IS475/675 Agenda: May 5, 2025

- Announcements.
- Answer questions.
- Review the different types of data stored in an organization and how they are managed.
- Review organizational data architecture.
  - Present how data is moved among databases
  - ETL (Extract, Transform, and Load)

HW#10 (due 05/09/2024 by 11:59PM)

Before starting the assignment, complete SQL lab exercise 9

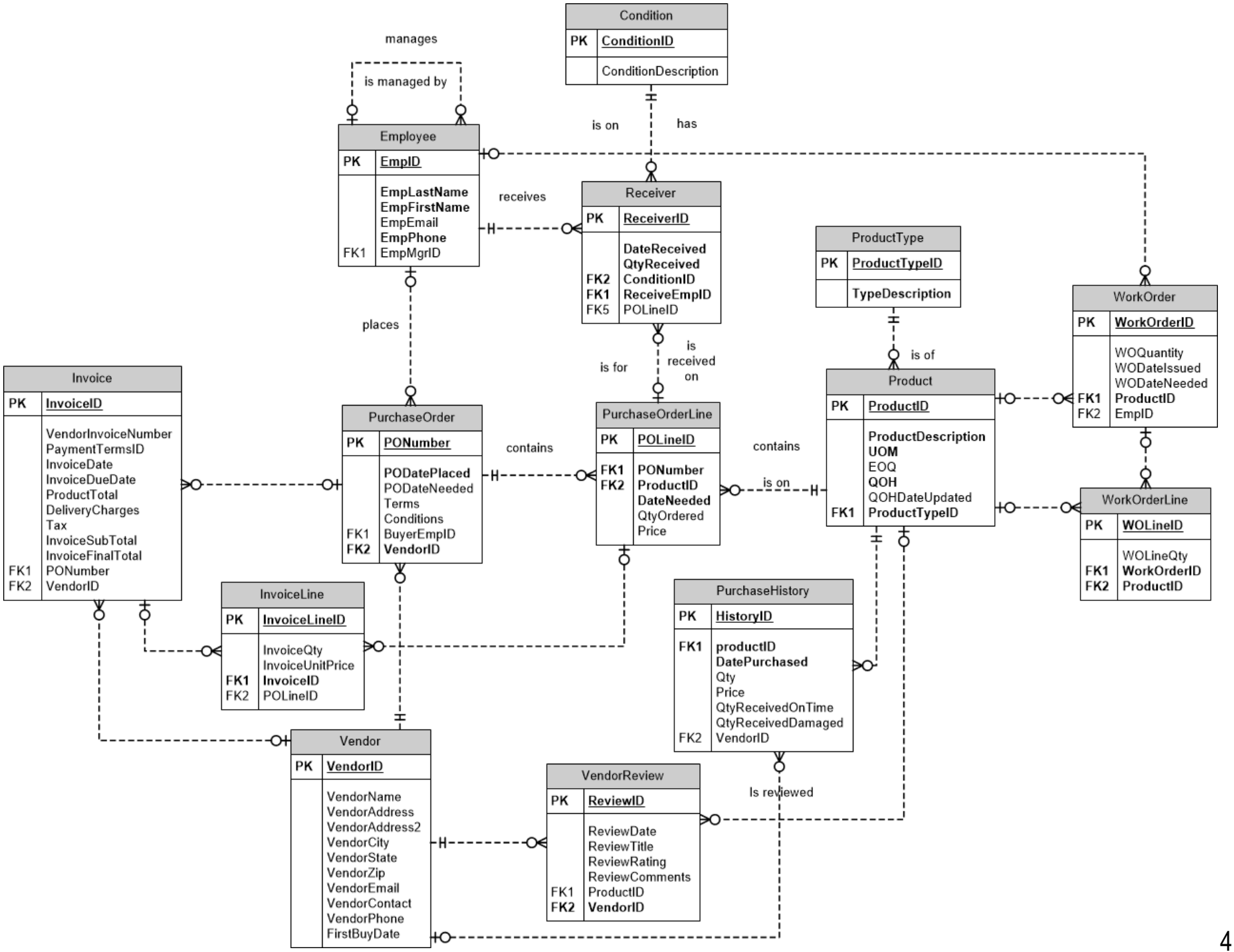
For the queries in Task 3, turn in all SQL code, including all code that composes views and CTE's.

Turn in all SQL code in a file that can be copied and pasted to SQL Server. Do not turn in "pictures" of your SQL code.

Be sure to include the names of the team on the document – also indicate the name of the database where we should test your work.

Final test

- Information available in a document on Canvas/WebCampus in Week 16 handout section.
- Section 1001 (meets normally at 2:30PM): 05/12/2025, 3:00PM-5PM, AB312
- Section 1002 (meets normally at 5:30PM): 05/12/2025, 5:30PM-7:30PM, AB301
- Includes two components:
  - SQL programming (70%)
  - Multiple Choice (30%)
- Uses the Mountain Design database as modified for HW#10. I will provide a script file to build the tables on 05/10/2025.


Condition
PK ConditionID
ConditionDescription
manages
is managed by
Employee
PK EmpID
EmpLastName
EmpFirstName
EmpEmail
EmpPhone
FK1 EmpMgrID
is on
has
receives
Receiver
PK ReceiverID
DateReceived
QtyReceived
FK2 ConditionID
FK1 ReceiveEmpID
FK5 POLineID
ProductType
PK ProductTypeID
TypeDescription
WorkOrder
PK WorkOrderID
WOQuantity
WODateIssued
WODateNeeded
FK1 ProductID
FK2 EmpID
places
is for
is received on
is of
Invoice
PK InvoiceID
VendorInvoiceNumber
PaymentTermsID
InvoiceDate
InvoiceDueDate
ProductTotal
DeliveryCharges
Tax
InvoiceSubTotal
InvoiceFinalTotal
FK1 PONumber
FK2 VendorID
PurchaseOrder
PK PONumber
PODatePlaced
PODateNeeded
Terms
Conditions
FK1 BuyerEmpID
FK2 VendorID
contains
PurchaseOrderLine
PK POLineID
FK1 PONumber
FK2 ProductID
DateNeeded
QtyOrdered
Price
contains
is on
Product
PK ProductID
ProductDescription
UOM
EOQ
QOH
QOHDateUpdated
FK1 ProductTypeID
WorkOrderLine
PK WOLineID
WOLineQty
FK1 WorkOrderID
FK2 ProductID
InvoiceLine
PK InvoiceLineID
InvoiceQty
InvoiceUnitPrice
FK1 InvoiceID
FK2 POLineID
PurchaseHistory
PK HistoryID
FK1 productID
DatePurchased
Qty
Price
QtyReceivedOnTime
QtyReceivedDamaged
FK2 VendorID
Vendor
PK VendorID
VendorName
VendorAddress
VendorAddress2
VendorCity
VendorState
VendorZip
VendorEmail
VendorContact
VendorPhone
FirstBuyDate
VendorReview
PK ReviewID
ReviewDate
ReviewTitle
ReviewRating
ReviewComments
FK1 ProductID
FK2 VendorID
Is reviewed

Organizations must manage data resources

Examples of data stored by an organization

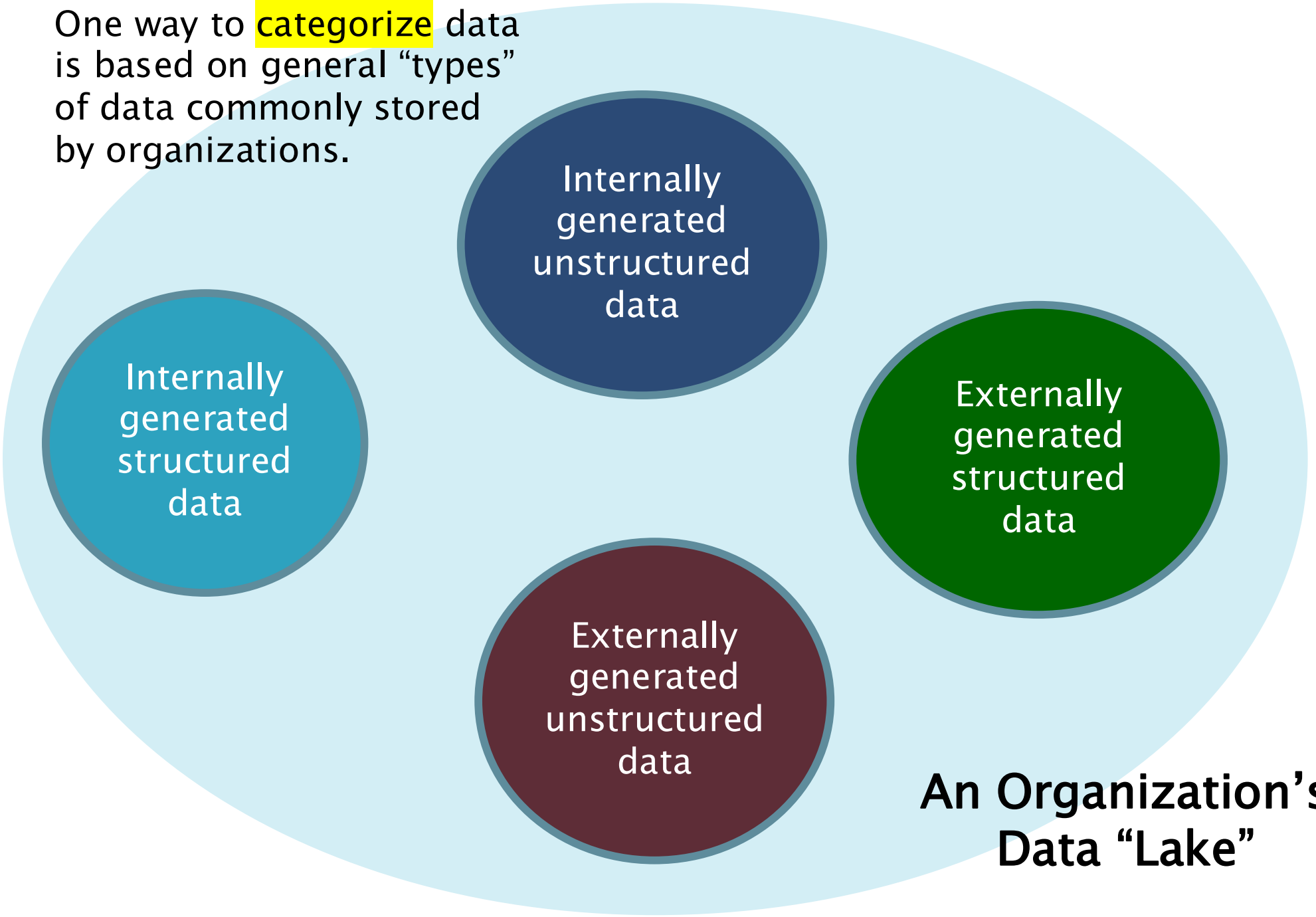
- Current transaction data.
- Historical data for decision making.
- External data of interest for decision making.

All must be designed, implemented and maintained.

All data does not have to be stored together.

Can improve efficiency by understanding what data is best stored in separate databases.

One way to categorize data is based on general "types" of data commonly stored by organizations.



The "V's" of Big Data

- **Volume :** The quantity of data collected. Volume alone does not define big data – database management systems have been handling large volumes of data for years.
- **Velocity:** The frequency of change. Big data arrives at high speed and from multiple sources.
- **Variety:** Different forms and sources of data. Both structured and unstructured data compose big data.
- **Veracity:** With big data, there is uncertainty of the accuracy of data. It isn't always clear who entered the data and whether it can be verified as accurate.
- **Value:** Storing data is just storing data if it doesn't provide value toward a meaningful goal. However, with big data the goal may not always be clear at first glance and may not have value until people have the data and decide how it can be used.

Examples of "Big Data" that an organization might store

- Data transmitted from sensors that may not be validated or calibrated – example: smart watch personal health data transmitted to physicians for evaluation and potential diagnosis.
- Data scraped from websites with product, personal, customer reviews.
- Data from traffic sensors, emergency dispatch (crime statistics), area demographics. May be used to determine placement of a new store.

How is big data stored?

Depends on the "V's".

Much can be stored in a relational database.

Other DBMS's and database "frameworks" can be used.

Need to consider the data management and application layering issues discussed for the last three weeks: Concurrency control, security, backup/recovery, and performance.

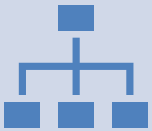
“Information Gap also called the Information Paradox”

Organizations are drowning in data, but people in those organizations are starving for information to help them solve problems.

Why do we have an information gap?

- **Fragmented development.** Today is not our first rodeo. Information systems and their supporting databases are built over many years frequently answering an urgent need for specific transaction processing.
  - Limited resources make many organizations focus on one-thing-at-a-time.
  - Inter-relationships among systems may not always be the major focus.
  - May buy differing hardware and software platforms to answer a specific need.
- **Accounting transaction focus.** Organizations focus on financial transactions. They focus on the need for transactional reporting rather than managerial decision making.
- **Too much data.** The vast quantity of data gathered from various sources has made many IS departments so constrained for resources, that the focus is on storing and protecting data rather than transforming it into information.

Another way to categorize data is based on the type of system that generates or uses the data



Operational system: a system that is used to run a business. Based on individual transactions. Frequently called a "system of record". Mountain Design database is an example of an operational system.

Informational system: A system designed to support decision making. Includes data stored over time so that it can be used as historical data for prediction analytic systems.

Operational Systems

- Synonyms: Transaction processing system, system of record.
- Operational systems:
  - Are frequently based on financial transactions.
  - Usually have a limited, pre-defined set of transactions.
  - Usually have a limited, pre-defined set of decisions that must be supported from the system.
  - Usually need fast answers to questions.
  - Serve as the source data for informational systems.

Informational Systems

- Synonyms: Business intelligence systems (BI), decision support systems (DSS), executive information systems (EIS).
- Informational systems:
  - Centralize data that are scattered throughout disparate operational systems.
  - Clean the data to prior to loading so that it is as accurate as possible for decision making.
  - Are separate from operational systems so that the complex queries required to gather the data for decision making do not impact the performance requirements of the operational system.
  - Support decisions/questions that can usually be answered slowly because they do not have a direct impact on operations.

We use data to answer management questions

Operational Questions

- What products are due to be received today?
- What products were received today?
- What is the price for ProductID 1224 on PO#0667?
- When is the next due date for ProductID 8992?
- Which employee received the products for PO#0667?
- Which employee placed the most purchase orders this month?

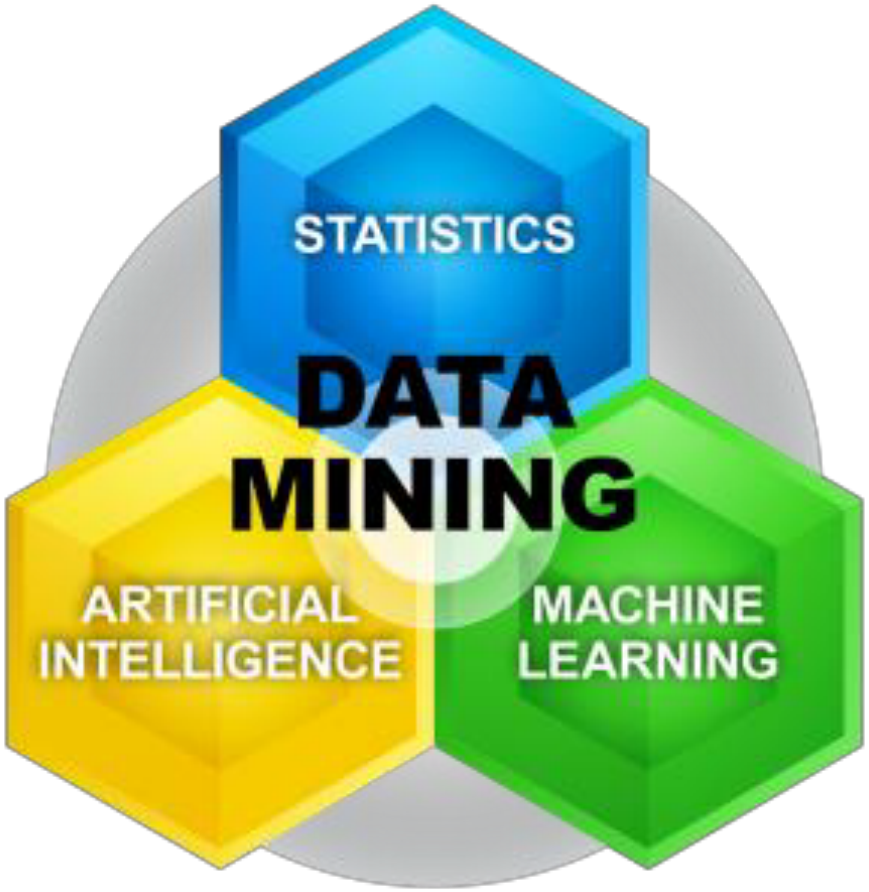
Informational Questions

- Which vendor gives us the best price for ProductID 8992?
- Which vendor delivers ProductID 8992 most reliably (on time and in best condition)?
- Which Product Type is increasing in price most steeply?
- Which Product Type is decreasing in price most quickly?
- Do employees place purchase orders with the same vendors, or do they differentiate based on price or reliability?

People in management want to be able to answer informational questions that require long term data storage; these questions sometimes require very long-term data storage with very large datasets.

OLAP (online analytical processing)	Data Mining
Which products did we buy the most last year?	Which types of products will we need the most next year?
Which vendors provided the best price for our most-used products?	What are the characteristics of the vendors that delivered the most products on-time and of the best quality?
What were our highest selling products in the western U.S. last year?	What products will be highest selling in the western U.S. next year?
Which geographic location bought the most profitable products from us last year?	What are the general characteristics of our customers vs. the demographics of various regions across the world?

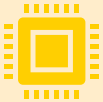
Data mining



- Data mining tools:
  - analyze the data;
  - uncover patterns hidden in the data;
  - form computer models based on the findings; and
  - use the models to predict business behavior.
- Proactive tools.
- Based on artificial intelligence software such as decision trees, neural networks, fuzzy logic systems, inductive nets and classification networking.

Create separate informational system database(s) from operational system database(s) because they are different based on these characteristics

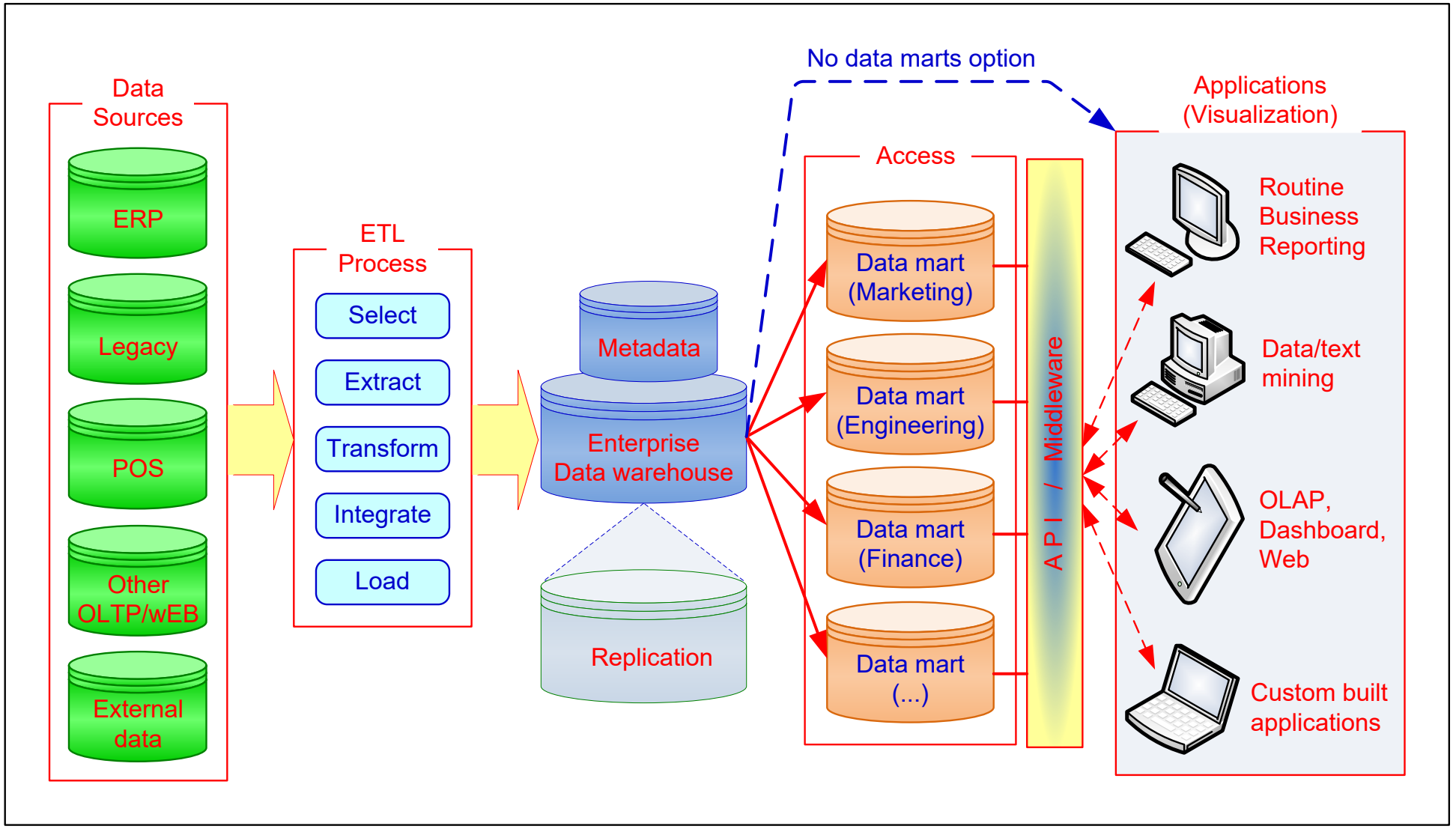
Volume: The quantity of data collected is significantly larger than what is stored to handle the day-to-day transaction processing in an organization. May slow processing to store both "old" and "current" data in the same tables.



Time: Storing data over time.

Data Capture: Data will come from pre-existing digital data; it isn't re-entered through human capture methods.

Use: Data will be used for more medium and long-term decision making.


Data Sources
ERP
Legacy
POS
Other OLTP/wEB
External data
ETL Process
Select
Extract
Transform
Integrate
Load
Metadata
Enterprise Data warehouse
Replication
No data marts option
Access
Data mart (Marketing)
Data mart (Engineering)
Data mart (Finance)
Data mart (...)
A P I / Middleware
Applications (Visualization)
Routine Business Reporting
Data/text mining
OLAP, Dashboard, Web
Custom built applications

What are these databases?

- **System of record database**: Represents the source data in an organization. Data is usually entered by people or automated data input (i.e. RFID, IOT, scanners).
- **Enterprise data warehouse**: A centralized, integrated database that is the single source of all data made available to end users for decision support applications (informational systems). Data is obtained from source data.
- **Data mart**: A data warehouse that is limited in scope whose data are obtained by selecting and summarizing data either from a data warehouse or from a separate ETL process from source data.

Quick Quiz!

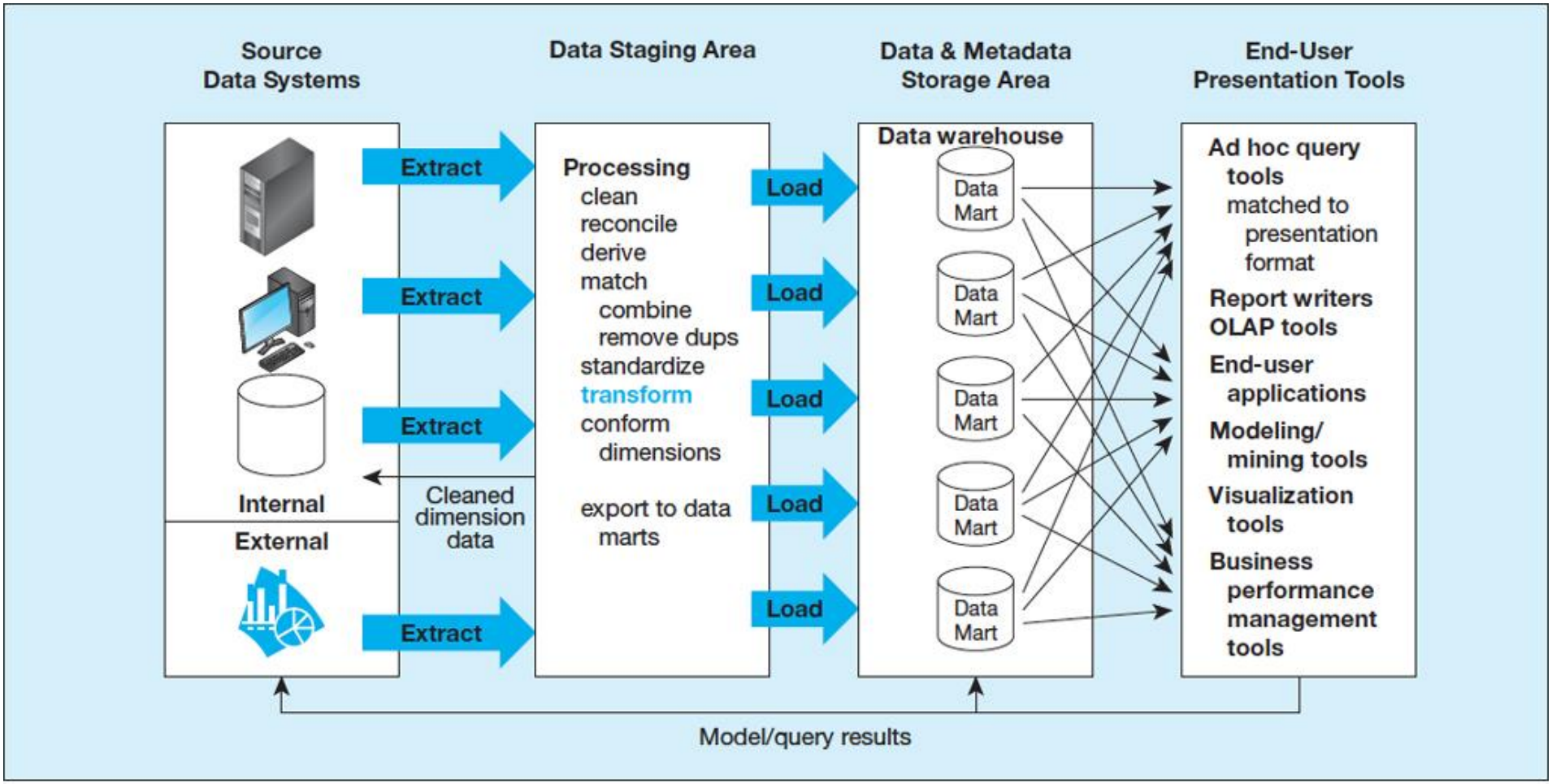
A “system of record”:

A. Is a term for the authoritative data source for a given data element.
B. Contains the data that people rely on as accurate transaction data.
C. Is the system where a data element is usually originally collected.
D. All of the above.

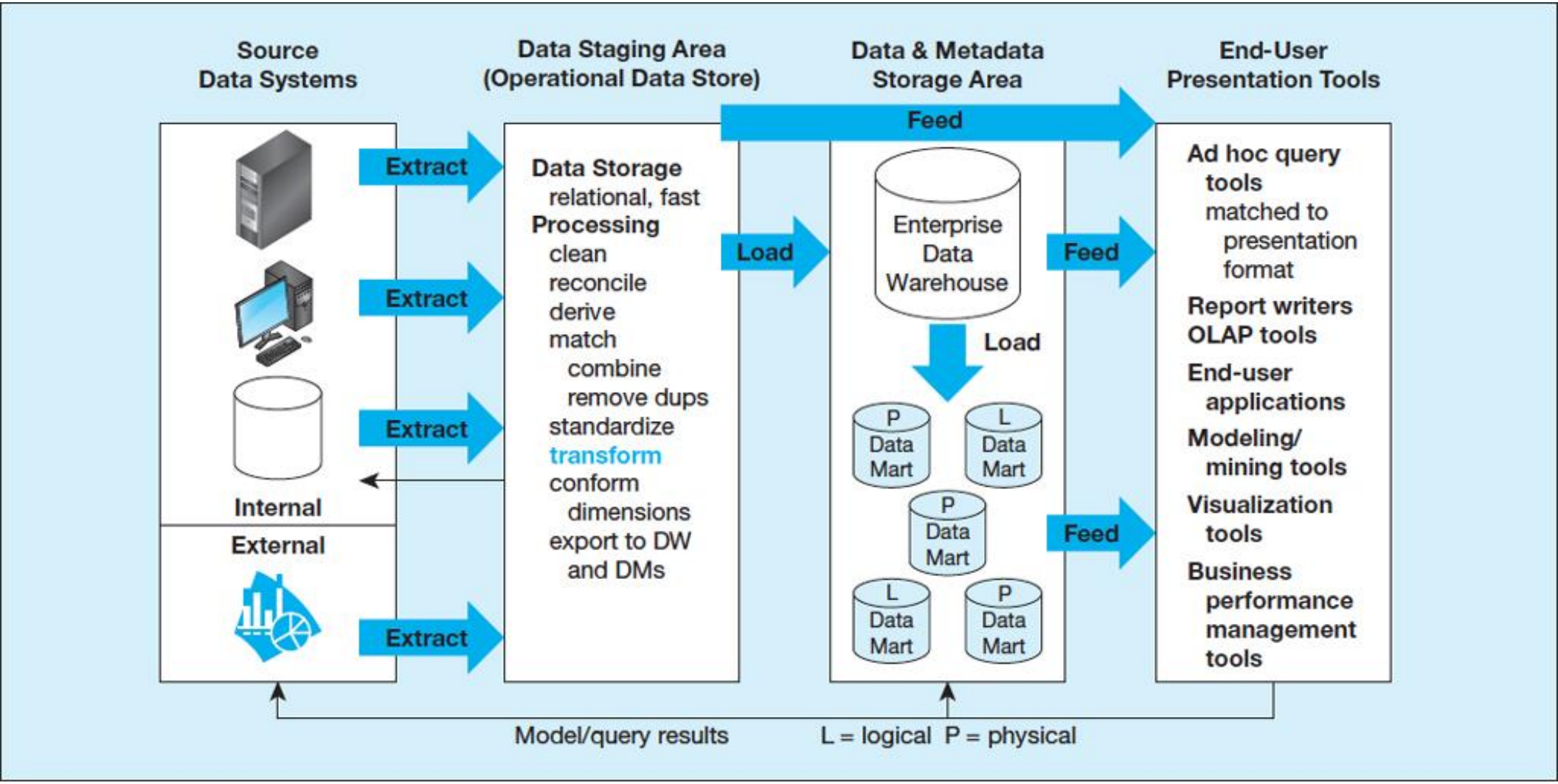
Organizations may create different databases to support operational questions and informational questions because:

a. Operational questions can be answered more slowly than informational questions.
b. Operational questions do not usually require data stored at a very detailed level of granularity.
c. Informational questions may require data integrated from a variety of different systems of record.
d. Operational questions tend to be very complex and require significant processing time and data stored over time to answer fully.

Data architecture option: Independent data mart



Data architecture option: 3 tier, dependent data mart



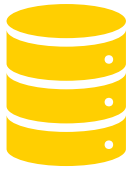
What is a data warehouse?

- A data warehouse is a database designed to support management decision-making.
- A data warehouse may be:
  - More Integrated: It is a centralized, consolidated database integrating data from an entire organization.
  - Subject-oriented: Data warehouse data are sometimes organized around key subjects. The data are often arranged by topic, such as customers, products, suppliers, etc.
  - Time-variant: Data in the warehouse contain a time dimension so that they may be used as a historical aggregation.
  - Non-volatile: Once data enter, they seldom leave. Data are appended rather than overwritten. Data are updated in batches.

Organizations specialize in data warehousing

- Snowflake
- Teradata
- Firebolt
- Redshift
- BigQuery
- General DBMS vendors also provide data warehousing capabilities: Oracle, IBM, Microsoft
- General ERP vendors also provide data warehousing capabilities: SAP, Salesforce

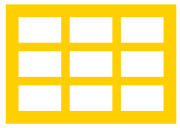
What is ETL (Extract, Transform, Load)?



Extract

Take data from source systems.

May require middleware to gather all necessary data.



Transformation

Put data into consistent format and content.

Validate data – check for accuracy, consistency using pre-defined and agreed-upon business rules.

Convert data as necessary.

Load

Use a batch (bulk) update operation that keeps track of what is loaded, where, when and how.

Keep a detailed load log to audit updates to the data warehouse.

Keep a detailed log of data that is rejected during the ETL process.

Data Cleansing (Transformation)

- Source systems frequently contain "dirty data" that must be cleansed. Examples:
  - Missing or incorrect data.
  - Data that should be null, or null data that should have values.
  - Data with embedded blanks.
  - Incorrect positioning of "decimal points" that aren't actually decimal points, but that are delineating parts of fields.
  - Composite data that should be stored in more granular fields.
  - Incorrect leading or trailing zeros.
  - Incorrect or unknown data types.
- ETL software contains rudimentary all the way to very sophisticated data cleansing capabilities
- Some organizations specialize in creating complex ETL software (Talend, Informatica, Matillion, StitchData)

Summary: Differing data stores

- There is more than one type of data required to be stored for an organization.
  - Structured vs. unstructured data.
  - Internal data vs. external data.
  - Transaction/operational data vs. informational data
- Most shared, structured data now is stored in a relational database.
- Each relational database must be designed with the needs/use of that database at the forefront.
- Many organizations differentiate transaction data vs. informational data. A database to store informational data is frequently termed a "data warehouse."
- A sub-set of a data warehouse is called a "data mart."
- All data stores must be managed to protect the integrity of the data.

People who manage data

- Chief data officer: Responsible for data governance. Creates policies, procedures, standards.
- Database administrator. DBMS expert and manager. Focuses on DBMS efficiency.
- Data security specialist. Protects, controls, secures data.
- Data administrator. Data modeler. Focuses on storing the right data to make decisions.
- Data architect. Develops and evolves the blueprint for an organization's data requirements and storage.
- Data engineer. Data transfer and storage. Builds the ETL processes.

People who use data

- Data scientist. Expert in using advanced forms of data analytics to support decision-making.
- Data analyst. Transforms and uses data to support decision-making.
- Business intelligence analyst. Focuses on using data from a data warehouse/data mart to support decision-making.
- Most analytic professions: Accountant, marketing analyst, financial analyst, general business analyst, economist.

Back to the first day: What is the objective of this class?

- The purpose of this course is to enhance your knowledge of database design, creation, and implementation. This course combines conceptual knowledge of database management systems (DBMS) with practical, hands-on skills using Microsoft's SQL Server DBMS.

Objective	Specific Learning Outcome
Learn how to design a relational database	• Recognize when potentially inaccurate data could be stored in a database. • Design a database that is stable and understand why the design is stable. • Evaluate the design of a database. • Use entity-relationship diagrams (ERD's) to document a database design.
Become a SQL programmer (intermediate-ish level)	• Create, relate, populate, and modify tables in a database using Microsoft's SQL Server database management system (DBMS). • Create result tables using SQL queries. • Use utilities and create custom procedures to import data into a database and transform that data into "clean" data that can be processed by other programs.
Describe how people and software manage data	• List the different types of data stored by organizations. • Describe the different layers of a business application. • Identify the objectives and design goals of a transaction database versus a data warehouse. • Describe the general components and features of a DBMS.