UNIT 01:

1. Consider a data warehouse with three dimensions date, vehicle, and road.
   There is one measure, toll, which is the money that the driver has to pay for using a particular road : =
   - Draw a Star schema assuming concept hierarchy for each dimension.
   - In your star schema, identify different types of dimension table.
   - Staring with the base cube and finest granularity [date, vehicle, road]. Which sequence of OLAP operations do you need to list the total toll collected on each road in the year 2022 ?
2. Explain schema integration and instance integration with the help of examples.
3. What is CUBE ? If we create CUBE for sales application with three dimension for time, location and item, illustrate with example how sub cubes in lattice can be created.
4. Consider the star schema of an automobile data warehouse :
   Autos (ModelId, modelname, serialNo, color)
   Dealers (DealerId, name, city state, phone)
   Time (TimeId, day, week, month, year)
   Sales (ModelId, DealerId, TimeId, QtySold, CountSold
   Where the attribute QtySold is intended to be the total price of all automobiles for the given model, color, date and dealer, while CountSold is the total number of automobiles in that category. Answer the following OLAP queries :—
   a. Find total sales generated for model name (Maruti, Honda) and dealer state (Maharashtra, Gujarat) in September 2017 and October 2017 using ROLL – UP across three dimensions – ModelId, DealerId and TimeId.
   b. Find total sales generated for model name (Maruti, Honda) and dealer state (Maharashtra, Gujarat) in September 2017 and October 2017 using CUBE across the dimensions – ModelId, DealerId and TimeId.
   c. Comment on difference in output using ROLL – UP and CUBE aggregation clause
5. What do you mean by ETL Process ? What is the purpose of 'refresh' in ETL process ?
6. Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.
   (i) Draw a Star schema for the above scenario.
   (ii) Write an SQL query assuming the data is stored in a relational database with the schema
   Fee (day, month, year, doctor, hospital, patient, count, charge).
   List the total Fee collected by each doctor in 2020?
7. A data cube, C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions

- What is the maximum number of cells possible in the base cuboid ?
- What is the minimum number of cells possible in the base cuboid ?
- What is the maximum number of cells possible (including both base cells and aggregate cells) in the C data cube ?
- What is the minimum number of cells possible in C ?

8. The data warehouse for wholesale furniture company has to allow analyzing the company's situation at least with respect to the Furniture, Customers and Time. Moreover, the company needs to analyze.
    - The furniture with respect to its type (chair, table, wardrobe, cabinet...), category (kitchen, living room, bedroom, bathroom, office...) and material (wood, marble...)
    - The furniture with respect to its type (chair, table, wardrobe, cabinet...), category (kitchen, living room, bedroom, bathroom, office...) and material (wood, marble...)

    The company is interested in learning at least the quantity, income and discount of its sales

    - Identify Facts, Dimensions and Measures.
    - Draw STAR and Snowflake Schema for the above Scenario.
    - Write SQL queries for the following
        - Find the quantity, the total income and discount with respect to each city, type of furniture and the month.
        - Find the average quantity, income and discount with respect to each country, furniture material and year.
9. Explain the Datawarehouse architecture and comment on importance of metadata repository.
10. A data warehouse for Hotel consists of three dimensions–hotel, customer, reservation and two measures amount_paid and customer_count Analyze these dimensions and list the possible attribute for each dimension tables. Also designate a primary key for each table. Construct snow flake schema for the above scenario. Make suitable assumptions. List the concept hierarchies.
11. What is factless fact table ? Design Star schema with factless fact table to track a patient by diagnostic procedure and time.
12. A data warehouse is subject oriented. Identify major critical business subjects for the following companies ?
    a. An international manufacturing company.
    b. A Hospital.
    c. A domestic hotel chain.
13. Describe data warehouse development life cycle with neat sketch.
14. What is CUBE? If we create CUBE for a retail application with three dimensions for time, product, and store, illustrate with an example how the subcubes in the lattice can be created.
15. Explain the following data warehouse models:

(a) Enterprise warehouse.
(b) Data Mart.
(c) Virtual warehouse.

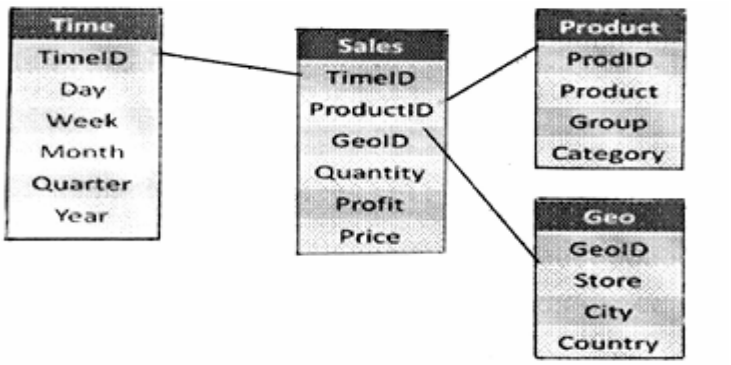
16. In a STAR schema to track the shipment for a distribution company, the following dimension tables are found :
    (i) Time,
    (ii) Customer ship–to,
    (iii) Ship–from,
    (iv) Product,
    (v) Type of deal, and,
17. Explain the difference between the following two SQL queries

```
SELECT gender, semester_year AS year, semester_month AS month,
SUM(num_of_student)AS total
 FROM instructor_summary
GROUP BY (gender,semester_year, semester_month);
SELECT gender, semester_year AS year, semester_month AS month,
SUM(num_of_students)AS total
FROM instructor_summary
GROUP BY ROLLUP(gender, semester_year, semester_month);
```

    - What changes when you use CUBE operator instead of ROLLUP ?
    - Rewrite the query to compute the following three result groups: (gender, Semester year), (semester month) and ().
    - What is the result of the following query

```
SELECT semester_year AS year, campus,
SUM(num_of_classes)AS num_of_classes
FROM instructor_summary
GROUP BY CUBE (semester_year, campus)
ORDER BY 1;
```

18. Why is metadata important in a data warehouse ? Explain the different components of metadata repository
19. Consider the following diagram



What kind of a schema is presented ? Considering that the product dimension is subject to change often, how would you transform the schema to accommodate this ? Draw the new schema.

20. Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need? Can such patterns be generated alternatively by data query processing or simple statistical analysis?
21. Write a short note on data warehouse development life cycle.
22. Explain the use of cube and rollup operators with an example.
23. What is a data warehouse? With the help of a neat sketch, explain the various components in a data warehousing system.
24. Differentiate between the following :
    a. ROLAP and MOLAP.
    b. Snowflake and fact constellation schema

UNIT 02:

1. A data cube, C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions : = What is the maximum number of cells possible in the base cuboid ? = What is the minimum number of cells possible in the base cuboid ? = What is the maximum number of cells possible (including both base cells and aggregate cells) in the C data cube ? What is the minimum number of cells possible in Cube.
2. Consider the following schema :
   SALES (time_key, item_key, branch_key, location_key, units_sold, dollars_sold)
   BRANCH (branch_key, branch_name, branch_type)
   LOCATION (location_key, street, city, state, country)
   ITEM (item_key, item_name, brand, type, supplier_type)
   TIME (time_key, day, day_of_week, month, quarter, year)
   - Identify the concept hierarchies in the given cube. Clearly state your assumptions.
   - Construct the following queries using advanced SQL :–
     - roll-up on total sales by year and by quarter.
     - roll-up on total sales by item brand and by item type (digital or analog).
     - drill down on total sales by month and by day.
     - drill down on total sales by street address.
3. Suppose two stocks infosys and TCS have the following values in one week : (3, 6), (4, 9), (6, 11), (5, 12), (7, 15). If the stocks are affected by the same industry trends, will their prices rise or fall together ?
4. The Restaurants 'SR' wholesale restaurant company supplies equipment to 55 different restaurants in Mumbai, such as tables, chairs, table cloths, napkin holders, cutlery and so on, as well as kitchen equipment such as saucepans, knives and chef clothing. They wish to analyze their daily sales in terms of revenue, unit sales, costs and profit for each product and customer. They also would like to know this information by product line and product group

- Design a STAR schema according to the given scenario.
- Convert STAR schema into Snowflake Schema.

Bring out the difference between STAR and Snowflake Schema

5. Given is the frequency of stop words in documents (The values are given in increasing order) :
   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Apply the following methods and show the results:—
   (i) Use smoothing by bin means with a depth of 3.
   (ii) Use Min – Max normalization to transform the value 30 into the range 0.0 to 1.0
   (iii) Use z – score normalization to transform the value 30 where the standard deviation of the above frequency is 12.94.
   (iv) Use normalization by decimal scaling to transform the value 30.
   (v) Plot an equi – width histogram of width 10 on graph paper.
6. Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.
7. Differentiate between the following :
   - ROLAP and MOLAP.
   - Snowflake and fact constellation schema.
8. Explain the computation of measures in a data cube.
   a. Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube.
   b. For a data cube with the three dimensions time, location and item, which category does the function variance belong to ? Describe how to compute it if the cube is partitioned into many chunks.
   c. Suppose the function "top 10 sales". Discuss how to efficiently compute this measure in data cube.
9. In data warehouse technology a multiple dimensional cube can be implemented either by a multi–dimensional database technique (MOLAP) or by a relational database technique (ROLAP). Briefly describe each implementation technique.
10. What is cube materialization ? Discuss its different types.
11. Explain various types of multidimensional models.
12. What are dimension hierarchies ? Give three examples.
13. Consider a data warehouse with three dimensions date, vehicle, road. There is one measure, toll, which is the money that the driver has to pay for using a particular road.
    a. Draw a simple star schema assuming some concept hierarchy for each dimension.
    b. Starting with the base cube and finest granularity [date, vehicle, road] which sequence of OLAP operations do you need to list the total toll collected on each road in the year 2011?
14. A data cube C, has n dimensions and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.

a. What is the maximum number of cells possible in the base cuboid ?
b. What is the minimum number of cells possible in the base cuboid ?
c. What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C ?
d. What is the minimum number of cells possible in the data cube C ?

15. What is the difference between slowly changing and rapidly changing dimensions? Give an example of each
16. Given are the fact table PropertySale (branchNo, PropertyType, YearMonth SaleAmount) and dimension table Branch (branchNo, city), Write SQL statement to answer the following query.
    "Retrieve total amount of sale in January and February 2007 in Manchester, Edinburgh, and Birmingham, with subtotals for each property type, month, and city (including all cross-tabular subtotals)" .
    - Write sample queries for : partial rollup, grouping set assuming suitable data.
    - What is the use of grouping function.
17. A data warehouse of a train company contains information about train segments. It consists of six dimensions, namely, departure station, arrival station, trip, train, arrival time, and departure time and three measures, namely, number of passengers, duration and number of kilometers. Write OLAP operations to be performed in order to answer the following queries. Propose the dimension hierarchies whenever needed.
    a. Total number of kilometers made by Alstom trains during 2012 departing from French or Belgian stations.
    b. Total duration of international trips during 2012, that is, trips departing from a station located in one country and arriving at a station located in another country.
18. You are data design specialist on the data warehouse project team for a manufacturing company. Design a STAR schema to track the production quantities. Production quantities are normally analyzed along the business dimensions of product, time, parts used, production facility and production run. State your assumptions and mention the concept hierarchies.

UNIT 03:

1. Create Index Organized Table (IOT) named Customer with attributes Cust_No, Cust_Name, Cust_Email and Cust_address. Cust_Email and Cust_address should be in Overflow block.
   If a secondary index is created on Cust_Name, List the contents of index clearly
2. Assuming suitable schema for EMP and DEPT tables, explain how the bitmap join index works. Also write the SQL command to create bitmap join index.
3. Explain query optimization with respect to data warehousing.
4. Explain hash index and bitmap index with an example.
5. Write a short note on load manager.
6. Write SQL command to create Index Organized Table Employee with the attributes empno, empname and salary in tablespace tsa as directed
   a. Empno is primary key for the table.

b. PCTTHRESHOLD is 20.
c. Specify Overflow and Including clause. Assume empname to be included in Including clause.
d. Give meaning of PCTTHRESHOLD, including and overflow clause. Mention advantages of IOT over B – tree indexes

7. State the advantages of data partitioning in data – warehouse. Write a SQL query to create composite List – Range partitioning for the following scenario :
Customer table having attributes cust_id, cust_name, cust_state and time_id. Perform list partitioning on state attributes and range partitioning on time –id.
Partition definitions for list are as below :
   - Partition East should accept values ('WB', 'JK')
   - Partition South should accept values ('TN', 'AP')
   - Partition North should accept values ('UP', 'HP')
   - Partition Temp should accept any other state.

   Partition definitions for range are as below for the year 2020 :

   - Partition P1 should accept values for Jan, Feb, March, April.
   - Partition P2 should accept values for May, June, July, August.
   - Partition P3 should accept values for September, October, November, and December.

8. Suppose a student collected the price and weight of 20 products in shop with the following result

Price	5·89	149	59·98	129	15·89	56·99	35·75	42·19	31	125·5
Weight	1·4	1·5	2·2	2·7	3·2	3·9	4·1	4·1	4·6	4·8
Price	4·5	22	52·9	61	33	328	122	142·19	229	89·4
Weight	4·9	5·1	5·5	5·8	5·8	8·9	9·6	18·0	36·9	38·2

   a. Give 5 number summary for price.
   b. Draw boxplot for price. Identify outliers, if any.
   c. Draw scatter plot based on these two variables.
   d. Calculate the Pearson correlation coefficient.
   e. Are these two variables positively or negatively correlated?

9. a. Explain the need to create function-based indexes. Write a command to create a function-based index on emp-name column of EMPLOYEE table.
   b. Explain query optimizer with respect to data warehousing.
   c. State the advantage of partitioning in data-warehouse. 3(CO1) Write a query to create composite List-Range partitioning for the following scenario :
   - Supplier table having attributes sup_id, sup_name, sup_state and time_id
     - Perform list partitioning on state attributes and range partitioning on time-id.
     - Partition definitions for list are as below
       - Partition East should accept values ('WB', 'JK').

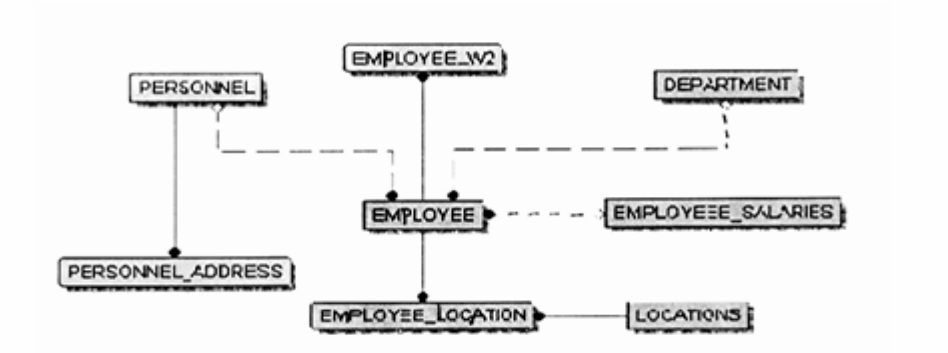
- Partition South should accept values ('TN', 'AP').
- Partition North should accept values ('UP', 'HP').
- Partition Temp should accept any other state.

○ Partition definitions for range are as below :

- Partition P1 should accept values less than 01-Jan.-2018.
- Partition P2 should accept values less than 01-April-2018.
- Partition P3 should accept values less than 01-July-2018.

Write query to access data from partition and subpartition.

10. State the advantage of storing partitions in different tablespaces ? Write a query to create range partitioned table for the following scenario:
    - Create a table named–Purchase consisting of four partitions, one for each quarter of Purchase for the year 2016. The column Purchase_Year is the partitioning column, while its values constitute the partitioning key of a specific row.
    - The other columns for table must be prod_id, cust_id, promo_id, quantity_purchased, amount_Purchased–all in number format and Year
    - Store each partition in different tablespaces.
    - Write a query to insert row in partition.
    - Write a query to merge partition 3 and 4.
11. Illustrate Index Organized Table and function based indexes with suitable examples.
12. Constrast beween Cost based and Rule based optimizer. Describe query optimization technique with materialized views in data warehouse. Take suitable example for illustration.
13. Consider the following model :



If PERSONNEL and PERSONNEL_ADDRESS are to be placed in a cluster and EMPLOYEE and EMPLOYEE_W2 are to be placed in another cluster, write commands to create these cluster and tables.

14. Explain the reason for error after the following SQL staement are executed:
    SQL>create table test (coll number, col2 varchar2(20));
    SQL>create index idx1 on test(coll);
    SQL>drop table test; SQL>drop index idx1;

Update col2 such that it stores values in allcaps. Write command to create a function based index on col2.

15. Consider a table called YEARLY_SALES with attributes (sales_month INTEGER, state VARCHAR2(2), sales_amount NUMBER).
    - Write the command to partition this table using range partitioning based on sales_month. Each partition is placed in a different tablespace.
    - Write the command to partition this table using list partitioning based on state. Each partition is placed in a different tablespace.
16. How are Index Oragnized Tables(IOT) different from B-Tree indexes ? Give syntax for creating a IOT with overflow area. Write SQL command to know which columns are in the overflow segment.
17. How is data stored in a hash cluster ? How is it retrieved ? Explain the purpose of the following clauses in the create cluster command :
    CLUSTER_KEY<datatype> , SIZE <size number., SINGLE TABLE, HASHKEYS<hash_key_number> , HASH IS <expr>.
18. State what is a Bitmap Join Index. List advantages of creating a bitmap join index over normal joins. Give the command for creating a bitmap join index.