



Metadata and Extract, Transform, Load (ETL)



COMP323 Chapter 4

Outline

▶ Metadata

- ▶ What is it?
- ▶ The use of metadata
- ▶ Sources of metadata

▶ ETL

- ▶ What is it?
- ▶ Data extraction
- ▶ Data transformation
- ▶ Key extraction and transformation steps
- ▶ Loading data into a data warehouse
- ▶ Introduction to ETL tools

What is metadata?

- ❖ Data about data;
- ❖ Data which provides information about resources;
- ❖ Metadata is used for indexing, discovery, and search;
- ❖ Metadata is about what sort of things an item contains;
- ❖ Metadata is not a summary or abstract;
- ❖ Metadata does not contain transaction data.



Wikipedia

What is metadata? (II)

IN SUMMARY

- ✓ Metadata helps you describe, use, find and manage content and data.
- ✓ Determine essential metadata properties needed to control and use business information.
- ✓ Use standardised metadata to support interoperability and information sharing.

<https://www.youtube.com/watch?v=3sLKVYYOM40>

Metadata example

All Sale
Shop a great range of styles on sale.

Metadata

RECOMMENDED

Refine

APPLIED FILTERS -

STYLE -

- ☐ Ankle Boots (17)
- ☐ Boots (26)
- ☐ Casuais (80)
- ☐ Comfort (5)
- ☐ Dress (37)
- ☐ Espadrilles (1)

SIZE -

EU Women

35	36	37	38
39	40	41	42

COLOUR -

PRICE -

Any

SALE

Lazoo Denim Suede
~~\$158.00~~ \$75.00

SALE

Lazoo Black Suede
~~\$158.00~~ \$75.00

SALE

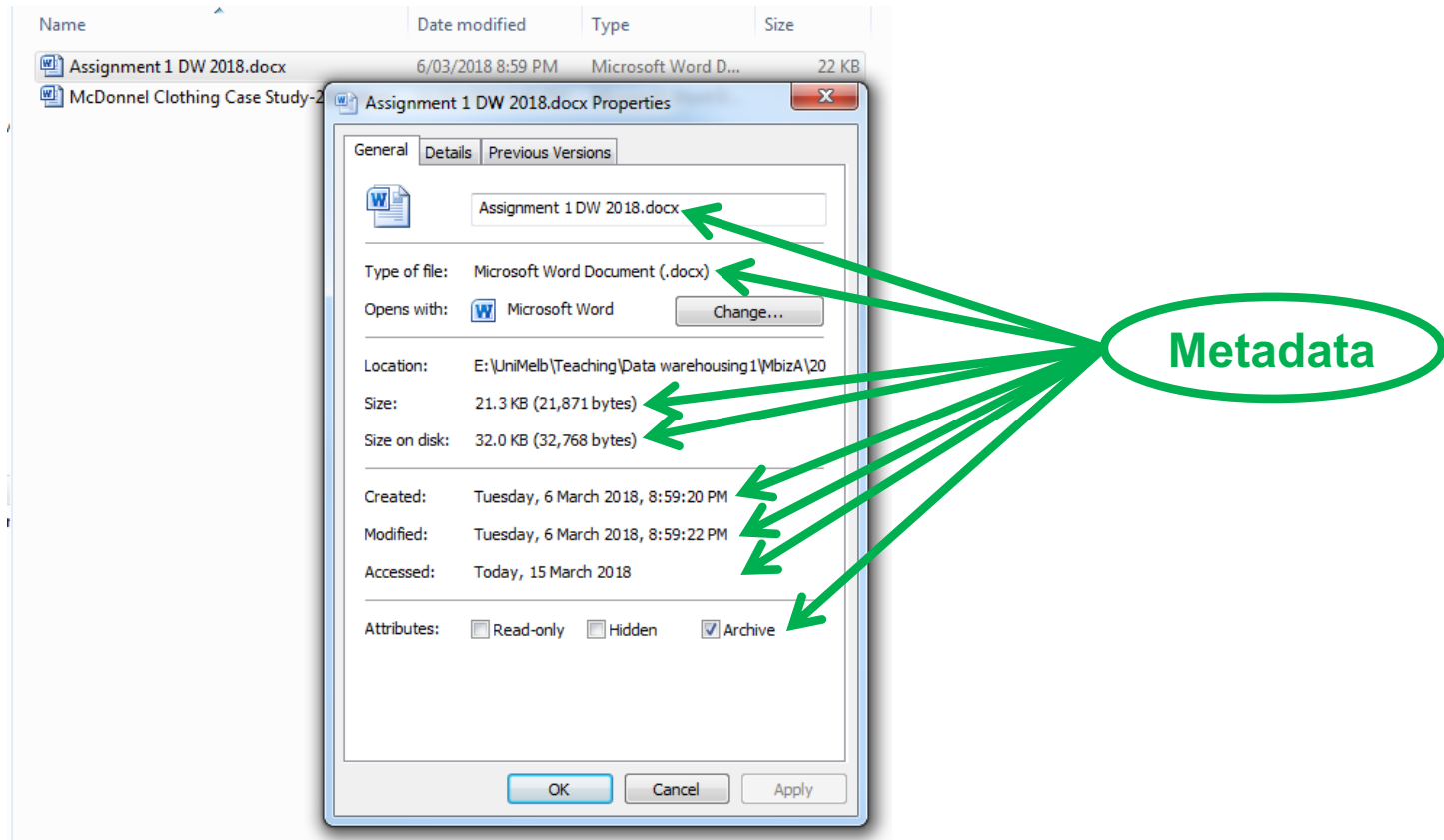
Cute Black Floral Multi Suede
~~\$298.00~~ \$120.00

SALE

SALE

SALE

Metadata example (II)



Why metadata is important?



Data
Warehouse



1. What are the elements of data in the warehouse?
2. Is there information about unit sales and unit costs by product?
3. How can I browse and see what is available?
4. From where did they get the data for the warehouse? From which source systems?
5. How did they merge the data from the telephone orders system and the mail orders system?
6. How old is the data in the warehouse?
7. When was the last time fresh data was brought in?
8. Are there any summaries by month and product?

Metadata: A Customer dimension example

Entity Name: Customer

Alias Names: Account, Client

Definition: A person or an organization that purchases goods or services from the company.

Remarks: Customer entity includes regular, current, and past customers.

Source Systems: Finished Goods Orders, Maintenance Contracts, Online Sales.

Create Date: January 15, 1999

Last Update Date: January 21, 2001

Update Cycle: Weekly

Last Full Refresh Date: December 29, 2000

Full Refresh Cycle: Every six months

Data Quality Reviewed: January 25, 2001

Last Deduplication: January 10, 2001

Planned Archival: Every six months

Responsible User: Jane Brown

(Ponniah 2011)

Metadata: A Customer dimension example (II)

- ❖ The metadata element describes the dimension called Customer residing in the data warehouse;
- ❖ It is not just description. It tells you more;
- ❖ It gives more than the explanation of the semantics and the syntax;
- ❖ Describes all the relevant aspects of the data in the data warehouse fully and precisely.

Q: Relevant to whom?

A: To the users and developers.

Entity Name: Customer
Alias Names: Account, Client

Definition:	A person or an organization that purchases goods or services from the company.
Remarks:	Customer entity includes regular, current, and past customers.
Source Systems:	Finished Goods Orders, Maintenance Contracts, Online Sales.
Create Date:	January 15, 1999
Last Update Date:	January 21, 2001
Update Cycle:	Weekly
Last Full Refresh Date:	December 29, 2000
Full Refresh Cycle:	Every six months
Data Quality Reviewed:	January 25, 2001
Last Deduplication:	January 10, 2001
Planned Archival:	Every six months
Responsible User:	Jane Brown

Why metadata?

❖ Using the Data Warehouse:

- Data Warehouse is different from operational systems as Data Warehouse users create own reports, analytics, etc.;
- Users need metadata to work out definitions, what is available in the system to query.

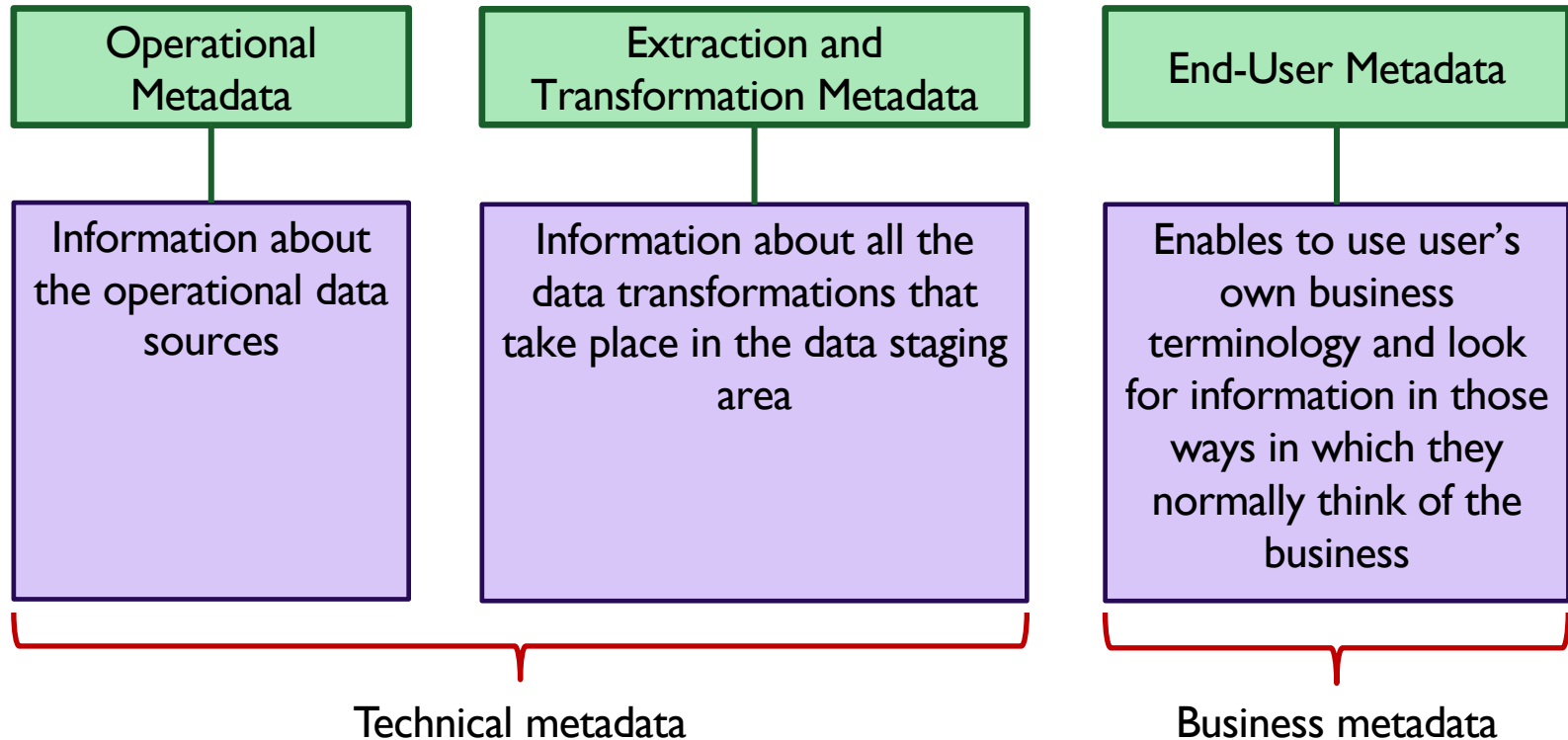
❖ Building the Data Warehouse:

- ETL specialists need to know data sources, and their transformations;
- Database administrators need to know the logical database structure, load cycles, etc.

❖ Administering the Data Warehouse:

- Administrators of data warehouses need lots of information on the management of the data warehouse, including any new data that must be loaded into it.

Metadata types



Metadata types (II)

Business Metadata

(external view of the data warehouse)

- ❖ Composed in simple business terms that users can easily understand;
- ❖ Is like a roadmap for the users to use the data warehouse.

Users: Managers, Business analysts, Power users, Regular users, Casual users, Senior managers/junior executives.

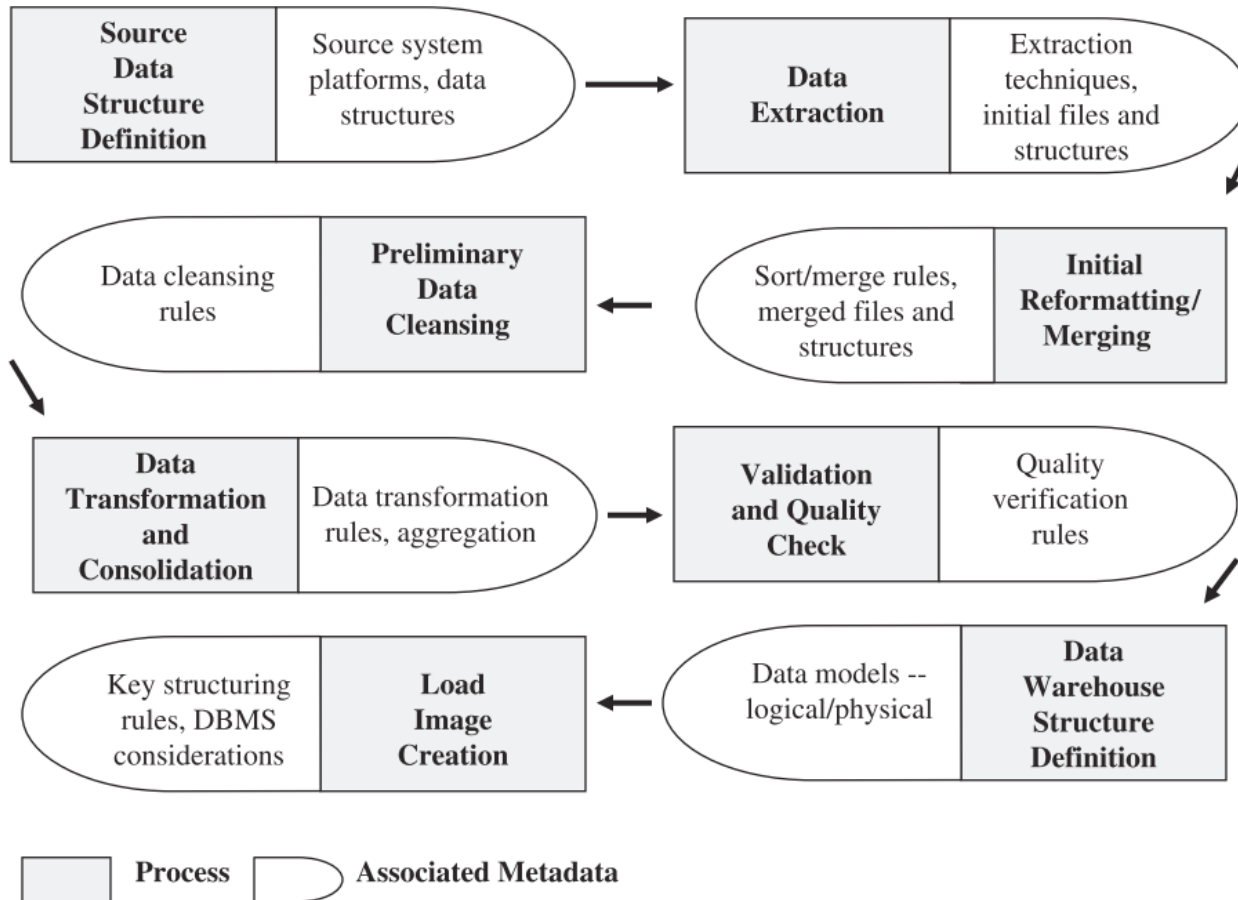
Technical Metadata

(internal view of the data warehouse)

- ❖ Information about the proposed structure and content of the data warehouse;
- ❖ A support guide for the IT professionals to build, maintain, and administer the data warehouse.

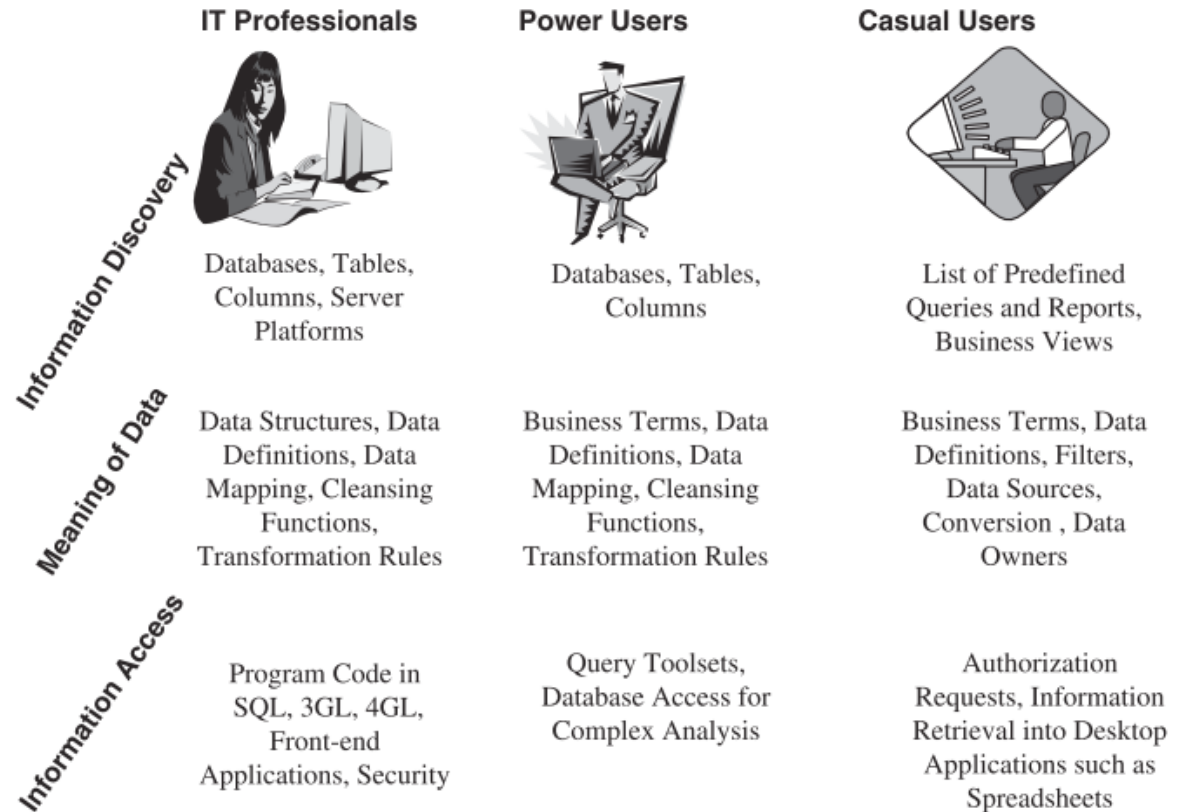
Users: Project manager, DW administrator, Database administrator, Metadata manager, DW architect, Data acquisition developer, Data quality analyst, Business analyst, System , administrator, Infrastructure specialist, Data modeler, Security architect.

Metadata drives data warehouse processes



(Ponniah 2011)

Who needs metadata?



(Ponniah 2011)

Why metadata is essential for IT

- ❖ Source data structures
- ❖ Source of platforms
- ❖ Data extraction methods
- ❖ External data
- ❖ Data transformation rules
- ❖ Data cleansing rules
- ❖ Staging area structures
- ❖ Dimensional models
- ❖ Initial loads
- ❖ Incremental loads
- ❖ Data summarization
- ❖ OLAP system
- ❖ Web-enabling
- ❖ Query/report design



Why metadata is essential for end-users

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



Providing metadata

- ❖ Metadata must serve as a roadmap to the Data Warehouse;
 - For users, and developers and administrators.
- ❖ Some metadata 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.

Sources of metadata

Many of the tools and techniques used at stages of the Data Warehouse development and implementation process generate metadata – some examples below:

- ❖ Data Models;
- ❖ Data Definitions (documentation and data dictionary);
- ❖ File layouts;
- ❖ Program specifications.

Metadata is produced during data extraction, for example:

- ❖ Data on source platforms, including layouts and definitions of selected data sources;
- ❖ Field definitions;
- ❖ Rules for standardising data types and lengths;
- ❖ Data extraction schedules, and extraction methods for incremental changes;
- ❖ Criteria for merging into initial extract files.

Sources of metadata (II)

Metadata is produced during data transformation and cleansing phases, for example:

- ❖ 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.

Metadata is produced during data loading phase, for example:

- ❖ 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 (III)

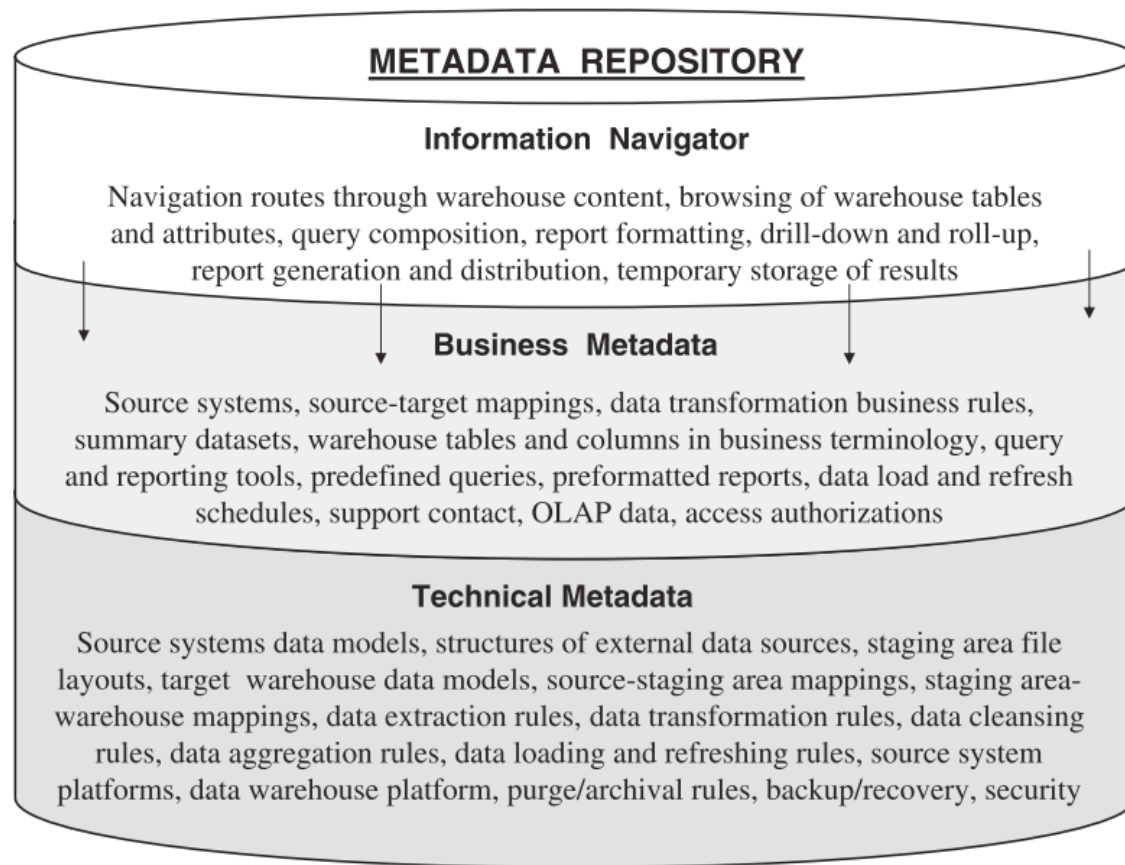
Metadata is produced when storing data, for example:

- ❖ Data models for the Data Warehouse;
- ❖ Subject groupings of tables;
- ❖ Physical files;
- ❖ Table and column definitions;
- ❖ Business rules for validity checking.

Metadata supports information delivery phase, for example:

- ❖ Lists of queries;
- ❖ Lists of reports;
- ❖ Lists of tools;
- ❖ Data model for OLAP;
- ❖ Schedules for data load into OLAP.

Metadata repository



(Ponniah 2011)

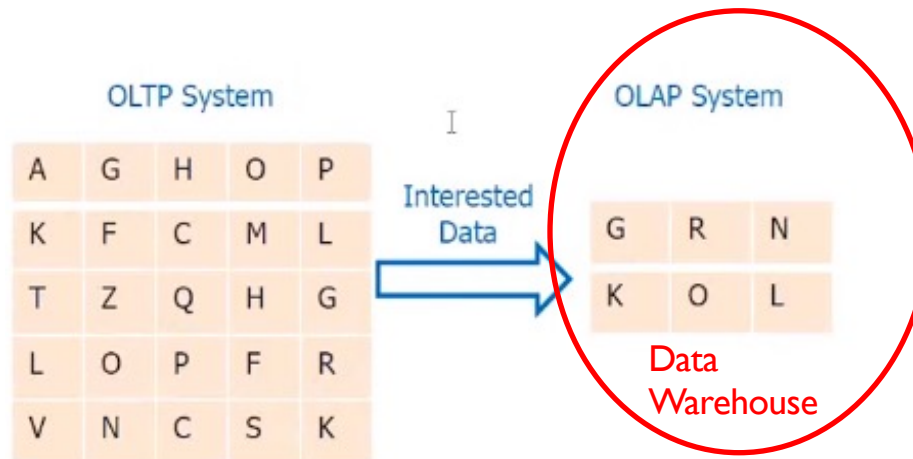
Metadata (Summary)


- ❖ Metadata is a critical need for using, building, and administering the data warehouse;
- ❖ For end-users, metadata is like a roadmap to the data warehouse contents;
- ❖ For IT professionals, metadata supports development and administration functions;
- ❖ Business metadata connects the business users to the data warehouse;
- ❖ Technical metadata is meant for the IT staff responsible for development and administration.

What is ETL?

- Online Transaction Processing (OLTP) systems cannot be used for analytics. Therefore, Online Analytical Processing (OLAP) is needed.
- Doing OLTP and OLAP in the same database system is often impractical:
 - ▶ Different performance requirements
 - ▶ Different data modelling requirements
 - ▶ Analysis queries require data from many sources
- Solution: Build a “data warehouse”
 - ▶ Copy data from various OLTP systems
 - ▶ Optimise data organisation, system tuning for OLAP
 - ▶ Transactions aren’t slowed by analysis queries
 - ▶ Periodically updated the data in the warehouse.

What is ETL?



- ▶ We have: Source systems (OLTP) -> Target systems (OLAP or Data Warehouse).
- ▶ How do we transfer the data from the source systems to target systems?
- ▶  This set of methods is called the **ETL** (Extract, Transform, and Load) process.

Data warehouse versus OLTP

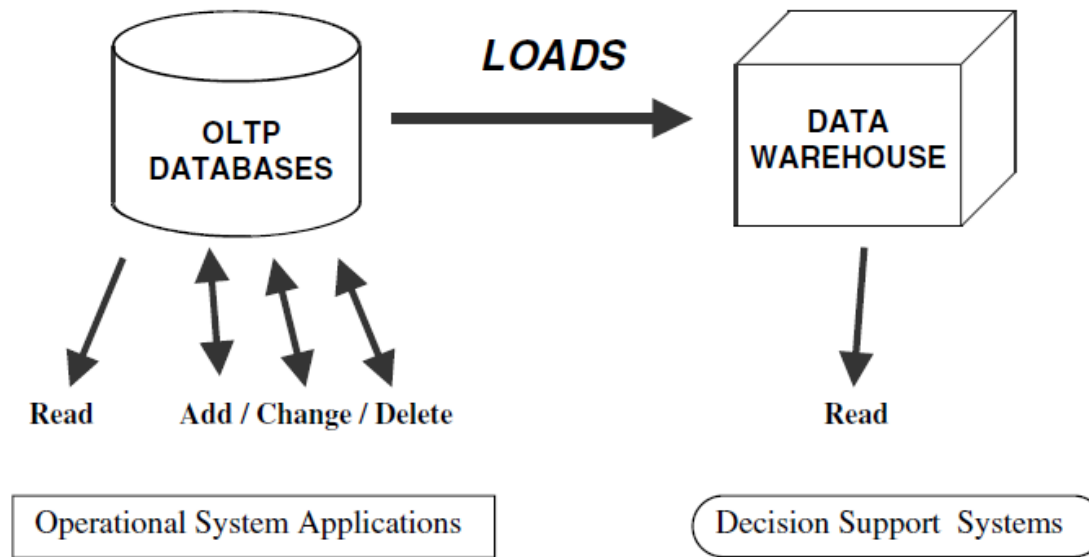


Figure 2-3 The data warehouse is nonvolatile.

Source: Ponniah 2011

The Role of ETL

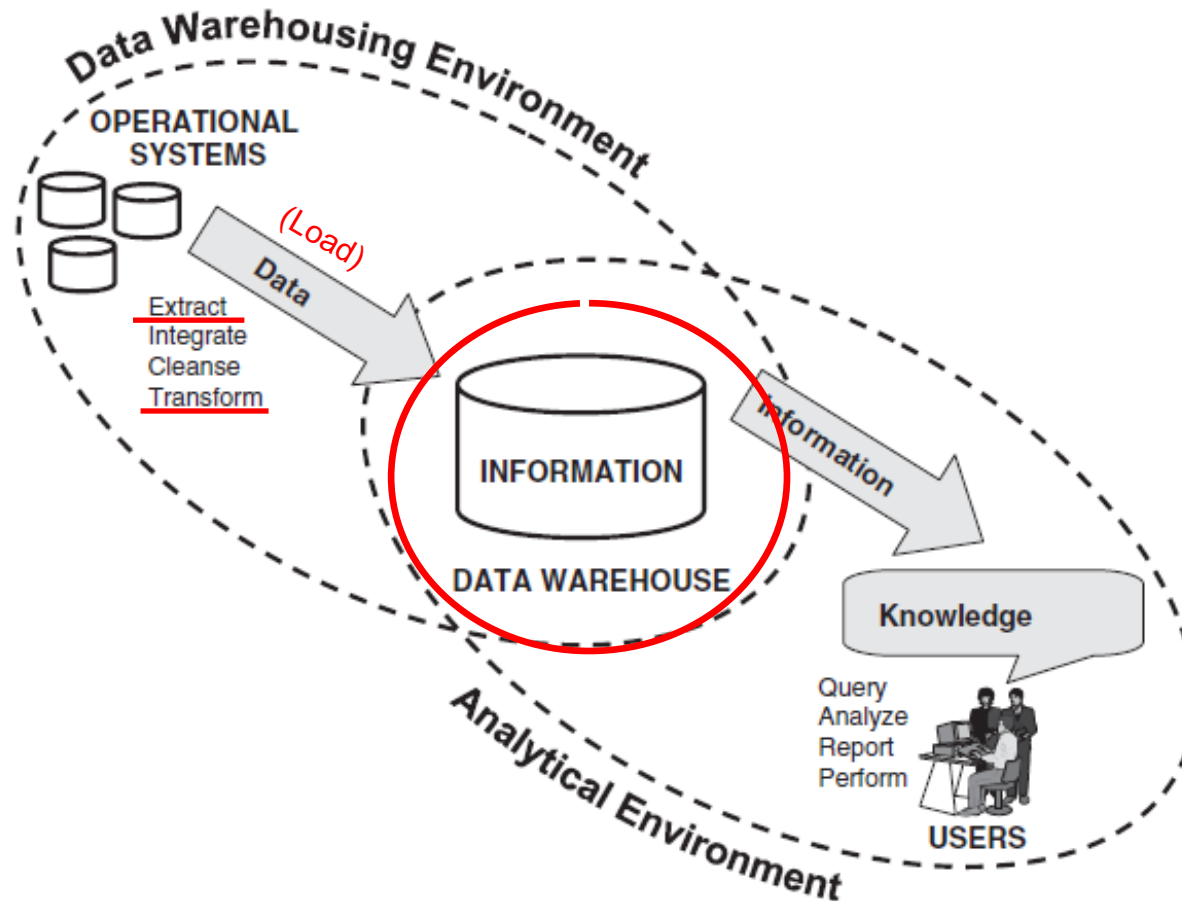
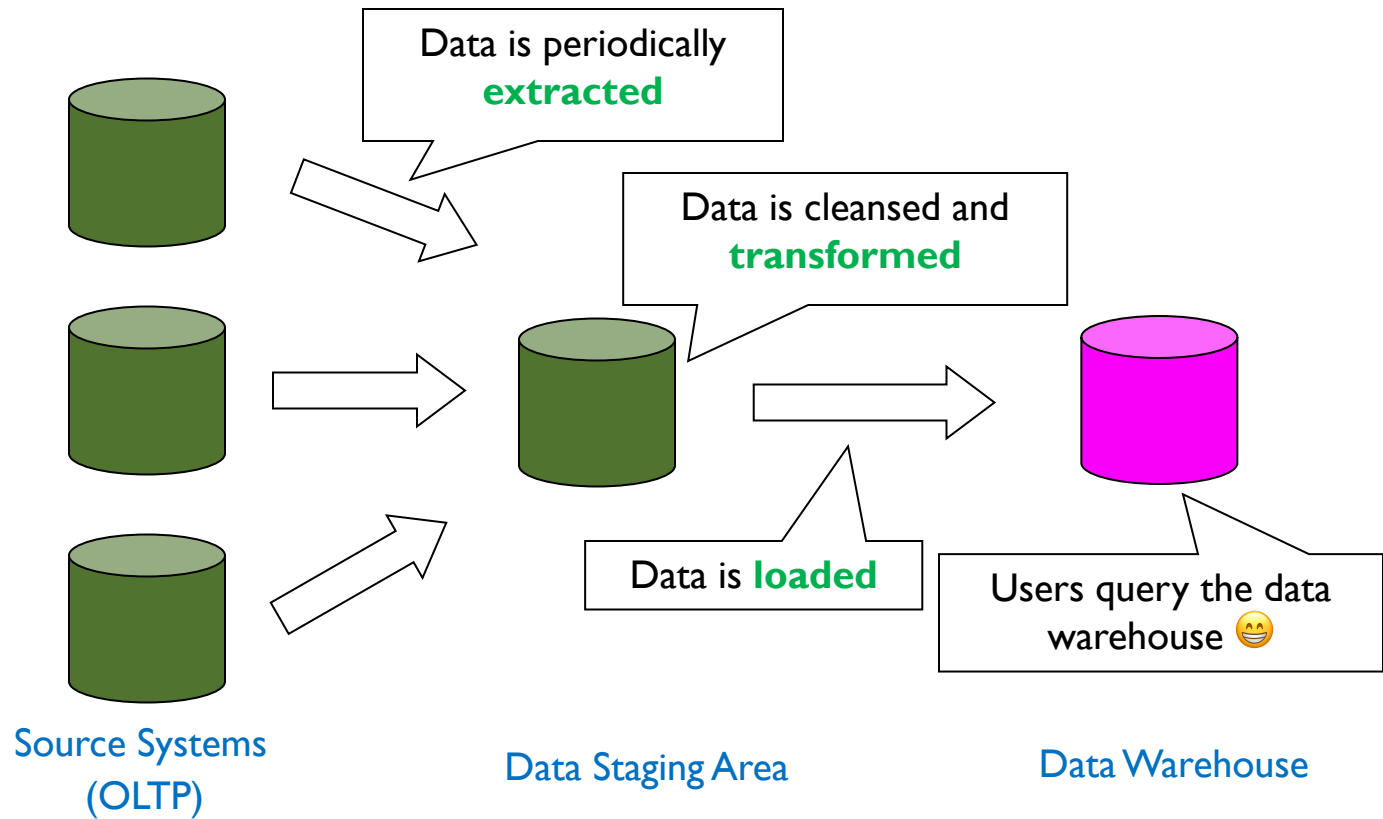


Figure 1-10 BI: data warehousing and analytical environments.

ETL Process



ETL process

- **Extract, Transform, Load**
- We are essentially talking about the integration of enterprise data
- 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

ETL challenges

- The complexity of the data warehouse
- Number of OLTP systems that data have to be extracted from
- The quality of data in the OLTP systems



- **Incremental load:** today's data is already loaded, no point to load the same data tomorrow.
- **Data duplication:** avoid loading the same data twice.
- Decide a **proper time slot** for loading data

Data extraction - Sourcing data

- ▶ What are the *proper* data sources?
 - ▶ Examine and verify - Can you get the necessary data for the DW?
 - ▶ The type of data extraction depends on how the data gets stored in the OLTP system.



Sourcing data for a retailer

► **Common Strategies:**

- Delivering superior customer service
- Satisfying customers' need

► **Data analytics help:**

- know customers, or customer segment
- understand customer buying preferences and patterns, historical transaction values, costs to serve
- provide information to make decisions on product mix, customer segment, optimising operations, lower cost to serve, etc.



What data are likely to be needed?

- Customer details
- Product information
- Transactions,
- Financial records,
- Costings,
- Competitors' offering, etc.

Sourcing data for a manufacturer

► **Common Strategies:**

- Optimising production operations
- Help promote better quality and consistency in production
- Improved work safety outcomes.

► **Data analytics help:**

- Report on operational on KPIs, and costing, etc.

What data types are likely to be needed?

- Production value chain data
- Procurement and financial data



Sourcing data



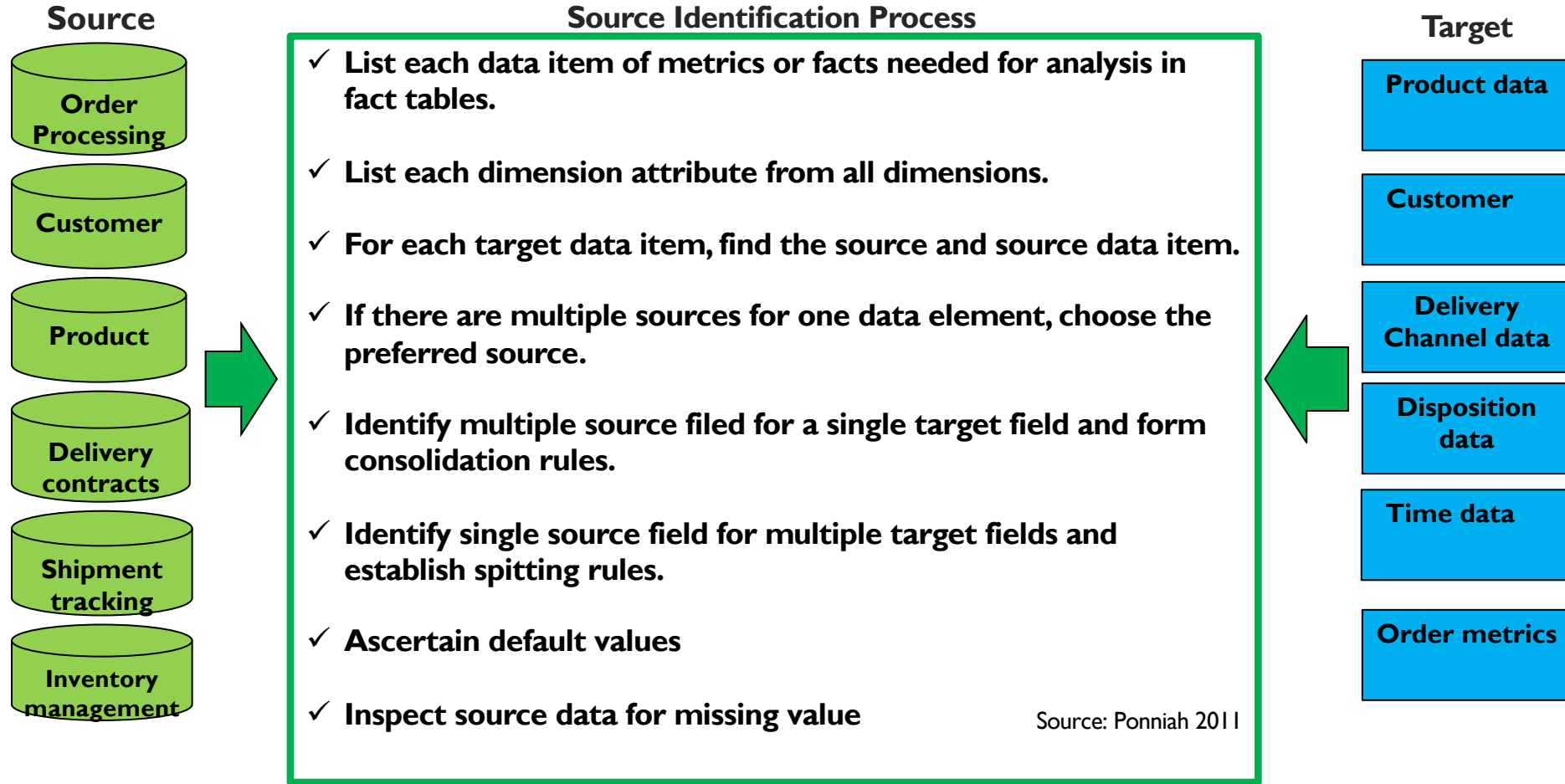
How can we decide?

- ✓ Depending on what analytics we need to build
- ✓ Depending on business needs and priorities.

Typical data sources

- ▶ **Internal data** sources: e.g. OLTP (customer master data store, HR, inventory, etc.)
- ▶ **External data** sources: e.g. economics data, weather data, Australian Bureau of Statistic Census, etc.
- ▶ **Big data**: e.g. from IoT sensors, social medial channels, etc.
- ▶ **Note:** Different data source types may require different mechanisms for getting and preparing data to load into the data warehouse

Sourcing data steps: mapping the sources to the targets



Data extraction: essential skills and knowledge



- ▶ What knowledge and skills do we need?

- ▶ I. Must have intimate knowledge of data sources
 - ▶ Time dependant data!
 - ▶ E.g. a person's address that may change over time.
 - ▶ Person ID A12345
 - ▶ From 1st Jan to 1st December 2005 – Lived in New York
 - ▶ From 2nd December 2005 to 20th Jan 2010 – Lived in Atlanta
 - ▶ From 21st Jan 2010 till now – is living in San Francisco
 - ▶ Person ID A12346
 - ▶ From 1st Jan to 1st December 2001 – California
 - ▶ From 2nd December 2002 till now – Canada
 - ▶ When do you update the DW?

Data extraction: Essential skills and knowledge

- ▶ 2. Also important to know how extracted data is used
 - ▶ When do we HAVE to update the data.
- ▶ 3. 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
 - ▶ 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 issues

- ▶ Think about 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
- ▶ 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)

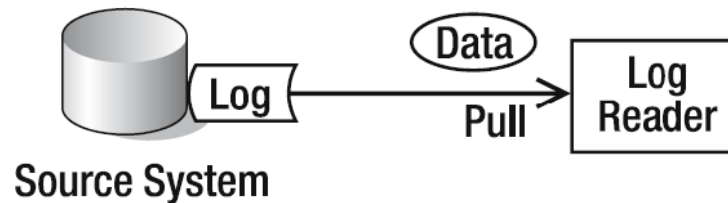


Immediate data extraction (real-time)

► Methods:

I. Capture via transaction logs

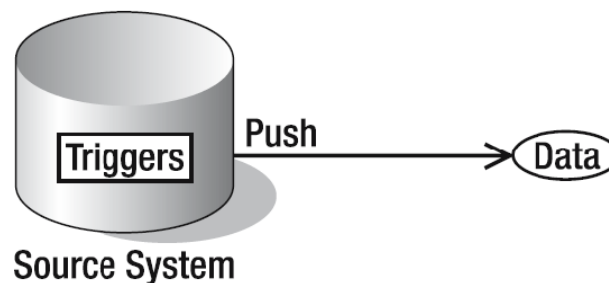
- Reads transaction logs and selects all committed transactions
- Must ensure you capture ALL logs
- Great if data comes from a database
- Could also use replication to get data into the ETL process



Immediate data extraction (real-time)

2. Capture via database triggers

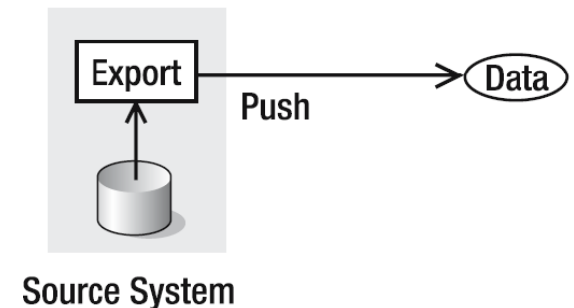
- ▶ Database only...
- ▶ Use triggers to generate data in separate file for all changes to data you want to track
- ▶ Additional burden on development effort – also changing source databases by adding triggers
 - ❖ Additional overhead...



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



Deferred data extraction (non real-time)

I. 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



Deferred data extraction (non 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 - Before moving extracted data to DW

- ▶ Data cleansing:
 - ▶ Clean the extracted data from each source: correction of mis-spellings, including resolution of conflicts between state codes and zip codes in the data sources, providing default values for missing data elements, or removing duplicated data
- ▶ Data standardisation:
 - ▶ Standardise data types and fields lengths for same data elements retrieved from the various sources
 - ▶ Semantic standardisation: resolve synonyms (2 or more terms from different source systems mean the same thing) and homonyms (a single term means many different things in different source systems)
- ▶ Data combination:
 - ▶ combining data from different sources, purging source data that is not useful and separating out source records into new combinations

Basic tasks of data transformation

1. 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

2. Splitting/Joining

- Manipulation of data
 - Splitting up records is uncommon
 - Joining info is very common (e.g. customer data)

3. Conversion

- Converting single fields
 - Standardise
 - Make fields understandable to users

Basic tasks of data transformation

4. Summarisation

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

5. 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
 - For example: 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 I, 2 or M, F – need to fix

Major Transformation Tasks

- Calculated and derived values
 - May need to calculate data points
 - For example: 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

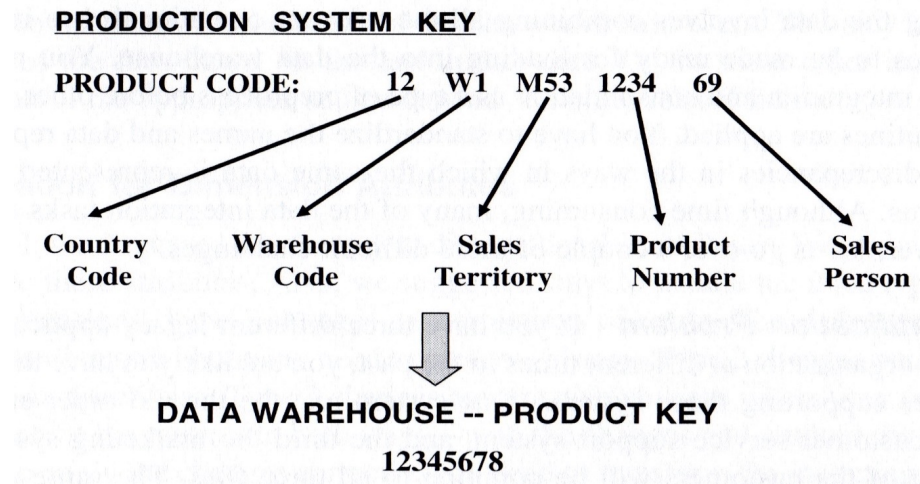
- Merging of information
 - Getting data about a particular thing all together in the DW
 - Merging info about a product from different sources
 - e.g. code, description, package types, cost
- Character set conversion
 - Different systems use different character sets (may not be compatible)
 - Must convert to DW character set
 - e.g. EBCDIC to ASCII, GB18030 to UTF-8
- Conversion of units of measurements
 - What is the standard of measurement for the organisation
 - May need to convert from imperial to metric

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

- ▶ Key restructuring
 - ▶ May 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
 - Data quality is often bad
 - Missing or default values
 - Multiple spellings of the same thing (Cal vs. UC Berkeley vs. University of California)
 - 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 like “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 from multiple source systems
 - For example: “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

- 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
 1. Load
 - Apply data directly to table, overwrites anything there
 2. Append
 - Adds data to the table, preserving what is already there
 3. Destructive Merge
 - Adds data to the table, if the key exists overwrite that record
 4. 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

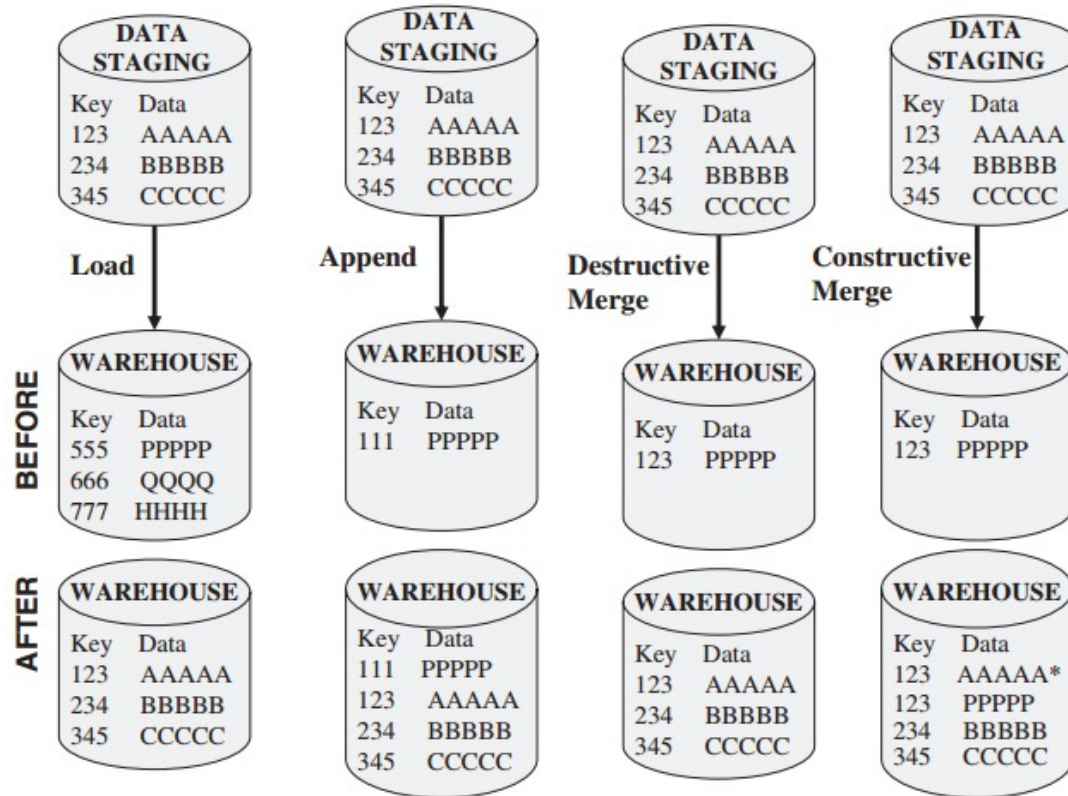


Figure 12-11 Modes of applying data.

Ponniah (2010) p304

ETL Summary

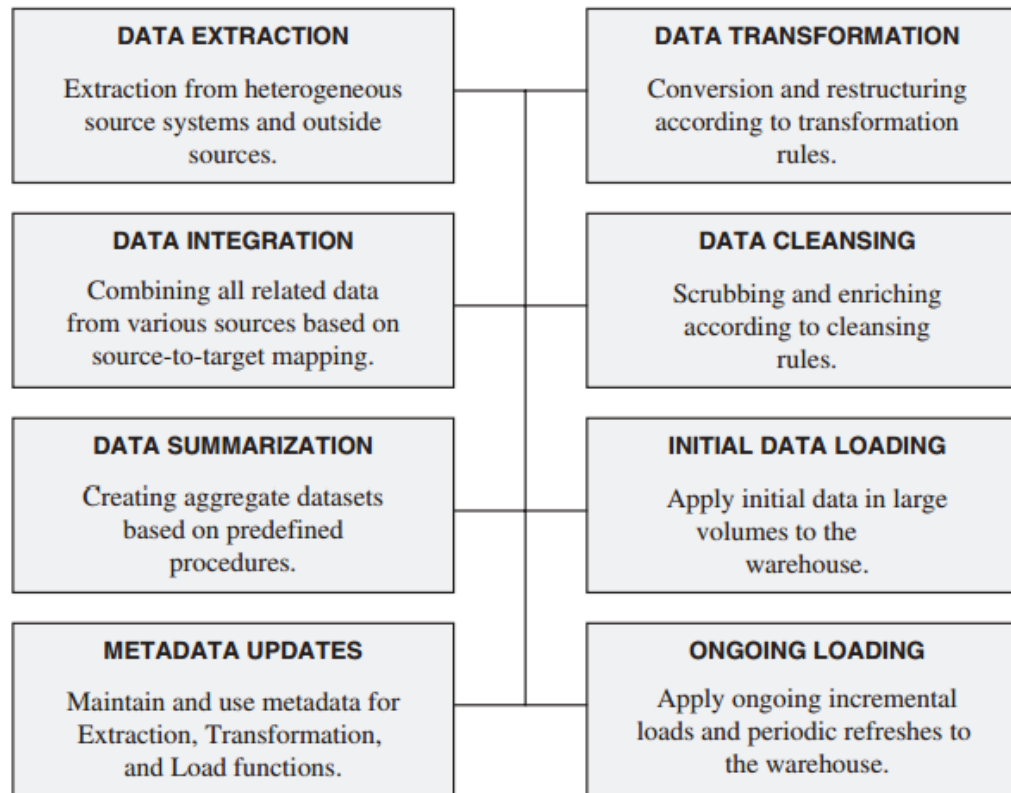


Figure 12-14 ETL summary.

Ponniah (2010) p310

ETL Tools

- 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](#) [Informatica](#)
 - [Oracle Data Integrator](#) [Pervasive Software](#)
 - [Pentaho](#) [SAS Data Integration Server](#)
 - [Safe Software](#) [SAP BusinessObjects Data Integrator](#)
 - [Benetl](#) [SQL Server Integration Services](#)
 - [Syncsort DMEexpress](#) [Talend Open Studio](#)

What can the tools do?

1. Data extraction from various relational databases, old databases, indexed files, and flat files
2. Data transformation from one format to another with variations in source and target fields
3. Performing of standard conversions, key reformatting, and structural changes
4. Provision of audit trails from source to target
5. Application of business rules for extraction and transformation
6. Combining of several records from the source systems into one integrated target record
7. Recording and management of meta-data

Review questions

- ▶ What is metadata and why it is important?
- ▶ What are the sources of metadata?
- ▶ What is ETL?
- ▶ For data extraction, where to source appropriate data?
- ▶ What are the ways of performing immediate data extraction and deferred data extraction?
- ▶ What are the 5 basic tasks of data transformation?
- ▶ What are the major data transformation tasks? Can you give some examples?
- ▶ What are the 4 ways to copy data to a data warehouse?