[According to the lecture on 22/03/2025, 3:00pm – 6:00pm]

UNIVERSITI MALAYA

FAKULTI SAINS KOMPUTER DAN TEKNOLOGI MAKLUMAT
Faculty of Computer Science and Information Technology

WQD7005 BIG DATA MANAGEMENT

OCCURRENCE 3

SEMESTER 2 SESSION 2024/2025

TUTORIAL 1

NAME:

BARKAVI A/P P CHEVEN (23093149)

LECTURER'S NAME:

TS. DR. MOHD SHAHRUL NIZAM BIN MOHD DANURI [msnizam@um.edu.my]

[According to the lecture on 22/03/2025, 3:00pm – 6:00pm]

Question 1 Differentiate between Big Data and Small Data, and support with appropriate examples.

Big Data refers to extremely large and complex datasets that cannot be managed, processed, or analyzed using traditional data processing tools. It encompasses massive volumes of information that require specialized systems and advanced analytical approaches. For example, Amazon's e-commerce platform generates Big Data through millions of daily transactions, customer browsing patterns, product reviews, and inventory movements across global warehouses. This data requires sophisticated processing systems and can reach petabytes in size. Similarly, social media platforms like Facebook handle enormous amounts of Big Data, processing billions of user interactions, photos, videos, and messages daily, requiring advanced infrastructure and analytical tools to derive meaningful insights. In contrast, Small Data refers to datasets that are manageable in size and can be easily processed using conventional tools and methods. These datasets are typically structured, easily comprehensible, and can be analyzed using standard software like spreadsheets or basic database systems. For instance, a local retail store's monthly sales records, containing information about daily transactions, inventory levels, and customer purchases, represent Small Data. Another example would be a school's student management system that tracks grades, attendance, and academic performance - data that can be easily managed using standard database software.

The key distinction lies in their processing requirements and complexity. Big Data requires specialized tools, significant computational power, and often distributed processing systems to handle its volume, variety, and velocity. For example, weather forecasting systems processing satellite imagery, sensor data, and historical patterns represent Big Data processing. Meanwhile, Small Data can be processed on standard computers using conventional software tools, like a small medical clinic managing patient records and appointment schedules. Furthermore, Big Data typically requires sophisticated storage solutions and specialized expertise to manage and analyze. Companies like Google handling search queries and YouTube managing video content deal with Big Data that needs cloud storage, distributed computing, and expert data scientists. In contrast, Small Data, such as a library's book inventory or a restaurant's menu pricing system, can be managed with basic storage solutions and analyzed by staff with fundamental analytical skills. The insights derived from Big Data often influence strategic, long-term business decisions, while Small Data typically supports day-to-day operational decisions.

Question 2 Explains the important characteristics of Big Data, including Volume, Velocity, Variety, Veracity and Variability.

The important characteristics of Big Data are commonly referred to as the "Five Vs" - Volume, Velocity, Variety, Veracity, and Variability. Each of these characteristics presents unique challenges and opportunities in the realm of data management and analysis. Volume refers to the sheer scale of data being generated and collected in today's digital world. This characteristic is perhaps the most obvious aspect of Big Data, as it deals with the massive quantities of information that exceed traditional database capabilities. By 2020, approximately 40 Zettabytes of data were created, representing a staggering 300-fold increase from 2015 levels. This exponential growth in data volume necessitates advanced storage solutions and processing techniques that can handle such immense scales efficiently. Organizations must develop infrastructure capable of storing, accessing, and analyzing these vast data repositories to extract meaningful insights.

Velocity addresses the unprecedented speed at which data is generated, collected, and processed in real-time. Modern data streams from sources like financial markets, IoT sensors, and network monitoring systems produce continuous flows of information that require immediate analysis. The content of these data streams is constantly changing through the absorption of complementary collections, previously archived information, and incoming data from multiple sources. These characteristic demands processing systems that can handle high-throughput data ingestion and provide real-time or near-real-time analysis capabilities. The ability to quickly process fast-moving data enables organizations to respond promptly to emerging trends or critical situations.

Variety encompasses the diverse forms and types of data that organizations must manage. Unlike traditional structured data that fits neatly into relational databases, Big Data includes unstructured and semi-structured information from numerous sources. This includes text documents, images, videos, social media posts, sensor readings, and complex records that don't conform to conventional database schemas. The heterogeneous nature of these data types presents significant challenges for integration, processing, and analysis. Organizations must develop flexible data architectures and analytics approaches that can accommodate and derive value from these diverse data formats.

Veracity concerns the reliability and accuracy of data, addressing the fundamental question: "How can we ensure the data is accurate?" Uncertainty exists in Big Data due to inconsistencies, incompleteness, ambiguities, and deception in the datasets. For example, marketing automation systems often contain false names and inaccurate contact information, compromising the quality of analysis. Studies indicate that one in three business leaders doesn't trust the information they use for decision-making. This characteristic highlights the importance of data quality management, validation processes, and statistical methods to handle uncertainty. Organizations must implement robust data governance frameworks to maintain data integrity and build confidence in their analytical outputs.

Variability refers to the changing meaning and interpretation of data over time or across different contexts. This is particularly evident in natural language processing, where the same words can have different meanings depending on context. Data variability creates challenges for consistent interpretation and analysis, as the significance of specific data points may shift based on temporal, cultural, or situational factors. Organizations must develop adaptive analytical approaches that can account for these semantic variations and contextual changes. This often involves sophisticated machine learning algorithms that can recognize patterns and adjust interpretations based on evolving contexts. Together, these five characteristics define the complex nature of Big Data and highlight the specialized approaches needed to effectively harness its potential for organizational insight and value creation.

[According to the lecture on 22/03/2025, 3:00pm – 6:00pm]

Question 3 Identify a potential big data problem and explain how the six phases of Big Data can be implemented to the aforementioned problem.

A significant big data problem facing **modern healthcare systems** is the effective management and utilization of patient data across multiple facilities to improve treatment outcomes and reduce costs. Healthcare organizations generate enormous volumes of diverse data, including electronic health records, medical imaging, laboratory results, wearable device readings, insurance claims, and pharmaceutical information. This fragmented data landscape presents challenges in delivering personalized care, identifying population health trends, and optimizing resource allocation.

The first phase in addressing this problem is Discovery, where healthcare organizations identify and catalog all relevant data sources. This involves mapping internal systems like electronic health records and laboratory information systems, as well as external sources such as insurance databases, public health registries, and patient-generated health data from wearables and mobile applications. During this phase, stakeholders from clinical, administrative, and IT departments collaborate to define specific objectives, such as reducing hospital readmissions or improving chronic disease management, and determine what data is necessary to achieve these goals.

The Data Preparation phase involves cleaning, standardizing, and integrating the diverse healthcare datasets. This critical step addresses challenges like inconsistent medical terminology, missing values, duplicate records, and varying data formats across different systems. Healthcare data engineers implement standardized medical coding systems (such as ICD-10, SNOMED CT, or LOINC) to ensure consistent representation of medical concepts. They also develop robust data pipelines that can handle the continuous influx of new patient information while maintaining patient privacy and regulatory compliance with frameworks like HIPAA. This phase establishes the foundation for reliable analysis by ensuring data quality and accessibility. In the Model Planning phase, data scientists and clinical experts collaborate to design analytical approaches tailored to specific healthcare objectives. They select appropriate statistical methods, machine learning algorithms, and visualization techniques based on the nature of the healthcare questions being addressed. For example, predictive models might be designed to identify patients at high risk for hospital readmission, while clustering algorithms could reveal patterns in disease progression. This phase includes defining evaluation metrics that align with clinical outcomes, such as reduction in adverse events or improvements in patient satisfaction scores and establishing a framework for validating model performance in a healthcare context.

The Model Building phase involves implementing and refining the analytical models using the prepared healthcare data. Data scientists develop predictive algorithms that can identify patients at risk for specific conditions, create classification systems for disease diagnosis, or build recommendation engines for treatment protocols. They employ techniques like deep learning for medical image analysis, natural language processing for extracting insights from clinical notes, and time-series analysis for monitoring patient vitals. Throughout this phase, models undergo rigorous testing and validation using historical healthcare data to ensure they meet performance requirements and can generalize to new patient populations.

The Communication phase focuses on translating complex analytical findings into actionable insights for healthcare stakeholders. Data visualization specialists create intuitive dashboards that present key metrics and trends to clinicians, administrators, and patients in accessible formats. Clinical decision support systems are developed to integrate predictive insights directly into workflow tools, providing recommendations at the point of care. Regular reports and presentations communicate system-wide trends to leadership teams, while patient portals deliver personalized health information directly to individuals. This phase ensures that big data insights are effectively communicated to drive informed decision-making across all levels of the healthcare organization. Finally, the Operationalization phase involves integrating the big data solutions into daily healthcare operations. IT teams deploy the analytical models into production environments, establishing automated data pipelines that continuously update predictions as new patient information becomes available. Clinical workflows are redesigned to incorporate data-driven insights at key decision points, such as discharge planning or medication management. Monitoring systems track both technical performance metrics and clinical outcomes to measure the real-world impact of the big data implementation. Continuous improvement processes are established to refine models based on feedback from healthcare providers and evolving medical knowledge, ensuring the system remains effective and relevant over time.

[According to the lecture on 22/03/2025, 3:00pm – 6:00pm]

Question 4 Identify one existing company that employs Big Data as one of their technologies and discuss the potential obstacles they might be facing while implementing Big Data in their system.

Netflix stands as a prime example of a company that has masterfully integrated Big Data into the core of its business model, transforming from a DVD rental service to a global streaming powerhouse through data-driven decision making. The company leverages massive datasets from its 200+ million subscribers worldwide, collecting and analyzing viewing habits, search queries, browsing patterns, device information, and even the specific moments when users pause, rewind, or abandon content. This wealth of information powers Netflix's sophisticated recommendation engine, which drives approximately 80% of content discovery on the platform and saves the company an estimated $1 billion annually through reduced customer churn. Additionally, Netflix utilizes Big Data analytics to inform content creation decisions, optimizing everything from script selection to casting choices for its original productions.

Despite its success, Netflix faces several significant obstacles in implementing and maintaining its Big Data systems. Data privacy and regulatory compliance represent major challenges as the company operates across different countries with varying data protection laws such as GDPR in Europe, CCPA in California, and other regional regulations. These frameworks impose strict requirements on how user data can be collected, processed, and stored, potentially limiting Netflix's analytical capabilities and necessitating complex, region-specific data governance policies. The company must continuously balance its data utilization with evolving global privacy standards to avoid regulatory penalties and maintain consumer trust.

Infrastructure scalability presents another substantial challenge for Netflix as its subscriber base and content library continue to expand globally. The company processes over 450 billion events daily across its data pipeline, requiring massive computational resources and sophisticated architecture. Netflix has addressed this through its migration to AWS cloud services, but still faces the ongoing challenge of optimizing costs while maintaining performance as data volumes grow exponentially. The company must continuously evolve its infrastructure to handle peak streaming periods, new market expansions, and increasingly complex analytical workloads without service degradation or unsustainable cost increases.

Data quality and integration obstacles also plague Netflix's Big Data implementation. The company collects information from diverse sources including different devices, operating systems, and network environments, creating inconsistencies in data formats and reliability. Merging this heterogeneous data into a coherent analytical framework requires sophisticated ETL (Extract, Transform, Load) processes and data validation mechanisms. Furthermore, as Netflix expands into new markets with different languages, cultural contexts, and viewing behaviors, the company struggles to maintain consistent data interpretation across regions, potentially affecting the accuracy of its recommendation algorithms and content decisions in emerging markets.

Finally, Netflix confronts significant talent and organizational challenges in maintaining its data-driven advantage. The company competes fiercely for skilled data scientists, engineers, and analysts in a tight labor market, particularly as specialized Big Data expertise becomes increasingly valuable across industries. Beyond recruitment, Netflix must foster effective collaboration between technical teams and creative content producers, bridging the gap between data insights and artistic decision-making. This cultural integration challenge requires developing a shared language and mutual respect between traditionally separate domains of expertise. Additionally, the company must continuously invest in employee training and development to keep pace with rapidly evolving Big Data technologies and methodologies, ensuring its workforce remains at the cutting edge of data science capabilities.