warehouse for sales and shipping)

DMQL (Data mining query language)

Data warehouses can be defined using two language primitives — cube definition.
dimension definition.

syntax of cube definition

define cube <cube_name> [<dimension_list>]: <measure_list>

syntax of dimension definition

define dimension <dimension_name> as (<attribute_or_dimension_list>)

Ex Star schema definition

define cube sales_star [time, item, branch, location]:

dollars_sold = sum(sales_in_dollars),
units_sold = count()

define dimension time as (time_key, day, day_of_week, month, quarter, year)

define dimension item as (item_key, item_name, brand, type, supplier_type)

define dimension branch as (branch_key, branch_name, branch_type)

define dimension location as (location_key, street, city, ~~state~~ state, country)

The define cube statement defines a data cube called sales_star, which corresponds to the central sales fact table.

Define dimension statement is used to define each of the dimensions.

Ex Snowflake schema.

define cube as sales_snowflake [time, item, branch, location]:
dollars_sold = sum(sales_in_dollars), units_sold = count()

define dimension time as (time_key, day, day_of_week, month, quarter, year)

define dimension item as (item_key, item_name, brand, type, supplier(supplier_key, supplier_type))

define dimension branch as (branch_key, branch_name, branch_type)

define dimension location as (location_key, street, city (city_key, city, state, country))

Example Fact constellation schema

define cube sales [time, item, branch, location]:
dollars_sold = sum(sales_in_dollars), units_sold = count()
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city, state, country)
define cube shipping [time, item, shipper, from_location, to_location]:
dollars_cost = sum(cost_in_dollars), units_shipped = count()
define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper_key, shipper_name, location as location in cube sales, shipper_type)
define dimension from_location as location in cube sales
define dimension to_location as location in cube sales.

Explanation

define cube statement is used to define data cubes for sales and shipping.

time, item and location dimensions of the sales cube are shared with the shipping cube.

So, define dimension time as time in cube sales.

concept Hierarchies

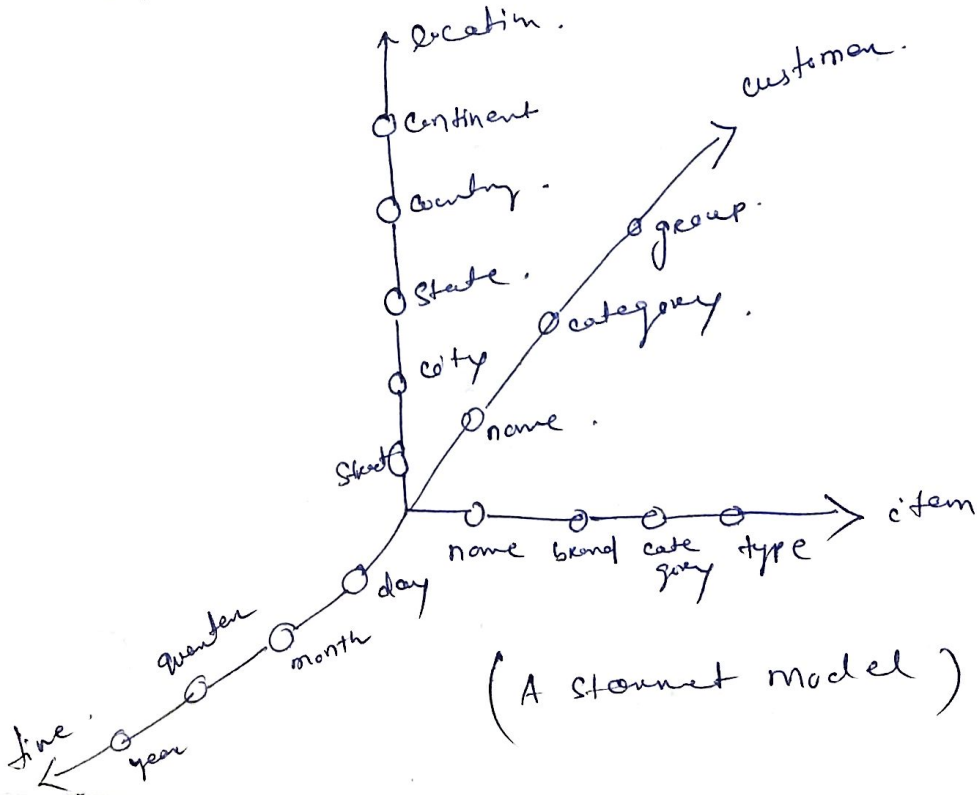
It defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies.

This organization provides users with the flexibility to view data from different perspectives.

Starnet query model

A starnet model consists of radial lines emanating from a central point, where each line represents a concept hierarchy for a dimension. Each abstraction level in the hierarchy is called a footprint.



(A starnet model)

From multiple heterogeneous sources, analytical

Measures: Their categorization and computation

Measure → is a numerical function that can be evaluated at each point in the data cube space.

A measure value is computed for a given point by aggregating the data corresponding to the respective dimension-value pairs defining the given points.

Measures can be divided into three categories

(1) distributive

(2) Algebraic

(3) Holistic.

Distributive

An aggregate function is distributive if it can be computed in a distributed manner.

Suppose the data are partitioned into n sets. We apply the function to each partition, resulting in aggregate value. If the result derived by applying the function to the n aggregate values is the same as that derived by applying the function to the entire data set, the function can be computed in a distributed manner.

Ex count()
sum()
min()
max()

For example, count() can be computed for a data cube by first partitioning the cube into a set of subcubes, computing count() for each subcube, and then summing up the counts obtained for each subcube.

Algebraic

An aggregate function is algebraic if it can be computed by an algebraic function with M arguments, each of which is obtained by applying a distributive aggregate function.

Ex $\quad avg() = \frac{sum()}{count()}$

where both sum() and count() are distributive aggregate functions.

Ex $\quad$ Standard_deviation().
min_N() → find N minimum values.
max_N() → find N maximum values.

Holistic

An aggregate function is holistic if there is no constant bound on the storage size needed to describe a subaggregate. That is there does not exist an algebraic function with M arguments that do the computation.

Ex $\quad$ median()
mode()
rank()

Concept Hierarchy

Star net model

The process of Data warehouse design

The design process consists of the following steps:

① Choose a business process to model, for example : orders, invoices, shipments, inventory etc.

If the process is organizational and involves multiple complex object collections, a data warehouse model should be followed.

If the process is departmental, a data mart model should be chosen.

(2) Choose the grain of the business process.

The grain is the fundamental, atomic level of data to be represented in the fact table for this process.

(3) Choose the dimensions that will apply to each fact table record. Ex:- time, item, customer, supplier etc.

(4) Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars_sold and units_sold.

A three-tier Data Warehouse Architecture.

(1) The bottom tier is a warehouse database server that is a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources. These tools and utilities perform data extraction, cleaning and transformation, as well as loading and refresh functions to update the data warehouse.

city.

Quick searching in data cube

Bitmap indexing (Indexing OLAP Data).

It is an alternative representation of the record-ID list.

In the bitmap index for a given attribute, there is a distinct bit vector, BV, for each value v in the domain of the attribute. If the domain of a given attribute consists of n values, then n bits are needed for each entry in the bitmap index.

If the attribute has the value v for a given row in the data table, then the bit representing that value is set to 1 in the corresponding row of the bitmap index. All other bits for that row are set to 0.

Base table.

RID	item	city
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

item bitmap index table.

RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

city bitmap index.

RID	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

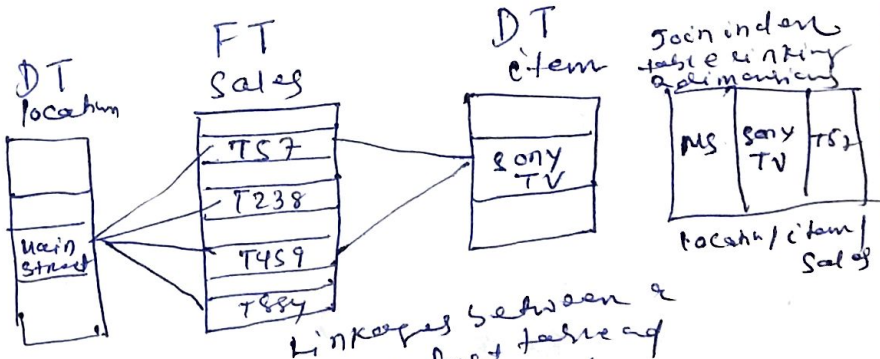
Join indexing

join index table for location/sales

location	sales_key
Main Street	T57
-do-	T238
-do-	T459
----	T564

For item/sales

item	sales_key
Sony TV	T57
-do-	T459



Linkages between a sales fact table and dimension tables.

The join index records can identify tuples without performing costly join operations. It is useful for maintaining relationship between a foreign key and its matching primary keys.

If two relations R(RID, A) and S(B, SID) join on the attributes A and B, then the join index record contains the pairing (RID, SID) where RID and SID are record identifiers from R and S, respectively.