Distributed Databases

School of Information Sciences, Manipal

Monsoon 2016

What is a Distributed System?

A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

- Leslie Lamport

Architectures for a Distributed System

Supported data models

Data placement

Communication protocols

Types of applications

Quality Attributes of a Distributed System

Efficiency of data manipulation

Latency

Throughput

Failure models

Security

Client-Server Model

Idetifies two roles in a context.

Partitions responsibilities.

Enables loosely-coupled communication.

Allows largely independent evolution of technical architectures.

Client-Server Model - Recent Developments

Clients have more computing power

- servers routinely off-load mundane computational tasks to clients

Clients operate within versatile runtime environments.

Security has taken a central role in client-server application development.

Parallel Databases

Multiple processorscontrol multiple disk units that host the database.

Database itself may be partitioned or replicated on disks.

Three models of parallel database management

- Shared memory
- Shared disk
- Shared nothing

Parallel Databases - Shared Memory Systems

In a shared memory system all processors share the main memory.

All processor share disks that contain the database

- When a processor requests data, database pages are transferred to main memory buffers that are shared across processors.

Parallel Databases - Shared Disk Systems

In a shared disk system each processor has exclusive access to private memory.

All processor share disks that contain the database

- When a processor requests data, database pages are brought to that processor’s memory.

Parallel Databases - Shared Nothing Systems

Each processor has an exclusive access over a set of disk units.

Each processor has access to private memory.

Processors may communicate over an interconnection network.

This architecture offers potentially linear scaleup. It also provides linear speedup.

Parallel Databases - Cluster Architecture

Multiple shared memory systems are wired over an interconnection network.

Distributed Databases

The distribution of data and control is *transparent* to the users.

A distributed system may be

Homogeneous

- software and hardware subsystems are more uniform

Heterogeneous

- Software and hardware could potentially represent disparate models
- translation of messages and data is mandatory

Components of a Distributed Database System

Local database management (LDBMS) component

Global data dictionary

- repository of location information
- list of data objects
- data locations
- data schema

Components of a Distributed Database System

Distributed database management (DDBMS) component

- enables location transparency.
- locates data leveraging the global data dictionary.
- processes queries (*local*, *remote*, and *compound*).
- provides network-wide concurrency control.
- provides network-wide recovery procedures.
- provides translation of queries and data in heterogeneous systems.

Data Distribution

These are the attributes to consider

- closer to the computation that requres it
- Reliability
- availability
- storage capacities and costs
- communication costs
- distribution of processing load

Data Placement Alternatives

Centralized

Replicated

Partitioned

Hybrid

Data Placement Alternatives

Centralized

- centralized database and clients are distributed
- no global data dictionary
- centralized resources are the bottleneck
- availability is poor if transaction requests are high
- locality of data reference is low

Replicated

Partitioned

Hybrid

Data Placement Alternatives

Centralized

Replicated

- database instance is replicated in distinct nodes
- improves reliability
- improves availability
- cost of updates are very high

Partitioned

Hybrid

Data Placement Alternatives

Centralized

Replicated

Partitioned

- database is partitioned into disjoint fragments
- columns or rows may be the basis of partiotion
    - projections must be lossless
- if organized properly, this scheme results in good performance

Hybrid

Data Placement Alternatives

Centralized

Replicated

Partitioned

Hybrid

- different partitions may be distributed in different modes
- Very careful analysis and design is required
    - data that is frequently updated is centralized
    - data which is frequently read is distributed

Transparency in Distributed Databases

Data distribution tansparency

- fragmentation transparency
- location transparency
- replication transparency

DBMS heterogeneity transparency

Transaction transparency

- concurrency
- recovery