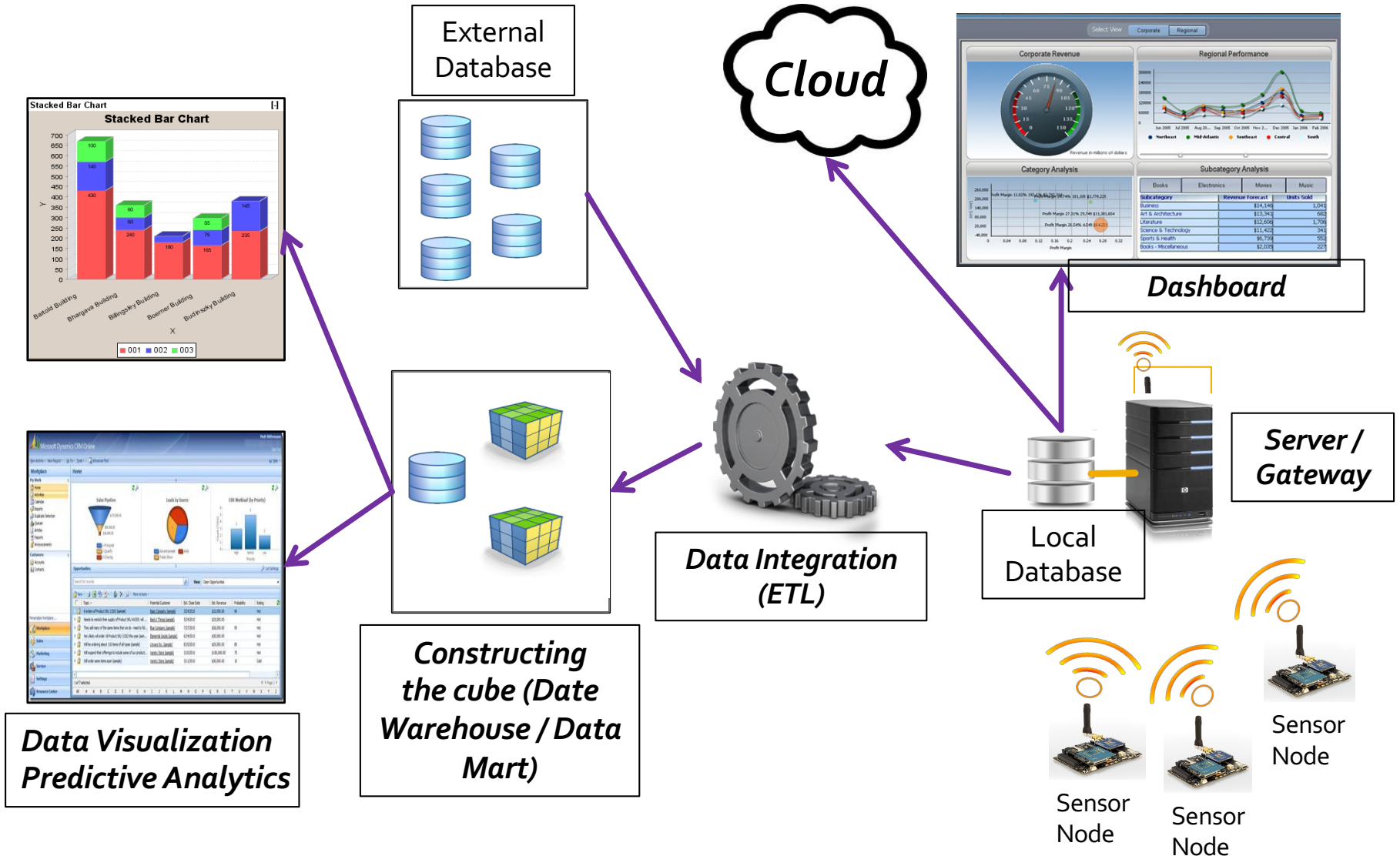
IT3779 Smart Object Technologies

L09: ETL, Data Warehouse and Mining
L10: Challenges and Emerging Trends

Objectives

1. Understand the need for multi-dimensional analysis
2. Define OLAP and its main characteristics
3. Define Data Warehouse (DW)
4. Describe the characteristics of data stored in DW
5. Differentiate between operational (OLTP) system and DW system
6. Distinguish between data warehouse and data mart
7. Outline the extraction, transformation, and loading/transportation (ETL) processes for building a data warehouse or data mart.
8. Future Trends

Overview IoT System Architecture



How business looks at info ?

- Need multi-dimensional view to answer complex questions like :

Marketing Vice President

How much did my new product generate

month by month, in the southern division, by user demographic, by sales office, relative to previous version, as compared to plan?

Marketing Manager

Give me sales statistics

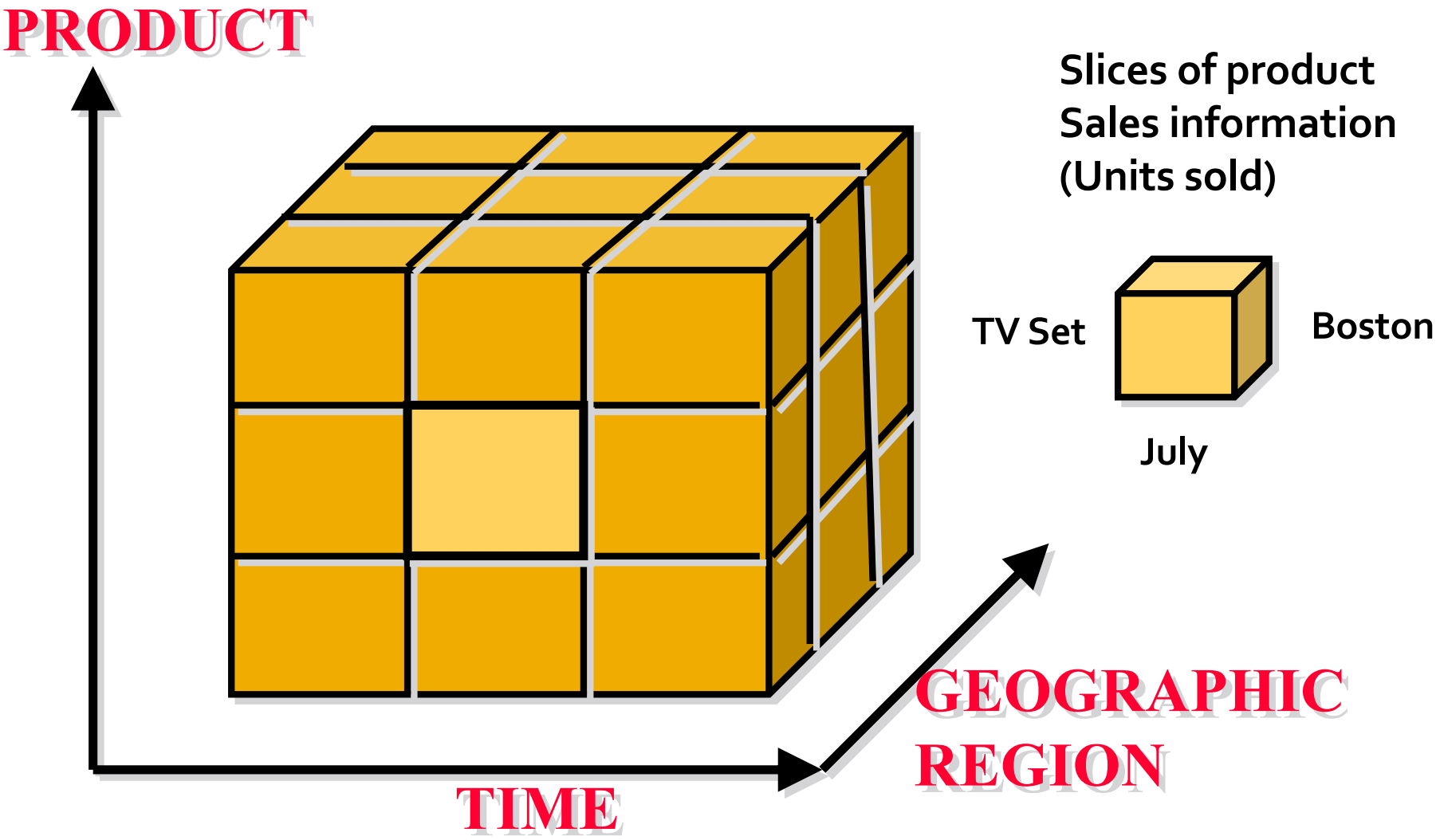
by products, summarized by product categories, daily, weekly, and monthly, by sale districts, by distribution channels.

Financial Controller

Show me expenses

listing actual vs. budget, by months, quarters, and annual, by budget line items, by district, division, summarized for the whole company.

Dimension Nature of Business Data



Multi-dimensional Concepts (1)

Sales Volume

Model	Color	Sales
Mini Van	Blue	6
Mini Van	Red	5
Mini Van	Gray	4
Sports Car	Blue	3
Sports Car	Red	5
Sports Car	Gray	5
Sedan	Blue	4
Sedan	Red	3
Sedan	Gray	2

Relational representation

Sales Volume

Model

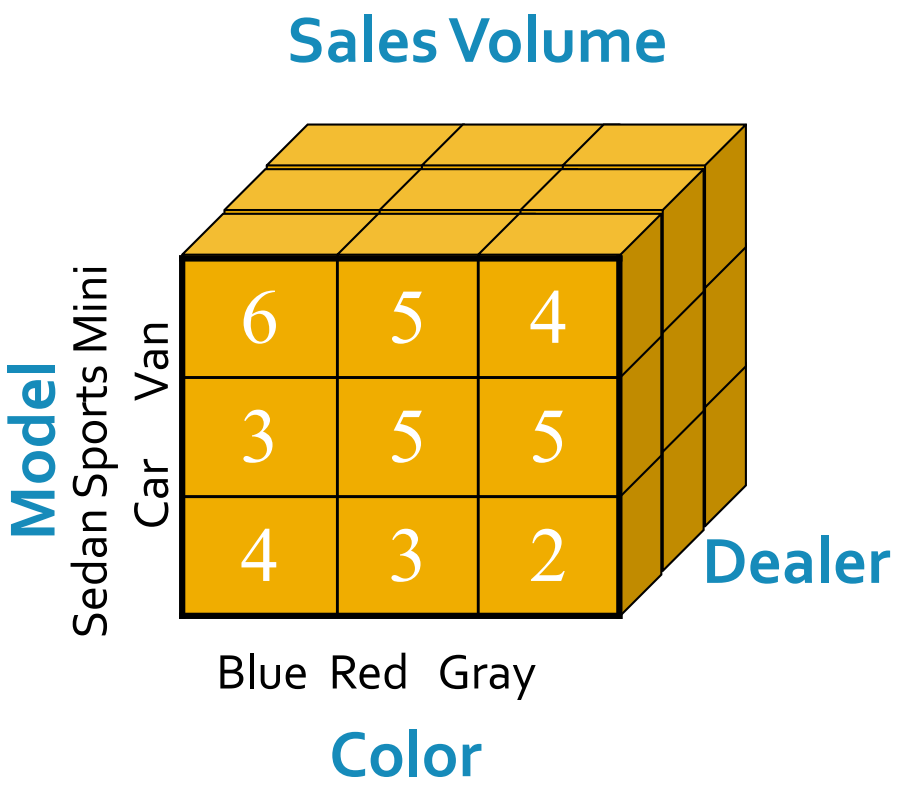
Sedan	6	5	4
Sports Car	3	5	5
Mini Van	4	3	2
	Blue	Red	Gray

Color

2-D representation

Multi-dimensional Concepts (2)

Model	Dealer	Color	Sales
Mini Van	SIN	Blue	6
Mini Van	SIN	Red	5
Mini Van	SIN	Gray	4
Mini Van	Penang	Blue	5
Mini Van	Penang	Red	5
Mini Van	Penang	Gray	5
Mini Van	KL	Blue	6
Mini Van	KL	Red	5
Mini Van	KL	Gray	4
Sports Car	SIN	Blue	3
Sports Car	SIN	Red	5
Sports Car	SIN	Gray	5
…	…	…	…



3-D representation

Online Analytical Processing (OLAP)

- The use of a set of graphical tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques.
- Major characteristics of OLAP are:
  - Examining many data items
  - Dealing with complex relationships
  - Looking for patterns, trends, exceptions
  - Allowing users flexibility in defining their questions

Data Warehouse Definition

A **Data Warehouse** is a collection of data in support of management's decisions.

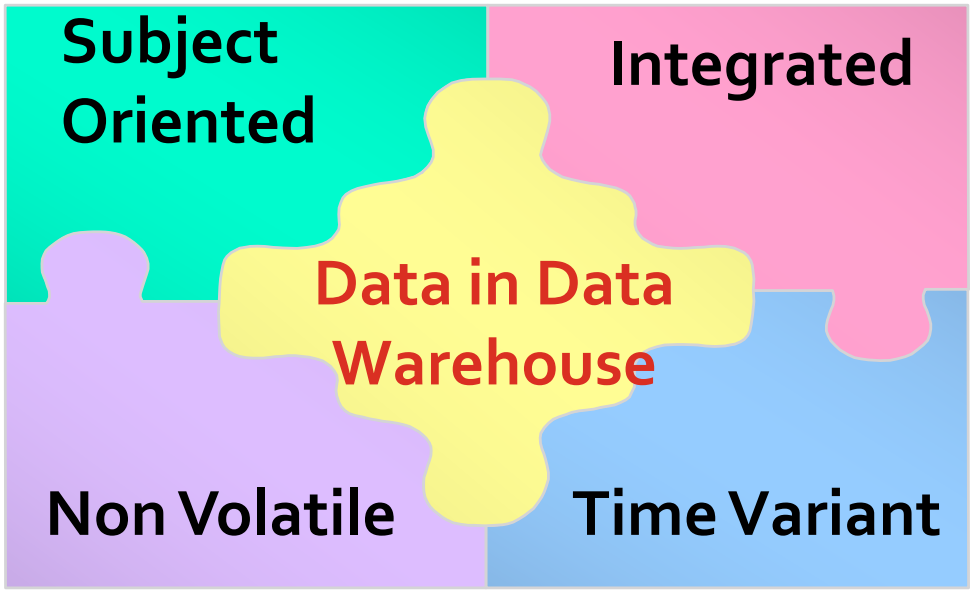
- It contains a lot of data
- It is organized for decision making purposes
- It provides tools for end users to access the data

Characteristics of Data in Data Warehouse

Data in Data Warehouse are structured and used for analytical processing activities.

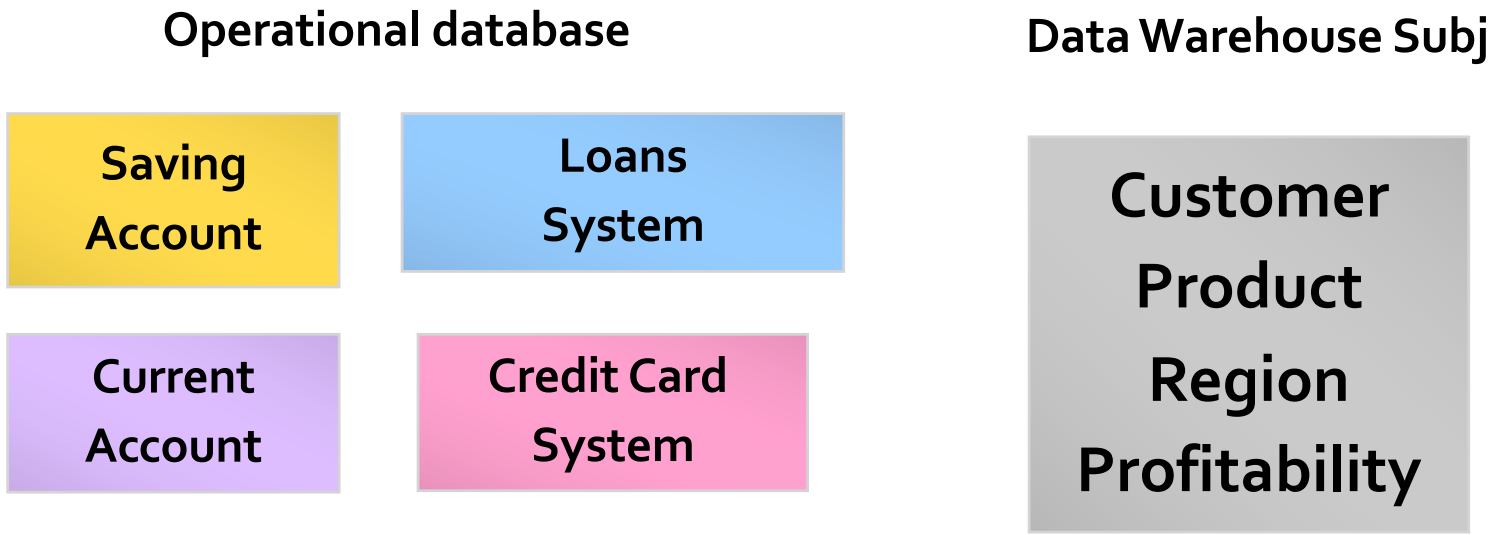
It has 4 characteristics:

- Subject-oriented
- Integrated
- Time-variant
- Nonvolatile



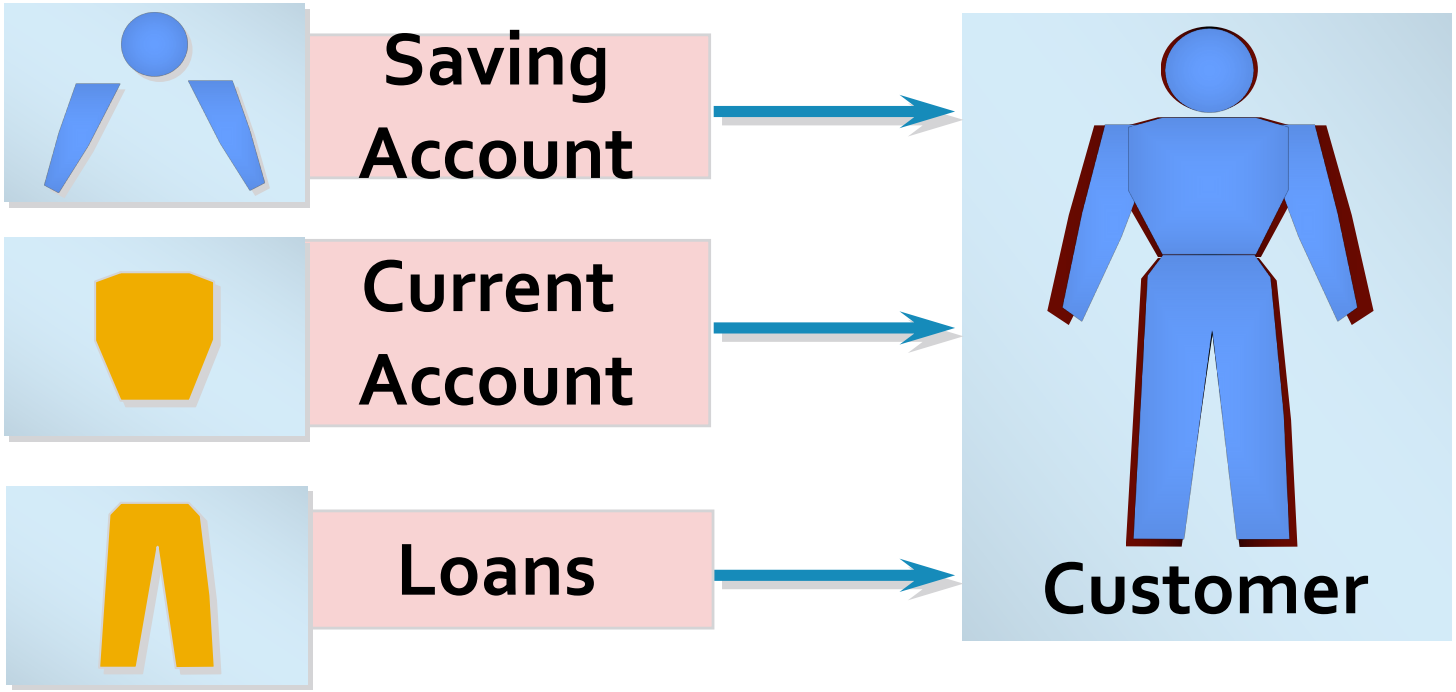
Subject-oriented Data

- Data in DW is organized by business subject areas, rather than by applications or functions.

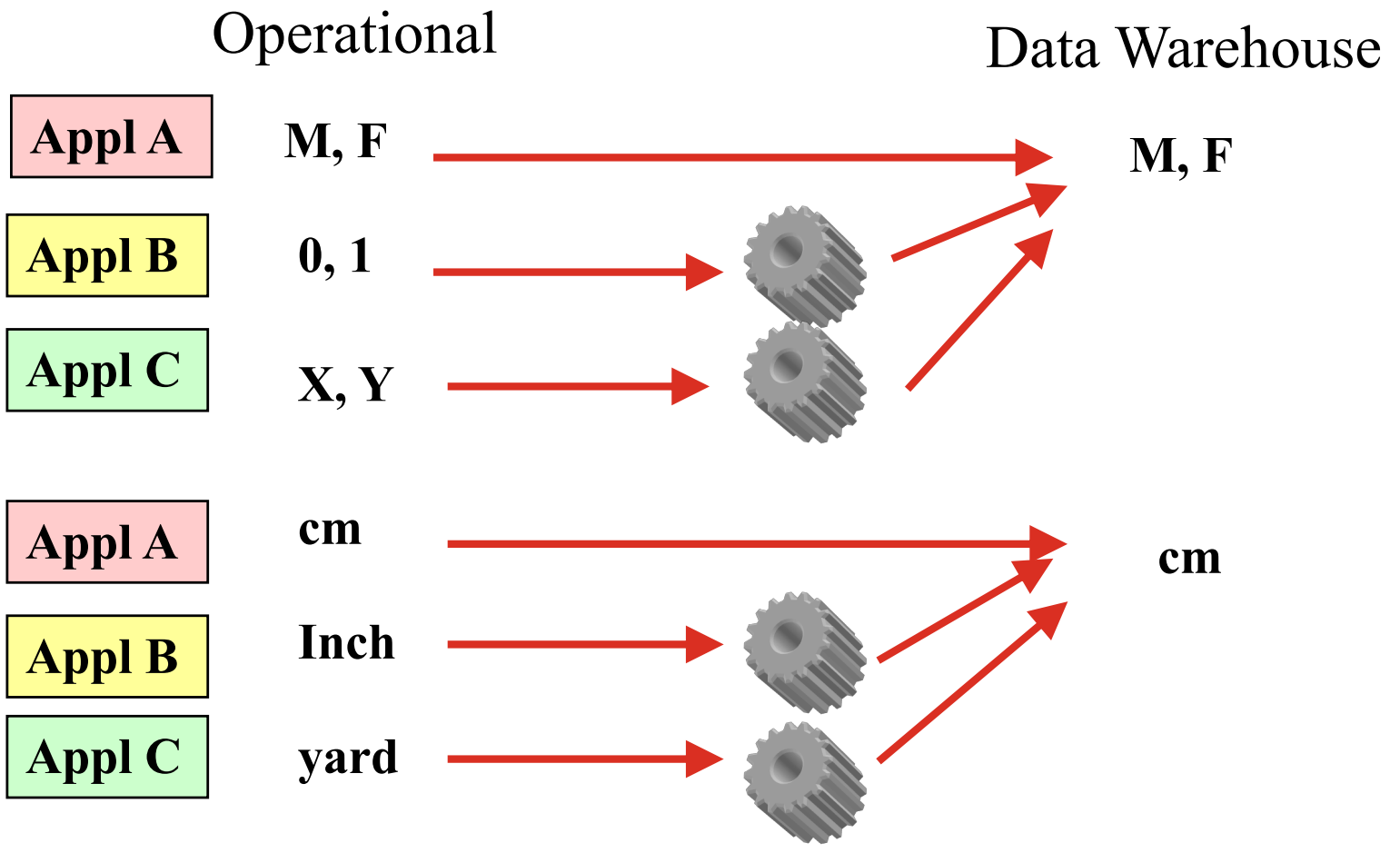


Integrated Data

- Data on a given subject is defined and stored once.

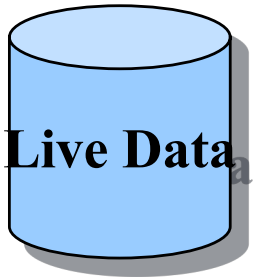


Examples of integration



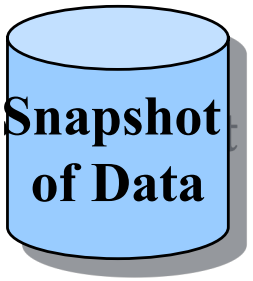
Time-Variant Data

- Data is stored as a series of snapshots, taken at some point in time.



Operational

- **Time horizon – current to 60-90 days**
- **Records can be updated**
- **Time may (not) be a key element**

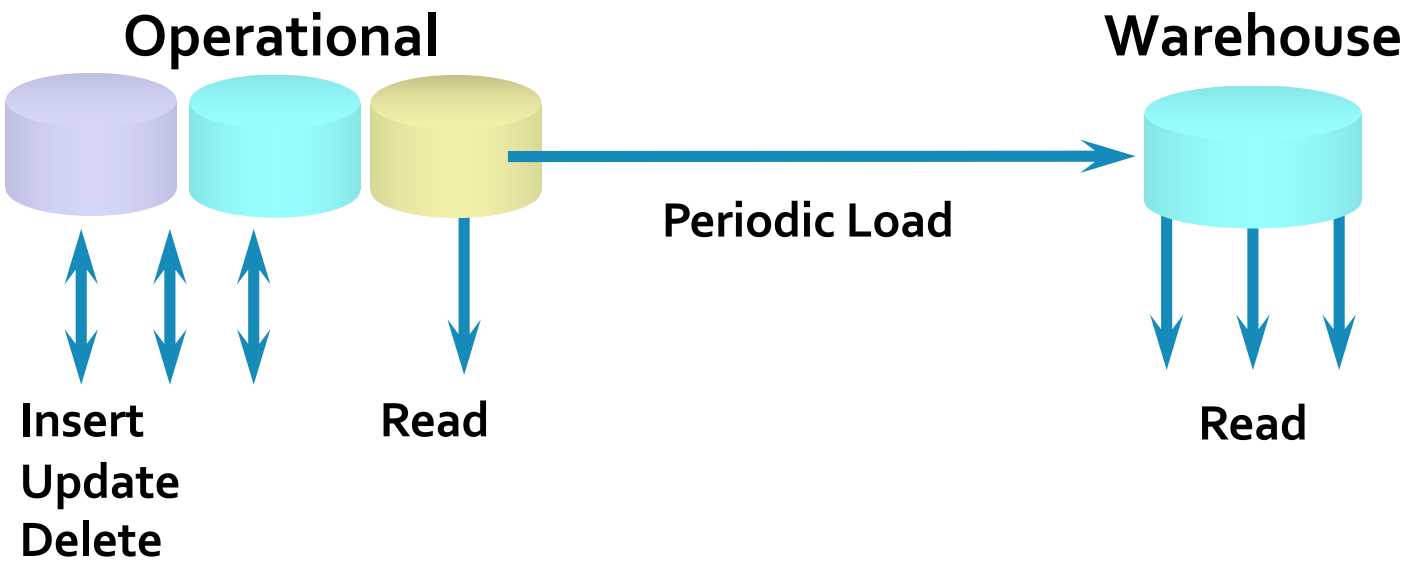


Data warehouse

- **Time horizon – 5-10 years**
- **Records cannot be updated**
- **Snapshots of data**
- **Time becomes a key to the data (dimension)**

Non-volatile Data

- Typically, once data is loaded (either on daily, weekly, monthly or annually basis) into the data warehouse, there are no updates to the data.

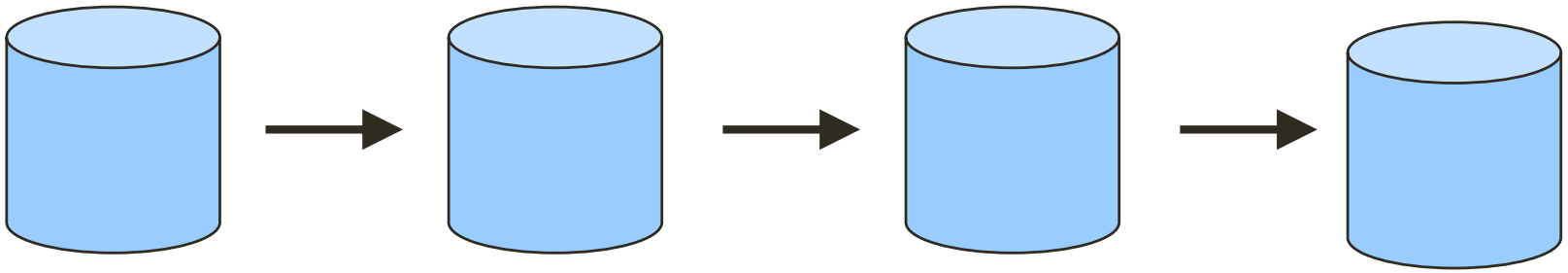


Record-by-record manipulation of data

Mass load/ access of data

Levels of Data Handling

Levels of the architecture



Operational

- Detailed
- Day to day
- Current valued
- High probability of access
- Application oriented

Data warehouse

- Most granular
- Time variant
- Integrated
- Subject oriented
- Some summary

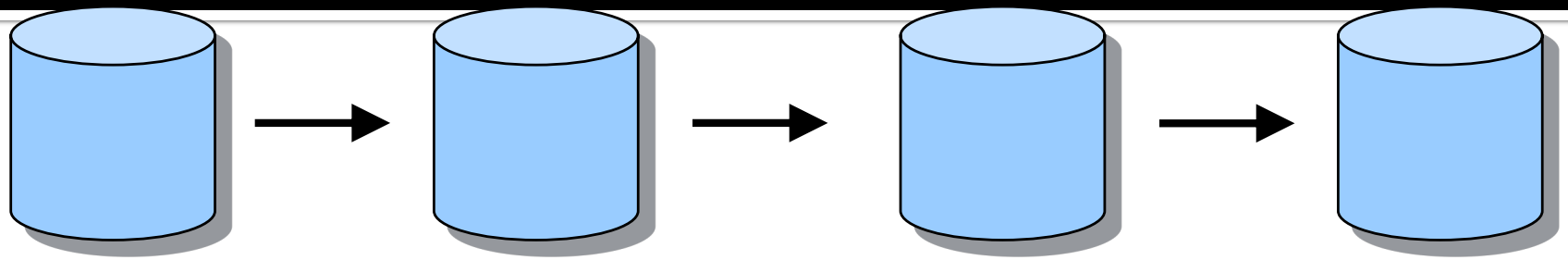
Departmental

- Parochial
- Some derived; some primitive
- Typical departments:
  - Accounting
  - Marketing
  - manufacturing

Individual

- Temporary
- Ad hoc
- Heuristic
- Non-repetitive
- PC, workstation based

A simple example – a customer (1)



Operational

J Jones

123 Main Street

Credit - AA

Data warehouse

J Jones
1986-1987
456 High St
Credit - B

J Jones
1987-1989
456 High St
Credit - A

J Jones
1989-pres
123 Main St
Credit - AA

Departmental

Customers by Month

Jan – 4101

Feb – 4209

Mar – 4175

Apr – 4215

……

Individual

Customers since 1982 with acct balances > 5000 and with credit ratings of B or higher

Temporary!

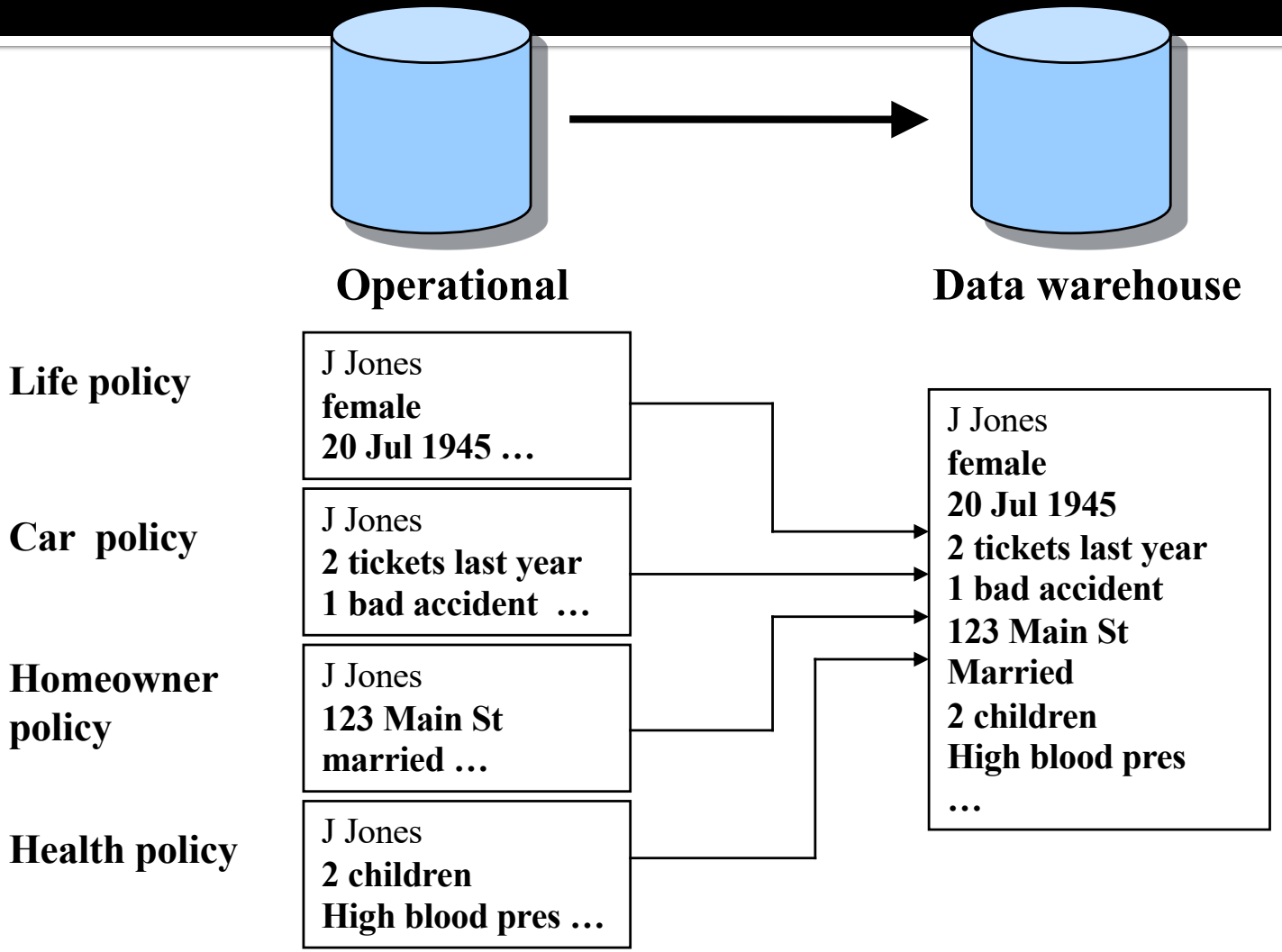
What is J Jones credit rating right now?

What has been the credit history of J Jones?

Are we attracting more or fewer customers over time?

What trends are there for the customers we are analyzing?

A simple example – a customer (2)



Operational (OLTP) System vs Data Warehouse

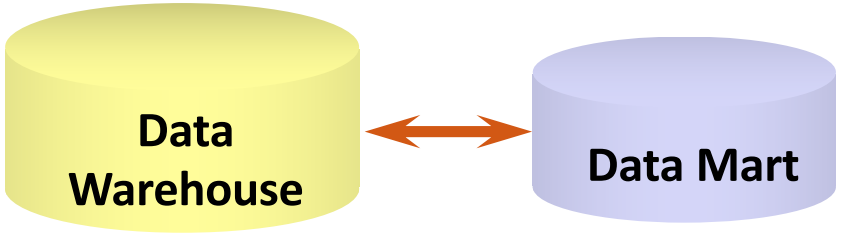
Operational	Data Warehouse
Holds current data	Holds historical data
Stores detailed data	Stores detailed, lightly, and highly summarised data
Data is volatile	Data is largely static (non-volatile)
Repetitive processing (Transaction-driven)	Ad hoc, unstructured, and heuristic processing (Analysis-driven)
High level of transaction throughput	Medium to low level of transaction throughput
Application –oriented	Subject –oriented
Supports day-to-day decision	Supports strategic decision
Serves large number of clerical/operational users	Serves relatively low number of managerial users

Data Mart

- Data mart is customized and/or summarized data derived from the data warehouse and tailored to support the specific analytical requirements of a business unit or function.

- It utilizes a common enterprise view of strategic data and provides business unit more flexibility, control and responsibility.

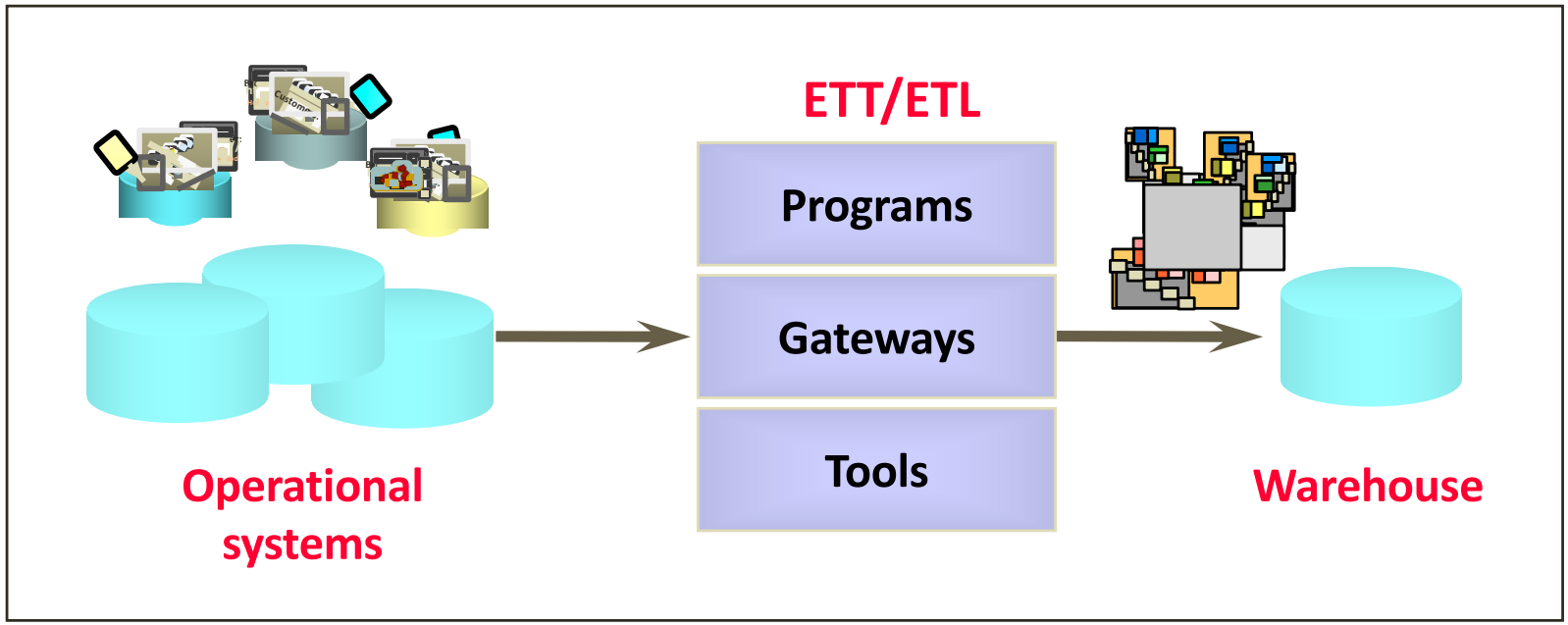
Data Warehouses vs Data Marts



Property	Data Warehouse	Data Mart
Scope	Enterprise	Department
Subjects	Multiple	Single-subject, LOB
Data Source	Many	Few
Size (typical)	100 GB to > 1 TB	< 100 GB
Implementation time	Months to years	Months

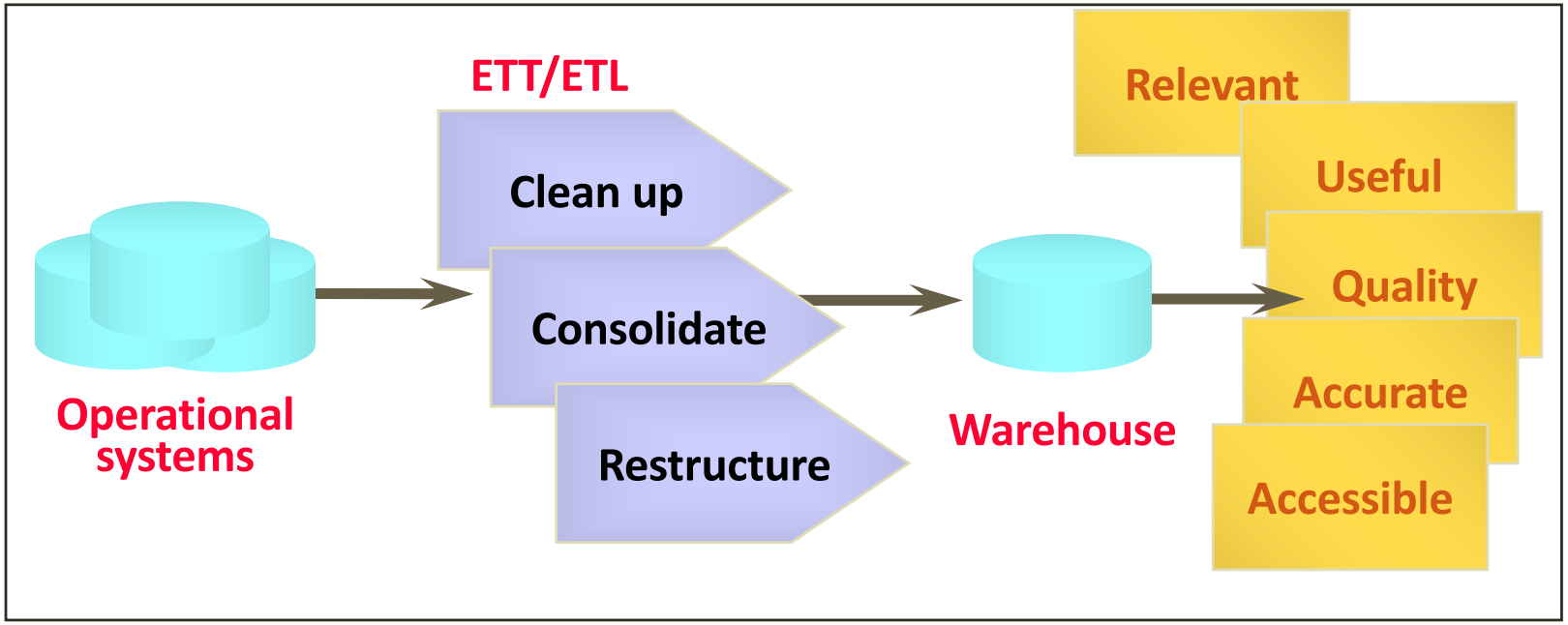
Extraction, Transformation, Transportation/Load (ETT/ETL) Processes

- Extract source data
- Transform/clean data
- Index and summarize
- Load data into DW
- Detect changes
- Refresh data



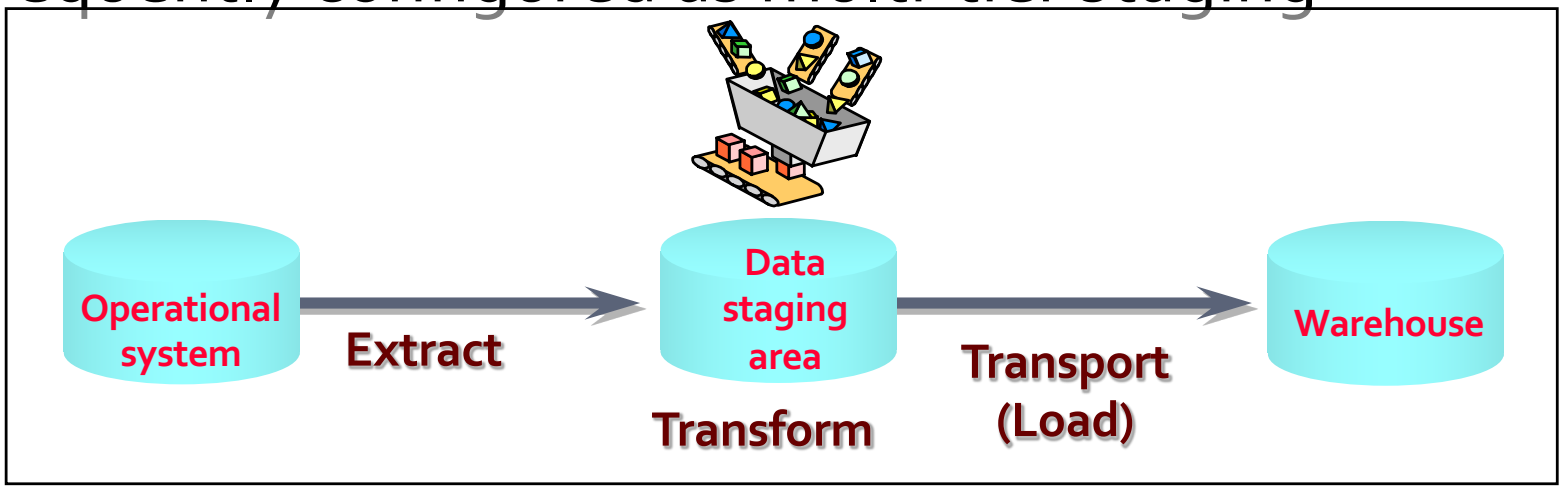
ETT/ETL Process Objectives

- Must result in data that is relevant, useful, high-quality, accurate, and accessible
- Require a large proportion of warehouse development time and resources

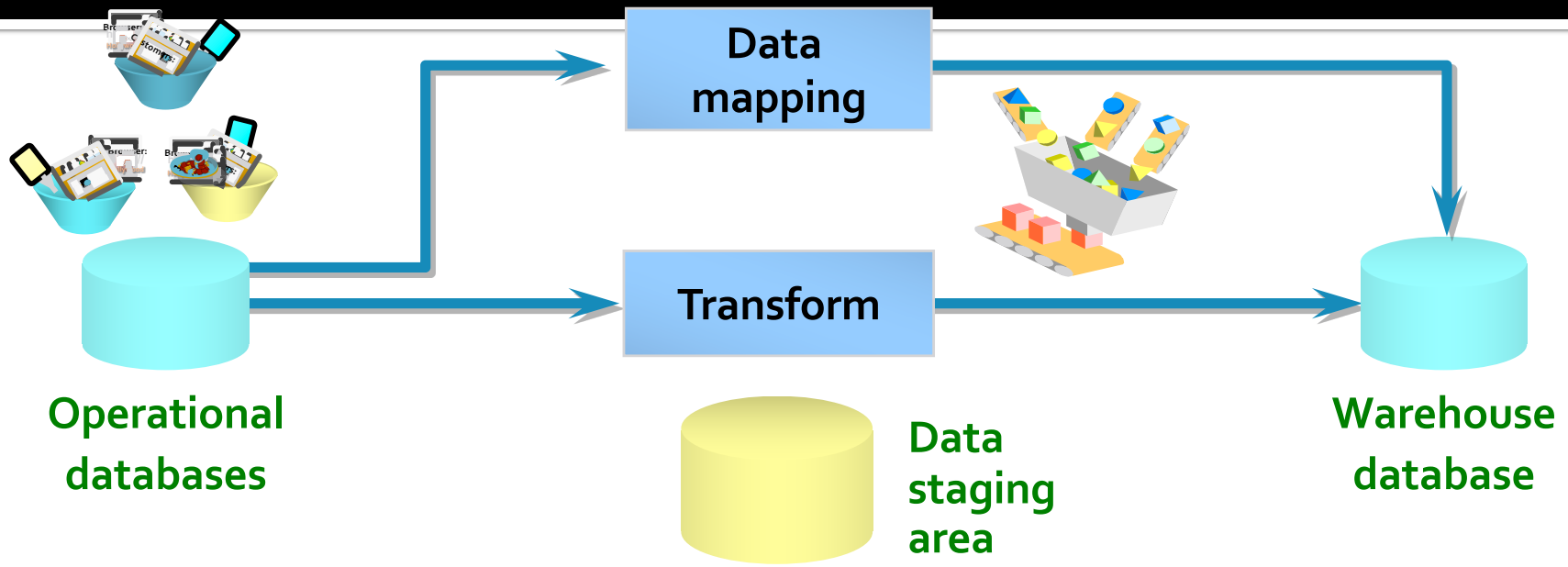


Data Staging Area

- The construction site for the warehouse
- Required by most implementations
- Composed of ODS, flat files, or relational server tables
- Frequently configured as multi-tier staging



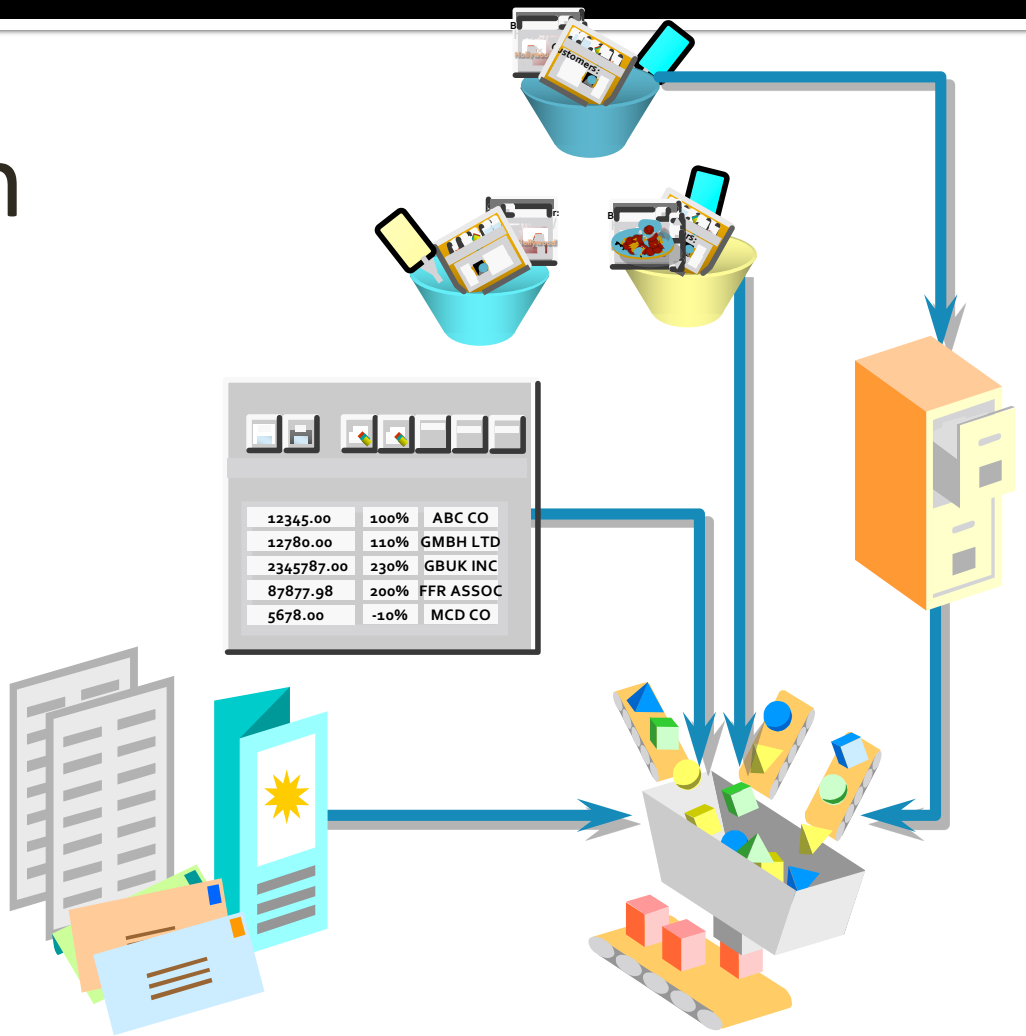
Extracting Data



- Routines developed to select fields from source
- Various data formats
- Rules, audit trails, error correction facilities

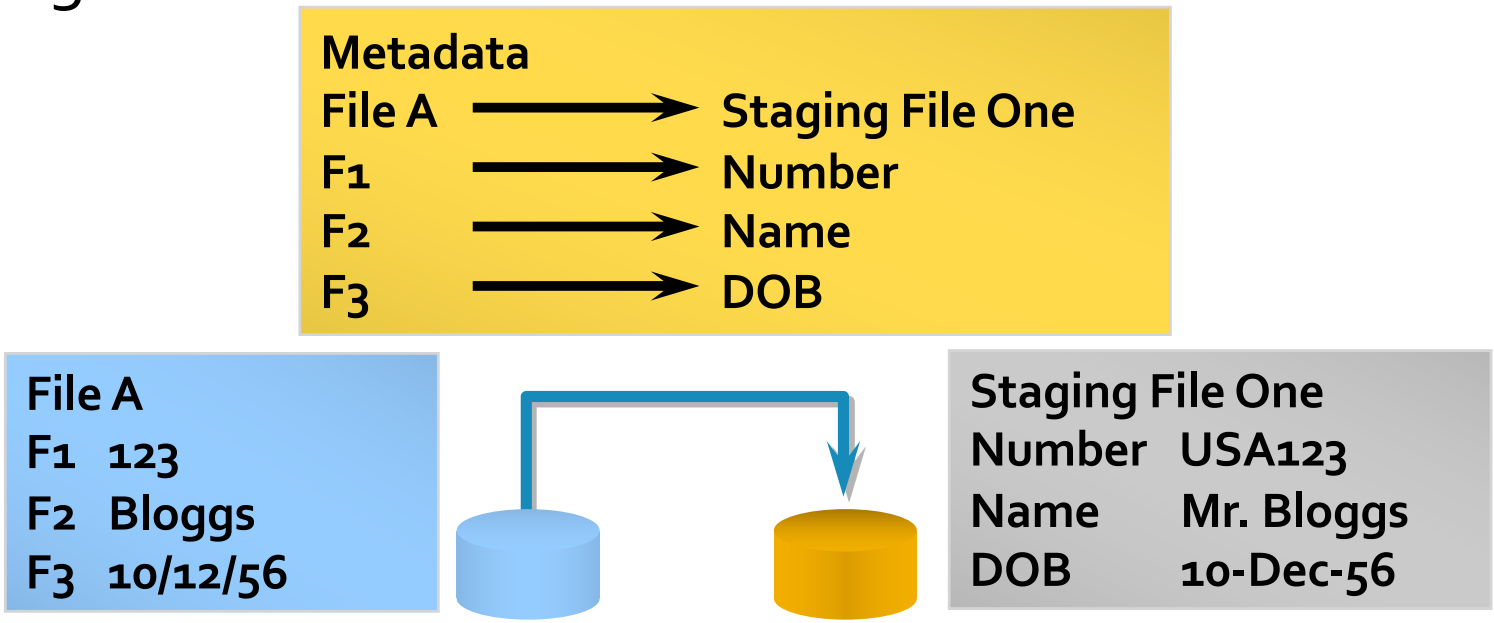
Source Systems

- Production
- Archive
- Internal
- External



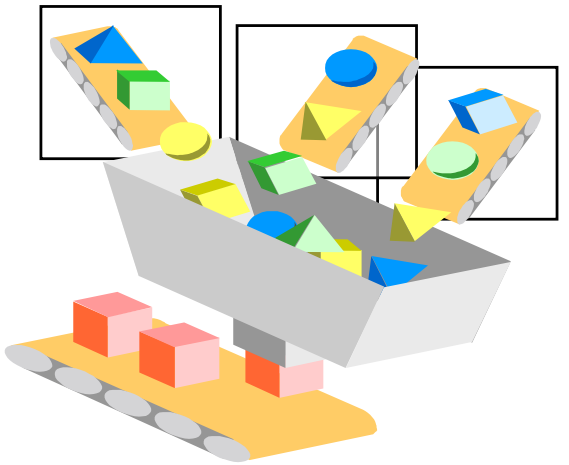
Mapping

- Defines which operational attributes to use
- Defines how to transform the attributes for the warehouse
- Defines where the attributes exist in the warehouse
- Mapping tools are available

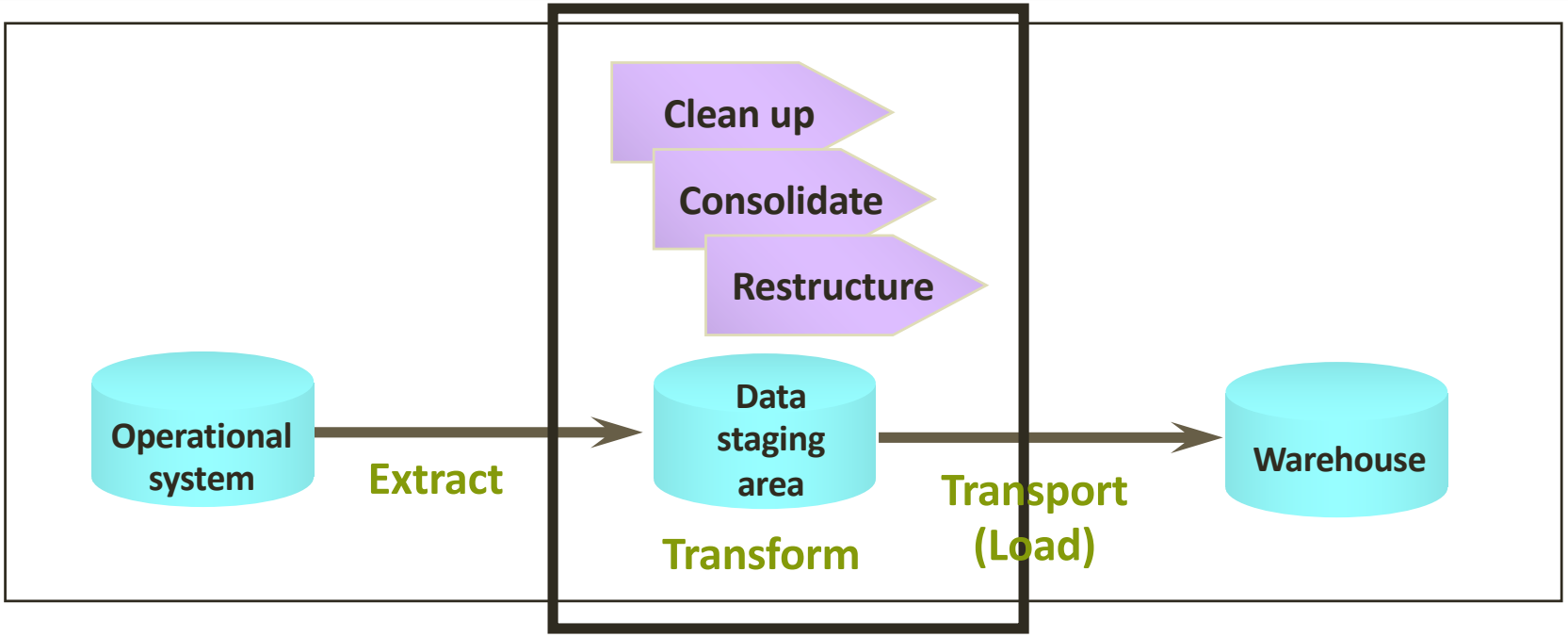


Extraction Techniques

- Programs: C, PL/SQL,
- Gateways: transparent database access
- In-house development is popular
- Tools
  - High initial cost
  - Ongoing automation
  - Data cleanup



Transformation Objectives



Transformation eliminates operational data anomalies

- Cleans
- Standardizes
- Presents subject-oriented data

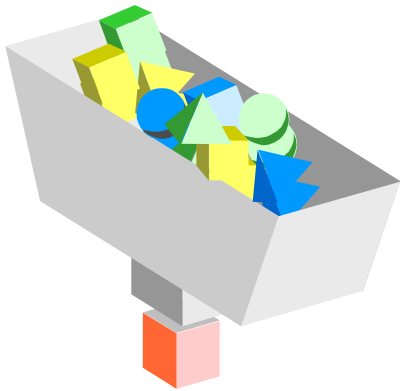
Source Data Anomalies

- No unique key
- Data naming and coding anomalies
- Data meaning anomalies between groups
- Spelling and text inconsistencies

CUSNUM	NAME	ADDRESS
90328575	Oracle Corp	100 NE 1st Street, Tampa
90328575	Oracle	100 NE. First St., Tampa
90238475	Oracle Services	100 North East 1st St., FLA
90233479	Oracle Limited	100 N.E. 1st St.
90233489	Oracle Computing	15 Main Road, Ft. Lauderdale
90234889	Oracle Corp. UK	15 Main Road, Ft. Lauderdale, FLA
90345672	Oracle Corp UK Ltd	181 North Street, Key West, FLA

Transformation Routines

- Cleaning data
- Eliminating inconsistencies
- Adding elements
- Merging data
- Integrating data
- Transforming data before load



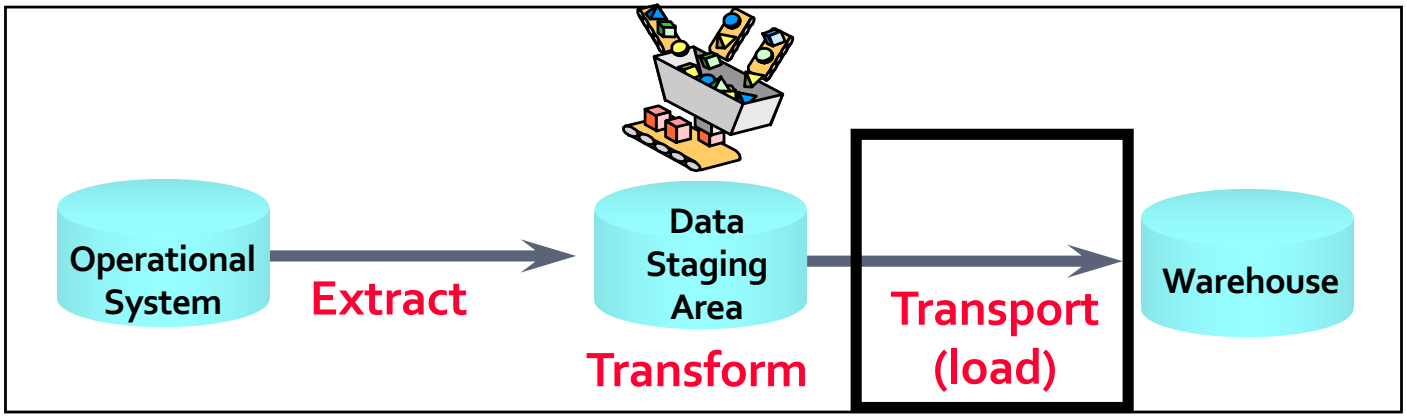
Importance of Data Quality

Quality data is the key to a successful warehouse, it is better to have no data at all than bad data.

- Dirty data must be removed
- Clean data is essential for:
  - Targeting customers
  - Determining buying patterns
  - Identifying householders: private or commercial
  - Matching customers
  - Identify historical data

Transporting Data into the Warehouse

- Loading moves the data into the warehouse
- Loading can be time-consuming:
  - Consider the load window.
  - Schedule the task; automate all processes.
- Initial load moves large volumes
- Subsequent refresh moves smaller volumes
- Business determines the cycle



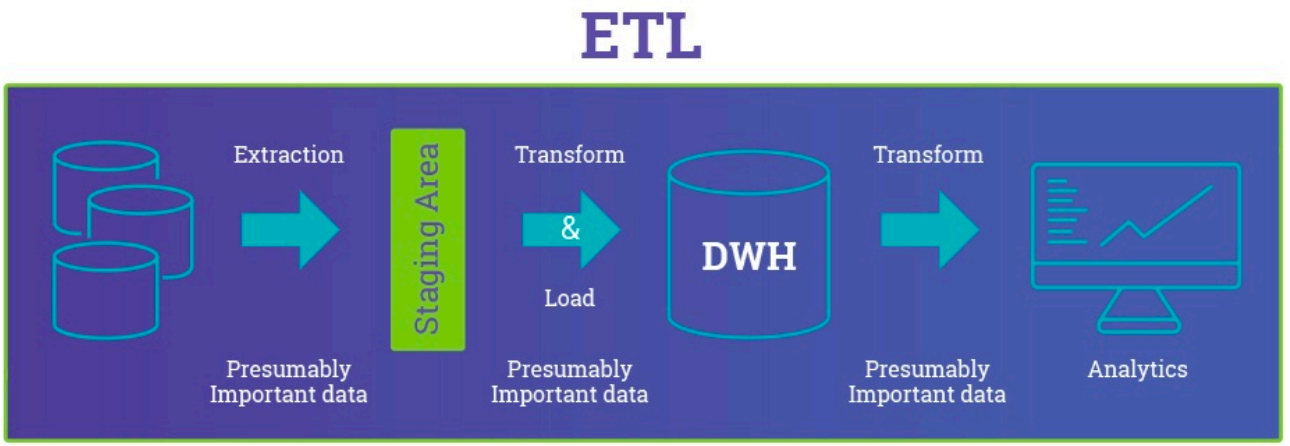
ETL vs ELT

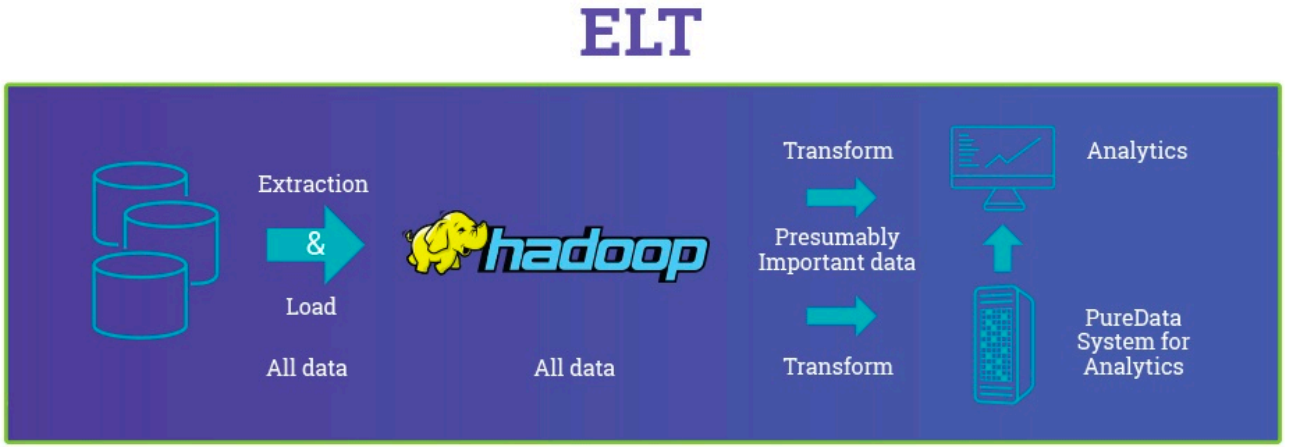
- For the last couple of decades ETL (extract, transform, load) has been the traditional approach for data warehousing and analytics. The ELT (extract, load, transform) approach changes the old paradigm.

ELT

- In the ELT approach, after you've extracted your data, you immediately start the loading phase - moving all the data sources into a single, centralized data repository. With today's infrastructure technologies such as Hadoop, or cloud storage, systems can now support large storage and scalable compute. Therefore, a large, expanding data pool and fast processing is virtually endless for maintaining all the extracted raw data.

ETL vs ELT

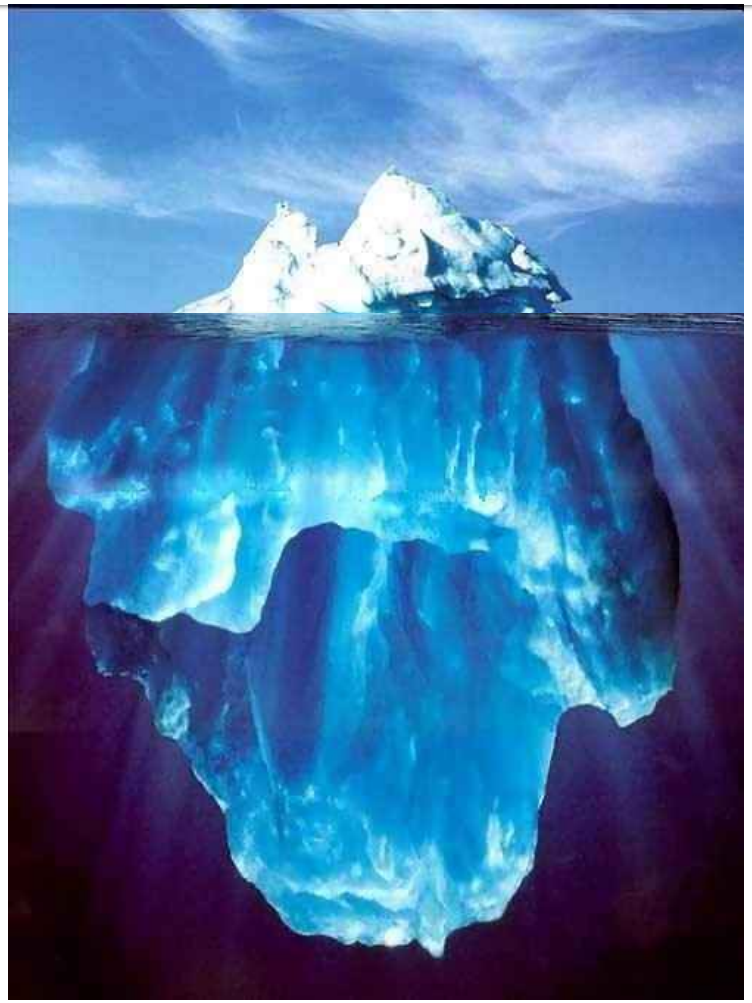

ETL
Extraction
Staging Area
Transform
&
Load
DWH
Transform
Analytics
Presumably Important data
Presumably Important data
Presumably Important data


ELT
Extraction
&
Load
All data
hadoop
All data
Transform
Presumably Important data
Transform
Analytics
PureData System for Analytics

Why ELT instead of ETL?

- **When ingestion speed is important.** Because ELT doesn't have to wait for the data to be worked off-site and then loaded, (data loading and transformation can happen in parallel) the ingestion process is much faster, delivering raw information considerably faster than ETL.

- **When more intel is better intel.** The advantage of turning data into business intelligence lay in the ability to surface hidden patterns into actionable information. By keeping all historical data on hand, organizations can mine along timelines, sales patterns, seasonal trends, or any emerging metric that becomes important to the organization. Since the data was not transformed before being loaded, you have access to all the raw data.



The majority of reports are based on known facts.

BUT

We don't know what we don't know

What is Driving Data Mining?

Changes in Technology:

- Increased usage of the Internet
- Appearance of data warehouses
- Increase in computing power
- Better modeling approaches

Changes in Competition:

- Evolution of strategies:
  - Mass marketing vs. One-to-One marketing
- Increased competition
- Fast-paced environment
- Emergence of niche players

Changes in Customer Behavior:

- Better informed
- More demanding
- Increased willingness to switch to competitors
- Evolution of needs: more complex, harder to satisfy

Why Data Mining?

- An exponential growth of the volume of data collected by organizations is taking place.
- Only a small fraction of this data is ever analyzed.
- Exploiting this 'knowledge mine' is crucial in today's fast changing environment in order to minimize the risk of missing critical emerging market trends.
- The vast amount of information available makes the traditional analysis methods obsolete.

Definition of Data Mining

•Data mining is the process of discovering meaningful new correlations, patterns and trends by "mining" large amounts of stored data using pattern recognition technologies, as well as statistical and mathematical techniques.
(Ashby, Simms (1998))

Data mining is also known as:

- **Knowledge discovery**
- **Data surfing**
- **Data harvesting**

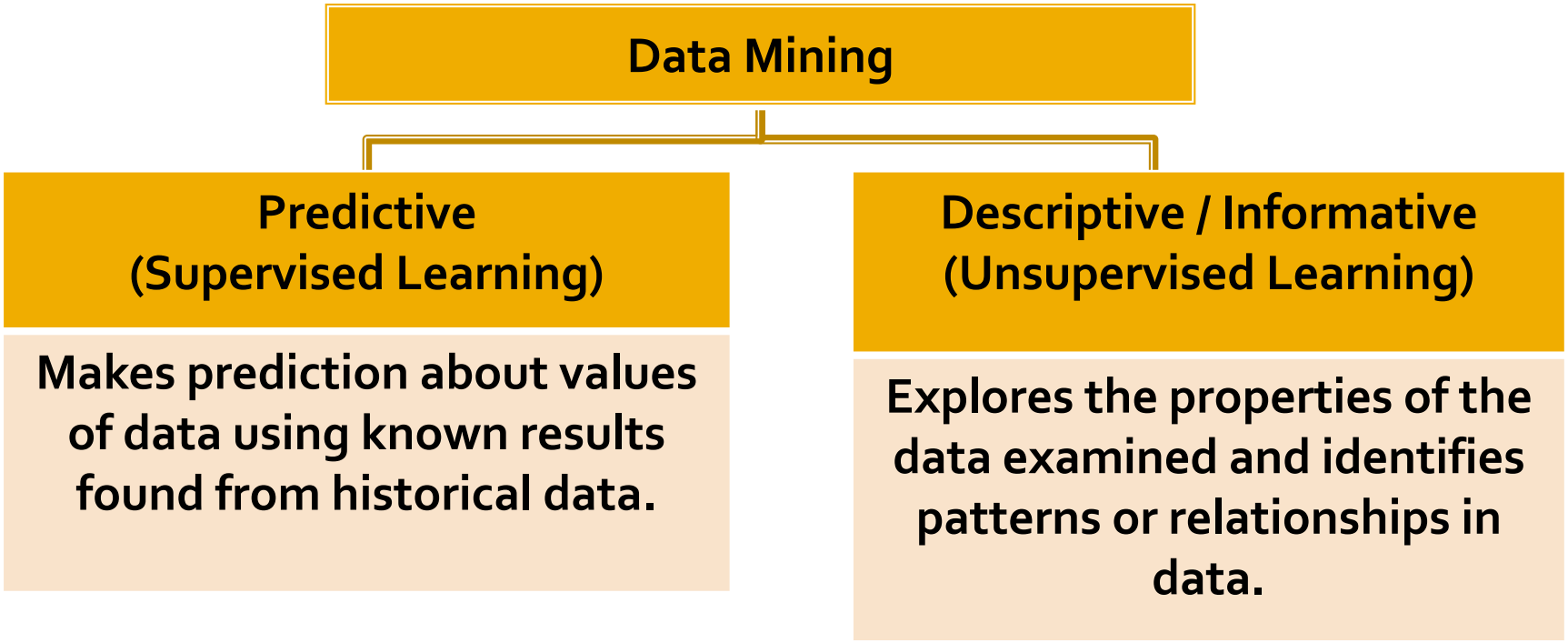
Data Mining Contributions

Data mining helps you analyze and understand … :

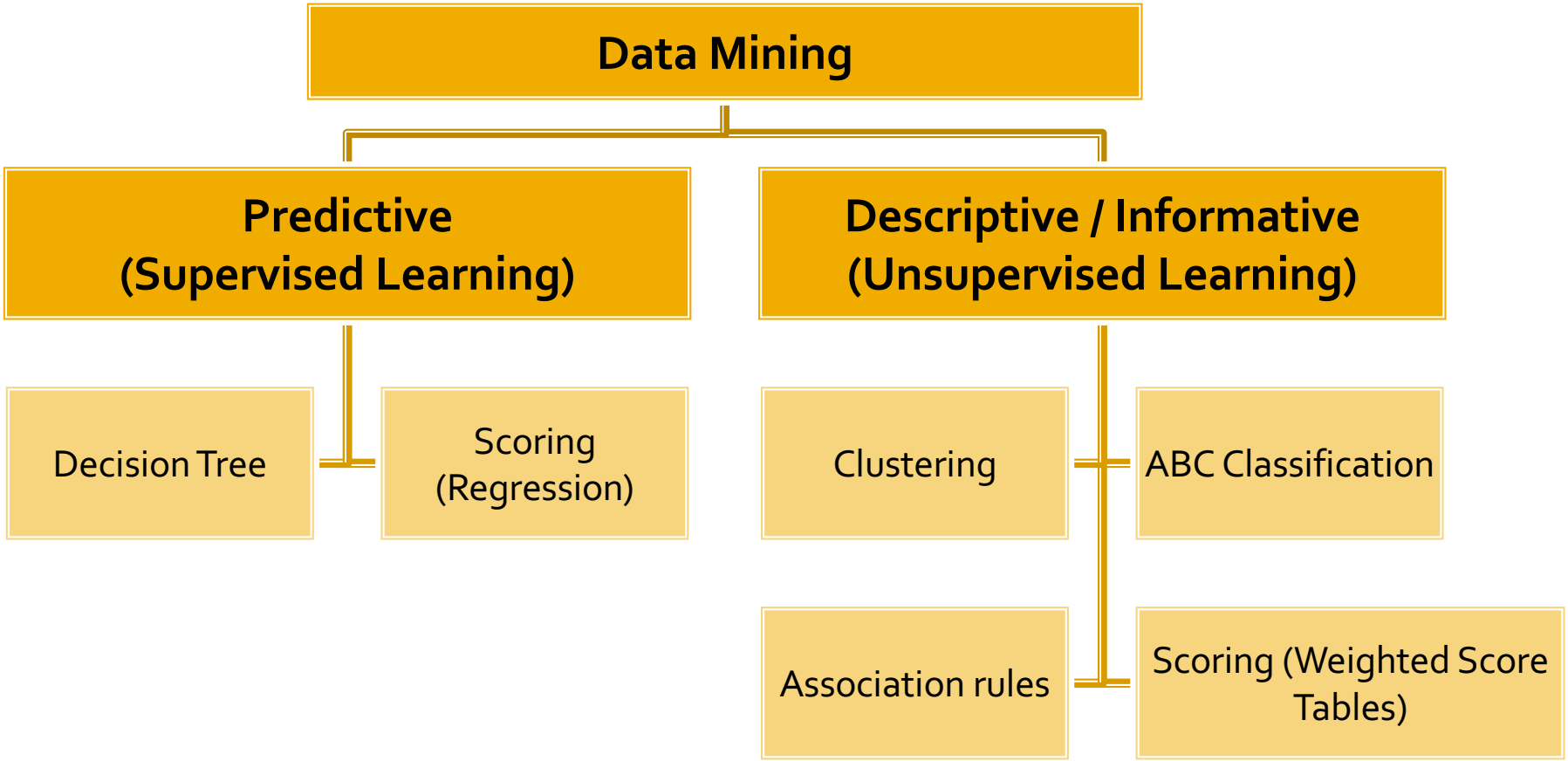
- Data mining is an analytical approach that looks for hidden data patterns and relationships in large databases
- Data mining not only provides insights by analyzing past data, but it is also capable of predicting future trends and behaviors
- Data mining allows organizations to make the critical jump from retrospective analysis to prospective decision-making

Data Mining Methods

- It can be divided into two broad categories:



Data Mining Methods



Future Trends

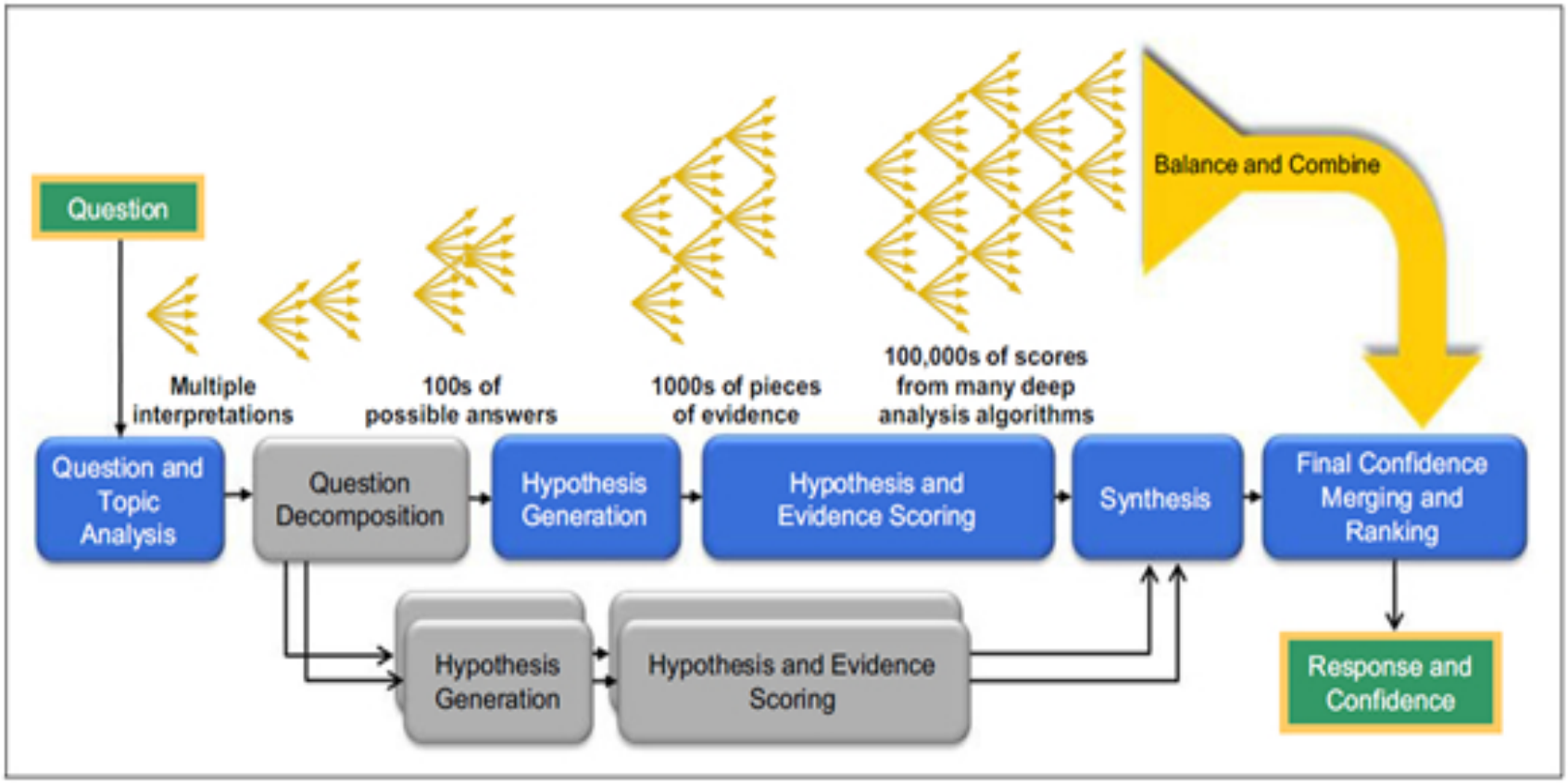
- Embedded Analytics (already prevalent today)
- Cognitive BI
  - the IBM-Watson effect
  - Siri, Cortana
  - Natural Query, etc.
- Automation of BI and Analytics
- Hyper-Individual Experiences
- Marketplaces
- Cloud (widespread growth)

Future Trends - Embedded Analytics

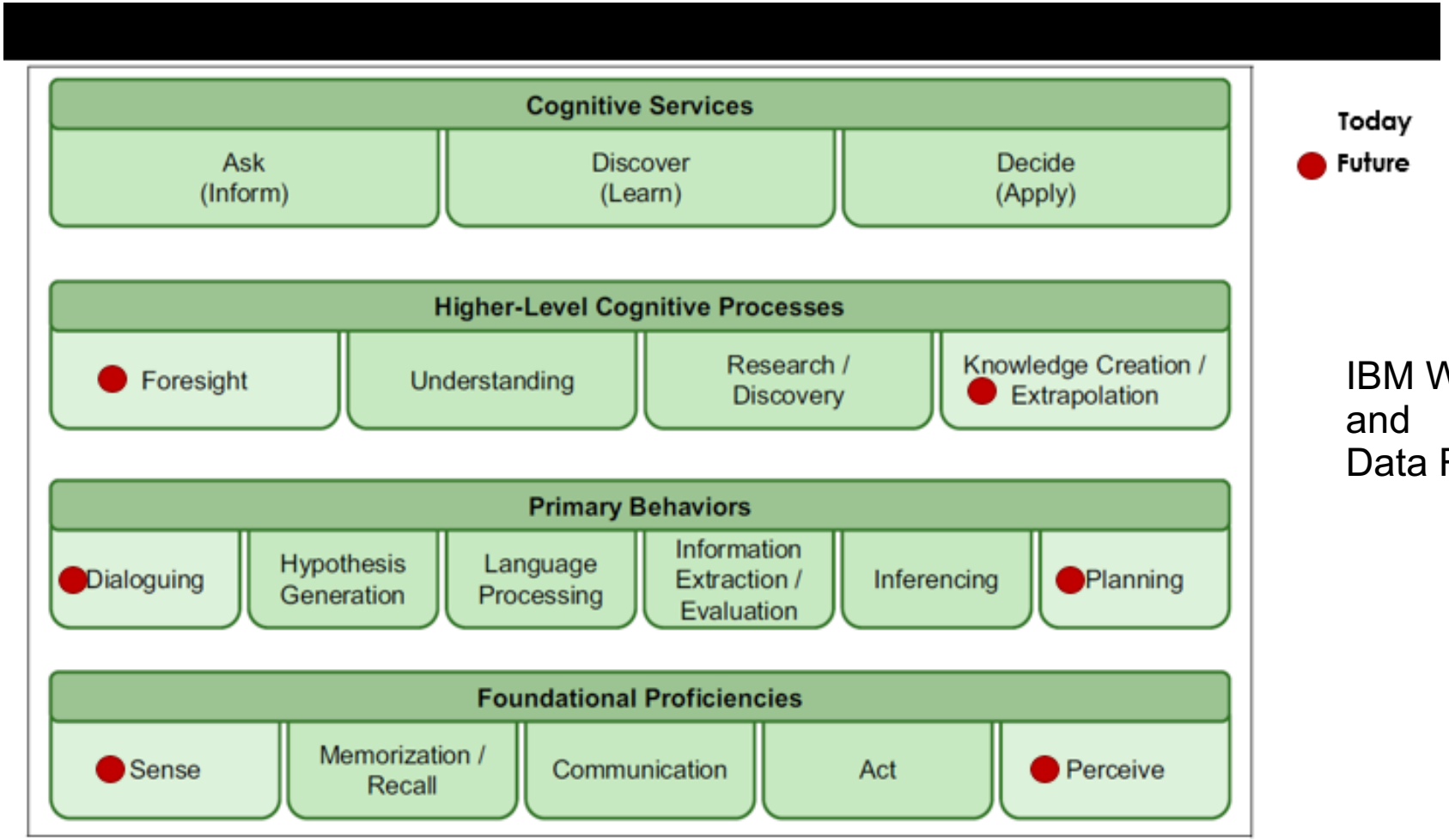
- Integration of analytic content and capabilities within business process applications.
- It provides relevant information and analytical tools designed for the task at hand so users can work smarter and more efficiently in the applications they use every day
- E.g. Google Analytics reporting tool

Future Trends - Cognitive BI

- "Smart machine models" designed to analyze unstructured data, video, images and human language via artificial intelligence and machine learning algorithms



Future Trends - Automation of BI and Analytics



IBM Watson
and
Data RPM

Future Trends - Hyper-Individual Experiences

- People learning and applying new methods of choice, efficiency, self-monitoring, data tracking and information-gathering in their everyday lives.

Future Trends - Marketplaces

- Online destination for developers, marketplace users and technology partners to create and sell innovative big data analytics solutions
- E.g. HP Vertica, Amazon AWS, Datameer