Principles of Distributed Database Systems - Background

Diego ROJAS

September 18, 2022

1 Overview of Relational DBMS

1.1 Relational Database Concepts

A *relational database* is a structured collection of data in the form of tables.

A *relation R* defined over n sets $D_1, D_2, \ldots, D_n$ is a set of *n-tuples* $\langle d_1, d_2, \ldots, d_n \rangle$ such that $d_1 \in D_1, d_2 \in D_2, \ldots, d_n \in D_n$.

Relation schemas are definitions of the *attributes*, *entities* and relations of a database.

$$\text{EMP(\underline{ENO}, ENAME, TITLE, SAL, \underline{PNO}, RESP DUR)}$$
$$\text{PROJ(\underline{PNO}, PNAME, BUDGET)} \tag{1}$$

The *key* of a relation is the minimum non-empty subset of its attributes which uniquely identify each tuple of the relation.

The number of attributes of a relation defines its *degree*, whereas the number of tuples of the relation defines its *cardinality*.

1.2 Normalisation

Normalisation aims to eliminate various undesirable anomalies of a relation.

- *Repetition anomaly*

- *Update anomaly*

- *Insertion anomaly*

- *Deletion anomaly*

A relation with any of the aforementioned anomalies is split into two or more relations of a higher *normal form*.

1.3   Relational Data Languages

Query languages for relational databases fall into two fundamental groups: *relational algebra* and *relational calculus*.

1.3.1   Relational Algebra

Comprised of the following fundamental operators:

- *Selection*

- *Projection*

- *Union*

- *Set difference*

- *Cartesian product*

From the set of fundamental operators we can derive:

- *Intersection*

- $\theta - join$

- *Natural join*

- *Semijoin*

- *Division*

And much more.

1.4 Relational Calculus

Relational calculus languages describe what the result should be by stating the relationships that are supposed to hold for that result. SQL is a great example of a relational calculus query language.

2 Review of Computer Networks

A *computer network* is a set of interconnected and autonomous computers that exchange information.

Computers on a network are referred to as *hosts*, *nodes*, *end systems* or *sites*.

Computers on a network are connected using *switches* that route messages through the network.

The most commonly used network is the internet which can be defined as a network of networks.

2.1 Types of Networks

2.1.1 Scale

The *scale* of a network refers to its geographic distribution.

2.1.2 Topology

The *topology* (or *interconnection structure*) of a network refers to the way nodes are connected to one another.

Common network topologies include.

- *Irregular*
- *Star*
- *Bus*

- *Ring*

- *Complete* (or *mesh*)

2.2 Communication Schemes

Either *unicast* or *broadcast* (optionally *multicasting*).

2.3 Data Communication Concepts

Data is transmitted through *links* that carry one or more *channels*.

Each communication channel has a *capacity* often called *bandwidth*. Usually it is expressed as bits-per-second (bps) within digital links.

The *data transfer rate* represents the rate at which data is transmitted across the network. It is usually less than the capacity of the network due to possible errors during transmission (messages that are truncated, corrupted or lost).

Messages are usually split into *packets*. Their size is usually dependent on the frame size.

2.4 Communication Protocols

Error-free, reliable and efficient communication between hosts requires the implementation of elaborate software systems that are generally called *protocols*.

Protocols are often stacked in what is called a *protocol suite/stack*.

The most commonly used protocols are *TCP* and *UDP*.