Data Science & Analytics

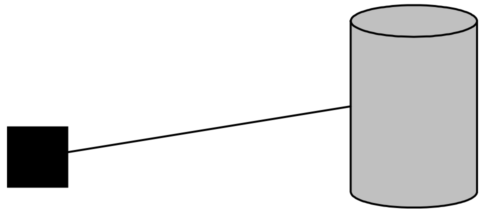
Hiren Deliwala & Dr. Jian ZHANG

Data Warehousing & Business Intelligence

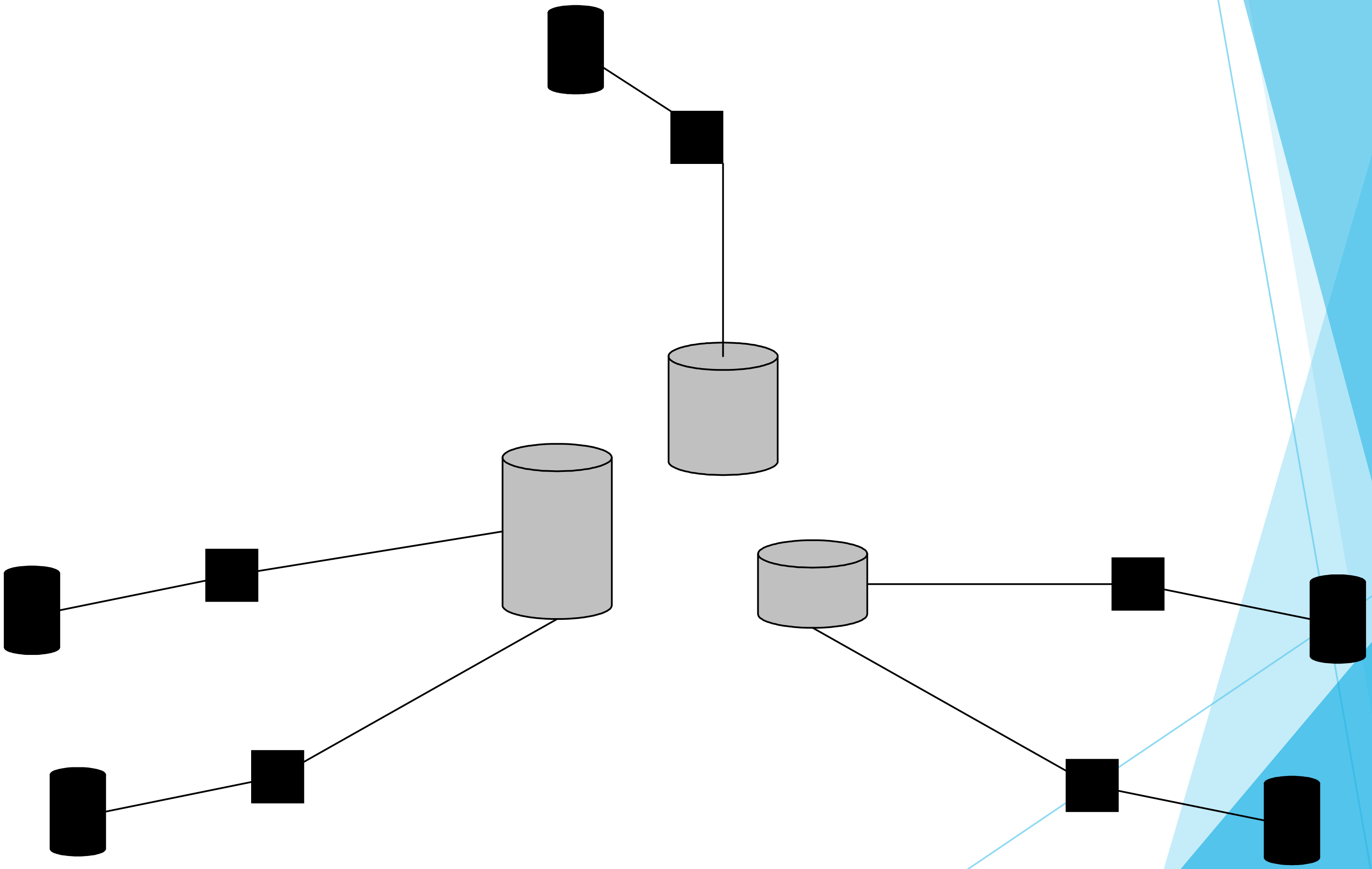
Agenda

- Data Warehousing
  - Introduction
  - OLTP Versus OLAP
  - Data Warehouse System – Components
  - Data Warehouse Design

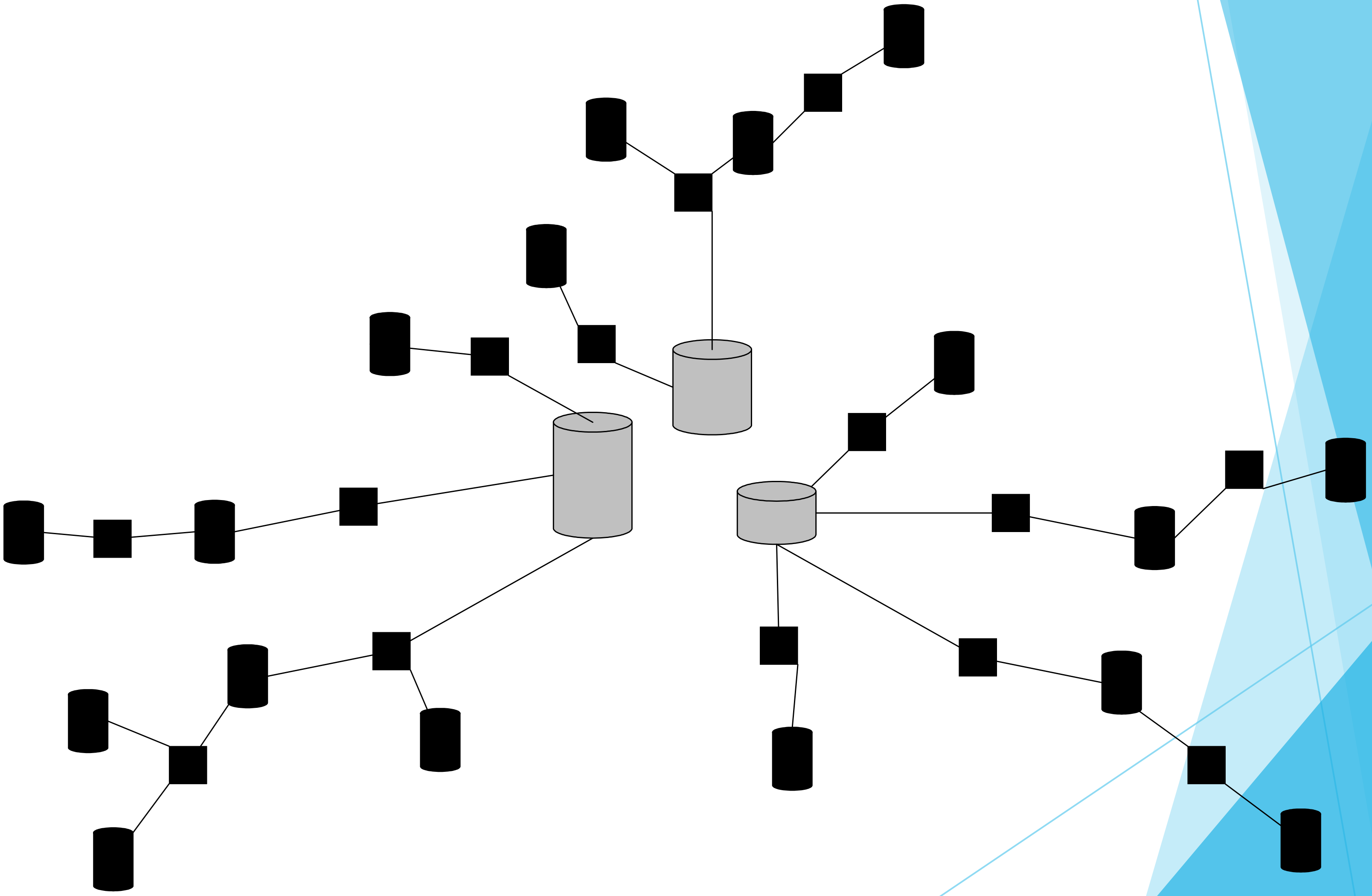
In the Beginning, life was simple…



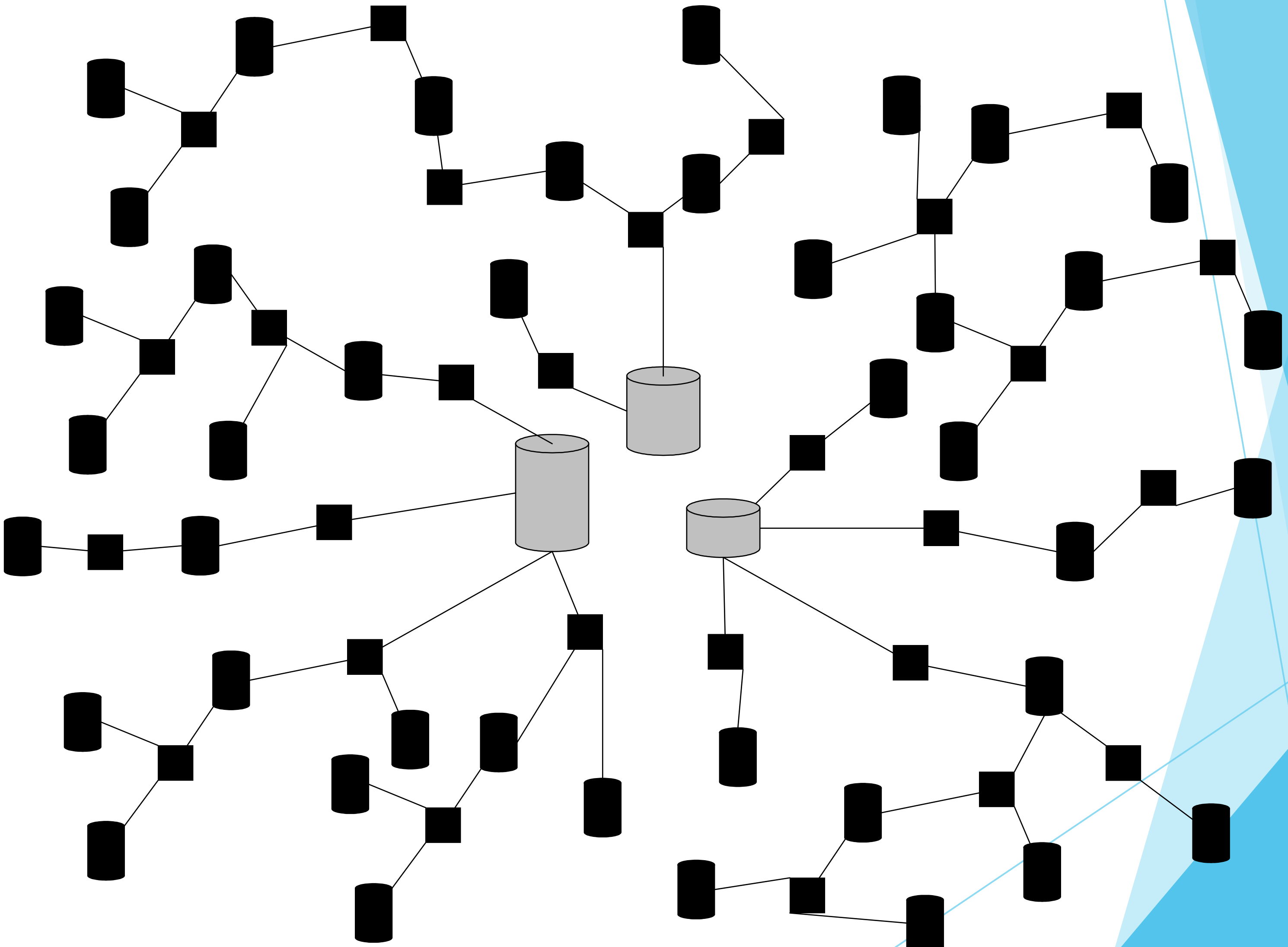
But…



Our information needs…



Kept growing. (The Spider web)



SOURCE: William H. Inmon

Why Separate Data Warehouse?

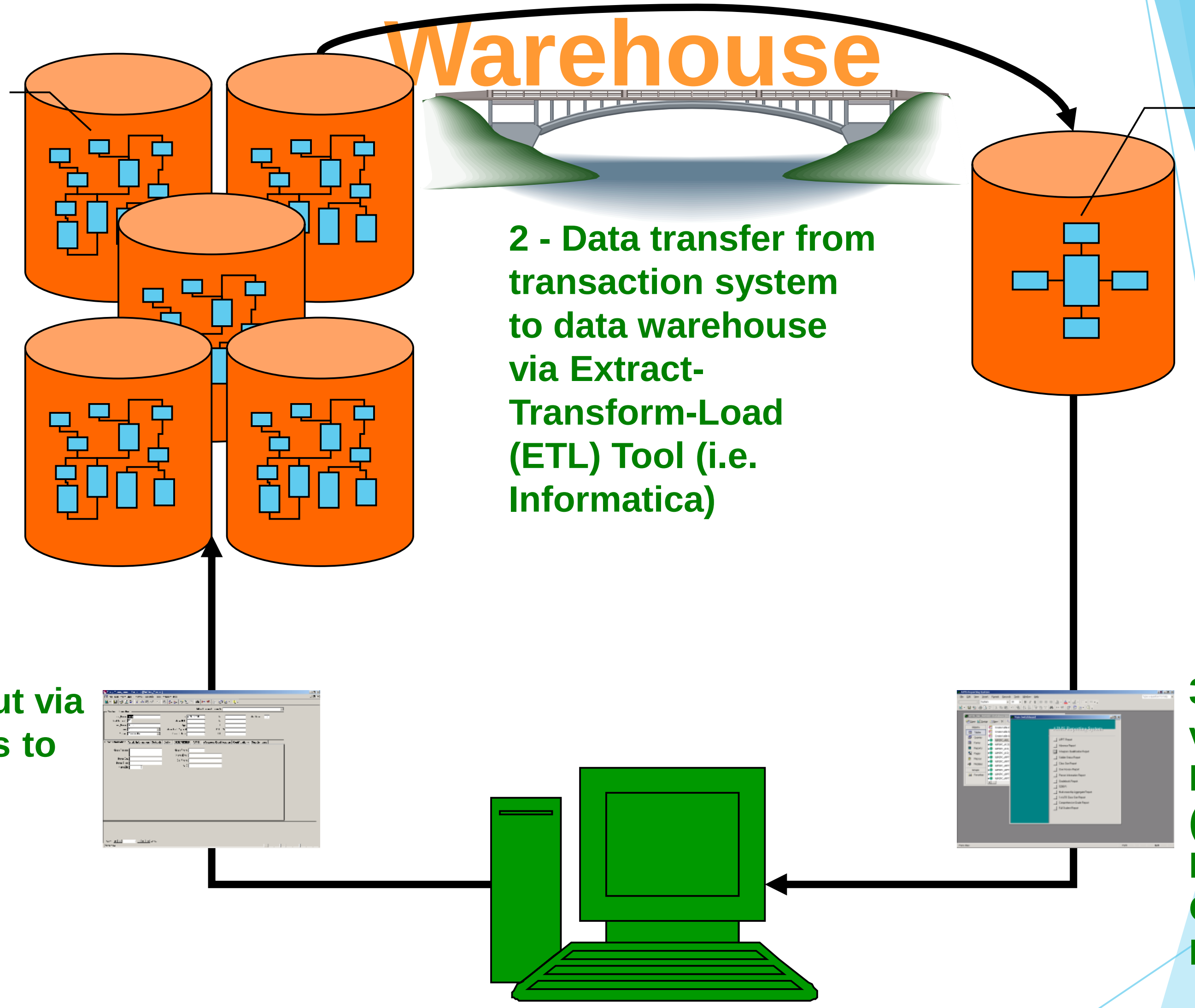
- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation.
- Different functions and different data:
  - missing data: Decision support(DS) requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

Data Warehousing (DW)

- Definition
  - A subject-oriented, integrated & non-volatile database updated on a typically rhythmic cycle from an enterprise's various transaction databases.
- Purpose
  - Accumulate data from disparate data sources for querying purposes
  - Separate reporting and analysis operations from transaction systems to maximize the performance of both

Commonly very large repositories that house historical data

Data Flow from Transaction to Warehouse



Complex Structure Necessary for Accurate Transactions

2 - Data transfer from transaction system to data warehouse via Extract-Transform-Load (ETL) Tool (i.e. Informatica)

Simplified Structure Necessary for Fast, Powerful Reports

1 - Data Input via Applications to transaction databases

3 - Data Output via Business Intelligence Tool (i.e. Cognos, Business Objects, Hyperion)

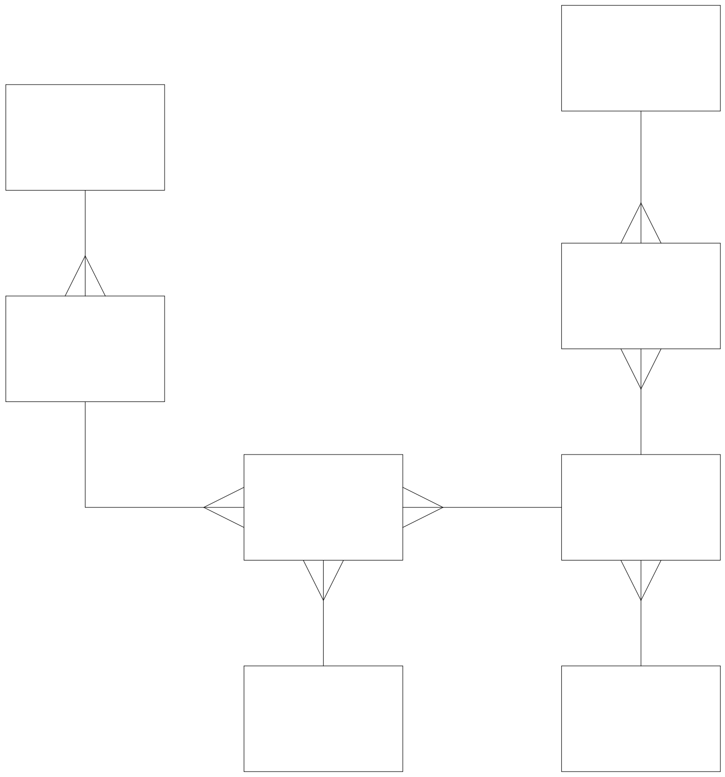
Separation of Transactions and Reporting Improves Performance and Enhances Capabilities

OLTP Versus OLAP

	OLAP	OLTP
Definition	Online Analytical Processing System	Online Transaction Processing System
Users	Analysts, Knowledge Workers	Line of business users
No. of Users	100s	1000s
Access	Read-only access; Lot of scans	Read/Write access
Records accessed	Millions GBs or TBs	Thousands MBs or GBs
DB design	Denormalized Data Model	Normalized Data Model
History	Maintains historical data for an extended period of time	Maintains recent history
Optimization	Optimized for queries against large data sets	Optimized for transaction processing
Data	Historical, summarized, multidimensional, integrated, consolidated	Current, up-to-date, detailed, flat relational isolated
Query Types	Adhoc-queries; Complex	Predefined queries; Repetitive; short, simple transactions

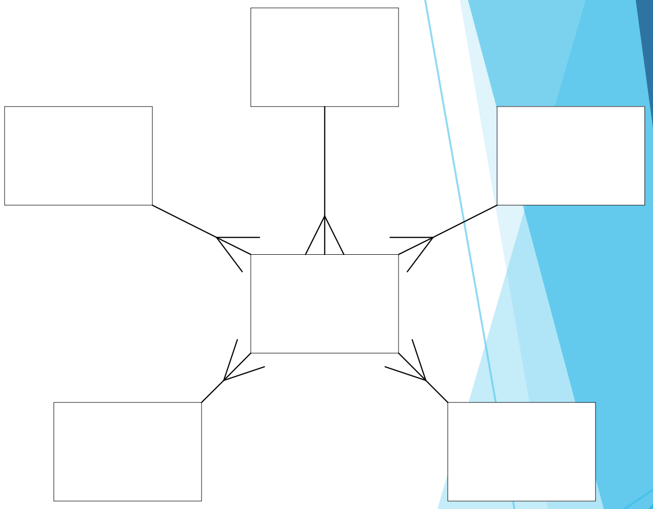
Design Differences

Operational System



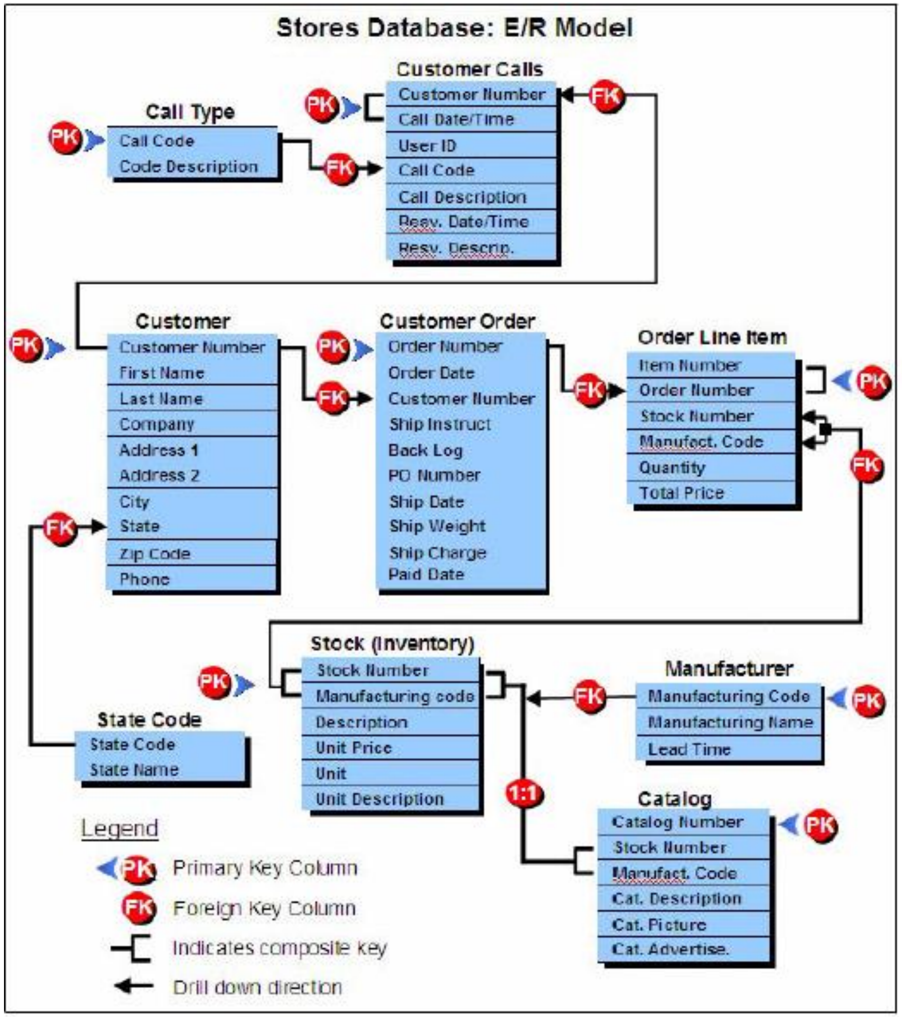
ER Diagram

Data Warehouse



Star Schema

Example OLTP Model



Stores Database: E/R Model

Customer Calls
Customer Number
Call Date/Time
User ID
Call Code
Call Description
Resv. Date/Time
Resv. Descrip.

Call Type
Call Code
Code Description

Customer
Customer Number
First Name
Last Name
Company
Address 1
Address 2
City
State
Zip Code
Phone

Customer Order
Order Number
Order Date
Customer Number
Ship Instruct
Back Log
PO Number
Ship Date
Ship Weight
Ship Charge
Paid Date

Order Line Item
Item Number
Order Number
Stock Number
Manufact. Code
Quantity
Total Price

Stock (Inventory)
Stock Number
Manufacturing code
Description
Unit Price
Unit
Unit Description

Manufacturer
Manufacturing Code
Manufacturing Name
Lead Time

State Code
State Code
State Name

Catalog
Catalog Number
Stock Number
Manufact. Code
Cat. Description
Cat. Picture
Cat. Advertise.

Legend
PK Primary Key Column
FK Foreign Key Column
Indicates composite key
Drill down direction

Figure 1-2 Sample OLTP schema

Source: IBM Redbook: Dimensional Modeling: In a Business Intelligence Environment.

Example OLAP Model

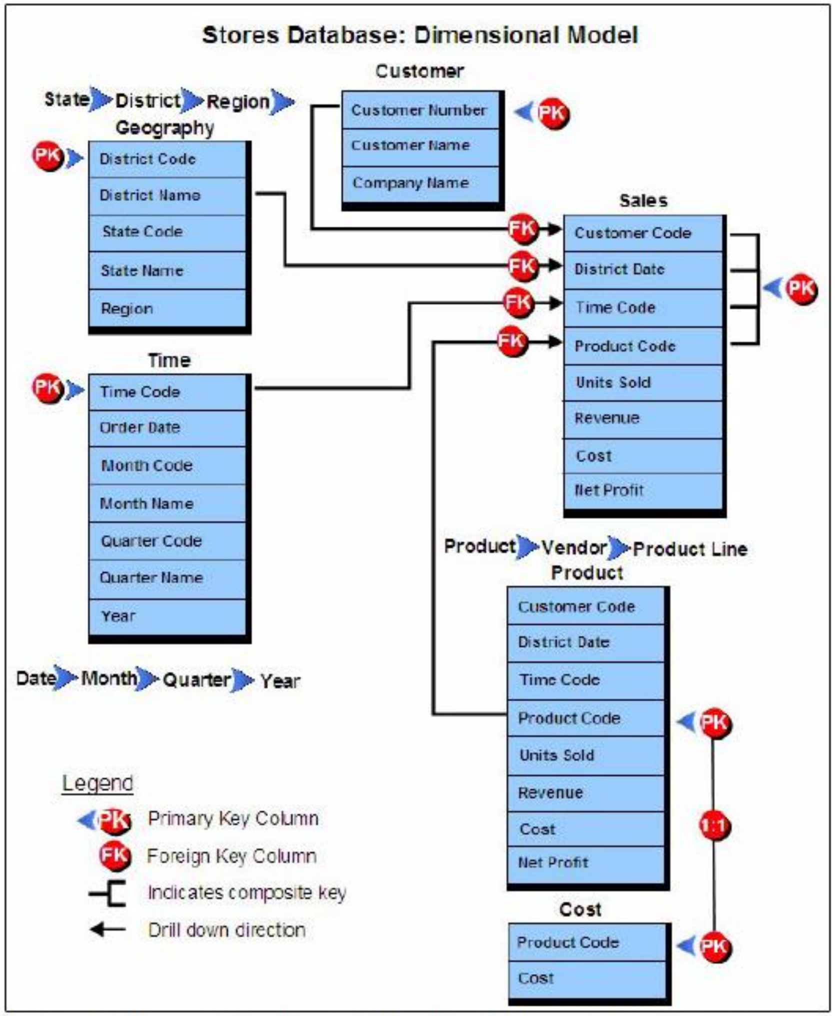
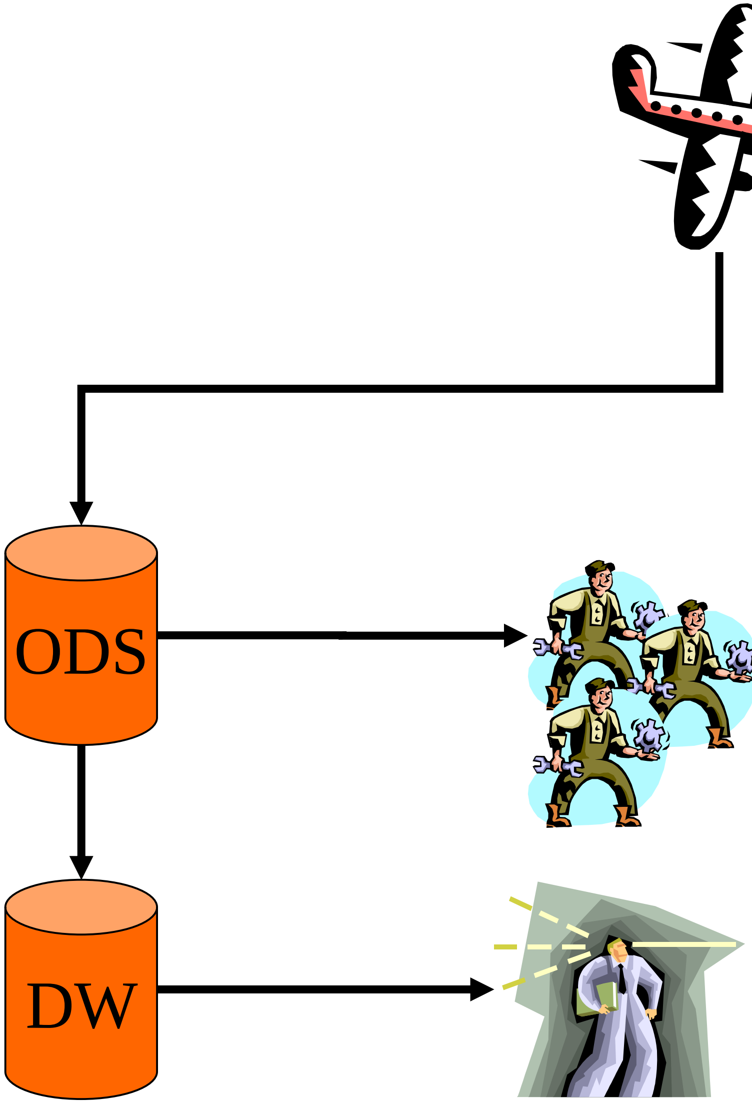


Figure 1-3 Dimensional model created from that shown in Figure 1-2 on page 12

Source: IBM Redbook: Dimensional Modeling: In a Business Intelligence Environment.

Why Is It Useful?



The data transmitted from the engine in flight can alert a service team of an engine component in need of repair so they can meet the plane at the gate.

Engineers can analyze the data to find ways of designing engine components with longer life spans.

Data Warehouse Benefits

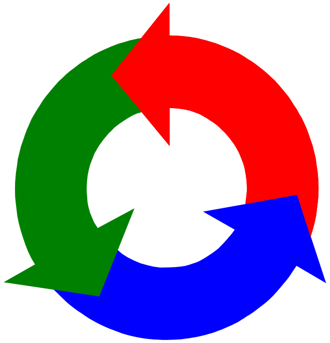
- Direct benefits of a data warehouse
  - Allows end users to perform extensive analysis
  - Allows a consolidated view of corporate data (single version of the truth)
  - Better and more timely information
  - Enhanced system performance
  - Simplification of data access

Data Warehousing Approaches

- Pre-eminent Data Warehousing Minds
  - Bill Inmon -> Normalization
    - Building the Data Warehouse
    - Corporate Information Factory
  - Ralph Kimball -> Dimensional
    - The Data Warehouse Lifecycle Toolkit
    - The Data Warehouse Toolkit

Stage 1: Analysis

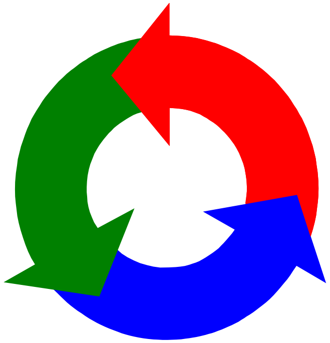
- Identify:
  - Target Questions
  - Data needs
  - Timeliness of data
  - Granularity
- Create an enterprise-level data dictionary
- Dimensional analysis
  - Identify facts and dimensions



✘ Analysis
– Design
– Import data
– Install front-end tools
– Test and deploy

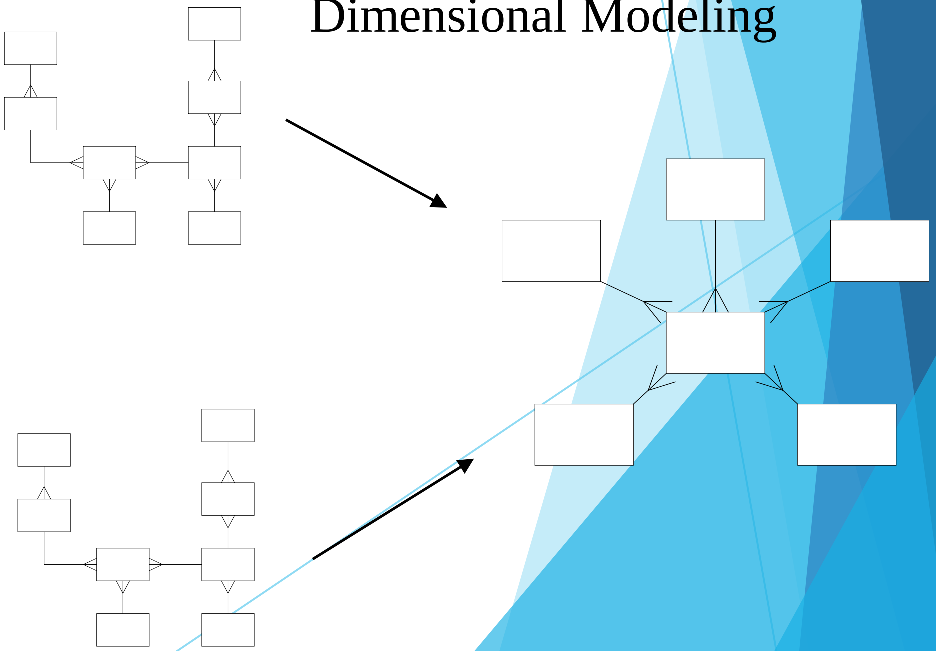
Stage 2: Design

- Star schema
- Data Transformation
- Aggregates
- Pre-calculated Values
- HW/SW Architecture



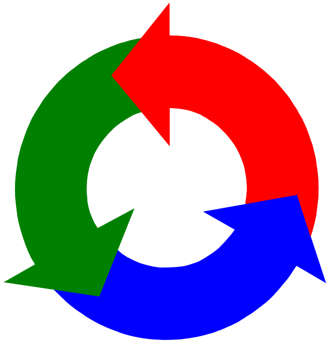
– Analysis
✘ Design
– Import data
– Install front-end tools
– Test and deploy

Dimensional Modeling



Stage 3: Import Data

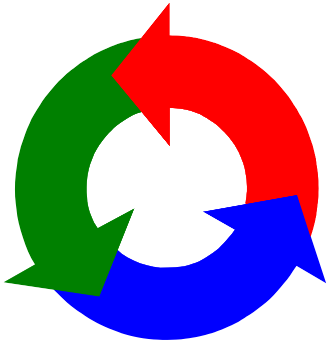
- Identify data sources
- Extract the needed data from existing systems to a data staging area
- Transform and Clean the data
  - Resolve data type conflicts
  - Resolve naming and key conflicts
  - Remove, correct, or flag bad data
  - Conform Dimensions
- Load the data into the warehouse



– Analysis
– Design
✘ Import data
– Install front-end tools
– Test and deploy

Stage 4: Install Front-end Tools

- Reporting tools
- Data mining tools
- GIS
- Etc.



– Analysis

– Design

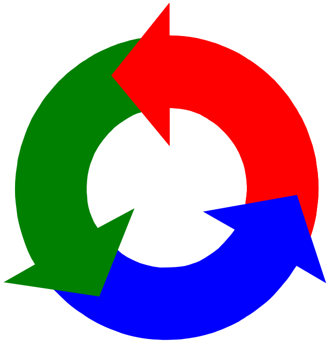
– Import data

✗ Install front-end tools

– Test and deploy

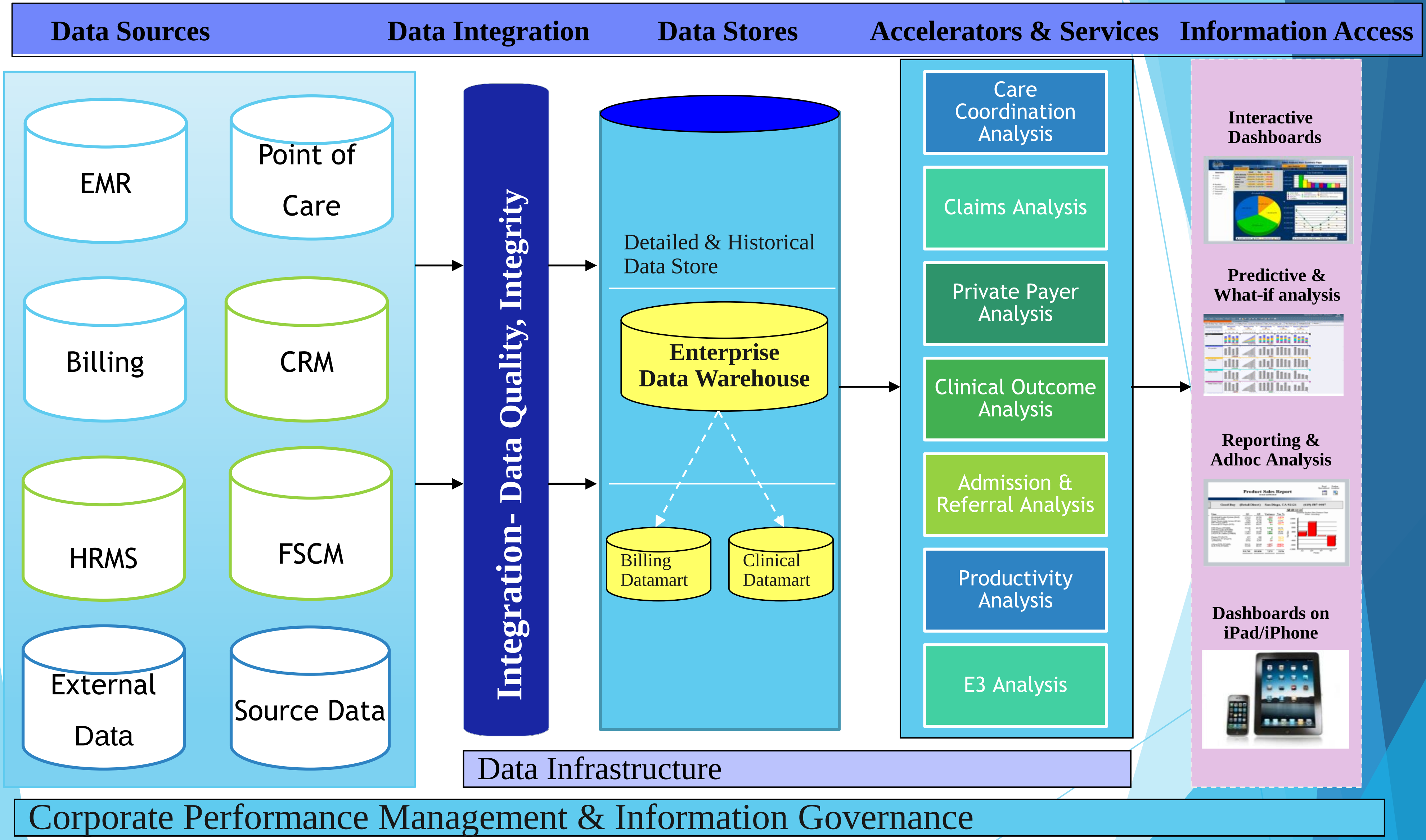
Stage 5: Test and Deploy

- Usability tests
- Software installation
- User training
- Performance tweaking based on usage



– Analysis
– Design
– Import data
– Install front-end tools
✗ Test and deploy

Data Warehouse Architecture



Data Sources | Data Integration | Data Stores | Accelerators & Services | Information Access

EMR

Point of Care

Billing

CRM

HRMS

FSCM

External Data

Source Data

Integration- Data Quality, Integrity

Detailed & Historical Data Store

Enterprise Data Warehouse

Billing Datamart

Clinical Datamart

Care Coordination Analysis

Claims Analysis

Private Payer Analysis

Clinical Outcome Analysis

Admission & Referral Analysis

Productivity Analysis

E3 Analysis

Interactive Dashboards

Predictive & What-if analysis

Reporting & Adhoc Analysis

Dashboards on iPad/iPhone

Data Infrastructure

Corporate Performance Management & Information Governance

Data Warehouse Architecture

- **Data Sources or component systems**
  - The operational systems or OLTP systems that provide the raw data for the data warehouse.
- **Staging Area**
  - A preparatory repository where transaction data can be transformed for use in the data warehouse
- **Operational Data Store (ODS)**
  - Modeled to support near real-time reporting needs
  - Contains traits of both relational and dimensional modeling techniques
- **Data Warehouse**
  - The database used to store abstracted and summarized operational data.
- **Data Mart**
  - Traditional dimensionally modeled set of dimension and fact tables
  - Per Kimball, a data warehouse is the union of a set of data marts

Data Warehouse Architecture

- **Extraction, Translation and Loading Tool** - Cleans operational data. Identifies out of bounds/incorrect data, discrepancies, missing data. Fills in NULLs.
- **User Interface** - An easy to use front end to facilitate querying and visualizing data in the data warehouse.

Data Warehousing Definitions and Concepts

- Characteristics of data warehousing
  - Subject oriented (sales, products, customers)
  - Integrated (consistent format)
  - Time variant (time series)
  - Nonvolatile (can’t change/update data)
  - Metadata (data about data)

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Metadata

- Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use
  - Document data about data elements or attributes, (name, size, data type, etc) and data about records or data structures (length, fields, columns, etc) and data about data (where it is located, how it is associated, ownership, etc.).
  - May include descriptive information about the context, quality and condition, or characteristics of the data.

Data Warehouse Development

- Data warehousing implementation issues
  - Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods
  - There are many facets to the project lifecycle, and no single person can be an expert in each area

Data Warehouse Development

- Massive data warehouses and scalability
  - The main issues pertaining to scalability:
    - The amount of data in the warehouse
    - How quickly the warehouse is expected to grow
    - The number of concurrent users
    - The complexity of user queries
  - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

Reference