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 4

NAME:

BARKAVI A/P P CHEVEN (23093149)

LECTURER'S NAME:

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

Question 1: Taking University of Malaya as example, discuss when we should use traditional database to store and access data, and when distributed computing such as Hadoop is more suitable to be used.

As for our University of Malaya, traditional databases and distributed computing frameworks serve distinct operational requirements based on data characteristics and processing demands. Traditional databases are optimal for University of Malaya when managing structured academic records, financial transactions, and administrative data with well-defined schemas. These systems excel in scenarios requiring ACID compliance, such as student registration, course enrollment, and examination records where data integrity is paramount. The relational model supports complex queries necessary for institutional reporting, academic analytics, and regulatory compliance. Traditional databases also provide superior performance for concurrent transaction processing with low latency requirements, essential for student information systems and administrative portals handling moderate data volumes (typically under 10TB).

Conversely, distributed computing frameworks like Hadoop become advantageous when the university confronts big data challenges. Research initiatives generating massive datasets from scientific instruments, genomic sequencing, or climate modeling benefit from Hadoop's scalable storage and processing capabilities. The framework excels in analyzing unstructured data from campus IoT sensors, research publications, and multimedia educational content. Hadoop's batch processing capabilities support complex analytical workloads for institutional research, student performance prediction, and resource optimization. Additionally, cost-effective storage architecture enables long-term preservation of research outputs and historical academic data, while the distributed architecture facilitates seamless horizontal scaling to accommodate exponential data growth from expanding digital initiatives across faculties. The university would benefit from hybrid architecture, deploying traditional databases for operational systems requiring transactional integrity while leveraging Hadoop ecosystems for analytical workloads and data-intensive research applications.

Question 2: Thunder Solution has been operating as a bank solution provider for the past 20 years, and they wish to start a big data initiative based on the data they have collected so far. Among distributed computing, parallel computing, grid computing and cloud computing, suggest one of the computing solution to the data manager of Thunder Solution. Justify your answer.

For Thunder Solution's big data initiative, cloud computing represents the optimal technological paradigm. As a banking solution provider with 20 years of accumulated data, Thunder Solution requires a computing infrastructure that balances innovation capacity with regulatory compliance and operational efficiency. Cloud computing offers compelling advantages for financial technology organizations transitioning to big data analytics. The model provides immediate access to enterprise-grade infrastructure without substantial capital expenditure, allowing Thunder Solution to allocate resources to value-generating analytics rather than hardware procurement. This elasticity enables dynamic resource allocation based on computational demands, efficiently handling periodic workloads like month-end processing and regulatory reporting without maintaining permanently provisioned infrastructure.

The financial services domain demands stringent security and compliance adherence, which modern cloud providers address through comprehensive security frameworks, encryption protocols, and compliance certifications (PCI-DSS, SOC 2, ISO 27001). These capabilities mitigate the operational risk associated with managing sensitive financial data while satisfying regulatory requirements. Cloud platforms deliver integrated big data services (data lakes, warehousing, ETL pipelines) and advanced analytics capabilities (machine learning, AI) as managed offerings, accelerating Thunder Solution's time-to-insight without specialized infrastructure expertise. This ecosystem approach facilitates the progressive modernization of legacy banking applications while enabling new analytical capabilities like fraud detection, customer segmentation, and risk modeling.

Furthermore, cloud computing's global infrastructure footprint supports Thunder Solution's potential geographic expansion while maintaining consistent performance and regulatory compliance across jurisdictions. The pay-as-you-go economic model aligns computing costs with business value, optimizing total cost of ownership compared to on-premises alternatives. While distributed, parallel, and grid computing offer specific technical advantages, cloud computing provides the comprehensive technological foundation and business flexibility essential for Thunder Solution's successful big data transformation in the banking solutions sector.

Question 3: What are the advantages and disadvantages of HBase, compared to MySQL?

HBase provides distinct advantages over MySQL in big data environments through its linear scalability architecture, enabling seamless horizontal expansion across commodity hardware clusters compared to MySQL's vertical scaling constraints. Its schema-less column-oriented design offers exceptional flexibility for diverse record structures, accommodating evolving data requirements without performance degradation unlike MySQL's rigid schema model. HBase implements log-structured merge trees that deliver superior write throughput for high-velocity data ingestion scenarios, while its automatic sharding capability transparently manages data distribution, eliminating complex manual partitioning required in large MySQL deployments.

However, these advantages present notable trade-offs. HBase lacks native SQL support and complex join operations, necessitating additional application logic compared to MySQL's comprehensive query capabilities. Its limited ACID compliance creates challenges for applications requiring strict transactional consistency across multiple operations. The distributed architecture demands substantial computational resources even for moderate workloads, resulting in higher infrastructure costs compared to MySQL's efficient operation in resource-constrained environments. Furthermore, HBase typically requires integration with supplementary processing frameworks to achieve analytical capabilities natively available in MySQL. The selection between these technologies should be guided by specific application requirements, with HBase excelling in scenarios demanding massive scalability and schema flexibility, while MySQL remains advantageous for structured, transaction-oriented workloads with complex query patterns.