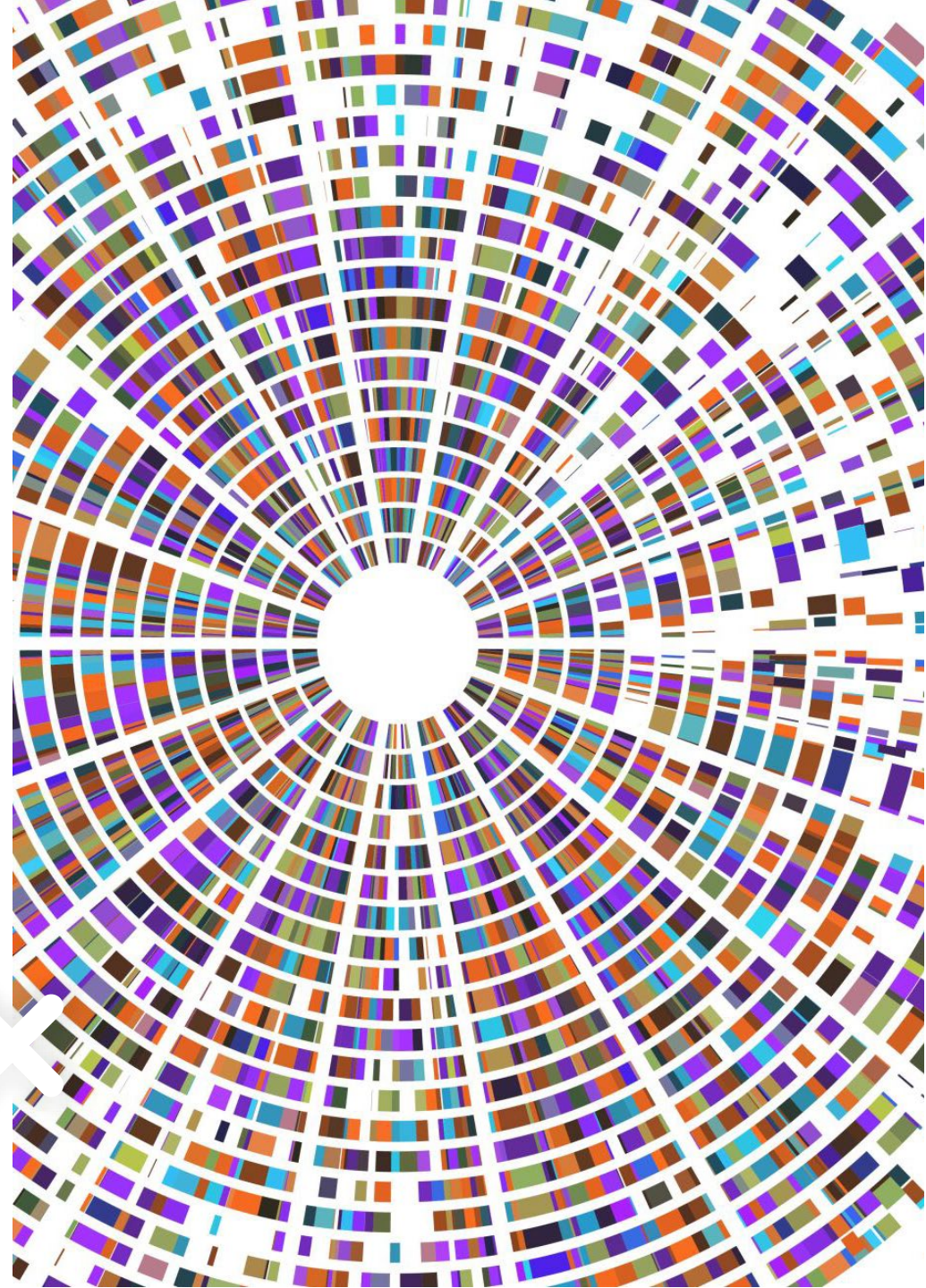# Advanced Database

Exam Preparation

Learning Outcomes

Design, create, and query

- A data warehouse
- Assessed via CA

Design, create, and query

- A distributed database
- Assessed via CA

Design, create, and query

- A document-oriented database
- Assessed via CA

Critically compare

- The strengths and limitations of different database technologies used in contemporary enterprise applications
- Derived from CA/**Assessed in Exam**

Discuss

- Recent developments and emerging trends in database technologies and their use in contemporary enterprise applications
- Derived somewhat from CA/**Assessed in Exam**

Learning Outcomes

Discuss

- Challenges of, and discriminate between approaches to, database modelling – conceptual, logical, and physical design
- **Assessed in Exam**

Discuss

- Database considerations for data integrity, integration, security, query optimisation, performance tuning and concurrency control in contemporary enterprise applications
- **Assessed in Exam**

Apply

- Techniques for data integrity, security, data optimisation, performance tuning and concurrency control
- Assessed in CA

Understand and apply

- Different approaches to data integration, both semantic and physical
- Apply assessed in CA/**Understand Assessed in Exam**

Understand and apply

- Data modelling techniques for logical and physical design to support contemporary enterprise applications
- Apply assessed in CA/**Understand Assessed in Exam**

Modelling

+ Relational v Non-Relational
  - Non-Relational Types
  - Strengths and Weaknesses
  - Uses

+ Conceptual V Logical V Physical
  - For each what is the focus
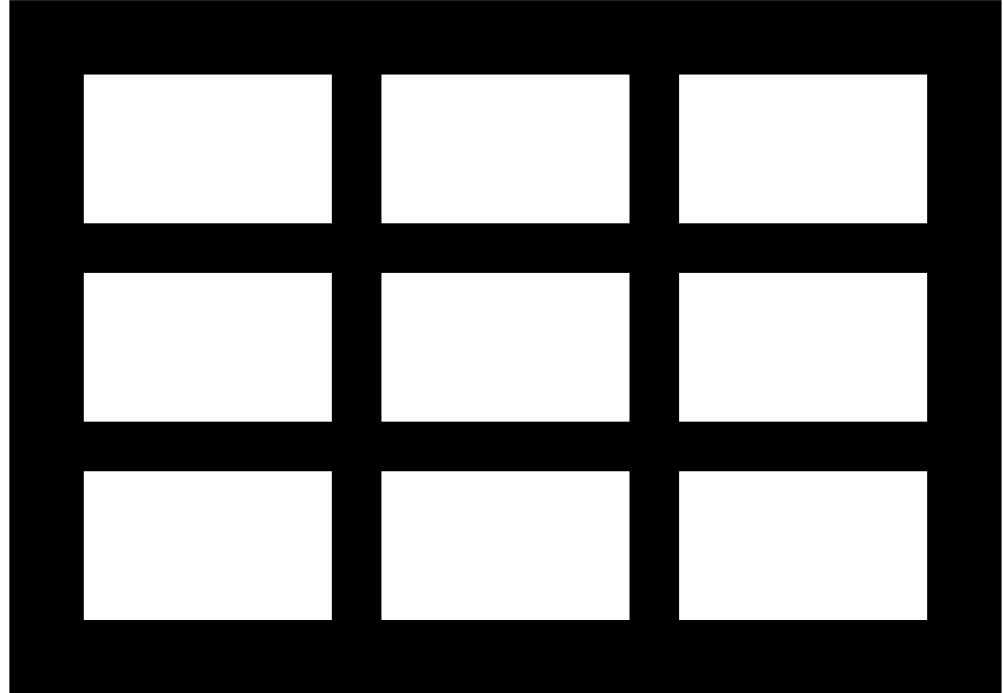 - For each in relational/non-relational what are the differences
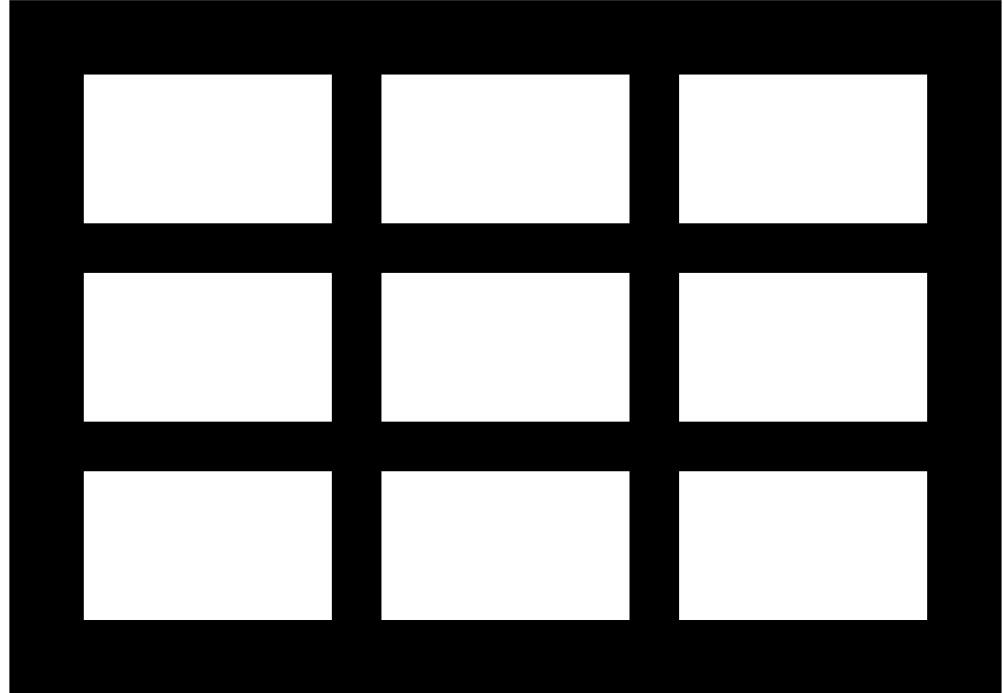
Normalization

+ Normalization
  - What?
  - Why?
  - How? Forms?
  - Advantages/disadvantages
 - Comparison with denormalization
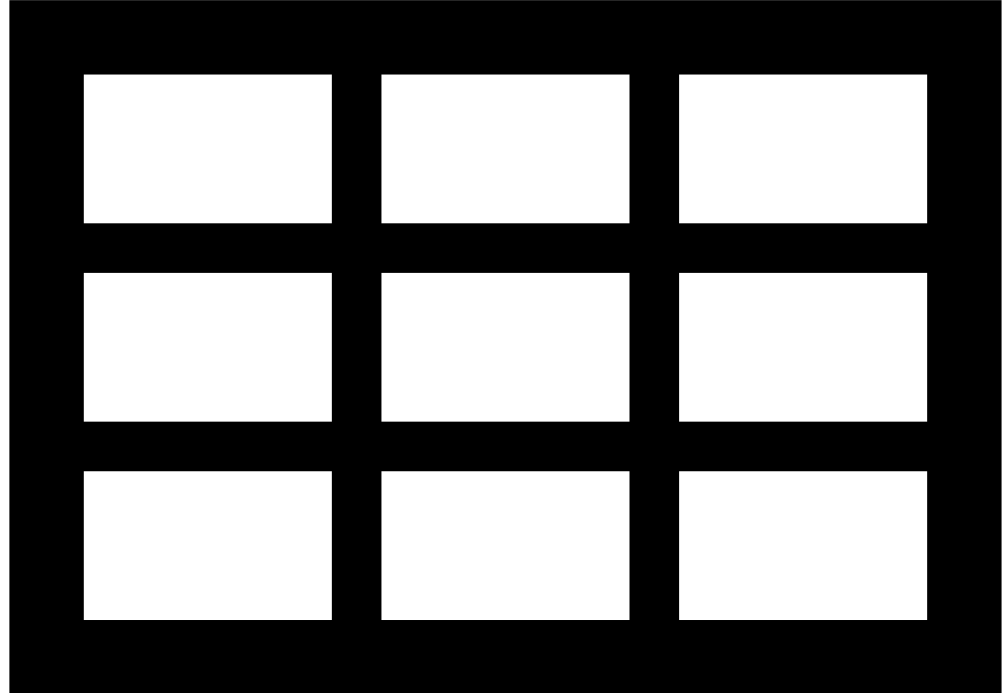
Denormalization

+ Denormalization
  - What?
  - Why?
  - How? Techniques
  - Advantages/disadvantages
  - Be able to work out what technique to use for a described situation

SQL V NoSQL

+ SQL v NOSQL (different types)
  o What?
  o Why?
  o Advantages/disadvantages
  o Query Processing
  o Query Optimization
  o Comparison
    + General
    + Databases used in module
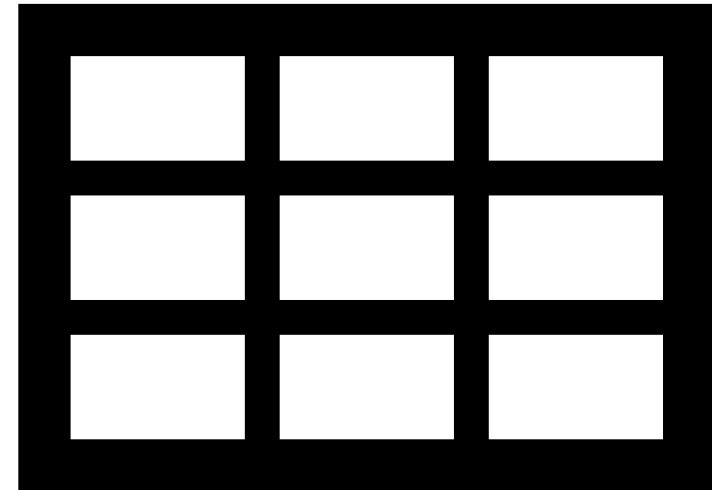 o Be able to work out for a described application
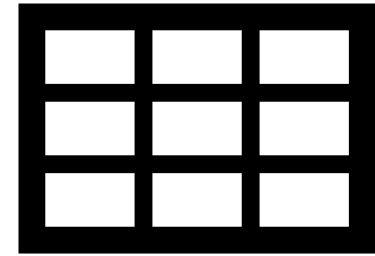
Database, Data Warehouse, Data Lake

+ Database v Data Warehouse v Data Lake
  o What?
  o Uses?
  o Uses?
  o Comparison
  o Be able to work out which is required for a described application
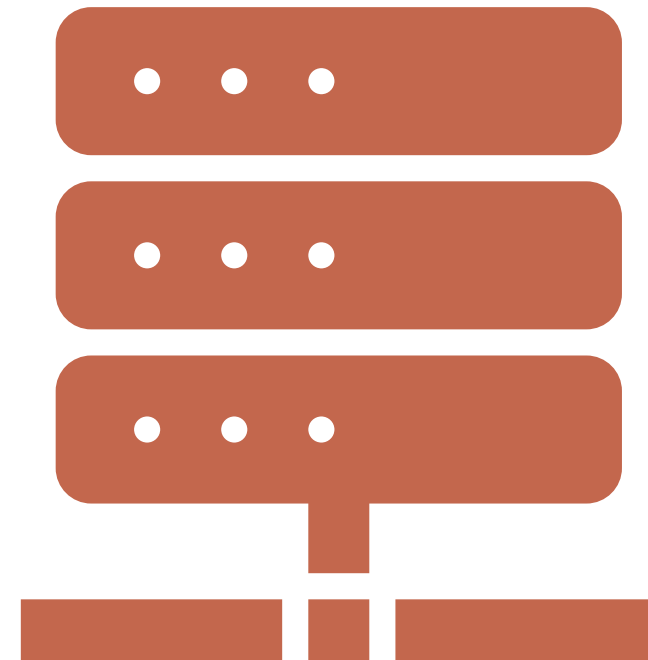
+ Hot v Warm v Cold data
  o What?
  o Uses?
  o Users?
  o Comparison
  o Be able to work out for a described application

ETL

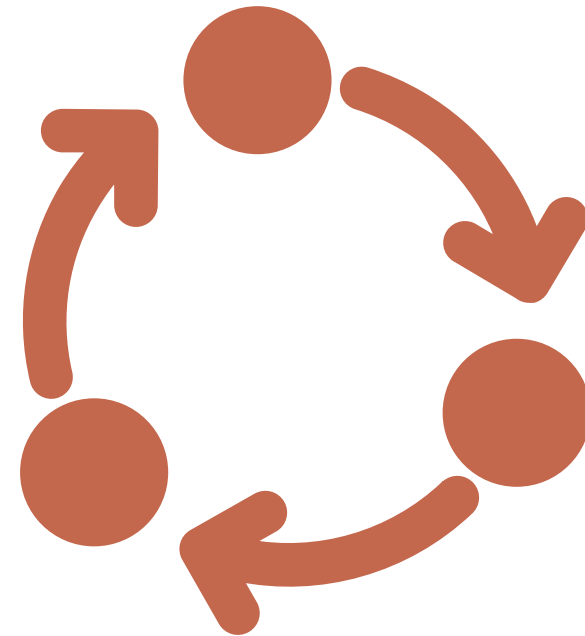+ Extract Transform and Load

+ What?

+ Why?

+ When?

+ What for?

+ How?
  o Different ways of handling SCDs
  o Be able to work out for a described situation

Transaction models

+ ACID v BASE
    o What?
    o When?
    o Comparison
        + For the databases used in module)
        + SQL v NoSQL
        + Database, data warehouse, data lake
        + Be able to work out for a described application

Performance Tuning/Query Optimization

+ Performance Tuning – Query Optimization

+ What?

+ When?

+ Why?

+ How?

+ Knowledge of and ability to describe approaches to optimization in different types of SQL/NoSQL data stores

+ Compare approaches in the databases used in CA

Distributed Data

+ Partitioning
  o Why?
    + Use Cases
  o Types (Vertical/Horizontal)
  o Differences
  o When to use each/Challenges of each
  o Relationship with replication
  o How it is done in the databases used in the module
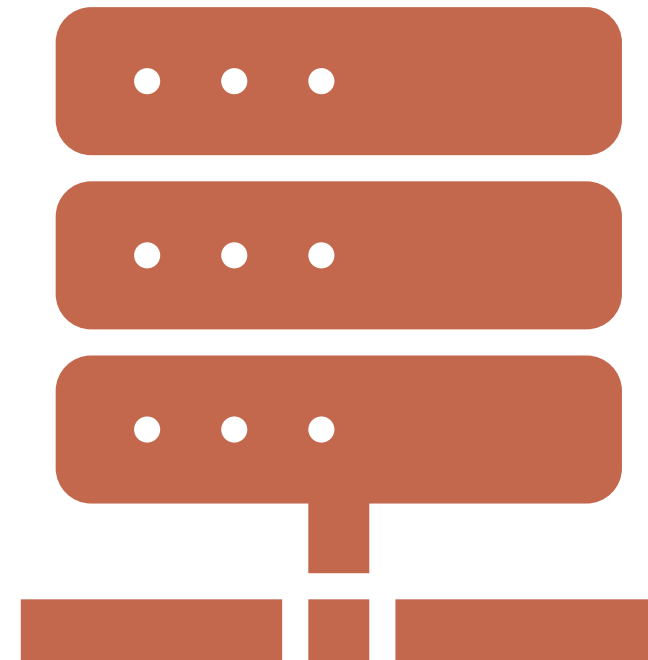
Distributed Data

+ Replication
  o Why?
    + Use Cases
  o Relationship with partitioning
  o How it is done in the databases used in the module
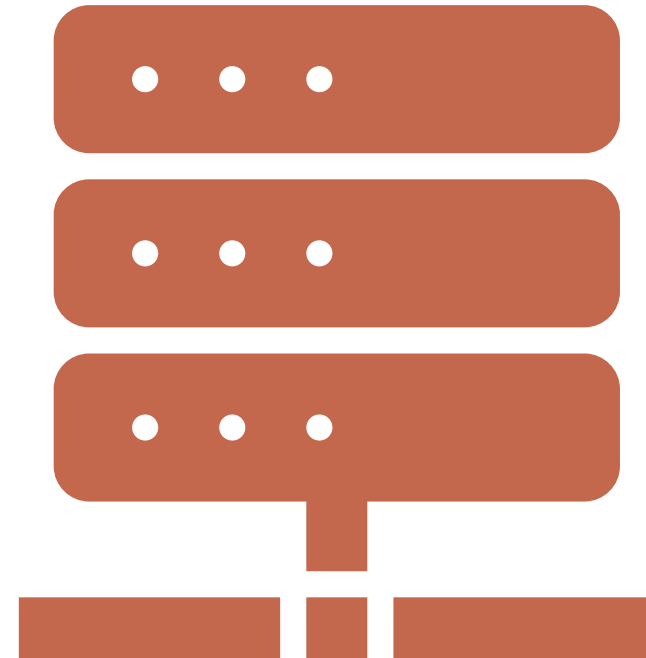
Query optimization

+ Indexes
  o Sparse/Dense
  o Primary/Secondary
  o B-Tree
  o Implementation in each database covered in the module
  o When to use different types of index
  o Issues related to tuning performance

Distributed Data

+ CAP theorem/PACELC Extension

+ What?
    ○ C, A, P, L

+ Why?

+ How to balance?

+ Be able to work out needs for a described application
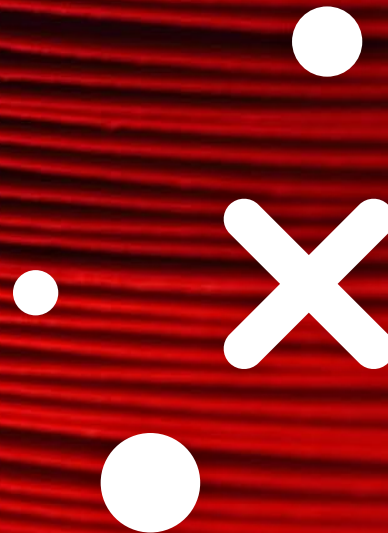
Sample
Paper

Question ONE

Sample Paper

Explain THREE (3) possible advantages and THREE (3) possible disadvantages of SQL compliant relational databases.

(6 x 2 marks)

Brainstorm

+The key here is the word explain – this means you cannot simply state an advantage or disadvantage you need to provide a bit more, explain why it is an advantage or disadvantage.

+Your answer must be in your own words.

Sample Paper

Describe THREE (3) possible advantages and THREE (3) disadvantages of NoSQL databases.

(6 x 2 marks)

Brainstorm

The key here is the word explain – this means you cannot simply state an advantage or disadvantage you need to provide a bit more, explain why it is an advantage or disadvantage.

Your answer must be in your own words.

Sample Paper

Suppose that you are tasked with implementing a distributed data solution for a retail enterprise who wishes to achieve the following:

a.   Store details of customers, their accounts, transactions against these accounts ensuring that all data is secure and consistent. The enterprise operates in several global regions. Customers are associated with a particular global region.

b.   Implement a chat utility for customer support which is available 24 x 7. Chat participants need to be able to view a full thread of chat conversations.

c.   Explore expansion into new regional markets, predicting expected profit levels, potential challenges to stock management using not only data owned by the retail enterprise but external sources such as regional regulatory information, taxation etc.

Your answer must be in your own words.

For each of the above, state whether you would implement a SQL or NoSQL solution. Justify your answer.

(3 x 3 marks)

Brainstorm

+This requires you to consider the answers you provided previously. Work out for each scenario which you think would work best SQL or NoSQL. You are making an overall choice – there may be some aspects of SQL that suit a scenario and some of NoSQL that suit a scenario. You need to make a judgment call, state it and discuss how you came to the conclusion.

Question Two

Sample Paper

(i)   Explain FOUR (4) key characteristics of a data warehouse.

(4 x 2 marks)

(ii)   Briefly compare a data warehouse and relational DBMS considering data design, data structure and access pattern.

(7 marks)

Brainstorm

+ (i) You will be able to find the characteristics in the lecture notes. The key here is the word explain – this means you cannot simply state a characteristic you need to provide a bit more about each characteristic.

+ Your answer must be in your own words.

+ (ii)   This is a straight-forward question. You will be able to work this out from the lecture notes. You need to construct you answer based on the things you are asked to consider – data design, data structure, and access pattern. For each of these discuss the issue for a data design then discuss the issue for a relational database – you can comment on similarities and differences.

Sample Paper

(i) Explain the ACID and BASE transaction models and when you would use each.

(12 marks)

Brainstorm

+The section of the course where they are discussed will allow you work out when you would use each. The key here is the word explain – this means you cannot simply state the properties for each model, you need to explain in your own words in a way that makes sense for the examples of use you provide.

Sample Paper

(ii)   Consider the following scenarios:

a.   Departmental managers in a retail company want to identify buying patterns of individual customers and different types of customers, analyse the impact of special sales promotions and determining future pricing policy for different products.

b.   A small marketing company wants to store data from social networks and conduct sentiment analysis on this data to explore the impact of its marketing campaigns, in particular involving TV advertising during prime time. Analysis will be differentiated between weekday and weekend sentiment.

For each, state whether you consider the ACID or BASE transaction model most suitable.

(2 x 3 marks)

Brainstorm

This requires you to consider the answer you provided previously. Work out for each scenario which you think would work best ACID or BASE. You are making an overall choice – there may be some aspects of ACID that suit a scenario and some of BASE that suit a scenario. You need to make a judgment call, state it and discuss how you came to the conclusion.

Question Three

Sample Paper

(i)   Explain what a secondary index is and its purpose in a database.

(4 marks)

(ii)   Suppose you are trying to improve query performance in a SQL compliant database and in a NonSQL data store. Should you add multiple indexes to the tables involved? Justify your answer.

(5 marks)

Brainstorm

(i) The key here is the word explain – this means you cannot simply provide a definition. You need to explain in your own words what an index is and why you would use one. This is a straight-forward question. You will be able to work this out from the lecture notes.

(ii) You will be able to find the answer in the lecture notes. The question asks for you to make a judgement but you should have a short discussion first. Why would you add multiple indexes? What would the impact be – positive and negative? What would you recommend – do it or not or do it with care? Your answer must be in your own words

Brainstorm

Suppose you are trying to improve query performance in a SQL compliant database and in a NoSQL data store. Should you add multiple indexes to the tables involved? Justify your answer

Can you do it in PostgreSQL? Can you do it in NoSQL?

Should you do it? What are the implications?

Sample Paper

Suppose you are implementing a chat system and are designing the database for a collection of messages. Each message has an author name, recipient name, content, sequence number and timestamp.

Explain how you would implement a secondary index to facilitate pattern matching on chat content in each of the following:

- PostgreSQL
- MongoDB
- Apache Cassandra

In your answer you must explain:

- The most appropriate type of index
- The potential disadvantages

You are not required to write any code.

(3 x 4 marks)

Brainstorm

You will have experience of doing this in the lab classes/CA. Pay attention to the detail of what you are being asked – the most appropriate type of index and the potential disadvantages. You are being asked specifically about PostgreSQL, MongoDB and Apache Cassandra so be specific about index types in those databases. Your answer must be in your own words.

Sample Paper

Suppose you are implementing a chat system and are designing the database for a collection of messages. Each message has a sender id, recipient id, chat content, sequence number and timestamp.

(i)   Other than using indexes, explain TWO (2) approaches which could be used to improve performance in a document NoSQL database and TWO (2) approaches which could be used in a distributed NoSQL wide column database.

(4 x 2 marks)

(ii)   Provide examples of situations in which you would use each approach.

(4 x 1 mark)

You do not need to provide any code.

Brainstorm

(i) Here you are looking at MongoDB and Apache Cassandra. Explain in your own words what else you can do to improve performance in each of these.

What could you do?

(ii) Using your answer to part (i) provide examples of how you would use each approach.

(iii) No code is required.

Question Four

Sample Paper

What is partitioning?

How can vertical and horizontal partitioning be used to improve database performance?

(10 marks)

Brainstorm

Explain what partitioning is then explain what vertical and horizontal partitioning are (this will explain the difference) then consider how each can improve database performance – so think about what could be partitioned in each approach.

Sample Paper

Explain the difference between partitioning and replication.
(4 marks)

Brainstorm

Explain what replication is first then then explain what is different to partitioning.

Sample Paper

(i)   Explain the CAP theorem.

(6 marks)

(ii)   Why is the PACELC extension to CAP important?

(3 marks)

Brainstorm

(i)   Explain in your own words what the CAP theorem is and then explain the C, A and P.

(ii)   Explain PACELC – why is it needed? What are the implications?

Sample Paper

Suppose that you are tasked with implementing a distributed data solution for a retail enterprise who wishes to achieve the following:

a.   Store details of customers, their accounts, transactions against these accounts ensuring that all data is secure and consistent. Online retail applications must be available 24 x 7.

b.   Implement a chat utility for customer support which is available 24 x 7. Chat participants need to be able to view a full thread of chat conversations.

c.   Explore expansion into new regional markets, predicting expected profit levels, potential challenges to stock management using not only data owned by the retail enterprise but external sources such as regional regulatory information, taxation etc.

Discuss the implications of each property of the CAP theorem for each scenario above and for each identify which properties are most important. Justify your answer.

(3 x 3 marks)

Brainstorm

Work out for each scenario which properties are most important. You are making an overall choice – there may be some aspects of each property that suit a scenario. You need to make a judgment call, state it and discuss how you came to the conclusion.

Tackling the Paper

Tackling the Exam Paper

+ Answer 3 of 4 questions

+ All questions are worth 33 marks (there is 1 free mark)

+ You have 2 hours to tackle this exam paper – 120 mins.

+ Use your time strategically

+ Attempt 3 Questions

+

Tackling the Exam Paper

+ How long should I spend on each question?

+ A general heuristic is to spend the same number of minutes answering a question as there are marks allocated.

+ In this case you have to answer 3 questions. Each is worth 33 marks so a general guide would be to spend 33 minutes answering each (this includes thinking time).

+ If we round that up to 35 marks per question. That works out at 105 minutes total.

+

Tackling the Exam Paper

+ When you reach 35 mins for a question stop and move on to another question.

+ Why?

    o The more questions you answer the better chance you have of gaining marks.

+ E.g.

    o If you answer two full questions that is a total of 66 marks available. If you do really well you may get 90% of those marks which will give you a result of 59.4% overall.

    o If you answer three questions there will be 100 marks available. If you answer 66% of each question then this is a better result.

+

Tackling the Exam Paper

+ How should I tackle answering a question?
    o Read the question
        + pay attention to the marks per part
    o Decide your answer
   + make notes
   o Write your answer.
    o Read your answer.

+
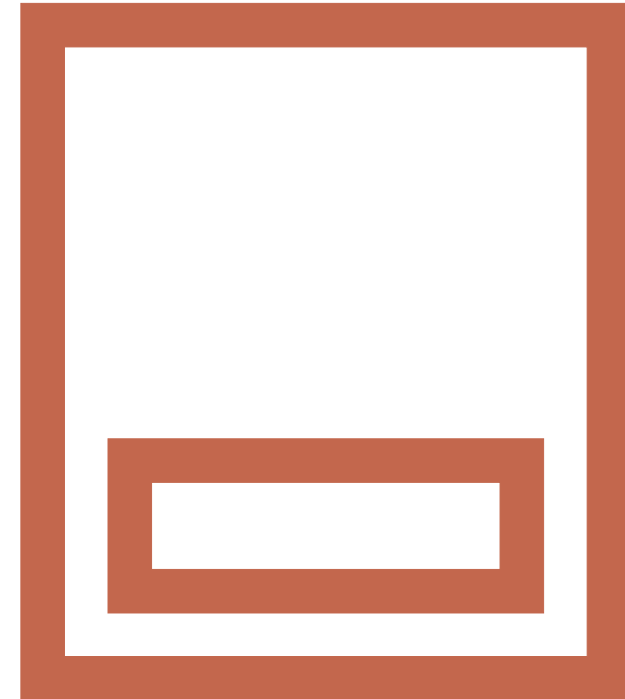
Tackling the Exam Paper

+ What should I do with the rest of the time?

+ If you follow the heuristic you are left with 15 mins.

+ A suggestion would be to use this time as follows:
  o At the start of the exam spend 5 mins reading the paper. Read every question and decide which questions you are going to answer.
  o When you are finished answering all the questions, revisit each question in turn
    + If you have left parts of the question unfinished quickly try to finish the ones that you think will be doable.
    + If you have finished answering all the questions read the questions again and make sure you have answered all parts of each question.
  o Max 10 mins

+
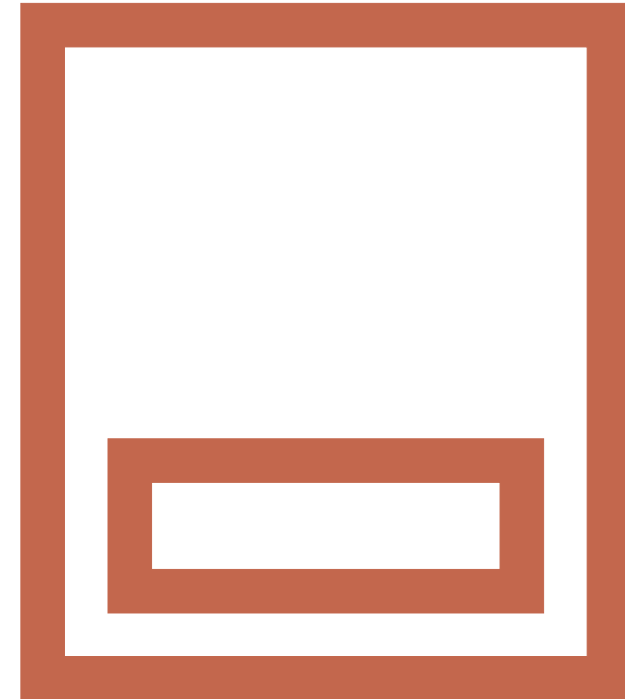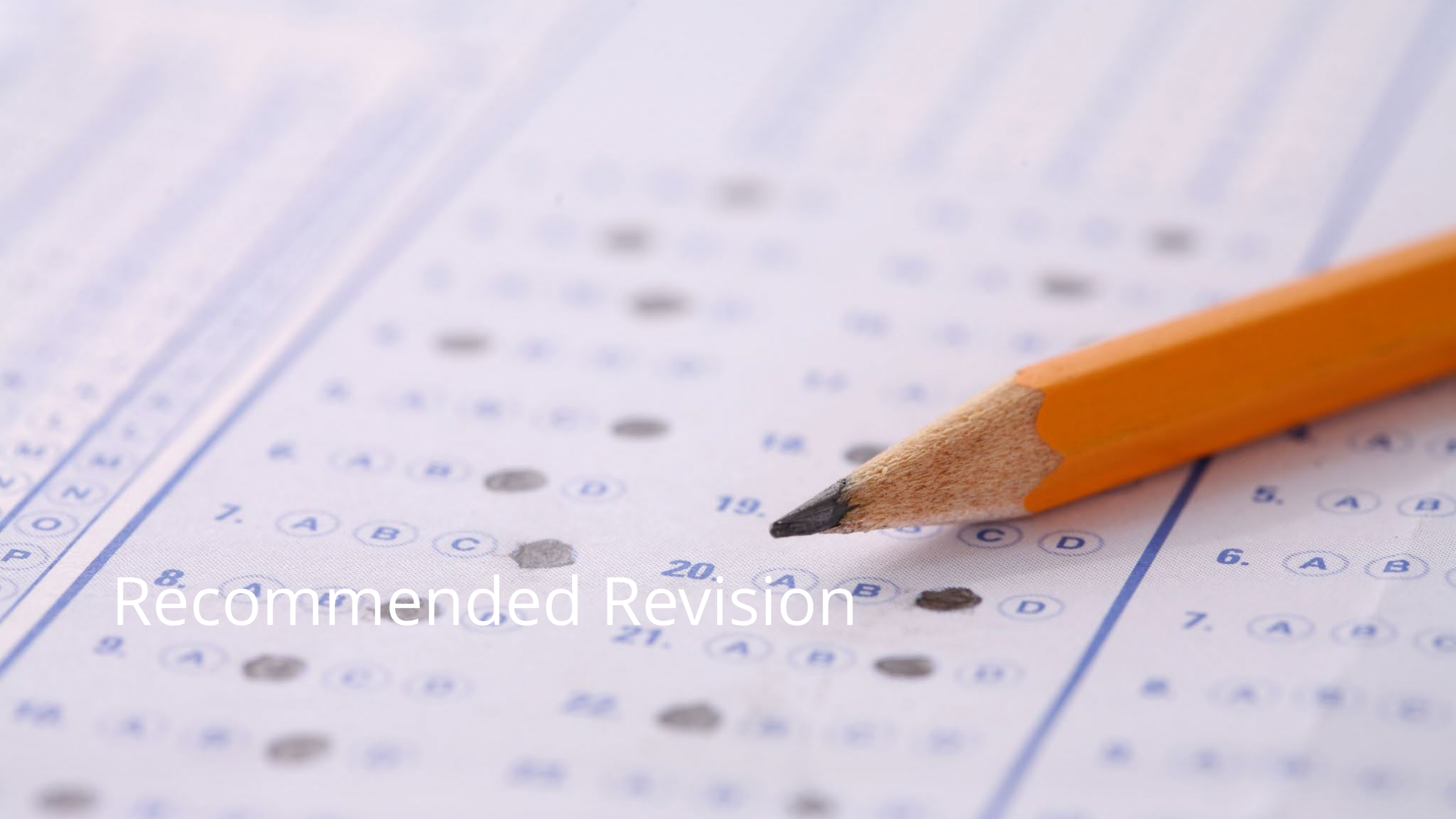
Recommended Revision

Questions to use in Revision

+ When would you use a SQL database? Why?
+ When would you use a NoSQL database? Why?
+ Users? Use Cases?
+ Advantages/Disadvantages

Questions to use in Revision

+ When would you use denormalization?
+ Why?
+ Advantages/Disadvantages?

Questions to use in Revision

+ When would you use a database/data warehouse/data lake?

+ Why?

+ Can you compare and contrast them?
  - OLTP/OLAP? Schema Flexibility? Freshness? Typical Users? Typical Use Cases? Access Pattern?

+ Key characteristics of each?

+ Advantages/Disadvantages of each?

+ Can you recommend one for a given example?

+ Hot/Cold/Warm Data
  - What are they? What type of database would you recommend?

Questions to use in revision

+ Indexes
    o What is an index? When would you use one?
    o What implications should you consider before adding?
    o Types (Sparse/Dense – Primary/Secondary)?
        + In relational? In non-relational?
    o B-tree
        + What? Why?
    o Indexes type in each database
        + PostgreSQL? Cassandra? MongoDB?

Questions to use in Revision

+ Query/Performance Optimization
  o What is query optimization?
  o What is a query optimizer?
    + What are the steps?
  o Other than indexes what could you do?
    + In PostgresQL? In Cassandra? In MongoDB?
    + Examples/Use Cases?

Questions to use in Revision

+ Partitioning
    o What is it?
    o What are the types?
    o When would you use each type? What would drive your decision?
    o How is it related to replication?

Questions to use in Revision

+ Replication
  - What is it?
  - When would you use it? What would drive your decision?
  - How is it related to partitioning?

Questions to use in Revision

+ CAP/PACELC
  - What is CAP? PACELC? Why are they needed?
  - What is C, A, P? How do you balance?
  - What is AC and LC ? When would you be concerned about each?
  - Given an example what would you choose and why?

Merry Christmas

Nollaig shona dhaoibh!

Click here for SQL Fun @ Christmas
https://youtu.be/3RsWqR9zSmw
Courtesy of Brendan Tierney