Big Data - Week 01

The key topics to learn include:
- Types of Databases
- Data Analytics
- Real-world Databased examples

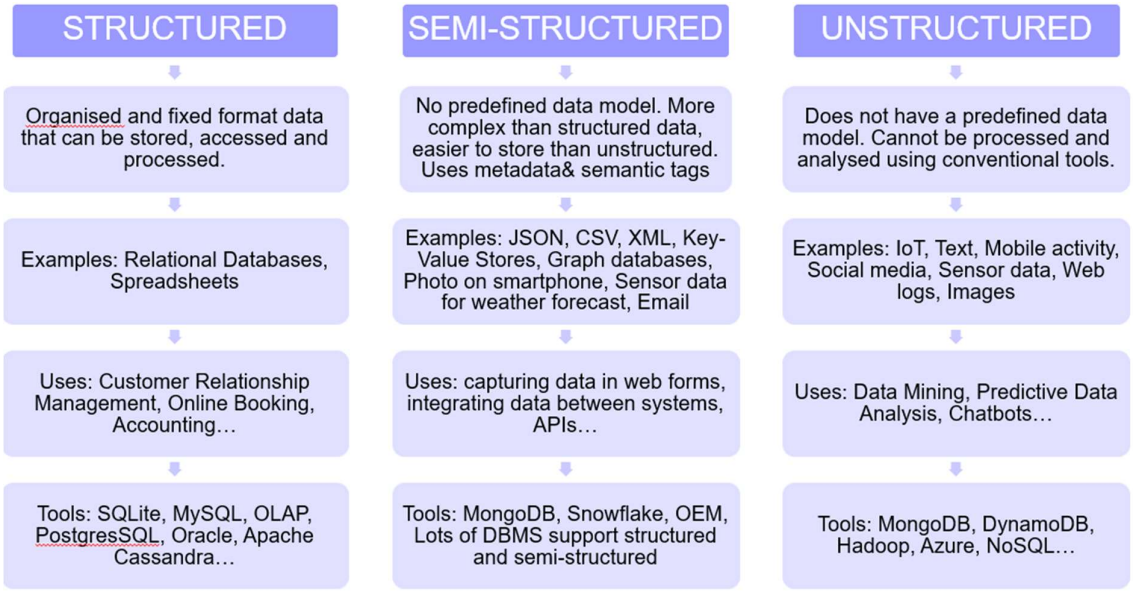
Lab Activity description: Form a small group of 2 - 3 members

BASIC TASKS

a) Read through the list of "Real World Database Examples" and choose several that you want to research.
b) For each scenario that you have chosen the aim is to produce a recommended database solution. This involves deciding what type of database (Relational, Network, Hierarchical, Object-Oriented, NoSQL[Key Value, Column-Based, Graph, Document]) is most suitable for each scenario as well as appropriate database software tools that could be used to implement it.

Consider each of the following when making your decision:

- **Data Model:** Consider what data will be stored in the database; is it structured, semi-structured or unstructured data? The type of data will impact the choice of database. The table from below explains the differences of each type. The data model defines the structure and organisation of your application's data. Different databases offer various data models, such as relational, document-oriented, key-value, graph, or columnar. Consider the nature of your data and how it will be accessed and queried by your application. If your data is highly structured and requires complex querying, a relational database might be suitable. On the other hand, if your data is unstructured or schema-less, a document-oriented or NoSQL database might be a better fit. Choosing the right data model ensures efficient data storage and retrieval, promoting flexibility and ease of development.



- **Scalability**: This is a vital factor when selecting a database, especially if your application is expected to handle a growing volume of data and users over time. Consider whether the database can handle increased workloads, higher data storage requirements, and additional user traffic. If your database needs to handle vast volumes of data either now or in the future then look for databases that offer horizontal scalability, allowing you to add

more servers to distribute the load effectively. A scalable database ensures that your application can grow seamlessly without compromising performance or reliability.

- **Data Consistency**: This is critical in maintaining the integrity of your application's information. When multiple users or processes concurrently access and modify data, it is crucial to ensure that changes are applied consistently. Choose a database that supports strong consistency guarantees, ensuring that all transactions are executed reliably, and that data is accurate and up to date. ACID (Atomicity, Consistency, Isolation, Durability) compliant databases provide robust consistency guarantees and are well-suited for applications where data integrity is paramount.

- **Performance:** Performance is a critical consideration when selecting a database, as it directly impacts the responsiveness and efficiency of your application. Evaluate the database's read and write speeds, as well as its ability to handle complex queries and large datasets. Look for features such as indexing, caching mechanisms, and query optimization capabilities that can enhance performance. Additionally, consider the database's ability to scale horizontally to accommodate increasing loads without sacrificing speed. Performance testing and benchmarking can provide valuable insights into a database's capabilities under realistic conditions.

- **Security:** Data security is of utmost importance, particularly if your application handles sensitive or personal information. Evaluate the security features offered by the database, such as access control mechanisms, encryption, auditing capabilities, and compliance with relevant data protection regulations. Ensure that the database provides robust authentication and authorisation mechanisms to control access to data. Regular security updates and a strong track record of addressing vulnerabilities are additional indicators of a secure database.

- **Cost:** Cost considerations play a significant role in the decision-making process when choosing a database. Evaluate both the upfront and ongoing costs associated with the database. Consider factors such as licensing fees, hardware requirements, maintenance and support costs, and the potential need for specialised database administrators. Additionally, factor in the scalability and growth costs as your application expands. Open-source databases can often provide cost advantages, but ensure they meet your requirements and have a reliable support ecosystem if needed.

- **Community Support**: Community support is crucial when selecting a database for your application. An active and vibrant community indicates a healthy ecosystem of developers, contributors, and users who can provide assistance, share best practices, and offer solutions to common challenges. A robust community ensures access to a wealth of resources, including documentation, forums, tutorials, and plugins or extensions. It also indicates that the database is likely to receive ongoing updates, bug fixes, and new features. When issues arise or when you need guidance, a supportive community can be invaluable in resolving problems quickly and efficiently. Consider the size and engagement of the community surrounding the database you are considering to ensure that you have access to the necessary support and resources.

LIST OF REAL-WORLD DATABASE EXAMPLES: (Example – Category)

Online Video Streaming – Entertainment. Streaming giants like Disney+ and Amazon Prime Video use databases to optimise your viewing experience. Whether it's NoSQL for Disney+ content or MySQL for Netflix billing information, databases catalogue countless hours of content and monitor your viewing patterns in exacting detail — from what you watch and when to where you pause and what you rewind. This data tracking enables them to personalise recommendations and tailor ads specific to you.

Online and Social Gaming – Entertainment. In the area of online and social gaming, there is a web of interconnected databases. Whether a sweeping MMORPG like World of Warcraft or a simple game of Hearts on Facebook, databases are crucial in shaping online gaming. Databases track your scores, your inventory, and your game state. They also track things like your friends' lists, in-game chats and transactions, and your interactions with other players. These databases not only support the game but also enable personalisation of the experience players have while playing.

Social Media – Entertainment. Social media companies track everything about you that they can. Yes, they track your friends and your posts, but they also track your engagement with the content displayed in your feed. Has Shi-ping-hao noticed you're more likely to comment on posts of puppy videos? Your feed will be more likely to display posts with pictures of cute animals. Do you often comment with an angry tone on posts relating to certain topics?

Broadcasting – Entertainment. Broadcasting is distributing video and audio content to a dispersed audience by television, radio, or other means. Broadcasting database stores data such as subscriber information, event recordings, event schedules, etc., so it becomes important to store broadcasting data in the database.

Grocery Stores – Shopping. You're probably aware that grocery stores use databases to manage inventories, track sales, and personalise recommendations/vouchers based on purchase histories. What you might not be aware of is that large chain grocery stores, like online streaming services, track everything. From the moment you walk through the door, they're recording data like which direction you headfirst, what music is being played through the overhead speakers while you shop, where products are in the store on that particular day, and whether you pay through self-checkout or go to a cash register (in which case, they record the gender and age of the cashier you chose). And if you have ever used a payment method other than cash, they can tie all of those aspects together and associate them with you, personally.

eCommerce – Shopping online. Any online organisation that sells on a platform has to use a database to operate properly. In this case, databases help organise products, pricing, customer information, and purchasing history. The eCommerce store owner can then leverage their database to recommend other potential products to customers. This data would be stored in highly secure databases.

Sports – Sports. Fan participation in national sports doesn't just utilise the power of the database — it depends on it. From fantasy football to March Madness brackets (Basketball), the sports industry depends on massive databases to keep track of everything. Such databases store and analyse player statistics, game performances, injury reports, and more — always calculating the odds of a win on a weekly basis.

Banking – Finance. It is one of the major applications of databases. Banks have a huge amount of data as millions of people have accounts that need to be maintained properly. The database keeps the record of each user in a systematic manner. Banking databases store a lot of information about account holders. It stores customer details, asset details, banking

transactions, balance sheets, credit card and debit card details, loans, fixed deposits, and much more. Everything is maintained with the help of a database.

Insurance - Finance. An insurance company needs a database to store large amounts of data. Insurance database stores data such as policy details, user details, buyer details, payment details, nominee details, address details, etc.

Education. Universities have so much data which can be stored in the database, such as student information, teacher information, non-teaching staff information, course information, section information, grade report information, and many more. University information must be kept safe and secure in the database. Anyone who needs information about the student, teacher, or course can easily retrieve it from the database. Everything needs to be maintained because even after a number of years, information may be required, and the information may be useful, so maintaining complete information is the primary responsibility of any university or educational institution.

Healthcare – Medical. Doctors' offices and healthcare organisations store extensive amounts of patient data for easy accessibility. The databases behind this collection of information are massive, with complex data structures and security protecting sensitive data. All of these organisations have to ensure they comply with the Data Protection for data management.

Health and Fitness – Medical. Apps for tracking health and fitness data; step counter, fitbit, diabetes tracker, calorie counter.

Weather. Predicting the weather across the globe is incredibly complex. Weather organisations use prediction models that depend on various factors; all gathered, stored, and analysed within databases. These databases allow weather data to be always accessible and easily delivered to your local TV station or smartphone app. The "Weather Company", for example, takes in over 20 terabytes of data per day.

Travel Reservation Systems (Flights and Train). The systems store information such as passenger name, mobile number, passenger check-in, passenger departure, flight schedule, number of flights, distance from source to destination, booking status, reservation details, transport schedule, employee information, account details, seating arrangement, route & alternate route details, etc. All the information needs to be maintained, and available for retrieval. The data must be secure.

Library Management System – Document Management. There are hundreds and thousands of books in the library, so it is not easy to maintain the records of the books in a register or diary, so a database management system is used which maintains the information of the library efficiently. The library database stores information like book name, issue date, author name, book availability, book issuer name, book return details, etc.

Telecommunication – Communication. We cannot deny that telecommunication has brought a remarkable revolution worldwide. The Telecom field has huge data, and it is very difficult to manage big data without a database; that is why a telecom database is required, which stores data such as customer names, phone numbers, calling details, prepaid & post-paid connection records, network usage, bill details, balance details, etc.

Human Resource Management – Organisations. Any organisation will have employees, and if there are a large number of employees, then it becomes essential to store data in a database as it maintains and securely saves the data, which can be retrieved and accessed when required. The human resource database stores data such as employee name, joining details, designation, salary details, tax information, benefits & goodies details, etc.

Government. Government organisations around the world are constantly collecting data for research, defence, legislation, and humanitarianism purposes, to name a few. This data is collected, stored and analysed using powerful and far-reaching database services.

Manufacturing. Manufacturing companies manufacture a large number of related products on a daily basis. They need to maintain a record of products, quantities, purchases, payments, invoices, employee data etc. Production / Inventory / Orders / Supply Chain.

MEDIUM TASKS

Work in a small group to produce a comprehensive comparison of Relational Vs NoSQL Databases. Do research and identify definitions, benefits and limitations of Relational and NoSQL database systems.

a) Create a comparison table of features for the two approaches. You should consider the following:
- Definitions of Relational and NoSQL databases
- Benefits of Relational and NoSQL databases
- Limitations of Relational and NoSQL databases
- Examples of software that implements each approach
- Use Cases for each type of Database System

b) Produce a comparison table which includes a selection of features and how each system differs/implements them. Some examples of features to compare:
- Database structures – Type of data and how it is stored
- Data Storage – volumes of data
- Do they support ACID transactions (Atomicity, Consistency, Isolation, Durability)?
- Is Normalisation supported?
- Integrity constraints; Data Accuracy
- Scalability; horizontal and vertical scaling
- Simplicity: ease of use, support available
- Complexity Cost
- Reliability
- Schema Flexibility
- Performance; read and write
- Storage Requirements

ADVANCED TASKS

a) Create a presentation (1-3 slides) with the following information on each DB system: 1 benefit and 1 limitation, up to 2 features to compare how they are different, 1 use case for each database system

b) Present in the class