Part 1: Distributed Databases (12 Marks）

Q1: Consider relation S(A, B, C, D), where A is the primary key attribute. S is vertically fragmented into S1(A, B) and S2(A, C, D) and allocated at site N1 and N2 respectively. How to insert a tuple (a, b, c, d) into S? This insert operation must meet the atomicity property.

A1:
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

Q2: What is data replication? What are the benefits of having data replications, at what costs?

A2:
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

Q3: Consider two relations R(A, B) and S(X, A, C), where S.A is the foreign key. Assume that R is horizontally fragmented based on its attribute A into R1 and R2. Please use the semijoin operation to define the derived horizontal fragmentation of S based on the fragmentation of R, and explain how your S fragmentation meets the reconstruction property.

A3: 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.

Q4: Assume that the relation R(A, B) is located on site 1 and that the relation S(X, A, C) is located on site 2. Consider a join query R ⋈A S at site 1. Please give a step-by-step query execution plan using semijoin operations to process this query.

A4: 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

Q5: When performing fragmentation on a relation, what properties should the fragmentation meet to ensure the correctness? List their names and meanings.

A5: 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

Q6: For any join query in a distributed database, does a semijoin always have less or equal join cost than a traditional join? Simply answer "Yes" or "No".

A6: No.
When the member is greater than the owner, semi-join has a greater join cost than an inter-join

Q7: Given a relation R(A, B) and two simple predicates (P1: R.A ≥ 75, P2: R.A < 50), generate a set of minterm predicates which satisfy the aforementioned properties.

A7:
m1 = P1 and P2 = empty
m2 = not P1 and P2 = R.A < 50
m3 = P1 and not P2 = R.A >= 75
m4 = not P1 and not P2 = 50 <= R.A < 75

Part 2: Distributed Transaction Management (8 Marks)

Q1: Voting-based approach is often used in a distributed system to maintain data consistency among data replications. Please explain how such a technique works to manage read and write operations on a data object that has N copies.

A1: 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.

Q2: If a database is read-intensive with rare updates, should we use a large number of write copies in the voting-based approach? Why or why not?

A2: 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.

Q3: List four types of replication strategies and point out whether they are synchronous replication or asynchronous replication.

A3: Synchronous Replication: 1. Voting 2. Read any Write all
Asynchronous Replication: 3. Primary Sit 4. Peer-to-Peer replication

Q4: What does ACID stand for?

A4: Atomicity; Consistency; Isolation; Durability.

Part 3: Data Warehouse Design (9 Marks)

Q1: A data warehouse can often make use of materialized views (e.g., using materialized data cubes). Discuss advantages and disadvantages of building materialized views in data warehouses.

A1:
Advantages of Data Warehouse (DWH):
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

Q2: It is not common for data warehousing systems to support update operations. List two reasons why supporting updates in data warehouses is not a good idea.

Briefly justify your answer.

A2:
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.

Q3: What are the main differences between an operational database and a data warehouse?

A3:
Operational systems focus on Data.
Data warehousing systems focuses on Information out.

It is used for Online Transactional Processing.
It is used for Online Analytical Processing.

Q4: The data warehouse can be modelled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences between the two models, and then analyse their advantages compared to one another.

A4:
1. 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.

2. Star schema results in high data redundancy and duplication. Snowflake schema ensures a very low level of data redundancy (because data is normalized).

Part 4: Data Warehouse Implementation (5 Marks)

Q1: Compare and contrast the following two data warehousing operations: SLICE and DICE.

A1: 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.

Q2: 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.

A2:
Advantages:
small size
speed up query.

Disadvantages:
Not suitable for attribute with high cardinality.
Not suitable for attributes that are updated frequently.

Q3: Explain what a data cube is in data warehousing systems.

A3: A data cube in a data warehouse is a multidimensional structure used to represent data along with some measure of interest.

Part 5: Database Integration and Data Linkage (10 Marks)

Q1: What is entity resolution?

A1: Entity resolution (ER) is the task of disambiguating records that correspond to real-world entities across and within datasets.

Q2:Edit distance and Jaccard coefficient are two common string similarity measures used in entity resolution. Please define these two measures.

A2: 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”)

Q3: What is semi-join?

A3: Semi join is a technique for processing a join between two tables that are stored sites.

Q4: Provide at least four examples of scenarios in which data integration is needed.

A4:
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.

Q5:
(a)List at least four possible challenges in data integration, and give one example of each challenge.
(b)List at least four challenges we need to address in data integration, and give one example for each challenge.

A5:
Schema heterogeneity:
S1: Employee (ID, name, address, position, salary from, until)
S2: Worker (EID name, address); Position (EID, PID, salary, from, until)

Data type heterogeneity:
Employee ID could be a string or an integer

Value heterogeneity:
The "cashier" position could be called "associate" in another system

Semantic heterogeneity:
Salary is hourly salary or is weekly salary with allowances

Q6: Record linkage is an important task in data quality management. Explain the meaning of record linkage

A6: Data linkage is an operation to identify records referring to the same real-world entity.

Q7: Why is data linkage so difficult in practice?

A7: The same real-world object can be represented as different strings.
The same string can represent different real-world objects.

Q8: In Edit distance, what operation is considered as an 'edit'/'transformation'?

A8: Operations: delete, insert or substitute one character

Q9: Discuss different roles that views play in the following systems. (You should give at least one type of use for each system)

A9:
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.

Q10: Efficiency of record linkage should also be considered in practice. Various techniques have been proposed to reduce the number of record comparisons, such as Blocking, Sorted Neighbourhood Approach, Clustering and Canopies, etc. Please explain one of these techniques.

A10: Canopies
Step 1: Data is divided into overlapping subsets, called canopies.
Step 2: Expensive distance measurement made among points within the same canopy.

Part 6: Data Quality Management (6 Marks)

Q1: Data quality can be measured from various dimensions. Please explain the meaning of the following data quality dimensions respectively, and give one example of data quality problems for each of these dimensions.

A1:

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: ITEE Vs Information Technology and Electrical Engineering.

Accessibility: Server down, privacy concerns.

Part 7: Data Privacy (4 Marks)

Q1: What is the K-anonymity?

A1: K-anonymity is a key concept that was introduced to address the risk of reidentification of anonymised data through linkage to other datasets.

Q2: Describe the general approach of K-anonymity.

A2: 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".

Q3: K-anonymity is still vulnerable in some situations. Explain possible problems of K-anonymity.

A3: 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.

Q4: Differential privacy.

A4:

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.

Q5: L-diversity is a method to reduce the vulnerability of K-anonymity. Describe the general approach of L-diversity, especially its difference with K-anonymity.

A5: 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.

Part 8: Advanced Topics (6 Marks)

Q1: What kind of values can be used as the key in key-values storage?

A1: Values that must be globally unique. Values that must not be empty

Q2: What is the "curse of dimensionality"? What is the impact of dimensionality to the data processing?

A2: Adding extra dimensions to a data space will exponentially increase the volume of data space. For a given set of data points, adding extra dimensions makes the data distribution becoming sparser, therefore it is more difficult to find regular patterns in processing data.

SQL:

(1):

```
CREATE VIEW MedicalVisit(personID, Name, DateOfBirth, VisitDate, VisitType) AS
    SELECT a.ID, a.Name, a.DateOfBirth, g.Date, "GP"
    FROM Visit g, Patient a
    WHERE g.Medicare# = a.Medicare#;
UNION
    SELECT a.ID, a.Name, a.DateOfBirth, h.InDate, "Hospital"
    FROM Admission h, Patient a
```

```
    WHERE h.pID = a.ID;
```

(2):

```
SELECT item, time, location, sum(sales)
FROM AllElectronics
GROUP BY CUBE(item, time, location);
```

(3):

```
SELECT SUM(sale)
FROM Table
WHERE product = 'Dell Laptop'
GROUP BY location;
```

(4):

```
SELECT avg(grade)
FROM grade
WHERE course = 'Advanced Database Systems'
GROUP BY student;
```

(5):

```
SELECT venue, sport
FROM FactTable
GROUP BY Venue,
    (SELECT p.Sport
    FROM Event e, Event_Class c, Subsport s, Sport p
    WHERE e.Event_Class = c.Event_Class AND c.Subsport = s.Subsport AND s.Sport
= p.Sport);
```