Supporting Technology for DS Data Security and Privacy, https://www.youtube.com/watch?v=vymcO-XgPN0

UNIT IV

MDSS : **Supporting Technology for DS** Session 1

Data Security and Privacy

Introduction

Data security is the process of safeguarding digital information throughout its entire life cycle to protect it from corruption, theft, or unauthorized access. It covers everything—hardware, software, storage devices, and user devices; access and administrative controls; and organizations' policies and procedures

Data security is the practice of safeguarding digital information from unauthorized access, accidental loss, disclosure and modification, manipulation or corruption throughout its entire lifecycle, from creation to destruction.

This practice is key to maintaining the confidentiality, integrity and availability of an organization's data.

Confidentiality refers to keeping data private,

Integrity to ensuring data is complete and trustworthy, and

Availability to providing access to authorized entities.

Known collectively as the CIA triad, if any of the three components is compromised, companies can face reputational and financial damage. The CIA triad is the basis upon which a data security strategy is built. Such a strategy must encompass policies, technologies, controls and procedures that protect data created, collected, stored, received and transmitted by an enterprise.

Why is Data Security Important?

Data is the lifeblood of every organization. It informs decision-making, finds solutions to problems, improves the efficiency and efficacy of operations, boosts customer service and informs marketing efforts, reduces risks, increases productivity, enhances collaboration and, in the end, is instrumental in increasing revenue and profit.

Data is often referred to as a company's crown jewels; for something so essential, its protection must be taken seriously.

Much like Coca-Cola's secret recipe that is locked away in a vault, Hershey's secret lab that concocts its famous Kisses and KFC's famous yet unknown 11 herbs and spices, it is crucial to keep certain data from prying eyes. It's not always as easy as putting something under lock and key -- especially in a digital environment. Multiple employees, stakeholders and partners need access to the data that enterprises value so highly. But more people having access means more chances for things to go wrong.

Data Breaches : which occur when data is accessed in an unauthorized manner, are a major concern for organizations of all shapes, sizes and industries. In fact, 63% of respondents to a KPMG study said they suffered a data breach or cyber incident in 2021 -- and that number is only projected to grow.

Data breaches are attributed to a number of cyber incidents, including the following:

- accidental leaks or exposures
- phishing attacks
- distributed denial-of-service attacks
- physical breaches
- lack of access controls
- backdoors

Data breaches can result in hefty remediation costs, as well as expenses stemming from downtime and lost business. Regulatory and legal fines may also be levied. In worst-case scenarios, companies can go bankrupt or out of business.

Data security is an important component in data compliance, the process that identifies governance and establishes policies and procedures to protect data. The process involves selecting applicable standards and implementing controls to achieve the criteria defined in those standards. Regulatory compliance, which refers to organizations following local, state, federal, international and industry laws, policies and regulations -- is related to data compliance. Regulatory compliance standards require the use of certain controls and technologies to meet the criteria defined in them. The following are some of the most common compliance regulations:

Europe's General Data Protection Regulation (GDPR)

California Consumer Protection Act (CCPA)

Health Insurance Portability and Accountability Act (HIPAA)

Sarbanes-Oxley Act (SOX)

Payment Card Industry Data Security Standard (PCI DSS)

HIPAA, for example, outlines provisions to safeguard medical information in the U.S. Among other mandates, healthcare organizations must adhere to standards for patient data security or else face noncompliance fines and penalties.

PCI DSS is a global standard aimed at protecting credit, debit and cash card transaction data. It sets guidelines for cardholder data, access controls and networks that process payment information.

Many regulations are subject to audits, during which organizations must prove they adhere to the policies set out in a given regulation.

Beyond preventing breaches and complying with regulations, data security is important to maintaining customer trust, building relationships and preserving a good company image. It is also key to sustaining a competitive advantage. After all, if everyone had the recipe and the means to make Hershey's Kisses, the chocolatier would be out a considerable amount of money.

Types of Data Security

Before an organization can secure data, it has to know what data it has. This is where a data inventory --a record of all the data created, used and stored by a company -- is key. The process starts with data discovery, or learning what and where the data is.

Data classification follows, which involves labelling data to make it easier to manage, store and secure. The four standard data classification categories are as follows:

- public information
- confidential information
- sensitive information
- personal information

Data is often further broken down by businesses using common classification labels, such as "business use only" and "secret.“

Sensitive data is often classified as confidential or secret. It includes these types of data:

- personally identifiable information
- protected health information
- electronic protected health information
- PCI data
- intellectual property

Compounding the difficulty of doing data inventory and classification is that data can reside in many locations -- on premises, in the cloud, in databases and on devices, to name a few. Data also can exist in three states:

1. in motion, meaning data that is being transported;

2. at rest, meaning data that is being stored, or data that is at its destination -- i.e., not transported or in use; and

3. in use, meaning data that is being written, updated, changed and processed -- i.e., not being transported or stored.

Because no single form of data exists, no single magic-bullet technique can secure all data. A defense-in-depth data security strategy is made up of a combination of tools, techniques and policies. Must-have data security technologies include the following:

- ❑ encryption
- ❑ data masking
- ❑ access control
- ❑ data loss prevention (DLP)
- ❑ data backup and resiliency

Encryption

Encryption is the process of converting readable plaintext into unreadable ciphertext using an encryption algorithm, or cipher. If encrypted data is intercepted, it is useless as it cannot be read or decrypted by anyone who does not have the associated encryption key.

Symmetric and asymmetric encryption are two commonly used ciphers:

Symmetric encryption uses a single secret key for both encryption and decryption. The Advanced Encryption Standard is the most commonly used algorithm in symmetric key cryptography.

Asymmetric encryption uses two interdependent keys: a public key to encrypt the data and a private key to decrypt the data. The Diffie-Hellman key exchange and Rivest-Shamir-Adleman are two common asymmetric algorithms.

Both symmetric and asymmetric encryption have pros and cons. Security expert Michael Cobb explains the differences between the ciphers and discusses why a combination of the two might be the fastest, most secure encryption option.

Data masking

Data masking involves obscuring data so it cannot be read. Masked data looks similar to the authentic data set but reveals no sensitive information. Legitimate data is replaced so the masked data maintains the characteristics of the data set as well as referential integrity across systems, thereby ensuring the data is realistic, irreversible and repeatable.

Below are some common data masking techniques:

- scrambling
- substitution
- shuffling
- data aging
- variance
- masking out
- nullifying

Data masking is useful when certain data is needed for software testing, user training and data analysis -- but not the sensitive data itself.

While the end result of encryption and masking are the same -- both create data that is unreadable if intercepted -- they are quite different.

Data loss prevention

An integral tool for any enterprise security strategy is a DLP platform. It monitors and analyzes data for anomalies and policy violations. Its many features can include data discovery, data inventory, data classification and analysis of data in motion, at rest and in use. Many DLP tools integrate with other technologies, such as SIEM systems, to create alerts and automated responses.

Data backup

Data backup involves creating copies of files and databases to a secondary, and often tertiary and quaternary, location. If the primary data fails, is corrupted or gets stolen, a data backup ensures it can be returned to a previous state rather than be completely lost. Data backup is essential to disaster recovery plans.

Resilience is another strategy growing in popularity. The ability of an organization to adapt and recover following a cyber incident equates to how resilient it is.

End of session 1

Data Security, Privacy & Policies,
https://www.youtube.com/watch?v=kCJL-LgIXkY

UNIT IV

MDSS : **Supporting Technology for DS** Session 2

Data security vs. Data Privacy vs. Data Protection

Data security, data privacy and data protection are overlapping but technically distinct concepts.

Data security. Data security has a broader scope, aiming to protect digital information not just from unauthorized access but also from intentional loss, unintentional loss and corruption. While data privacy primarily focuses on the confidentiality part of the CIA triad, data security is equally concerned with information's integrity and accessibility.

For example, imagine threat actors obtain a confidential file, but encryption successfully prevents them from reading the data. The information itself stays inaccessible, and data privacy remains intact. The attackers are still able to corrupt or destroy the illegible file, however, which is a security failure.

Data privacy. The goal of data privacy is to make sure the ways an organization collects, stores and uses sensitive data are responsible and in compliance with legal regulations. Privacy policies and measures prevent unauthorized parties from accessing data, regardless of their motivation and whether they are internal end users, third-party partners or external threat actors.

Data protection. Data protection ensures digital information is backed up and recoverable if it's lost, corrupted or stolen. Data protection is an important part of a larger data security strategy, serving as a last resort if all other measures fail.

What are data security risks and challenges?

In 2017, *The Economist* declared "The world's most valuable resource is no longer oil, but data." Unfortunately, data is more difficult to protect and easier to steal, and it presents enormous opportunity to not just businesses but also criminals.

Today's enterprises face an uphill battle when it comes to securing their data. Consider the following perennial risks and challenges.

Insider threats. One of the biggest threats to data security is the enterprise end user, whether that's a current or former employee, third-party partner or contractor. Malicious insiders sometimes use their legitimate access privileges to corrupt or steal sensitive data, either for profit or to satisfy personal grudges.

Unintentional insider threats are no less dangerous. An innocent click on a link in a phishing email could compromise a user's credentials or unleash ransomware or other malware on corporate systems. In fact, in nearly 40% of data breaches, attackers used either compromised credentials or phishing as initial attack vectors, according to the Ponemon Institute's 2021 "Cost of a Data Breach" report, sponsored by IBM.

Simple end-user negligence or carelessness -- absent any malicious threat actor -- can also result in accidental exposure of sensitive data. An employee might email confidential information to the wrong person, for example, or upload it to an unprotected cloud account. In addition, someone could lose a laptop and fail to report it to IT, leaving the device vulnerable to whoever happens to find it.

Misconfigurations. Technical misconfigurations pose another major threat, regularly resulting in accidental exposure of confidential data sets. The Ponemon Institute found cloud misconfigurations alone were responsible for 15% of data breaches in 2021.

Third-party risk. An organization is arguably only as secure as its least secure third-party partner, whether that's a supplier, contractor or customer. Consider the infamous Solarwinds supply chain attack, which enabled threat actors to target the vendor's customers' networks. Organizations point to vulnerable third-party software as the initial attack vector in 14% of data breaches, according to the Ponemon Institute.

Other top data security challenges organizations face today include mushrooming enterprise data footprints, inconsistent data compliance laws and increasing data longevity.

Data security best practices: How to secure data

To effectively mitigate risk and grapple with the challenges listed above, enterprises should follow established data security best practices.

Organizations must start with an inventory of what data they have, where it is and how their applications use it. Only once they understand what needs protecting can they effectively protect it.

Formal data risk assessments and regular security audits can help companies identify their sensitive data, as well as how their existing security controls might fall short.

Next, enterprises should weigh how they will close any data security gaps they have flagged. Experts recommend considering tools, technologies and techniques such as the following:

Access control. Regardless of data's location and state, the ability to limit who can read, edit, save and share it is the bedrock of data security.

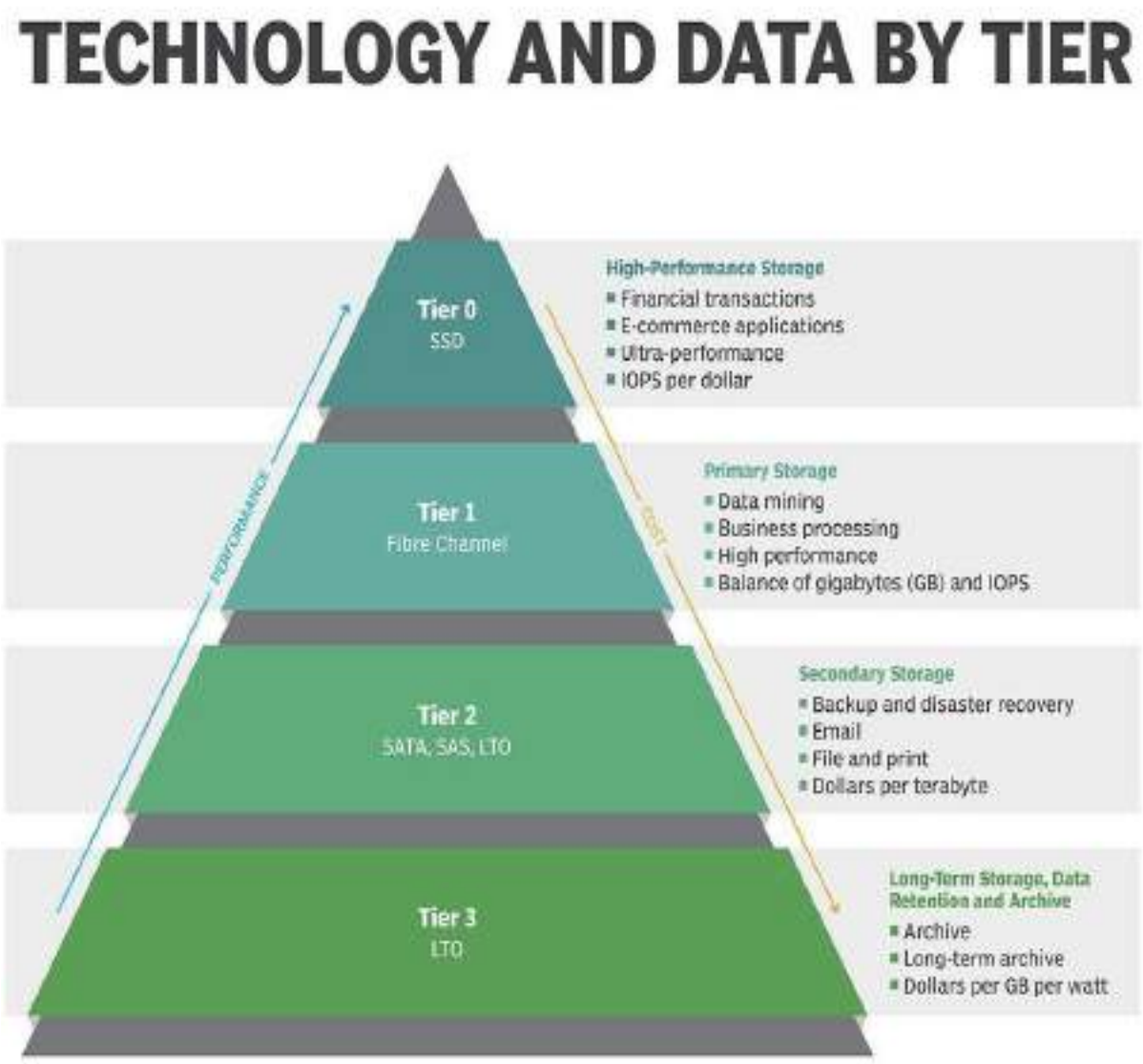
Cloud security. While cloud use has significant benefits, such as scalability and cost savings, it also carries plenty of risk. Enterprises that use SaaS, IaaS and PaaS must contend with a number of cloud security concerns, including credential and key management, data disclosure and exposure, and cloud storage exfiltration.

Data backup. The best advice is to expect the best and plan for the worst. Data backup acts as an insurance policy in case digital information is corrupted, lost or stolen, as in the case of a ransomware attack.

Data lifecycle management. DLM is an automated approach to keeping massive amounts of digital information accurate, confidential, secure and available -- and destroying it in a safe and timely fashion, in keeping with enterprise policies -- all while meeting relevant compliance requirements. DLM policies are based on data attributes such as type, size, age and classification. The main phases of the data lifecycle in a DLM framework include the following:

- generation and collection
- processing and storage
- usage

- archiving
- destruction



DLM tools can automatically sort data into separate tiers based on specified policies.

This lets enterprises use storage resources efficiently and effectively by assigning top-priority data to high-performance storage, for example.

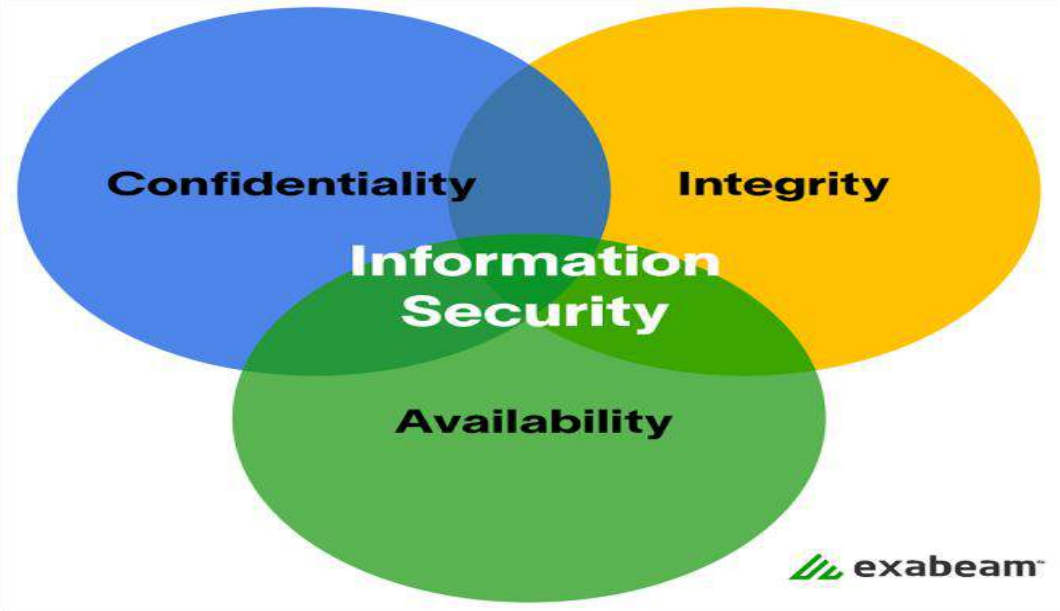
Patch management. Leaving a known vulnerability unpatched is like failing to fix a broken lock on the side door of an otherwise secure home. Patch software quickly and often to limit the ways attackers can gain access to enterprise property.

Security awareness training. Intentional and unintentional mistakes of staff, contractors and partners represent one of the greatest threats to data security. Security awareness training is therefore of utmost importance to educate users on organizational security policies and topics such as phishing attacks.

User behavior analytics. UBA, also known as user and entity behavior analytics (UEBA), flags attempts to gain unauthorized or unusual levels of access to sensitive data. Among top UEBA use cases, the technology can help detect lateral network attacks, identify compromised user accounts and uncover insider threats.

A security policy is a document that outlines the rules and expectations for maintaining an organization's data **confidentiality, integrity, and availability**. Security policies are important because they:

1) Guide the implementation of technical controls
2) Set clear expectations
3) Help meet regulatory and compliance requirements
4) Improve organizational efficiency
5) Protect an organization's assets, both physical and digital
6) Identify all company assets and all threats to those assets



Security policies are living documents that are continuously updated and changing as technologies, vulnerabilities, and security requirements change. It is important to tailor IT security policies to meet the specific needs of an organization. An organization's size, nature, structure, and culture should inform the development of its security policies

Here are some examples of security policies:

Acceptable Use Policy (AUP)

Defines how computer systems can be used, including computers, mobile devices, networks, email systems, and the internet

Remote Access Policy

Lowers the risk of unauthorized access and extends the responsibility of your employee while accessing the company's network system outside the work premise

Network Security Policy

Ensures the confidentiality, integrity, and availability of data on company's systems by

following a specific procedure for conducting information system and network activity review on a periodic basis

Password Policy

Governs the creation, management, and use of passwords within the network

The importance of an data (information) security policy

Information security policies can have the following benefits for an organization:

Facilitates data integrity, availability, and confidentiality —effective information security policies standardize rules and processes that protect against vectors threatening data integrity, availability, and confidentiality.

Protects sensitive data — Information security policies prioritize the protection of intellectual property and sensitive data such as personally identifiable information (PII).

Minimizes the risk of security incidents — An information security policy helps organizations define procedures for identifying and mitigating vulnerabilities and risks. It also details quick responses to minimize damage during a security incident.

Executes security programs across the organization — Information security policies provide the framework for operationalizing procedures.

Provides a clear security statement to third parties — Information security policies summarize the organization's security posture and explain how the organization protects IT resources and assets. They facilitate quick response to third-party requests for information by customers, partners, and auditors.

Helps comply with regulatory requirements — Creating an information security policy can help organizations identify security gaps related to regulatory requirements and address them.

12 Elements of an Data (Information) Security Policy

A security policy can be as broad as you want it to be, from everything related to IT security and the security of related physical assets, but enforceable in its full scope. The following list offers some important considerations when developing a security policy:

1. Purpose

First state the purpose of the policy, which may be to:

Create an overall approach to information security, especially as touches standards, security requirements, and best practices adopted by the organization.

Detect and preempt information security breaches such as misuse of networks, data, applications, and computer systems.

Maintain the reputation of the organization, and uphold ethical and legal responsibilities and applicable governance.

Respect employee and customer rights, including how to react to inquiries and complaints about non-compliance.

2. Audience

Define the audience to whom the information security policy applies. You may also specify which audiences are out of the scope of the policy (for example, staff in another business unit which manages security separately may not be in the scope of the policy).

3. Information security objectives

Guide your management team to agree on well-defined objectives for strategy and security. Information security focuses on three main objectives:

Confidentiality — Only authenticated and authorized individuals can access data and information assets.

Integrity — Data should be intact, accurate and complete, and IT systems must be kept operational.

Availability — Users should be able to access information or systems when needed.

4. Authority and Access Control Policy

Hierarchical Pattern — A senior manager may have the authority to decide what data can be shared and with whom. The security policy may have different terms for a senior manager vs. a junior employee or contractor. The policy should outline the level of authority over data and IT systems for each organizational role.

Network Security Policy — Critical patching and other threat mitigation policies are approved and enforced. Users are only able to access company networks and servers via unique logins that demand authentication, including passwords, biometrics, ID cards, or tokens. You should monitor all systems and record all login attempts.

5. Data Classification

The policy should classify data into categories, which may include :

“top secret,” “secret,” “confidential,” and **“public.”**

The objectives for classifying data are:

- To understand which systems and which operations and applications touch on the most sensitive and controlled data, to properly design security controls for that hardware and software (see next point 6.)
- To ensure that sensitive data cannot be accessed by individuals with lower clearance levels

To protect highly important data, and avoid needless security measures for unimportant data.

6. Data support and operations

Data protection regulations — systems that store personal data, or other sensitive data — must be protected according to organizational standards, best practices, industry compliance standards, and relevant regulations. Most security standards require, at a minimum, encryption, a firewall, and anti-malware protection.

Data backup — Encrypt data backup according to industry best practices, both in motion and at rest. Securely store backup media, or move backup to secure cloud storage.

Movement of data — Only transfer data via secure protocols. Encrypt any information copied to portable devices or transmitted across a public network.

7. Security awareness and behavior

Share IT security policies with your staff. Conduct training sessions to inform employees of your security procedures and mechanisms, including data protection measures, access protection measures, and sensitive data classification.

Social engineering — Place a special emphasis on the dangers of social engineering attacks (such as phishing emails or informational requests via phone calls). Make all employees responsible for noticing, preventing, and reporting such attacks.

Clean desk policy — Secure laptops with a cable lock. Shred sensitive documents that are no longer needed. Keep printer areas clean so documents do not fall into the wrong hands.

Work with HR to define how the internet should be restricted both on work premises and for remote employees using organizational assets.

Do you allow YouTube, social media websites, etc.? Block unwanted websites using a proxy.

8. Encryption Policy

Encryption involves encoding data to keep it inaccessible to or hidden from unauthorized parties. It helps protect data stored at rest and in transit between locations and ensure that sensitive, private, and proprietary data remains private. It can also improve the security of client-server communication. An encryption policy helps organizations define:

- The devices and media the organization must encrypt
- When encryption is mandatory
- The minimum standards applicable to the chosen encryption software

9. Data Backup Policy

A data backup policy defines rules and procedures for making backup copies of data. It is an integral component of overall data protection, business continuity, and disaster recovery strategy. Here are key functions of a data backup policy:

- Identifies all information the organization needs to back up
- Determines the frequency of backups, for example, when to perform an initial

full backup and when to run incremental backups

- Defines a storage location holding backup data
- Lists all roles in charge of backup processes, for example, a backup administrator

and members of the DS (IT) team

10. Responsibilities, Rights, and Duties of Personnel

Appoint staff to carry out user access reviews, education, change management, incident management, implementation, and periodic updates of the security policy. Responsibilities should be clearly defined as part of the security policy.

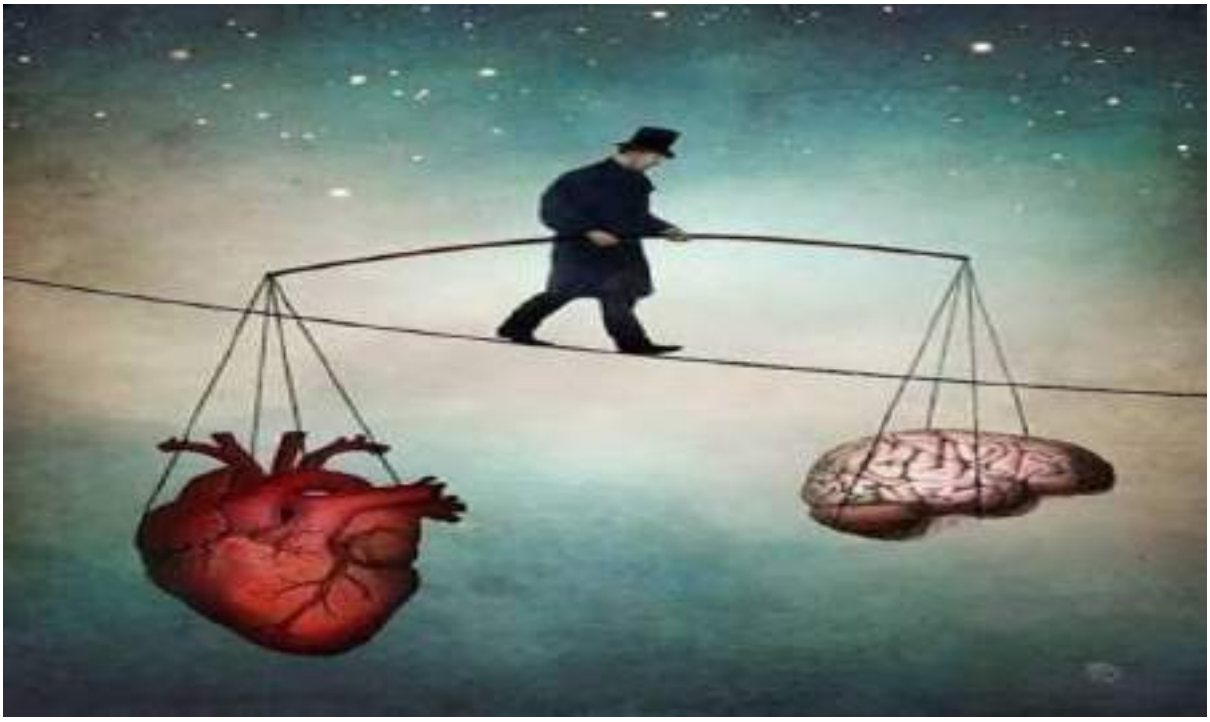
11. System Hardening benchmarks

The information security policy should reference security benchmarks the organization will use to harden mission-critical systems, such as the Center for Information Security (CIS) benchmarks for Linux, Windows Server, AWS, and Kubernetes.

12. References to regulations and compliance standards

The information security policy should reference regulations and compliance standards that impact the organization, such as :

- General Data Protection Regulation (GDPR),
- California Consumer Privacy Act (CCPA),
- Payment Card Industry Data Security Standard (PCI DSS),
- Sarbanes-Oxley Act (SOX), and
- Health Insurance Portability and Accountability Act (HIPAA).



End of session 2

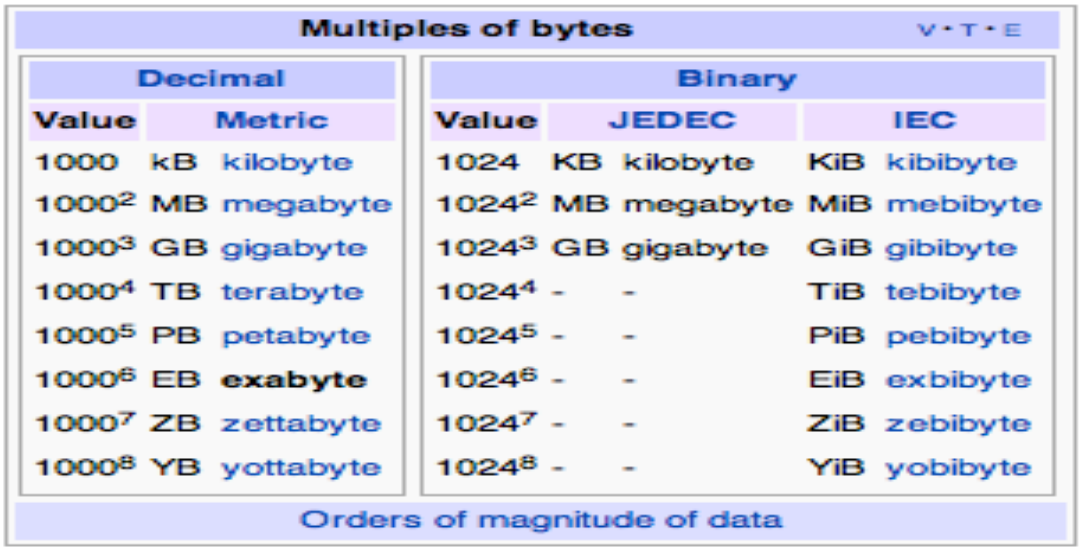
Data Mining Techniques in MDSS - Introduction,
https://www.youtube.com/watch?v=LtjqoFEyWJA

UNIT IV

MDSS : Data Mining - Concepts and Techniques Session 3

Data Mining Techniques in MDSS - Introduction

Big Data Age



Multiples of bytes							
Decimal			Binary				
Value	Metric		Value	JEDEC		IEC	
1000	kB	kilobyte	1024	KB	kilobyte	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MB	megabyte	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GB	gigabyte	GiB	gibibyte
1000^4	TB	terabyte	1024^4	-	-	TiB	tebibyte
1000^5	PB	petabyte	1024^5	-	-	PiB	pebibyte
1000^6	EB	**exabyte**	1024^6	-	-	EiB	exbibyte
1000^7	ZB	zettabyte	1024^7	-	-	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	-	-	YiB	yobibyte

Orders of magnitude of data

Domain for Non-Personal Big Data are extensive

- Physics
- Astronomical
- Genomics
- Cognitive (brain mapping)
- Meteorological
- Intelligent Machines

Domain for Personal Big Data

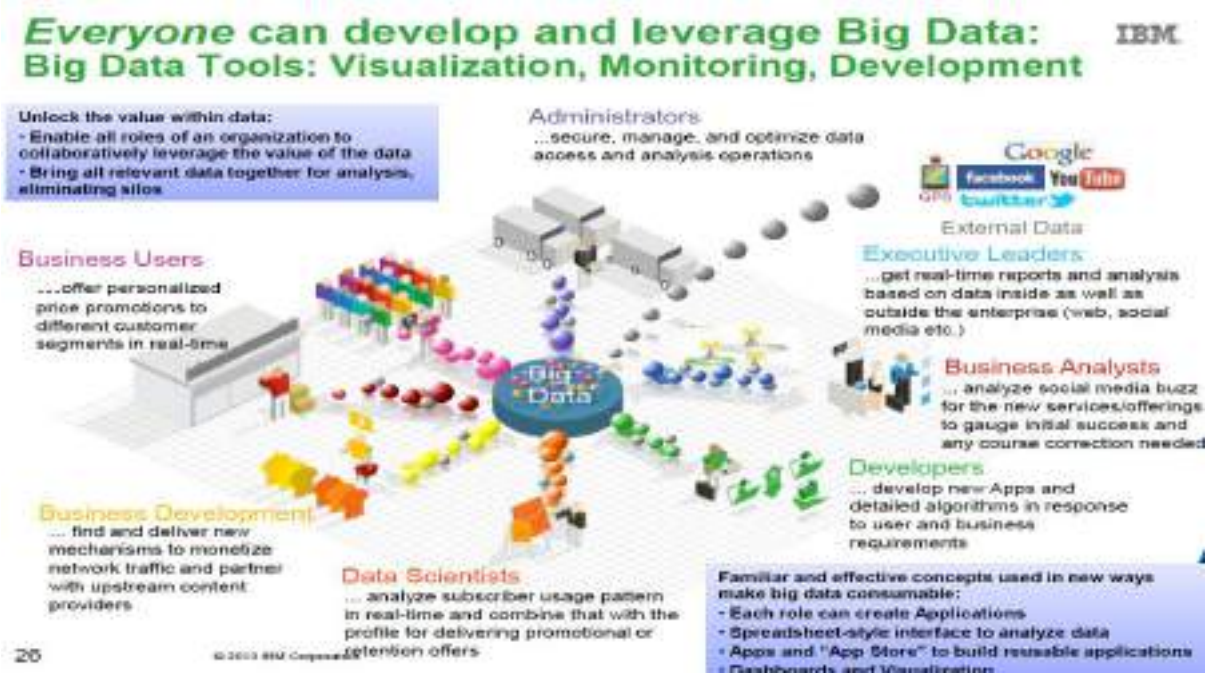
- Political
- Social (Facebook, Linked, etc.)

Domain for Global Big Data cross non-personal and personal domains

- Surveillance
- Operational (electric grid)

Big Data Applications

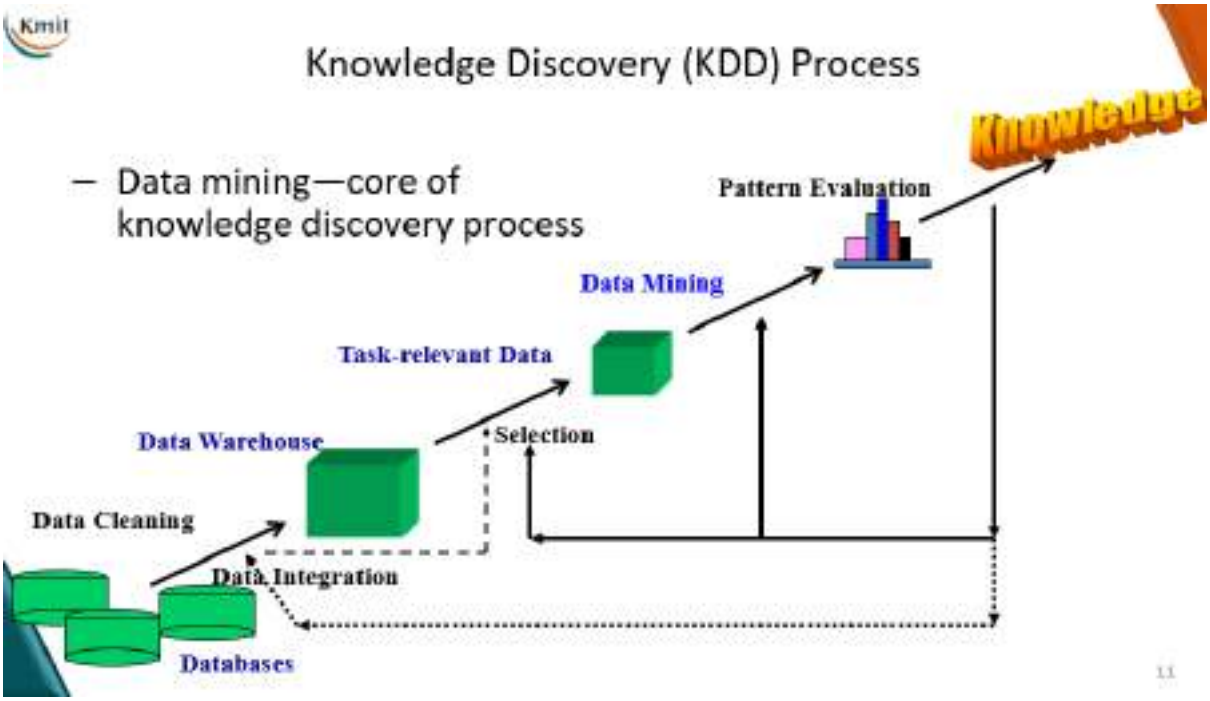
- Smart Healthcare
- Traffic Control
- Finance
- Log Analysis
- Homeland Security
- Telecom industry
- Search Quality
- Trading Analysis
- Fraud and Risk detection
- Retailing



DWDM- Introduction

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems



End of session 3

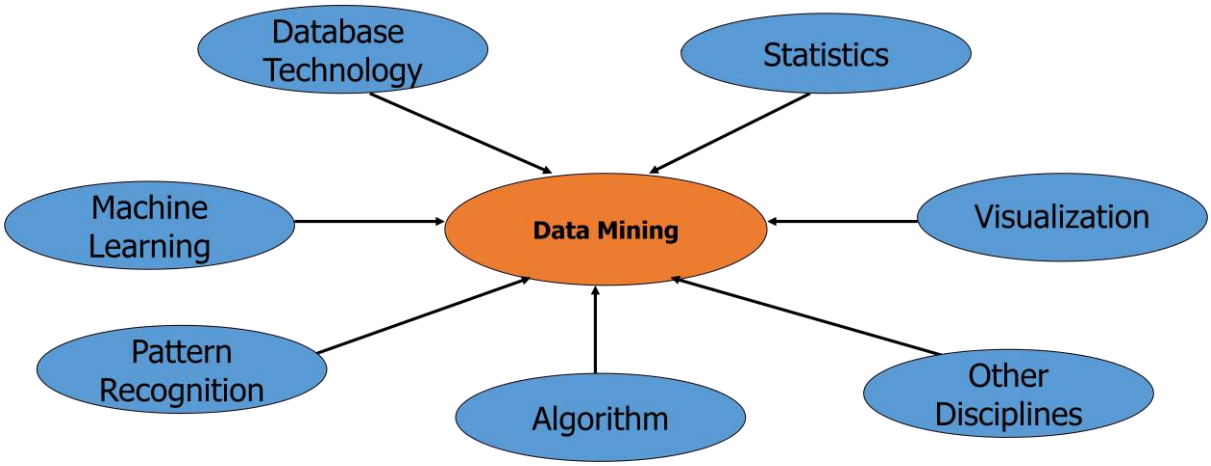
Mining Patterns, Functions and Issues,
https://www.youtube.com/watch?v=FDovRHZL_8Q

UNIT IV

MDSS : **Supporting Technology for DS** Session 4

Mining Patterns, Functions and Issues

Data Mining: Confluence of Multiple Disciplines



Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data

-
    - Heterogeneous databases and legacy databases
    - Spatial, spatiotemporal, multimedia, text and Web data
    - Software programs, scientific simulations
- New and sophisticated applications

Multi-Dimensional View of Data Mining

- **Data to be mined**
    - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
    - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
    - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
    - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
    - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining (BJP <-> NON BJP)
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database

- **Text databases**
- **The World-Wide Web**

Data Mining Functionalities

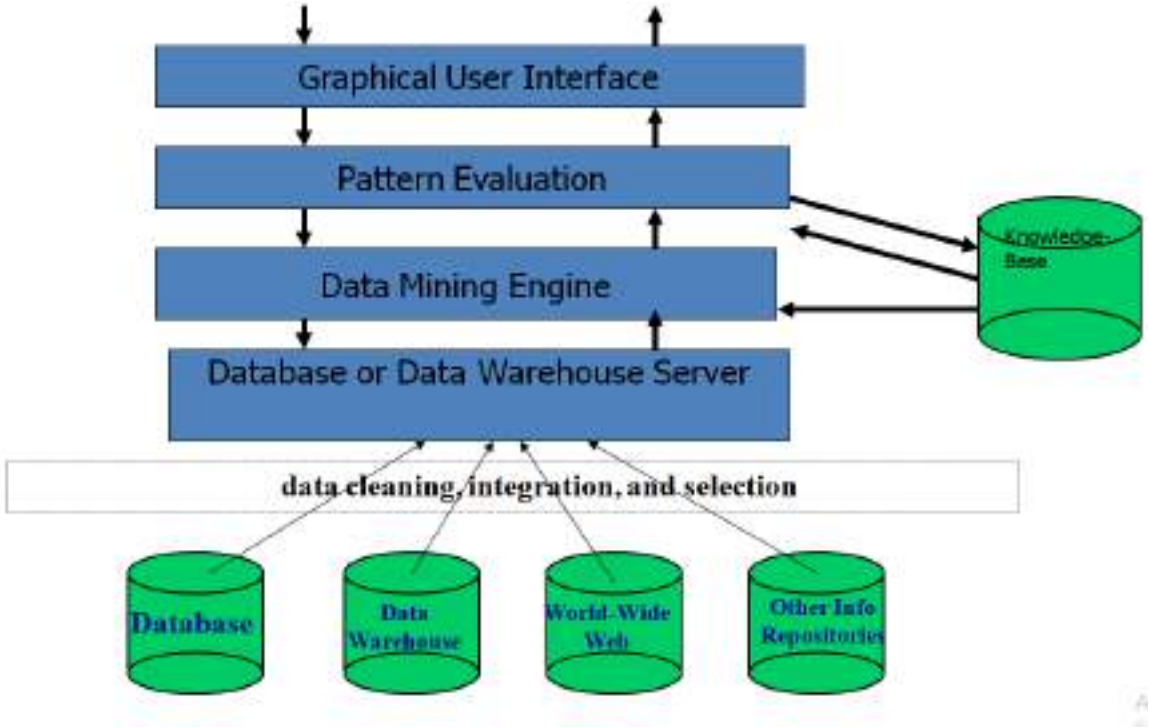
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., human and monkey?
  - Good income VS poor?
- Frequent patterns, association, correlation vs. causality
  - Diaper → Beer [0.5%, 75%], Education -> Income
- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (economy), cars based on (gas mileage), internet news (Google News), product (Amazon)
  - Predict some stock price, traffic jam
- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns, terrain images?
  - Maximizing intra-cluster similarity & minimizing intercluster similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis

- Trend and deviation: e.g., political polls? Who will win republican nomination in 2016? Who will win presidential election?
- Sequential pattern mining: e.g., video mining -> identify objects
- Periodicity analysis: climate change?
- Similarity-based analysis: future of the Apple company?

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining
  - Protection of data security, integrity, and privacy

Architecture: Typical Data Mining System



Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A natural evolution of database technology and machine learning, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining

End of session 4

Association, Correlation and Frequent Patterns, https://www.youtube.com/watch?v=7MLBtX4U-uI

UNIT IV

MDSS : **Supporting Technology for DS** Session 5

Frequent Patterns, Association and Correlations

Imagine that you are a Sales Manager at APANA BAZAR and you are talking to a customer who recently bought a Laptop with built-in digital camera.

What should you recommend to him | her next?

Information about which products are frequently purchased by your customers following their purchases of a Laptop with built-in digital camera in sequence would be very helpful in making your recommendation.

Frequent patterns and association rules are the knowledge that you want to mine in such a scenario.

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset.

A subsequence, such as buying first a PC(laptop), then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.

A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern.

Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.

Moreover, it helps in data classification, clustering, and other data mining tasks.

Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research and has paved ways into DS to mine the various pattterns.

Market Basket Analysis: A Motivating Example

Frequent itemset mining leads to the discovery of associations and correlations among

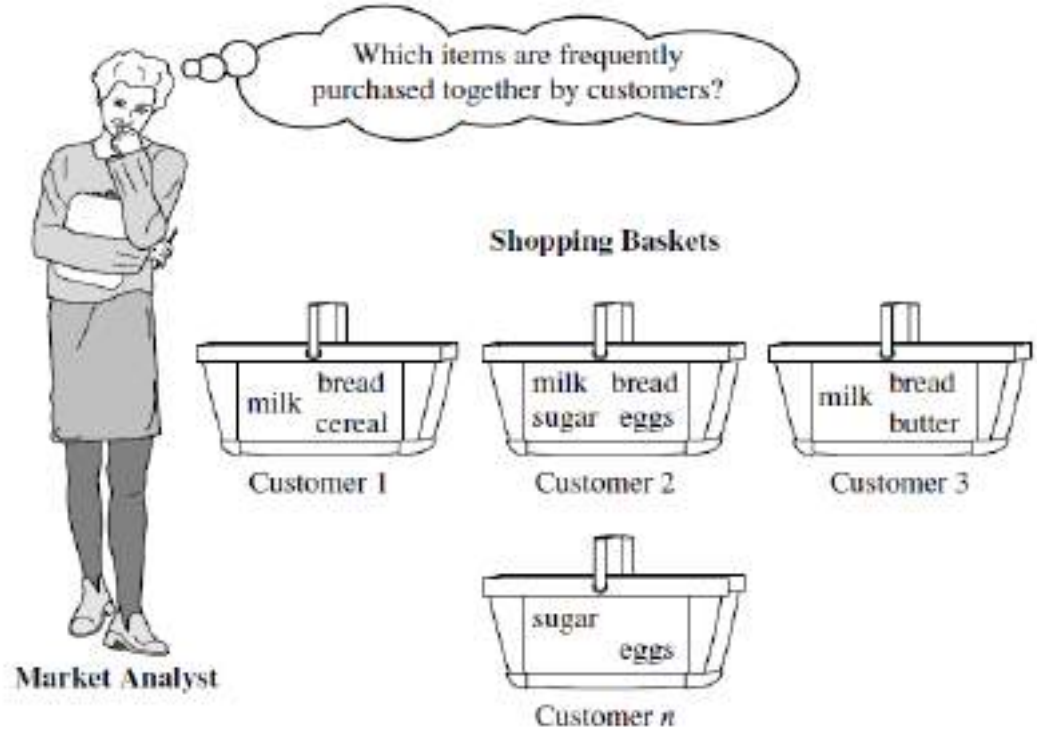
items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining

such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.

A typical example of frequent itemset mining is **market basket analysis.** This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" (Fig.).

The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. **For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket?**

This information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.



Example : Market basket analysis.

Suppose, as manager of ***Apna Bazar ,*** *you would* like to learn more about the buying

habits of your customers. Specifically, you wonder,

"Which groups or sets of items are customers likely to purchase on a given trip to the store?"

To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity to further encourage the combined sale of such items. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items.

In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software, and may decide to purchase a home security system as well.

Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.

If we think of the universe as the set of items available at the store, then each item has a

Boolean variable representing the presence or absence of that item.

Each basket can then be represented by a Boolean vector of values assigned to these variables.

The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently *associated* or purchased together.

These patterns can be represented in the form of ***association rules*.**

E.g., Information that customers who purchase computers(Laptop) also tend to buy antivirus software at the same time is represented in the following association rule:

$$computer \Rightarrow antivirus\_software\ [support = 2\%, confidence = 60\%].$$

Rule **support** and **confidence** are **two measures of rule interestingness.**

They respectively reflect the usefulness and certainty of discovered rules.

A support of 2% for Rule means that 2% of all the transactions under analysis show that computer(Laptop) and antivirus software are purchased together.

A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold.** These thresholds can be a set by users or

domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.

End of session 5

Association - Support and Confidence Rules, https://www.youtube.com/watch?v=VcJ05bbjUmQ

Association - Support and Confidence Rules

If we think of the universe as the set of items available at the store, then each item has a

Boolean variable representing the presence or absence of that item.

Each basket can then be represented by a Boolean vector of values assigned to these variables.

The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.

These patterns can be represented in the form of association rules.

E.g., Information that customers who purchase computers(Laptop) also tend to buy antivirus software at the same time is represented in the following association rule:

$$computer \Rightarrow antivirus\_software\ [support = 2\%, confidence = 60\%].$$

Rule **support** and **confidence** are **two measures of rule interestingness.**

They respectively reflect the usefulness and certainty of discovered rules.

A support of 2% for Rule means that 2% of all the transactions under analysis show that computer(Laptop) and antivirus software are purchased together.

A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold.** These thresholds can be a set by users or

domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.

Transaction	Items
1	Milk, Bread, Butter
2	Bread, Cheese
3	Bread, Jam
4	Milk, Bread, Cheese
5	Milk, Jam
6	Bread, Jam
7	Milk, Jam
8	Milk, Bread, Jam, Butter
9	Milk, Bread, Jam

	Milk	Bread	Butter	Beer	Diapers
0	Yes	Yes	Yes	No	No
1	No	Yes	Yes	No	No
2	Yes	Yes	No	No	Yes
3	No	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes	Yes
5	Yes	Yes	Yes	No	Yes

{Bread, Butter} → Milk (3/5 confidence)

Beer → Diapers (1/2 confidence)

$$support(A \Rightarrow B) = P(A \cup B)$$

$$confidence(A \Rightarrow B) = P(B|A).$$

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ be an itemset. Let D, the task-relevant data, be a set of database transactions where each transaction T is a nonempty itemset such that $T \subseteq \mathcal{I}$. Each transaction is associated with an identifier, called a *TID*. Let A be a set of items. A transaction T is said to contain A if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, $A \neq \emptyset$, $B \neq \emptyset$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with **support** s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the *union* of sets A and B say, or, both A and B). This is taken to be the probability, $P(A \cup B)$.[1] The rule $A \Rightarrow B$ has **confidence** c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is,

Rules that satisfy both a minimum support threshold (***min sup****) and a minimum confidence* threshold (***min conf*** *) are called* ***strong.*** *By convention, we write* ***support*** *and* **confidence** values so as to occur between 0% and 100%, rather than 0 to 1.0.

A set of items is referred to as an **itemset.** An **itemset** that contains ***k*** *items is a* ***k-itemset.***

The set **{computer, antivirus software}** is *a* ***2-itemset. The occurrence frequency***

of an **itemset** is the ***number of transactions that contain the itemset*****.**

This is also known, simply, as the **frequency, support count, or count** of the **itemset.**

Note that the **itemset** support defined

is sometimes referred to as ***relative support,***

whereas the occurrence frequency is called the **absolute support.**

If the relative support of an itemset I satisfies a prespecified minimum support threshold

(i.e., the absolute support of I satisfies the corresponding minimum support count threshold), then I is a frequent itemset. The set of frequent k-itemsets is commonly denoted by Lk.

$$support(A \Rightarrow B) = P(A \cup B)$$
$$confidence(A \Rightarrow B) = P(B|A).$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Equation (6.4) shows that the confidence of rule $A \Rightarrow B$ can be easily derived from the support counts of A and $A \cup B$. That is, once the support counts of A, B, and $A \cup B$ are found, it is straightforward to derive the corresponding association rules $A \Rightarrow B$ and $B \Rightarrow A$ and check whether they are strong. Thus, the problem of mining association rules can be reduced to that of mining frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

1. **Find all frequent itemsets:** By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, *min_sup*.
2. **Generate strong association rules from the frequent itemsets:** By definition, these rules must satisfy minimum support and minimum confidence.

Frequent Pattern

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?

- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

End of session 6

Apriory Algorithm,
https://www.youtube.com/watch?v=D7LUdSmP_98

UNIT IV

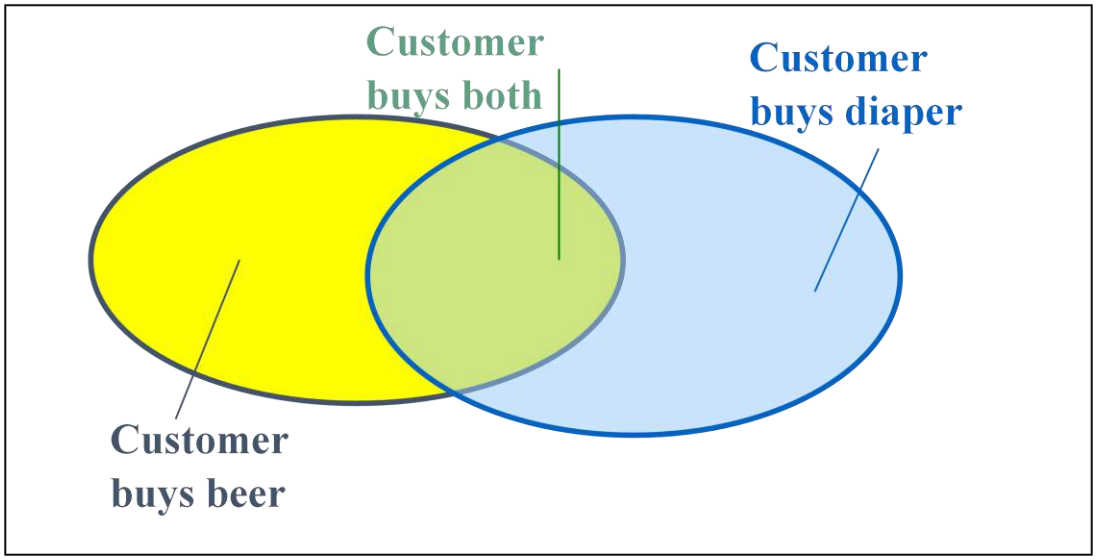
MDSS : **Supporting Technology for DS** Session 7

Apriory Algorithm

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications

Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold

Additional interestingness measures can be applied for the discovery of correlation relationships between associated items, as will be discussed in Section 6.3. Because the second step is much less costly than the first, the overall performance of mining association rules is determined by the first step.

A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support (*min_sup*) threshold, especially when *min_sup* is set low. This is because if an itemset is frequent, each of its subsets is frequent as well. A long itemset will contain a combinatorial number of shorter, frequent sub-itemsets. For example, a frequent itemset of length 100, such as $\{a_1, a_2, \ldots, a_{100}\}$, contains $\binom{100}{1} = 100$ frequent 1-itemsets: $\{a_1\}, \{a_2\}, \ldots, \{a_{100}\}$; $\binom{100}{2}$ frequent 2-itemsets: $\{a_1, a_2\}, \{a_1, a_3\}, \ldots, \{a_{99}, a_{100}\}$; and so on. The total number of frequent itemsets that it contains is thus

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}. \quad (6.5)$$

This is too huge a number of itemsets for any computer to compute or store. To overcome this difficulty, we introduce the concepts of *closed frequent itemset* and *maximal frequent itemset*.

- A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{1}{1}\binom{0}{0}\binom{0}{0} = 2^{100} - 1 = 1.27*10^{30}$ sub-patterns!
- Solution: *Mine closed patterns and max-patterns instead*
- An itemset X is closed if X is *frequent* and there exists *no super-pattern* Y ⊃ X, *with the same support* as X
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ⊃ X
- Closed pattern is a lossless compression of freq. patterns
  - Reducing the # of patterns and rules

Apriori: A Candidate Generation & Test Approach

- **Apriori pruning principle: If there is any itemset which is infrequent,**

its superset should not be generated/tested!

- **Method:**
  - **Initially, scan DB once to get frequent 1-itemset**
  - **Generate length (k+1) candidate itemsets from length k frequent itemsets**
  - **Test the candidates against DB**
  - **Terminate when no frequent or candidate set can be generated**

Frequent Item Set Example

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

I1 = Bread
I2 = Butter
I3 = Milk
I4 = Umbrella
I5 = Shoe

T100	**I1**, **I2**, **I5**
T200	**I2**, **I4**
T300	**I2**, **I3**
T400	**I1**, **I2**, **I4**
T500	**I1**, **I3**
T600	**I2**, **I3**
T700	**I1**, **I3**
T800	**I1**, **I2**, **I3**, **I5**
T900	**I1**, **I2**, **I3**

Transaction database : ***D***.

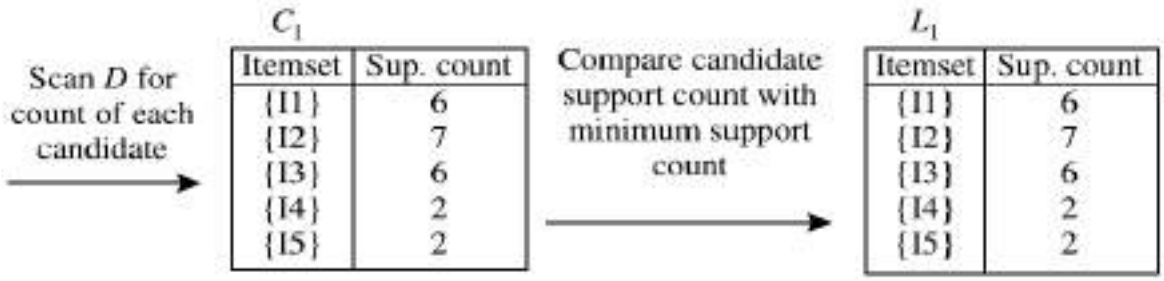
There are nine transactions in this database,

that is, ***| D | = 9.***

We use this to illustrate the Apriori algorithm for finding frequent itemsets in ***D***.

1. In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemsets, *C1.*

The algorithm simply scans all of the transactions to count the number of occurrences of each item.



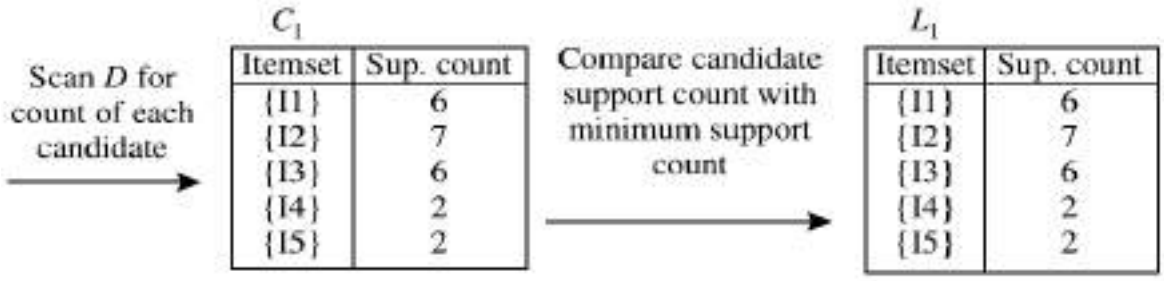
2. Suppose that the minimum support count required is 2,

that is, ***min_sup = 2.***

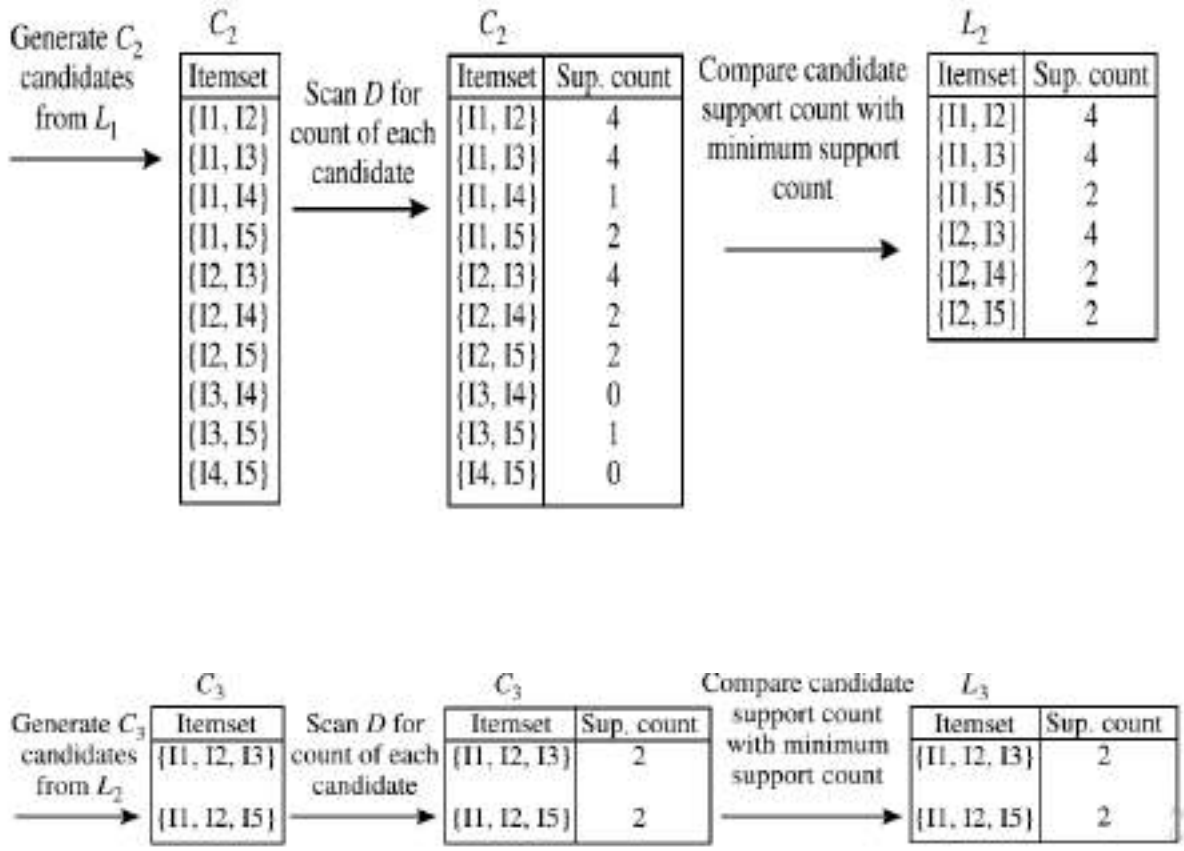
(Here, we are referring to *absolute support because we are using a support count. The corresponding*

relative support is 2/9 = 22%.)

The set of frequent 1-itemsets, *L1, can then* be determined. It consists of the candidate 1-itemsets satisfying minimum support. In our example, all of the candidates in *C1 satisfy minimum support.*



3. To discover the set of frequent 2-itemsets, L_2, the algorithm uses the join $L_1 \bowtie L_1$ to generate a candidate set of 2-itemsets, C_2.[7] C_2 consists of $\binom{|L_1|}{2}$ 2-itemsets. Note that no candidates are removed from C_2 during the prune step because each subset of the candidates is also frequent.



4. Next, the transactions in D are scanned and the support count of each candidate itemset in C_2 is accumulated, as shown in the middle table of the second row in Figure 6.2.

5. The set of frequent 2-itemsets, L_2, is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.

6. The generation of the set of the candidate 3-itemsets, C_3, is detailed in Figure 6.3. From the join step, we first get $C_3 = L_2 \bowtie L_2 = \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}$. Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that the four latter candidates cannot possibly be frequent. We therefore remove them from C_3, thereby saving the effort of unnecessarily obtaining their counts during the subsequent scan of D to determine L_3. Note that when given a candidate k-itemset, we only need to check if its $(k-1)$-subsets are frequent since the Apriori algorithm uses a level-wise

(a) Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$\bowtie \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

(b) Prune using the Apriori property: All nonempty subsets of a frequent itemset must also be frequent. Do any of the candidates have a subset that is not frequent?

- The 2-item subsets of {I1, I2, I3} are {I1, I2}, {I1, I3}, and {I2, I3}. All 2-item subsets of {I1, I2, I3} are members of L_2. Therefore, keep {I1, I2, I3} in C_3.
- The 2-item subsets of {I1, I2, I5} are {I1, I2}, {I1, I5}, and {I2, I5}. All 2-item subsets of {I1, I2, I5} are members of L_2. Therefore, keep {I1, I2, I5} in C_3.
- The 2-item subsets of {I1, I3, I5} are {I1, I3}, {I1, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I1, I3, I5} from C_3.
- The 2-item subsets of {I2, I3, I4} are {I2, I3}, {I2, I4}, and {I3, I4}. {I3, I4} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I4} from C_3.
- The 2-item subsets of {I2, I3, I5} are {I2, I3}, {I2, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I5} from C_3.
- The 2-item subsets of {I2, I4, I5} are {I2, I4}, {I2, I5}, and {I4, I5}. {I4, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I4, I5} from C_3.

(c) Therefore, $C_3 = \{\{I1, I2, I3\}, \{I1, I2, I5\}\}$ after pruning.

End of session 7