Data Warehousing Architectures

Ten factors that potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **Data integration**

  Integration that comprises three major processes: data access, data federation, and change capture

- **Enterprise application integration (EAI)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse
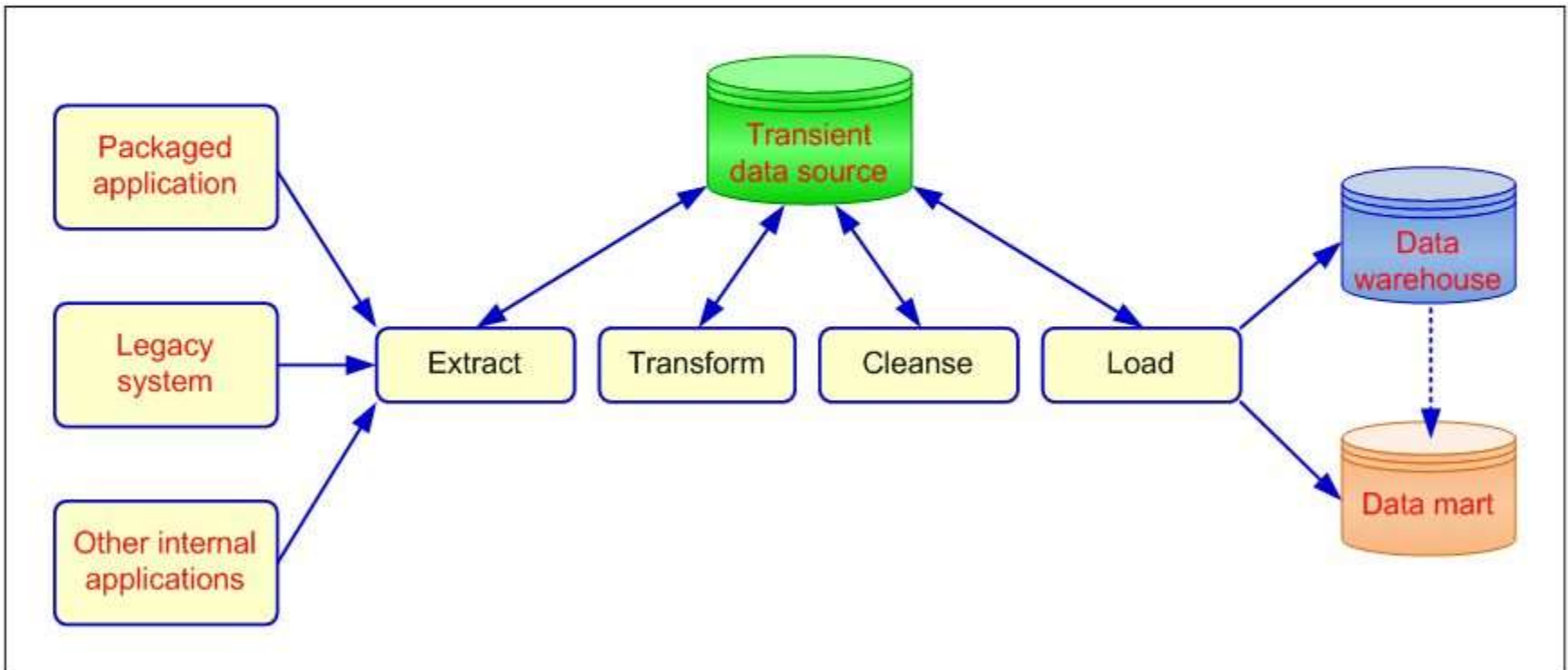
- **Enterprise information integration (EII)**

  An evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases

Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, transformation, and load (ETL)

ETL

- Issues affecting the purchase of ETL tool
    - Data transformation tools are expensive
    - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
    - Ability to read from and write to an unlimited number of data sources/architectures
    - Automatic capturing and delivery of metadata
    - A history of conforming to open standards
    - An easy-to-use interface for the developer and the functional user

Data Warehouse Development

- Data warehouse development approaches
  - Inmon Model: EDW approach (top-down)
  - Kimball Model: Data mart approach (bottom-up)
  - Which model is best?
    - There is no one-size-fits-all strategy to DW
  - One alternative is the hosted warehouse
- Data warehouse structure:
  - The Star Schema vs. Relational
- Real-time data warehousing?

Hosted Data Warehouses

- Benefits:
    - Requires minimal investment in infrastructure
    - Frees up capacity on in-house systems
    - Frees up cash flow
    - Makes powerful solutions affordable
    - Enables powerful solutions that provide for growth
    - Offers better quality equipment and software
    - Provides faster connections
    - Enables users to access data remotely
    - Allows a company to focus on core business
    - Meets storage needs for large volumes of data

Representation of Data in DW

- **Dimensional Modeling** – a retrieval-based system that supports high-volume query access
- **Star schema** – the most commonly used and the simplest style of dimensional modeling
    - Contain a fact table surrounded by and connected to several dimension tables
    - Fact table contains the descriptive attributes (numerical values) needed to perform decision analysis and query reporting
    - Dimension tables contain classification and aggregation information about the values in the fact table
- **Snowflakes schema** – an extension of star schema where the diagram resembles a snowflake in shape

Multidimensionality

- ## Multidimensionality

    The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)
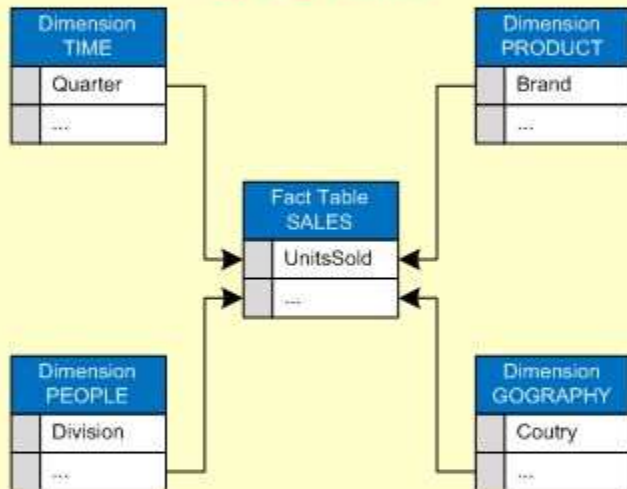
- ## Multidimensional presentation

    - **Dimensions:** products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry

    - **Measures:** money, sales volume, head count, inventory profit, actual versus forecast

    - **Time:** daily, weekly, monthly, quarterly, or yearly

Star vs Snowflake Schema