MIS781 Business Intelligence and Database

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

DEAKIN BUSINESS SCHOOL

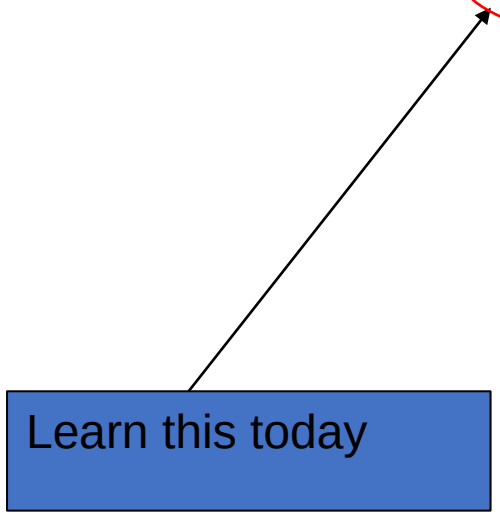
Recap: Data warehouse vs Data warehousing

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.



Data Extraction, Transformation and Loading (ETL)

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

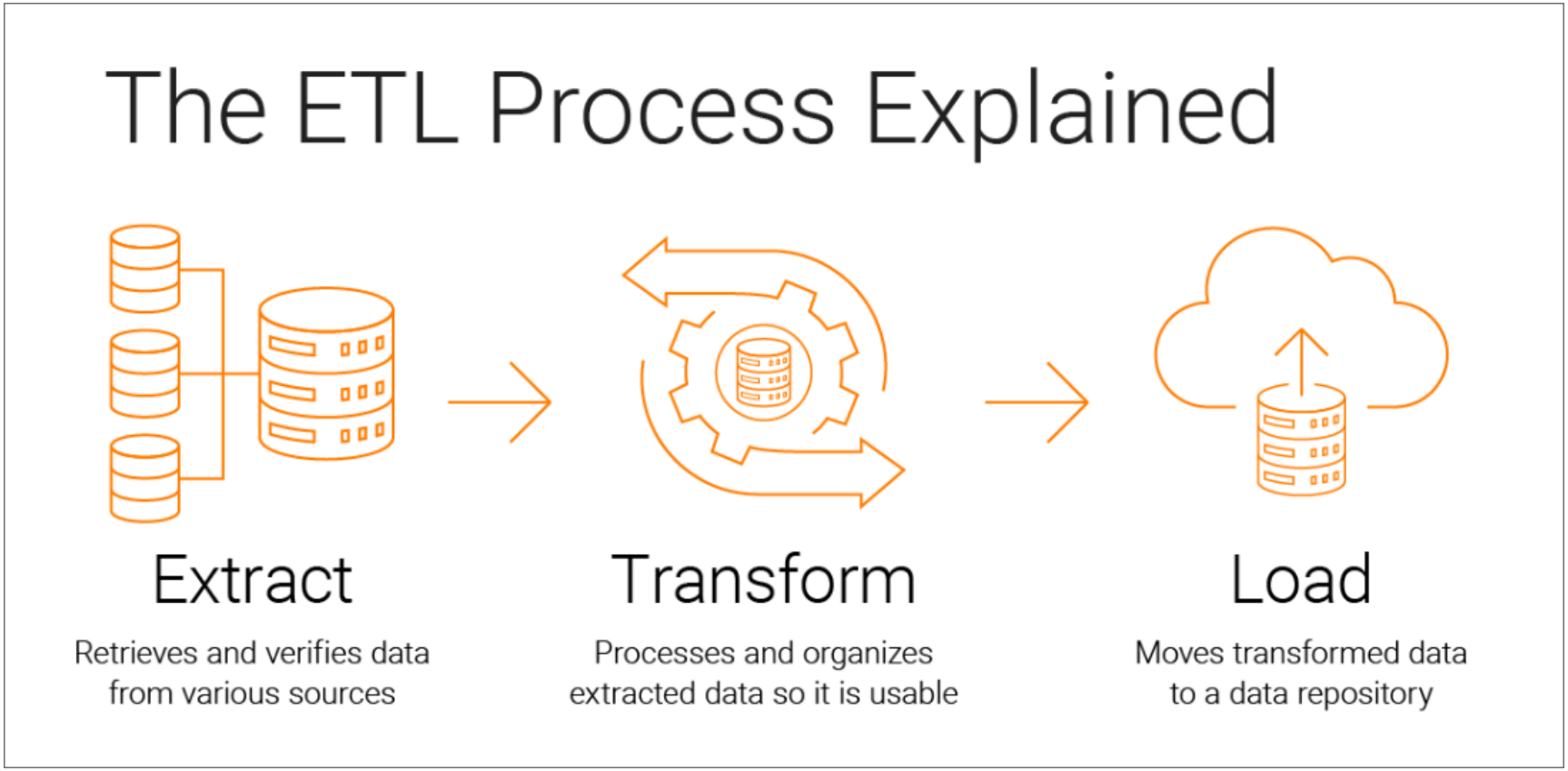
Learning objectives

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.

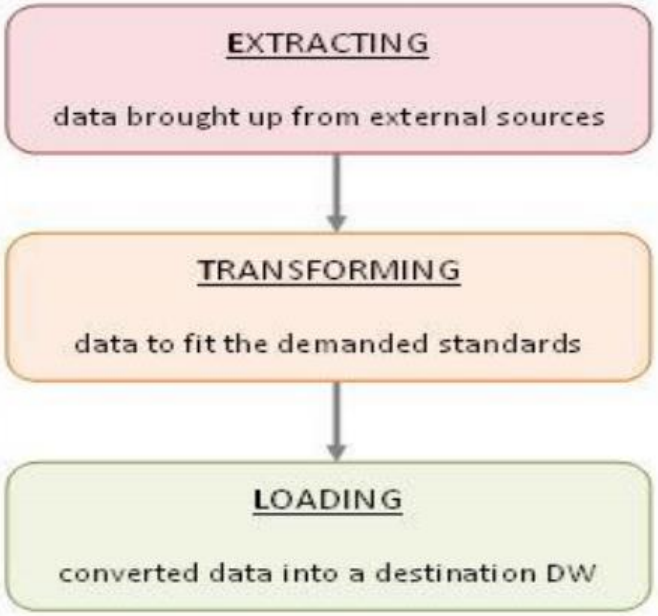
What is ETL?

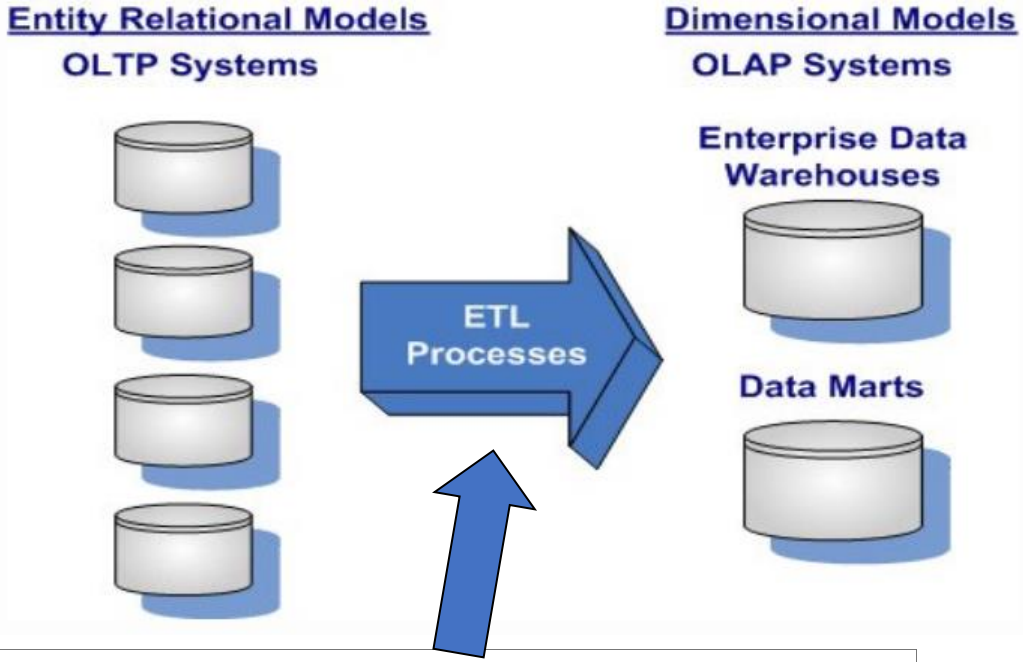
- ETL stands for extract, transform, and load. It's a three-step data integration process used to combine raw data from multiple data sources into a data warehouse.

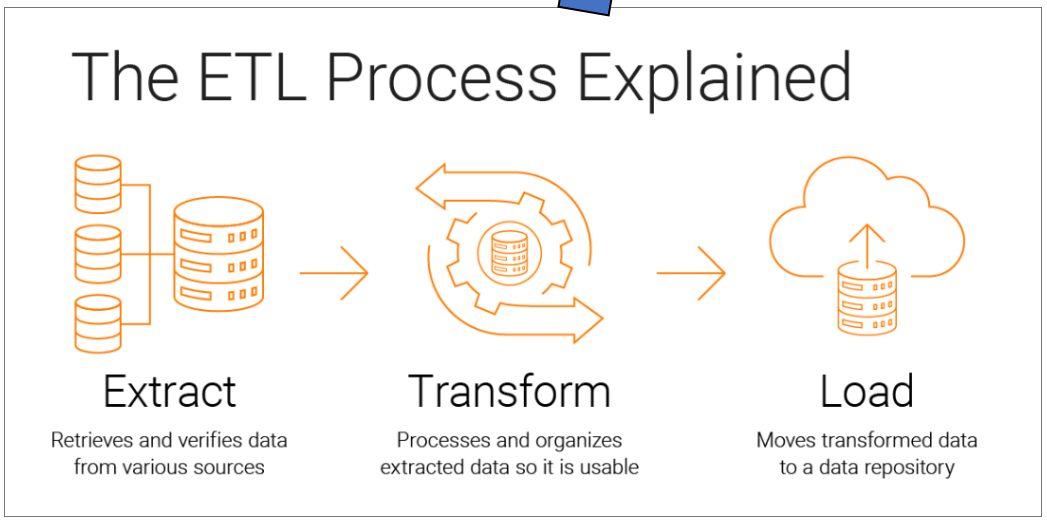


Source: informatica.com/au/resources/articles/what-is-etl.html

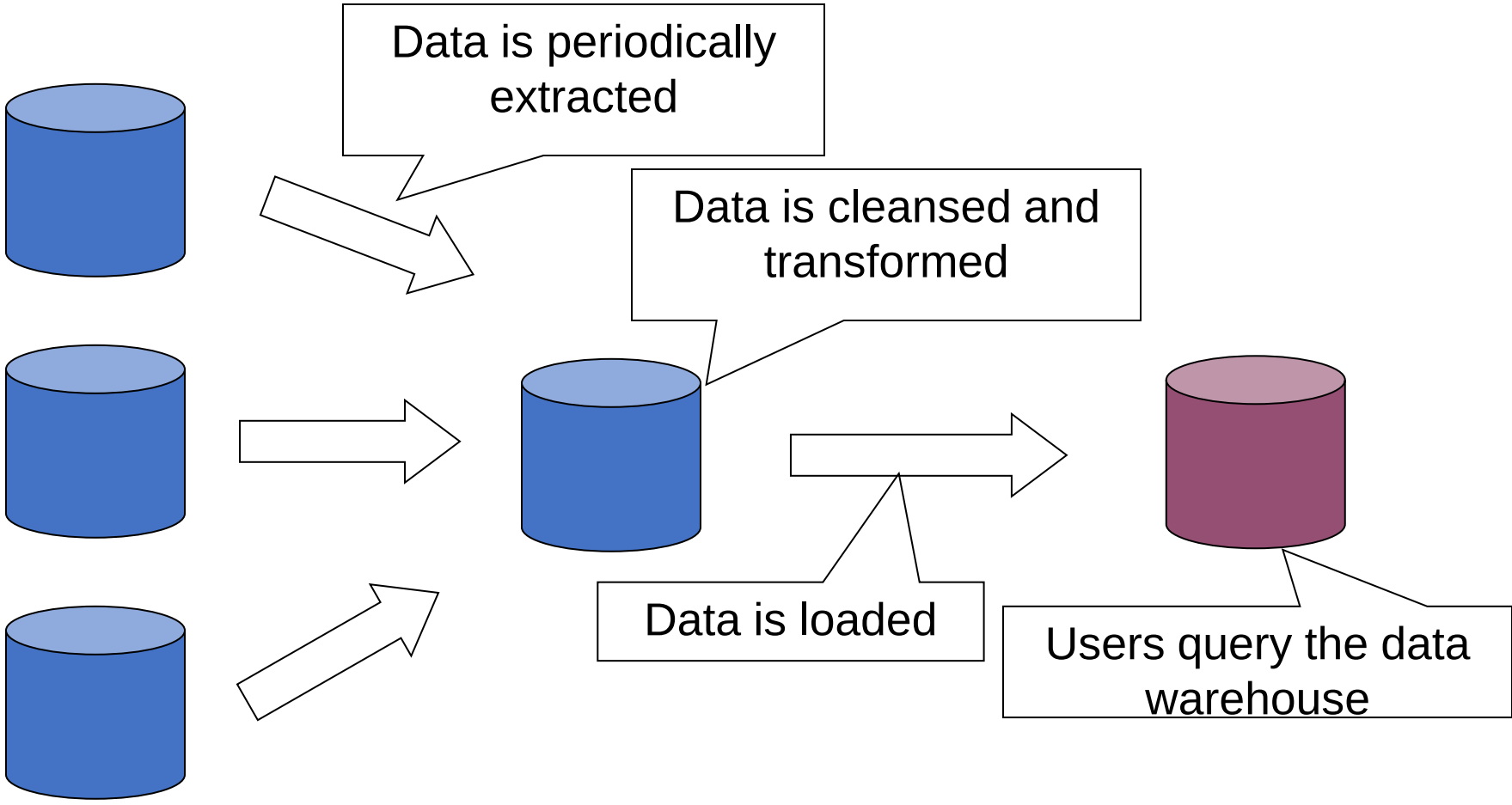
ETL Process







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

What is ETL?

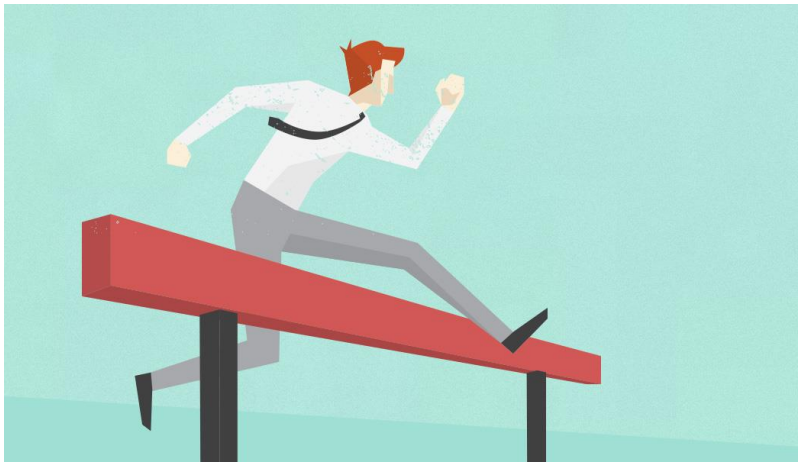
- Online Transaction Processing (OLTP) systems cannot be used for analytics. Therefore, Online Analytical Processing (OLAP) is needed.
- Doing OLTP and OLAP in the same database system is often impractical :
  - Different performance requirements
  - Different data modelling requirements
  - Analysis 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.

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

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

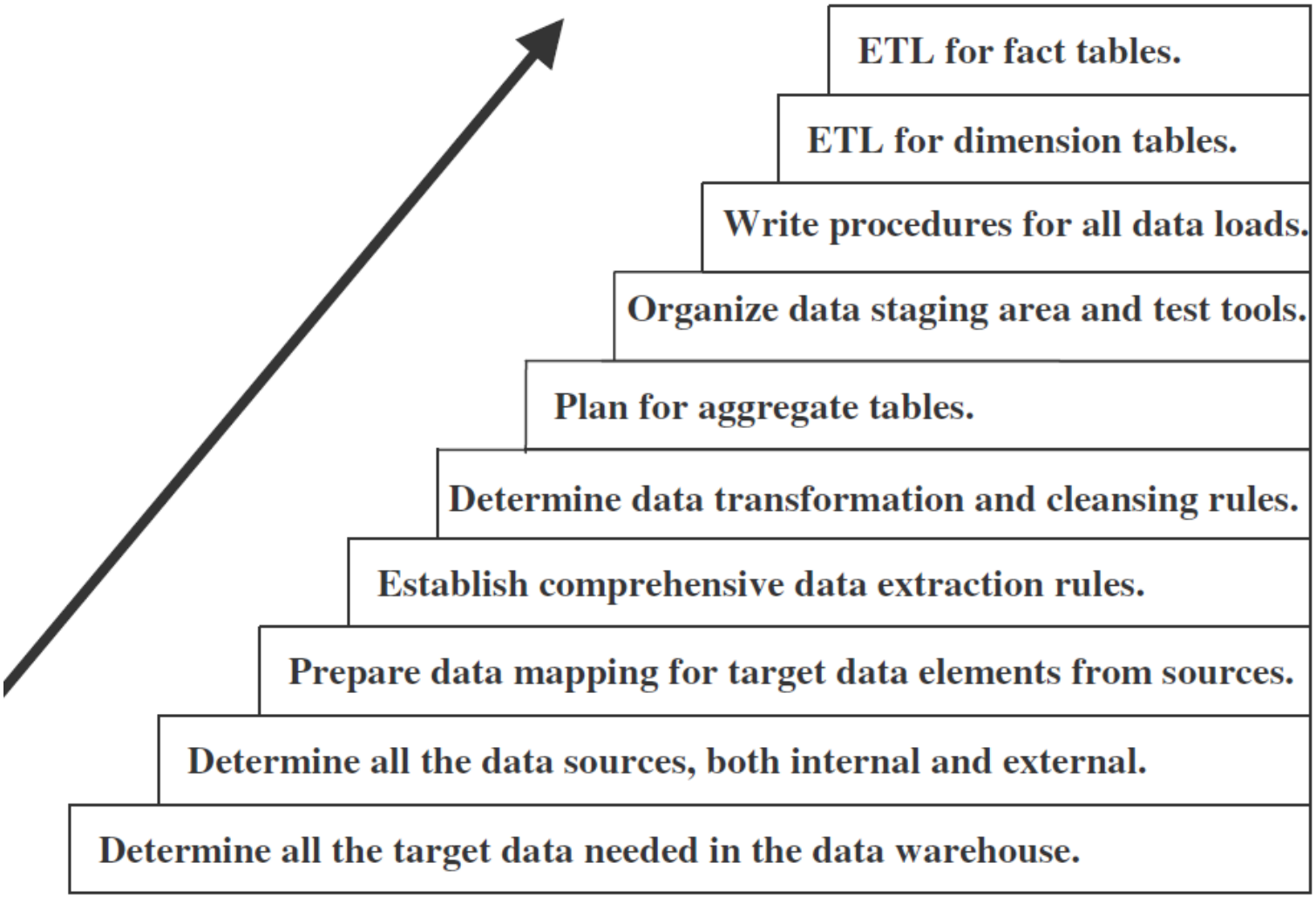


Figure 12-1 Major steps in the ETL process.

Source Ponniah 2011

Video: [What is an ETL?]

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?
- See next slides example

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. https://www.abs.gov.au/
- **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

Source Ponniah 2011

Target

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

Data extraction: Essential skills and knowledge

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?

- What knowledge and skills do we need?

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** (see next slide)
   - Lots of issues around this

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

Data Extraction Types: Immediate data extraction – REAL TIME!

- - Capture via transaction logs
  - Reads transaction logs and selects all committed transactions
  - Must ensure you capture ALL logs
  - Could also use replication to get data into the ETL process

- Capture in source applications
  - Source applications are modified to ALSO capture data warehouse data

Results | Messages

	Current LSN	Transaction ID	Operation	Transaction Name	CONTEXT	AllocUnitName	Page ID	Slot ID	Begin Time
1	00000016:00000132:0009	0000:00000451	LOP_BEGIN_XACT	SplitPage	LCX_NULL	NULL	NULL	NULL	2013/09/27
2	00000016:0000014a:0009	0000:00000455	LOP_BEGIN_XACT	SplitPage	LCX_NULL	NULL	NULL	NULL	2013/09/27
3	00000016:0000014a:0013	0000:00000456	LOP_BEGIN_XACT	SplitPage	LCX_NULL	NULL	NULL	NULL	2013/09/27

	Current LSN	Transaction ID	Operation	Transaction Name	CONTEXT	AllocUnitName	Page I
1	00000016:00000132:0009	0000:00000451	LOP_BEGIN_XACT	SplitPage	LCX_NULL	NULL	NULL
2	00000016:00000132:000a	0000:00000451	LOP_MODIFY_ROW	NULL	LCX_PFS	sys.sysobjvalues.clst	0001
3	00000016:00000132:000b	0000:00000451	LOP_HOBT_DELTA	NULL	LCX_NULL	NULL	NULL
4	00000016:00000132:000c	0000:00000451	LOP_FORMAT_PAGE	NULL	LCX_CL...	sys.sysobjvalues.clst	0001
5	00000016:00000132:000d	0000:00000451	LOP_INSERT_ROWS	NULL	LCX_CL...	sys.sysobjvalues.clst	0001
6	00000016:00000132:000e	0000:00000451	LOP_DELETE_SPLIT	NULL	LCX_CL...	sys.sysobjvalues.clst	0001
7	00000016:00000132:000f	0000:00000451	LOP_MODIFY_HEADER	NULL	LCX_HEAP	sys.sysobjvalues.clst	0001
8	00000016:00000132:0010	0000:00000451	LOP_MODIFY_HEADER	NULL	LCX_HEAP	sys.sysobjvalues.clst	0001
9	00000016:00000132:0011	0000:00000451	LOP_INSERT_ROWS	NULL	LCX_IND...	sys.sysobjvalues.clst	0001
10	00000016:00000132:0012	0000:00000451	LOP_COMMIT_XACT	NULL	LCX_NULL	NULL	NULL

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.
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

Data Transformation

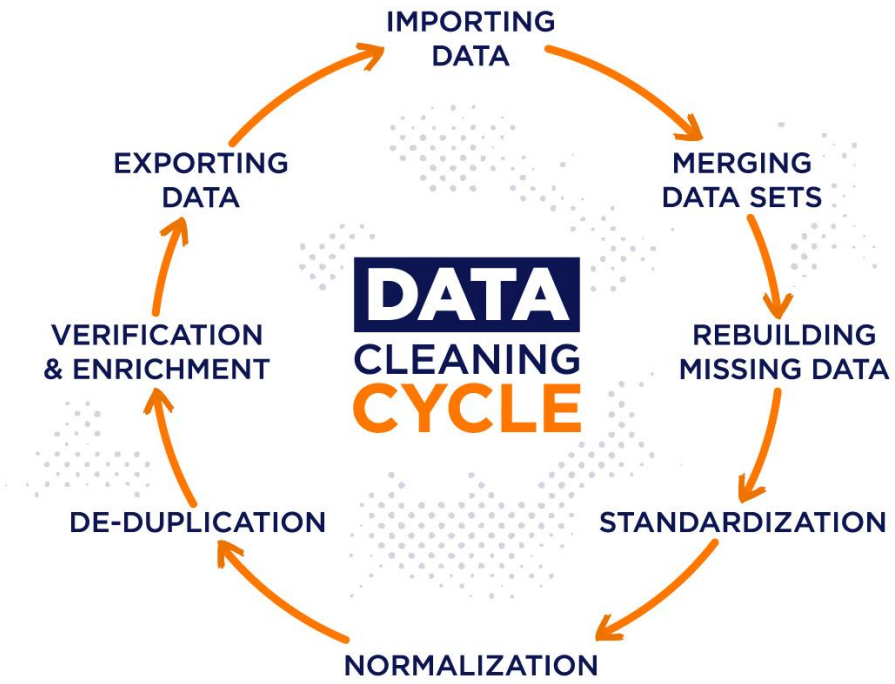
We have the RAW data...

Not good enough for the DW

- Quality
- Format

Before moving extracted data to DW

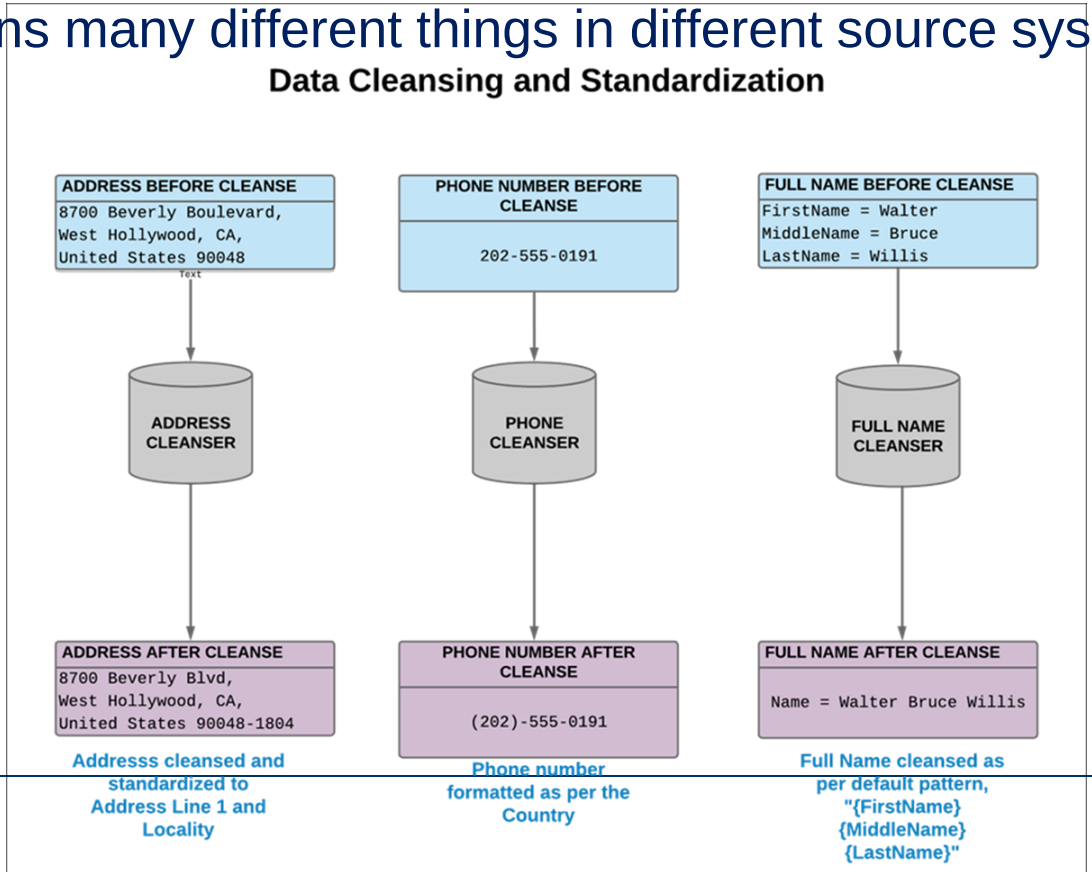
- **Data cleansing:**
  - Clean the extracted data from each source: correction of mis-spellings, including resolution of conflicts between state codes and post codes in the data sources, providing default values for missing data elements, or removing duplicated data



Source: https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/

Before moving extracted data to DW

- **Data standardisation:**
  - Standardise data types and fields lengths for same data elements retrieved from the various sources
  - Sematic 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



Source:
https://docs.reltio.com/datacleanse/cleanseoverview.html

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 (8 bit) to ASCII (7 bit)
- 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.)

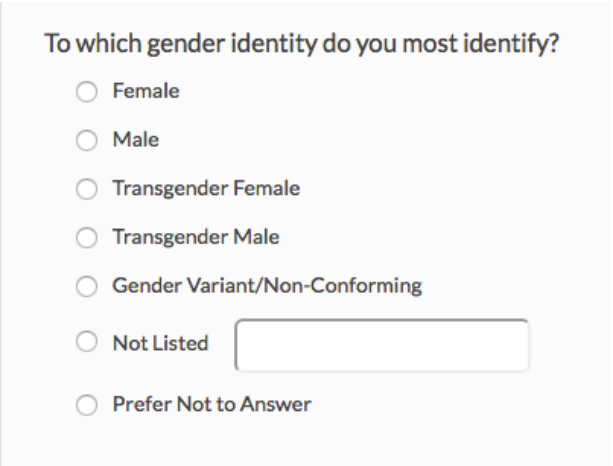
EBCDIC to ASCII Conversion Chart

EBCDIC is an 8-bit coding scheme. Valid hex values for an EBCDIC character are 00 to FF. The 16 rows in the chart below correspond to the first hex digit of an EBCDIC character (0 to F). The 16 columns correspond to the second hex digit of the character (0 to F). The contents of the cells shows the ASCII value (in hex) that corresponds to that EBCDIC character. For example, to convert the EBCDIC character C1 (which is the letter "A" in EBCDIC), look in the row labelled C, and in the column labelled 1. There you will find that the corresponding ASCII character is 41 (which is the letter "A" in ASCII).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	00	01	02	03	1A	09	1A	F1	1A	1A	1A	0B	0C	0D	0E	0F
1	10	11	12	13	1A	1A	08	1A	18	19	1A	1A	1C	1D	1E	1F
2	1A	1A	1A	1A	1A	0A	17	1B	1A	1A	1A	1A	1A	05	06	07
3	1A	1A	16	1A	1A	1A	1A	04	1A	1A	1A	1A	14	15	1A	1A
4	20	1A	1A	1A	1A	1A	1A	1A	1A	1A	5B	2E	3C	28	2B	21
5	26	1A	1A	1A	1A	1A	1A	1A	1A	1A	5D	24	2A	29	3B	5E
6	2D	2F	1A	1A	1A	1A	1A	1A	1A	1A	7C	2C	25	5F	3E	3F
7	1A	1A	1A	1A	1A	1A	1A	1A	1A	60	3A	23	40	27	3d	22
8	1A	61	62	63	64	65	66	67	68	69	1A	1A	1A	1A	1A	1A
9	1A	6A	6B	6C	6D	6E	6F	70	71	72	1A	1A	1A	1A	1A	1A
A	1A	7E	73	74	75	76	77	78	79	7A	1A	1A	1A	1A	1A	1A
B	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A	1A
C	7B	41	42	43	44	45	46	47	48	49	1A	1A	1A	1A	1A	1A
D	7D	4A	4B	4C	4D	4E	4F	50	51	52	1A	1A	1A	1A	1A	1A
E	5C	1A	53	54	55	56	57	58	59	5A	1A	1A	1A	1A	1A	1A
F	30	31	32	33	34	35	36	37	38	39	1A	1A	1A	1A	1A	1A

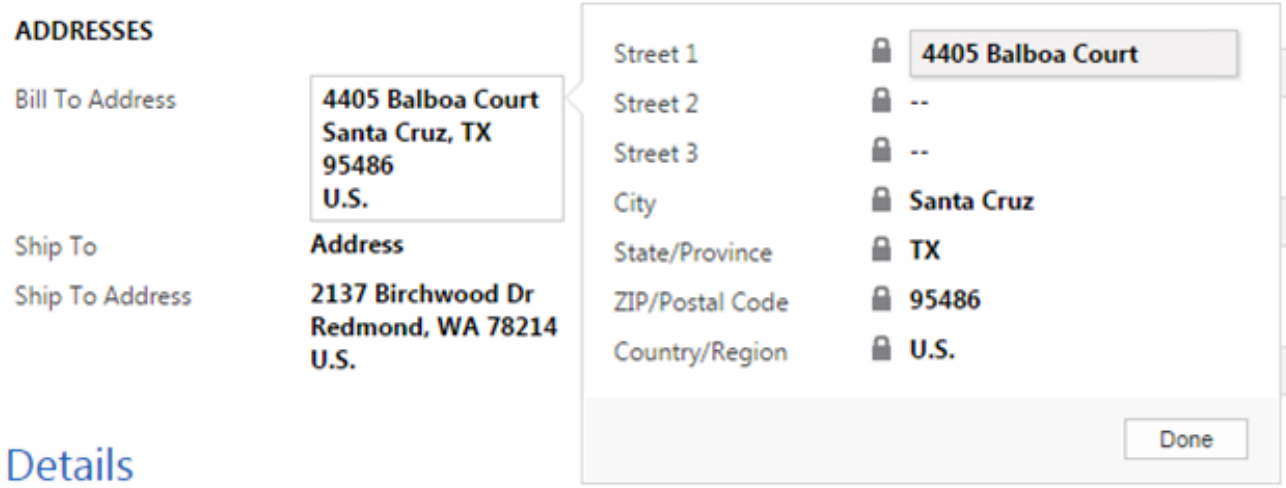
Major Transformation Tasks

- 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

- 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
- De-duplication -Get rid of the duplicate records that you find

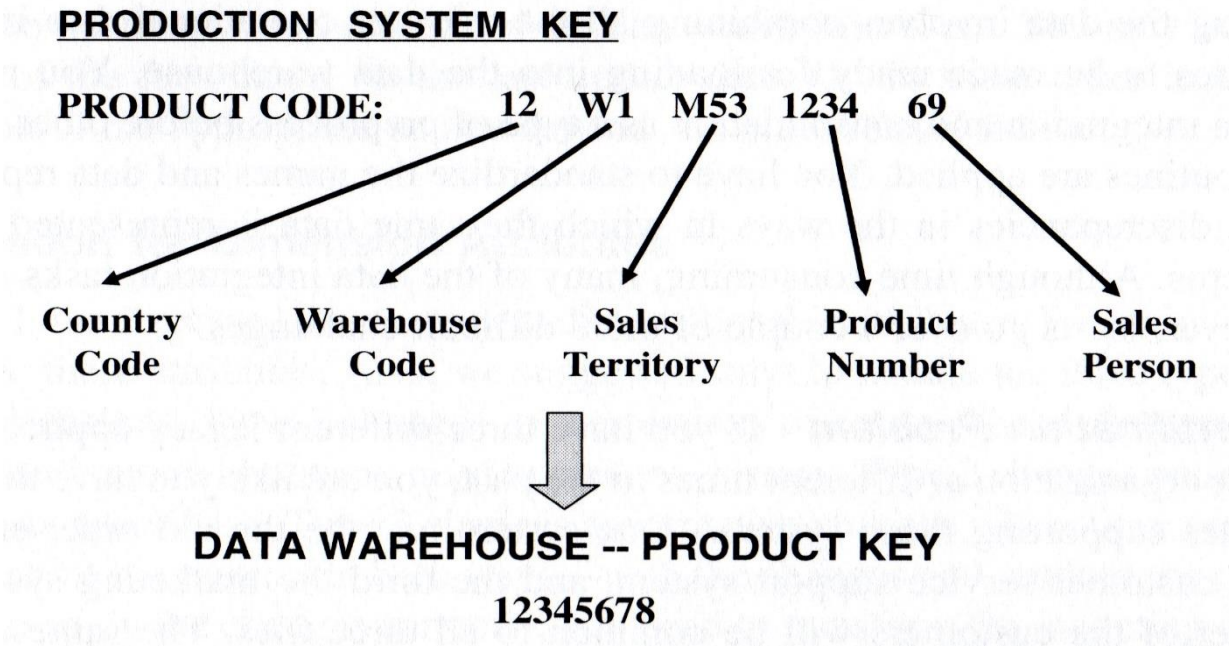


Major Transformation Tasks

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

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



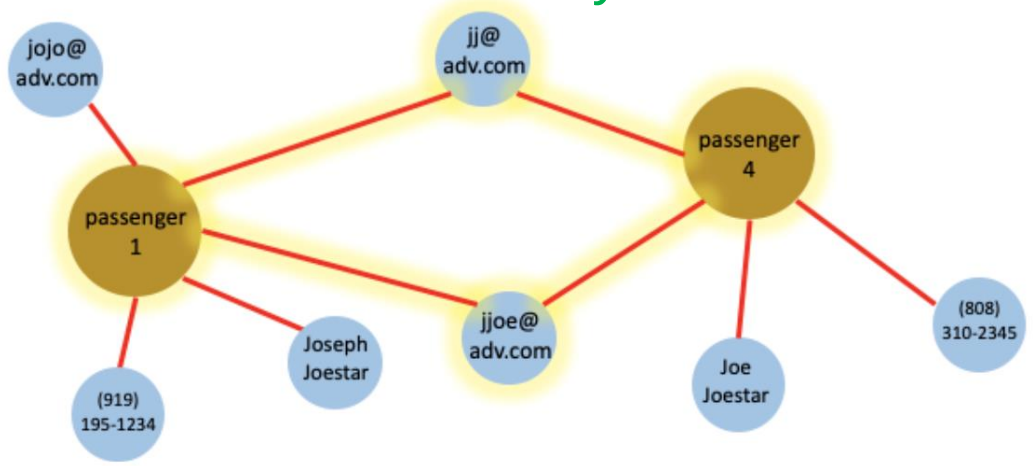
Source: Ponniah (2010)

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...



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!

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

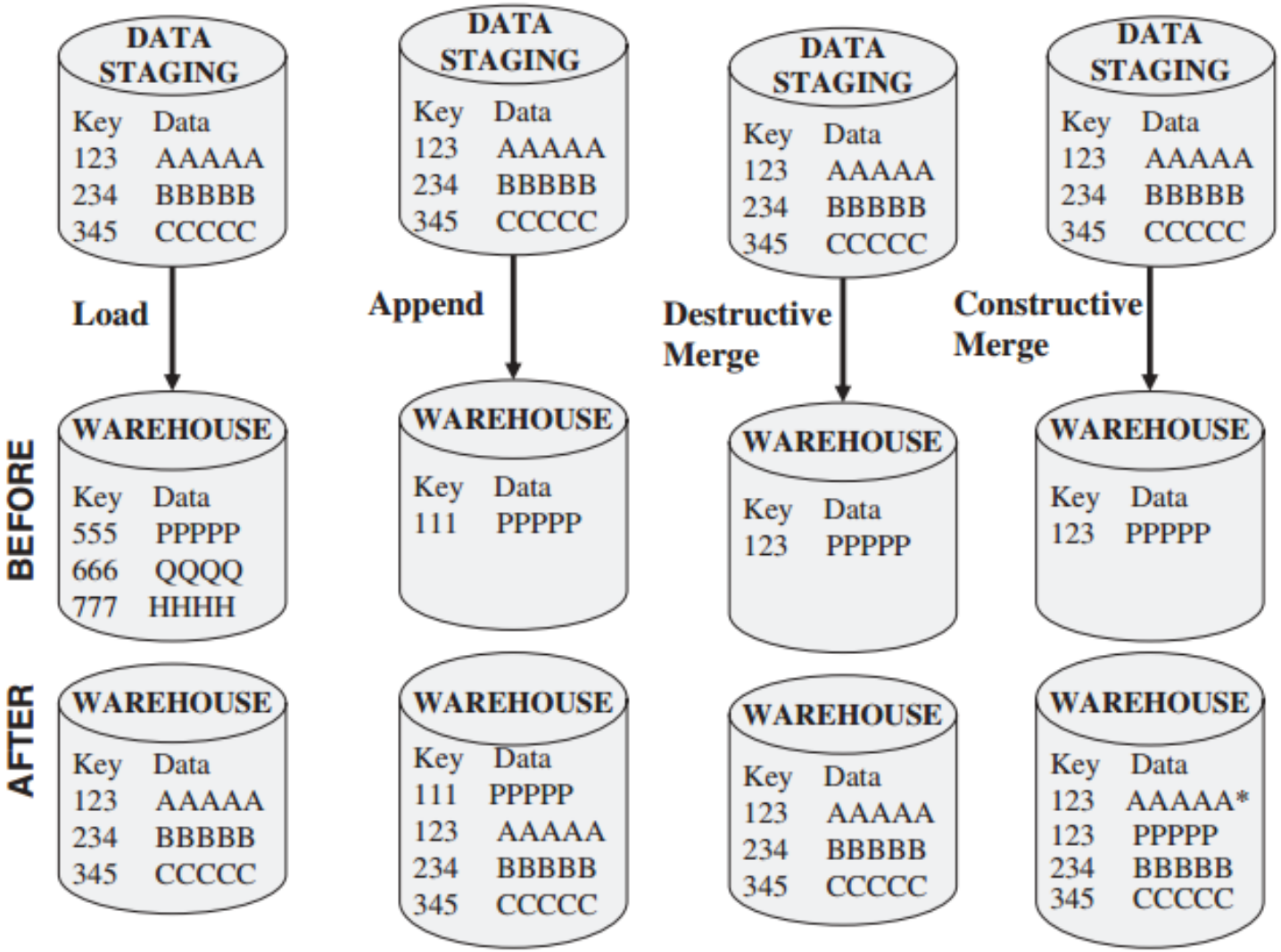


Figure 12-11 Modes of applying data.

Ponniah (2010) p304

ETL Summary

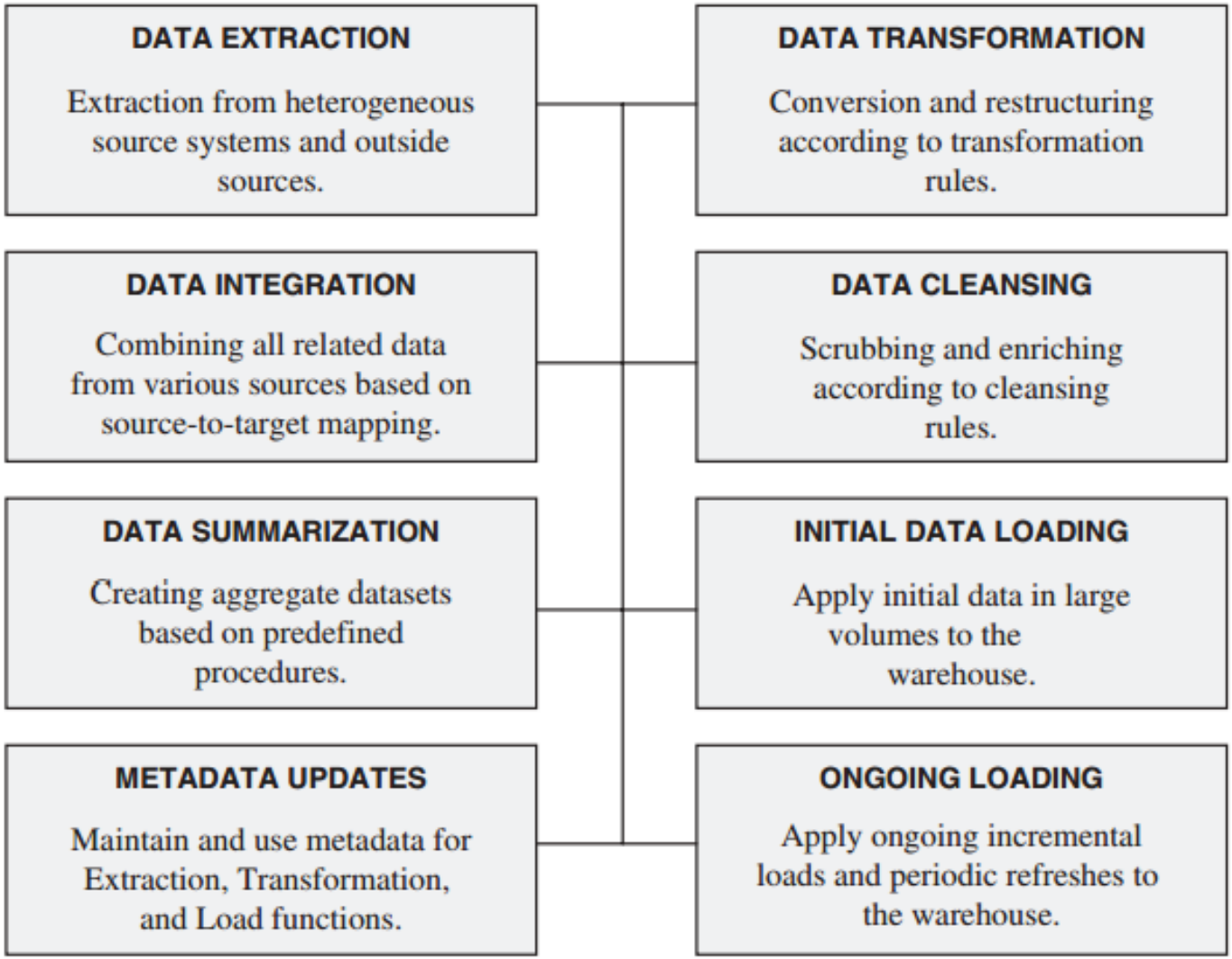


Figure 12-14 ETL summary.

Ponniah (2010) p310

ETL Tools

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
  - 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

Video ETL Process and tools

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

Power BI Tutes

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

Power BI Practical Assignment Discussion & Tutorial