Final Review

INF 551

Wensheng Wu

Foundations

- Storage systems
  - Hard drive, SSD
- File systems
  - Standalone, network

Not in final

XML, Xpath will NOT be in the final

- File formats
  - Unicode, UTF-8/16, JSON, XML, XPath

RDBMS

- Data modeling
  - ER, relational, conversion

- Query language
  - SQL
  - Relational algebra
  - Constraints (esp. FK)
  - Using views to answer queries

RDBMS

- Data organization & external sorting
  - Multiway merging & I/O costs

- B+-tree indexing
  - Searching: equality and range
  - Insertion (split may propagate to ancestor nodes)
  - Deletion (first try borrowing, then merging)

Big data

- 3V's
  - Volume, variety, and velocity
- HDFS
  - Hadoop distributed file system
  - Concept of replication

Big Data

- Cloud data storage
  - Amazon S3: data model
  - Eventual consistency model

- CAP theorem

Big data

- NoSQL
  - Different types
  - Scale up vs. scale out

- Key features, e.g.,
  - Flexible data model
  - High availability
  - Scalability

Big data

- Amazon DynamoDB
  - Data model, partition & sort key
  - Data types (string, number, set, map, list)
  - Consistent hashing
- Apache Cassandra
  - Write & read path
  - Upsert
  - Minor & major compaction

Big data

- Hadoop MapReduce
  - Architecture: job tracker, task tracker
  - Map and reduce functions
  - Concept of combiner
  - Shuffling

Big data

- Apache Spark
  - Concept of RDD
  - Transformations and actions
  - Lazy transformation
  - RDD reuse
  - Data frame, MySQL integration, SQL support

Big data

- Apache Hive
  - HiveQL: SQL-like language
  - Analyze data stored in HDFS
  - Queries compiled into MapReduce jobs

Workload

- Hive
  - Analytical workload (~ OLAP)
  - A query may need to process terabytes of data

- Cassandra & DynamoDB
  - Key-based (~ OLTP)
  - Processing a small amount of data per query

NoSQL

- MongoDB
  - Manage JSON documents
  - Key concepts: document, collection, primary key (_id)
  - Query language: insert, find, update, remove, aggregate
  - Sharding

Topics NOT covered in final

- Lectures 2 – 5
  - Storage system (hard drive, SSD, disk scheduling)
  - File system
  - Network file system

- Lecture 8: XML & XPath

- Lectures 16-17
  - Query execution (except for simple sort-based join algorithm: slides #41-43 in Lecture 16)
  - Data warehousing