Final Review

INF 551

Wensheng Wu

RDBMS

- SQL
- Constraints & views
- Data representation & external sorting
- Indexing
- Query execution

Big data

- NoSQL databases
  - MongoDB
  - DynamoDB
- Hadoop MapReduce
- Apache Spark

Advanced topics

- Data warehousing
  - Warehousing vs query-driven
  - OLAP vs OLTP
  - Data model & schema
  - Cube aggregation using SQL
  - Operators: rollup, drill down

SQL

- Select… from… where… group by… having…
- Subquery
  - (not) in
  - all/any
  - (not) exists
- Set operations: union, intersect, except
- distinct vs. union all
- Join: natural, outer, …
- Aggregations

Constraints & views

- PK, FK
  - Options for enforcing FKs
- Views
  - define
  - Using views to answer queries

Data representations

- Storing individual records
  - Fixed-length vs variable-length

- Storing records in a block

External sorting

- How to sort 1TB of data using 1GB of memory?

- Merge-sort:
  - 2-way
  - Multi-way
  - Cost

Indexing

- Clustered/un-clustered, dense/sparse

- B+-tree
  - Order/degree
  - Search, insertion, deletion, and their costs

Query execution

- Selection/projection
- Duplicate elimination, group-by
- Set operations: union, intersect, except
- Join

Query execution

- One-pass, 2-pass, k-pass algorithm
  - Cost, memory requirements
- Join algorithms & costs
  - Nested-loop
  - Sort-merge
  - Simple sort-based
  - Partitioned-hash
  - Index-based

MongoDB

- Data format?
- find():
  - pattern matching
  - operators ($gt, $and, $or, …)
- aggregate()
  - pipeline
- upsert

DynamoDB

- Structure of a table
  - Primary key
  - Heterogeneous items

MapReduce

- Architecture:
  - job tracker, task tracker
  - Shuffling (partition, sort, merge & group by)
- Map & reduce function
- Combiner:
  - Saving in communication cost?
- Operations: sum, max/min, count, avg., etc.
  - Commutative & associative?

Spark

- Creation: textFile, parallelize

- Transformations
  - map, flatMap, mapPartitions, mapValues, flatMapValues
  - filter
  - reduceByKey, groupByKey, aggregateByKey, sortByKey
  - distinct
  - Join (& outer joins)
  - Set operations (union, subtract, intersection, etc.)

Spark

- Actions
  - reduce, sum, min, max
  - count, mean, aggregate
  - countByKey
  - collect
- Which requires shuffling and why?
- Implement derived operations

Final exam

- 12/10, Monday
  - Morning section: 8-10am
  - Afternoon section: 2-4pm
  - Same classroom

- Closed-notes and books