Exam2 Study Guide CSE30246 - Database Concepts

Indexing

How is data stored on disk?

Types of indexes

- Clustered vs Unclustered, Sparse vs Dense
- Composite Indexes, Why attribute order matters

B+ Trees (How are they different/better than B-Trees),

When and how do indexes help with Table Joins.

What is UTF-8 how is it different than ASCII?

Query Optimization

How does a DB decide what physical algorithms to run?

Cost Parameters, M, T, B, V

When can sorting, hashing, indexing help? How do we estimate this?

Rules based optimization

- Push down Optimization
- Pull up, Push Down Optimization

Cost based optimization

- Calculating a query plan cost

Transaction Management

ACID Properties

- What are they, why are they important?

Serial Schedules vs Serializable Schedules

2PL to enforce Isolation and serializability

Logging

How is the log written under various regimes:

- UNDO
- REDO
- UNDO/REDO

What are the differences in recovery?

How does checkpointing work in the log

What is (non)quiescent checkpointing? How does it work?

Hadoop

How does Hadoop split apart large files?

Why is replication important? Be able to perform an example.

What happens when a node in HDFS fails? What happens when an entire rack goes down?

Map Reduce

- What are the inputs and outputs of the mapper?
- What are the inputs and outputs of the reducer?
- What does the MapReduce subsystem do in between map and reduce?

How can Map Reduce be used to answer large SQL queries

Be prepared to design a map reduce program (pseudocode) that performs some SQL function

NOSQL

Describe the CAP theorem, name some databases that might apply to the different regimes

Why do columnar databases store their data in columns, and why is that congruent with HDFS?

What is SPARK? How is it different from MapReduce/HDFS systems?