VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

Exam and Lecture Overview

Lecturer : Dr. Pavle Mogin

SWEN 432
Advanced Database Design and Implementation

Plan for Final Lecture

- What you may expect to be asked in the Exam?
- The answer:
  - All what we have learned is examinable, except the topic of the essay
- Exam questions will cover all the important topics we learned in lectures and applied in assignments
- 120 marks, 120 minutes
- Since we covered many topics, quiet a number of questions will be asking for a rather short answer
- Use Lecture Notes and Assignments as your main source for making the preparation

Exam – Structure and Content

- Exam structure will follow the structure of lectures:
  - Cloud Databases
    - General features of Cloud Databases (~23%)
    - Cassandra (~44%)
    - MongoDB (~16%)
  - Data Warehousing and OLAP (~17%)

Cloud Databases – General Features

- Basic features
  - Scalability
  - Availability,
  - Shared nothing,
  - Data partitioning and replication,
    - Consistent Hashing,
    - Master – Slave,
    - Membership changes
  - Data versioning,
  - Gossip protocol
- Trade-offs in cloud databases
  - CAP theorem
  - BASE
    - Range of consistency levels
    - Availability and consistency trade-offs

Cassandra (1)

- Data model:
  - Column families with a CQL abstraction layer
- CQL:
  - Keyspace (column families having the same replication factor),
  - Tables (abstracting column families) (have non default features),
  - Query syntax: very close to the relational SQL, but not that powerful and with many restrictions not present in SQL

Cassandra (2)

- Data modeling
- Storage Engine
  - Table Primary Key and Partitioning
  - Storage Engine Rows
  - Log Structured Merge Trees (mem table, log file, SSTables)
  - Write Paths for Insert and Update
  - About Reads (use of Bloom filters)
  - About Deletes (use of tombstones)
  - Compaction (deleting obsolete and deleted column values, uniting row fragments)
- Consistency levels:
  - Spread from strict via strong to eventual
  - Adjustable per a DML statement

Cassandra (3)

- Architecture
  - Internode Communication
    - Gossip Protocol
    - Seed Nodes
  - Partitioning
    - Consistent hashing
    - A partitioner maps a partition key value into a token
  - Snitches (files containing topology information)
  - Keyspace and the replication strategy
  - `cassasndra.yaml`
  - `cassandra-topology.properties`
- Light weight transactions (CAS):
  - To avoid overwrite of other people’s work

Cassandra (4)

- Repair mechanisms
  - Read repair
  - Hinted Handoff Writes
  - Extreme write availability
  - Anti-entropy Node Repair (Merkle Trees)

MongoDB (1)

- Data Model:
  - Data representation by documents
  - Document implementation by:
    - Embedding or Referencing
  - A rich query language:
  - Shell methods:
    - find(), insert(), update()
    - Simple aggregation (count(), distinct()),
    - Pipelined Aggregation with several stages, each having several expressions,
    - Powerful when combined with java scripts

MongoDB (2)

- Architecture
  - Sharding:
    - Shard Key,
    - Balancing Data Distribution,
  - Sharding architecture
    - Drivers (accept client's requests),
    - Routers (mogos processes - rout user requests to appropriate replica set using meta data),
    - Configuration servers (contain routing information)
    - Replica set (contain shard data)
  - Replication:
    - Master - Slave Mode,
    - Replica Set Operation,
    - Failover - Election of a new Master (HA Cluster)

Data Warehousing and OLAP (1)

- OLAP Database Structures
  - Star Schema
    - Facts,
    - Dimensions,
    - Attributes and attribute hierarchies,
    - Snowflakes, and
    - Constellations
  - Hyper Cube (Multidimensional Cube)
- Basic OLAP Queries
  - Roll-up (hierachical and dimensional)
  - Drill-down,
  - Slice and Dice, and
  - Pivoting.

Data Warehousing (2)

- OLAP Specific Extensions of SQL
  - CUBE,
  - ROLLUP,
  - WINDOW, and
  - RANK
- Materialized Views
- Dimension Hierarchy
  - A syntax for defining attribute hierarchies and functional dependencies between attributes
- Query Rewriting
  - Text match, and
  - General query rewrite with:
    - Join compatibility check,
    - Data sufficiency check,
    - Grouping compatibility check, and
    - Aggregate computability check

Data Warehousing (3)

- Aggregate Functions
  - Distributive (SUM, COUNT, MIN, MAX),
  - Algebraic (AVG, VAR), and
  - Holistic (Median, RANK, TopN)
- OLAP Architectures
  - Virtual Datawarehouse,
  - Data Mart in a Box,
  - Stovepipe Data Mart,
  - Architected Data Mart,
  - Enterprise Data Warehouse with HOLAP