DATA MINING AND WAREHOUSING – CS402

Prof. Linda Sara Mathew

DATA MINING AND WAREHOUSING

Course Objectives

- To introduce the concepts of data mining and its applications
- To understand investigation of the data using practical data mining tools
- To introduce Association Rules Mining
- To introduce advanced Data Mining techniques

DATA MINING AND WAREHOUSING

Expected Outcome

Students will be able to :

- Identify the key process of Data mining and Warehousing
- Apply appropriate techniques to convert raw data into suitable format for practical data mining tasks
- Analyze and compare various classification algorithms and apply in appropriate domain
- Evaluate the performance of various classification methods using performance metrics
- Make use of the concept of association rule mining in real world scenario
- Select appropriate clustering algorithm for various applications and extend data mining methods to the new domains of data

DATA MINING AND WAREHOUSING

Syllabus

Data Mining, Applications, Data Mining Models, Data Warehousing and OLAP, Challenges, Tools, Data Mining Principles, Data Preprocessing: Data Preprocessing Concepts, Data Visualization, Data Sets and Their Significance, Classification Models, Multi Resolution Spatial Data Mining, Classifiers, Association Rules Mining, Cluster Analysis, Practical Data Mining Tools, Advanced Data Mining Techniques, Web Mining, Text Mining, CRM Applications and Data Mining, Data warehousing.

DATA MINING AND WAREHOUSING

Text Books

- Dunham M H, “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, 2003.
- Jaiwei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Elsevier, 2006.

DATA MINING AND WAREHOUSING

References

- M Sudeep Elayidom, “Data Mining And Warehousing”, 1st Edition, 2015, Cengage Learning India Pvt. Ltd.
- Mehmed Kantardzic, “Data Mining Concepts, Methods And Algorithms”, John Wiley And Sons, Usa, 2003
- Pang-ning Tan And Michael Steinbach, “Introduction To Data Mining”, Addison Wesley, 2006.

Module 1

Concepts

- Data Mining?
  - Extracting or mining knowledge from large amounts of data.
  - The amount of data kept in computer files and database is growing at a phenomenal rate. The users of these data are expecting more sophisticated information from them.
  - *Knowledge discovery from Data(KDD)*

Data mining applications

1. Classification

   Eg: In loan database, to classify an applicant as a prospective or defaulter , given his various personal and demographic features along with previous purchase characteristics.

2. Estimation

   Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student , whose previous marks are already known.

3. Prediction

   Predictive model predicts a future outcome rather than the current behaviour. Eg: Predict next week's closing price for the Google share price per unit.

Data mining applications

4. Market basket analysis(association rule mining)

   Analyses hidden rules called association rule in a large transactional database.

   {pen, pencil-> book} – whenever pen and pencil are purchased together, book is also purchased.

5. Clustering

   Classification into different classes based on some similarities but the target classes are unknown.

Areas where data mining is applied:

Business intelligence

Business data analytics

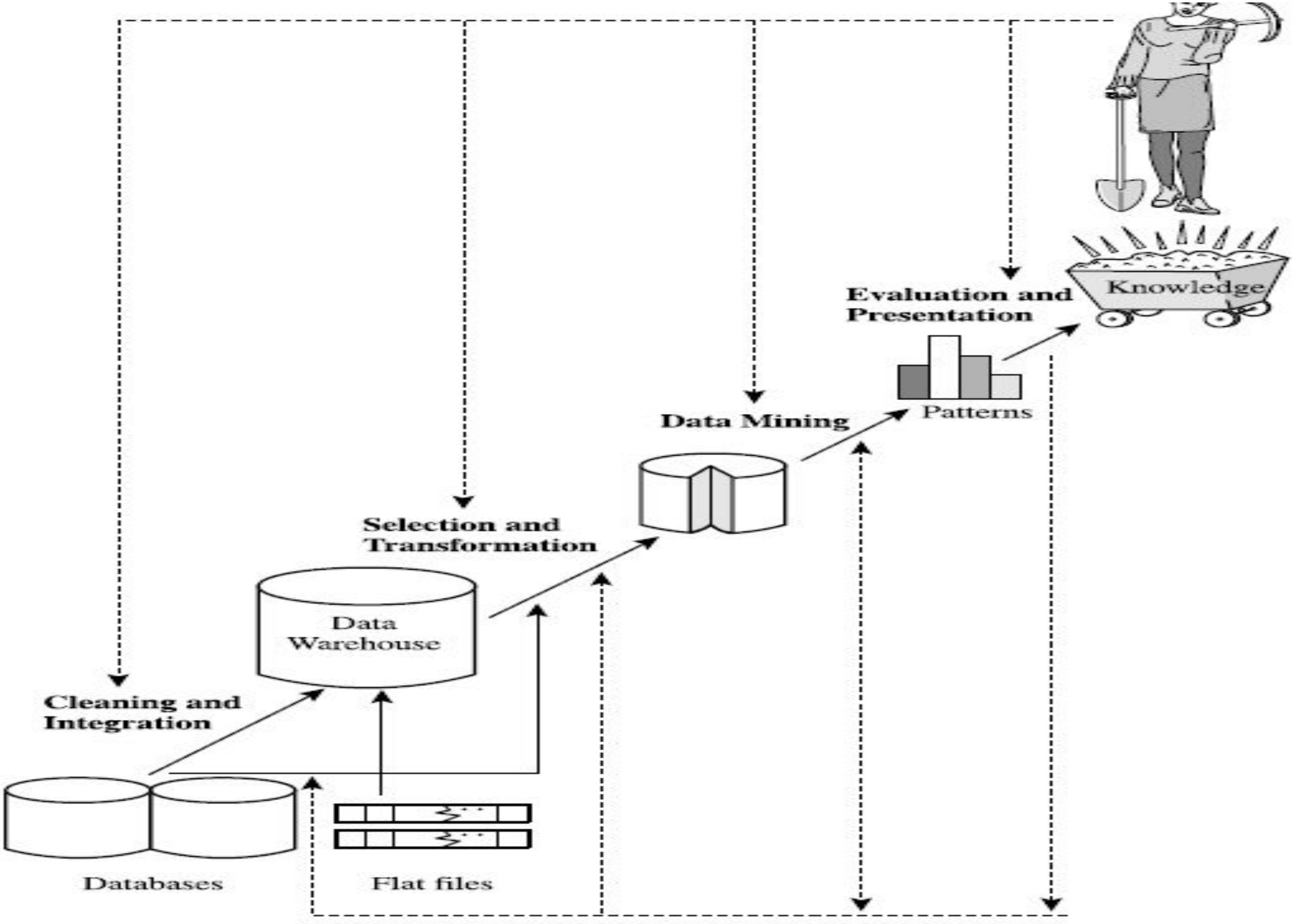
Bioinformatics

Web mining

Text mining

Social network data analysis

Data mining as a process in knowledge discovery



Data mining stages

1. **Data cleaning :**

   to remove noise and inconsistent data

2. **Data integration**

   where multiple data sources may be combined

3. **Data selection**

   where data relevant to the analysis task are retrieved from the database

4. **Data transformation**

   where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations

5. **Data mining :**

   an essential process where intelligent methods are applied in order to

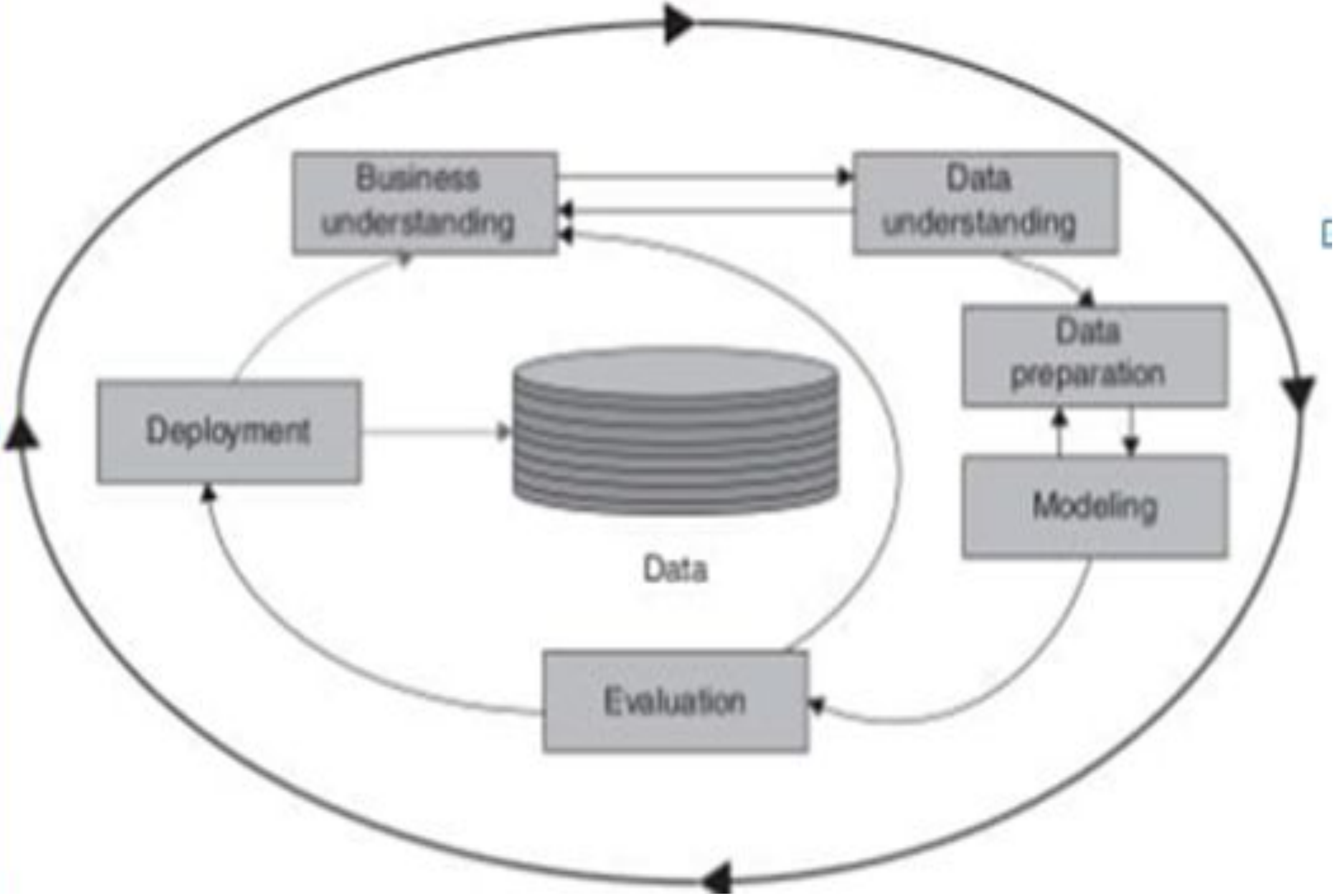
   extract data patterns

6. **Pattern evaluation**

   to identify the truly interesting patterns representing knowledge based on some interestingness measures

7. **Knowledge presentation**

   where visualization and knowledge representation techniques are used to present the mined knowledge to the user


Business understanding
Data understanding
Data preparation
Deployment
Modeling
Data
Evaluation

Data mining models

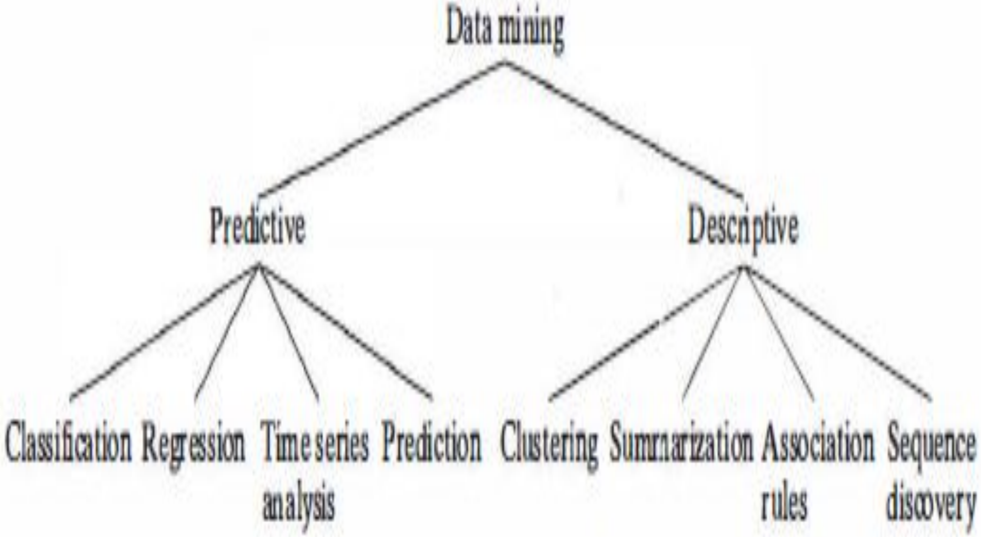


FIGURE 1.2: Data mining models and tasks.

Data mining models

1. **Predictive model**
   - makes a prediction about values of data using known results found from different data.
   - may be made based on the use of other historical data.
2. **Descriptive model**
   - identifies patterns or relationships in data.
   - serves as a way to explore the properties of the data examined, not to predict new properties.
   - Clustering, summarization, association rules, and sequence discovery are usually viewed as descriptive in nature.

Data mining models

1.1 Classification

- maps data into predefined groups or classes.
- It is often referred to as supervised learning because the classes are determined before examining the data.
- Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes
- Eg: Naïve Bayes Classifier

Data mining models

1.2 Regression

- Regression is used to map a data item to a real valued prediction variable.
- regression involves the learning of the function that does this mapping.
- Regression assumes that the target data fit into some known type of function (e.g., linear, logistic, etc.) and then determines the best function of this type that models the given data.
- Some type of error analysis is used to determine which function is "best."

Data mining models

1.3 Time Series Analysis

- The value of an attribute is examined as it varies over time.
- The values usually are obtained as evenly spaced time points (daily, weekly, hourly, etc.).
- A time series plot , is used to visualize the time series.
- There are three basic functions performed in time series analysis:
    - distance measures are used to determine the similarity between different time series.
    - the structure of the line is examined to determine its behavior.
    - use the historical time series plot to predict future values.

Data mining models

1.3 Prediction

- Many real-world data mining applications can be seen as predicting future data states based on past and current data.
- Prediction can be viewed as a type of classification.
- Prediction is predicting a future state rather than a current state.
- Prediction applications include flooding, speech recognition, machine learning, and pattern recognition.

Data mining models

2.1 Clustering

- similar to classification except that the groups are not predefined, but rather defined by the data alone.
- unsupervised learning or segmentation.
- It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed.
- The clustering is usually accomplished by determining the similarity among the data on predefined attributes.
- The most similar data are grouped into clusters.

Data mining models

2.2 Summarization

- Summarization maps data into subsets with associated simple descriptions.
- also called characterization or generalization.
- It extracts or derives representative information about the database.
- This may be accomplished by actually retrieving portions of the data.
- summary type information (such as the mean of some numeric attribute) can be derived from the data.
- The summarization characterizes the contents of the database

Data mining models

2.3 Association Rules

- Link analysis, alternatively referred to as affinity analysis or association, refers to the data mining task of uncovering relationships among data.
- An association rule is a model that identifies specific types of data associations.
- These associations are often used in the retail sales community to identify items that are frequently purchased together.
- Associations are also used in many other applications such as predicting the failure of telecommunication switches.

Data mining models

2.4 Sequence Discovery

- Sequential analysis or sequence discovery is used to determine sequential patterns in data.
- These patterns are based on a time sequence of actions.
- These patterns are similar to associations in that data (or events) are found to be related, but the relationship is based on time.
- In sequence discovery the items are purchased over time in some order.
- For example, most people who purchase CD players may be found to purchase CDs within one week.

Data mining Functionalities

1.Class/Concept Description:
Characterization and Discrimination

2.Mining Frequent Patterns,
Associations and Correlations

3.Classification and Regression for
Predictive Analysis

4.Clustering

5.Outlier Analysis

Data warehousing

- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
- Data warehouse refers to a database that is maintained separately from an organization’s operational databases.

Data warehouse

- “A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision making process”
  - Subject-oriented:
    - A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.
    - A data warehouse focuses on the modelling and analysis of data for decision makers(not on day to day transaction).
    - Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
  - Integrated:
    - A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.

Data warehouse

- Time-variant:
  - Data are stored to provide information from a historical perspective
  - Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
- Non-volatile:
  - A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: *initial loading of data and access of data.*

Data warehousing

- Data warehousing is the process of constructing and using data warehouses.
  - The construction of a data warehouse requires data cleaning, data integration, and data consolidation.
  - The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows “knowledge workers” (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse.

Data warehousing

- Data warehousing is very useful from the point of view of *heterogeneous database integration*.
  - The traditional database approach to heterogeneous database integration was a ‘query- driven’ approach
  - data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis

Operational Database Systems Vs DataWare houses

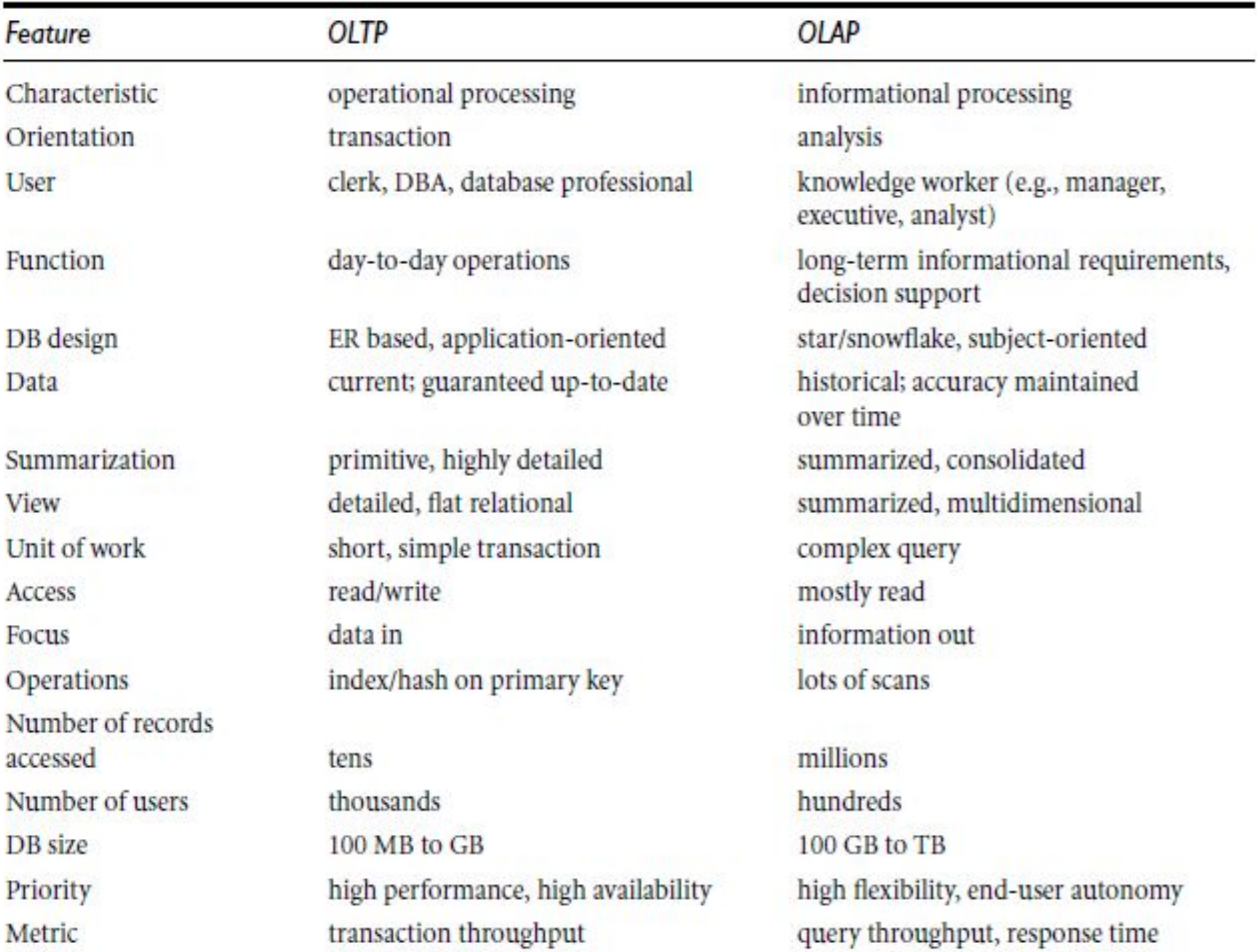
- Operational Database systems
  - Main task is to perform on-line transaction and query processing. These systems are called **on-line transaction processing (OLTP)** systems.
  - They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.
- Data Warehouse
  - serve users or knowledge workers in the role of data analysis and decision making.
  - Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as **on-line analytical processing (OLAP)** systems.

OLTP Vs OLAP

- Users and system orientation:
  - OLTP system is *customer-oriented and is used for* transaction and query processing by clerks, clients, and information technology professionals.
  - OLAP system is *market-oriented and is used for data analysis by knowledge* workers, including managers, executives, and analysts.
- Data contents:
  - OLTP system manages current data
  - OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
- Database design:
  - An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
  - An OLAP system typically adopts either a *star or snowflake model and a subjectoriented* database design.

OLTP Vs OLAP

- View:
  - An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.
  - An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
  - OLAP systems also deal with information that originates from different organizations.
  - OLAP data are stored on multiple storage media.
- Access patterns:
  - The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.
  - Accesses to OLAP systems are mostly read-only operations although many could be complex queries.



Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Need for Data Warehousing

- To promote the high performance of both online transaction processing and online analytical processing
- Data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views. Processing OLAP queries in operational databases would substantially degrade the performance of operational tasks.
- An operational database supports the concurrent processing of multiple transactions. Concurrency control techniques are required in OLTP. But such measures will degrade the performance of OLAP.
- Structures, contents, and uses of the data in these two systems are different.

Data Mining Issues

- **Mining methodology and user interaction issues**
  - *Mining different kinds of knowledge in databases:*

    Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery task. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

  - *Interactive mining of knowledge at multiple levels of abstraction:*

    Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

  - *Incorporation of background knowledge:*

    Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.

Data Mining Issues

- **Mining methodology and user interaction issues**
  - *Data mining query languages and ad hoc data mining:*

    Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval. Such a language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

  - *Presentation and visualization of data mining results:*

    Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

  - *Handling noisy or incomplete data:*
    - *The data stored in a database may reflect noise,* exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to overfit the data. As a result, the accuracy of the discovered patterns can be poor

Data Mining Issues

- **Mining methodology and user interaction issues**
  - *Pattern evaluation—the interestingness problem:*

    A data mining system can uncover thousands of patterns. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

Data Mining Issues

- **Performance issues**
  - *Efficiency and scalability of data mining algorithms:*

    To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. The running time of a data mining algorithm must be predictable and acceptable in large databases.
  - *Parallel, distributed, and incremental mining algorithms:*

    The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

Data Mining Issues

- **Issues relating to the diversity of database types:**
  - *Handling of relational and complex types of data:*
    - *Because relational databases and* data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. Specific data mining systems should be constructed for mining specific kinds of data.
  - *Mining information from heterogeneous databases and global information systems:*
    - Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semistructured, or unstructured data with diverse data semantics poses great challenges to data mining

Data Warehousing Challenges

- **Data Quality**

  When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.

- **Understanding Analytics**

  The powerful analytics tools and reports available through integrated data will provide credit union leaders with the ability to make precise decisions that impact the future success of their organizations. When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed. Envisioning these reports will be difficult for someone that hasn’t yet utilized a BI strategy and is unaware of its capabilities and limitations.

- **Quality Assurance**

  – The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader could make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate. The credit union will have to develop all of the steps required to complete a successful Software Testing Life Cycle (STLC), which will be a costly and time intensive process.

Data Warehousing Challenges

- Data Structuring and Systems Optimization
  - The correct processing of data requires structuring it in a way that makes sense for your future operations. As you add more and more information to your warehouse, structuring data becomes increasingly difficult and can slow down the process significantly. In addition, it will become difficult for the system manager to qualify the data for analytics. In terms of systems optimization, it is important to carefully design and configure data analysis tools. This will provide better results, making development decisions easier.
- Choosing the Right Type of Warehouse
  - Which one you choose will depend on your business model and specific goals.

Data Warehousing Challenges

- Balancing Resources
  - To receive the most benefit from data warehouse deployment, most businesses choose to allow multiple departments to access the system. This can add stress to the warehouse and decrease efficiency. However, implementing access control and security measures can help you balance the usefulness and performance of warehouse systems.
- Data Governance and Master Data
  - One mistake that some businesses make is a lack of investment in data governance and master data. Because information is one of your most important assets, it should be closely monitored. Implementing data governance allows you to clearly define ownership and ensures that shared data is both consistent and accurate.

Data Warehousing Challenges

- **Performance**
  - A data warehouse must be carefully designed to meet overall performance requirements. While the final product can be customized to fit the performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start.
- **Designing the Data Warehouse**
  - People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. Usually, there is a high level perception of what they want out of a data warehouse. However, they don't fully understand all the implications of these perceptions and, therefore, have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians building the data warehouse. The typical end result is a data warehouse which does not deliver the results expected by the user. Since the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery. The unfortunate outcome is greatly increased development fees.

Data Warehousing Challenges

- **User Acceptance**
  - People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization. Having a comprehensive user training program can ease this hesitation but will require planning and additional resources.
- **Cost**
  - A frequent misconception among credit unions is that they can build data warehouse in-house to save money. As the foregoing points emphasize, there are a multitude of hidden problems in building data warehouses. Even if a credit union adds a data warehouse “expert” to their staff, the depth and breadth of skills needed to deliver an effective result is simply not feasible with one or a few experienced professionals leading a team of non-BI trained technicians. The harsh reality is an effective do-it-yourself effort is very costly.

OLTP vs Data Warehouse

- **OLTP systems**
  - designed to maximize the transaction processing capacity.
  - commonly used in clerical data processing tasks, structured repetitive tasks, read update a few records.
  - isolation, recovery and integrity are critical.
- **Data warehouse**
  - holds data that is historical, detailed, and summarized to various levels and rarely subject to change.
  - designed to support relatively low numbers of transactions that are unpredictable in nature and require answers to queries that are *ad hoc,* unstructured, and heuristic.

DIFFERENCES BETWEEN OLTP AND DWH

OLTP	DWH
Designed to support business transactions support	Designed to support decision making process
Data is volatile	Data is non volatile
It holds current data	It holds historical data (5-10 years)
Detailed data	Summarized data
Normalized data	Denormalized data
Designed for running the business	Designed for analyzing the business
Clerical/End user access	Managerial access
E-R Modeling	Dimensional modeling
Transaction processing	Query processing
10MB-100GB Database size	100GB-2TB Database size

Applications of DWH

- **Banking Industry**
  - In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.
  - Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.
  - Analysis of card holder’s transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity.
- **Finance Industry**
  - revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

Applications of DWH

- **Consumer Goods Industry**
  - They are used for prediction of consumer trends, inventory management, market and advertising research.
  - In-depth analysis of sales and production is also carried out.
- **Government and Education**
  - The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.
  - The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers.
  - Criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.
  - Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management.

Applications of DWH

- **Healthcare**
  - All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.
- **Hospitality Industry**
  - A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services.
  - They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

Applications of DWH

- **Insurance**
  - The warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants.
  - The design of tailor-made customer offers and promotions is also possible through warehouses.
- **Manufacturing and Distribution Industry**
  - A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.
  - They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.
  - For the distributions, the supply chain management of products operates through data warehouses.

Applications of DWH

- **The Retailers**
  - Retailers serve as middlemen between producers and consumers.
  - They use warehouses to track items, their advertising promotions, and the consumers buying trends.
  - They also analyze sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.
- **Services Sector**
  - Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources.

Applications of DWH

- **Telephone Industry**
  - The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.
  - Analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.
- **Transportation Industry**
  - In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.
  - To analyze customer feedback, performance, manage crews on board as well as analyze customer financial reports for pricing strategies.

End of Module 1