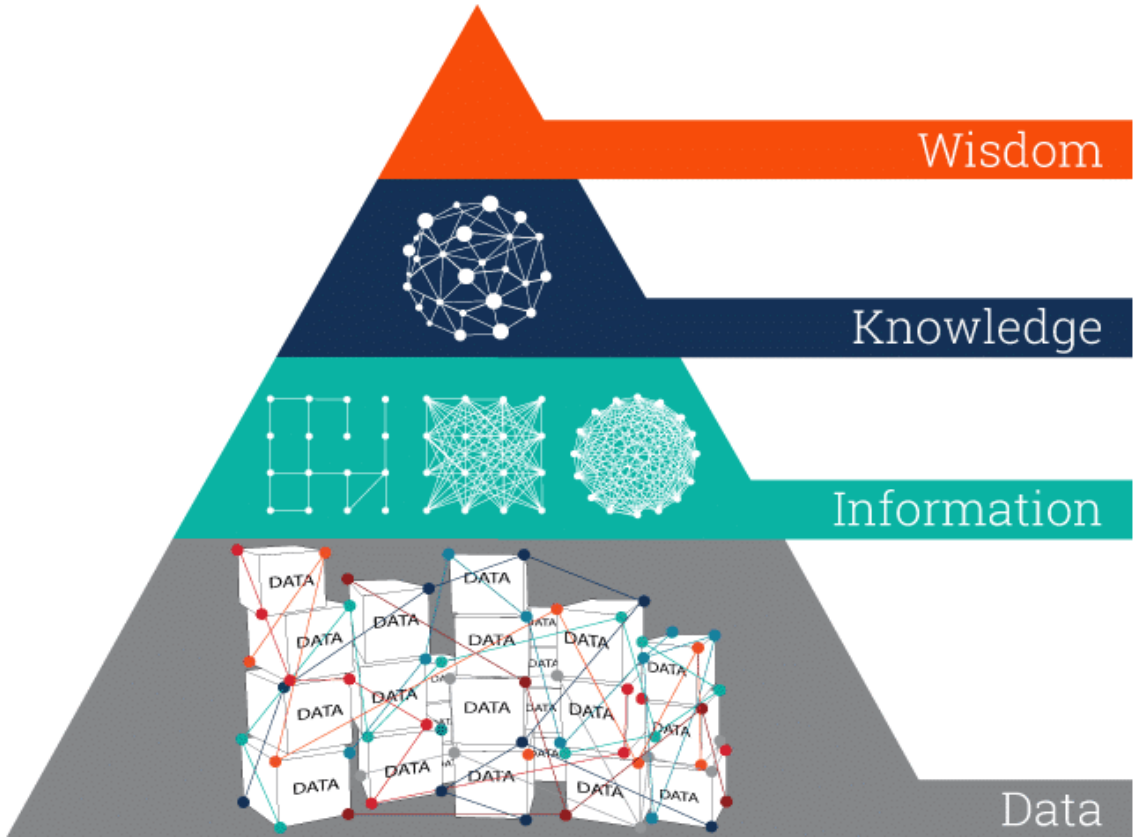
Data Analysis

CS 341 Database Systems

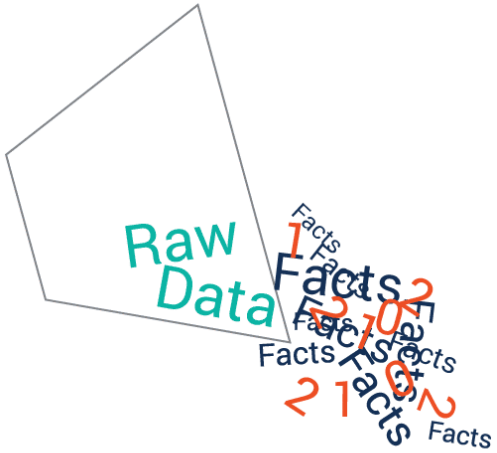
"The goal is to transform data into information, and information into insight"

Carly Fiorina (Executive and president of Hewlett-Packard Co. in 1999. Chairwoman in 2000)

D-I-K-W Pyramid

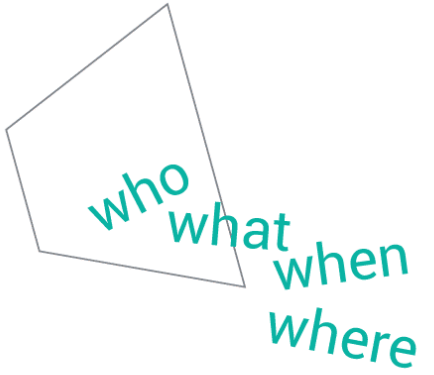


Each step up the pyramid answers questions about and adds value to the initial data.



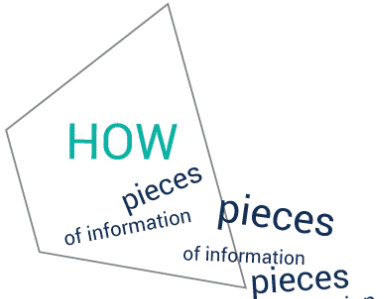
= a collection of facts in a raw or unorganized form

Base building block - Raw Data



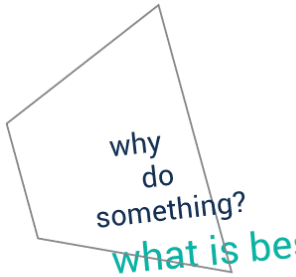
= easier to measure, visualize and analyze data for a specific purpose

Second building block - Derived Information



connected to other pieces help us understand how to apply information to achieve our goal

Third building block - Relevant Knowledge



Wisdom is knowledge applied in action

The top of the DIKW hierarchy - Guiding Wisdom

Need for Data Analysis

- By tapping into the operational database, management can develop an understanding of how the company is performing and evaluate whether the current strategies meet organizational goals.
- Analyzing the company data can provide insightful information about short term tactical evaluations and strategic questions, such as:
  - *Are our sales promotions working?*
  - *What market percentage are we controlling?*
  - *Are we attracting new customers?*

Decision Makers

- Decision makers can no longer wait a couple of days for a report to be generated; they are compelled to make quick decisions if they want to remain competitive.
- The key is in having the right data at the right time to support the **decision-making process**.

Data Driven Decision Making

Ask questions and define the problem.

Prepare data by collecting and storing the information.

Process data by cleaning and checking the information.

Analyze data to find patterns, relationships, and trends.

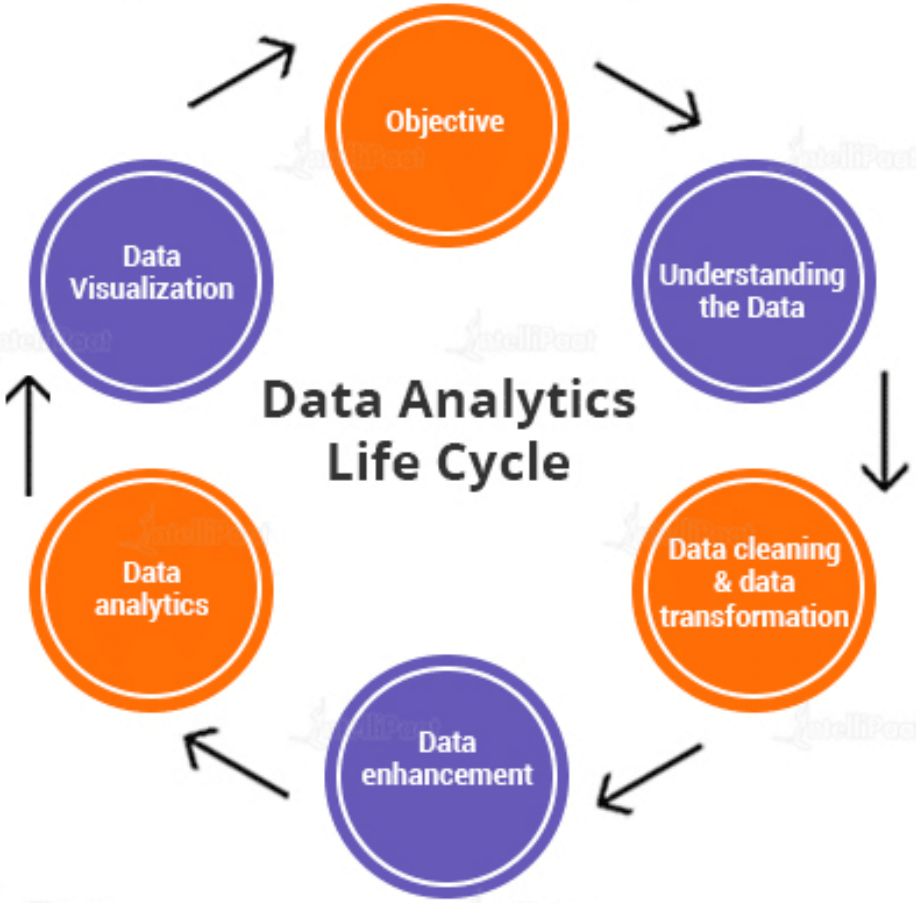
Share data with your audience.

Act on the data and use the analysis results.

Business Problems

- **Profit/ Loss Analysis:** What are the key reasons of losses in the business and how do we fix them?
- **Devising Marketing Strategy:** Who are my customers and how do I design targeted marketing campaigns for each segment?
- **Employee Performance Analysis:** How are my employees performing? What remedies are needed to improve the performance and productivity?
- What else?
- Sales Analysis, Quality Analysis, Demand Forecasting Analysis, Supply chain efficiency, Inventory Management, etc.

Data Analytics Lifecycle



Analyze, Report and Present

- Goal is to effectively communicate to the business executives.
- How do we do it?
  - Create effective data visualizations.
  - Design comprehensive dashboards.
  - Communicate your findings.

Business Intelligence (BI)

Business intelligence (BI) is a term that describes a comprehensive, cohesive, and integrated *set of tools and processes* used to capture, collect, integrate, store, and analyze **data** with the purpose of generating and presenting information to **support business decision making.**

Unlocking Wisdom

- BI is a *framework* that allows a business to transform data into information, information into knowledge, and knowledge into wisdom.
- BI is a complex proposition that requires a deep understanding and alignment of the business processes, business data, and information needs of users at all levels in an organization.
- BI is not a product by itself, but a framework of concepts, practices, tools, and technologies that help a business
  - better understand its core capabilities,
  - provide snapshots of the company situation,
  - and identify key opportunities to create competitive advantage

Figure 13.1 Business Intelligence Framework

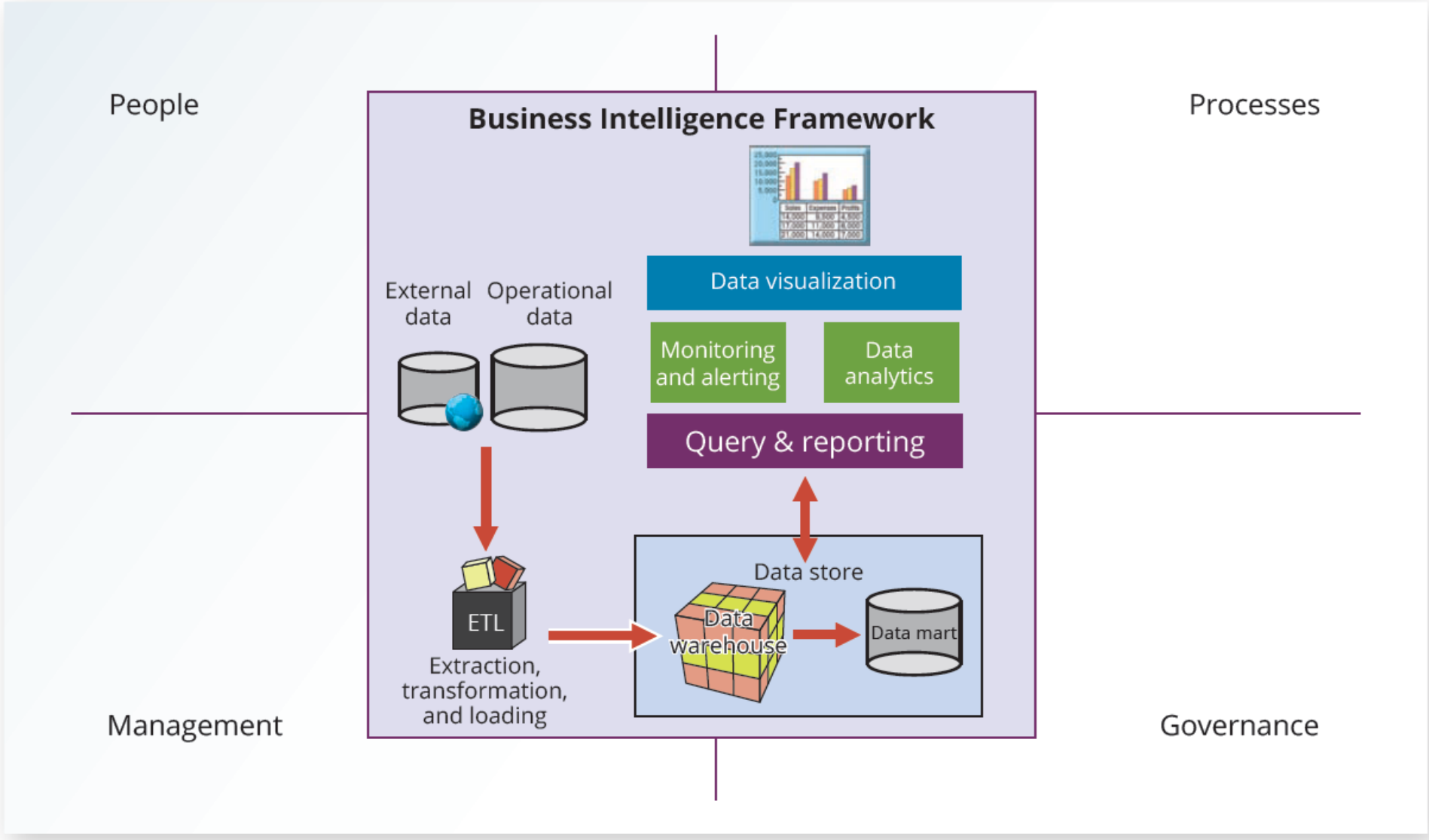


Figure 13.2 Evolution of BI Information Dissemination Formats

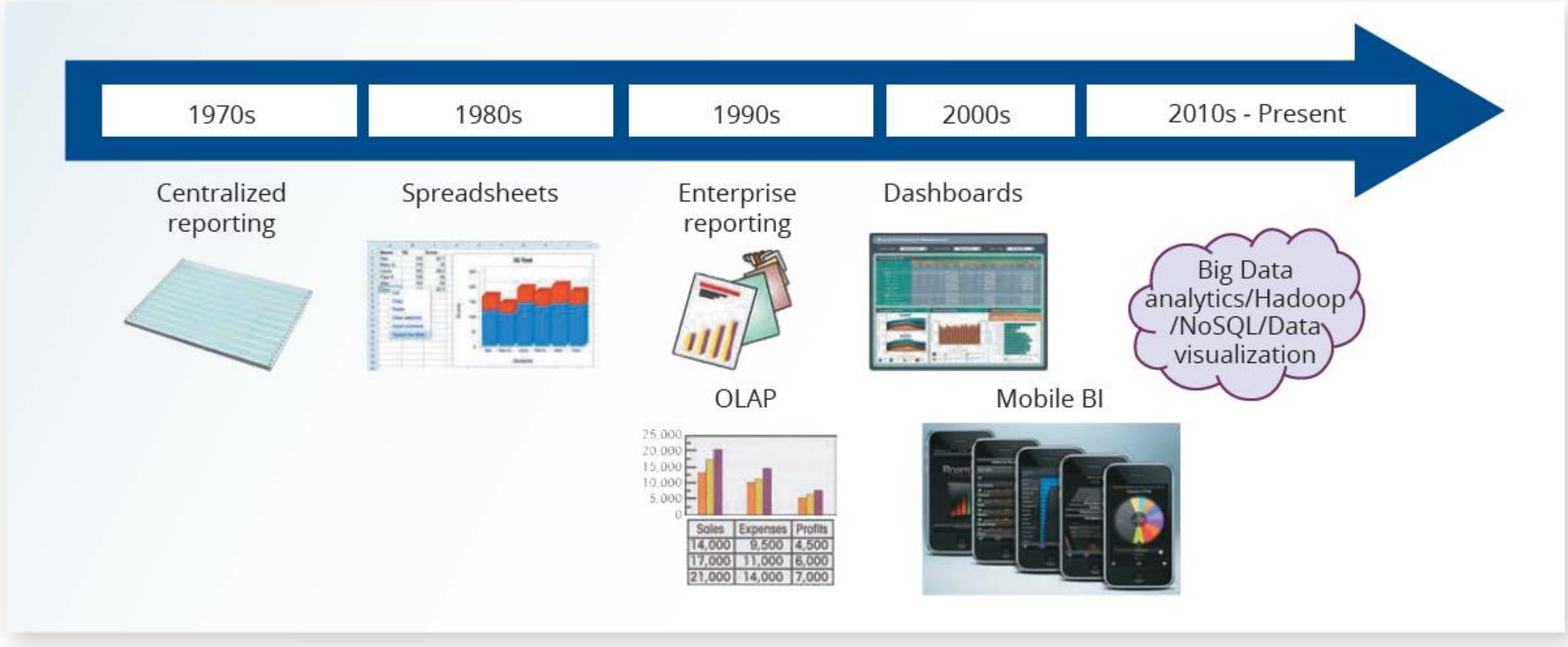
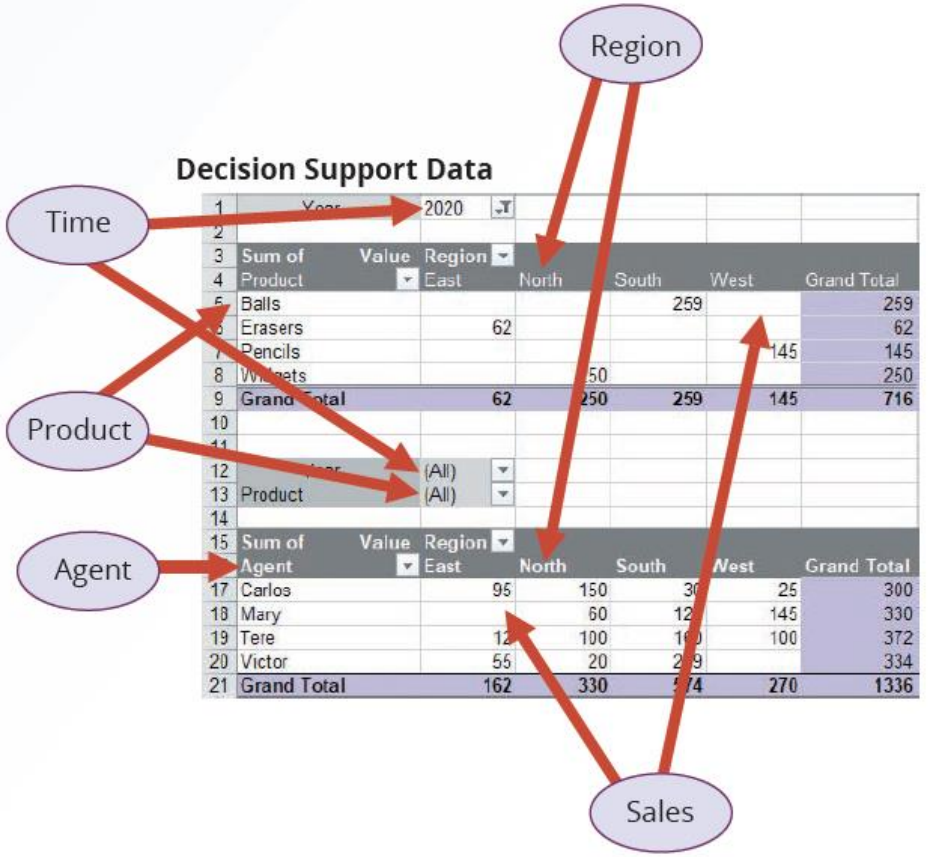


Figure 13.3 Transforming Operational Data Into Decision Support Data

Operational Data

	Year	Region	Agent	Product	Value
2	2020	East	Carlos	Erasers	50
3	2020	East	Tere	Erasers	12
4	2020	North	Carlos	Widgets	120
5	2020	North	Tere	Widgets	100
6	2020	North	Carlos	Widgets	30
7	2020	South	Victor	Balls	145
8	2020	South	Victor	Balls	34
9	2020	South	Victor	Balls	80
10	2020	West	Mary	Pencils	89
11	2020	West	Mary	Pencils	56
12	2021	East	Carlos	Pencils	45
13	2021	East	Victor	Balls	55
14	2021	North	Mary	Pencils	60
15	2021	North	Victor	Erasers	20
16	2021	South	Carlos	Widgets	30
17	2021	South	Mary	Widgets	75
18	2021	South	Mary	Widgets	50
19	2021	South	Tere	Balls	70
20	2021	South	Tere	Erasers	90
21	2021	West	Carlos	Widgets	25
22	2021	West	Tere	Balls	100

Operational data has a narrow time span, low granularity, and single focus. Such data is usually represented in tabular format, in which each row represents a single transaction. This format often makes it difficult to derive useful information.



Decision support system (DSS) data focuses on a broader time span, tends to have high levels of granularity, and can be examined in multiple dimensions. For example, note these possible aggregations:

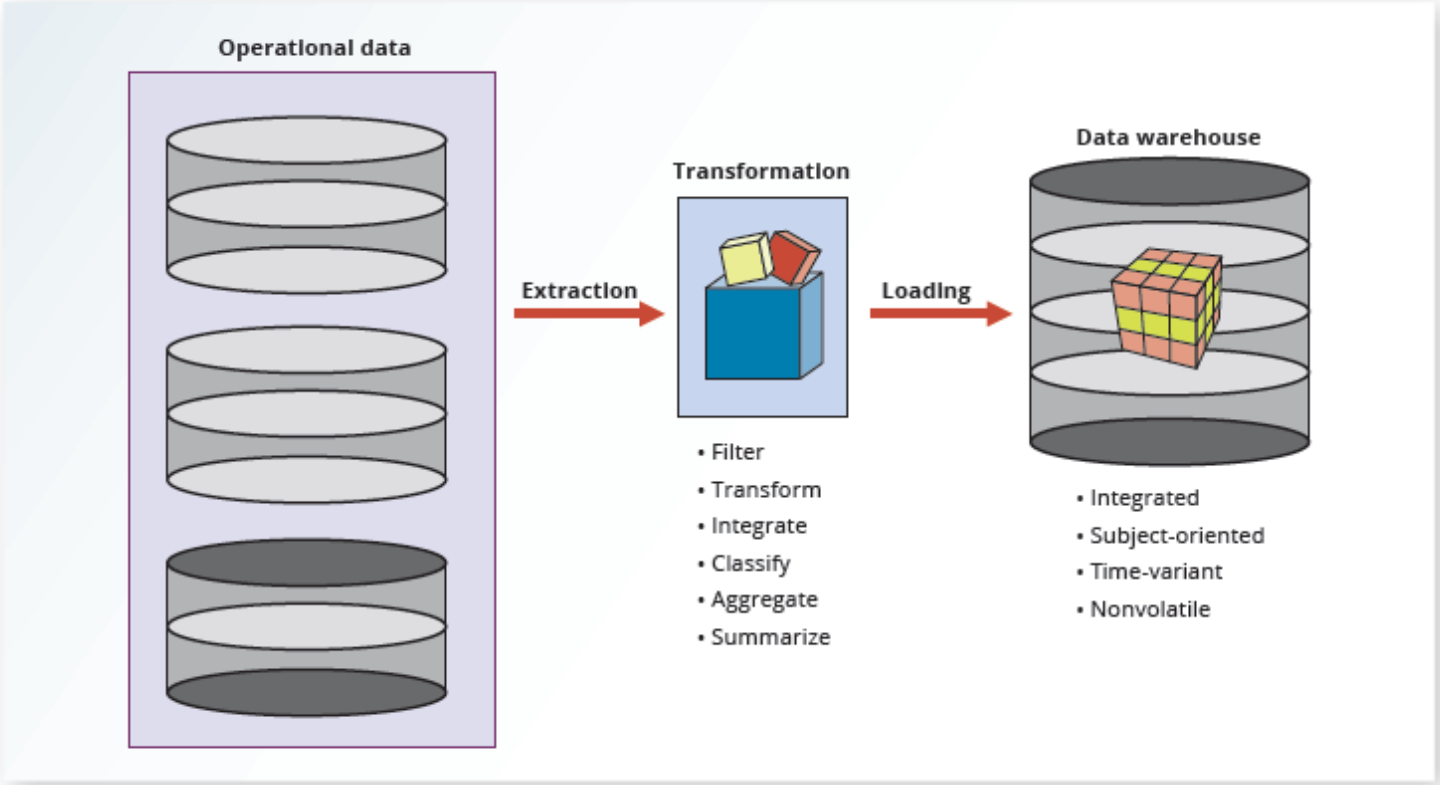
- Sales by product, region, agent, and so on
- Sales for all years or only a few selected years
- Sales for all products or only a few selected products

Data Warehouse

Bill Inmon, the acknowledged "*father*" of the data warehouse, defines the term as…

"an *integrated, subject-oriented, time-variant, nonvolatile* collection of data that provides support for decision making

Figure 13.4 The ETL Process



Integrated

Data has been standardized regardless of how it is stored in the source systems.

Subject-Oriented

Data collected relates to a particular subject (e.g sales, customer, etc.)

Data Warehouse

Time Variant

The data collected changes with time. Newer snapshots record the updates.

Non-Volatile

Data in the DW is hardly ever over-written or deleted - once committed, the data is static, read-only, and retained for future reporting

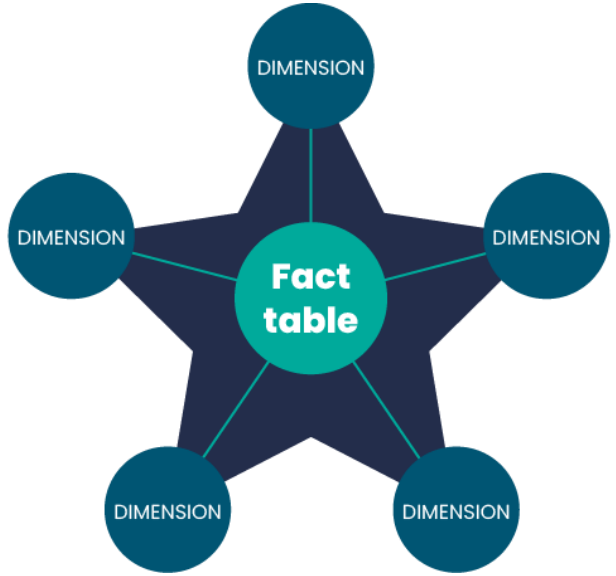
Defining Data Warehouse

Long Complicated JOIN Operations

- *"There was a **Colgate promotion** recently, directed to people who live in small towns. How much toothpaste did we sell in those towns yesterday? "*

```
select sum(sales.quantity_sold) from sales, products, product_categories, manufacturers,
stores, cities where manufacturer_name = 'Colgate'
and product_category_name = 'toothpaste'
and cities.population < 40 000
and trunc(sales.date_time_of_sale) = trunc(sysdate-1)
and sales.product_id = products.product_id
and sales.store_id = stores.store_id
and products.product_category_id = product_categories.product_category_
and products.manufacturer_id = manufacturers.manufacturer_id
and stores.city_id = cities.city_id
```

Analyzing Facts Across Dimensions



Restructuring as STAR SCHEMA

Figure 13.10 Star Schema for Sales

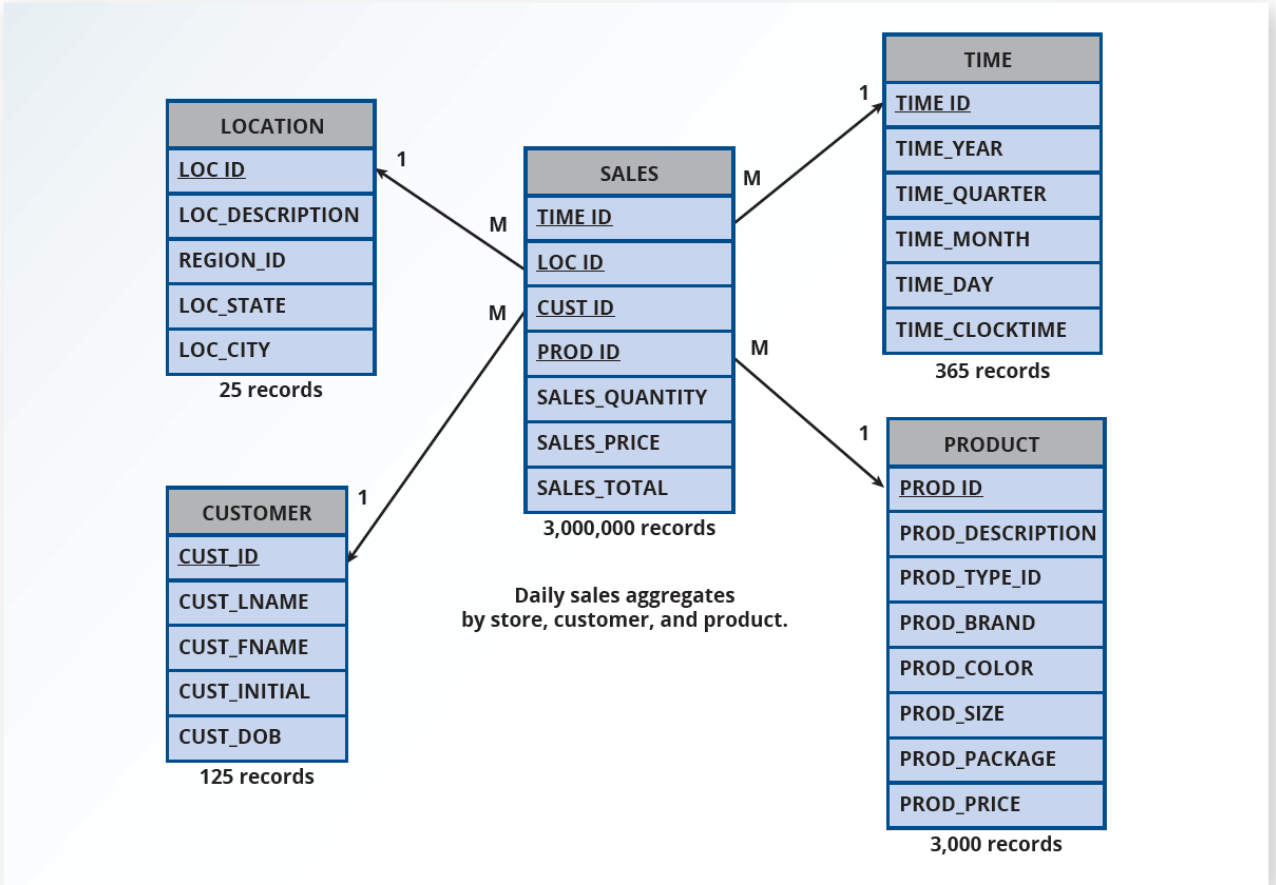
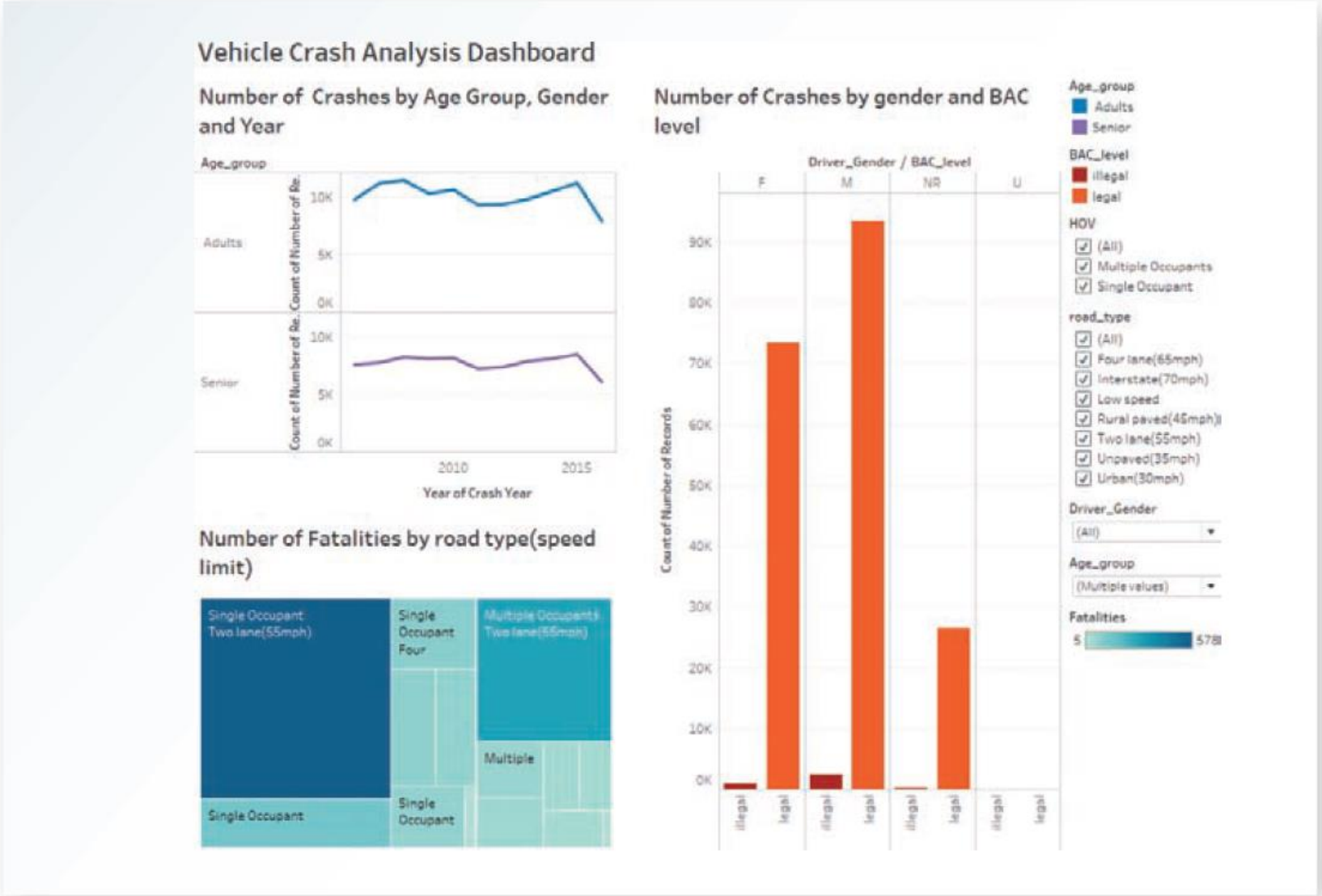
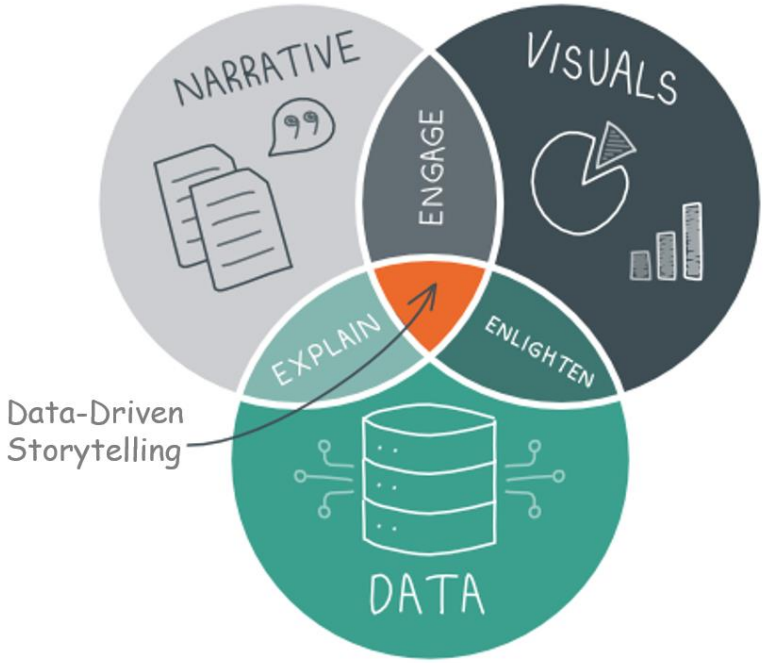


Figure 13.29 Vehicle Crash Analysis





Art of Story Telling

"...and the customers, shareholders, and employees lived happily ever after!"

- Data storytelling is the concept of building a compelling narrative based on complex data and analytics that help tell your story and influence and inform a particular audience.
- *Data storytelling is the best way to use data to create new knowledge and new decisions or actions.*

The End.

CS 341 – Database Systems