Why need data warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context. A process of transforming data into information and making it available to users in a timely enough manner to make a difference Information Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible.Data warehousing technology comprises a set of new concepts and tools which support the executives, managers and analysts with information material for decision making.The fundamental reason for building a data warehouse is to improve the quality of information in the organization.The main goal of data warehouse is to report and present the information in a very user friendly form.

What is Data warehousing?

A DW is a DB used for query, analysis and reporting . By definition DW is a subject oriented, intergrated, non volatile and time variant

Subject Oriented:- Represents a subject Area like sales, Mktg

Integrated :- Data Collated from multiple source systems integrated into a user readable unique format Ex:- male, female ,0,1, M,F, T, F

Non Volatile :-Data warehouse stores historical data

Time Variant :- Stores data time wise like weekly,monthly,quarterly, yearly

A data warehouse is a database designed to support a broad range of decision tasks in a specificorganization. It is usually batch updated and structured for rapid online queries and managerialsummaries. Data warehouses contain large amounts of historical data which are derived fromtransaction data, but it can include data from other sources also. It is designed for query andanalysis rather than for transaction processing.It separates analysis workload from transaction workload and enables an organization toconsolidate data from several sources.The term data warehousing is often used to describe the process of creating, managing and usinga data warehouse.

Explain the DW life cycle?

Data warehouses can have many different types of life cycles with independent data marts. Thefollowing is an example of a data warehouse life cycle.In the life cycle of this example, four important steps are involved.

Extraction - As a first step, heterogeneous data from different online transaction processingsystems is extracted. This data becomes the data source for the data warehouse.

Cleansing/transformation - The source data is sent into the populating systems where the data iscleansed, integrated, consolidated, secured and stored in the corporate or central data warehouse.

Distribution - From the central data warehouse, data is distributed to independent data martsspecifically designed for the end user.

Analysis - From these data marts, data is sent to the end users who access the data stored in thedata mart depending upon their requirement.

What is the life cycle of DW?

Getting data from OLTP systems from diff data sources

Analysis & staging - Putting in a staging layer- cleaning, purging, putting surrogate keys, SCM ,dimensional modeling

Loading

Writing of metadata

What approach to be followed for creation of Data Warehouse?

Top Down Approach (Data warehousing first) , Bottom Up (data marts), Enterprise Data Model (combines both)

What is ER model and Dimensional Model?

ER Model - Relational

Dimensional - Star Schema(central table fact table with numeric data , all others arelinked to central table, faster ,

but denormalised )

Snowflake Schema(one fact table,Normalizing the dimension tables ,

Fact Constellation(Different fact tables andcombined from one data mart to other)

What is Metadata?

Information about domain structure of data warehouse

What are different types of Dimensional Modeling?

Dimensional - Star Schema(central table fact table with numeric data , all others arelinked to central table, faster ,

but denormalised )

Snowflake Schema(one fact table,Normalizing the dimension tables ,

Fact Constellation(Different fact tables andcombined from one data mart to other)

Why need staging area database for DWH?

Staging area needs to clean operational data before loading into data warehouse. Cleaning in the sense your merging data which comes from different source

What is a data mart?

A data mart is a selected part of the data warehouse which supports specific decision supportapplication requirements of a company’s department or geographical region. It usually containssimple replicates of warehouse partitions or data that has been further summarized or derivedfrom base warehouse data. Instead of running ad hoc queries against a huge data warehouse, datamarts allow the efficient execution of predicted queries over a significantly smaller database.

What is difference between data mart and data warehouse?

A data mart designed for a particular line of business, such as sales, marketing, or finance. Whereas data warehouse is enterprise-wide/organizational The data flow of data warehouse depending on the approach.A data warehouse is for very large databases (VLDBs) and a data mart is for smaller databases.The difference lies in the scope of the things with which they deal.A data mart is an implementation of a data warehouse with a small and more tightly restrictedscope of data and data warehouse functions. A data mart serves a single department or part of anorganization. In other words, the scope of a data mart is smaller than the data warehouse. It is adata warehouse for a smaller group of end users.

Difference between OLTP and DWH?

OLTP system is basically application orientation (e.g., purchase order it is functionality of an application)

Where as in DWH concern is subject orient (subject in the sense customers, product, item, time)

OLTP

• Application Oriented

• Used to run business

• Detailed data

• Current up to date

• Isolated Data

• Repetitive access

• Clerical User

• Performance Sensitive

• Few Records accessed at a time (tens)

• Read/Update Access

• No data redundancy

• Database Size 100MB-100 GB

DWH

• Subject Oriented

• Used to analyze business

• Summarized and refined

• Snapshot data

• Integrated Data

• Ad-hoc access

• Knowledge User

• Performance relaxed

• Large volumes accessed at a time(millions)

• Mostly Read (Batch Update)

• Redundancy present

• Database Size 100 GB - few terabytes

What is star schema? And what is snowflake schema?

The center of the star consists of a large fact table and the points of the star are the dimension tables. Snowflake schemas normalized dimension tables to eliminate redundancy. That is, theDimension data has been grouped into multiple tables instead of one large table. Star schema contains demoralized dimension tables and fact table, each primary key values in dimension table associated with foreign key of fact tables. Here a fact table contains all business measures (normally numeric data) and foreign key values, and dimension tables has details about the subject area. Snowflake schema basically a normalized dimension tables to reduce redundancy in the dimension tables.

What is a Star Schema?

A star schema is a set of tables comprised of a single, central fact table surrounded by de-normalized dimensions. Each dimension is represented in a single table. Star schema implementdimensional data structures with de- normalized dimensions. Snowflake schema is an alternativeto star schema. A relational database schema for representing multidimensional data. The data isstored in a central fact table, with one or more tables holding information on each dimension.Dimensions have levels, and all levels are usually shown as columns in each dimension table.

What is a Snowflake Schema?

A snowflake schema is a set of tables comprised of a single, central fact table surrounded bynormalized dimension hierarchies. Each dimension level is represented in a table. Snowflakeschema implements dimensional data structures with fully normalized dimensions. Star schema isan alternative to snowflake schema.An example would be to break down the Time dimension and create tables for each level; years,quarters, months; weeks, days… These additional branches on the ERD create ore of aSnowflake shape then Star.

Difference between Star & Snowflake Schema?

Snowflaking is a star schema design technique to separately store logical attributes usually of lowcardinality along a loosely normalization technique. For example, you could snowflake thegender of your customers in order for you to track changes on these attributes if your customerdimension is too large to SCD's.The technique s not quite recommendable if you are going to use OLAP tools for your front enddue to speed issues.snowflaking allows for easy update and load of data as redundancy of data is avoided to someextent, but browsing capabilities are greatly compromised. But sometimes it may become anecessary evil.To add a little to this, snowflaking often becomes necessary when you need data for which thereis a one-to-many relationship with a dimension table. To try to consolidate this data into thedimension table would necessarily lead to redundancy (this is a violation ofsecond normal form, which will produce a Cartesian product). This sort of redundancy can causemisleading results in queries, since the count of rows is artificially large (due to the Cartesianproduct). A simple example of such a situation might be a "customer" dimension for which thereis a need to store multiple contacts. If the contact information is brought in to the customer table,there would be one row for each contact (i.e., one for each customer/contact combination). In thissituation, it is better just to create a "contact" snowflake table with a FK to the customer. Ingeneral, it is better to avoid snowflaking if possible, but sometimes the consequences of avoidingit are much worse.In star schema, all your dimensions will be linked directly with your fact table. On the other handin Snowflake schema, dimensions maybe interlinked or may have one to many relationship withother tables. As previous mails said this isn't a desirable situation but you can make best choiceonce you have gathered all the requirements.The snowflake is a design like a star but with a connect tables in the dimensions tables is arelation between 2 dimensions.

Why can’t I use a copy of my transactional system for my data warehouse?

This is one of the absolute worst things you can do. A lot of people initially go down this roadbecause a tool vendor will support the idea when making their sales pitch. Many of theseattempts will even experience success for a short period of time. It’s not until your data sets growand your business questions begin to be complex that this design mistake will really come out tobite you.

Which is better, Star or Snowflake?

Strict data warehousing rules would have you use a Star schema but in reality most designstend to become Snowflakes. They each have their pros and cons but both are far better thentryingto use a transactional system third-normal form design.

What is slowly changing dimension. What kind of scd used in your project?

Dimension attribute values may change constantly over the time. (Say for example customerdimension has customer\_id,name, and address) customer address may change over time.

How will you handle this situation?

There are 3 types,

one is we can overwrite the existing record,

second one is create additionalnew record at the time of change with the new attribute values.

Third one is create new field to keep new values in the original dimension table.

What is bitmap index why it’s used for DWH?

A bitmap for each key value replaces a list of rowids. Bitmap index more efficient for datawarehousing because low cardinality, low updates, very efficient for where clause.

What is difference between view and materialized view?

Views contains query whenever execute views it has read from base tableWhere as M views loading or replicated takes place only once which gives you better queryperformance

Refresh m views 1.on commit and 2. on demand

(Complete, never, fast, force).

Can 2 Fact Tables share same dimensions Tables? How many Dimension tables are associated with one Fact Table ur project?

Yes

What is ROLAP, MOLAP, and DOLAP?

ROLAP(Relational OLAP),MOLAP(Multidimensional OLAP), and DOLAP (DesktopOLAP).

In these three OLAParchitectures, the interface to the analytic layer is typically the same; what is quite differentis how the data is physically stored.

In MOLAP,the premise is that online analytical processing is best implemented by storingthe data multidimensional; that is,data must be stored multidimensional in order to be viewed in a multidimensional manner.

In ROLAP,architects believe to store the data in the relational model; for instance, OLAPcapabilities are best providedBusiness Intelligence against the relational database.

In DOLAP,is a variation that exists to provide portability for the OLAP user. It createsmultidimensional datasets that can betransferred from server to desktop, requiring only the DOLAP software to exist on the targetsystem. This provides significantadvantages to portable computer users, such as salespeople who are frequently on the road

and do not have direct access totheir office server.

What is an MDDB? and What is the difference between MDDBs and RDBMSs?

Multidimensional Database

There are two primary technologies that are used for storingthe data used in OLAP applications. These two technologies are multidimensional databases (MDDB) and relational databases(RDBMS). The major difference

between MDDBs and RDBMSs is in how they store data.

Relational databasesstore theirdata in a series of tables andcolumns. Multidimensionaldatabases, on the other hand, store their data in a largemultidimensional arrays.

Forexample, in an MDDB world, you might refer to a sales figure as Sales with Date,Product, and Location coordinates of12-1-2001, Car, and south, respectively.

Advantages of MDDB:

* Retrieval is very fast because
* The data corresponding to any combination of dimension members can be retrieved witha single I/O.
* Data is clustered compactly in a multidimensional array.
* Values are calculated ahead of time.
* The index is small and can therefore usually reside completely in memory.
* Storage is very efficient because The blocks contain only data.
* A single index locates the block corresponding to a combination of sparse dimensionnumbers.

What is the difference between OLTP & OLAP?

* OLTP stand for Online Transaction Processing. This is standard, normalized database structure. OLTP is designed for Transactions, which means that inserts, updates, and deletes must be fast. Imagine a call center that takes orders. Call takers are continually taking calls and entering orders that may contain numerous items. Each order and each item must be inserted into a database. Since the performance of database is critical, we want to maximize the speed of inserts (and updates and deletes). To maximize performance, we typically try to hold as few records in the database as possible.
* OLAP stands for Online Analytical Processing. OLAP is a term that means many things to many people. Here, we will use the term OLAP and Star Schema pretty much interchangeably. We will assume that star schema database is an OLAP system .( This is not the same thing that Microsoft calls OLAP; they extend OLAP to mean the cube structures built using their product, OLAP Services ). Here, we will assume that any system of read-only, historical, aggregated data is an OLAP system. A data warehouse(or mart) is way of storing data for later retrieval. This retrieval is almost always used to support decision-making in the organization. That is why many data warehouses are considered to be DSS (Decision-Support Systems ). Both a data warehouse and a data mart are storage mechanisms for read-only, historical, aggregated data. By read-only , we mean that the person looking at the data won’t be changing it. If a user wants at the sales yesterday for a certain product, they should not have the ability to change that number. The “historical” part may just be a few minutes old, but usually it is at least a day old. A data warehouse usually holds data that goes back a certain period in time, such as five years. In contrast, standard OLTP systems usually only hold data as long as it is “current” or active. An order table, for example, may move orders to an archive table once they have been completed, shipped, and received by the customer. When we say that data warehouses and data marts hold aggregated data , we need to stress that there are many levels of aggregation in a typical data warehouse.

Why did u choose STAR SCHEMA only? What are the benefits of STAR SCHEMA?

Because it’s denormalized structure, i.e., Dimension Tables are denormalized. Why todenormalize means the first (and often only) answer is : speed. OLTP structure is designed fordata inserts, updates, and deletes, but not data retrieval. Therefore, we can often squeeze somespeed out of it by denormalizing some of the tables and having queries go against fewer tables.These queries are faster because they perform fewer joins to retrieve the same record set.Joins are also confusing to many End users. By denormalizing, we can present the user with aview of the data that is far easier for them to understand.

Benefits of STAR SCHEMA:

• Far fewer Tables.

• Designed for analysis across time.

• Simplifies joins.

• Less database space.

• Supports “drilling” in reports.

• Flexibility to meet business and technical needs.

What is staging Area and Work Area?

Staging Area : -

- Holding Tables on DW Server.

- Loaded from Extract Process

- Input for Integration/Transformation

- May function as Work Areas

- Output to a work area or Fact Table

Work Area: -

-Temporary Tables

- Memory

What the difference is between a database, a data warehouse and a data mart?

Adatabaseis an organized collection of information.

Adata warehouseis a very large database with special sets of tools to extract andcleanse data from operational systemsand to analyze data.

Adata martis a focused subset of a data warehouse that deals with a single area ofdata and is organized for quickanalysis.

What is Data Mart, Data Warehouse and Decision Support System explain briefly?

Data Mart:

Adata martis a repository of data gathered from operational data and other sources that isdesigned to serve a particularcommunity of knowledge workers. In scope, the data may derive from an enterprise-widedatabase or data warehouse or be more specialized.The emphasis of a data martis onmeeting the specific demands of a particular group of knowledge users in terms of analysis,content, presentation, and ease-of-use. Users of a data mart can expect to have data presentedin terms that are familiar.In practice, the terms data mart and data warehouse each tend to imply the presence of theother in some form. However, most writers using the term seem to agree that the

design of adata mart tends to start from an analysis of user needsand that adata warehouse tendsto start from an analysis of what data already existsandhow it can be collected in such away that the data can later be used.

A data warehouse is a central aggregation of data(which can be distributed physically); a data mart is a data repository that may derive from adata warehouse or not and that emphasizes ease of access and usability for a particulardesigned purpose. In general, a data warehouse tends to be a strategic but somewhatunfinished concept; a data mart tends to be tactical and aimed at meeting an immediate need.

Data Warehouse:

Adata warehouseis a central repository for all or significant parts of the data that anenterprise's various business systems collect. The term was coined by W. H. Inmon.IBMsometimes uses the term "information warehouse."Typically, adata warehouseis housed on an enterprise mainframe server. Data from variousonline transaction processing (OLTP) applications and other sources is selectively extractedand organized on the data warehouse database for use by analytical applications and userqueries.

Data warehousing emphasizesthe capture of data from diverse sources for usefulanalysis and access, but does not generally start from the point-of-view of the end user orknowledge worker who may need access to specialized, sometimes local databases. The latteridea is known as the data mart.data mining, Web mining, and a decision support system (DSS)are three kinds ofapplications that can make use of a data warehouse.

Decision Support System:

A decision support system (DSS) is a computer program application that analyzes businessdata and presents it so that users can make business decisions more easily. It is an"informational application" (in distinction to an "operational application" that collects thedata in the course of normal business operation).Typical information that a decision support application might gather and present wouldbe:

Comparative sales figures between one week and the nextProjected revenue figures based on new product sales assumptionsThe consequences of different decision alternatives, given past experience in a context that isdescribed

A decision support system may present information graphically and may include an expertsystem or artificial intelligence (AI). It may be aimed at business executives or some othergroup of knowledge workers.

What r the differences between Heterogeneous and Homogeneous?

Heterogeneous :Stored in different Schemas

Homogeneous :Common structure

Heterogeneous :Stored in different file or db types

Homogeneous : Same database type

Heterogeneous :Spread across in several countries

Homogeneous : Same data center

Heterogeneous :Different platform n H/W config.

Homogeneous : Same platform and H/Ware configuration.

What are the responsibilities of a data warehouse consultant/professional?

The basic responsibility of a data warehouse consultant is to ‘publish the right data’.

Some of the other responsibilities of a data warehouse consultant are:

1. Understand the end users by their business area, job responsibilities, and computertolerance

2. Find out the decisions the end users want to make with the help of the data warehouse

3. Identify the ‘best’ users who will make effective decisions using the data warehouse

4. Find the potential new users and make them aware of the data warehouse

5. Determining the grain of the data

6. Make the end user screens and applications much simpler and more template driven

What are Stars and Cubes?

The star schema and OLAP cube are intimately related. Star schemas are most appropriate forvery large data sets. OLAP cubes are most appropriate for smaller data sets where analytic toolscan perform complex data comparisons and calculations. In almost all OLAP cube environments,it’s recommended that you originally source data into a star schema structure, and then usewizards to transform the data into the OLAP cube.

What is the necessity of having dimensional modeling instead of an ER modeling?

Compared to entity/relation modeling, it's less rigorous (allowing the designer more discretion inorganizing the tables) but more practical because it accommodates database complexity andimproves performance.

What are Dimensions and Facts?

Dimensional modeling begins by dividing the world intomeasurements and context.

Measurements are usually numeric and taken repeatedly.

Numeric measurements arefacts

Factsare always surrounded by mostly textual context that's true at the moment the fact is recorded.Facts are very specific, well-defined numeric attributes. By contrast, the context surrounding thefacts is open-ended and verbose. It's not uncommon for the designer to add context to a set offacts partway through the implementation.Dimensional modeling divides the world of data into two major types: Measurements andDescriptions of the context surrounding those measurements. The measurements, which aretypically numeric, are stored in fact tables, and the descriptions of the context, which are typicallytextual, are stored in the dimension tables.A fact table in a pure star schema consists of multiple foreign keys, each paired with a primarykey in a dimension, together with the facts containing the measurements.Every foreign key in the fact table has a match to a unique primary key in the respectivedimension (referential integrity). This allows the dimension table to possess primary keys thataren’t found in the fact table. Therefore, a product dimension table might be paired with a salesfact table in which some of the products are never sold.Dimensional models are full-fledged relational models, where the fact table is in third normalform and the dimension tables are in second normal form.The main difference between second and third normal form is that repeated entries are removedfrom a second normal form table and placed in their own “snowflake”. Thus the act of removingthe context from a fact record and creating dimension tables places the fact table in third normalform.

E.g. for Fact tables Sales, Cost, Profit

E.g. for Dimensions Customer, Product, Store, Time

What are Additive Facts? Or what is meant by Additive Fact?

The fact tables are mostly very huge and almost never fetch a single record into our answer set.We fetch a very large number of records on which we then do, adding, counting, averaging, ortaking the min or max. The most common of them is adding. Applications are simpler if theystore facts in an additive format as often as possible. Thus, in the grocery example, we don’t needto store the unit price. We compute the unit price by dividing the dollar sales by the unit saleswhenever necessary.

What is meant by averaging over time?

Some facts, like bank balances and inventory levels, represent intensities that are awkward toexpress in an additive format. We can treat these semi additive facts as if they were additive – butjust before presenting the results to the end user; divide the answer by the number of time periodsto get the right result. This technique is called averaging over time.

What is a Conformed Dimension?

When the enterprise decides to create a set of common labels across all the sources of data, theseparate data mart teams (or, single centralized team) must sit down to create master dimensionsthat everyone will use for every data source. These master dimensions are called ConformedDimensions.Two dimensions are conformed if the fields that you use as row headers have the same domain.

What is a Conformed Fact?

If the definitions of measurements (facts) are highly consistent, we call them as Conformed Facts. Fact conformation means that if two facts exist in two separate locations, then they must have the same name and definition. As examples, revenue and profit are each facts that must be conf ormed. By conforming a fact, then all business

processes agree on one common definition for the revenue and profit measures. Then, revenue and profit, even when taken from separate fact tables, can be mathematically combined.

Establishing conformity

Developing a set of shared, conformed dimensions is a significant challenge. Any dimensions that are common across the business processes must represent the dimension information in the same way. That is, it must be conformed. Each business process will typically have its own schema that contains a fact table, several conforming dimension tables, and dimension tables unique to the specific business function. The same is true for facts.

Degenerate dimensions

Before we discuss degenerate dimensions in detail, it is important to understand the following:

A fact table may consist of the following data:

\_ Foreign keys to dimension tables

\_ Facts which may be:

– Additive

– Semi-additive

– Non-additive

– Pseudo facts (such as 1 and 0 in case of attendance tracking)

– Textual fact (rarely the case)

– Derived facts

– year-to-date facts

\_ Degenerate dimensions (one or more)

What are the 3 important fundamental themes in a data warehouse?

The 3 most important fundamental themes are:

1. Drilling Down

2. Drilling Across and

3. Handling Time

What is meant by Drilling Down?

Drilling Down in a relational database means “adding a row header” to an existing SELECTstatement. For instance, if you are analyzing the sales of products at a manufacturer level, theselect list of the query reads:

SELECT MANUFACTURER, SUM(SALES).

If you wish to drill down on the list of manufacturers to show the brand sold, you add theBRAND row header:

SELECT MANUFACTURER, BRAND, SUM(SALES).

Now each manufacturer row expands into multiple rows listing all the brands sold. This is theessence of drilling down.We often call a row header a “grouping column” because everything in the list that’s notaggregated with an operator such as SUM must be mentioned in the SQL GROUP BY clause. Sothe GROUP BY clause in the second query reads, GROUP BY MANUFACTURER, BRAND.

What is meant by Drilling Across?

Drilling Across adds more data to an existing row. If drilling down is requesting ever finer andgranular data from the same fact table, then drilling across is the process of linking two or morefact tables at the same granularity, or, in other words, tables with the same set of groupingcolumns and dimensional constraints.A drill across report can be created by using grouping columns that apply to all the fact tablesused in the report.The new fact table called for in the drill-across operation must share certain dimensions with thefact table in the original query. All fact tables in a drill-across query must use conformeddimensions.

What is the significance of handling time?

Example, when a customer moves from a property, we might want to know:

1. who the new customer is

2. when did the old customer move out

3. when did the new customer move in

4. how long was the property empty etc

What is meant by Drilling Up?

If drilling down is adding grouping columns from the dimension tables, then drilling up issubtracting grouping columns.

What is meant by Drilling Around?

The final variant of drilling is drilling around a value circle. This is similar to the linear valuechain that I showed in the previous example, but occurs in a data warehouse where the relatedfact tables that share common dimensions are not arranged i n a linear order. The best example isfrom health care, where as many as 10 separate entities are processing patient encounters, and aresharing this information with one another.

E.g. a typical health care value circle with 10 separate entities surrounding the patient.

When the common dimensions are conformed and the requested grouping columns are drawnfrom dimensions that tie to all the fact tables in a given report, you can generate really powerfuldrill around reports by performing separate queries on each fact table and outer joining theanswer sets in the client tool.

What are the important fields in a recommended Time dimension table?

Time\_key

Day\_of\_week

Day\_number\_in\_month

Day\_number\_overall

Month

Month\_number\_overall

Quarter

Fiscal\_period

Season

Holiday\_flag

Weekday\_flag

Last\_day\_in\_month\_flag

Why have timestamp as a surrogate key rather than a real date?

The timestamp in a fact table should be a surrogate key instead of a real date because the rare timestamp that is inapplicable, corrupted, or hasn’t happened yet needs avalue that cannot be a real datemost end-user calendar navigation constraints, such as fiscal periods, end-of-periods,holidays, day numbers and week numbers aren’t supported by database timestampsinteger time keys take up much less disk space than full dates.

Why have more than one fact table instead of a single fact table?

We cannot combine all of the business processes into a single fact table because the separate fact tables in the value chain do not share all the dimensions. Yousimply can’t put the customer ship to dimension on the finished goods inventorydataeach fact table possesses different facts, and the fact table records are recorded atdifferent time along the chain.

What is mean by Slowly Changing Dimensions and what are the different types ofSCD’s?

Dimensions don’t change in predicable ways. Individual customers and products evolve slowlyand episodically. Some of the changes are true physical changes. Customers change theiraddresses because they move. A product is manufactured with different packaging. Other changesare actually corrections of mistakes in the data. And finally, some changes are changes in how welabel a product or customer and are more a matter of opinion than physical reality. We call thesevariations Slowly Changing Dimension (SCD).The 3 fundamental choices for handling the slowly changing dimension are:Overwrite the changed attribute, thereby destroying previous history

e.g. Useful when correcting an error

Issue a new record for the customer, keeping the customer natural key, but creating a newsurrogate primary key

Create an additional field in the existing customer record, and store the old value of theattribute in the additional field. Overwrite the original attribute field

A Type 1 SCD is an overwrite of a dimensional attribute. History is definitely lost. We overwritewhen we are correcting an error in the data or when we truly don’t want to save history.

A Type 2 SCD creates a new dimension record and requires a generalized or surrogate key for thedimension. We create surrogate keys when a true physical change occurs in a dimension entity ata specific point in time, such as the customer address change or the product packing change. Weoften add a timestamp and a reason code in the dimension record to precisely describe the change.The Type 2 SCD records changes of values of dimensional entity attributes over time. Thetechnique requires adding a new row to the dimension each time there’s a change in the value ofan attribute (or group of attributes) and assigning a unique surrogate key to the new row.

A Type 3 SCD adds a new field in the dimension record but does not create a new record. Wemight change the designation of the customer’s sales territory because we redraw the salesterritory map, or we arbitrarily change the category of the product from confectionary to candy.In both cases, we augment the original dimension attribute with an “old” attribute so we canswitch between these alternate realities.

What are the techniques for handling SCD’s?

Overwriting

Creating another dimension record

Creating a current value filed

What is a Surrogate Key and where do you use it?

Surrogate key used in slowly changing dimension table to track old and new values and it’sderived from primary key

A surrogate key is an artificial or synthetic key that is used as a substitute for a natural key. It isjust a unique identifier or number for each row that can be used for the primary key to the table.It is useful because the natural primary key (i.e. Customer Number in Customer table) can changeand this makes updates more difficult.Some tables have columns such as AIRPORT\_NAME or CITY\_NAME which are stated as theprimary keys (according to the business users) but ,not only can these change, indexing on anumerical value is probably better and you could consider creating a surrogate key called, say,AIRPORT\_ID. This would be internal to the system and as far as the client is concerned you maydisplay only the AIRPORT\_NAME.Another benefit you can get from surrogate keys (SID) is in Tracking the SCD - Slowly ChangingDimension.

A classical example:

On the 1st of January 2002, Employee 'E1' belongs to Business Unit 'BU1' (that's what would bein your Employee Dimension). This employee has a turnover allocated to him on the BusinessUnit 'BU1' But on the 2nd of June the Employee 'E1' is muted from Business Unit 'BU1' toBusiness Unit 'BU2.' All the new turnover has to belong to the new Business Unit 'BU2' but theold one should Belong to the Business Unit 'BU1.'If you used the natural business key 'E1' for your employee within your data warehouseeverything would be allocated to Business Unit 'BU2' even what actually belongs to 'BU1.'If you use surrogate keys, you could create on the 2nd of June a new record for the Employee 'E1'in your Employee Dimension with a new surrogate key.This way, in your fact table, you have your old data (before 2nd of June) with the SID of theEmployee 'E1' + 'BU1.' All new data (after 2nd of June) would take the SID of the employee 'E1'+ 'BU2.'You could consider Slowly Changing Dimension as an enlargement of your natural key: naturalkey of the Employee was Employee Code 'E1' but for you it becomesEmployee Code + Business Unit - 'E1' + 'BU1' or 'E1' + 'BU2.' But the difference with the naturalkey enlargement process is that you might not have all part of your new key within your facttable, so you might not be able to do the join on the new enlarge key so you need another id.Every join between dimension tables and fact tables in a data warehouse environment should bebased on surrogate key, not natural keys.

What is the necessity of having surrogate keys?

Production may reuse keys that it has purged but that you are still maintaining. Production might legitimately overwrite some part of a product description or acustomer description with new values but not change the product key or the customerkey to a new value. We might be wondering what to do about the revised attributevalues (slowly changing dimension crisis)Production may generalize its key format to handle some new situation in thetransaction system.

E.g. changing the production keys from integers to alphanumericor may have 12-byte keys you are used to have become 20-byte keys

Acquisition of companies

What are the advantages of using Surrogate Keys?

We can save substantial storage space with integer valued surrogate keys

Eliminate administrative surprises coming from production

Potentially adapt to big surprises like a merger or an acquisition

Have a flexible mechanism for handling slowly changing dimensions

What are Factless Fact tables?

Fact tables which do not have any facts are called factless fact tables. They may consist ofnothing but keys.There are two kinds of fact tables that do not have any facts at all.The first type of factless fact table is a table that records an event. Many event-tracking tables indimensional data warehouses turn out to be factless.

E.g. A student tracking system that detects each student attendance event each day.

The second type of factless fact table is called a coverage table. Coverage tables are frequentlyneeded when a primary fact table in a dimensional data warehouse is sparse.

E.g. A sales fact table that records the sales of products in stores on particular days under eachpromotion condition. The sales fact table does answer many interesting questions but cannotanswer questions about things that did not happen. For instance, it cannot answer the question,“which products were in promotion that did not sell?” because it contains only the records ofproducts that did sell. In this case the coverage table comes to the rescue. A record is placed inthe coverage table for each product in each store that is on promotion in each time period.

What are Causal dimension?

A causal dimension is a kind of advisory dimension that should not change the fundamental grainof a fact table.

E.g. why the customer bought the product? It can be due to promotion, sales etc.

What is Operational Data Store?

Operating Data Source - directly connects to application database

What is BI? And why do we need BI?

Business Intelligence, it is an ongoing process of various integration packages to analyze data.

What is Slicing and Dicing ? How we can do in Impromptu (We cannot do)? It is done only in

Powerplay.

GENERAL

What are the characteristics of a data warehouse?

Data in a data warehouse is organized as subject oriented rather than application oriented. It isdesigned and constructed as a non-volatile store of business data, transactions and events. Datawarehouse is a logically integrated store of data originating from disparate operational sources.It is the only source for deriving information needed by the end users. Several temporal modelingstyles are usually used in different areas of the data warehouse.

What are the characteristics of the data in a data warehouse?

Data in the DWH is integrated from various, heterogeneous operational systems (like databasesystems, flat files, etc.) and further external data sources (like demographic and statisticaldatabases, WWW, etc.). Before the integration, structural and semantic differences have to bereconciled, i.e., data have to be “homogenized” according to a uniform data model. Furthermore,data values from operational systems have to be cleaned in order to get correct data into the datawarehouse.The need to access historical data (i.e., histories of warehouse data over a prolonged period oftime) is one of the primary incentives for adopting the data warehouse approach. Historical data

are necessary for business trend analysis which can be expressed in terms of understanding thedifferences between several views of the real-time data (e.g., profitability at the end of eachmonth). Maintaining historical data means that periodical snapshots of the correspondingoperational data are propagated and stored in the warehouse without overriding previouswarehouse states. However, the potential volume of historical data and the associated storagecosts must always be considered in relation to their potential business benefits.Furthermore, warehouse data is mostly non-volatile, i.e., access to the DWH is typically read-oriented. Modifications of the warehouse data takes place only when modifications of the sourcedata are propagated into the warehouse.Finally, a data warehouse contains usually additional data, not explicitly stored in the operationalsources, but derived through some process from operational data (called also derived data).

Forexample, operational sales data could be stored in several aggregation levels (weekly, monthly,quarterly sales) in the warehouse.

When should a company consider implementing a data warehouse?

Data warehouses or a more focused database called a data mart should be considered when asignificant number of potential users are requesting access to a large amount of related historicalinformation for analysis and reporting purposes. So-called active or real-time data warehousescan provide advanced decision support capabilities.

What data is stored in a data warehouse?

In general, organized data about business transactions and business operations is stored in a datawarehouse. But, any data used to manage a business or any type of data that has value to abusiness should be evaluated for storage in the warehouse. Some static data may be compiledfor initial loading into the warehouse. Any data that comes from mainframe, client/server, orweb-based systems can then be periodically loaded into the warehouse. The idea behind a datawarehouse is to capture and maintain useful data in a central location. Once data is organized,managers and analysts can use software tools like OLAP to link different types of data togetherand potentially turn that data into valuable information that can be used for a variety of businessdecision support needs, including analysis, discovery, reporting and planning.

What is difference between data scrubbing and data cleansing?

Scrubbing data is the process of cleaning up the junk in legacy data and making itaccurate and useful for the next generationsof automated systems. This is perhaps the most difficult of all conversion activities.Very often, this is made more difficult when the customer wants to make good data out of bad data. This is the dog work. It is alsothe most important and cannot be donewithout the active participation of the user. Data scrubbing is a process of filtering, merging, decoding and translating the source datato create the validation data for data warehouse.Data cleansing is a process of removing errors and resolving inconsistencies insource data before loading data into targets. Data Cleansing is a two step process includingDETECTIONand thenCORRECTIONof errors in a data set.

What is the difference between OLTP and data warehouse?

Operational System :Transaction Processing

Data Warehouse: Query Processing

Operational System :Time Sensitive

Data Warehouse: History Oriented

Operational System :Operator View

Data Warehouse: Managerial View

Operational System :Organized by transactions (Order,

Data Warehouse: Organized by subject (Customer, Input, Inventory)

Operational System :Relatively smaller database

Data Warehouse: Large database size

Operational System :Many concurrent users

Data Warehouse: Relatively few concurrent users

Operational System :Volatile Data

Data Warehouse: Non Volatile Data

Operational System :Stores all data

Data Warehouse: Stores relevant data

Operational System :Not Flexible

Data Warehouse: Flexible

What is the Architecture of a data warehouse?

Adata warehouse system(DWS) comprises the data warehouse and all components used forbuilding, accessing and maintaining the DWH (illustrated in Figure 1). The center of a datawarehouse system is the data warehouse itself. The data import and preparation component isresponsible for data acquisition. It includes all programs, applications and legacy systemsinterfaces that are responsible for extracting data from operational sources, preparing and loadingit into the warehouse. The access component includes all different applications (OLAP or data

mining applications) that make use of the information stored in the warehouse.Additionally, a metadata management component (not shown in Figure 1) is responsible for themanagement, definition and access of all different types of

metadata. In general, metadata isdefined as “data about data” or “data describing the meaning of data”. In data warehousing, thereare various types of metadata, e.g., information about the operational sources, the structure and

semantics of the DWH data, the tasks performed during the construction, the maintenance andaccess of a DWH, etc. The need for metadata is well known. Statements like “A data warehousewithout adequate metadata is like a filing cabinet stuffed with papers, but without any folders orlabels” characterize the situation. Thus, the quality of metadata and the resulting quality ofinformation gained using a data warehouse solution are tightly linked.Implementing a concrete DWS is a complex task comprising two major phases. In the DWSconfigurationphase, a conceptual view of the warehouse is first specified according to userrequirements (data warehouse design). Then, the involved data sources and the way data will beextracted and loaded into the warehouse (data acquisition) is determined. Finally, decisions aboutpersistent storage of the warehouse using database technology and the various ways data will beaccessed during analysis are made.After the initial load (the first load of the DWH according to the DWH configuration), during theDWSoperationphase, warehouse data must be regularlyrefreshed, i.e., modifications ofoperational data since the last DWH refreshment must be propagated into the warehouse such that

data stored in the DWH reflect the state of the underlying operational systems. Besides DWHrefreshment, DWS operation includes further tasks like archiving and purging of DWH data orDWH monitoring.

What are the functional requirements for a data warehouse?

A data warehouse must be able to support various types of information applications.Decision support processing is the principle type of information application in a data warehouse,but the use of a data warehouse is not restricted to a decision support system.It is possible that each information application has its own set of requirements in terms of data,the way that data is modeled, and the way it is used.The data warehouse is where these applications get their "consolidated data."A data warehouse must consolidate primitive data and it must provide all facilities to deriveinformation from it, as required by the end-users. Detailed primitive data is of prime importance,but data volumes tend to be big and users usually require information derived from the primitivedata. Data in a data warehouse must be organized such that it can be analyzed or explored fromdifferent angles.Analysis of the historical context (the time dimension) is of prime importance.Examples of other important contextual dimensions are geography, organization, products,suppliers, customers, and so on.

Database administrators (DBAs) have always said that having non-normalized or de-normalized data is bad. Why is de-normalized data now okay when it's used for DecisionSupport?

Normalization of a relational database for transaction processing avoids processing anomalies andresults in the most efficient use of database storage. A data warehouse for Decision Support is notintended to achieve these same goals. For Data-driven Decision Support, the main concern is toprovide information to the user as fast as possible. Because of this, storing data in a de-normalized fashion, including storing redundant data and pre-summarizing data, provides the bestretrieval results. Also, data warehouse data is usually static so anomalies will not occur from

operations like add, delete and update a record or field.

How often should data be loaded into a data warehouse from transaction processing and othersource systems?

It all depends on the needs of the users, how fast data changes and the volume of information thatis to be loaded into the data warehouse. It is common to schedule daily, weekly or monthlydumps from operational data stores during periods of low activity (for example, at night or onweekends). The longer the gap between loads, the longer the processing times for the load when itdoes run. A technical IS/IT staffer should make some calculations and consult with potentialusers to develop a schedule to load new data.

What are the benefits of data warehousing?

Some of the potential benefits of putting data into a data warehouse include:

1. Improving turnaround time for data access and reporting;

2. Standardizing data across the organization so there will be one view of the "truth";

3. Merging data from various source systems to create a more comprehensive informationsource;

4. Lowering costs to create and distribute information and reports;

5. Sharing data and allowing others to access and analyze the data;

6. Encouraging and improving fact-based decision making.

What are the limitations of data warehousing?

The major limitations associated with data warehousing are related to user expectations, lack ofdata and poor data quality. Building a data warehouse creates some unrealistic expectations thatneed to be managed. A data warehouse doesn't meet all decision support needs. If needed data isnot currently collected, transaction systems need to be altered to collect the data. If data quality isa problem, the problem should be corrected in the source system before the data warehouse isbuilt. Software can provide only limited support for cleaning and transforming data. Missing andinaccurate data cannot be "fixed" using software. Historical data can be collected manually,coded and "fixed", but at some point source systems need to provide quality data that can beloaded into the data warehouse without manual clerical intervention.

How does my company get started with data warehousing?

Build one! The easiest way to get started with data warehousing is to analyze some existingtransaction processing systems and see what type of historical trends and comparisons might beinteresting to examine to support decision making. See if there is a "real" user need forintegrating the data. If there is, then IS/IT staff can develop a data model for a new schema andload it with some current data and start creating a decision support data store using a databasemanagement system (DBMS). Find some software for query and reporting and build a decisionsupport interface that's easy to use. Although the initial data warehouse/data-driven DSS mayseem to meet only limited needs, it is a "first step". Start small and build more sophisticatedsystems based upon experience and successes.

Why should the OLTP database different from data warehouse database?

OLTP and data warehousing require two very differently configured systemsIsolation of Production System from Business Intelligence SystemSignificant and highly variable resource demands of the data warehouseCost of disk space no longer a concernProduction systems not designed for query processingData warehouse usually contains historical data that is derived from transaction data, but it caninclude data from other sources. Having separate databases will separate analysis workload fromtransaction workload and enables an organization to consolidate data from several sources.

What is the main difference between Data Warehousing and Business Intelligence?

The differentials are:

DW - is a way of storing data and creating information through leveraging data marts. DM's aresegments or categories of information and/or data that are grouped together to provide'information' into that segment or category. DW does not require BI to work. Reporting tools cangenerate reports from the DW.

BI - is the leveraging of DW to help make business decisions and recommendations. Informationand data rules engines are leveraged here to help make these decisions along with statisticalanalysis tools and data mining tools.

What is a Physical data model?

During the physical design process, you convert the data gathered during the logical design phaseinto a description of the physical database, including tables and constraints.

What is a Logical data model?

A logical design is a conceptual and abstract design. We do not deal with the physicalimplementation details yet; we deal only with defining the types of information that we need.The process of logical design involves arranging data into a series of logical relationships calledentities and attributes.

What are an Entity, Attribute and Relationship?

An entity represents a chunk of information. In relational databases, an entity often maps to atable.An attribute is a component of an entity and helps define the uniqueness of the entity. Inrelational databases, an attribute maps to a column.The entities are linked together using relationships.

What are the different types of Relationships?

Entity-Relationship.

What is an ETL or ETT? And what are the different types?

ETL is the Data Warehouse acquisition processes of Extracting, Transforming (or Transporting)and Loading (ETL) data from source systems into the data warehouse.

E.g. Oracle Warehouse Builder, Power mart.

What is data mining?

Data Mining is the process of automated extraction of predictive information from largedatabases. It predicts future trends and finds behavior that the experts may miss as it lies beyondtheir expectations. Data Mining is part of a larger process called knowledge discovery;specifically, the step in which advanced statistical analysis and modeling techniques are appliedto the data to find useful patterns and relationships.Data mining can be defined as "adecision support process in which we search for patterns ofinformation in data." This search may be done just by the user, i.e. just by performing queries, inwhich case it is quite hard and in most of the cases not comprehensive enough to reveal intricatepatterns. Data mining uses sophisticated statistical analysis and modeling techniques to uncoversuch patterns and relationships hidden in organizational databases – patterns that ordinarymethods might miss. Once found, the information needs to be presented in a suitable form, withgraphs, reports, etc.

What is an OLAP?

OLAP is software for manipulating multidimensional data from a variety of sources. The data isoften stored in data warehouse. OLAP software helps a user create queries, views, representationsand reports. OLAP tools can provide a "front-end" for a data-driven DSS.On-Line Analytical Processing (OLAP) is a category of software technology that enablesanalysts, managers and executives to gain insight into data through fast, consistent,interactive access to a wide variety of possible views of information that has beentransformed from raw data to reflect the real dimensionality of the enterprise asunderstood by the user.OLAP functionality is characterized by dynamic multi-dimensional analysis of consolidatedenterprise data supporting end user analytical and navigational activities

What are the Different types of OLAP's? What are their differences?

OLAP -

Desktop OLAP(Cognos),

ROLAP,

MOLAP(Oracle Discoverer)

ROLAP, MOLAP and HOLAP are specialized OLAP (Online Analytical Analysis) applications.ROLAP stands for Relational OLAP. Users see their data organized in cubes with dimensions,but the data is really stored in a Relational Database (RDBMS) like Oracle. The RDBMS willstore data at a fine grain level, response times are usually slow.MOLAP stands for Multidimensional OLAP. Users see their data organized in cubes withdimensions, but the data is store in a Multi-dimensional database (MDBMS) like Oracle ExpressServer. In a MOLAP system lot of queries have a finite answer and performance is usuallycritical and fast.HOLAP stands for Hybrid OLAP, it is a combination of both worlds. Seagate Software's “Holos”is an example HOLAP environment. In a HOLAP system one will find queries on aggregateddata as well as on detailed data.DOLAP

What is the difference between data warehousing and OLAP?

The terms data warehousing and OLAP are often used interchangeably. As the definitionssuggest, warehousing refers to the organization and storage of data from a variety of sources sothat it can be analyzed and retrieved easily. OLAP deals with the software and the process ofanalyzing data, managing aggregations, and partitioning information into cubes for in-depthanalysis, retrieval and visualization. Some vendors are replacing the term OLAP with the termsanalytical software and business intelligence.

What are the Data Warehouse Center administration functions?

The functions of Visual Warehouse administration are:

Creating Data Warehouse Center security groups.

Defining Data Warehouse Center privileges for that group.

Registering Data Warehouse Center users.

Adding Data Warehouse Center users to security groups.

Registering data sources.

Registering warehouses (targets).

Creating subjects.

Registering agents.

Registering Data Warehouse Center programs.

What types of data sources does Data Warehouse Center support?

The Data Warehouse Center supports a wide variety of relational and non relational data sources.

You can populate your Data Warehouse Center warehouse with data from the following

databases and files:

Any DB2 family database

Oracle

Sybase

Informix

Microsoft SQL Server

IBM DataJoiner

Multiple Virtual Storage (OS/390), Virtual Machine (VM), and local area network (LAN) files

IMS and Virtual Storage Access Method (VSAM) (with Data Joiner ClassicConnect)

What is the Data Warehouse Center control database?

When you install the warehouse server, the warehouse control database that you specify duringinstallation is initialized. Initialization is the process in which the Data Warehouse Center createsthe control tables that are required to store Data Warehouse Center metadata. If you have morethan one warehouse control database, you can use the Data Warehouse Center -->Control Database Management window to initialize the second warehouse control database.However, only one warehouse control database can be active at a time.

What databases need to be registered as system ODBC data sources for the Data Warehouse Center?

The Data Warehouse Center database that needs to be registered as systemODBC data sources are:

source

target

control databases

What is Dynamic Data Store?

The need to share data is just as pressing as the need to share metadata. Often, several data martsin the same organization need the same information. For example, several data marts may need toread the same product data from operational sources, perform the same profitability calculations,and format this information to make it easy to review.If each data mart reads, transforms, and writes this product data separately, the throughput for theentire organization is lower than it could be. A more efficient approach would be to read,transform, and write the data to one central data store shared by all data marts. Transformation isa processing-intensive task, so performing the profitability calculations once saves time.Therefore, this kind of dynamic data store (DDS) improves throughput at the level of the entireorganization, including all data marts. To improve performance further, you might want tocapture incremental changes to sources. For example, rather than reading all the product data eachtime you update the DDS, you can improve performance by capturing only the inserts, deletes and updates that have occurred in the PRODUCTS table since the last time you updated the DDS.The DDS has one additional advantage beyond performance: when you move data into the DDS,you can format it in a standard fashion. For example, you can prune sensitive employee data thatshould not be stored in any data mart. Or you can display date and time values in a standardformat. You can perform these and other data cleansing tasks when you move data into the DDS

instead of performing them repeatedly in separate data marts.

When should you create the dynamic data store? Do you need a DDS at all?

To decide whether you should create a dynamic data store (DDS), consider the following issues:

How much data do you need to store in the DDS?

The one principal advantage of datamarts is the selectivity of information included in it. Instead of a copy of everythingpotentially relevant from the OLTP database and flat files, data marts contain only theinformation needed to answer specific questions for a specific audience (for example,sales performance data used by the sales division). A dynamic data store is a hybrid ofthe galactic warehouse and the individual data mart, since it includes all the data neededfor all the data marts it supplies. If the dynamic data store contains nearly as muchinformation as the OLTP source, you might not need the intermediate step of the dynamicdata store. However, if the dynamic data store includes substantially less than all the datain the source databases and flat files, you should consider creating a DDS staging area.

What kind of standards do you need to enforce in your data marts?

Creating a DDSis an important technique in enforcing standards. If data marts depend on the DDS forinformation, you can provide that data in the range and format you want everyone to use.For example, if you want all data marts to include the same information on customers,you can put all the data needed for this standard customer profile in the DDS. Any datamart that reads customer data from the DDS should include all the information in thisprofile.

How often do you update the contents of the DDS?

If you plan to frequently updatedata in data marts, you need to update the contents of the DDS at least as often as youupdate the individual data marts that the DDS feeds. You may find it easier to read datadirectly from source databases and flat file systems if it becomes burdensome to updatethe DDS fast enough to keep up with the needs of individual data marts. Or, if particulardata marts need updates significantly faster than others, you can bypass the DDS forthese fast update data marts.

Is the data in the DDS simply a copy of data from source systems, or do you plan toreformat this information before storing it in the DDS?

One advantage of the dynamicdata store is that, if you plan on reformatting information in the same fashion for severaldata marts, you only need to format it once for the dynamic data store. Part of thisquestion is whether you keep the data normalized when you copy it to the DDS.

How often do you need to join data from different systems?

On occasion, you mayneed to join records queried from different databases or read from different flat filesystems. The more frequently you need to perform this type of heterogeneous join, themore advantageous it would be to perform all such joins within the DDS, then make theresults available to all data marts that use the DDS as a source.

What is a Materialized View?

Materialized views, also called snapshots, are schema objects that can be used to summarize,precompute, replicate, and distribute data. They are suitable in various computing environmentsespecially for data warehousing.

From a physical design point of view, Materialized Views resembles tables or partitioned tablesand behave like indexes.Materialized views in these environmentsare typically referred to as summaries, because they store summarized data. Theycan also be used to precompute joins with or without aggregations. A materializedview eliminates the overhead associated with expensive joins or aggregations for alarge or important class of queries.

What is the need of Materialized View?

Materialized views are used in data warehouses to increase the speed of queries onvery large databases. Queries to large databases often involve joins between tablesor aggregations such as SUM, or both. These operations are very expensive in termsof time and processing power.

What is the significance of Materialized Views in data warehousing?

In data warehouses, materialized views are used to precompute and store aggregated data such assums and averages. Materialized views in these environments are typically referred to assummaries because they store summarized data. They can also be used to precompute joins withor without aggregations.Cost-based optimization can use materialized views to improve query performance byautomatically recognizing when a materialized view can and should be used to satisfy a request.The optimizer transparently rewrites the request to use the materialized view. Queries are thendirected to the materialized view and not to the underlying detail tables or views.

How does MV’s work?

The query optimizer can use materialized views byautomatically recognizing when an existing materialized view can and should beused to satisfy a request. It then transparently rewrites the request to use thematerialized view. Queries are then directed to the materialized view and not to theunderlying detail tables. In general, rewriting queries to use materialized viewsrather than detail tables results in a significant performance gain.If a materialized view is to be used by query rewrite, it must be stored in the samedatabase as its fact or detail tables. A materialized view can be partitioned, and youcan define a materialized view on a partitioned table and one or more indexes onthe materialized view.

The types of materialized views are:

Materialized Views with Joins and Aggregates

Single-Table Aggregate Materialized Views

Materialized Views Containing Only Joins

What is the major difference between an index and Materialized view?

Unlike indexes, materialized views can be accessed directly using a SELECT statement.

What are the procedures for refreshing Materialized views?

Oracle maintains the data in materialized views by refreshing them after changes are made totheir master tables.

The refresh method can be:

a) incremental (fast refresh) or

b) complete

For materialized views that use the fast refresh method, a materialized view log or direct loaderlog keeps a record of changes to the master tables.Materialized views can be refreshed either on demand or at regular time intervals.

Alternatively, materialized views in the same database as their master tables can be refreshedwhenever a transaction commits its changes to the master tables.

What are materialized view logs?

A materialized view log is a schema object that records changes to a master table’s data so that amaterialized view defined on the master table can be refreshed incrementally. Another name formaterialized view log is snapshot log.

Each materialized view log is associated with a single master table. The materialized view logresides in the same database and schema as its master table.

Why Constraints are Useful in a Data Warehouse?

Constraints provide a mechanism for ensuring that data conforms to guidelinesspecified by the database administrator. The most common types of constraintsinclude unique constraints (ensuring that a given column is unique), not-nullconstraints, and foreign-key constraints (which ensure that two keys share aprimary key-foreign key relationship).

What are the tools used in ur environment for Data Mining and Data Modeling?

* Data Modeling :- Erwin from platinum, Visio from Microsoft
* Data Mining: - A component of Data Warehousing Solution, used to discover patterns and relationships in ur data in order to help u make better business decisions. Also called knowledge discovery in databases (KDD), Oracle’s sol to this business pr of data mining is known as Oracle Darwin or as oracle’s Data mining Suite. Used to help u build predictive models, Data Mining is a discovery process that allows users to understand the substance of and the relationships between their data Ex: - What is likely to happen to the Boston Sales next month and why? , Characteristics of Data Mining are Prospective, Proactive information delivery

What is diff between data warehousing and enterprise data warehousing?

Widely used by whole enterprise data warehousing .Total GE warehouse is called enterprise data warehouse. Data warehouse build for all business in an enterprise is enterprise data warehousing andfor a single business is data warehouse and module in a data warehouse is data martEx:- GE capital , GE appliances, GE electrical each has their own data warehouse.

What is Slowly Growing Dimension?

The dimension, which does not have many changes for the warehouse, e.g. Region dimensiondoes not have many changes it may add a row in a year. You have a wizard in Informatica just gothroughThe Target mapping filters source rows based on user-defined comparisons, and then inserts onlythose found to benewto the target. Use the Target mapping to determine which source rows arenew and to load them to an existing target table. In the Target mapping, all rows are current. Usethe Target mapping to load a fact or dimension table, one in which existing data does not requireupdates.For example, you have a site code dimension table that contains only a store name and a

corresponding site code that you update only after your company opens a new store. Althoughlisted stores might close; you want to keep the store code and name in the dimension for historicalanalysis. With the Target mapping, you can load new source rows to the site code dimension tablewithout deleting historical sites.

What is a cube?

A set of related factual measures, aggregates, and dimensions for a specific dimensionalanalysis problem. Ex: regional product sales.

Explain linking universes?

To link the universes u must have exported to where the kernel universe at least once otherwisethe designer does not allow the link.

What is Click Stream In Data Warehousing?

Click stream is basically web based data warehousing analysis basically e-web intelligenceWhen to index a particular column on what percentageGenerally if you are creating ordinary b-tree index.........if you know that that columns will bemostly reference in select statementif your select statement is retrieving more than 30 % rows then...it's better not to use indexwill be used out of the table when you are retrieving smaller no. of rowsif you retrieving more no. of rows. you can read the table directly rather than using the index(Fulltable scan)

What is STAR QUERY transformation?

Generally used to improve the query performance in generally starschema models.The star transformation is a cost-based query transformation aimed at executingstar queries efficiently. Whereas the star optimization works well for schemas with asmall number of dimensions and dense fact tables, the star transformation may beconsidered as an alternative if any of the following holds true:

In order to get the best possible performance for star queries, it is important tofollow some basic guidelines:

- A bitmap index should be built on each of the foreign-key columns of the facttable(s).

-The initialization parameter STAR\_TRANSFORMATION\_ENABLED should beset to TRUE.

This enables an important optimizer feature for star-queries; it isset to FALSE by default for backwards-

compatibility.

- The cost-based optimizer should be used. [This does not apply solely to star1 schemas: all data warehouses

should always use the cost-based optimizer].

What is the procedure to load the fact table. Give in detail?

Based on the requirement to your fact table choose the sources and data and transform it based on your business needs. For the fact table you need a primary key so use a sequence generator transformation to generate a unique key and pipe it to the target (fact) table with the foreign keys from the source tables. we use the 2 wizards (i.e.) the getting started wizard and slowly changing dimension wizard to load the fact and dimension tables by using these 2 wizards we can create different types of mappings according to the business requirements and load into the star schemas(fact and dimension tables). first dimension tables need to be loaded then according to the specifications the fact tables should be loaded. Don’t think that fact tables r different in case of loading it is general mapping as we do for other tables. specifications will play important role for loading the fact. usually source records are looked up with the records in the dimension table.DIM tables are called lookup or reference table. all the possible values are stored in DIM table. e.g. product all the existing prod\_id will be in DIM table. when data from source is looked up against the dim table the corresponding keys are sent to the fact table. this is not the fixed rule to be followed it may vary as per ur requirments and methods u follow. sometimes only the existence check will be done and the prod\_id itself will be sent to the fact.

What is a degenerate dimension?

A degenerate dimension sounds a bit strange, but it is a dimension without attributes. It is a transaction-based number which resides in the fact table. There may be more than one degenerate dimension inside a fact table.

Identifying garbage dimensions

What is a garbage or junk dimension?

A garbage dimension is a dimension that consists of low-cardinality columns such as codes, indicators, and status flags. The garbage dimension is also referred to as a junk dimension.The attributes in a garbage dimension are not

related to any hierarchy.

What is a Non-Aditive Facts?

Non-additive facts

Non-additive facts are facts which cannot be added meaningfully across any dimensions.

What is a Textual Facts?

Textual facts:

Adding textual facts does not result in any number. However, counting textual facts may result in a sensible number.

Per-unit prices:

Adding unit prices does not produce any meaningful

Percentages and ratios:

Measures of intensity:

Measures of intensity such as the room temperature

Averages:

What is a Semi-Additivel Facts?

Semi-additive facts are f acts which can be summarized across some dimensions but not others. Examples of semi-additive facts include the following:

\_ Account balances

\_ Quantity-on-hand

adding the monthly balances across the different days for the month of January results in an incorrect balance figure.

However, if we average the account balance to find out daily average balance during each day of the month, it would be valid.

What is a Event Based Facts tables?

Event-based fact tables are tables that record events. For example, event fact tables are used to record events such as Web page clicks and employee or student attendance. Events, such as a W eb user clicking on a Web page of a Web site, do not always result in facts. In other words, millions of such Web page click events do not always result in sales. If we are interested in handling such event-based scenarios where there are no facts, we use event fact tables which consist of either pseudo facts or these tables have no facts (factless) at all. From a conceptual perspective, the event-based fact tables capture the many-to-many relationships between the dimension tables.

What are the Ad hoc quries, Canned Quries/Reports? and How do u create them?

The data warehouse will contain two types of query. There will be fixed queries that are clearly defined and well understood, such as regular reports, canned queries (standard reports) and common aggregations . There will also be ad hoc queries that are unpredictable, both in quantity and frequency.

Ad Hoc Query:

Ad hoc queries are the starting point for any analysis into a database. Any business analyst wants to know what is inside the database. He then proceeds by calculating totals, averages, maximum and minimum values for most attributes within the database. These are unpredictable element of a data warehouse. It is exactly that ability to run any query when desired and expect a reasonable response that makes the data warhouse Worthwhile, and makes the design such a significant challenge. The end-user access tools are capable of automatically generating the database query that answers any Question posed by the user. The user will typically pose questions in terms that

they are familier with (for example, sales by store last week); this is converted into the database query by the access tool, which is aware of the structure of information within the data warehouse.

Canned queries:

Canned queries are predefined queries. In most instances, canned queries contain prompts that allow you to customize the query for your specific needs. For example , a prompt may ask you for a School, department, term, or section ID. In this instance you would enter the name of the School, department or term, and the query will retrieve the specified data from the Warehouse.You can measure resource requirements of these queries, and the results can be used for capacity palnning and for database design. The main reason for using a canned query or report rather than creating your own is that your chances of misinterpreting data or getting the wrong answer are reduced. You are assured of getting the right data and the right answer.