① a) Explain about Determining a Consistent view of the Network.

Consistency: Consistency is the agreement between multiple nodes in a distributed system to achieve a certain value.

» It can be divided into strong consistency and weak consistency.

Strong Consistency: The data in all nodes is the same at any time. At the same time, you should get the value of key1 in node A and the value of key1 in node B.

Weak Consistency: There is no guarantee that all nodes have the same data at any time, and there are many different implementations. The most widely achieved is the ultimate consistency. The so-called final consistency means that the same data on any node is the same at any time, ~~but~~ but as time passes, the same data on different nodes always changes in the direction of convergence. It can also be simple understood that after a period of time, the data b/w nodes will eventually reach a consistent state.

Distributed and consistent application scenarios:

Multi-node provides read & write services to ensure high availability and scalability.

Problems faced by distributed systems

» Message asynchronous: The real network is not a reliable channe there are message delays, loss, and inter-node messaging cannot be synchronized.

» node-fail-stop: The node continues to crash and will not recover.

» node down recovery: Node recovery after a period of time, the most common in distributed systems.

* The basic theory of distributed systems, the FLP Theorem, is when only the node is down, the availability and strong consistenc cannot be satisfied at the same time. Another point of view is CAP theory, namely strong consistency, availability and partition fault tolerance, only two of which can be guaranteed.

There are many agreements to ensure consistency, including 2PC.

2PC : 2-stage lock commit protocol to guarantee the atomicity of operations on multiple data slices.

The nodes are divided into coordinators and participants, and the execution is divided into two phases.

Phase 1 : The coordinator initiates a proposal to ask whether each participant is accepted. Participate performs transaction operations, writes undo & redo information to the transaction log, and replies yes or no to the coordinator.

Phase 2 : The coordinator submits or aborts the transaction according to the feedback of the participant. If the participant is all yes, it is submitted, as long as there is a participant reply no. Participate formally commits or terminates the transaction according to the commit/ rollback information of coordinator, and releases the occupied resources and returns ack.

b) Explain briefly about Authorization and Protection.

A distributed system needs additional security measures than centralized system, since there are many users, diversified data, multiple sites and distributed control.

Security measures encompass security in communications, security in data and data auditing.

Communications Security

In a distributed database, a lot of data communication takes place owing to the diversified location of data, users and trasact-ions. So, it demands secure communication b/w users and databases and b/w the different database environments.

Security in communication encompasses the following -

» Data should not be corrupt during transfer.

» The communication channel should be protected against both passive ~~eva~~ eavesdroppers and active attackers.

» In order to achieve the above stated requirements, well-defined security algorithms and protocols should be adopted.

Data Security

In distributed systems, it is imperative to adopt measure to secure data apart from communications. The data security measures are –

Authentication and authorization – These are the access control measures adopted to ensure that only authentic users can use the database. To provide authentication digital certificates are used. Besides, login is restricted through username/password combination.

Data encryption – The two approaches for data encryption in distributed systems are –

» Internal to distributed database approach:

The user applications encrypt the data and then store the encrypted the data and in the database.

» External to distributed database:

The distributed database system has its own encryption capabilities The user applications store data and retrieve them without realizi that the data is stored in an encrypted form in the database.

Data Auditing

A database security system needs to detect and monitor security viola violations, in order to ascertain the security measures it should adopt. It is often very difficult to detect breach of security at the time of occurences. One method to identify security violations is examine audit logs. Audit logs contain information such as –

» Date, time and site of failed access attempts

» Details of successful access attempts

» vital modifications in the database system.

All the above information gives an insight of the activities in th database. A periodical analysis of the log helps to identify any unnatural activity along with its site and time of occurence. This log is ideally stored in a separate server so that it is inaccessible to attackers.

② Explain object query processing in detail.

Query processing in a distributed DBMS requires the transmission of data b/w the computers in a network. A distribution strategy for a query is the ordering of data transmissions and local data processing in a database system. Generally, a query in distributed DBMS requires data from multiple sites, and this need for data from different sites is called the transmission of data that causes communication costs. Query processing in DBMS is different from query processing in centralized DBMS due to the communication cost of data transfer over the network. The transmission cost is low when sites are connected through high-speed Networks and is quite significant in other networks.

1) Costs of Distributed Query processing:

In Distributed query processing, the data transfer cost of transferring intermediate files to other sites for processing and therefore the cost of transferring the ultimate result files to the location where that results required. Let's say that a user sends a query to site S1, which requires data from its own and also from another site S2. Now, there are three strategies to process this query which are given below.

1) We can transfer the data from S2 to S1, and then process query.
2) We can transfer the data from S1 to S2, and then process query
3) We can transfer the data from S1 and S2 to S3 and then process the query. So the choice depends on various factors like, the size of relations and the results, the communication cost between different sites, and at which the site result will be utilized.

Commonly, the data transfer cost is calculated in terms of the size of the messages. By using the below formula, we can calculate the data transfer cost:

Data transfer cost = C * size

where C refers to the cost per byte of data transferring and size R the no. of bytes transmitted.

Ex: Consider the following table EMPLOYEE and DEPARTMENT

Site 1 : EMPLOYEE

EID	NAME	SALARY	DID

EID - 10 bytes
SALARY - 20 bytes
DID - 10 bytes
Name - 20 bytes
Total records - 1000
Record Size - 60 bytes

Site 2 : DEPARTMENT

DID	DNAME

DID - 10 bytes
Dname - 20 bytes
Total records - 50
Record Size - 30 bytes

Ex: Find the name of employees and their department names. Also, find the amount of data transfer to execute this query when the query is submitted to Site 3.

Ans: Considering the query is submitted at site 3 and neither of the two relations is an EMPLOYEE and the DEPARTMENT not available at site 3. So, to execute this query, we have three strategies:

1) Transfer both the tables that are EMPLOYEE and DEPARTMENT at site 3 then join the tables there. The total cost in this is,

$1000 * 60 + 50 * 30 = 60,000 + 1500 = 61500$ bytes.

2) Transfer the table EMPLOYEE to site 2, join the table at site 2 and then transfer the result at site 3. The total cost in this is, $60 * 1000 + 60 * 1000 = 120000$ bytes since we have to transfer 1000 tuples having NAME and DNAME from site 1.

3) Transfer the table DEPARTMENT to site 1, join the table at site 2 join the table at site 1 and then transfer the result at site 3. The total cost is $30 * 50 + 60 * 1000 = 61500$ bytes since we have to transfer 1000 tuples having NAME and DNAME from site 1 to site 3 which is 60 bytes each.

(3) What is (a) Schema translation (b) Schema Integration

(a) Schema Translation

A database system consists of three components: schema, data and programs.

Database reengineering starts with the schema, which defines the meaning of data and their relationship in different models.

Only after a schema has been redefined can data and programs be reengineered into a new database system, which makes use of the translated schema.

Schema translation is the process of changing a schema expressed in one data model into an equivalent schema expressed in a different data model.

(b) Schema Integration.

Schema integration is used to merge two or more database schemas into a single schema that can store data from both the original databases. For large databases, with many expected users and applications, the integration approach of designing individual schema and then merging them can be used. Because the individu views can be kept relatively small and simple.

Schema Integration is divided into the following subtask.

1) Identifying correspondences and conflicts among the schema :-

As the schemas are designed individually it is necessary to specify constructs in the schemas that represent the same real-word concept. We must identify these correspondences before proceeding with the integration. During this process, several types of conflicts may occur such as:

(1) Naming conflict – Naming concepts are of two types synonyms and homonyms. A synonym occurs when two schemas use differen names to describe the same concept. A homonym occurs when two schemas use the same name to describe different concept

(2) Type conflicts – A similar concept may be represented in two schemas by different modeling constructs.

(3) Domain conflicts – A single attribute may have different domains in different schemas.

(4) Conflicts among constraints – Two schemas may impose different constraints

2) Modifying views to conform to one another :

Some schemas are modified so that they conform to other schemas more closely. Some of the conflicts that may occur during the first steps are resolved in this step.

3) Merging of views and Restructuring :

The global schemas are created by merging the individual schemas Corresponding concepts are represented only once in the global schema and mapping b/w the views and the global schemas are specified. This is the hardest step to achieve in real-world databases which involve hundreds of entities and relations. It involves a considerable amount of human intervention and negotiation to resolve conflicts and to settle on the most reasonable and acceptable solution for a global schema.

Restructuring as a final optional step the global schemas may be analyzed and restructured to remove any redundancies or unnecessary complexity.