॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Distributed Data Storage and Management

Saptarshi Pyne
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur, Rajasthan, India 342030

CSL4030 Data Engineering Lectures 11, 12
August 25th, 28th, 2023

What we discussed in the last class

- NoSQL databases
    - The graph-like data model: Suitable when
        - we have documents having many-to-many relationships
    - Graph database management systems
        - Neo4j by Neo4j Inc.
        - RDF4J by Eclipse Foundation (open source)

Today we will discuss

- Distributed data storage and management

What is a distributed database?

- A distributed database is a database in which data is stored across multiple **interconnected computers** that might be at different geographical locations a.k.a. **sites**.

Homogeneous vs. Heterogeneous distributed databases

- **Homogeneous** distributed database: All computers have identical DBMS and follow the same schema (or at least aware of each other's schemas).

- **Heterogeneous** distributed database: Different computers may have different DBMS and different schemas.

Data storage

- Fragmentation
  - Horizontal fragments or ‘shards’
  - Vertical fragments
  - Fragments of fragments

- Replication
  - The ‘primary copy’ and its replicas
  - *Pros:* Increases availability of read-only data
  - *Cons:* Complicates write operations

- Combined (replicas of fragments): When?

Query processing with fragments

- Select + union
- Project + natural join
  - Over any superkeys, e.g., the primary key
  - Over ‘tuple-id’

References

- A. SILBERSCHATZ, H.F. KORTH, S. SUDARSHAN (2011), Database System Concepts, McGraw Hill Publications, 6th Edition.
  - Chapter 19. Distributed Databases

- Paper: Bronson et al., “TAO: Facebook’s Distributed Data Store for the Social Graph”, 2013 USENIX Annual Technical Conference (USENIX ATC ‘13).
  - Video: https://www.usenix.org/conference/atc13/technical-sessions/presentation/bronson

Thank you