ADBMS Viva

Parallel database

The parallel database is designed to take advantages of a such architectures by running multiple instances which is share a single physical database

Types:

1.share memory:- system has the large amount of cashe memory at each proccesser so the referencing share memory is avoided

2.share disk:- each proccesser has own share memory so the data sharing is efficient

3.share nothing :- each proccesser has own local memory and local disk

Distributed database architecture types

Database that not limited to only one system

1.client server system :- has no of clients and server connected on net

2.collaborating server system :- single query is breaks on the multiple small queries and run on multiple server

3.middle ware server system :- uses local server to runs the local queris.

Types of distributed database

1.homogenous: sites uses similar software so its easy to handle

    1.autonomous 2.non autonomous

2.heterogenous : sites uses dissimilar software so the transcation processes is complex

    1.federated 2.non federated

Replication

copy the several copies of relation of the fragment

1.synchonous : write data to both primary and secondary area at the same time in that time the data remains current and identical

2.asynchronous : write data to both primary and secondary area at the same time but in that process there is delay to copy the data one location to another

OODBMS	ORDBMS
Data store in files,documents,graph	Data store in table
Does not support to distributed database	support to distributed database
Xml file system	Mysql , sql , oracle

Datawarehouse

Where we store the data and integrated In diff department

Steps: extracting data , cleans data , info save in individsual repositories, makes a reports

ETL

1.extraction : extracting from diff sources and store in staging area

2.tranformation : transformed data in standard format

3.loading : transformed data load in data warehouse

OLAP	OLTD
Online analytic processing	Online transcational processing
Uses historical data	Uses current data

Data mart

Access layer of the data warehouse it is provides business user with relavant data . uses for particular department

1.dependent data mart:- Extract data from central data warehouse

2.independent data mart :- created without the data warehouse

Rolap	**Molap**
Relational online analystic proccessing	Multi dimensional online analytic proccessing
Data store in relational tables	Data store in multi dimensional array

ADT (abstract data type)

It is mathematical model and it is set of objects and operations such as insert ,search , delete

1.list adt :- sorted data in key sequence

Get(),insert(),remove(),removeall(),isempty(),isfull()

2.stack adt:-allocate memory for store the data

Push(),pop(),peek(),isempty(),isfull()

3.queue adt :- follows basic design of stack adt

Enqueu(),dequeue(),peek(),isempty(),isfull()

Data mining

Is extracting meaningful data , data cleaning ,data transformation ,represent the pattern

Applications :

Customer segmentation , fraud detection , market basket analysis , text mining , web minig ,risk management

KDD(knowledge discovery from database)

Step:

Selection , preprocessing ,transformation ,data minig, pattern evolution ,knowledge presentation

Data preprocessing

Remove the inconsistent data , missing values and noisy data

Steps:

1.data cleaning

2.data integration

3.data transformation

4.data reduction

Data cleaning

Cleaning the data by filling the missing values and correct the inconsistent data and also remove redundancy

Using binning ,regression ,clustering

Data reduction

Use when data warehouse store the terabytes of data

Techniques:

1.data sampling

2.dimensionally reduction

3.data compression

4.feature selection

K-Nearest Neighbour

It is versatile algorithm that Use in classification to Finding the k closest relatives. Handle the numerical , categorical data

Regression

Provide the value of dependent variable from the independent variable

Linear regression

Linear regression is a supervised learning algorithm that compares input (X) and output (Y) variables based on labeled data. It's used for finding the relationship between the two variables and predicting future results based on past relationships.

Types :

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.

Multiple Linear Regression

This involves more than one independent variable and one dependent variable.

Decision tree

It is use to perform the classification and regression

ID3

Interactive dichotomister 3

It is smallest decision tree with the top-down approach. it is predecessor of c45 algorithm

C45

Improvement of id3 . it is handle the continous and missing attributes .use to build the decision tree

CART

Suitable for classification and regression and handles numerical and categorical data

Clustering algorithm application

1.data mining

2.image segmentation

3.speech recognization.

Hierachical Clustering

Use to group the unlabled dataset into cluster.in that develop dendrogram

Use t identify fake news and criminal activity

Types:

1.Agglomerative clustering

It is bottom-up approach , merge the single objects and join the cluster stepwise

2. divise clustering

It is top-down approach ,divide the cluster into single object ,separating stepwise

Text retrieval method

Transforming the unstructured text into the structure format

Methods:

1.document selection

Uses Boolean retrieval model and It is provides Boolean expressions of keywords

2.document ranking

It is uses vector space model and rank all the records, included the algebra ,probability , logic

Web mining

Use to extracting the content from the web ,useful for ecommerce and e services

Types

1.web content mining : extracting the content from web like image, text ,list ,audio , video

2.web structure mining : extracting the structure of documents like

Hyperlink ,document structure

3.web usage mining : extracting the useful patterns (web server logs, application server logs)

Applications : ecommerce , fraud detection ,customer service, healthcare

Naïve Bayesian

Used to determined the conditional probability of event A when event B is already happens and the simplified version of the bays theorm known as naïve bays theorm

Why clustering is unsupervised

In that the data is unlabeled so it call unsupervised

K-means

It is unsupervised algorithm that use to solve the clustering problems.it is identifies centroid

Example : partition model

Olap operation

1.rollup

2.drill down

3.dice

4.slice

4.pivot

Range Partitioning:

In this method, Tables is partitioned according to the specific date and number range

List Partitioning

In this method, Tables is partitioned according to the listed values.

Rollup

The ROLLUP option allows you to include extra rows that represent the subtotals

CUBE

CUBE allows you to generate subtotals like the ROLLUP extension.

FIRST

FIRST() function returns the first value of selected column.

Last

last() function returns the last value of selected column.

Lead

We can access data of the next row

Lag

We can access data of the previous row

Rank

RANK() function use to find the rank of each row

DENSE_RANK

DENSE_RANK() function to rank rows in partitions with no gaps in ranking values.

Cume_list

calculates the cumulative distribution of a value in a set of values

PERCENT_RANK

PERCENT_RANK() function to calculate the percentile rankings of rows

TRANSFORMATION 1: Copy data from Source & store to
Target
Step 1: In the data integration folder open “Spoon (Windows Batch File)”.
Step 2: Go to File→New→Transformation
Step 3: Import SQL table to pentaho-
Design tab →Input folder → Drag and drop Table input
Double click on table input .

Connect to the Database:
Fill in the details as below. Here enter User Name and Password same as your database username and password. Then click on Test.
Click OK→
again Ok
Get SQL select statement… in table input window

Import table:
In t1, under tables, select the required table (In this case emp58).
Click on OK→Click on Yes
Preview in Table input window →OK
Step 4: Show output:
drag and drop table output
Hold the mouse pointer on Table input and select and drag the output connector to the Table output.
Double Click on Table Output.
In the Table Output Window, give name to the Target table, check the check boxes and click on Get fields.
Click on SQL.
Click on Execute→OK
Click on Close →OK
click on Debug this transformation
Click on Quick Launch
click on yes
If the Transformation is successful, you will see green ticks.

Close → close
You will also see the table created in the database with the name same as the target table table name.
Step 5: Run query in SQL Plus.
SQL> SELECT * FROM output;

TRANSFORMATION 2: Adding sequence
Step 1: Repeat Steps 2 and 3 from TRANSFORMATION 1.
Step 2: Perform transformation (Add sequence).
Drag and drop Add Sequence from the transform folder under the Design tab

Hold the mouse pointer on Table input and select and drag the output connector to the Add sequence
Double click on Add sequence and fill in the details as shown below→Click on OK.
Step 3: Perform transformation (Sort rows)
Drag and drop Sort rows from the transform folder under the Design tab.

Hold the mouse pointer on Add sequence and select and drag the output connector to the Sort rows.
Double click on Sort rows and fill in the details as shown below→Click on OK.
Step 4: Repeat Step 4 from TRANSFORMATION 1.
If the Transformation is successful, you will see green ticks.
Step 5: Run query in SQL.

TRANSFORMATION 3: Adding Calculator

Repeat Steps 1 to 3 from TRANSFORMATION 1.
Step 4: Perform Transformation
Drag and drop Calculator from Transform folder under Design tab.

TRANSFORMATION 4: Concatenation of two Fields

Step 1: Repeat Steps 2 and 3 from TRANSFORMATION 1.
Step 2: Perform Transformation.
Drag and drop Concat Fields from Transform folder under Design tab.
Hold the mouse Pointer on Table input and select and drag the output connector to the Concat Fields.
Double Click on Concat Fields and fill in the details as shown below→Click on OK.

TRANSFORMATION 5: Splitting of two Fields

Step 1: Repeat Steps 2 and 3 from TRANSFORMATION 1 (Import output table of concat fields transformation as Table input).
Step 2: Perform Transformation.
Drag and drop Concat Fields from Transform folder under Design tab.
Hold the mouse Pointer on Table input and select and drag the output connector to the Split Fields.
Double Click on Split Fields and fill in the details as shown below→Click on OK.
Step 3: Repeat Step 4 from TRANSFORMATION 1.
If the Transformation is successful, you will see green ticks
Step 4:
SQL> SELECT * FROM split;

TRANSFORMATION 6: Number Range

SQL> SELECT * FROM percent;
Step 1: Repeat Steps 2 and 3 from TRANSFORMATION 1.
Step 2: Perform Transformation.
Drag and drop Number Range from Transform folder under Design tab.
Hold the mouse Pointer on Table input and select and drag the output connector to the Number range.
Double Click on Number range and fill in the details as shown below→Click on OK.

TRANSFORMATION 7: String Operations

Step 1: Repeat Steps 2 and 3 from TRANSFORMATION 1.
Step 2: Perform Transformation.
Drag and drop Number Range from Transform folder under Design tab.
Hold the mouse Pointer on Table input and select and drag the output connector to the String operations.
Double Click on String operations and fill in the details as shown below→Click on OK.
Step 3: Repeat Step 4 from TRANSFORMATION 1.

If the Transformation is successful, you will see green ticks
step4:Run SQL query.

TRANSFORMATION 8: Merge Join
Step 1: Drag and drop 2 Data Grid from Input folder under Design tab.
Rename them as Employee and Department.
Step 2: Double click on them and insert records into respective grids→ Click on OK.
Step 2: Perform Sort rows transformation for both data grids respectively. Click on OK.
Step 3: Drag and Drop Merge join from joins folder under Design tab.
Hold the mouse Pointer on both the sort rows and select and drag the output connector to the Merge join as shown below.
Step 4: Double click on Merge join and fill in the details as shown below to perform INNER join→Click on OK.
Debug the transformation and perform Quick launch
Step 5: Double click on Merge join and fill in the details as shown below to perform LEFT OUTER join→Click on OK
Debug the transformation and perform Quick launch
Step 6: Double click on Merge join and fill in the details as shown below to perform RIGHT OUTER join→Click on OK.
Debug the transformation and perform Quick launch
Step 6: Double click on Merge join and fill in the details as shown below to perform FULL OUTER join→Click on OK.
Debug the transformation and perform Quick launch

TRANSFORMATION 9: Data validations
Step 1: Drag and drop Data Grid from Input folder under Design tab.
Rename it as Product.
Step 2: Double click on Product data grid and insert records as shown below→ Click on OK.
Step 3: Drag and drop Data validator from Validation folder under Design tab.
Hold the mouse pointer on Product data grid and select and drag the output connector to the Data validator.
Double clickc on Data validator→New Validattion
Give Valiadtion Name and click OK.
Select the validation to edit and fill in the details as shown below.
Click on add to set validation, set it 65 to Shifted and press Enter and click on ok.
Click on OK.
Step 4: Drag and drop Dummy from Flow folder under the Design tab.
Hold the mouse pointer on Data validator and select and drag the output connector to the Dummy. Select Main output of step.
Step 5: Drag and drop another Dummy from Flow folder under the Design tab and connect it to the data validator. Select Error handling of step. In the next window click in Copy.
Step 6: Quick Launch the transformation selecting one dummy file each.

Data Warehouse

A data Warehouse is based on analytical processing. A Data Warehouse maintains historical data

A Data Warehouse is integrated generally at the organization level, by combining data from different databases.