Topic 01: Introduction to Data Management
Business Data Formation
and Data Storage

BDM3302: Data Management

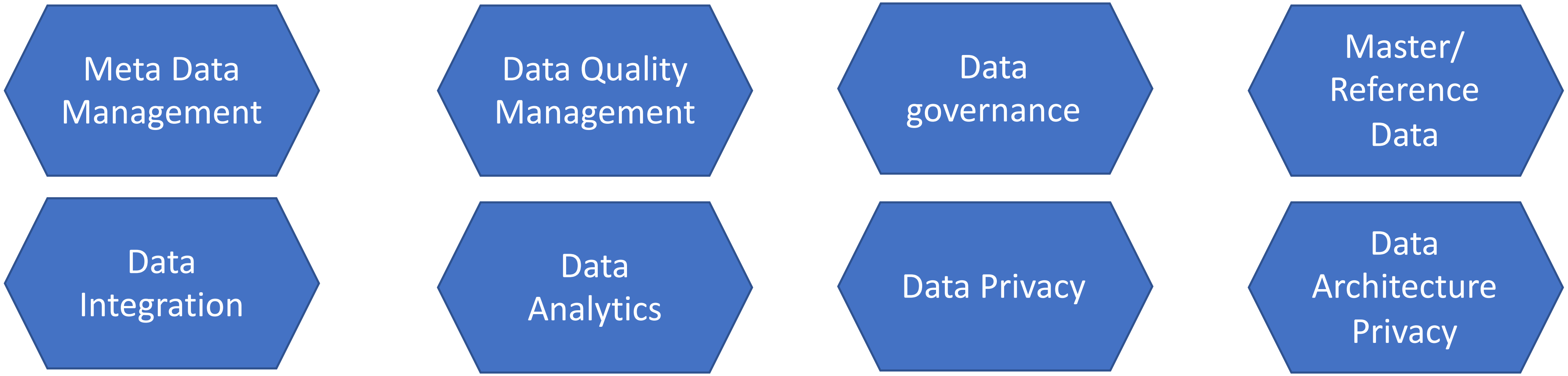
Do you trust your data?

- Each digital touchpoint provides an opportunity to gain new insight that you can use to propel your business forward. But do you trust your data?
- If your business have so much data and it's impossible to tell what's important and what's not. Or your data may be stuck in different systems in inconsistent formats, which makes it difficult to trust or share with teams that need it. Or, worse yet, you have outdated and inaccurate data. How quickly and accurately you can resolve these challenges will determine whether your data is truly the asset your business needs to succeed.
- Most organizations, large or small, struggle with data management issues such as quality, speed, availability, and privacy. This is because data is commonly replicated across multiple silos with few or no data governance processes to manage or maintain it.
- you might have a variety of customer data sources, such as loyalty programs, sales records, service calls, or surveys. Those data sources reside in systems across the enterprise and extended value chain.

What is Data Management?

ปฏิวัติ

- Data Management refers to the development and execution of architectures, policies, practices and procedures, in order to manage the information lifecycle of an enterprise in an effective manner.
- There are eight subject areas (capabilities) in Data Management:



Meta Data Management

Data Quality Management

Data governance

Master/ Reference Data

Data Integration

Data Analytics

Data Privacy

Data Architecture Privacy

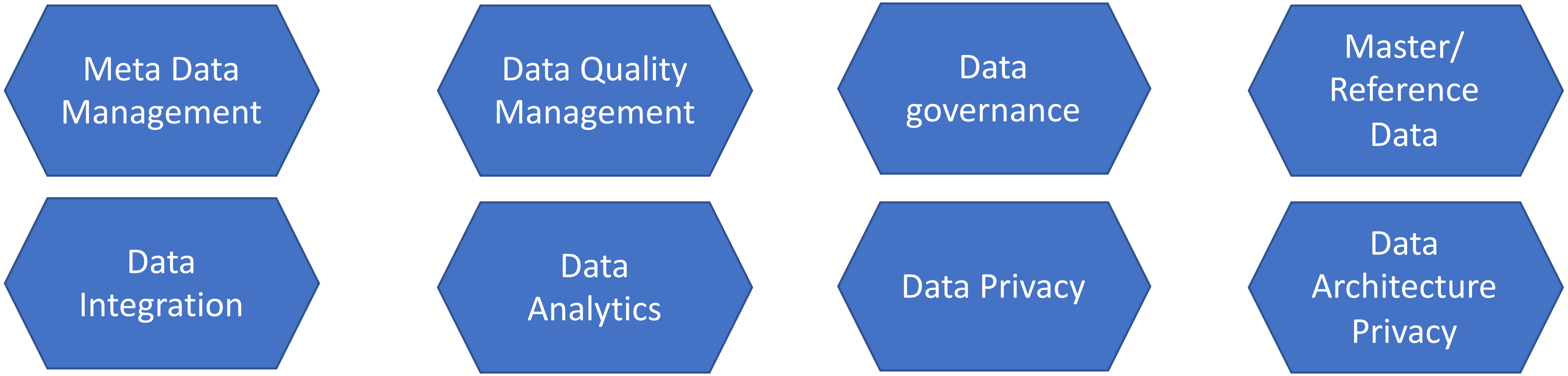
Data Management Framework

- **People** refers to organizational aspects the roles and responsibilities required for each subject areas.
- **Process** refers to activities that are associated with the subject areas.
- **Technology** refers to technologies and tools required to support capability business processes.

People

Process

Technology

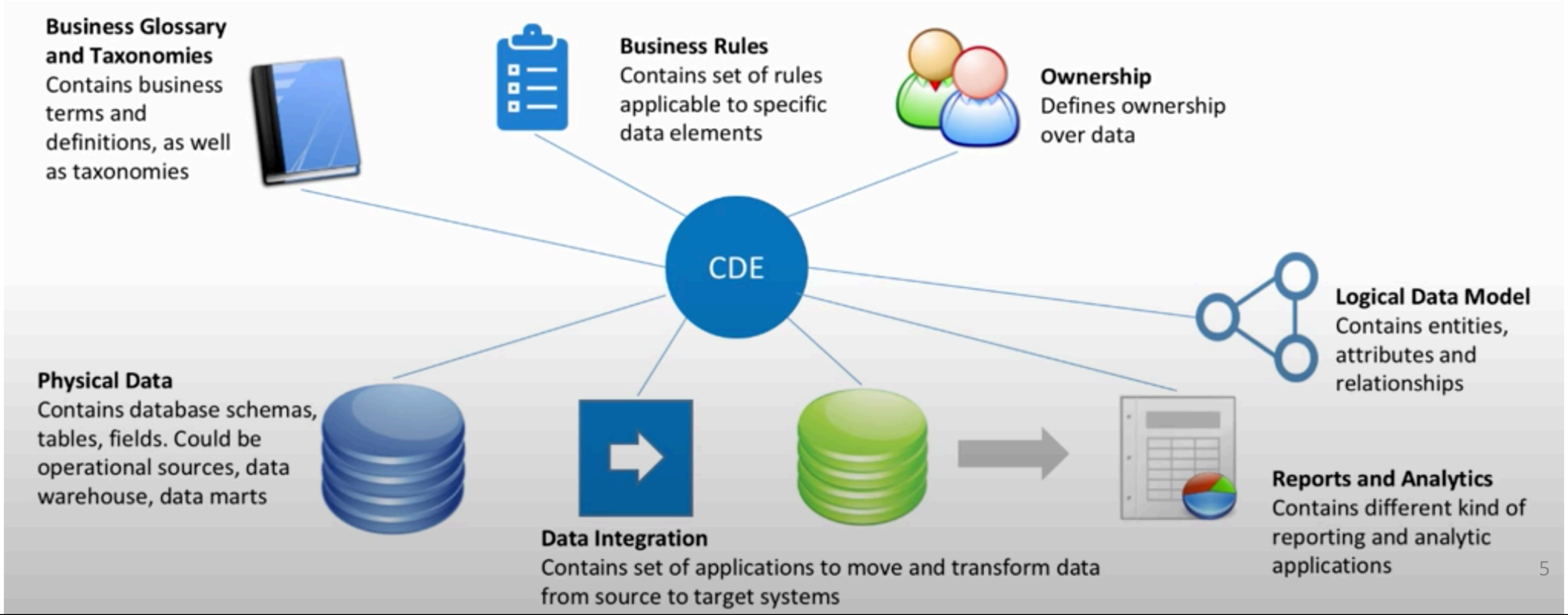


Metadata Management

- What is Data Element?
- Data Element (DE) is a unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes. ↳ ทุกอย่างยกเช่น consumer ID
- Critical Data Element (CDE) is the data element that is “critical to success” in a specific business area or business process.
- Criteria for Data Element to become Critical
  - Business facts that are deemed critical to the organization
  - Support critical business processes across an organization and its components
  - Data used to derive values that appear in key reports
  - Unique identifiers of things important to the business (such as Customer ID)
  - Any data element that is required for the execution of a key business process is a Critical Data Element

What is Metadata Management?

- Meatadata Management involves managing data about other data, whereby this “other data” is generally referred to data models and structures, not the content. It includes managing information about data structures from different models and their mutual associations (like business terms in glossary, attributes in logical data model, or tables and columns in the database, as well as their associations).



Business Glossary and Taxonomies
Contains business terms and definitions, as well as taxonomies

Business Rules
Contains set of rules applicable to specific data elements

Ownership
Defines ownership over data

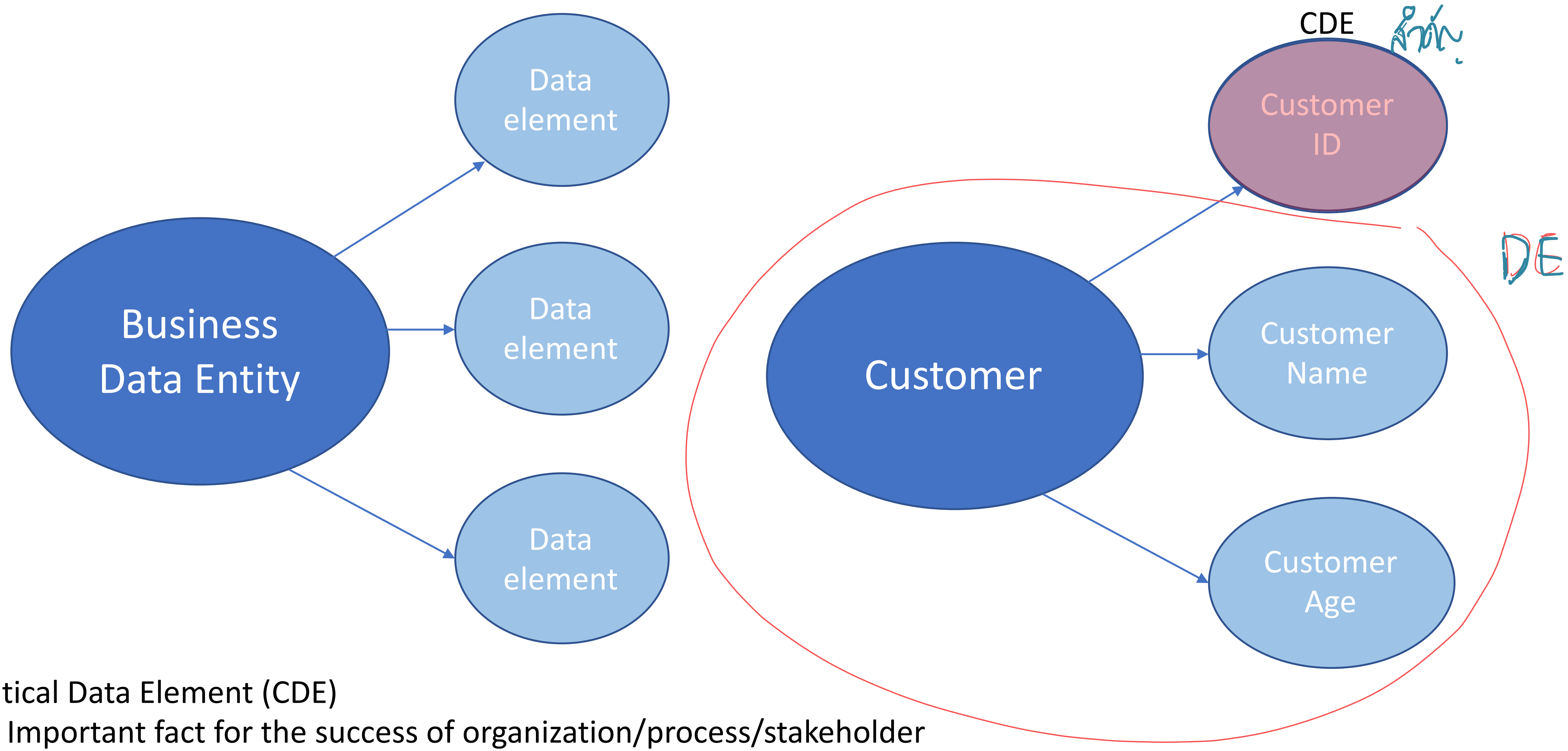
Logical Data Model
Contains entities, attributes and relationships

Physical Data
Contains database schemas, tables, fields. Could be operational sources, data warehouse, data marts

Data Integration
Contains set of applications to move and transform data from source to target systems

Reports and Analytics
Contains different kind of reporting and analytic applications

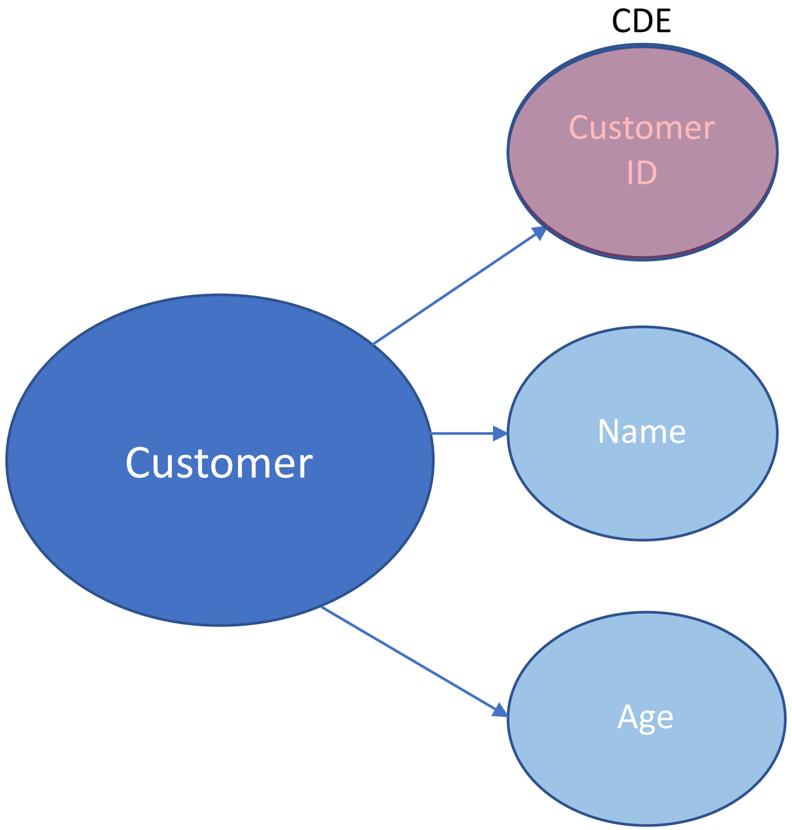
Data Management> Meta Data Management> Business Data Entity & Data Element



Critical Data Element (CDE)

- Important fact for the success of organization/process/stakeholder
- Appears in Key report
- Important for a set of business proccesses

Data Management> Meta Data Management> Data Element Charcteristic



Data element characteristic
DE : has size (What?) ✓
DE : has data type (What?) ✓
CDE : purpose of data serves (Why?)
CDE : critical to whom/stakeholder (Who?)
CDE : unique identification (text, or number) (What?) ✓
CDE : time stamp (When?)
CDE : Data source (Where?)

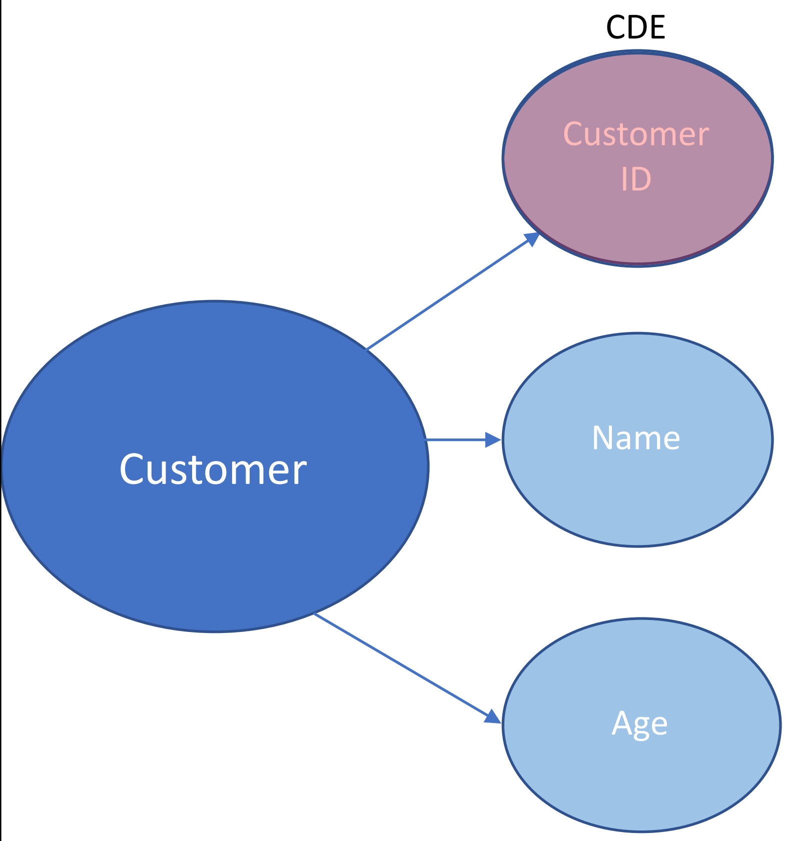
PK → ข้อมูลแค่ 1
เพราะรรมทุกทาศ

FK ลิงก์หลายๆตาราง

Data Name	Custome ID	Name	Age
type	int ตัวเลข	Varchar(20) text	int
role	PK,DE	DE	DE
sample	101	John	40

double ทศนิยม

Data Management> Meta Data Management> Data Element Charcteristic



Data element characteristic

DE : has size (What?) ✓

DE : has data type (What?) ✓

DE : purpose of data serves (Why?)

DE : critical to whom/stakeholder (Who?) ✓

DE : unique identification (text, or number) (What?) ✓

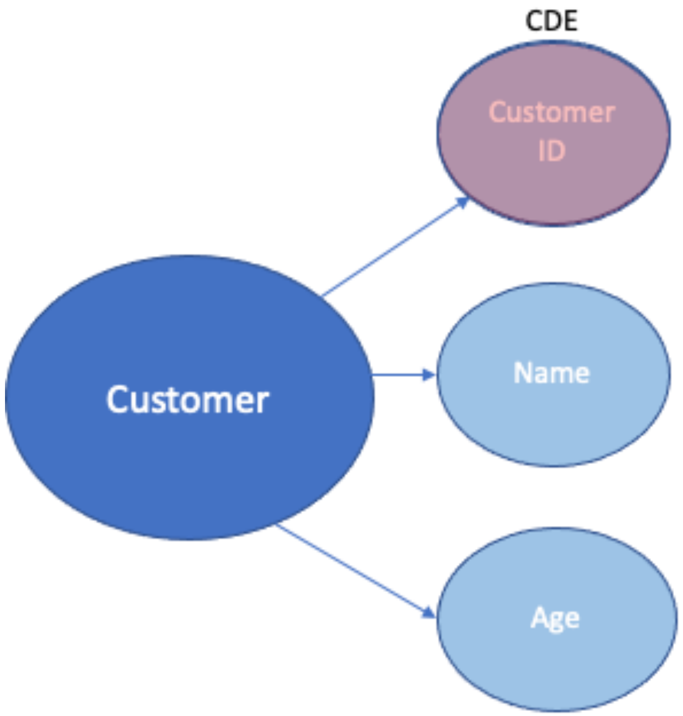
DE : time stamp (When?) ✓

DE : data source (Where?) ✓

DE : constraints (PK, Nullity, Autoincrement, FK, etc)

Key processes	**Product order, customer support, product enquery, promotion etc.**			
Table	**Customer**			
Data Name	**Customer ID**	**Name**	**Age**	**Date**
type	int	Varchar (20) text	int	DateTime
constriant	PK, auto_inc	Not null	Not null	Data field
sample	101	John	40	15 Nov 2021

Data Management> Meta Data Management> Data Element Charcteristic for technical and business perspective



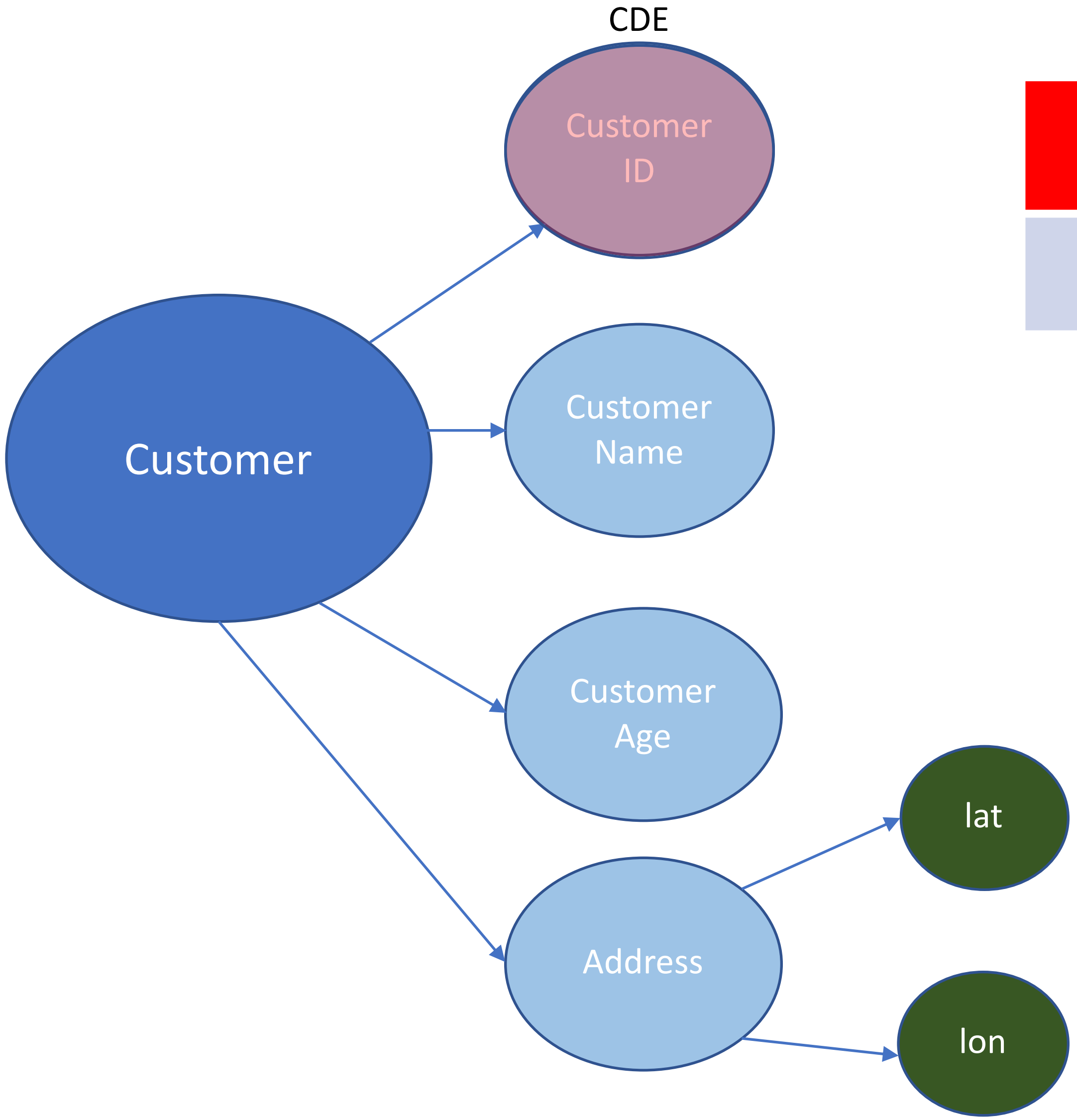
Key processes	Product order, customer support, product enquery, promotion etc.			
Data Name	Customer ID	Name	Age	Date
type	int	Varchar (20)	int	DateTime
constriants	PK, DE	DE	DE	Data field
sample	101	John	40	15 Nov 2021

technical

Data Name	Name
Description	represents customer full name
Business rule	Thai full name consist of First name and Last Name Chinese name has middle name in between
Data Owner/ Producer	Sale Team
Data Consumer	Sale Team, Marketing Team, Shipment Team, etc
Data Policy	Marketing Team can have access to the data prior to the year 2016
Report	Sale Report, Daily shipment Report, Promotion Report

business

Data Management> Meta Data Management> Data Type, Table, and Normalization Process



Customer ID (int)	Name (varchar)	Age (int)	Address ?
101	John	40	???

Data Management> Meta Data Management> Data Type, Table, and Normalization Process

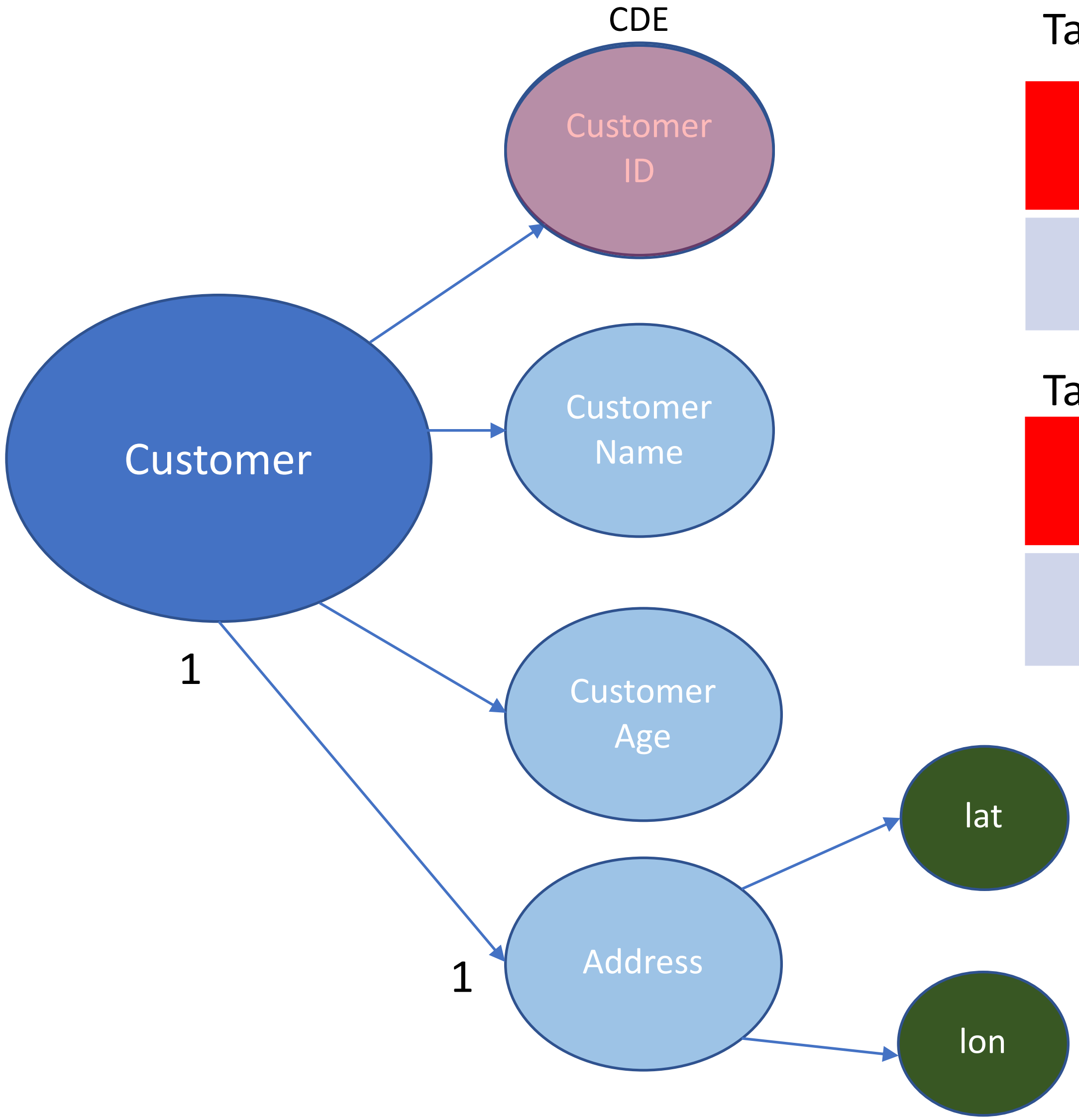


Table: Customer

Customer ID (int)	Name (varchar)	Age (int)	Address Id (int)
101	John	40	101

Table: Address

Address ID (int)	Lat (double)	Long (double)
101	13.12345	101.1234

Object notation

```
{
     CustomerId : 101   ,
     Name : "John"   ,
     Age : 40  ,
     Address : { lat : 13.12345    ,    lon : 101.1234  }


}
```

Data Management> Meta Data Management> Data Type, Table, and Normalization Process

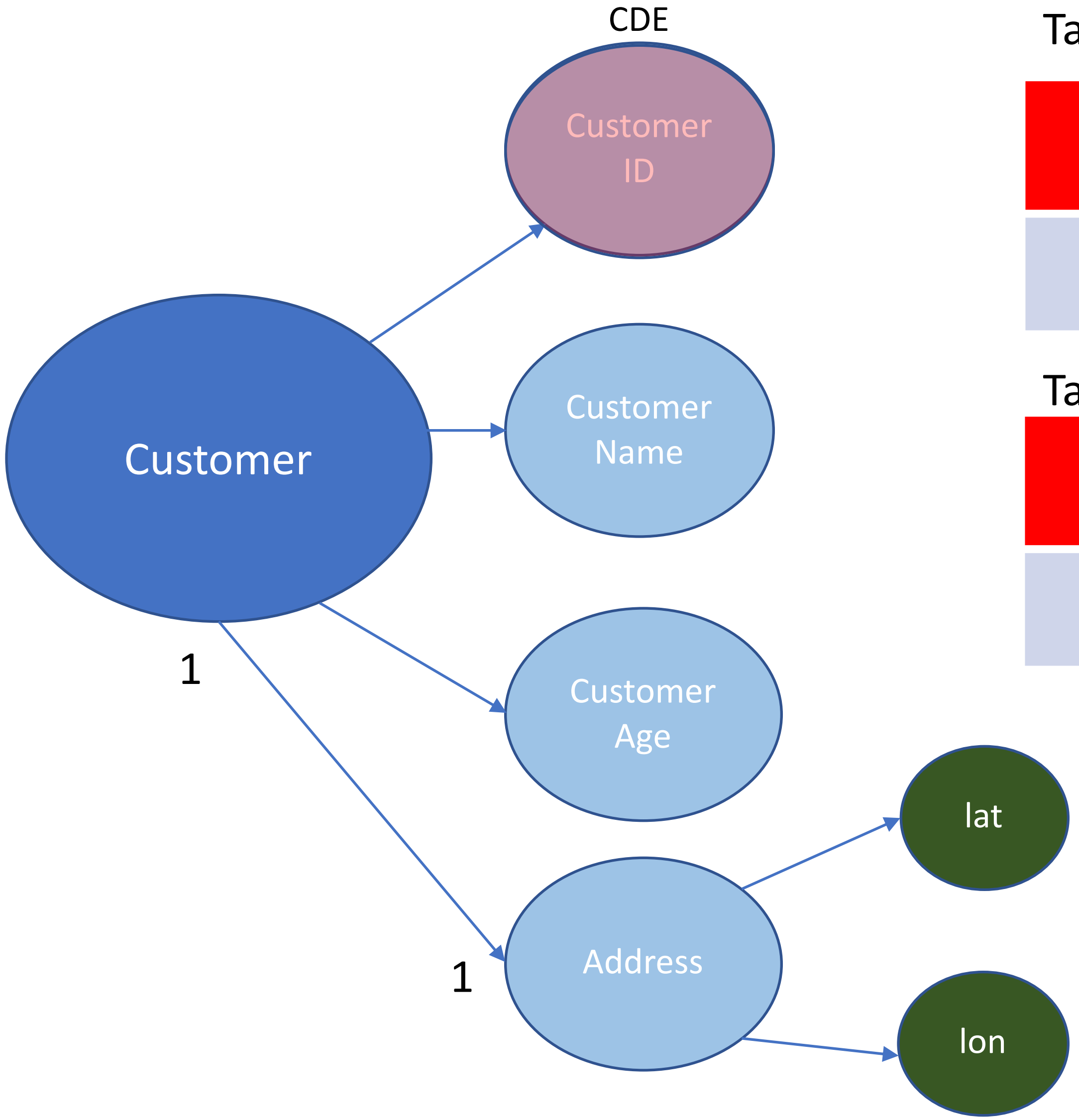


Table: Customer

Customer ID (int)	Name (varchar)	Age (int)	Address Id (int)
101	John	40	101

Table: Address

Address ID (int)	Lat (double)	Long (double)
101	13.12345	101.1234

```
{
      "CustomerId" : 101   ,
      "Name " : "John" ,
      "Age" : 40  ,
      "Address" :{ lat : 13.12345 ,     lon : 101.1234 }


}
```

JSON format

Data Management> Meta Data Management> Data Type & Normalization Process

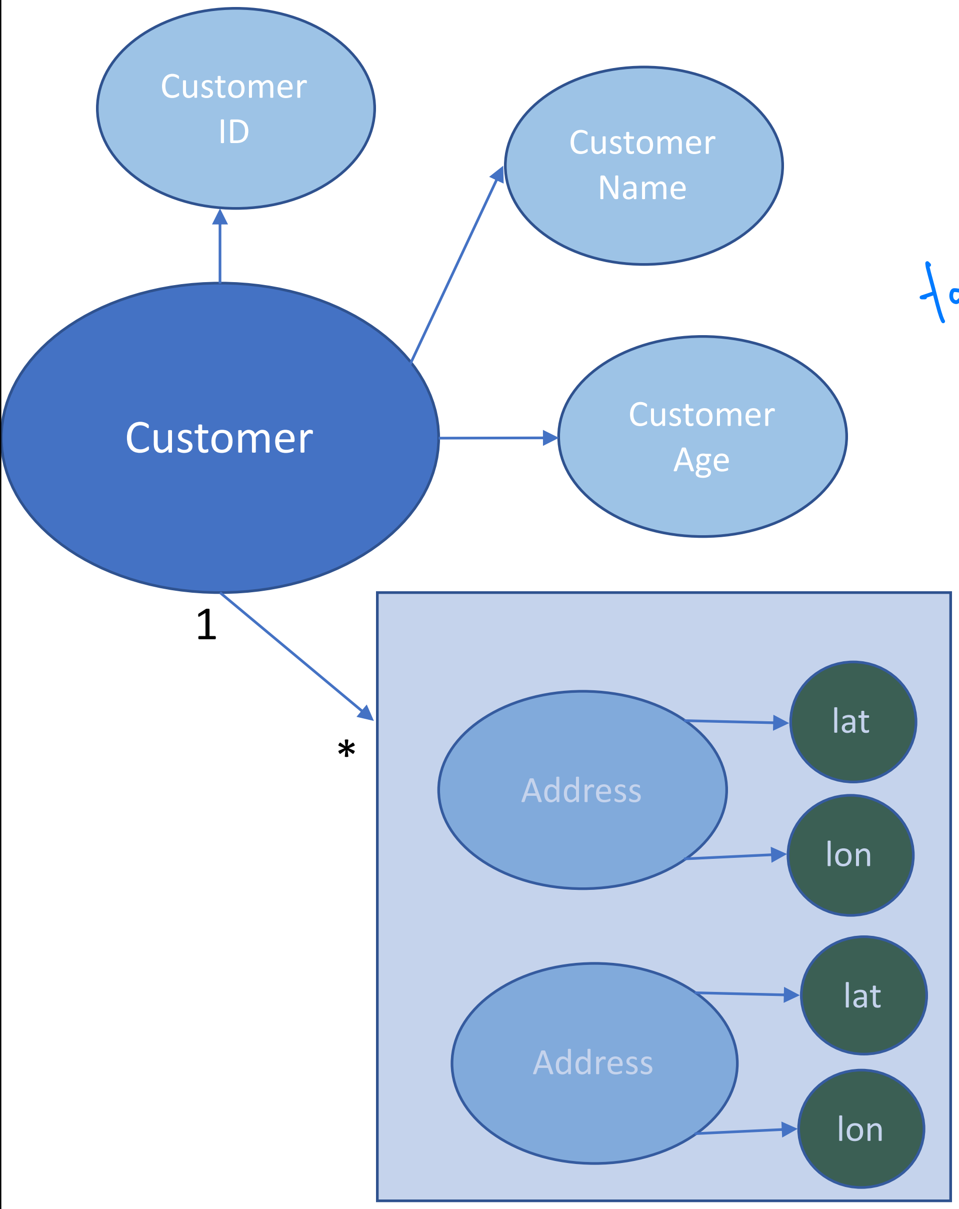


Table: Customer

Customer ID (int)	Customer (varchar)	Customer Age (int)
101	John	40

table แรก

~~101~~	John	40	102

Address ID PK (int)	Lat (double)	Long (double)	Customer ID (int) FK
101	13.12345	101.1234	101
102	13.22345	103.1234	101

ทศนิยม

เชื่อมตาราง

```
{
     CustomerId : 101   ,
     Name : "John"   ,
     Age : 40  ,
     Addresses :  [    { lat : 13.12345  ,   lon : 101.1234 },
                       { lat : 13.22345  ,   lon : 103.1234 }
                  ]
}
```

Data Management> Meta Data Management> Data Type & Normalization Process

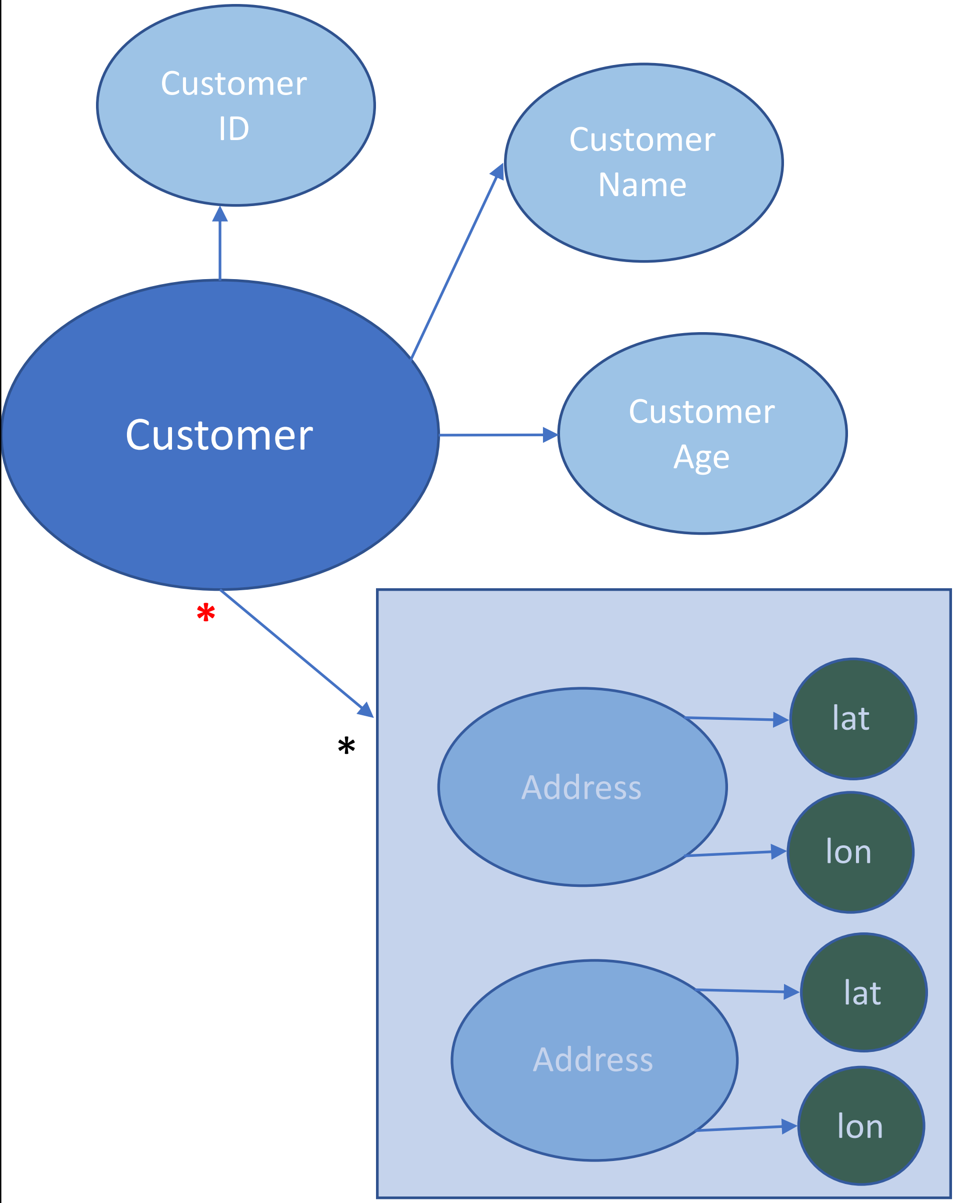


Table: Customer

Customer ID (int)	Customer (varchar)	Customer Age (int)
101	John	40
102	Sam	20

Table: Address

Address ID (int)	Lat (double)	Long (double)
101	13.12345	101.1234
102	13.22345	103.1234

Table: CustomerAddress

No	Customer ID (int)	Address ID (int)
1	101	101
2	101	102
3	102	101

2

Data Quality Management

- What is Data Quality (DQ)?
- Data Quality refers to the methodical approach, policies and processes by which an organization manages the **accuracy, validity, timeliness, completeness, uniqueness, and consistency** of its data in systems and data flows.
- Not all Data Quality Dimensions are applicable on particular CDE (e.g. Date of Birth can be assessed against Validity and Completeness only)

Data Quality Dimensions

- **Accuracy** refers to error-free records that can be used as a reliable source of information.
- **Validity** refers to information that does not conform to a specific format or does not follow business rules. (DOB format)
- **Timeliness** refers to the time expectation for accessibility and availability of information. It can be measured as the time between when information is expected and when it is readily available for use.
- **Completeness** refers to the degree to which all data in a data set is available. (all required fields like *first name* must have data.)
- **Uniqueness** refers to a measure of unwanted duplication existing within or across systems for a particular field, record, or data set.
- **Consistency** refers to data values that are the same for all instances of an applications. (an employee does not work anymore but still receiving a check or money transfer.)

3

What is Data Governance?

- Data Governance refers to a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives. (Wikipedia.org)
- The key focus areas of data governance include availability, usability, consistency, data integrity and data security, standard compliance and includes establishing processes to ensure effective data management throughout the enterprise.

ภายใน —internal

4 What is Master and Reference Data Management?

external

- Master data management (MDM) refers to a process that creates a uniform set of data on customers, products, suppliers and other business entities from different IT systems. One of the core disciplines in the overall data management process, MDM helps improve data quality by ensuring that identifiers and other key data elements about those entities are accurate and consistent enterprise-wide. (TechTarget.com)
- Once created, this master data serves as a trusted view of business-critical data that can be managed and shared across the business to promote accurate reporting, reduce data errors, remove redundancy, and help workers make better-informed business decisions. (Informatica.com)

รวม

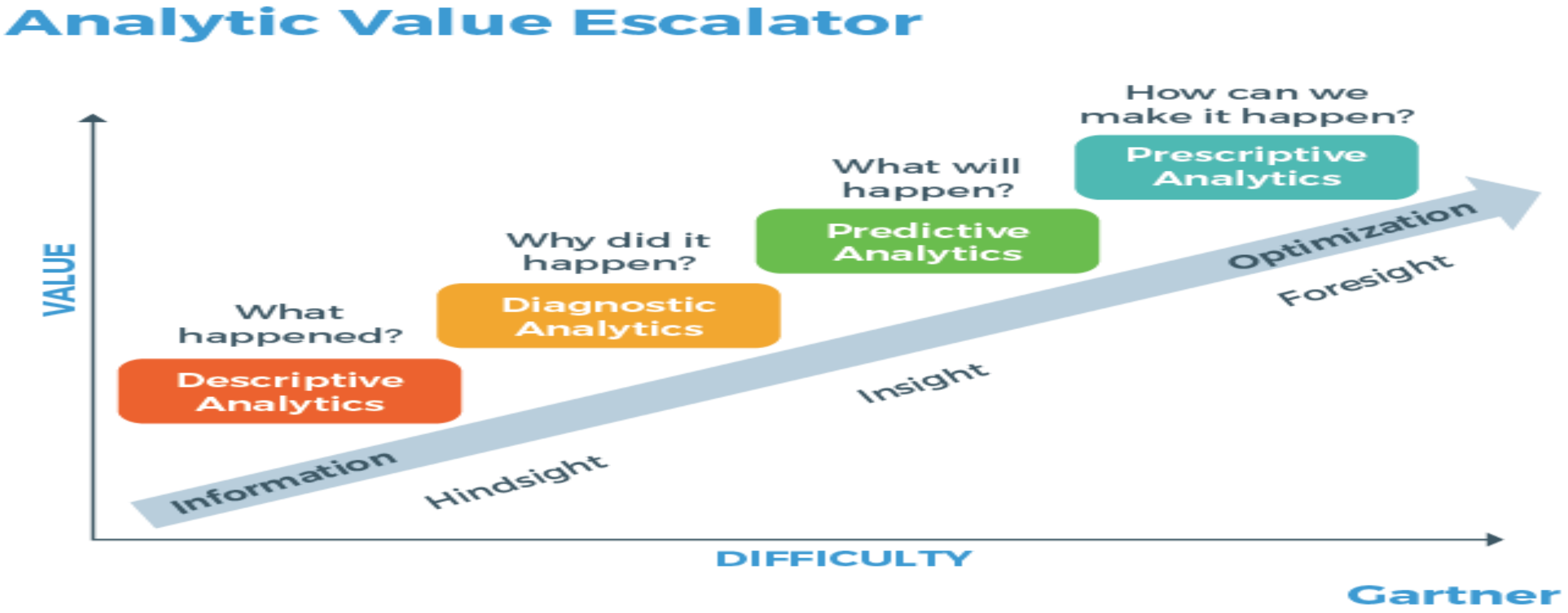
5 What is Data Integration?

- Data Integration (DI) refers to the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process, and includes steps such as cleansing, ETL mapping, and transformation. Data integration ultimately enables analytics tools to produce effective, actionable business intelligence. (Talend.com)
- Extract, Transform, Load (ETL) is a process within data integration wherein data is taken from the source system and delivered into the warehouse. This is the ongoing process that data warehousing undertakes to transform multiple data sources into useful, consistent information for business intelligence and analytical efforts.
- Batch vs Real Time Data

6

What is Data Analytics?

- Data Analytics is the process of examining data sets in order to find trends and draw conclusions about the information they contain. Increasingly, data analytics is done with the aid of specialized systems and software. (TechTarget.com)
- Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions.



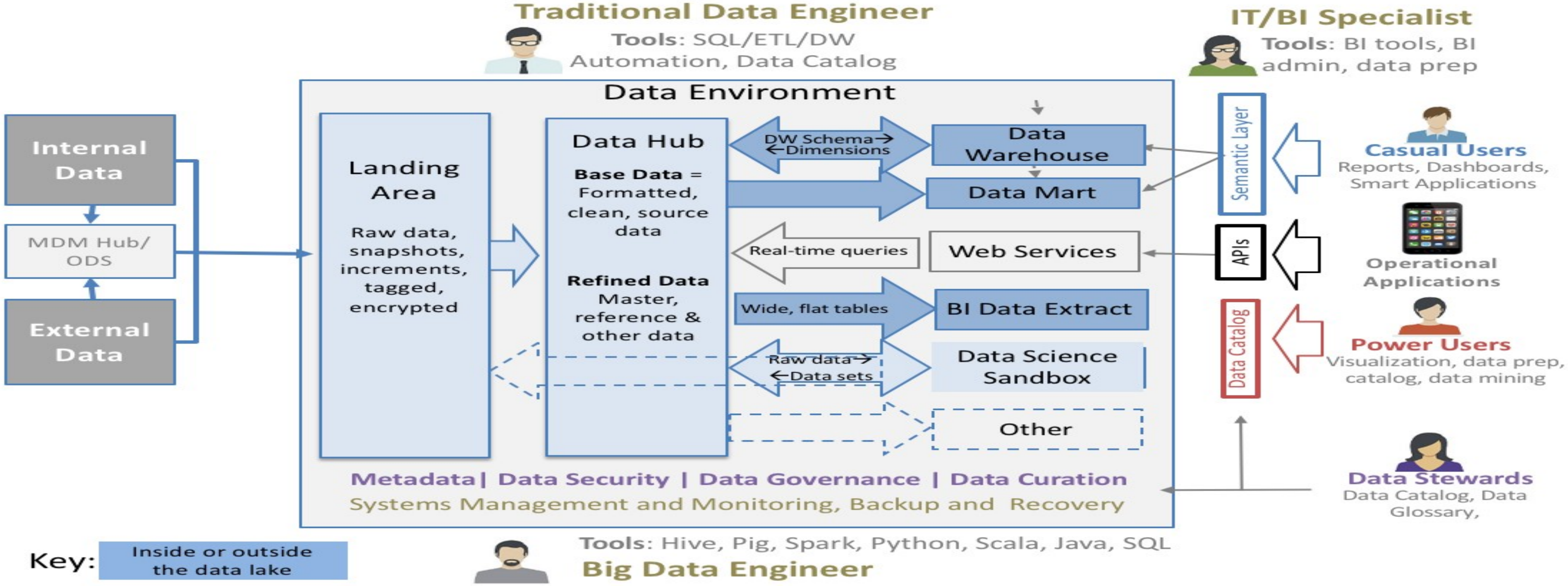
7

What is Data Privacy?

- Data privacy generally refers to the ability of a person to determine for themselves when, how, and to what extent personal information about them is shared with or communicated to others. This personal information can be one's name, location, contact information, or online or real-world behavior. (Cloudflare.com)
- Privacy is considered a fundamental human right, and data protection laws exist to guard that right. Data privacy is important because when individuals to be willing to engage online, they have to trust that their personal data will be handled with care.
- Organizations use data protection practices to demonstrate to their customers and users that they can be trusted with their personal data.
- General Data Protection Regulation (GDPR) - Regulates how the personal data of European Union (EU) data subjects, meaning individuals, can be collected, stored, and processed, and gives data subjects rights to control their personal data.
- California Consumer Privacy Act (CCPA) - Requires that consumers be made aware of what personal data is collected and gives consumers control over their personal data, including a right to tell organizations not to sell their personal data.

What is Data Architecture?

- Data Architecture is the models, policies, rules, and standards that govern which data is collected and how it is stored, arranged, integrated, and put to be used in data systems and in organizations. (Wikipedia.org)
- Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.


Traditional Data Engineer
Tools: SQL/ETL/DW Automation, Data Catalog
IT/BI Specialist
Tools: BI tools, BI admin, data prep
Data Environment
Internal Data
MDM Hub/ ODS
External Data
Landing Area
Raw data, snapshots, increments, tagged, encrypted
Data Hub
Base Data = Formatted, clean, source data
Refined Data Master, reference & other data
DW Schema→
←Dimensions
Data Warehouse
Data Mart
Real-time queries
Web Services
Wide, flat tables
BI Data Extract
Raw data→
←Data sets
Data Science Sandbox
Other
Semantic Layer
APIs
Data Catalog
Casual Users
Reports, Dashboards, Smart Applications
Operational Applications
Power Users
Visualization, data prep, catalog, data mining
Data Stewards
Data Catalog, Data Glossary,
Metadata | Data Security | Data Governance | Data Curation
Systems Management and Monitoring, Backup and Recovery
Key:
Inside or outside the data lake
Tools: Hive, Pig, Spark, Python, Scala, Java, SQL
Big Data Engineer

Data Formation

- Text: XML, PDF/A, HTML, ASCII, UTF-8 (not Word)
- Tabular Data: CSV (not Excel)
- MS Excel: XLS, XLSX
- Databases: XML, CSV

Multimedia formats:

- Still Images: TIFF, JPEG 2000, PDF, PNG, BMP (not GIF or JPG)
- Moving Images: MOV, MPEG, AVI, MXF (not Quicktime)
- Sounds: WAVE, AIFF, MP3, MXF

Statistics formats:

- Statistics: ASCII, SAS, SAV

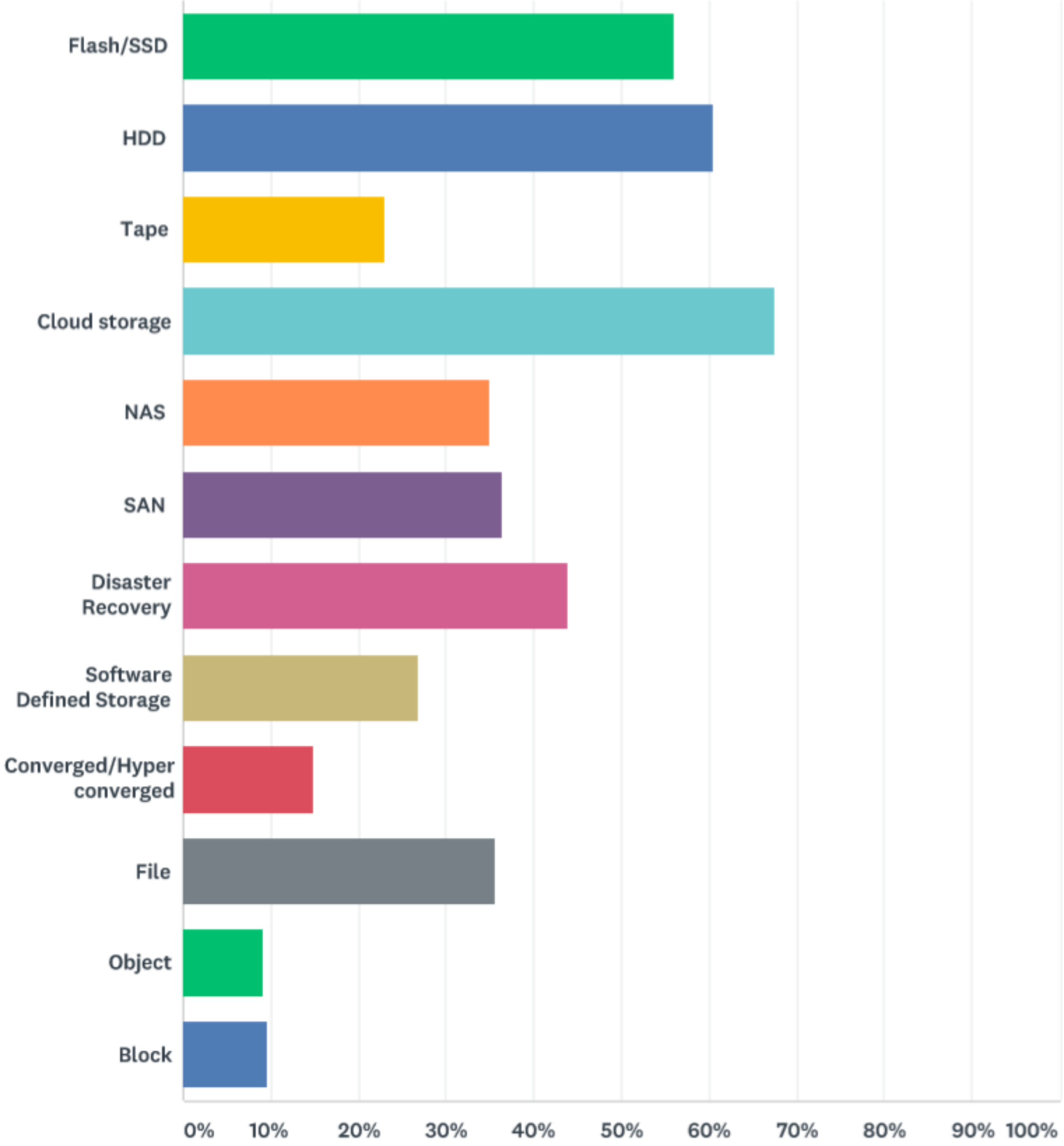
Data Types

- String (text)
  - Varchar - last name
- Numeric (number)
  - Int – student ID
  - Double - weight ทศนิยม
- Date
  - Date – enroll date
- Datetime
  - Clock-in/ Clock-out at workplace (Timestamp)

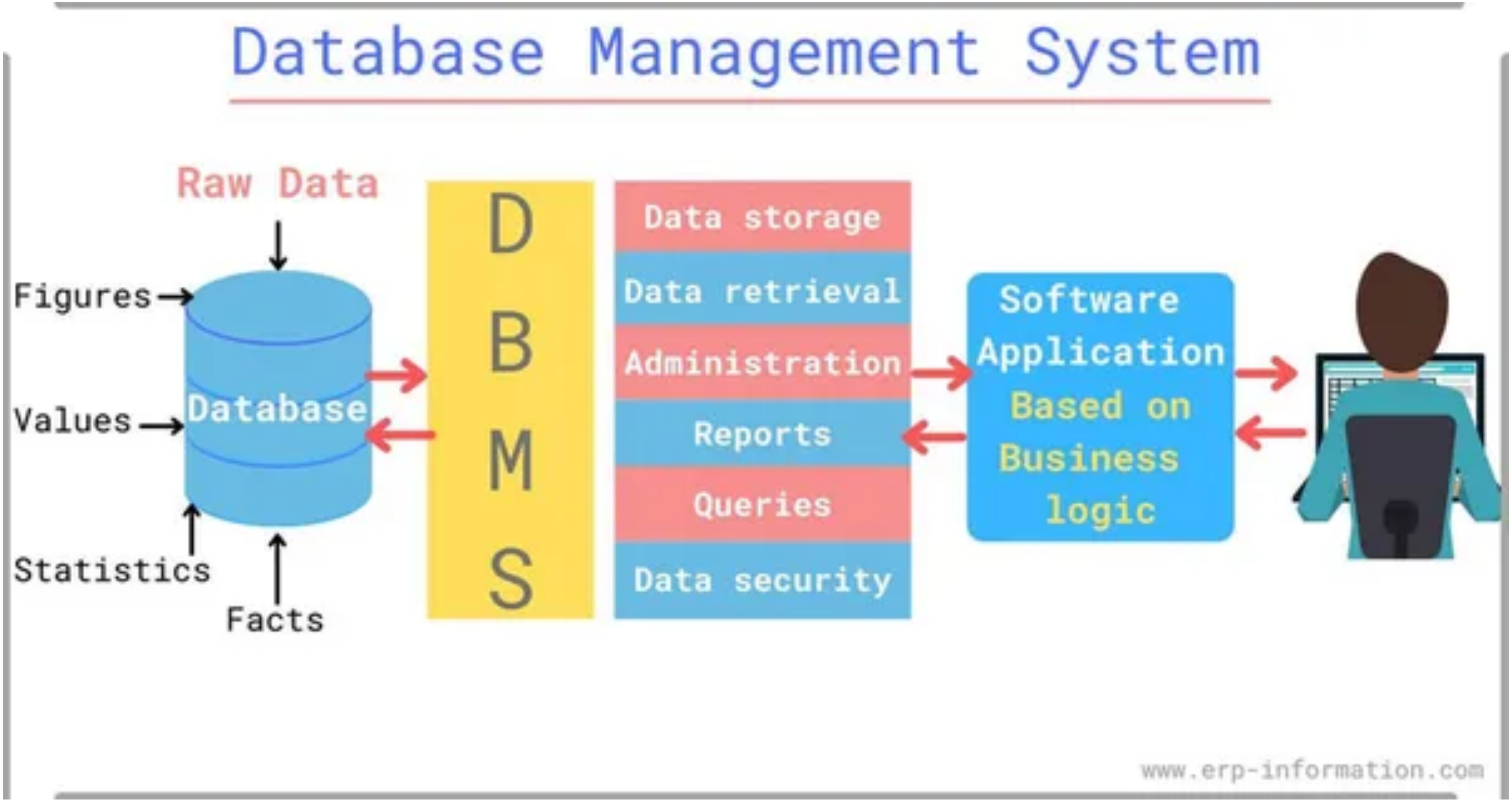
Data Storage

เก็บข้อมูลสั้นๆ

- Computer storage
    - Primary storage
    - Secondary storage
- Cloud storage - a cloud computing model that stores data on the Internet through a cloud computing provider who manages and operates data storage as a service. (AWS)



Source: https://www.enterprisestorageforum.com/management/survey-reveals-tech-trends-reshaping-data-storage/



Database Management System

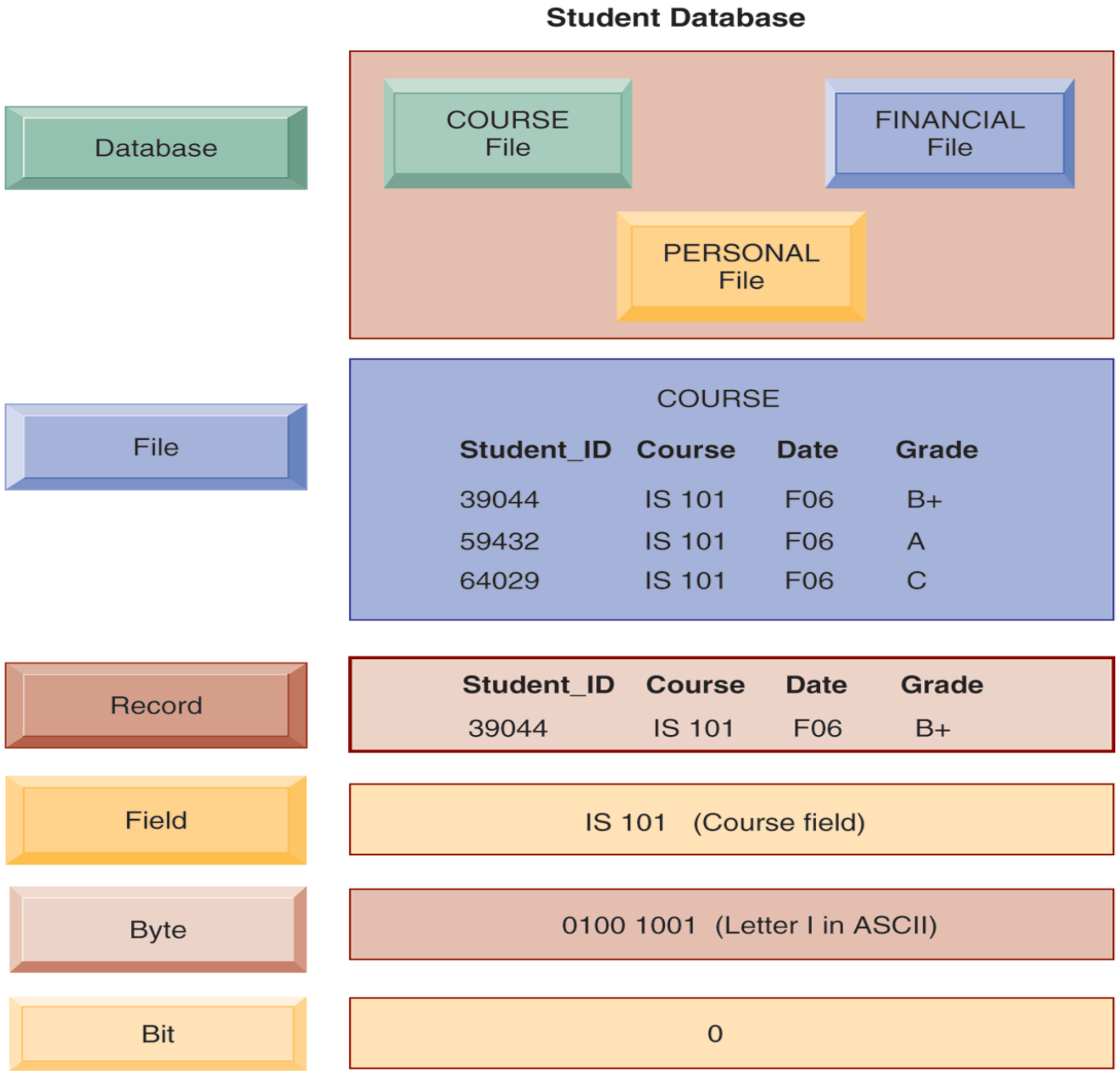
Organizing Data in a Traditional File Environment

- File organization concepts
  - Database: Group of related files
  - File: Group of records of same type
  - Record: Group of related fields
  - Field: Group of characters as word(s) or number
    - Describes an **entity** (person, place, thing on which we store information)
    - **Attribute:** Each characteristic, or quality, describing entity
      - E.g., Attributes Date or Grade belong to entity COURSE

Organizing Data in a Traditional File Environment

THE DATA HIERARCHY

A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.



Student Database

Database

COURSE File

FINANCIAL File

PERSONAL File

File

COURSE

Student_ID	Course	Date	Grade
39044	IS 101	F06	B+
59432	IS 101	F06	A
64029	IS 101	F06	C

Record

Student_ID	Course	Date	Grade
39044	IS 101	F06	B+

Field

IS 101 (Course field)

Byte

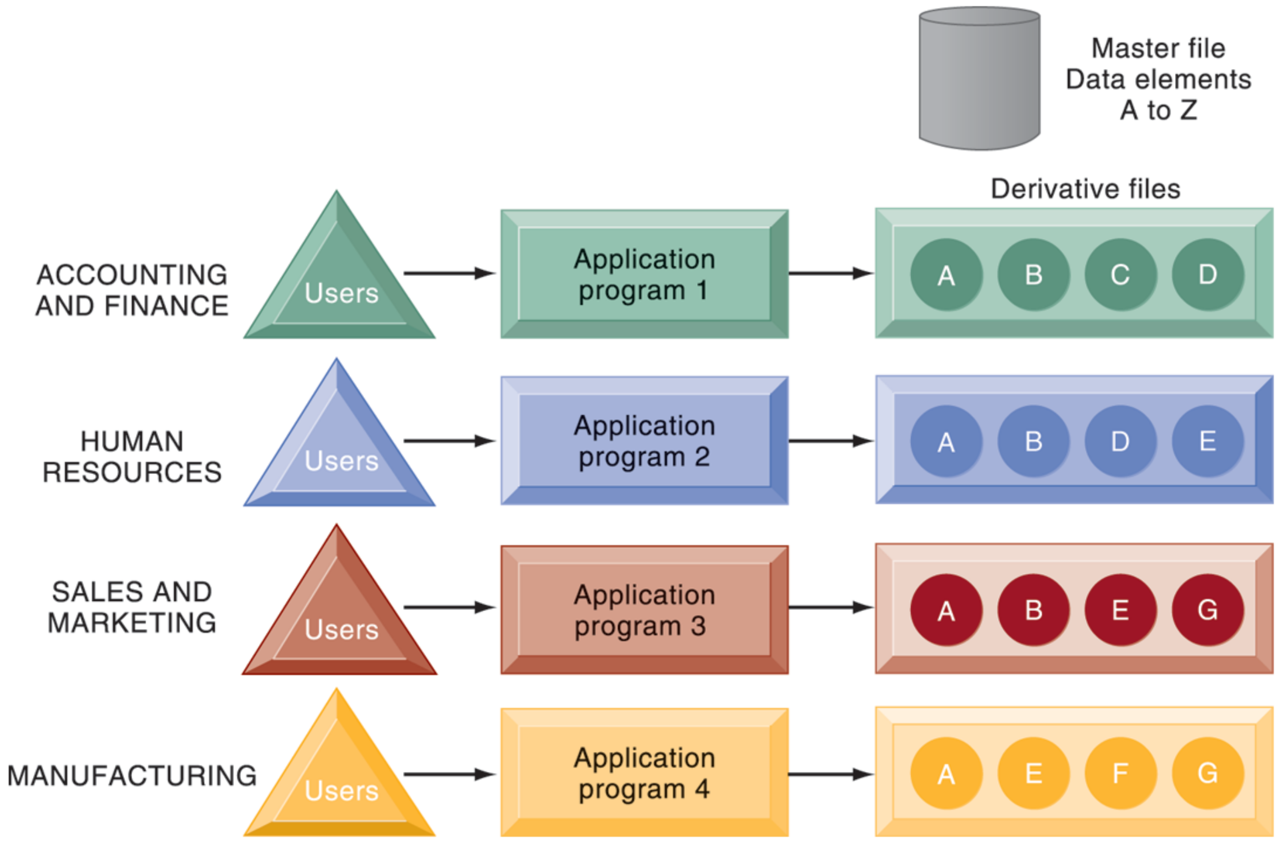
0100 1001 (Letter I in ASCII)

Bit

0

Organizing Data in a Traditional File Environment

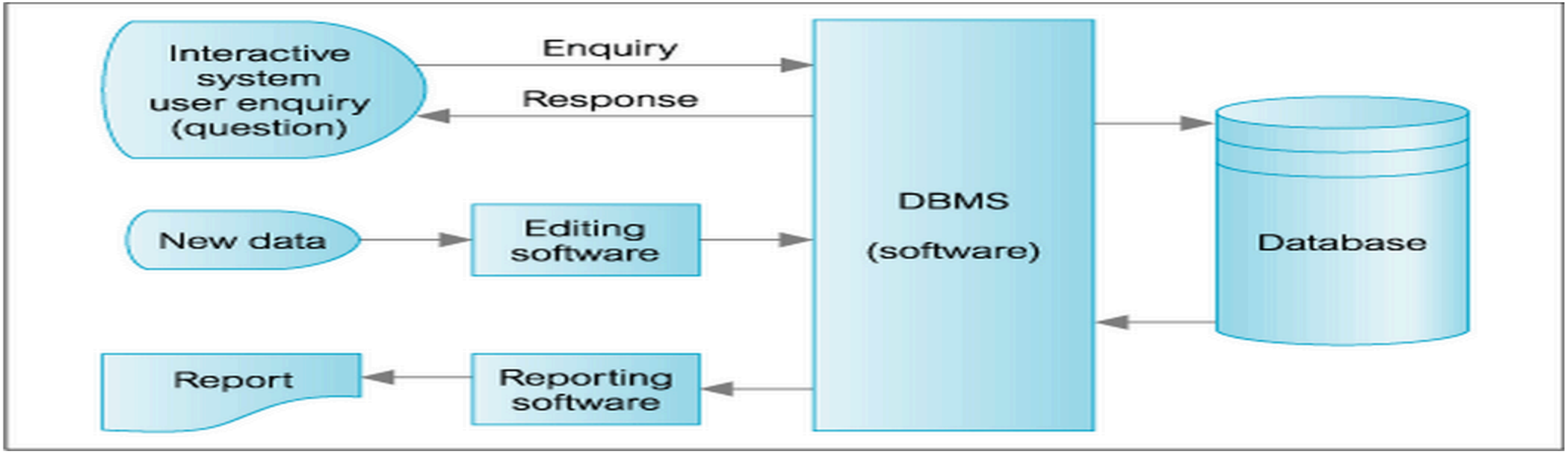
TRADITIONAL FILE PROCESSING



What kinds of data might be shared between sales and marketing and accounting?

Database Approach to Data Management

- Database
  - Eliminates many of the problems of traditional file organization by organizing data, centralizing data and controlling redundant data, and serve many applications and different groups at the same time.


Interactive system user enquiry (question)
Enquiry
Response
DBMS
(software)
Database
New data
Editing software
Report
Reporting software