BASICS OF DATA INTEGRATION

Imagine that you are

1. Sales manager for a large retail organization.
2. Productivity manager
3. Petroleum minister
4. Supply chain manager
5. Customer retention manager
6. Customer care manager
7. Education minister
8. Chief of hospital
9. HR manager
10. Restaurant owner
11. LIC branch head
12. Printing press owner
13. TV channel owner
14. Film maker
15. Owner of Airline

and have
BI tool with you.
What you will
do to improve
existing s/m.

NEED for DATA WAREHOUSE

- Lack of Information sharing(Example)
- Lack of information credibility(EXAMPLE)
- Reports take longer time to be prepared(EXAMPLE)
- Queries that require historical data.(EXAMPLE)

Williams Definition of DW

- A data warehouse is a subject oriented, integrated, time variant and non-volatile collection of data in support of management’s decision making process.

Ralph Kimbell’s definition of DW

- A data warehouse is made up of all the data marts in an enterprise.

WHAT IS

- DATA MART
- Operational data store

Goals of data warehouse

- Information accessibility
- Information credibility
- Flexible to change
- Support for more fact based decision making
- Support for the data security
- Information consistency

What constitutes a data warehouse

- Operational source systems
- Data staging area
- Data presentation area

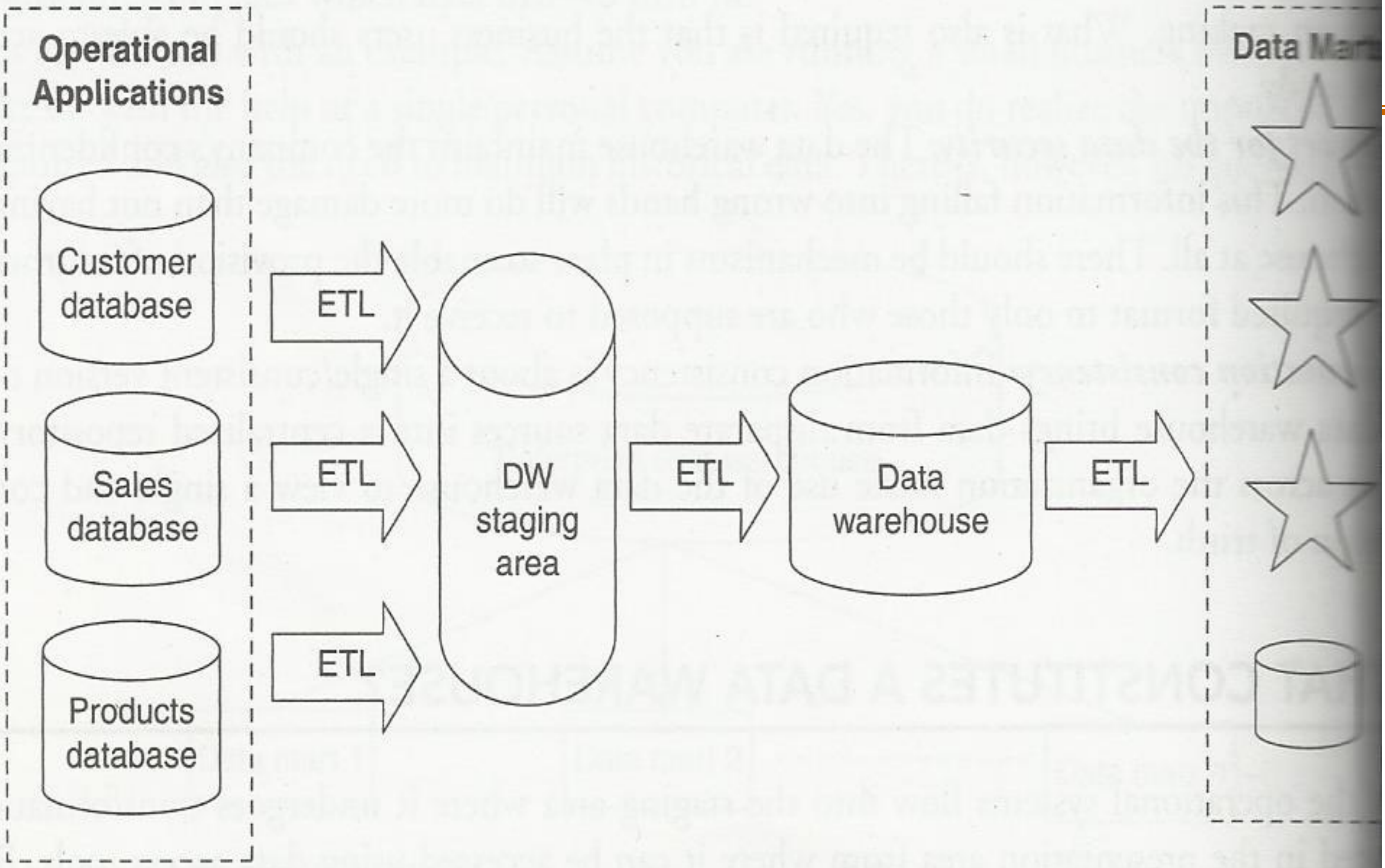

Figure 6.4 Operational Data Sources → Data Warehouse → Data Marts.

ETL

- DATA MAPPING
- DATA STAGING
  - Data extraction
  - Data transformation
  - Data loading

What is Data Integration

- Schema Integration
- Instance Integration

Common Approaches of Data Integration

- Federated databases
- Data warehouses

Federated database	Data Warehouse
Preferred when databases are geographically decentralized	Preferred when source of information can be taken from one location
Data would be present in various servers	The entire data warehouse would be present in one server
Requires high speed network connection	Requires no network connection
It is easier to create as compared to data warehouse	Creation is difficult
Requires no creation of new database	Data warehouse must be created from scratch
Requires network experts to set up the network connection	Requires database expert as data steward.

Data Integration Technologies

- Data interchange
- Object Brokering
- Modeling techniques
  - – ER MODELING(Removes data redundancy but creates lots of table)
  - – Dimensional modeling

Problems posed by ER Modeling

- End users find it difficult to comprehend and traverse through the ER model.
- Not too many software exist which can query a general ER model.
- ER modeling cannot be used for data warehousing since it degrades the performance and cannot answer ad hoc queries.

Steps to drawing an ER model

- Identify entities
- Identify relationships between various entities
- Identify the key attribute
- Identify other relevant attributes
- Draw the ER diagram
- Review the ER diagram with the business users and get their sign off.

ER to Dimensional

- Separate out the various business processes and represent each as a separate dimensional model.
- Identify all many to many relationship in the ER diagram, containing additive and non key attributes, and construct them into fact tables.
- De-normalize all the remaining tables into single part key tables
- In the cases where same dimension table connects to more than one fact table, the dimensional table is represented in each schema.

Why data quality matters?

- CD company promotion
- Cell company’s cross sell opportunities
- Mislabeled shipment
- Phone number column

Definition of data integrity

- Data integrity reflects the degree to which the attributes associated with certain entity accurately describe the occurrence of that entity.

Examples of data integrity

- Primary key: A column or a collection of columns designated as primary key imposes the unique and not null constraint.

Examples of data integrity

- A column designated as the foreign key column means it can only have values that are present in the primary key column of the same or different table that it refers to. A foreign key column can have a null or duplicate value.

Examples of data integrity

- Not null : not null constraint on the column means that the column must be given a value. It cannot have null value.

Check constraint

- A check constraint allows imposing a business rule on a column or a collection of column.

Definition of data quality

- Data quality is measured with reference to appropriateness for purpose as defined by the business users of data and conformance to enterprise data quality standards as formulated by system architects and administrators.

Characteristics of data quality

- Accuracy or correctness
- Consistency
- Completeness
- Timeliness
- Metadata

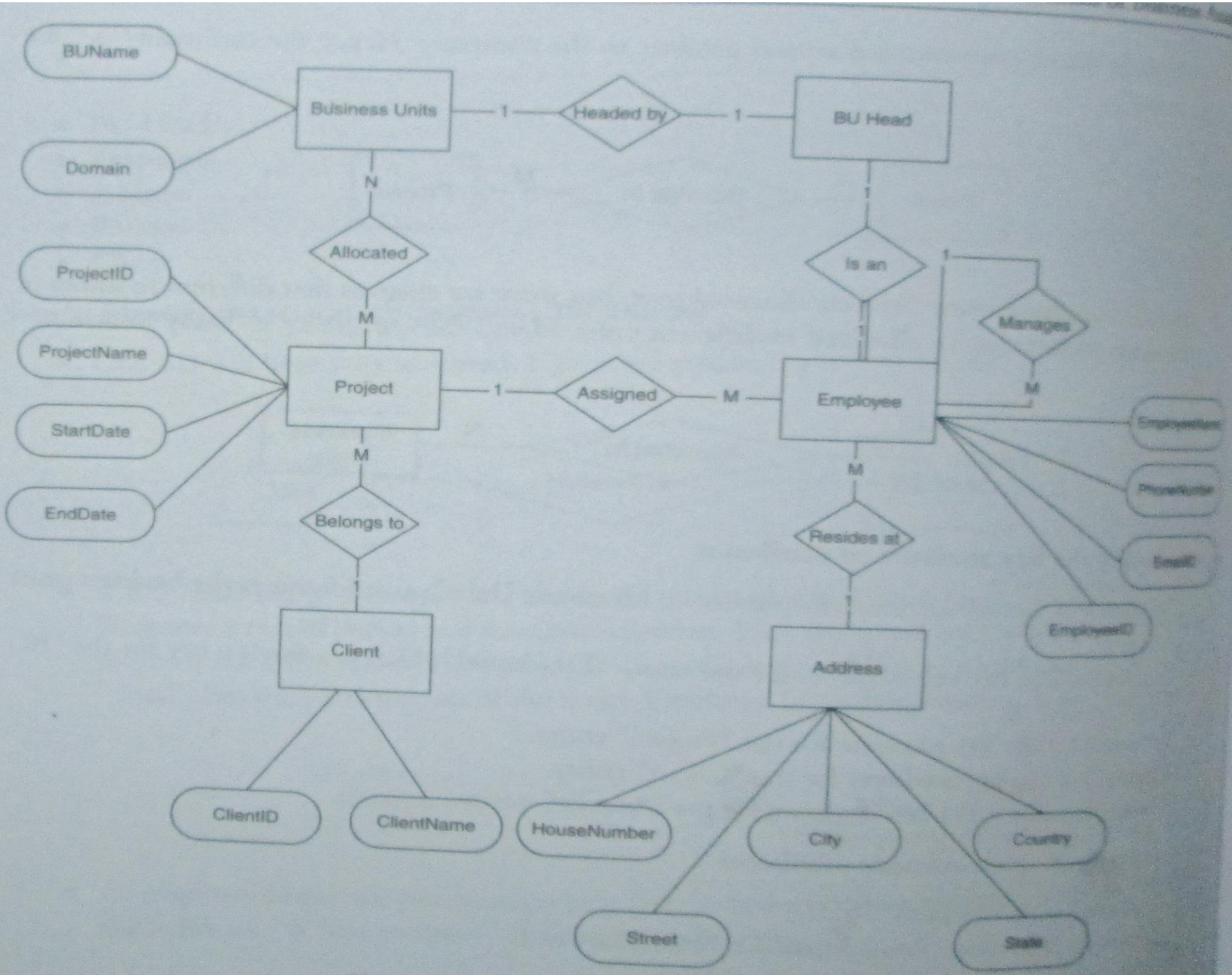

Figure 7.4 ER model for "InfoMechanists".

Data Profiling

- Data profiling is the process of statistically examining and analyzing the content in a data source.
- Data profiling is the initial step that defines the data quality rules and lays the basic groundwork for the task of data quality profiling.

- Data profiling helps us make a thorough assessment of data quality.
- It assists the discovery of anomalies in data.
- It helps us know whether the existing data can be applied to other areas or purposes.
- It helps us to understand content, structure, relationships about the data in the data source which is to be analyzed.

- It helps us in understanding the various challenges/issues we may face in a database project much before the actual work begins.
- It helps add quality to data by converting from the format in which it is stored or categorizing it.
- It helps assess the risks associated with integrating data with other applications.
- It is also used to assess and validate metadata.

- Data profiling can be either –
- Data quality profiling – It refers to analyzing the data from a data source or database against certain specified business rules or requirements.
- The analysis can be represented as –
- Summaries – Counts and percentages that give information on the completeness of data sets, the uniqueness of the column, problem distribution in a data set, etc.

- Details – Involves lists that contain information about missing data records or data problems in individual records, etc.
- Database profiling – It refers to the procedure of analysis of a database, with respect to its schema structure, relationships between tables, columns used, data-type of columns, keys of the tables, etc.

When to conduct data profiling

- Requirement gathering phase
- Just before the dimensional modeling process
- During ETL package design.

How to conduct data profiling

- Data quality
- Null values
- Candidate keys
- Primary key
- Empty string values
- String length
- Numeric length and type
- Identification of cardinality
- Data format

References

- David Loshin, “Business Intelligence”, Morgan Kaufmann Publishers, 2003
- Mike Biere, “Business Intelligence for the Enterprise”, 2nd edition, IBM Press,2003.
- R N Prasad, Seema Acharya, “Fundamentals of Business Analytics”, Wiley India, 2011