1, Check if "a" exists in S.A; 2, Check with N1 and N2 to see if S1 and S2 can be updated; 3, Insert (a, b) to S1 and insert (a, c, d) to S2

Q: What is data replication?; A: Data replication is the process of making multiple copies of data and storing them at different locations to improve their overall accessibility across a network.; Advantage: Return queried records faster; Data backup; Disadvantage: Extra storage cost; Update cost

Because S.A is a foreign key, we fragment the S table just in the same way as R, and make sure that the fragmentation after fragmentation is refactored is still the S table.

Approach 1: stie 2 S -> site 1; Approach 2: site 1 all unique values of R.A -> site2; site 2 S semijoin R.A -> site 1

Horizontal Fragmentation: ; Completeness: if any t belongs to R, exist Fi belongs to F, then t belongs to Fi; Disjointness: if any Fi, Fj belong to F, i is not equal to j, then Fi intersecting Fj is empty; Reconstruction: R is equal to F1 union F2 union F3 union ... union Fn; Vertical Fragmentation; Completeness: A is equal to A1 union A2 union A3 ... union An

Suppose we have N copies.; Whenever there is a need to update them, we update the majority of copies, say m(m > N/2) copies. The version number of each copy is updated at the same time too.; When we read, we need to read at least n copies such that n + m > N.

Yes. In the voting approach, we need to read more than one copy. Adopt the Read-any Write-All approach so we only need to read one copy at a time. We rarely need to write because update queries are rare.

Synchronous Replication: 1. Voting 2. Read any Write all; Asynchronous Replication: 3. Primary Sit 4. Peer-to-Peer replication

Atomicity; Consistency; Isolation; Durability.

Delete, insert or substitute one character

Advantages of Data Warehouse:; OLAP Queries are usually aggregate queries.; Materialized views make queries much faster and response times are interactive.; Disadvantages of Data Warehouse:; Data Warehouse can be outdated relatively quickly; Increased storage costs; Materialized views must be updated when underlying data tables are modified.

Database updates must lock data resources. Large scale aggregation reports lock many resources for a long time.; Because the data warehouse is historical, the existing business data in the past will not be updated generally, but will always be there waiting for various access select.

Operational systems focus on Data.; Data warehousing systems focus on Information out.; Operational system is used for Online Transactional Processing.; Data warehousing system is used for Online Analytical Processing.

In a star schema, all information is placed in the fact table and the lookup tables that have a direct reference to the fact table.; In a snowflake schema, the first-level lookup tables may have their lookup tables. So, the information is dispersed over the entire system.; Star schema results in high data redundancy and duplication. Snowflake schema ensures a very low level of data redundancy (because data is normalized).;

The slice is an operation that selects one specific dimension from a given data cube and provides a new subcube.; The dice is an operation that selects two or more dimensions from a given data cube and provides a new subcube.; Pivoting operating is one of the OLAP operations, it is essentially rotating the data cube in order to perceive the cube from a different dimension.

Q: Bitmap indexing is a useful technique in data warehousing. Taking this cube as an example, briefly discuss the advantages and problems of using a bitmap index structure.; A: Advantages: small size; speed up query.; Disadvantages: Not suitable for attribute with high cardinality.; Not suitable for attributes that are updated frequently.

A data cube in a data warehouse is a multidimensional structure used to represent data along with some measure of interest.

Entity resolution is the task of disambiguating records that correspond to real-world entities across and within datasets.

Semi join is a technique for processing a join between two tables that are stored sites.

Data linkage is an operation to identify records referring to the same real-world entity.

Jaccard coefficient. When measuring string similarity using JD, strings are broken into a set of Q-grams. The order of elements in a set does not matter. The edit distance between the same names is large if the first and last names are written in a different order.; Edit Distance. Last and first names are written in the same order. ED is defined to be the minimum number of operations to transform one string to another. However, a minor type can cause a huge change in Jaccard distance since their Q-grams are completely different. e.g. “Emily Smith” VS “Emmily Smith”

Schema heterogeneity:; S1: Employee (ID, name, position, salary); S2: Worker (EID name); Position (EID, PID, salary); Data type heterogeneity:; Employee ID could be a string or an integer; Value heterogeneity:; The "Employer" could be called "Worker" in another system; Semantic heterogeneity:; Salary is hourly salary or is weekly salary with allowances

Q: Why is data linkage so difficult in practice?; A: The same real-world object can be represented as different strings.; The same string can represent different real-world objects.

Relational DB: You can query a view like you can a table. A view can combine data from two or more table, using joins, and also just contain a subset of information.; Distributed database system: used in the bottom-up approach of database design; Data

warehousing system: Store pre-calculated expensive joins to speed up online OLAP queries.; Federated database systems: Provide a virtual view of integrated data without actually bringing data into a physical centralized database.

Step 1: Data is divided into overlapping subsets, called canopies.; Step 2: Expensive distance measurement made among points within the same canopy.

Accuracy is defined as the closeness between a value v and a value v' considered as the correct representation of the phenomenon that v aims to represent.; eg: Postcode "4109" is typed "4019".; Completeness is defined as the sufficiency of data for the task at hand.; eg: Students don't have to declare a major till graduation, so major is missing in most enrolments.; Currency is data that has not been updated on time and is obsolete.; eg: Old phone numbers.; Consistency is that refers to when two values that are supposed to represent idea/value are represented in different ways.; eg: OLAP Vs Online Analytical Processing.; Accessibility: Server down, privacy concerns.

K-anonymity is a key concept that was introduced to address the risk of reidentification of anonymised data through linkage to other datasets.

Individual values of attributes are replaced with a broader category. For example, the value "19" of the attribute "age" can be replaced by "under 20".

After processing the data with the k-anonymity technique, each record from the adversary's knowledge corresponds to at least k records of the table containing sensitive information. However, the sensitive values of the k records can be identical. In such a case, the sensitive value can still be predicted.

We do not publish information that does highly depend on any particular individual record.; We can introduce some randomness/noises to the data, which does not affect the data utility while giving each individual refutability.; We may use a Laplace Distribution to introduce the noises.

L-diversity introduces some intra-group diversity to the k records containing sensitive information. L-diversity specifies that the sensitive column of the k records must have at least l different values, whereas k-anonymity specifies that each record known by the adversary can linked to at k records of the table containing sensitive information.

Q: What kind of values can be used as the key in key-values storage?; A: Values that must be globally unique. Values that must not be empty

Q: Big Data is of Three Utilities:; A: Connecting Dots From small to big; Discovering Specifics From big to small"; Data Inferencing Knowing unknown"

Q: Operations of a schedule are in conflict if they satisfy all the following three conditions:; A: they belong to different transactions; they access the same item X; at least one is a write item (X)

Schema mapping: mapping of structures; Data mapping: matching based on content; Data fusion: reconciliation of mismatching content; Schema mapping: e.g. Name = Title; Domain mapping: e.g. Integer = String; Value mapping: e.g. 'UK' = 'United Kingdom'

Healthcare：Treating patients require utmost care as well as information.; Retail：Brick and mortar stores and online retailers deal with tons of data.; Finance: Banks have started integrating data, which is allowing them to determine, eradicate, and prevent instances of fraud.; Marketing: Managing information on potentially millions of customers is impossible without proper integration channels and tools for data integration.

Model inversion attack; Statistical attack; Poisoning attack: Evasion attack; Model poisoning attack.