Comp 4522 Advanced Databases Final Review

MongoDB

1. What is it?
2. Architecture
3. Advantages/Disadvantages
4. Query Languages
5. Data Model
6. Data Storage

Distributed Databases

1. Concepts
   1. Multiple interconnected databases spread across different geographical locations that appear to the user as a single database.
   2. Has a Distributed DBMS that manages the DDB providing an access mechanism for users.
   3. Transparency – The separation of the higher-level semantics of a system from the lower-level implementation issues.
      1. Data Independence
      2. Network Transparency (or distribution transparency)
         1. Location transparency
         2. Fragmentation transparency
      3. Fragmentation transparency
         1. Horizontal Fragmentation = Selection
         2. Vertical Fragmentation = Projection
      4. Replication transparency
2. Architectures
   1. Types:
      1. Homogenous (Same DBMS)
      2. Heterogenous (Different DBMS)
   2. Structures:
      1. Client-Server
      2. Peer-to-peer
3. Challenges with distributed
   1. Scalability
      1. Adding Processing and storage power
      2. Scale-out
         1. Scale-up: increase the capacity of one server
   2. Design, such as how it is distributed, replicated and non-replicated database distribution.
   3. Distributed query processing for conversion of user transactions to data manipulation. Has an optimization problem.
   4. Distributed concurrency control.
4. Protocols
   1. Read one, write all protocol (ROWA)
   2. Eager centralized
   3. Eager distributed
   4. Lazy centralized
   5. Lazy distributed
5. Sync types
   1. Synchronous
   2. Deferred
6. Replication styles
   1. Eager Replication, changes occur within the scope of the transaction
   2. Lazy Replication, changes occur after the transaction executes on one copy, then changes propagate to the remaining copies.
   3. Centralized, there is only one copy which can be updated (the master) all other copies (slave copies) are updated reflecting the changes to the master copy.
   4. Distributed, changes can be initiated at any o the copies. That is, any of the sites which owns a copy can update the value of the data item.
7. CAP Theorem
   1. Stands for Consistency, Availability, and Partition tolerance; highlights the trade-offs involved in distributed database design.
   2. States that any distributed system or data store can simultaneously provide only two of three guarantees: consistency, availability, and partition tolerance
8. Advantages/Disadvantages
9. Horizontal/Vertical Scalability
   1. Horizontal scalability involves adding more machines or nodes to a pool, vertical scalability involves adding more power (CPU, RAM) to existing machine.

NoSQL/NewSQL Databases

1. Concepts
2. Architectures
3. Challenges with distributions
4. Protocols
5. Sync types
6. Replication styles
7. Advantages/Disadvantages
8. In-Memory DBs
9. In-Column DBs

Big Data

1. Architectures
2. Hadoop (Videos)
3. File System
4. Distributed Processing Mechanisms
   1. Map/Reduce
5. Storage Systems
6. Parallel Systems
7. Data Lake
8. Data Warehouse

Roles, Responsibilities & Standard Procedures

RDBMS SQL Query Processing and Optimization

RDBMS SQL Physical Storage