

**Final Exam Study Notes**

[Not guaranteed to be all-inclusive.]

This is a summary of the main topics covered throughout the course. Please see the lecture notes, the assignment descriptions, and assigned readings from the text.

See also the summary lecture notes (course wrap up) that were posted.

There will be a variety of questions on the final exam. Some will be conceptual in nature, so be sure you understand the main concepts of the course. There will be SQL questions. Besides writing the queries, you will be asked to interpret the questions in terms of real-world applications. You should also be able to read and interpret conceptual models. See examples done in class and from the text.

**Topics**

* Conceptual modeling representations
  + What is it and why is it important?
  + Representation: Chen’s, Crow’s Feet. Be able to read/interpret either
* Relational model and translation of conceptual models into logical models
  + Know the transformation rules (what they are, why they are needed, how to apply them)
* Data mining: Why is it important? What results might you get? What are its challenges?
  + Understand the inclass examples (e.g., diaper, coffee store).
* SQL queries: DDL vs DML (When is each used?)
* Disruptive technologies. Sustaining vs disruptive and role of data.
* Big data: why does big data exist? Why is its management important? How is such data acquired? Example of customer segmentation.
* Privacy – Ways to manage privacy. Challenges of data breaches.
* Implications of proper data management (from both technical and managerial points of view).

1. Conceptual Modeling Representations:

a. Conceptual modeling represents the high-level abstraction of a database system, capturing essential aspects of the data structure, relationships, and constraints without delving into implementation details. It serves as a blueprint for designing the database schema, ensuring alignment with the organization's requirements and facilitating communication between stakeholders.

b. Representation: Chen's and Crow's Feet notations are graphical techniques used for conceptual modeling in entity-relationship diagrams (ERDs). Chen's notation emphasizes the distinction between entities, attributes, and relationships, while Crow's Feet notation emphasizes cardinality and participation constraints. Mastery of these representations enables effective communication and interpretation of database designs among developers, designers, and users.

1. Relational Model and Translation of Conceptual Models into Logical Models: a. Transformation rules are guidelines used to translate conceptual models, typically represented in ERDs, into logical models, specifically relational schemas. These rules facilitate the conversion process by defining how entities, attributes, and relationships in the conceptual model map to tables, columns, and constraints in the relational model. They ensure data integrity and normalization while preserving the semantics of the original conceptual design.
2. Data Mining:

a. Data mining is crucial for extracting valuable insights and patterns from large datasets, enabling informed decision-making and predictive analytics. Results may include identifying trends, associations, clusters, or anomalies within the data, which can inform business strategies, optimize processes, and detect fraudulent activities.

b. Challenges in data mining include processing vast amounts of data efficiently, ensuring the quality and reliability of results, addressing privacy concerns, and interpreting complex patterns accurately. In-class examples such as the diaper and coffee store illustrate how data mining techniques can reveal unexpected correlations and actionable insights from seemingly unrelated data sources.

1. SQL Queries:

DDL vs DML:

a. Data Definition Language (DDL) is used to define and modify the structure of database objects such as tables, indexes, and constraints. It includes commands like CREATE, ALTER, and DROP.

b. Data Manipulation Language (DML) is used to retrieve, insert, update, and delete data within the database. It includes commands like SELECT, INSERT, UPDATE, and DELETE. c. DDL is typically used during database schema design and administration, while DML is used for day-to-day data manipulation operations.

1. Disruptive Technologies:

a. Disruptive technologies are innovations that significantly alter existing markets, industries, or processes by introducing novel approaches, products, or business models. They often challenge established players and paradigms, leading to industry transformation.

b. Sustaining technologies, on the other hand, are incremental improvements to existing products or processes that sustain the competitiveness of established players within an industry.

c. Data plays a crucial role in both sustaining and disruptive technologies, serving as the foundation for insights, innovation, and competitive advantage.

1. Big Data:

a. Big data refers to datasets that are too large or complex to be processed using traditional database management tools and techniques. It exists due to the proliferation of digital devices, sensors, and online interactions, generating vast amounts of structured and unstructured data.

b. Managing big data is important for organizations to extract value, gain competitive insights, and enhance decision-making. Techniques such as distributed computing, parallel processing, and advanced analytics are employed to store, process, and analyze big data effectively.

c. Big data acquisition involves collecting data from diverse sources, including social media, IoT devices, transaction records, and sensor networks. Customer segmentation is an example where big data analytics can identify distinct customer groups based on behavior, preferences, and demographics.

1. Privacy:

a. Managing privacy involves implementing policies, technologies, and practices to protect individuals' sensitive information from unauthorized access, use, or disclosure.

b. Challenges of data breaches include financial losses, reputational damage, legal consequences, and loss of customer trust. Organizations must invest in robust security measures, encryption techniques, and access controls to mitigate these risks.

1. Implications of Proper Data Management:

a. From a technical standpoint, proper data management ensures data accuracy, consistency, and integrity, supporting efficient storage, retrieval, and processing operations. It also facilitates scalability, interoperability, and data governance.

b. From a managerial perspective, effective data management enables evidence-based decision-making, fosters innovation, and enhances organizational agility. It also enables compliance with regulatory requirements, risk management, and strategic planning.

**ER-Relational Model**

* Translation of ER Model to Relational Model
  + Every entity becomes a separate relation
  + For relationships there are two options:
    - Foreign key: for 1:N relationships (with some variations for optional relationships)
    - Separate relation: with the key of the relation the concatenation, or joining together, of the two keys of the corresponding entities. Relationship attributes become non-key attributes.
  + Reverse engineering. Logical model to conceptual model. Given a logical model, can you answer questions about the corresponding conceptual model from which it came? This is important to identify whether existing relational models are missing concepts.
  + Conceptual modeling is important because it dictates the structure of the final relational model.

**Translation of ER Model to Relational Model:**

1. Every entity in the Entity-Relationship (ER) model becomes a separate relation (table) in the relational model. Each attribute of the entity corresponds to a column in the relation.
2. For relationships, there are two options:

a. For 1: N relationships (one-to-many), a foreign key is added to the "many" sides of the relationship, referencing the primary key of the "one" side.

b. For M: N relationships (many-to-many), a separate relation (junction table) is created, with the primary key being the concatenation or joining together of the keys of the corresponding entities. Relationship attributes become non-key attributes in this separate relation.

1. Reverse Engineering from Logical Model to Conceptual Model involves deducing the conceptual model from an existing logical model. This process is crucial for identifying whether existing relational models adequately capture all necessary concepts. By examining the logical model, one can infer the entities, relationships, and constraints that constitute the conceptual model.
2. Conceptual modeling is important because it serves as the foundation for designing the relational model. The conceptual model captures the essential aspects of the data structure, relationships, and constraints without delving into implementation details. It dictates the structure of the final relational model, ensuring alignment with the organization's requirements and facilitating effective database design and development.

**SQL**

* Understand what a query language is
  + Why are query languages important for data management?
  + Helps us to retrieve useful data from the database.
  + SQL – Structured Query Language, nonprocedural language, tells what to retrieve, not how to do so. Used for data administration, data manipulation and to query a database.
* SQL – DDL (data definition language) and DML (data manipulation language)
  + Understand the Create Statement and its usefulness (create the tables)
  + Appreciate how to populate a database. Ensure referential integrity in the data.
* SQL – DML for Basic SQL queries
  + Be able to write an SQL query that involves multiple tables and multiple joins (e.g., chef example)
  + Be able to interpret an SQL query. That is, given an SQL query, be able to provide a corresponding business interpretation of it.
  + Basic form of SQL DML command: Select – From – Where
  + Know how to insert data into a set of tables
  + Understand the requirements for specifying a data type for each attribute
  + See examples in lecture notes and text
* SQL queries on single versus multiple tables
  + See lecture notes and the examples posted and reviewed in class.
  + Understand the concept of “join” on common attributes when queries involve more than one table.
* Be able to answer short answer questions similar to those reviewed in class. This includes providing real-world interpretations of results and the implications for managerial decision making.

SQL:

1. Understanding Query Languages:

a. A query language is a specialized language used to interact with databases, enabling users to retrieve, manipulate, and manage data stored in a database system.

b. Query languages are essential for data management as they provide a standardized method for accessing and retrieving information from databases. They facilitate data administration, manipulation, and querying tasks, allowing users to retrieve useful data without needing to understand the underlying database structure or implementation details.

1. SQL - DDL (Data Definition Language) and DML (Data Manipulation Language):

a. The CREATE statement in SQL is used within the Data Definition Language (DDL) to define and create database objects such as tables, views, indexes, and constraints. It is useful for establishing the structure of the database, including defining the attributes, data types, and relationships of tables.

b. Populating a database involves inserting data into the tables. Referential integrity in the data ensures that relationships between tables are maintained, preventing inconsistencies or errors. This is typically achieved using foreign key constraints, which enforce relationships between tables.

1. SQL - DML for Basic SQL Queries:

a. Writing SQL queries that involve multiple tables and multiple joins is essential for retrieving data from complex database structures. For example, in a chef example, one might join tables representing chefs, recipes, and ingredients to retrieve information about which chefs prepare which recipes using which ingredients.

b. Interpreting an SQL query involves understanding the business context and objectives behind the query. Given an SQL query, one should be able to provide a corresponding business interpretation, explaining what information is being retrieved and why it is relevant. c. The basic form of SQL DML commands typically involves the SELECT-FROM-WHERE structure, where the SELECT clause specifies the columns to retrieve, the FROM clause specifies the tables to query, and the WHERE clause specifies any conditions or filters to apply. d. Inserting data into tables involves specifying values for each attribute, ensuring they adhere to the specified data types and constraints.

1. SQL Queries on Single Versus Multiple Tables:

a. Understanding the concept of "join" is crucial when querying multiple tables. Joins allow data to be retrieved from multiple tables based on common attributes, enabling the combination of related information from different tables into a single result set.

1. Short Answer Questions:

a. Being able to answer short answer questions similar to those reviewed in class involves providing real-world interpretations of SQL query results and discussing the implications for managerial decision-making. This requires understanding the business context, the data being queried, and the potential insights or actions that can be derived from the results.

**Data Warehouses and Data Mining**

* Know the difference between query processing, OLTP and OLAP
* Understand how / why data is represented or considered as an OLAP cube to handle the multi-dimensional aspects of it. Think of each piece of data as represented as one piece of a (multi-valued) cube.
* Appreciate that data is input from multiple sources into a data warehouse.
* Data mining applications in: customer segmentation, marketing and promotion targeting, market basket analysis, collaborative ﬁltering, customer churn, fraud detection, financial modeling, and hiring and promotion. Recall also the separate example on market segmentation. This is the matrix of customers, who are categorized based on their status as customers. The managerial implications are that this provides insights into whether you should put resources for obtaining or retaining customer.

Data Warehouses and Data Mining:

1. Difference between Query Processing, OLTP, and OLAP:

a. Query processing refers to the execution of queries against a database system to retrieve or manipulate data. It involves parsing, optimizing, and executing queries to produce results. b. OLTP (Online Transaction Processing) systems are designed for real-time transactional processing, handling a large volume of short, frequent transactions. They prioritize data integrity, concurrency, and quick response times for day-to-day operations.

c. OLAP (Online Analytical Processing) systems are optimized for complex analytical queries involving aggregated data from multiple dimensions. They facilitate decision support and business intelligence by providing multidimensional analysis capabilities, such as drill-down, roll-up, and slicing-and-dicing.

1. Representation of Data as OLAP Cubes:

a. Data is represented as an OLAP cube to handle its multi-dimensional aspects effectively. Each piece of data is conceptualized as one cell within a multi-valued cube, where each axis represents a different dimension (e.g., time, geography, product category). OLAP cubes enable analysts to explore data from various perspectives and dimensions, facilitating advanced analysis and reporting.

1. Input of Data from Multiple Sources into Data Warehouse:

a. Data warehouses serve as central repositories for storing and integrating data from various sources, including transactional databases, external sources, and legacy systems. This integrated data provides a comprehensive view of the organization's operations, enabling in-depth analysis, reporting, and decision-making.

1. Data Mining Applications:

a. Customer segmentation involves categorizing customers into distinct groups based on their characteristics, behaviors, or purchasing patterns. This enables targeted marketing strategies, personalized promotions, and tailored services to meet the diverse needs of different customer segments.

b. Marketing and promotion targeting utilizes data mining techniques to identify potential customers who are most likely to respond positively to specific marketing campaigns or promotions. This improves marketing ROI and enhances customer engagement.

c. Market basket analysis examines the relationships between products frequently purchased together by customers. It helps retailers optimize product placement, cross-selling, and promotional strategies to increase sales and customer satisfaction.

d. Collaborative filtering recommends items or products to users based on their preferences and behaviors, leveraging similarities with other users' preferences. It enhances personalization and improves user experience in recommendation systems.

e. Customer churn prediction identifies customers who are likely to stop using a service or product. It enables proactive retention efforts, customer loyalty programs, and targeted interventions to reduce churn rates and increase customer lifetime value.

f. Fraud detection utilizes data mining algorithms to detect anomalous patterns or behaviors indicative of fraudulent activities. It helps organizations mitigate risks, prevent financial losses, and safeguard their assets.

g. Financial modeling involves using data mining techniques to analyze financial data, predict market trends, and optimize investment strategies. It supports informed decision-making and risk management in finance and investment industries.

h. Hiring and promotion applications of data mining involve analyzing employee performance, skills, and attributes to identify top performers, succession planning, and talent management. It aids in making objective, data-driven decisions in human resources management.

1. Managerial Implications:

a. Market segmentation provides insights into resource allocation for customer acquisition and retention efforts. It helps organizations prioritize investments and tailor strategies to meet the needs of different customer segments effectively.

**Data and Databases**

* Data is an important asset in any organization. Understand the difference and uses of public versus private data.
* Why must you ensure data consistency? Data integrity? Data updates?
* Understand how data supports decision making; must be correct and database must be properly designed.

Data and Databases:

1. Difference and Uses of Public Versus Private Data:

a. Public data refers to information that is freely available and accessible to the general public. This data can include government records, open-source datasets, and publicly available research findings. Public data is often used for research, analysis, and transparency purposes.

b. Private data, on the other hand, refers to sensitive information that is restricted and protected from unauthorized access. This includes personal, financial, and proprietary data owned by individuals, organizations, or businesses. Private data is typically governed by privacy regulations and used for internal operations, decision-making, and competitive advantage.

1. Importance of Ensuring Data Consistency, Integrity, and Updates:

a. Data Consistency: Ensuring data consistency involves maintaining uniformity and accuracy of data across different parts of the database. Consistent data prevents discrepancies and ensures that all users access the same information, enhancing reliability and trustworthiness.

b. Data Integrity: Data integrity refers to the accuracy, completeness, and reliability of data within the database. It ensures that data is valid, consistent, and conforms to defined rules or constraints. Maintaining data integrity is crucial for making informed decisions and preventing errors or inconsistencies.

c. Data Updates: Regular updates to data are necessary to reflect changes in the real world and ensure the relevance and accuracy of information. Updates may include adding new records, modifying existing data, or deleting obsolete information. Timely updates support decision-making processes by providing up-to-date and relevant data for analysis and planning.

1. How Data Supports Decision Making:

a. Data plays a critical role in supporting decision-making processes across various domains and industries. Correct and reliable data, combined with a properly designed database, provides a foundation for informed and effective decision-making.

b. Correct data ensures that decisions are based on accurate information, reducing the risk of errors or misinterpretations. Properly designed database’s structure and organize data in a way that facilitates retrieval, analysis, and interpretation, enabling users to derive meaningful insights and make informed choices.

c. Decision makers rely on data to assess performance, identify trends, evaluate risks, and formulate strategies. Whether it's in business, healthcare, finance, or government, data-driven decision making is essential for achieving organizational goals, improving efficiency, and staying competitive in today's dynamic environment.

**Disruptive technologies**

* Still require data. Many other related topics including cybersecurity (another course).
* Know characteristics of a true disruptive technology; see examples.
* Just because a technology is based on large amounts of data, does not mean it is disruptive and vice versa.
* Examples of truly disruptive technologies (e.g., impact an industry or industries)

Disruptive Technologies:

1. Requirement of Data:

a. Disruptive technologies, despite their transformative nature, still heavily rely on data. Data serves as the fuel that powers these innovations, enabling them to deliver novel solutions, insights, and functionalities. Whether it's collecting, analyzing, or leveraging data, disruptive technologies often depend on robust data ecosystems to fulfill their potential.

1. Characteristics of a True Disruptive Technology:

a. Disruptive technologies exhibit several key characteristics:

* Accessibility: They democratize access to services or products, making them available to a broader audience or previously underserved markets.
* Affordability: They offer cost-effective alternatives to existing solutions, challenging established pricing models and market dynamics.
* Scalability: They have the potential to scale rapidly and disrupt traditional industry structures, often leveraging network effects or cloud-based platforms.
* Innovation: They introduce novel approaches, functionalities, or business models that redefine industry standards and practices.
* Market Impact: They fundamentally change the competitive landscape, challenging incumbents and reshaping industry dynamics.

b. Examples of disruptive technologies include:

* Uber and Lyft: Disrupted the transportation industry by introducing on-demand ride-sharing services that leverage mobile apps and crowdsourced drivers.
* Airbnb: Disrupted the hospitality industry by enabling individuals to rent out their homes or spare rooms to travelers, challenging traditional hotel chains.
* Netflix: Disrupted the entertainment industry by offering streaming services that bypass traditional cable TV subscriptions, leading to the decline of DVD rentals and traditional TV viewing habits.
* Amazon Web Services (AWS): Disrupted the IT industry by providing cloud computing services, allowing businesses to scale infrastructure on-demand without significant upfront investments in hardware.
* Tesla: Disrupting the automotive industry with electric vehicles and autonomous driving technology, challenging traditional car manufacturers and reshaping the future of transportation.

1. Relationship Between Disruptive Technologies and Data:

a. While disruptive technologies often rely on data for innovation and functionality, the mere presence of large amounts of data does not guarantee disruptive potential. Similarly, a technology can be disruptive even if it doesn't rely solely on large datasets.

b. Disruptive technologies leverage data strategically to differentiate themselves, enhance user experiences, and create new market opportunities. However, the disruptive impact stems from their ability to transform industries, business models, and consumer behaviors, rather than solely from their data-centric nature.

**Privacy**

* What is it? The right to be left alone.
* Courts: personal privacy balanced with society’s right to know (legal)
* Why is digital privacy an issue? Data can be easily shared. Inferences can be made from data collected from different places.
* Issues: anonymity, control, sharing with 3rd parties, personal identifying information
* Policy protection: personal, technical, legal or policy
* Laws (many), vary by country and challenging for international organizations
* Many privacy and security concerns
  + Need to protect against them
  + Security is evolving problem
  + Do notions of privacy change over time?
* Why is ethics of great concern in our digital world? Ease of sharing of data. Personal information may be explicitly or implicitly shared.

Privacy:

1. Definition:

a. Privacy refers to the fundamental right of individuals to control their personal information and to be left alone without interference or intrusion. It encompasses the ability to maintain autonomy, confidentiality, and secrecy in one's personal affairs.

1. Legal Considerations:

a. Courts often weigh personal privacy rights against society's right to access information, particularly in legal contexts. Balancing individual privacy with public interest or societal needs can be challenging, leading to ongoing debates and legal interpretations.

1. Digital Privacy Concerns:

a. Digital privacy is a significant issue due to the ease of sharing data and the potential for inferences to be drawn from disparate sources of information. With the proliferation of digital technologies and online platforms, personal data can be collected, aggregated, and analyzed without individuals' awareness or consent.

1. Key Issues:

a. Anonymity: Preserving individuals' anonymity to prevent the identification or tracking of their online activities.

b. Control: Ensuring individuals have control over how their personal information is collected, used, and shared by third parties.

c. Sharing with Third Parties: Addressing concerns about unauthorized sharing or disclosure of personal data to external entities without individuals' consent.

d. Personal Identifying Information (PII): Safeguarding sensitive information such as social security numbers, addresses, and financial details to prevent identity theft or fraud.

1. Policy Protection:

a. Privacy protection strategies encompass personal, technical, legal, and policy measures aimed at safeguarding individuals' privacy rights. These may include data encryption, access controls, privacy policies, and regulatory compliance frameworks.

1. Legal Frameworks:

a. Numerous privacy laws exist globally, with variations across countries and regions. Compliance with these laws presents challenges for international organizations operating in multiple jurisdictions.

1. Privacy and Security Concerns:

a. Privacy and security concerns abound in the digital age, necessitating robust measures to protect against data breaches, unauthorized access, and malicious activities.

b. Security remains an evolving problem, requiring continuous adaptation to emerging threats and vulnerabilities.

1. Evolution of Privacy Notions:

a. Notions of privacy may evolve over time in response to technological advancements, societal norms, and cultural shifts. The ease of sharing personal information and the blurring of boundaries between public and private spheres in the digital world may influence changing perceptions of privacy.

1. Ethics in the Digital World:

a. Ethics is of great concern in the digital world due to the ease of sharing data and the potential for personal information to be shared explicitly or implicitly. Ethical considerations encompass issues of transparency, consent, fairness, and accountability in data collection, use, and dissemination practices.

**Big data**

Large amounts of data being collected, stored, and used. Why do we have so much data and hence the need to store, manage, and use it? Advances in database and other technologies; automated collection of data from many sources; trend to data-driven, real-time decision making.

* Reliable telecommunications enable us to share data easily.
* Value from big data includes identifying patterns from which predictions and decisions can be made. Also, to identify operation concerns or anomalies (e.g., Walmart cookie example).
* Effective use of data for data mining can require large databases, statistics/mathematics, professional, e.g., data analysts working with domain-specific experts (finance, marketing, human resources).

Big Data:

1. Reasons for the Proliferation of Data:

a. Advances in Database and Other Technologies: Innovations in storage, processing, and analytical technologies have enabled organizations to collect, store, and analyze vast amounts of data more efficiently and cost-effectively.

b. Automated Data Collection: The widespread adoption of sensors, IoT devices, social media platforms, and other digital technologies has led to the automated collection of data from diverse sources, generating massive datasets.

c. Trend Towards Data-Driven Decision Making: Organizations increasingly rely on data-driven insights to inform strategic decisions, optimize processes, and gain competitive advantages in dynamic markets.

1. Role of Reliable Telecommunications:

a. Reliable telecommunications infrastructure facilitates the seamless sharing and transfer of data across distributed networks, enabling real-time collaboration, communication, and data exchange between individuals, organizations, and devices.

1. Value Derived from Big Data:

a. Identifying Patterns for Predictive Analytics: Big data analytics enables organizations to identify patterns, trends, and correlations within large datasets, which can be leveraged for predictive modeling and forecasting. By analyzing historical data, organizations can make informed predictions and decisions about future events or outcomes.

b. Detection of Operational Concerns or Anomalies: Big data analytics can uncover operational inefficiencies, anomalies, or irregularities within systems or processes. For example, the Walmart cookie example demonstrates how analyzing shopping patterns and behaviors can reveal operational concerns or anomalies that require attention.

1. Effective Use of Data for Data Mining:

a. Data mining involves the extraction of valuable insights and knowledge from large datasets through statistical and mathematical techniques. This process often requires expertise in database management, statistics, mathematics, and domain-specific knowledge.

b. Data analysts collaborate with domain-specific experts in fields such as finance, marketing, and human resources to effectively analyze and interpret data within their respective domains. By combining domain expertise with analytical skills, organizations can derive actionable insights and make data-driven decisions that drive business success.

Good luck on the exam!