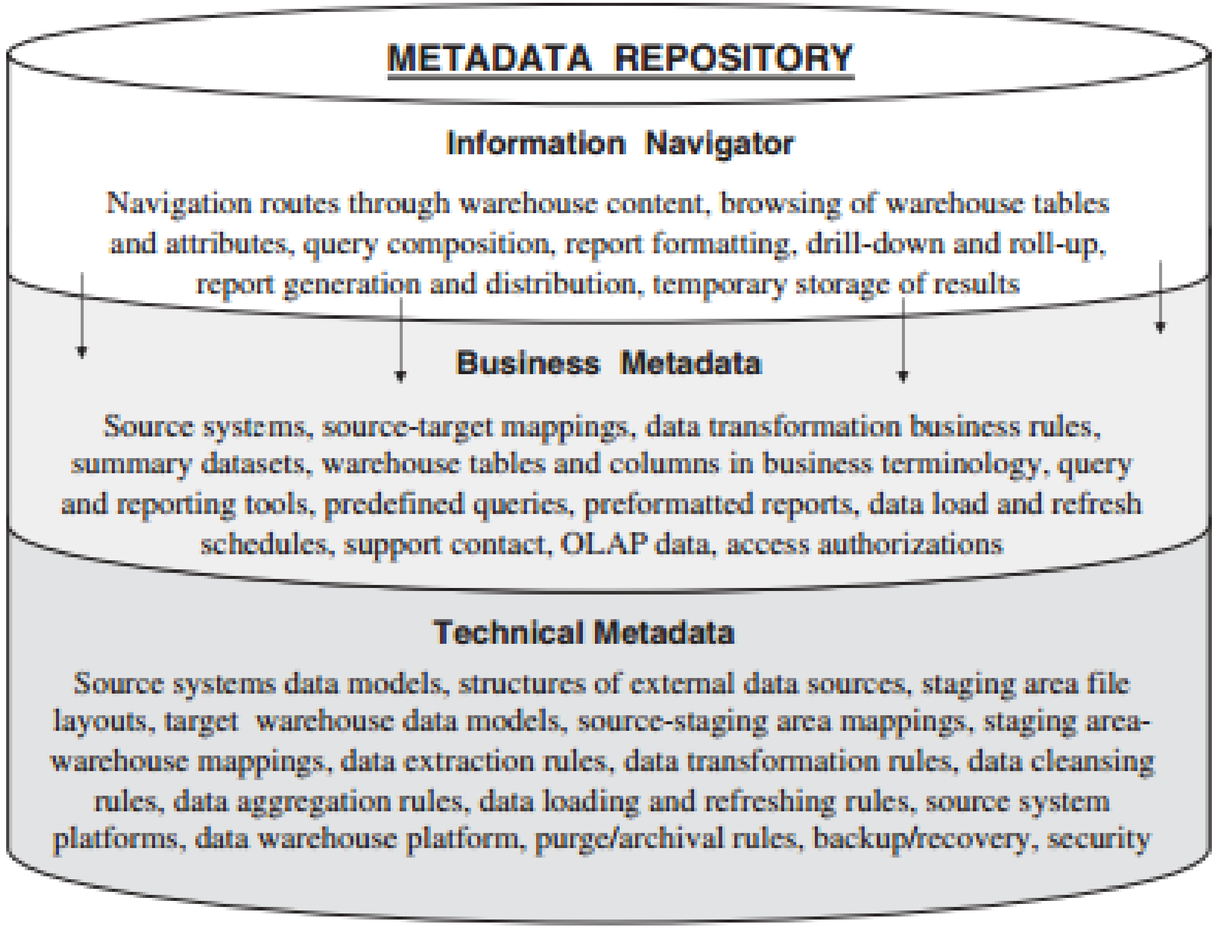
- Data Storage
  - Data models for DW
  - Subject groupings of tables
  - Physical files
  - Table and column definitions
  - Business rules for validity checking

Sources of Metadata

- Information Delivery
  - Lists of queries
  - Lists of reports
  - Lists of tools
  - Data model for OLAP
  - Schedules for data load into OLAP

The Metadata Repository

- Stores both technical and business metadata
- The Ideal:



Sourced from Ponniah (2010)

Figure 9-11 Metadata repository.

Metadata Management Challenges

- Software tools provide proprietary metadata
- No industry wide standard
- No widely accepted methods for sharing metadata amongst systems (and from the various DW components)
- Version control is tedious and difficult
- Unifying metadata is a huge task
  - Conflicting standards, formats, data naming conventions, data definitions, attributes, values etc.

Metadata Standards

- 1 Main standard (for DW)
  - Object management Group
    - Common Warehouse Metamodel
      - http://www.omg.org/spec/CWM/1.1/PDF/
    - The main purpose of CWM is to enable easy interchange of warehouse and business intelligence metadata between warehouse tools, warehouse platforms and warehouse metadata repositories in distributed heterogeneous environments.
    - CWM is based on three key industry standards:
      - UML - Unified Modelling Language
      - MOF - Meta Object Facility
      - XMI - XML Metadata Interchange

What is Examinable:

- What is ETL and Why?
- Data transformations with examples
- Definition and importance of metadata
- Examples of metadata

Next Seminar

Next Seminar

- **BA Architectures**
- **Assignment Q&A and Group Work**