TECHNOLOGY INFORMATION SYSTEMS

6

Foundations of Business Intelligence: Databases and Information Management

Managing Data in a Traditional File Environment

File organization concepts

- **Database: Group of related files**
- **Database: A collection of related information stored in an organized way so that specific items can be selected and retrieved quickly.**
- **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
    - Example: Attributes DATE or GRADE belong to entity COURSE

Database – advantages

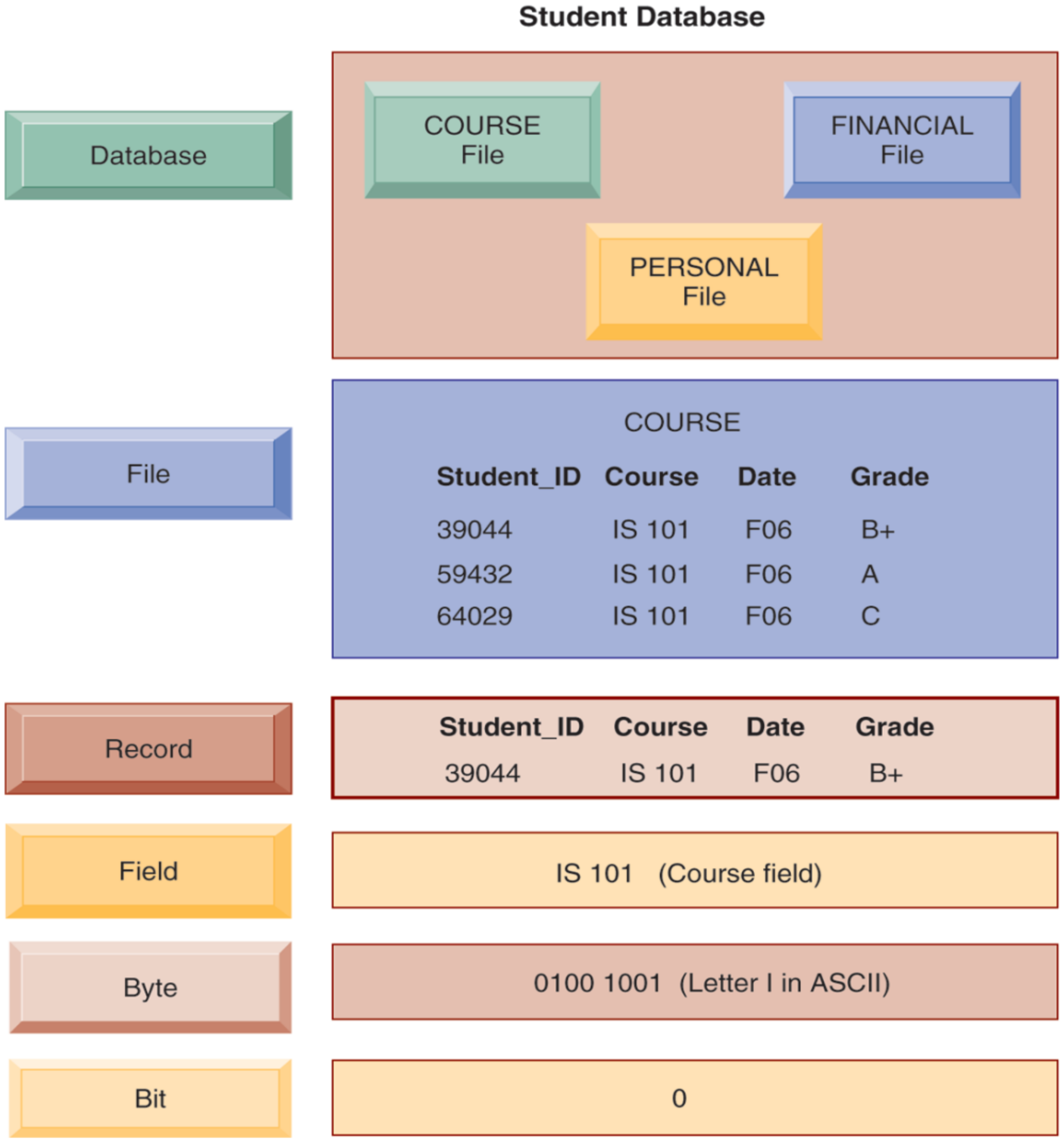
- ***Multi-user access*** – allowing different people in the business access to the same data simultaneously such as a manager and another member of staff accessing a single customer's data.
- ***Distributed access*** – users in different departments of the business can readily access data.
- ***Speed*** – for accessing large volumes of information, such as the customers of a bank, only databases are designed to produce reports or access the information rapidly about a single customer.
- ***Data quality*** – sophisticated validation checks can be performed when data are entered to ensure their integrity.
- ***Security*** – access to different types of data can readily be limited to different members of staff. In a car dealership database, for example, the manager of a single branch could be restricted to sales data for their branch.
- ***Space efficiency*** – by splitting up a database into different tables when it is designed, less space is needed

Key database concepts

- **Field**: The data in an electronic database are organised by fields and records. A field is a single item of information, such as a name or a quantity.
- **Record**: In an electronic database, a record is a collection of *related* fields. See *Field*.
- **Table**: In an electronic database, data are organised within structures known as tables. A table is a collection of many records.
- **Relationship**: In a relational database, data can be combined from several different sources by defining relationships between tables.
- **Compound key**: In a relational database, it is possible to retrieve data from several tables at once by using record keys in combination, often known as a compound key.
- **Foreign (secondary) key fields**: These fields are used to link tables together by referring to the primary key in another database table.

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

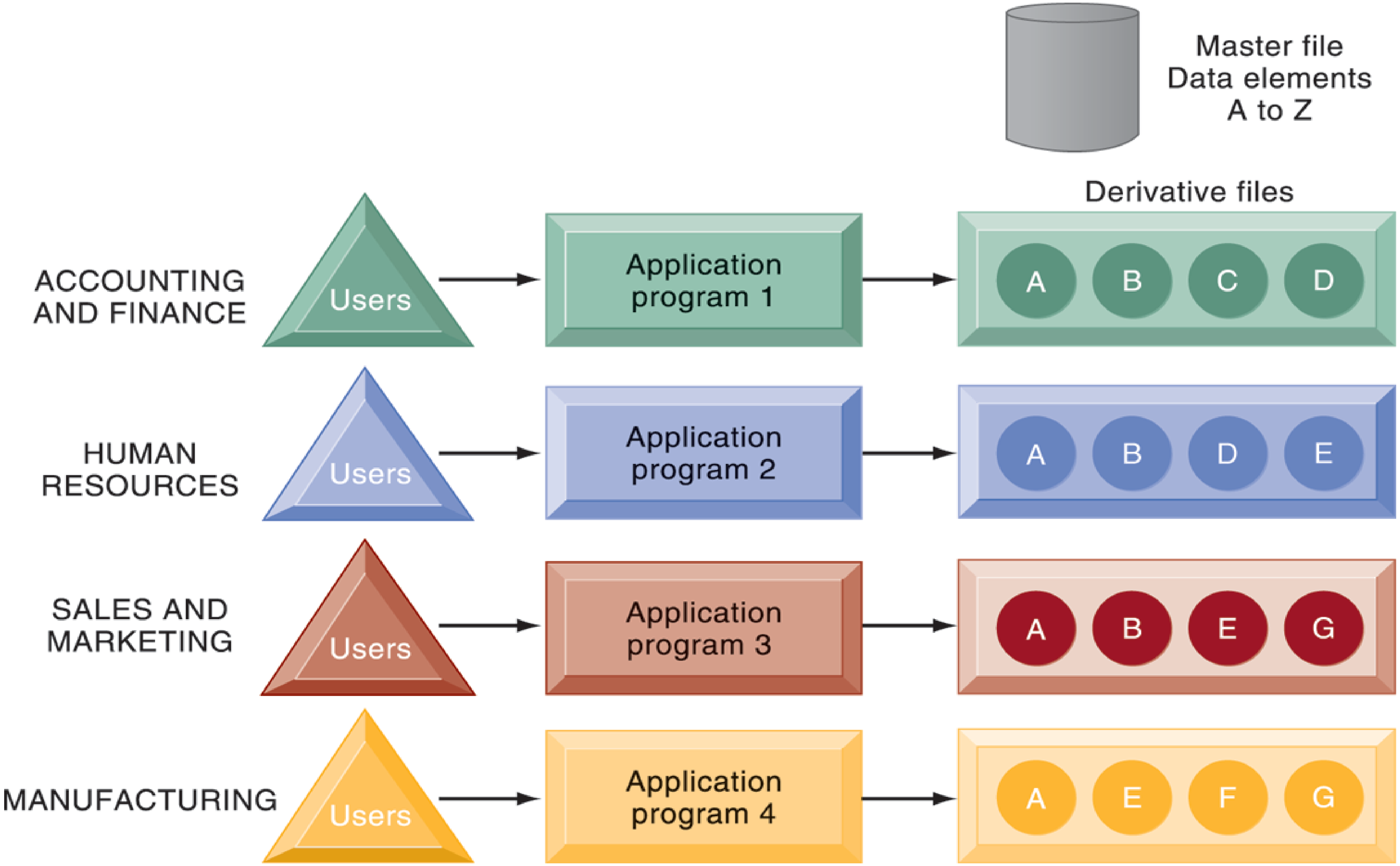
Managing Data in a Traditional File Environment

Problems with the traditional file environment (files maintained separately by different departments)

- **Data redundancy:**
  - Presence of duplicate data in multiple files
- **Data inconsistency:**
  - Same attribute has different values
- **Program-data dependence:**
  - When changes in program requires changes to data accessed by program
- **Lack of flexibility**
- **Poor security**
- **Lack of data sharing and availability**

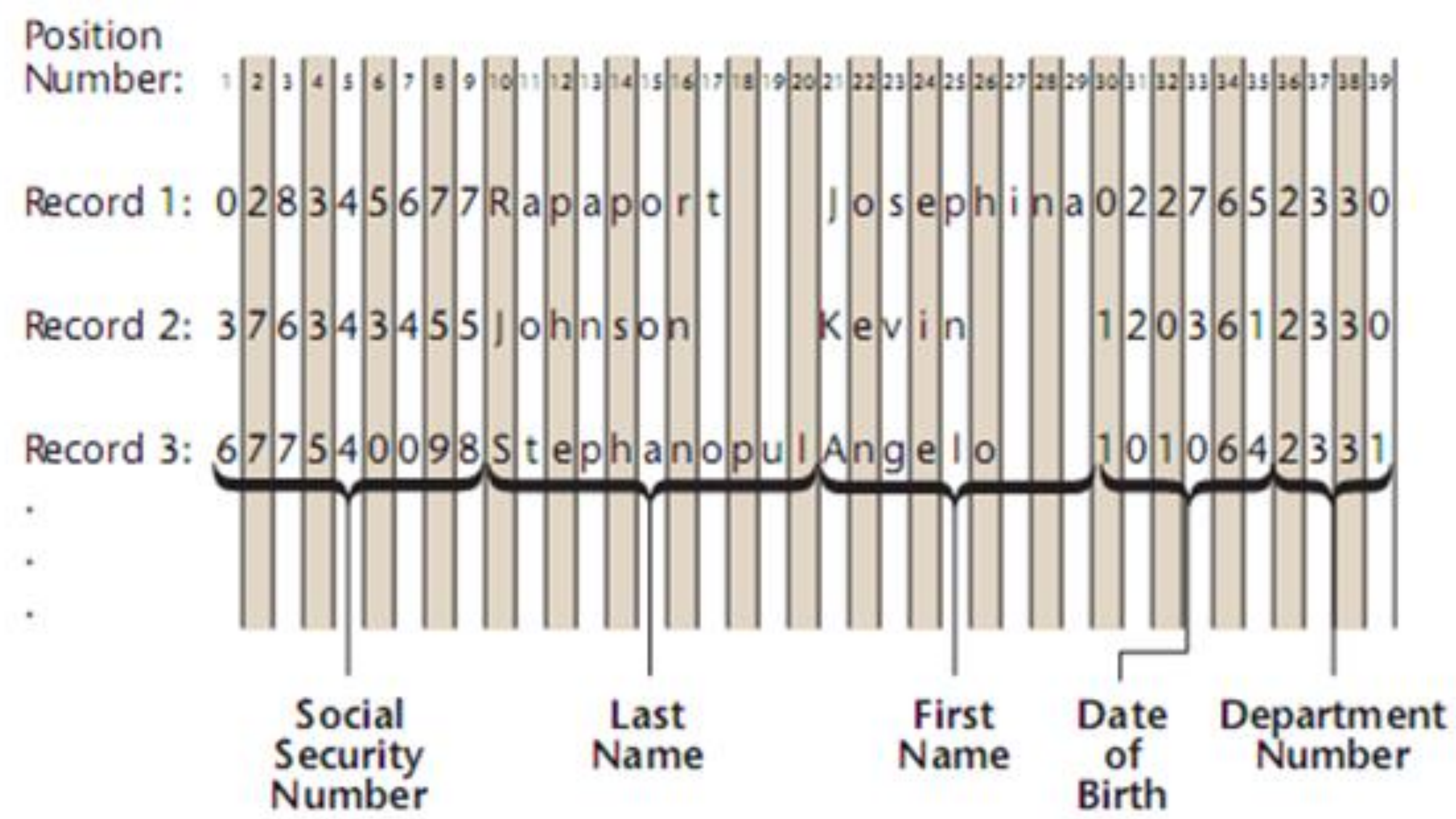
TRADITIONAL FILE PROCESSING

The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications. Each application requires a unique data file that is likely to be a subset of the master file. These subsets of the master file lead to data redundancy and inconsistency, processing inflexibility, and wasted storage resources.



TRADITIONAL FILE APPROACH

Information file in Human Resources Department



Position Number: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

Record 1: 028345677Rapaport Josephina0227652330

Record 2: 376343455Johnson Kevin 1203612330

Record 3: 677540098StephanopulAngelo 1010642331

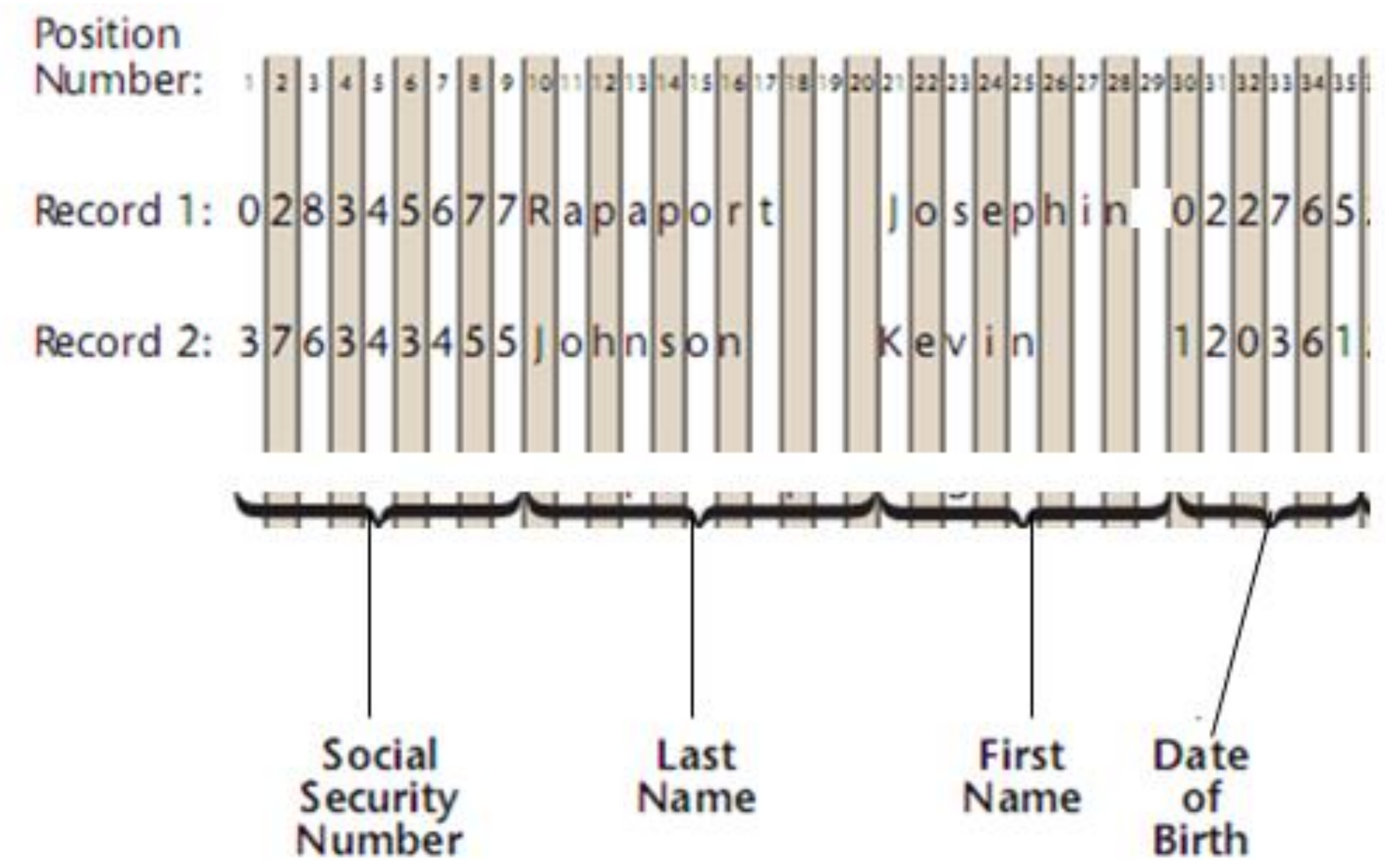
.

.

.

TRADITIONAL FILE APPROACH

Information file in Marketing Department



Position Number: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

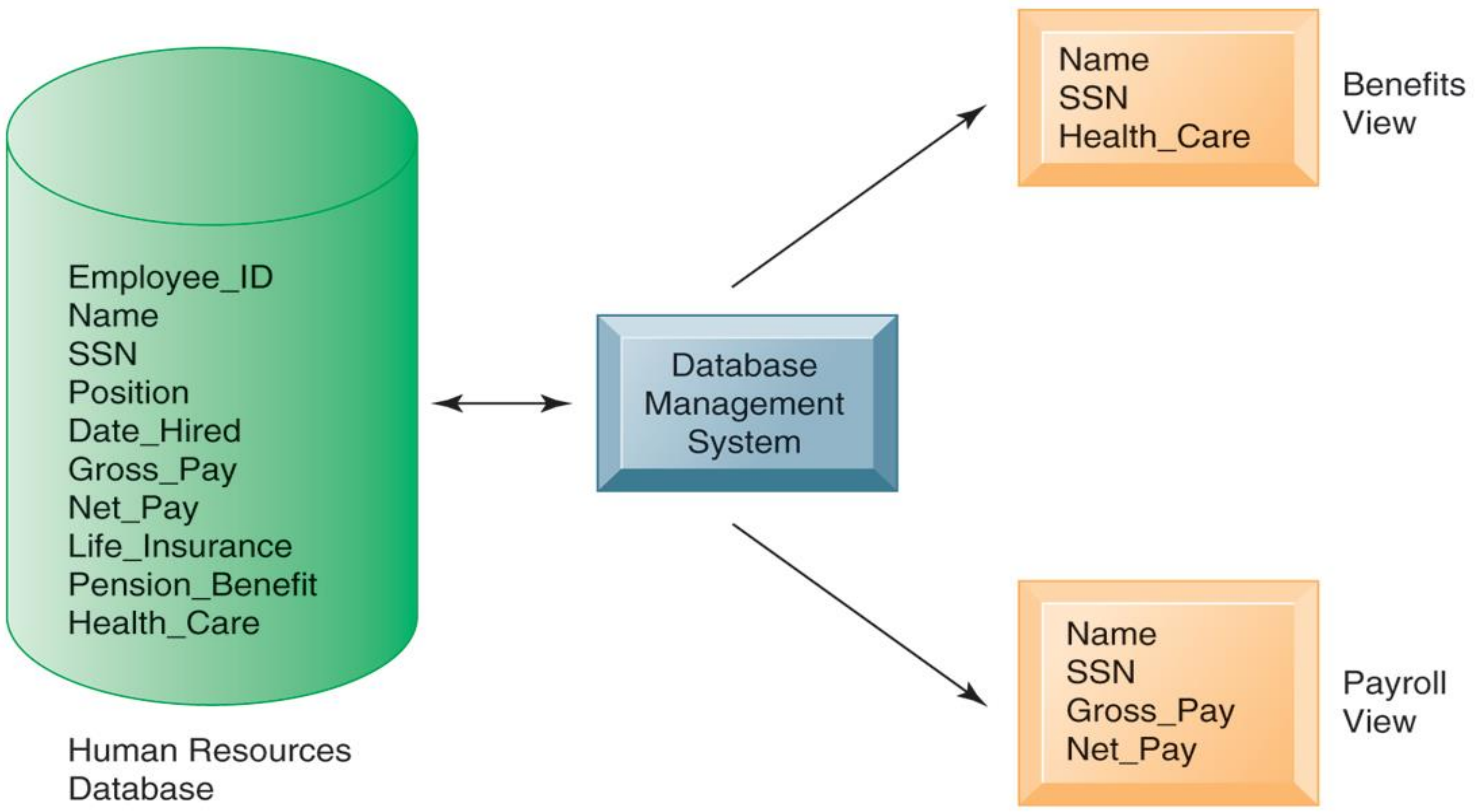
Record 1: 0 2 8 3 4 5 6 7 7 R a p a p o r t J o s e p h i n 0 2 2 7 6 5

Record 2: 3 7 6 3 4 3 4 5 5 J o h n s o n K e v i n 1 2 0 3 6 1

Capabilities of Database Management Systems (DBMSs)

- Database
  - Serves many applications by centralizing data and controlling redundant data
- Database management system (DBMS)
  - Interfaces between applications and physical data files
  - Separates logical and physical views of data
  - Solves problems of traditional file environment
    - Controls redundancy
    - Eliminates inconsistency
    - Uncouples programs and data
    - Enables organization to central manage data and data security

HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS

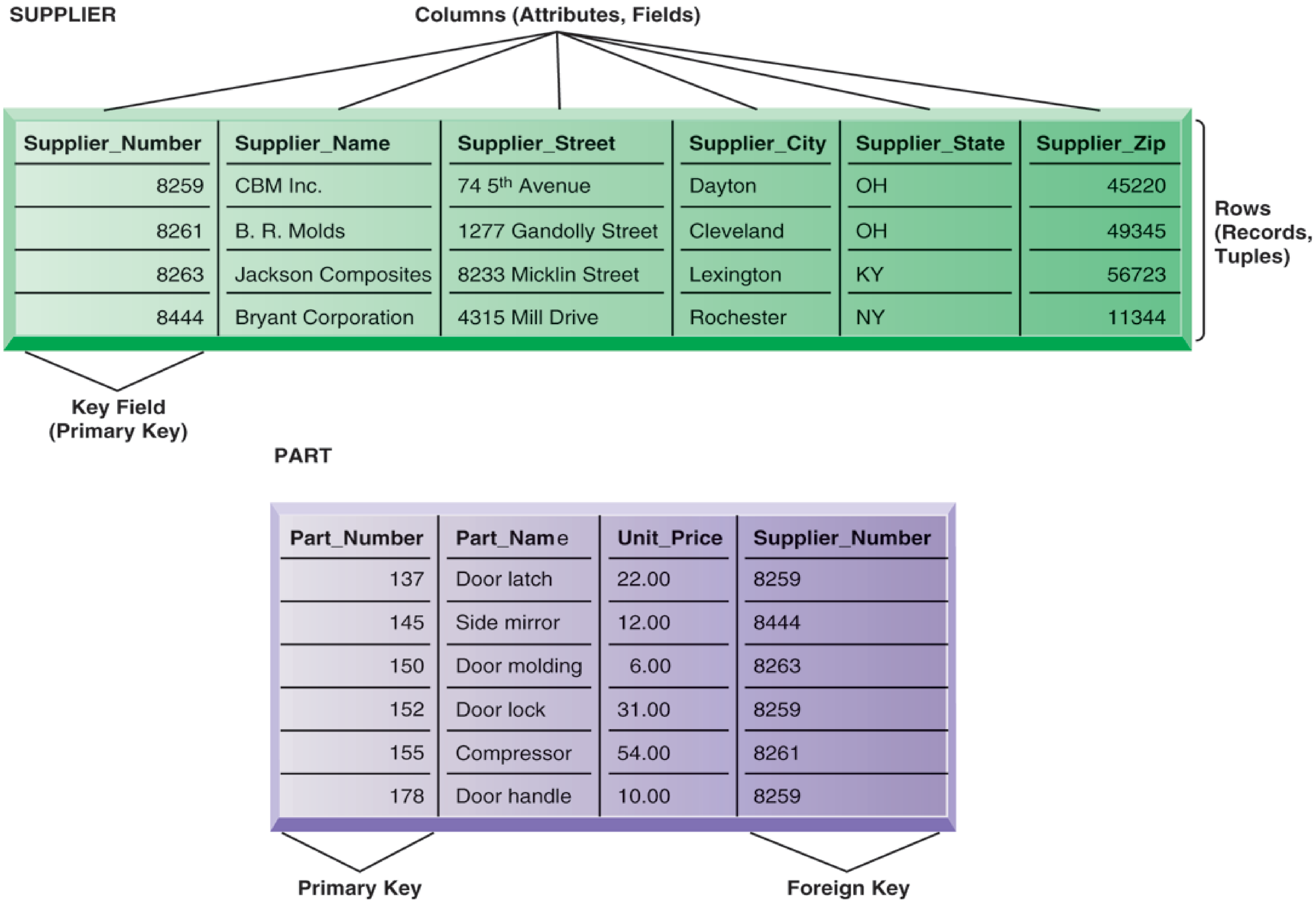


A single human resources database provides many different views of data, depending on the information requirements of the user. Illustrated here are two possible views, one of interest to a benefits specialist and one of interest to a member of the company's payroll department.

CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS (DBMSS)

- **Relational DBMS**
  - Represent data as two-dimensional tables
  - Each table contains data on entity and attributes
  - Oracle, IBM DB2, Microsoft SQL Server 2014, Microsoft Access, etc.
- **Table: grid of columns and rows**
  - Rows (tuples): Records for different entities
  - Fields (columns): Represents attribute for entity
  - Key : A field or a combination of fields that **can be** used to uniquely identify each record. There can be more than one key (e.g. studentID, HKID)
  - Primary key: A key **selected** to uniquely identify each record. It **must** have a value in a record
  - Foreign key: Primary key used in second table as look-up field to identify records from original table

RELATIONAL DATABASE TABLES



SUPPLIER

Columns (Attributes, Fields)

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Rows (Records, Tuples)

Key Field (Primary Key)

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Primary Key

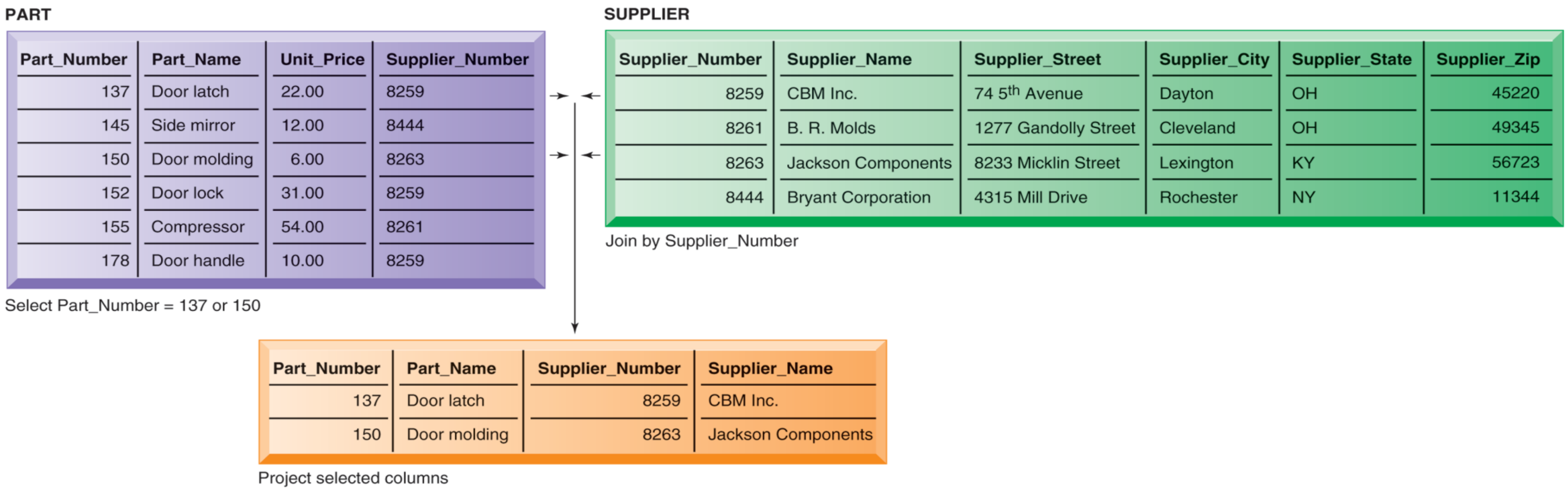
Foreign Key

A relational database organizes data in the form of two-dimensional tables. Illustrated here are tables for the entities SUPPLIER and PART showing how they represent each entity and its attributes. Supplier Number is a primary key for the SUPPLIER table and a foreign key for the PART table.

Capabilities of Database Management Systems (DBMSs)

- Operations of a Relational DBMS
  - Three basic operations used to develop useful sets of data
    - **SELECT:** Creates subset of data of all records that meet stated criteria
    - **JOIN:** Combines relational tables to provide user with more information than available in individual tables
    - **PROJECT:** Creates subset of columns in table, creating tables with only the information specified

THE THREE BASIC OPERATIONS OF A RELATIONAL DBMS



PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Select Part_Number = 137 or 150

SUPPLIER

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Components	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Join by Supplier_Number

Part_Number	Part_Name	Supplier_Number	Supplier_Name
137	Door latch	8259	CBM Inc.
150	Door molding	8263	Jackson Components

Project selected columns

The select, join, and project operations enable data from two different tables to be combined and only selected attributes to be displayed.

Capabilities of Database Management Systems (DBMSs)

- Non-relational databases: "NoSQL"
  - More flexible data model
  - Data sets stored across distributed machines
  - Easier to scale
  - Handle large volumes of unstructured and structured data (Web, social media, graphics)
- Databases in the cloud
  - Typically, less functionality than on-premises DBs
  - Amazon Relational Database Service, Microsoft SQL Azure
  - Private clouds

CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS (DBMSs)

- Capabilities of database management systems
  - Data definition capability: Specifies structure of database content, used to create tables and define characteristics of fields
  - Data dictionary: Automated or manual file storing definitions of data elements and their characteristics
  - Data manipulation language: Used to add, change, delete, retrieve data from database
    - Structured Query Language (SQL)
    - Microsoft Access user tools for generating SQL
  - Many DBMS have report generation capabilities for creating polished reports (Crystal Reports)

Capabilities of Database Management Systems (DBMSs)

- **Designing Databases**
  - Conceptual (logical) design: abstract model from business perspective
  - Physical design: How database is arranged on direct-access storage devices
- **Design process identifies:**
  - Relationships among data elements, redundant database elements
  - Most efficient way to group data elements to meet business requirements, needs of application programs
- **Normalization**
  - Streamlining complex groupings of data to minimize redundant data elements and awkward many-to-many relationships

CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS (DBMSS)

- **Referential integrity rules**
  - Used by RDMS to ensure relationships between tables remain consistent
- **Entity-relationship diagram**
  - Used by database designers to document the data model
  - Illustrates relationships between entities

➢ **Caution: If a business doesn’t get data model right, system won’t be able to serve business well**

Tools for Improving Business Performance and Decision Making

- Big data
  - Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, and so on
  - Petabytes, exabytes of data
    - Volumes too great for typical DBMS
  - Can reveal more patterns and anomalies

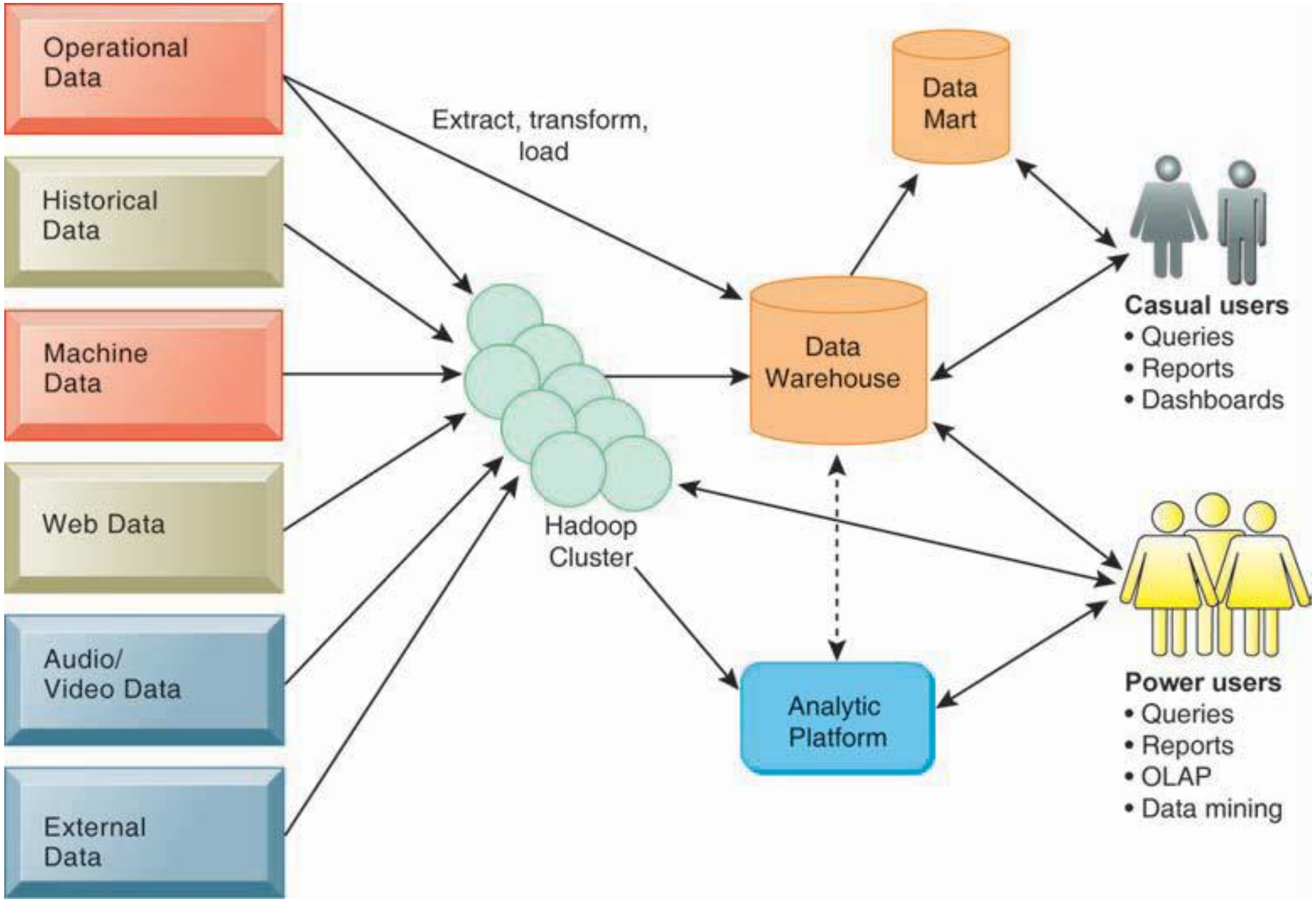
Tools for Improving Business Performance and Decision Making

- **Business intelligence infrastructure**
  - Today includes an array of tools for separate systems, and big data
- **Contemporary tools:**
  - Data warehouses
  - Data marts
  - Hadoop
  - In-memory computing
  - Analytical platforms

Tools for Improving Business Performance and Decision Making

- **Data warehouse:**
  - Stores current and historical data from many core operational transaction systems
  - Consolidates and standardizes information for use across enterprise, but data cannot be altered
  - Provides analysis and reporting tools
- **Data marts:**
  - Subset of data warehouse
  - Summarized or focused portion of data for use by specific population of users
  - Typically focuses on single subject or line of business

CONTEMPORARY BUSINESS INTELLIGENCE INFRASTRUCTURE



A contemporary business intelligence infrastructure features capabilities and tools to manage and analyze large quantities and different types of data from multiple sources. Easy-touse query and reporting tools for casual business users and more sophisticated analytical toolsets for power users are included.

- **Hadoop**
  - **Enables distributed parallel processing of big data across inexpensive computers**
  - **Key services**
    - Hadoop Distributed File System (HDFS): data storage
    - MapReduce: breaks data into clusters for work
    - Hbase: NoSQL database
  - **Used by Facebook, Yahoo, NextBio**

Tools for Improving Business Performance and Decision Making

- **In-memory computing**
  - Used in big data analysis
  - Uses computers main memory (RAM) for data storage to avoid delays in retrieving data from disk storage
  - Can reduce hours/days of processing to seconds
  - Requires optimized hardware
- **Analytic platforms**
  - High-speed platforms using both relational and non-relational tools optimized for large datasets

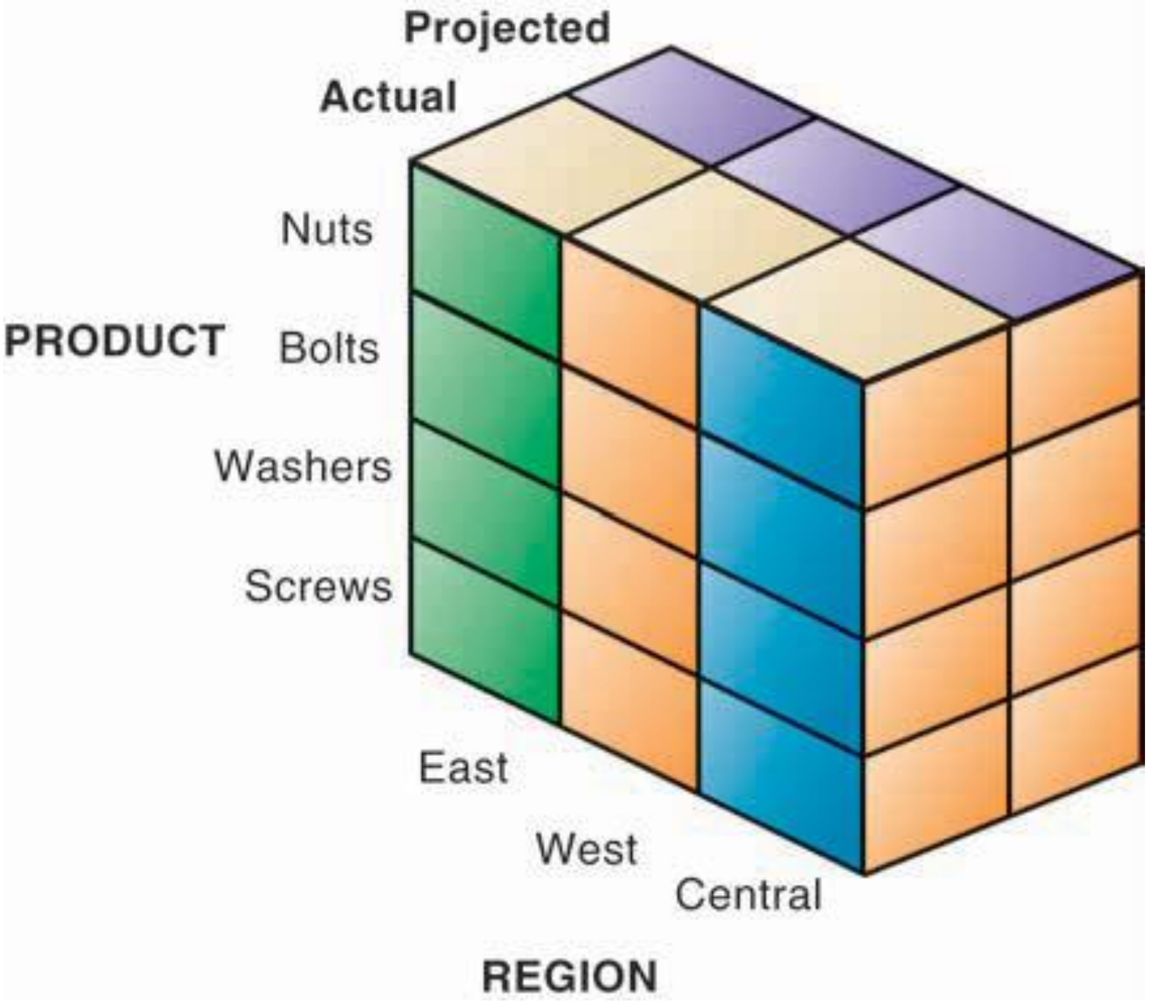
Tools for Improving Business Performance and Decision Making

- **Analytical tools: Relationships, patterns, trends**
  - Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
    - Multidimensional data analysis (OLAP)
    - Data mining
    - Text mining
    - Web mining

TOOLS FOR IMPROVING BUSINESS PERFORMANCE AND DECISION MAKING

- **Online analytical processing (OLAP)**
  - **Supports multidimensional data analysis**
    - Viewing data using multiple dimensions
    - Each aspect of information (product, pricing, cost, region, time period) is different dimension
    - Example: How many washers sold in the East in June compared with other regions?
  - **OLAP enables rapid, online answers to ad hoc queries**

MULTIDIMENSIONAL DATA MODEL



This view shows product versus region. If you rotate the cube 90 degrees, the face that will show is product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. Other views are possible.

Tools for Improving Business Performance and Decision Making

- Data mining:
  - Finds hidden patterns, relationships in datasets
    - Example: customer buying patterns
  - Infers rules to predict future behavior
  - More discovery driven than OLAP
  - Types of information obtainable from data mining:
    - Associations
    - Sequences
    - Classification
    - Clustering
    - Forecasting

Tools for Improving Business Performance and Decision Making

- **Text mining**
  - **Extracts key elements from large unstructured data sets**
    - Stored e-mails
    - Call center transcripts
    - Legal cases
    - Patent descriptions
    - Service reports, and so on
  - **Sentiment analysis software**
    - Mines e-mails, blogs, social media to detect opinions

Tools for Improving Business Performance and Decision Making

- **Web mining**
  - **Discovery and analysis of useful patterns and information from Web**
    - Understand customer behavior
    - Evaluate effectiveness of Web site, and so on
  - **Web content mining**
    - Mines content of Web pages
  - **Web structure mining**
    - Analyzes links to and from Web page
  - **Web usage mining**
    - Mines user interaction data recorded by Web server

Managing the Firm's Data Resources

- **Establishing an information policy**
  - **Firm's rules, procedures, roles for sharing, managing, standardizing data**
  - **Data administration**
    - Establishes policies and procedures to manage data
  - **Data governance**
    - Deals with policies and processes for managing availability, usability, integrity, and security of data, especially regarding government regulations
  - **Database administration**
    - Creating and maintaining database

- **Ensuring data quality**
  - **More than 25 percent of critical data in Fortune 1000 company databases are inaccurate or incomplete**
    - Redundant data
    - Inconsistent data
    - Faulty input
  - **Before new database in place, need to:**
    - Identify and correct faulty data
    - Establish better routines for editing data once database in operation

Managing the Firm's Data Resources

- **Data quality audit:**
  - **Structured survey of the accuracy and level of completeness of the data in an information system**
    - Survey samples from data files, or
    - Survey end users for perceptions of quality
- **Data cleansing**
    - Software to detect and correct data that are incorrect, incomplete, improperly formatted, or redundant
    - Enforces consistency among different sets of data from separate information systems

Source:

>> Management Information Systems, Managing the Digital Firm, 13 Edition (2014), Laudon and Laudon.