Advance Database

Short Notes based on ***Exit Exam***

Zemzem labs

1. Distributed Database Systems (DDBMS)

What is DDBMS?

A database whose data is distributed across multiple sites but appears as a **single logical database**.

Types of Distribution

1. **Fragmentation**
    - **Horizontal fragmentation:** rows divided
    - **Vertical fragmentation:** columns divided
    - **Mixed fragmentation**
2. **Replication**
    - Full replication
    - Partial replication
    - No replication (partitioning only)

Advantages

- Reliability & availability
- Faster local access
- Scalability

Disadvantages

- Complex design
- Increased overhead
- Data consistency challenges

2. Distributed Transaction Management

Two-Phase Commit Protocol (2PC)

Ensures distributed transactions commit atomically.

Steps:

1. **Prepare Phase:** coordinator asks all sites to ready
2. **Commit Phase:** if all agree → commit; else rollback

Three-Phase Commit (3PC)

Non-blocking version of 2PC, with an intermediate "pre-commit" phase.

3. Concurrency Control in Distributed Systems

Techniques

- Distributed 2PL (Two-Phase Locking)
- Timestamp ordering
- Validation-based control

Deadlock Handling

- Centralized deadlock detection
- Distributed deadlock detection
- Timeout-based

4. Parallel Databases

Types of Parallelism

- **Inter-query** parallelism
- **Intra-query** parallelism
- **Intra-operation** parallelism
- **Inter-operation** parallelism

Architecture

- Shared memory
- Shared disk
- Shared nothing (best for scaling)

5. Data Warehousing

Data Warehouse Definition

A large store of historical data:

- Subject-oriented
- Integrated
- Time-variant

- Non-volatile

ETL Process

1. Extract
2. Transform
3. Load

Schemas

- **Star Schema (denormalized):**
  - Fact table + dimension tables
- **Snowflake Schema (normalized):**
  - Dimension tables normalized
- **Fact Constellation (Galaxy Schema):**
  - Multiple fact tables

6. OLAP (Online Analytical Processing)

Used for multi-dimensional analysis.

OLAP Operations

- **Roll-up:** summarize
- **Drill-down:** detailed view
- **Slice:** fix one dimension
- **Dice:** select sub-cube
- **Pivot:** rotate data cube

Types of OLAP

- MOLAP → Multidimensional
- ROLAP → Relational (SQL-based)
- HOLAP → Hybrid

7. Big Data & Hadoop Ecosystem

Big Data Characteristics (3V or 5V)

- **Volume**
- **Velocity**
- **Variety**
- **Veracity**
- **Value**

Hadoop Components

- **HDFS:** distributed file system
- **YARN:** resource management
- **MapReduce:** parallel processing

MapReduce

Consists of:

- **Map function** → processes input into key–value pairs
- **Reduce function** → aggregates/interprets results

8. NoSQL Databases (Advanced Perspective)

Types & Examples

- **Document Stores:** MongoDB, CouchDB
- **Key–Value Stores:** Redis, DynamoDB
- **Column–Family:** Cassandra, HBase
- **Graph DB:** Neo4j

CAP Theorem

A distributed system cannot guarantee all 3 simultaneously:

- **Consistency**
- **Availability**
- **Partition tolerance**

Must choose any 2.

9. Advanced Indexing

Index Structures

- **B-Tree, B+ Tree** – standard for relational DB
- **Bitmap Index** – best for low-cardinality attributes
- **Hash Index** – fast exact-match retrieval
- **R-Tree** – spatial data indexing (GIS)
- **Trie-based indexes** – text search

Clustered vs Non-Clustered

- **Clustered:** table ordered by index
- **Non-clustered:** separate index structure

10. Query Optimization (Advanced)

Optimization Techniques

- Cost-based optimization
- Rule-based optimization
- Join reordering
- Index selection
- Materialized views

Query Execution Plan

- Chosen by optimizer
- Includes:
    - Join method
    - Index usage
    - Operation sequence

Join Algorithms

- Nested loop join
- Sort-merge join
- Hash join
- Block nested loop join

11. Temporal, Spatial, and Object Databases

Temporal Databases

- Store time-related data

- Valid-time & transaction-time attributes

Spatial Databases

- Store geometric data
- Use **R-tree indexes**

Object-Oriented Databases

- Support objects, inheritance, encapsulation

12. XML & JSON Databases

XML DB

- Uses DTD, XML Schema
- XPath, XQuery for querying

JSON DB

- Used by NoSQL systems
- Flexible schemas

🌟 Ultra-Condensed Exam Summary

- DDBMS = fragmentation + replication
- 2PC ensures atomic distributed commits
- OLAP operations = slice, dice, roll-up, drill-down
- Data warehouse = subject-oriented, time-variant
- Star schema = denormalized; Snowflake = normalized
- Indexes: B+ tree, Bitmap, R-tree
- CAP theorem = choose 2 of CA, AP, CP
- Hadoop = HDFS + MapReduce + YARN
- NoSQL = document, key-value, column, graph