MIS781 Business Intelligence and Database

MODULE 5: EXTRACT, TRANSFORM, AND LOAD (ETL)

DEAKIN BUSINESS SCHOOL

To You

Hi Will,

I just wanted to say thank you, I've really enjoyed your BI lectures this semester. They are always fun and engaging, and your focus on refreshing the tutorial content and discussing real world applications has been great.

It was due to what I've learnt in this course that I was able to land a role with PetCircle as a data engineer. Part of the application process involved a case study for where I had to handle everything we've covered in this subject, responding to business needs, explaining my methodology for ETL, and discussing the various reports that would be generated to meet the business needs. While it was rather terrifying (I had only a few days to write the report while still working), it was great fun to be able apply what I've learnt directly.

So again, thanks for your lectures, they've been great.

Kind regards,

Data warehouse vs Data warehousing

Data warehouse

- A data warehouse is a collection of data created to support decision-making applications

Data warehousing

- Data warehousing is the entire process of data extraction, transformation, and loading of data to the warehouse and the access of the data by end users and applications.

Learn this today

Source: Watson 2017, Teradata University Network

Data Extraction, Transformation and Loading (ETL)

One of the most important and time consuming tasks in the DW space.

Learning objectives

By the end of this class, you should be able to:

- Understand what an ETL is.
- Understand, explain, and interpret the steps in ETL process.

Data Warehouse versus OLTP

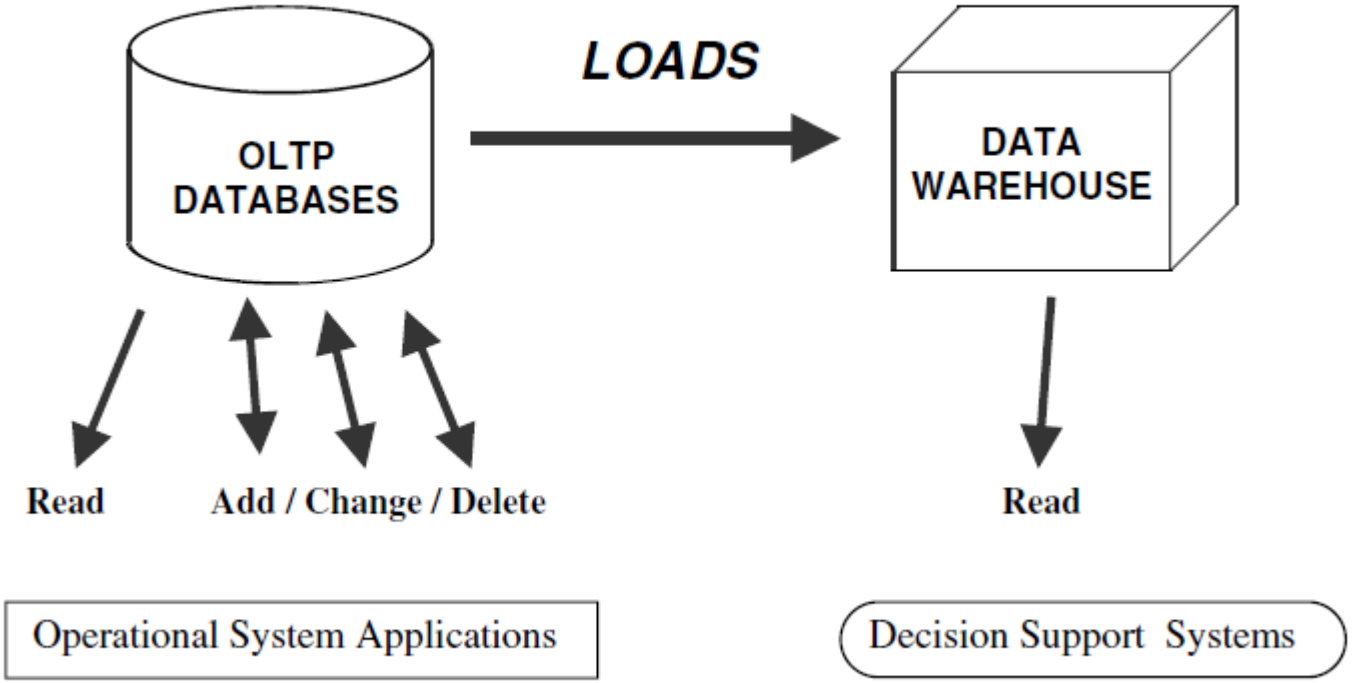
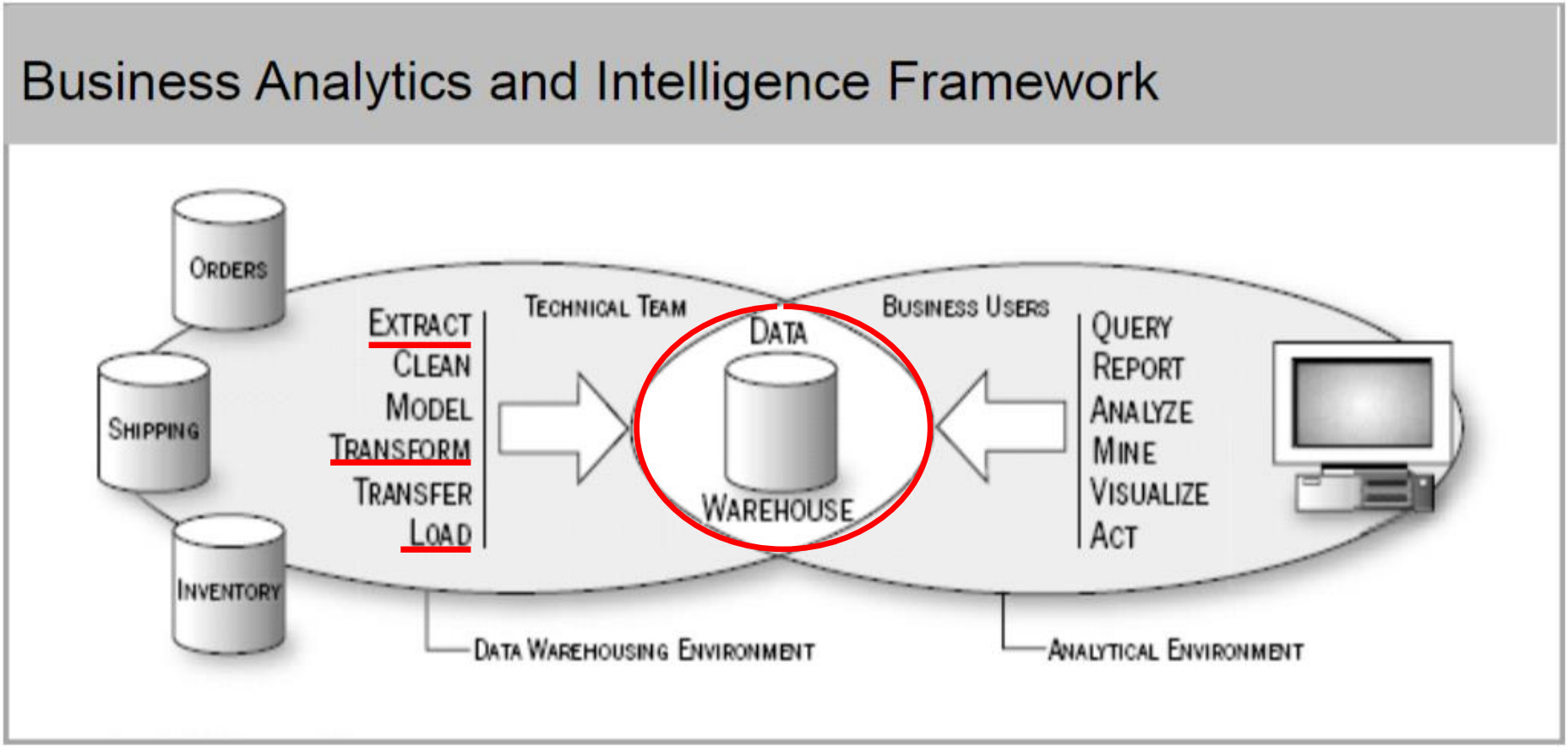


Figure 2-3 The data warehouse is nonvolatile.

Data Warehouse

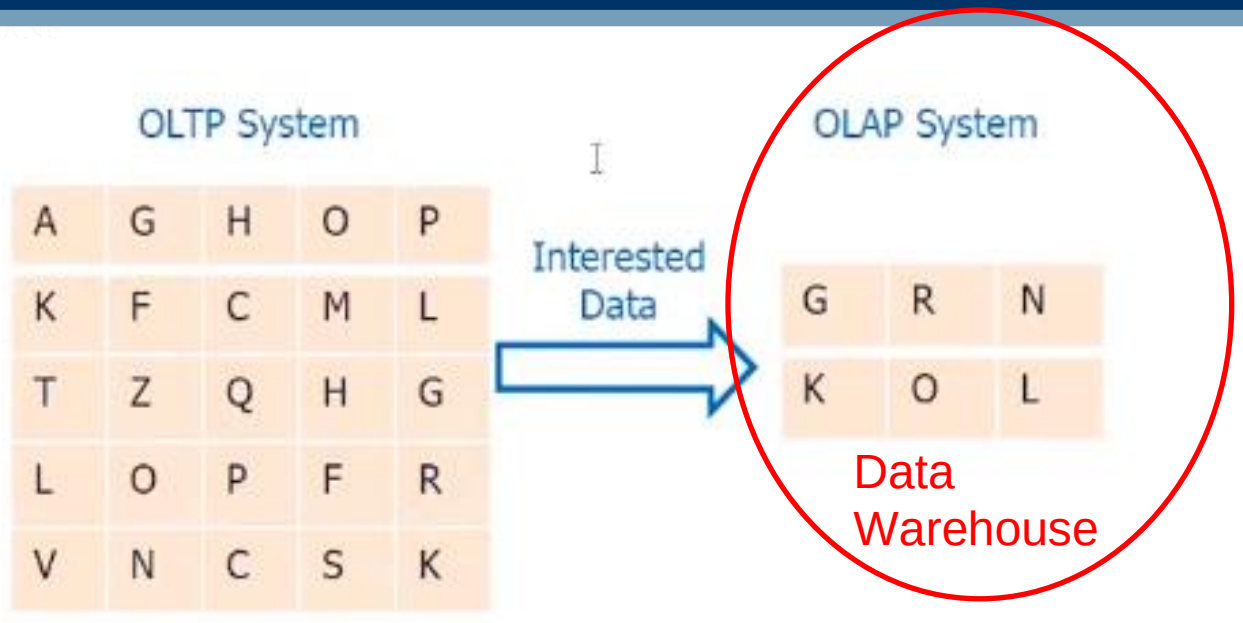


What is ETL?

- 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.

Video: What is an ETL?

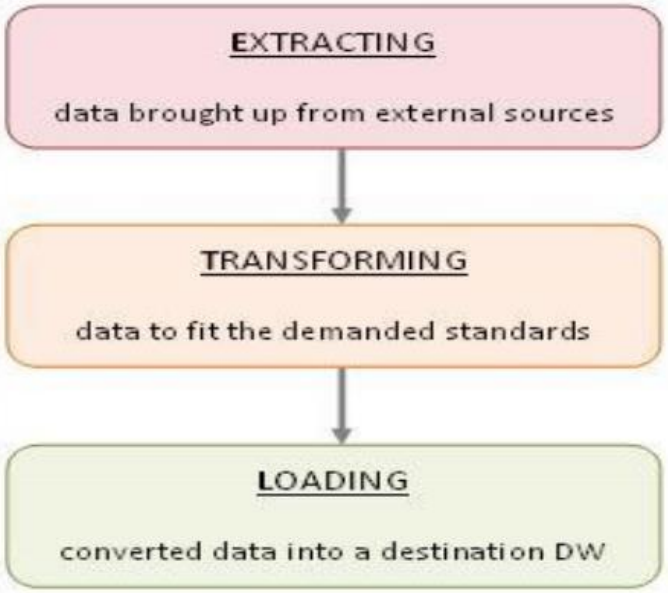
What is ETL?

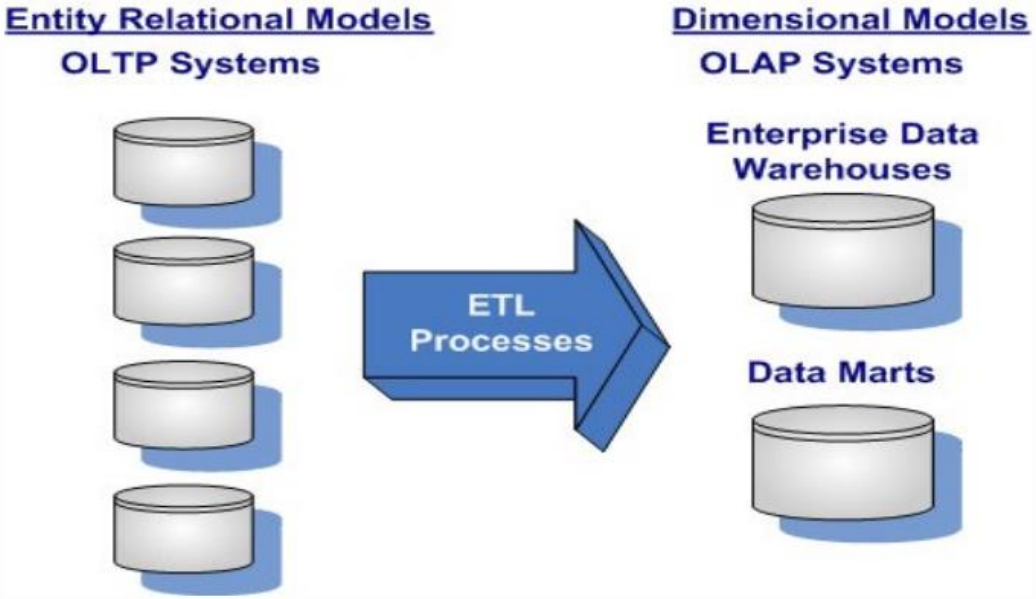


- We have: Source systems (OLTP) -> Target systems (OLAP or Data Warehouse).
- How do we transfer the data from the source systems to target systems?

This set of methods is called the ETL (Extract, Transform, and Load) process.

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

Data is periodically extracted

Data is cleansed and transformed

Data is loaded

Users query the data warehouse

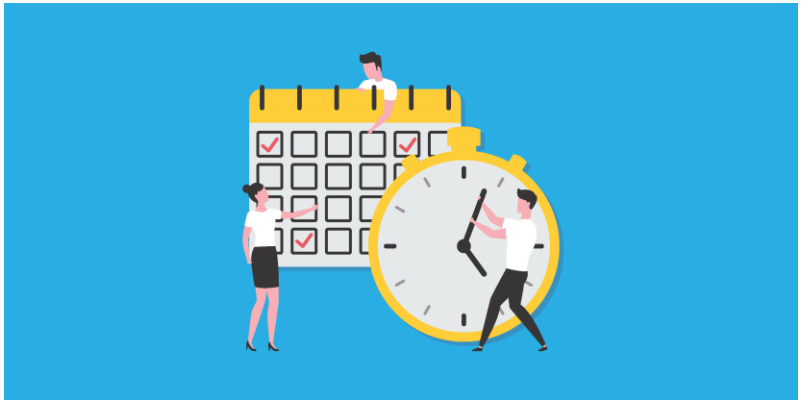
Source Systems (OLTP)

Data Staging Area

Data Warehouse

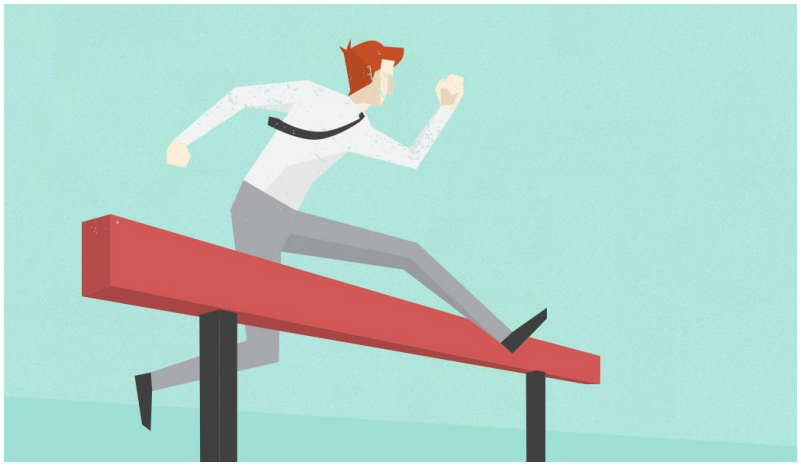
ETL process

- 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 has 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

Major steps in ETL Process

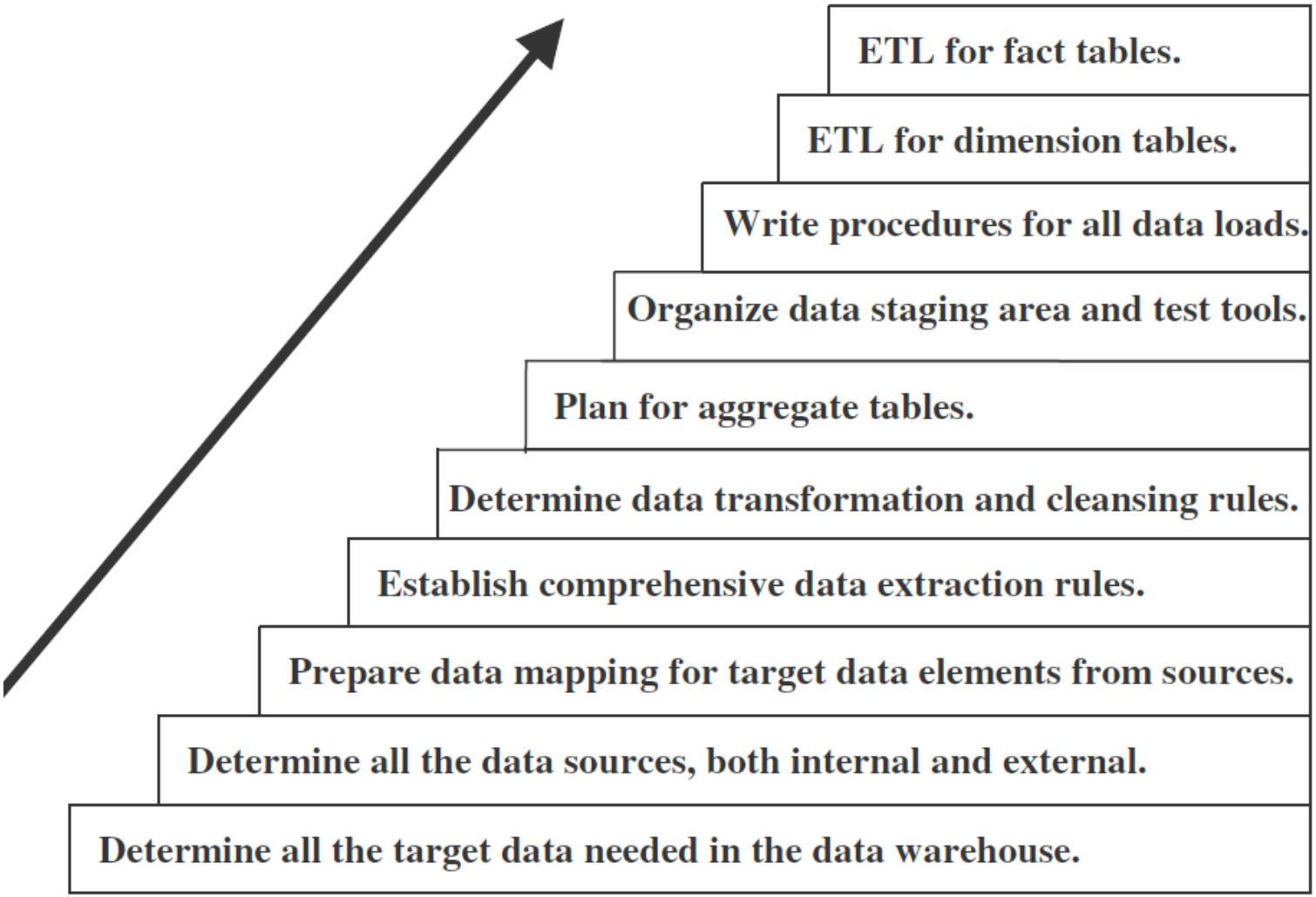


Figure 12-1 Major steps in the ETL process.

Source Ponniah 2011

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.

- What drives data sourcing decisions at the start a business analytics journey?

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 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 Statistics, 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

Source

- Order Processing
- Customer
- Product
- Delivery contracts
- Shipment tracking
- Inventory management

Source Identification Process

- ✓ List each data item of metrics or facts needed for analysis in fact tables.
- ✓ List each dimension attribute from all dimensions.
- ✓ For each target data item, find the source and source data item.
- ✓ If there are multiple sources for one data element, choose the preferred source.
- ✓ Identify multiple source filed for a single target field and form consolidation rules.
- ✓ Identify single source field for multiple target fields and establish spitting rules.
- ✓ Ascertain default values
- ✓ Inspect source data for missing value

Target

- Product data
- Customer
- Delivery Channel data
- Disposition data
- Time data
- Order metrics

Data extraction: Essential skills and knowledge

- What knowledge and skills do we need?

1. 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 (again slowing changing dimensions?)
  - History is preserved in operational system
    - Thus easy to get history into DW

Slowly Changing Dimension (SCD) concept

- "Slowly Changing Dimension" is a common issue in data warehousing, because attribute for a record varies over time

E.g.:

Christina is a customer with XYZ Inc. She first lived in Chicago, Illinois. So, the original entry in the customer lookup table has the following record:

Customer Key	Name	State
1001	Christina	Illinois

At a later date, she moved to Los Angeles, California on 1 January, 2016. How should XYZ Inc now modify its customer table to reflect this change? This is the SCD problem.

Source: http://www.1keydata.com/datawarehousing/slowly-changing-dimensions.html

Slowly Changing Dimension (SCD)

- Data Warehouse designers have sorted out **three major approaches to SCDs**. These are called TYPE 1, TYPE 2 and TYPE 3.

1. A **Type 1** SCD is an **overwrite** of a dimensional attribute. The new record replaces the original record. No trace of the old record exists.
2. A **Type 2** SCD **creates a new record** for each change. A new record is added into the customer dimension table. Therefore, the customer is treated essentially as two people.
3. A **Type 3** SCD **adds a new field** in the dimension record but does not create a new record. The original record is modified to reflect the change

Customer Key	Name	State
1001	Christina	Illinois

Read: http://www.1keydata.com/datawarehousing/slowly-changing-dimensions.html

SCD Example

Type 1 SCD

Customer Key	Name	State
1001	Christina	California

Type 2 SCD

Customer Key	Name	State
1001	Christina	Illinois
1005	Christina	California

Type 3 SCD

Customer Key	Name	Original State	Current State	Effective Date
1001	Christina	Illinois	California	1 January, 2016

Slowly changing dimension (SCD) – Example 2

E.g. 2:

- Universities (and other organisations) are not static. Faculties are created/disbanded, schools are opened/closed. Courses are modified, Campuses open/close.
- People want to see their data with the relationships which existed at the time it was current – eg the school which is now closed.
- People also want to see their data with the relationships which exist today – eg units history – regardless of the fact it was in a different school.
- People want to see data in ways they haven't thought of yet!

Example of a real-world problem

- We have a school – Health and Behaviour sciences (HBS).
- The school has two units – HL84 (Health), and BS92 (Behaviour Science).
- HL84 started with 50 EFTSL in 2000 and BS92 started with 60 EFTSL. Each increased by 5 EFTSL a year.

Simple Load Report by School

	2000	2001	2002	2003	2004
HBS	110	120	130	140	150

Example of a real-world problem

- The School is now split into Health (HHL) and Behaviour Sciences (BSS) from 2005. The Two units are allocated accordingly to the new schools.

The Simple Load Report by School now looks like

	2000	2001	2002	2003	2004	2005
HBS	110	120	130	140	150	
HHL						75
BSS						85

Example of a real-world problem

- But of course, while the users/clients agree that the report is accurate.. What they really want is..

A "amended history" Simple Load Report by School

	2000	2001	2002	2003	2004	2005
HHL	50	55	60	65	70	75
BSS	60	65	70	75	80	85

- Oh.. But don't change the data or anything.. We might need to report it the other way as well!

Type 1,2,3 Problems

- These three types of slowly changing dimensions handle most of the situations faced by the DW Designer.

Like this….

	2000	2001	2002	2003	2004	2005
HBS	110	120	130	140	150	
HHL						75
BSS						85

Or like this….

	2000	2001	2002	2003	2004	2005
HHL	50	55	60	65	70	75
BSS	60	65	70	75	80	85

Using Standard Type 2 to represent the data

- If we use a standard type 2 approach to represent this data we would have the following.

Dimension Table

Key	Code	Description
1	HBS	Health and Behaviour sciences
2	HHL	Health
3	BSS	Behaviour Sciences

Fact Table

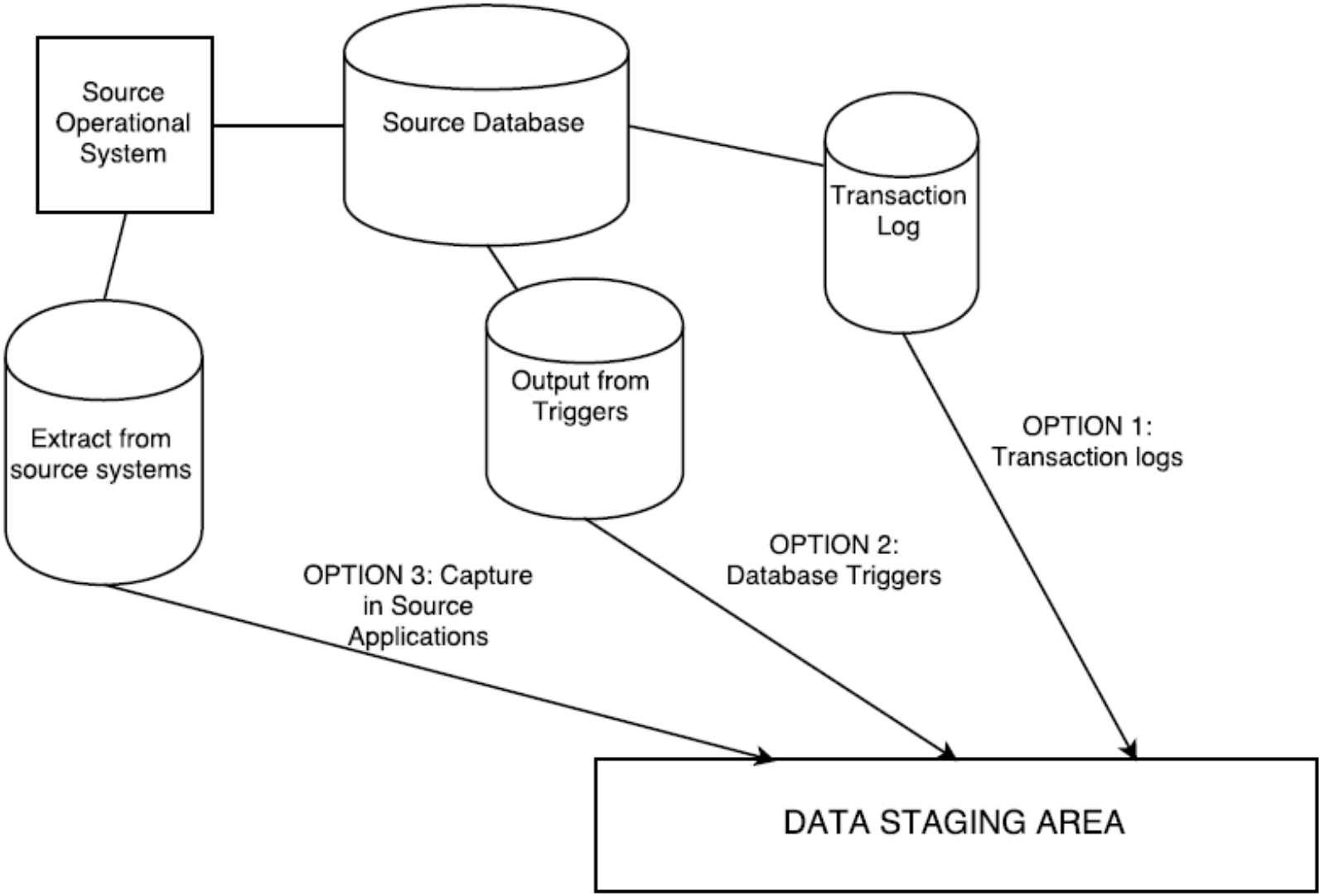
Key	Year	Course	EFTSL
1	2000	HL84	50
1	2000	BS92	60
1	2001	HL84	55
1	2001	BS92	65
1	2002	HL84	60
1	2002	BS92	70
1	2003	HL84	65
1	2003	BS92	75
1	2004	HL84	70
1	2004	BS92	80
2	2005	HL84	75
3	2005	BS92	85

Using Standard Type 3 to represent the data

- We could use a type 3 dimension and store the Original value of the records.. Eg..

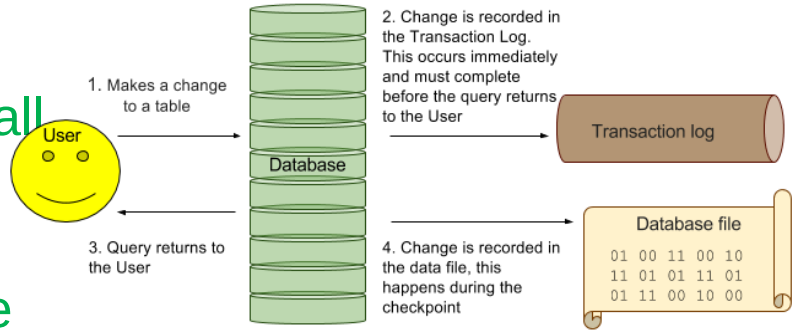
Key	Code	Desc	Old Code	Old Desc
1	HBS	Health and Behaviour sciences	HBS	Health and Behaviour sciences
2	HHL	Health	HBS	Health and Behaviour sciences
3	BSS	Behaviour Sciences	HBS	Health and Behaviour sciences

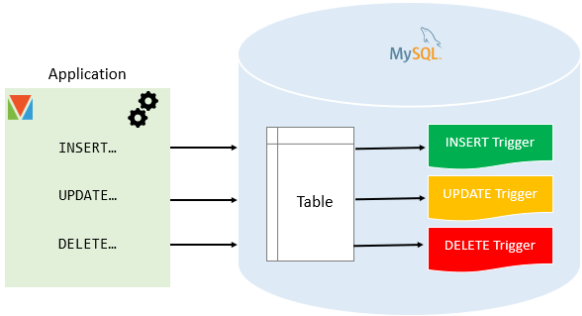
Data Extraction Types: Immediate data extraction – REAL TIME: 3 Options



Data Extraction Types: Immediate data extraction – REAL TIME!

- Methods:
  1. 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
  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...



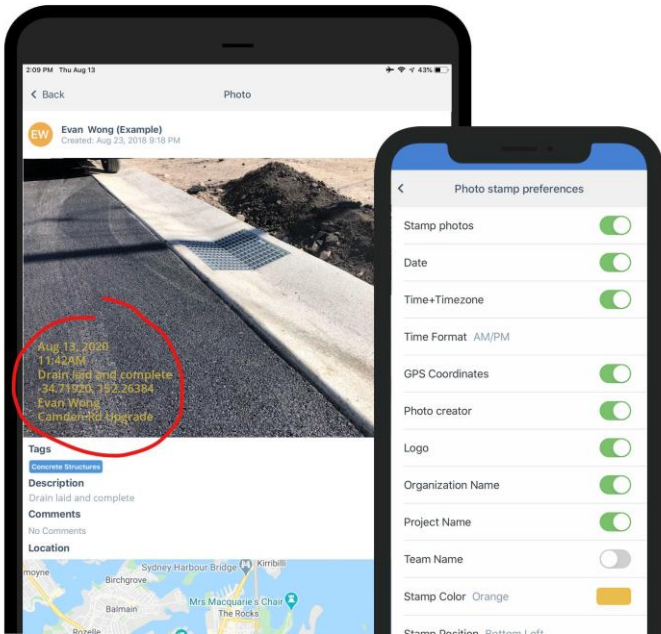


Data Extraction Types: Immediate data extraction – REAL TIME!

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 ELT 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

We have the RAW data...

Not good enough for the DW

- Quality
- Format

Before moving extracted data to DW

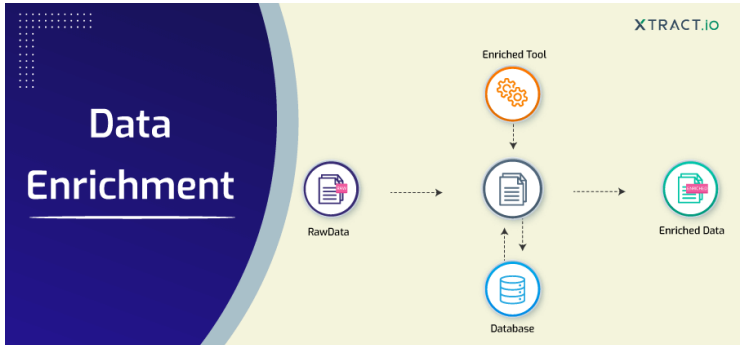
- **Data cleansing:**
  - Clean the extracted data from each source: correction of mis-spellings, including resolution of conflicts between state codes and zip/post codes in the data sources, providing default values for missing data elements, or removing duplicated data
- **Data standardisation:**
  - 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 outsourced records into new combinations

Initial and Basic Tasks in ETL

- 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

Initial and Basic Tasks in ETL

- Summarisation
  - Depending on level of detail required some data can be summarised
    - Eg:
      - 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

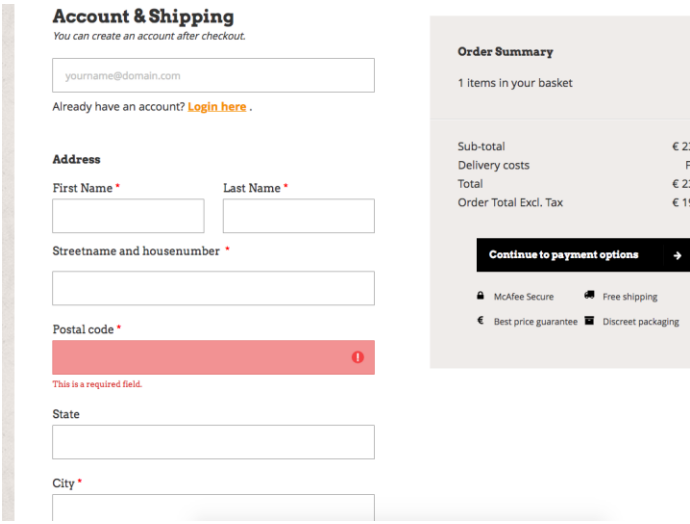


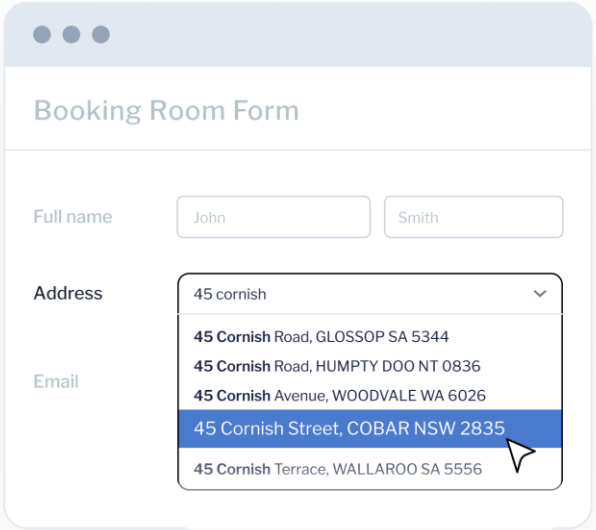
Major Transformation Tasks

- 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, postcode
    - Customer name breakdowns also
  - Important
    - Can index things like postcode
    - Allows for analysis on components



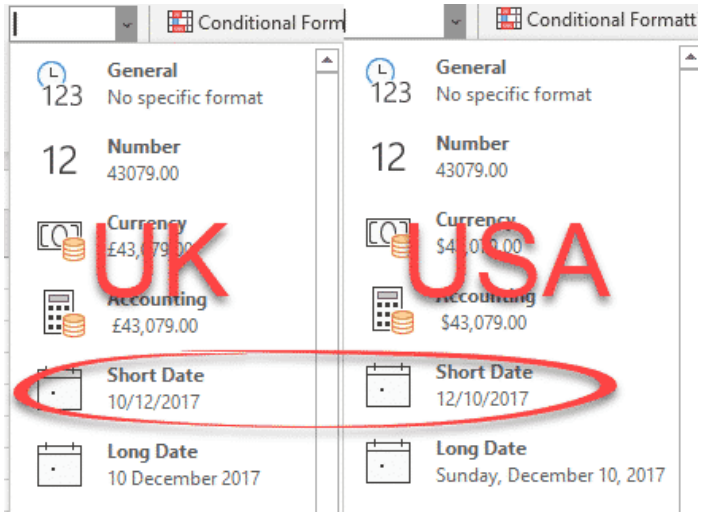


Major Transformation Tasks

- 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 (e.g. ounce, pound, inch, foot etc.) to metric (kg., km., etc.)

Major Transformation Tasks

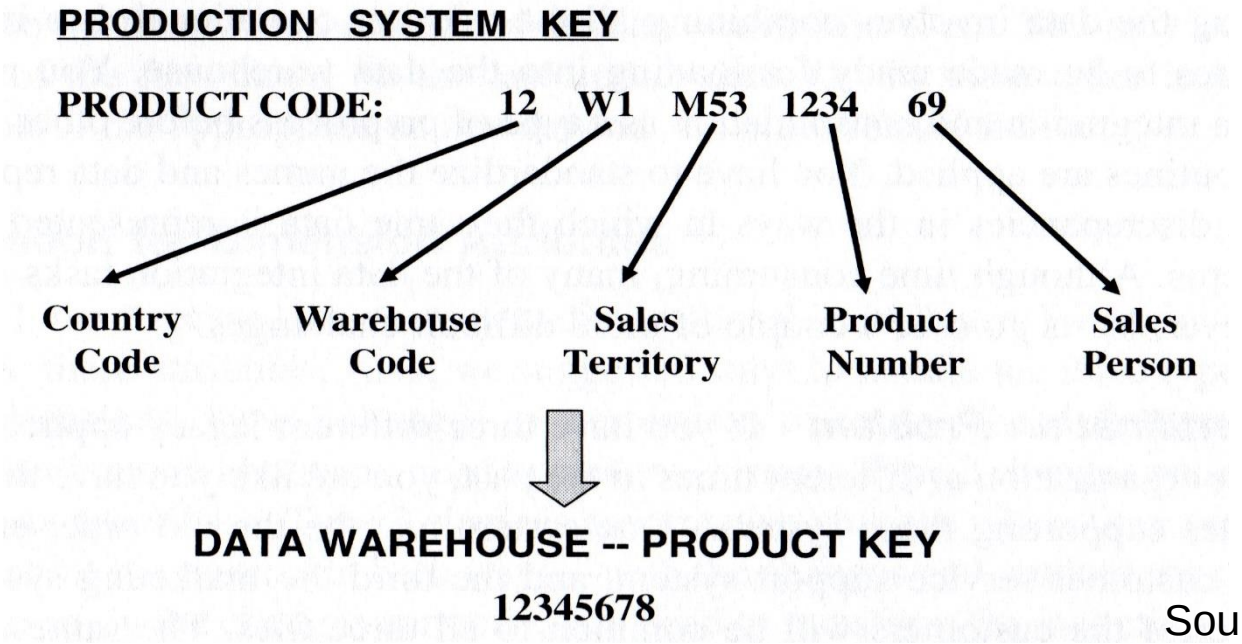
- Date/time conversion
  - Different systems may use different formats
  - Need to be clear
    - 11/12/2011
      - 11 Dec 2011 or 12 Nov 2011
      - Store it in a standard format
        - » 11 DEC 2011



- 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 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

Once the transformation of data is complete the load can start!

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 (see next slide)
  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

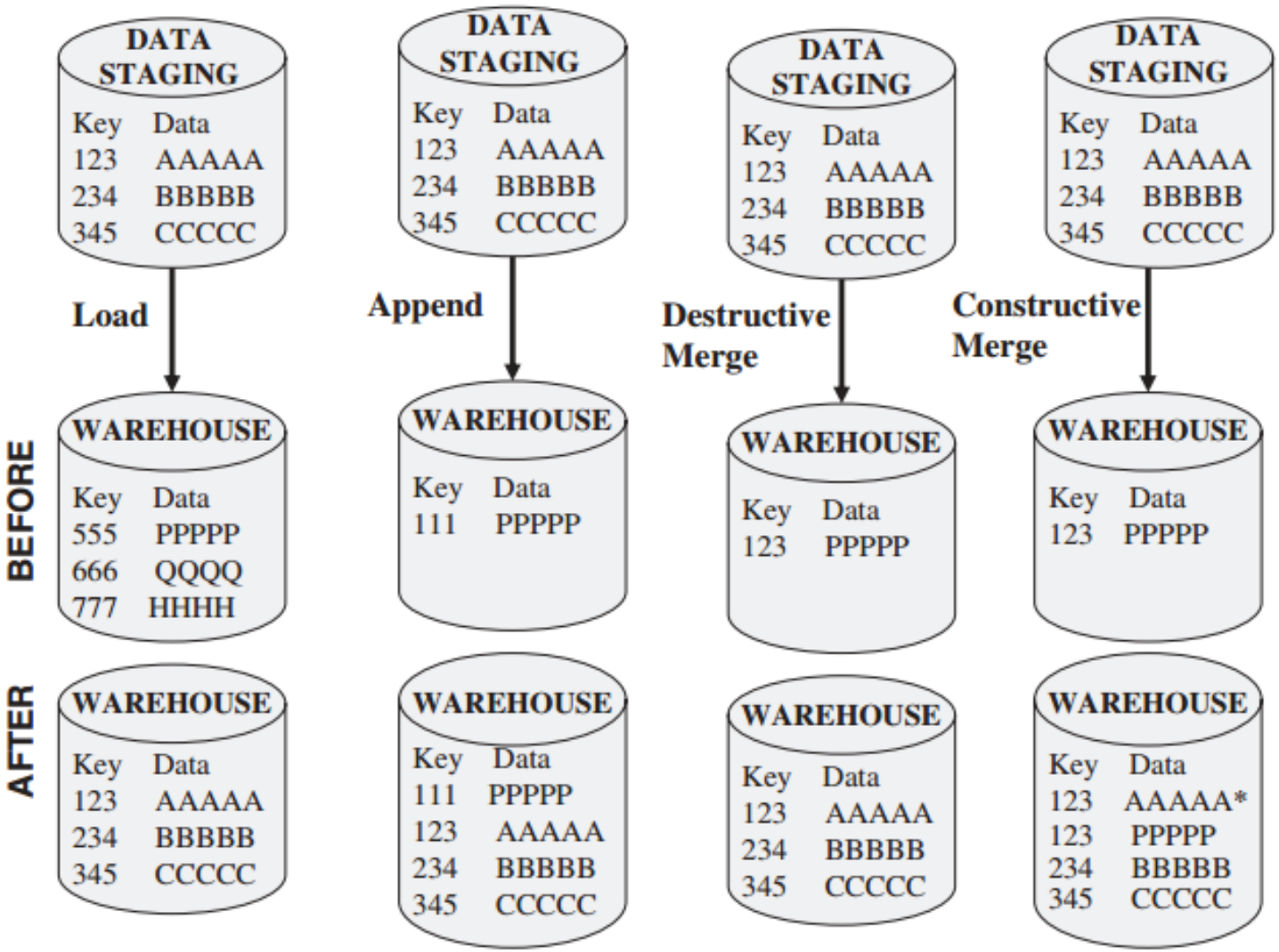


Figure 12-11 Modes of applying data.

Ponniah (2010) p304

ETL Summary

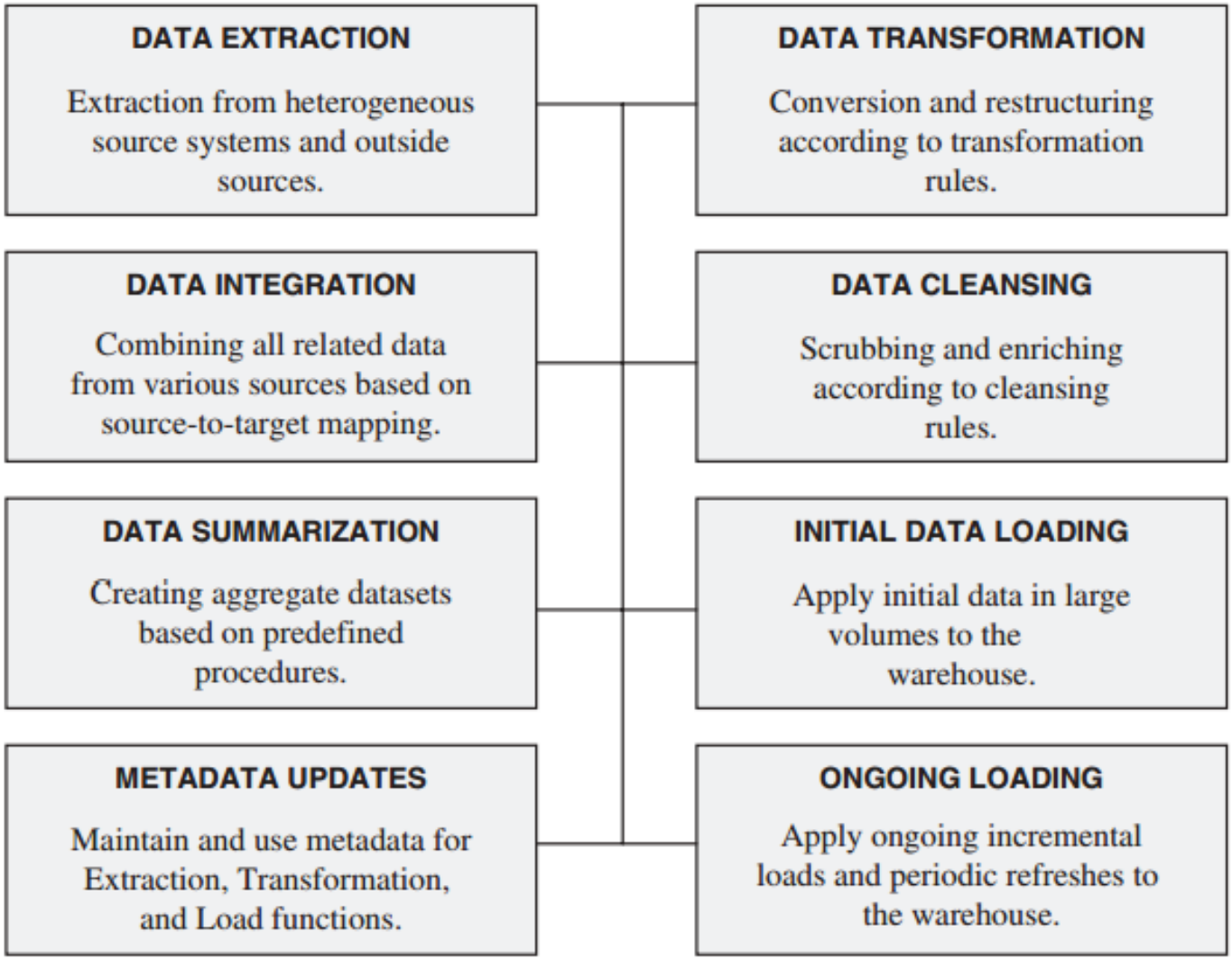


Figure 12-14 ETL summary.

ETL Tools

Its not all manual labour after all...

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
  - Power BI
  - 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

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

ETL Tools: Video ETL Process and tools

Video: What is an ETL?

If you want to be a Power BI expert this site has lots of videos tutorials on it: https://learn.microsoft.com/en-us/training/browse/?products=power-bi

Practical Assignment Discussion