Extract Transform Load (ETL) pp. 403++

Page 403:
"One of the most difficult tasks in data warehousing is **ETL** development"

"The process of loading a **dimension** table is made complex by the need to handle slowly changing dimensions"

"Loading a **fact table** presents its own fundamental processing challenges. The most significant of these involves foreign keys. As supplied by a source system, transactions will be accompanied by natural keys. The load process must replace these with surrogate keys."

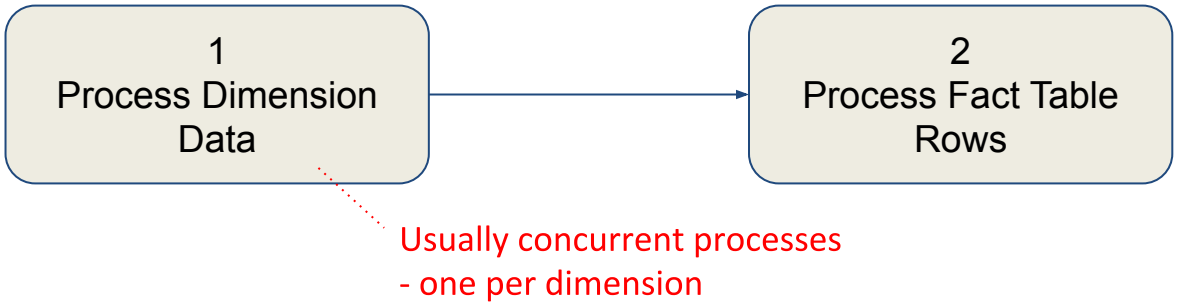
"The load process must also deal with **poor** source data and is often called upon to **clean it up**."

Process of loading a star involves two major activities:

- processing data for dimension tables
- processing data for fact tables

A fundamental dependency:

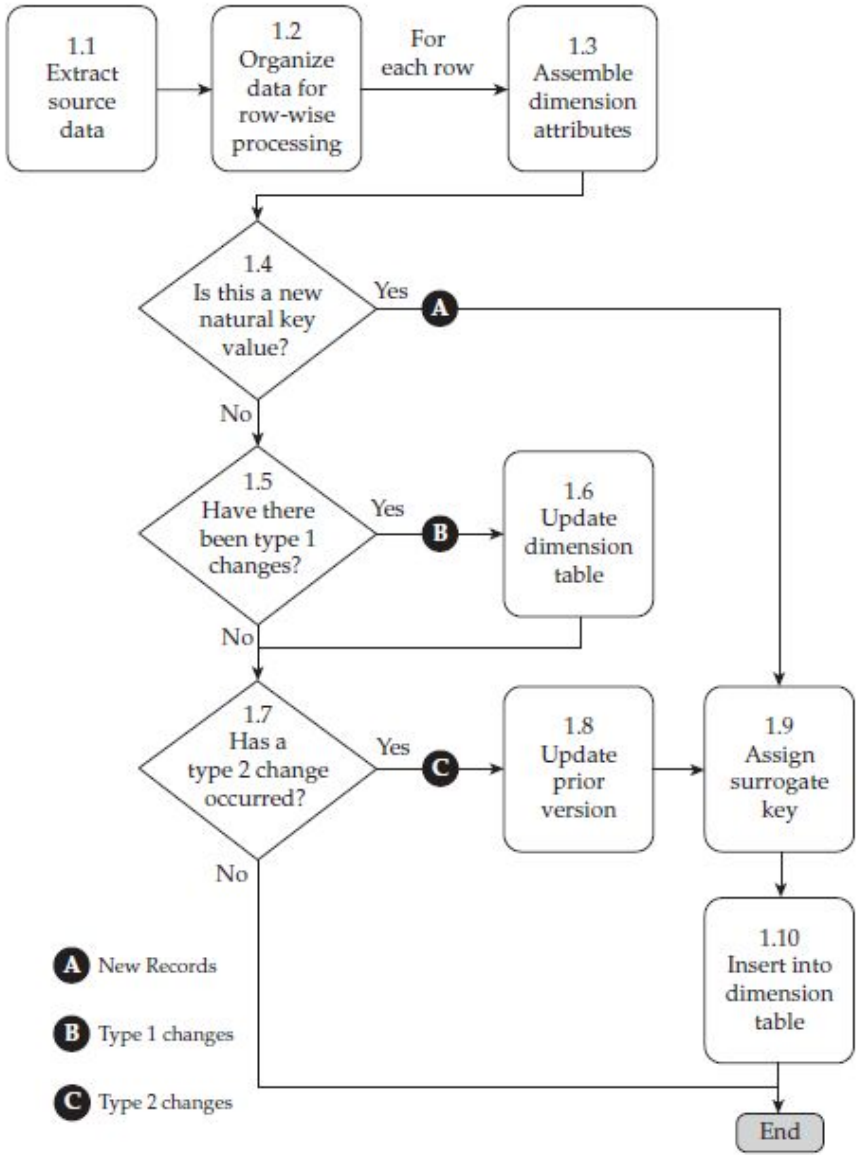
- each fact table row contains foreign key references to dimension table rows.
- hence, before a row can be inserted into a fact table, the appropriate dimension rows must exist.



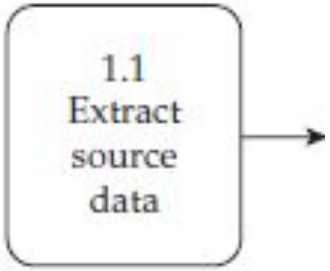
Loading a Dimension

pp. 406++

See figure 17-2 and following slides



Loading a Dimension

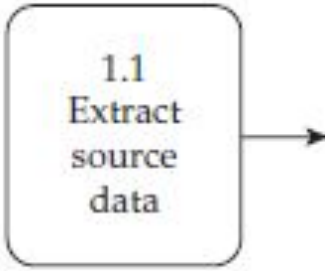


From text, p 408:

Source data may be acquired through a number of means

- programmatically into a file or through specialized utilities
- extracted directly via SQL or provided through an export file
- messaging system

Loading a Dimension



From text, page 418:

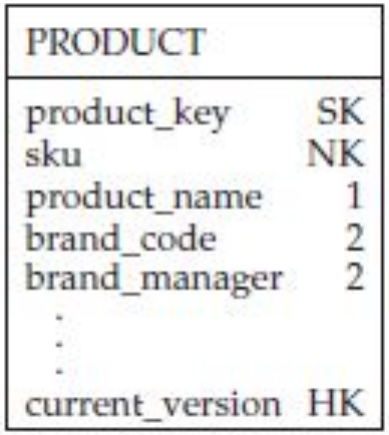
- New and changed data can be detected in many ways.
- Some source systems do so automatically, logging and time-stamping changes to key entities.

- For relational systems that do not do this, it may be possible to make use of triggers or stored procedures to generate a change log for key entities.

- Some databases offer native logging capabilities that can be exploited for the same effect.

Loading a Dimension

Figure 17.3



← An example dimension with surrogate key, natural key, type 1 and type 2 changes.

Legend:

SK Surrogate Key
NK Natural Key
1 Type 1 Change (Overwrite)
2 Type 2 Change (Insert new record)
HK Housekeeping Column (Internal use)

Existing content of PRODUCT Dimension Table:

SK	*NK*	*1*	*2*	*2*	*HK*
product_ key	sku	product_ name	brand_ code	brand_ manager	current_ version
22	A-0033	Flat Box 12"	110	Lei, Michelle	Current
344	A-1011	Packing Tape	221	Jones, Paul	Current
1001	B-3691	Twine Spool	501	Smith, Dei	Not Current
1201	B-3691	Twine Spool	501	Klein, Pete	Not Current
2700	B-3691	Twine Spool	702	Jones, Fay	Current
2711	Z-3320	Envelope A5	109	Harvey, Ned	Current

← Existing data

Loading a Dimension

Figure 17.3 Source changes

	sku	product_ name	brand_ code	brand_ manager
1	S-8472	Envelope B4	109	Harvey, Ned
2	B-3691	TwineSpool 12	702	Jones, Fay
3	A-0033	Flat Box 12"	110	Cook, Dan
4	A-1011	Clear Packing Tape	221	Davis, Jon
5	Z-3320	Envelope A5	109	Harvey, Ned

A new record

Type 1 change affecting 3 dimension rows

Type 2 change

Type 1 and type 2 changes

An unchanged record

Brand manager has changed (type 2) which results in one new row

Product name changes (type 1) which results in changes to 3 related dimension rows

A new row in the dimension must be created

Loading a Fact Table p. 413

The fact table load is governed by the following requirements

- Extract data from the source systems
- Compute the facts
- Aggregate facts to match the grain of the fact table
- Obtain surrogate keys for each of the dimensions
- Load fact table records into the warehouse