SOFTWARE RE-ENGINEERING

Unit 5

DATA RE-ENGINEERING

- So far, most of the discussion of software evolution has focused on the problems of changing programs and software systems.
- However, in many cases, there are associated problems of data evolution.
- The storage, organisation and format of the data processed by legacy programs may have to evolve to reflect changes to the software.
- The process of analysing and re-organising the data structures and, sometimes, the data values in a system to make it more understandable is called *data re-engineering*

- *Data degradation* Over time, the quality of data tends to decline.
  - errors, duplicate values,

- *Inherent limits that are built into the program:* When originally designed, developers of many programs included built-in constraints on the amount of data which could be processed. However, programs are now often required to process much more data than was originally envisaged by their developers.

- *Architectural evolution:* If a centralised system is migrated to a distributed architecture it is essential that the core of that architecture should be a data management system that can be accessed from remote clients. This may require a large data re-engineering effort to move data from separate files into the server database management system.

- Data re-engineering may be required to remove the limitations.
- 1- Data cleanup: The data records and values are analysed to improve their quality.

  -Duplicates are removed.

  -Redundant information is deleted.

  -Consistent format applied to all records.

  -This should not normally require any associated program changes.

- 2- Data extension:

  -Remove limits on the data processing.

  -This may require changes to programs to increase field lengths.

  -Modify upper limits on the tables, etc.

- 3-Data migration:

  -Data is moved into the control of a modern DBMS.

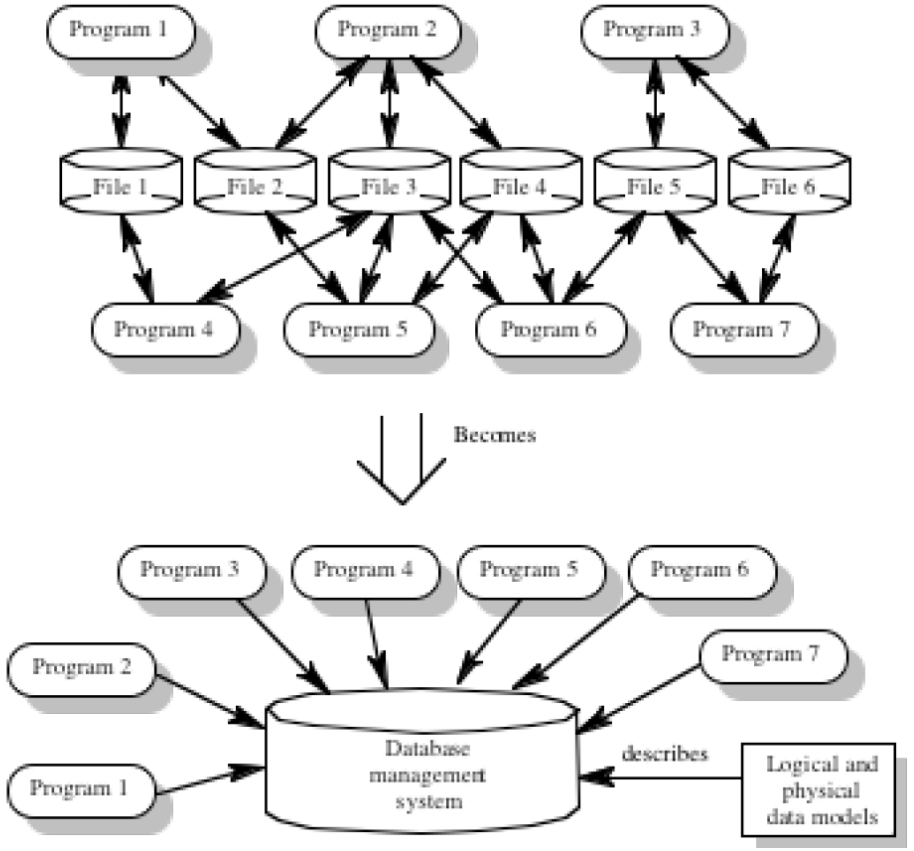
  -The data may be stored in separate files or may be managed by an older type of DBMS.

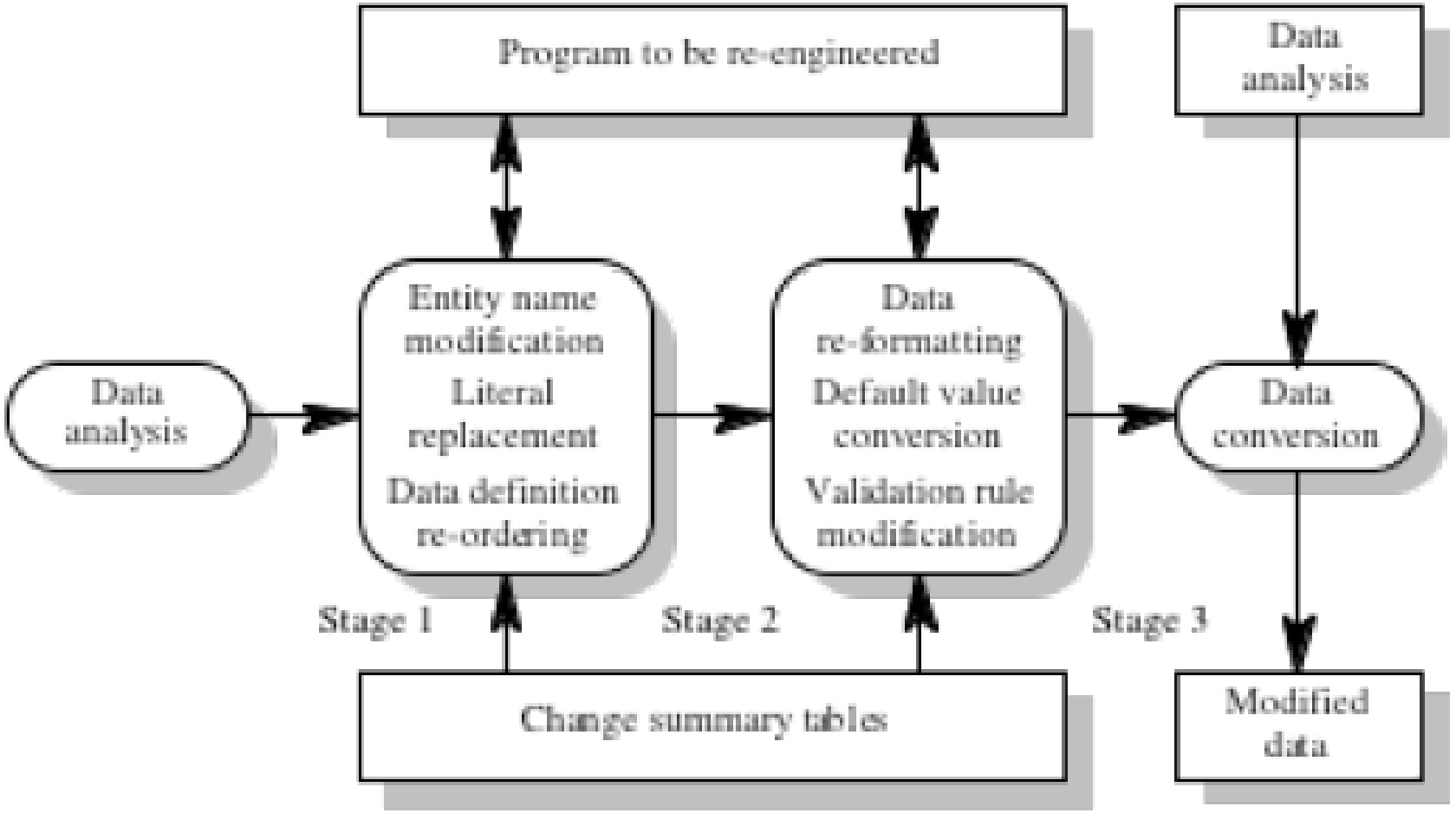
SOME OF THE PROBLEMS WITH DATA

- *Data naming problems* Names may be cryptic and difficult to understand.
  - The same name may be used in different programs to mean different things.
- This is a problem when field lengths in records are explicitly assigned in the program.
- Records representing the same entity may be organised differently in different programs.
- *Hard-coded literals* Literal (absolute) values.
- *No data dictionary*

SOME OF THE PROBLEMS WITH DATA

- Inconsistent data definitions, data values may also be stored in an inconsistent way.
- **Inconsistent Defaults**: *Different programs assign different default values to the same logical data items.*
- **Inconsistent Units.** *(lbs vs kg)*
- **Inconsistent Validation rules:** *Data written by one program may be rejected by another.*
- **Inconsistent representation semantics.**
- **Inconsistent handling of negative values**




Program to be re-engineered
Data analysis
Entity name modification
Literal replacement
Data definition re-ordering
Data re-formatting
Default value conversion
Validation rule modification
Data analysis
Data conversion
Stage 1
Stage 2
Stage 3
Change summary tables
Modified data