UNIT-3

What is a data warehouse?

The data warehouse is a informational environment that
_Provides an integrated and total view of the enterprise
_makes the enterprises current historical information easily available for decision making
_Makes decision support transactions possible without hindering operational systems
_Renders the organization information consistent
_Presence the flexible and interactive source of strategic information.

A data warehouse is a subject oriented, integrate, time-variant, and nonvolatile collection of data in support of management's decision making process "Data warehousing: The process of constructing and using data warehouses"

DATA WAREHOUSE_SUBJECT ORIENTED

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple concise view around particular subject issues by excluding data that are not useful in the decision support process.

- Data warehouse-integrated

  Constructed by integrating multiple, heterogeneous data sources-relational databases, flat files, online transaction records
- Data cleaning and data integration techniques are applied-Ensure consistency in naming conventions, encoding structures, attribute measures, etc.
- E.g., hotel price: currency, tax, breakfast covered, etc.

_when data is moved to warehouse, it is converted

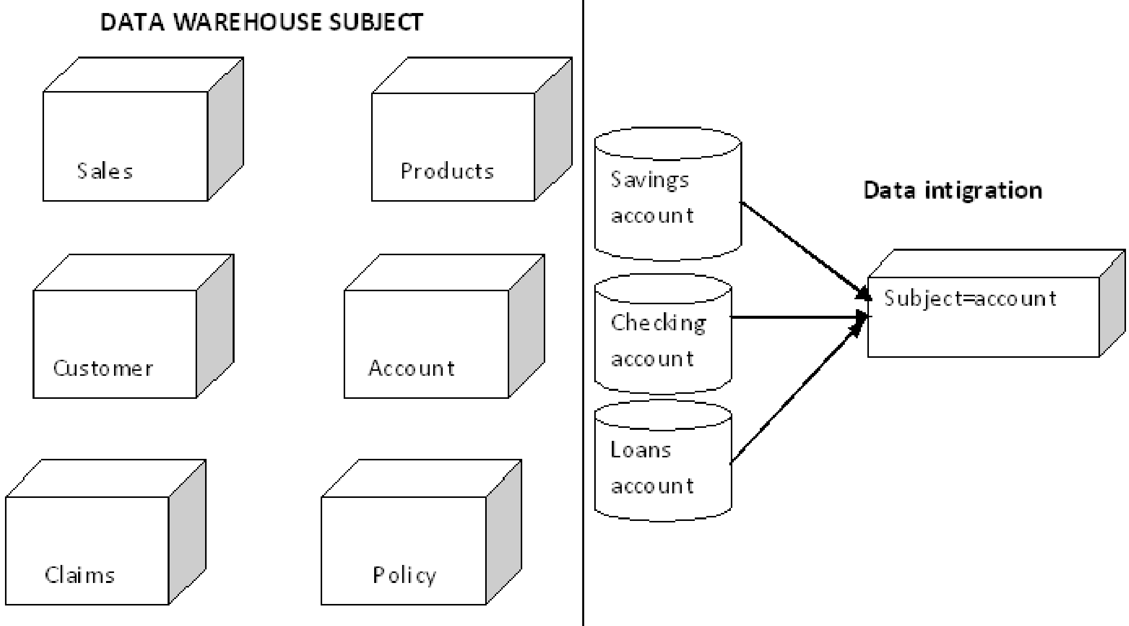


Figure1.11 Data warehouse is subjects oriented

Data warehouse-Time variant

- The time horizon for the data warehouse is significantly longer than that of operational systems

_operational database: currents value data
_data warehouse data: provide information from a historical perspective (e.g.,past5-10yars)

- Every key structure in the data warehouse
- Contains an element of time ,explicitly or implicitly
- But the key of operational data may or may not contain “time element “

Data warehouse-nonvolatile

- A physically separate storage of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Requires only to operations in data accessing:

Initial loading of data and access of data

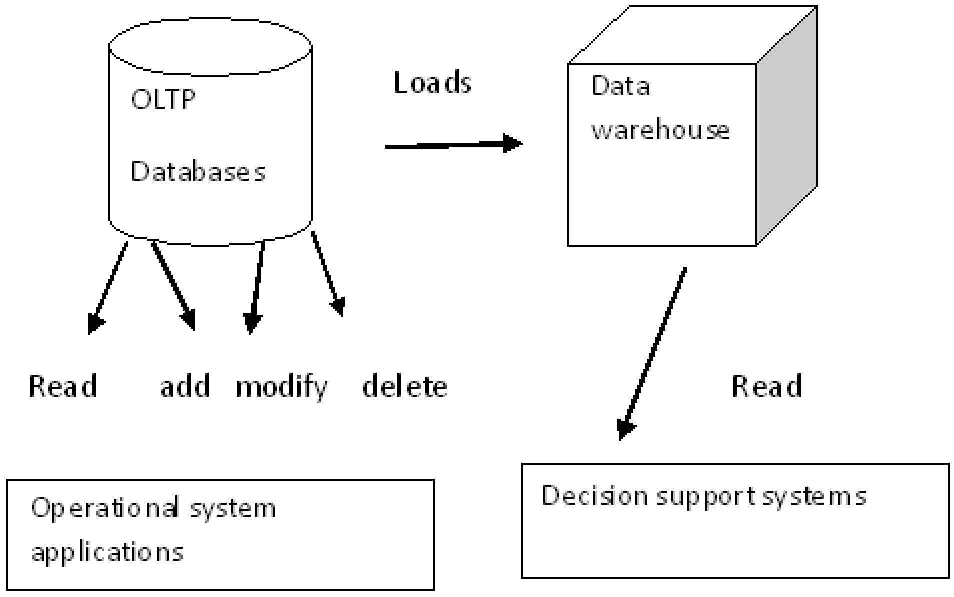


Figure1.13 data warehouse in non volatile

Data warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration:
- Build wrappers/mediators on the top of heterogeneous databases
- Query driven approach
- When a query is pose to a client site, a meta dictionary is use to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
- Complex information filtering ,complete for resources

- Data ware house: update -driven ,high performance

  -Information from heterogeneous sources is integrated in advance and stored in ware house for direct query and analysis

  Data ware house vs. Operational DBMS
  - OLTP(On-line transaction processing)
  - Major task of traditional relational DBMS
  - Day to day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting etc.
  - OLAP(online analytical processing)
  - Major task of warehouse system
  - Data analysis an decision making
  - Distinct features (OLTP VS .OLAP):
  - User and system orientation :customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read only but complex queries

Data warehouse VS Data Mart

Data Warehouse	Data Mart
• Corporate or enterprise wide • Union of all data marts • Data received from staging area • Queries on presentation resource • Structure for corporate view of data	• Departmental • A single business process • Start-join(facts& dimensions) • Technology optical for data access and analysis • Structure to suite the departmental view of data

OILP vs. OLAP

Feature	*OLTP*	*OLAP*
Orientation	Transaction	Analysis
Users	Clerk, IT professional	Knowledge worker
Function	Day to day operations	decision support
Db design	Application-orient	Subject-orient
Data	Current, up-to-date	Historical
View	Detailed, flat relational isolated	Summarized, multidimensional integrated consolidated
Usage	Repetitive	Ad-hoc
Access	Read/write index/hash on prim. Key	Lots of scans
Focus	Data in	Information out
Unit of work	Short, simple transaction	Complex query
# records accessed	Tens	Millions
# users	Thousands	Hundreds
DB size	100MB-GB	100GB-MB

Metric	Transaction throughput	Query throughput, response

1.9 DATA WAREHOUSE COMPONENTS

Data Warehouse Database

The central data warehouse database is the cornerstone of the date ware housing environment. This database is almost always implemented on the relational database management system (RDBMS) technology. However, this kind of implementation is often constrained by the fact that traditional RDMBS products or optimized for traditional database processing. Certain data warehouse attributes, such as very large database size, and hoc query processing and the need for flexible user view creation including aggregates, multi-table joins and drill-downs have became drivers for different technological approaches to data warehouse, database.

- Parallel relational database designs for scalability that include shared memory, shared disk, or shared nothing models implemented on various multiprocessor Configurations (symmetric multi processors or SMP, massively parallel processors or MPP, and/or clusters of uni-or multi processors).
    - An innovative approach to speed up a traditional RDBMS by using new index structures to bypass relational table scans.
    - Multidimensional data bases (MDDBs) that are based on proprietary database technology. MDDBs enable on-line analytical processing (OLAP) tools that architecturally belong to a group of data warehousing components jointly categorized as the data query, reporting, analysis and mining tools.

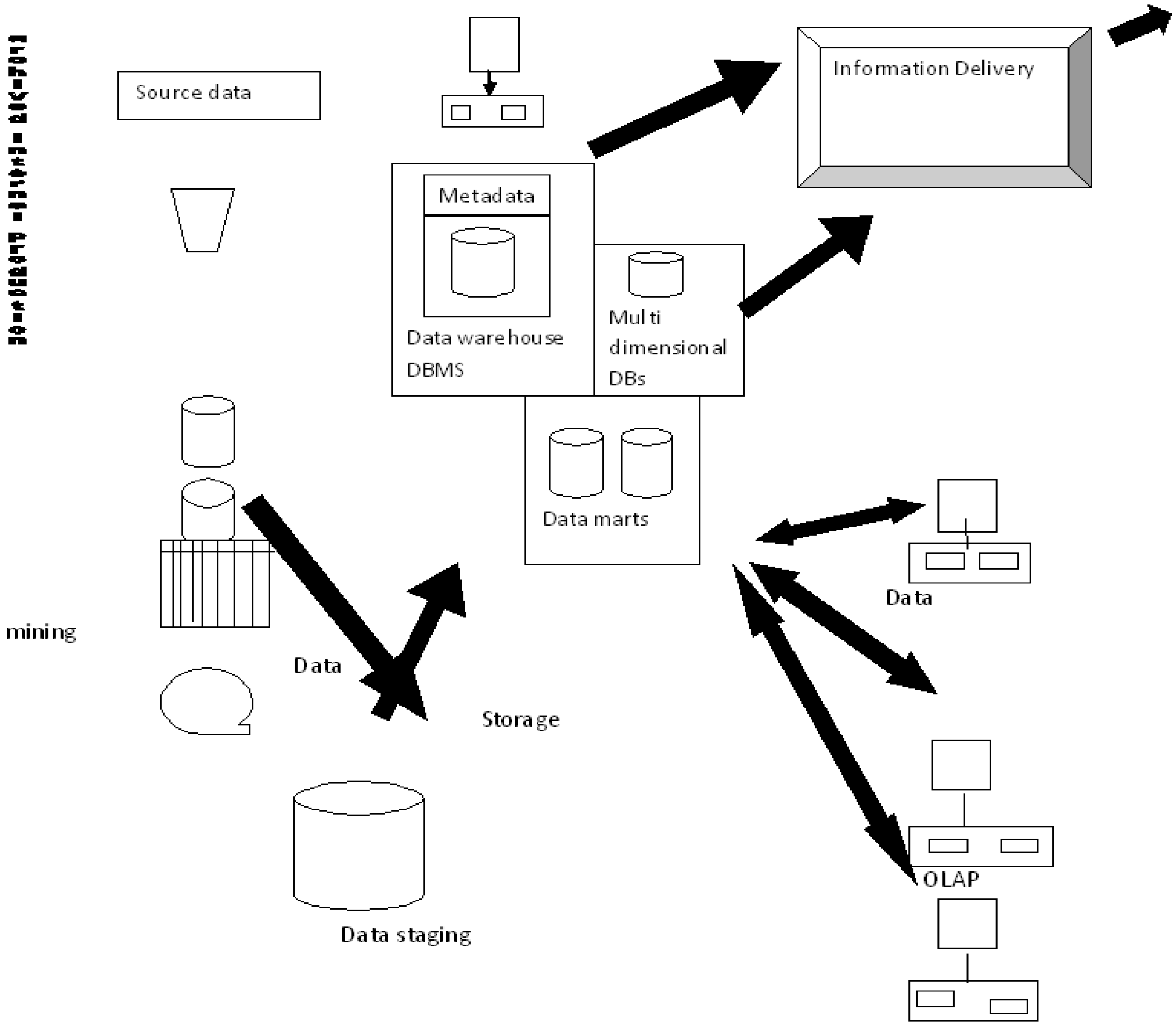


Fig 1.14.Data ware house building blocks or components

Sourcing, acquisition, cleanup and transformation tools:

The data sourcing, cleanup, transformation and migration tools perform all of the conversions, summarizations, key changes, structural changes and condensations needed to transform disparate data into information that can be used by decision support tool. They produce the programs and control statements; including the COBOL programs, MVS job control language (JCL), UNIX scripts, and SQL data definition language (DDL) needed to move data into the data warehouse for multiple operational systems. These tools also maintain the Meta data.

The functionality includes:

- Removing unwanted data from operational databases
- Converting to common data names and definitions
- Establishing defaults for missing data
- Accommodating source data definition changes

The data sourcing, cleanup, extract, transformation and migration tools have to deal with some significant issues including:

- Database heterogeneity. DBMSs e very different in data models, data access language, data navigation, operations, concurrency, integrity, recover etc.

- Data heterogeneity. This is the difference in the way data is defined and used in different models-homonyms, synonyms, unit compatibility, different attributes for the same entity and different ways of modeling same fact

Meta Data

Meta data is data about data that describes the data ware house. It is used for building, maintaining, managing and using the data ware house.
Meta data can be classified into:

- Technical meta data, which contains information about warehouse data for use by warehouse designers and administrators when carrying out warehouse development and management tasks
- Business meta data, which contains information that gives users an easy to understand perspective of the information stored in the data ware house

Equally important Meta data provides interactive access to the users to help understand content &find data. One of the issues dealing with Meta data relates to the fact that many data extraction tool capabilities together Meta data remain fairly immature .Therefore; there is often the need to create a mete data interface for users, which may involve some duplication of effort.

Meta data management is provided via a Meta data repository accompanying software. Meta data repository management software, which typically runs on a work station, can be used to map data to the target database: generate code for data Transformation integrate and transform the data, and control moving data to the warehouse.

The principle purpose of data Ware housing is to provide information to business users for strategic decision-making. These users interact the data warehouse using front-and tools. Tools fall into for main categories: Query and reporting tools, Application development tools, online analytical processing tools, a data mining tools.

Query and reporting tools can be divided into two groups: reporting tools and managed query tools. Reporting tools can be further divided into production reporting tools and report writers production reporting tools let companies generate regular operational paychecks. Report writers, on the other hand, are inexpensive desktop tools designed for end-users managed query tools shield and users from the complexities of SQL and database structures by inserting a Meta layer between users and database. These tools are designed for easy to use, point and click; operations either accept SQL or generate SQL database queries. Often, the analytical needs of the data warehousing user community exceed the built in capabilities of query and reporting tools. in these cases, organizations will often rely on the tried-and-true approach of in-house application development using graphical development environments such as power Builder, visual basic and forte.

These application development platforms integrate well with popular OLAP tools an access all major database systems including Oracle, Sybase, and Informix. LAP tools are based on the concepts of dimensional data models and corresponding databases, and allow users to analyze the data using elaborate, multidimensional views. Typical business applications include product performance and profitability, effectiveness of a sales program or marketing campaign, sales forecasting and capacity planning. These tools assume that the data is organized in a multidimensional model.

A critical success factor for any business today is the ability to use information effectively. Data Mining is the process of discovering meaningful new correlations, pattern and trends by digging into large amounts of data stored in the warehouse using artificial intelligence, statistical and mathematical techniques.

Data marts

Data marts is a data stored that is subsidiary to a data warehouse of integrated data. The data mart is directed at a position of data (often called a subject area) that is created for the use of a dedicated group of users. A data mart might, in fact, be a set of renormalized, summarized or aggregated data. The data mart is a physically separate store of data an is resident on separate database server, often a local area network serving a dedicated user group.

In dependent data marts, data is sourced from the data warehouse, have a high value because no matter how they are developed a how many different enabling technologies are used, different users are all accessing the information views derived from the single integrated version of data.

Unfortunately the misleading statements about the simplicity and low cost of data marts sometimes result in organizations or vendors incorrectly them as an alternative to data warehouse. The view point defines independent data marts that in fact, represent fragmented point solutions to a range of business problems in the enterprise. This type of implementation s should be rarely developed in the contest of an overall technology or application architecture. Moreover, the concept of an independent of data mart is dangerous because as soon as the first data mart is created, other organizations, groups, and the subject areas within the enterprises embark on the task of building their own data marts .As a result, you create an environment where multiple optional systems feed multiple non-integrated data marts that are often overlapping in data content, job scheduling, connectivity and management.

Data warehouse administration and management

Data warehouses tend to be as much as four times as large as relate operational databases, reaching terabytes in size depending on how much history needs to be saved. They are not synchronized in real time to the associated operational data but are updated as often as once a day if the application requires it. In addition, almost all data ware house products include gateways to transparently access multiple enterprise data or sources without having to rewrite applications to interrupt and utilize the data. Furthermore, in a heterogeneous data warehouse environment, the various databases reside on disparate systems, associated systems thus requiring inter-net working tools.

The nee manage this environment is obvious. Managing data warehouse includes security and priority management, monitoring updates from the multiple sources, data quality checks, managing an updating meta data, auditing an reporting data warehouse usage and status, purging data, replicating, sub setting an distributing data, backup and data warehouse storage management.

Information delivery system

The information delivery component is used to enable the process of subscribing for data warehouse information and having it delivered to one or more destinations according to some user-specified and scheduling algorithm. Delivery of information

may be based on time of day on the completion of an external event. The rational for the delivery systems component is based on fact that once the date warehouse is installed and operational, it's users don't view of data at a specific point in time.
In order to provide information to wide community if data warehouse users the information delivery components includes different methods of information delivery.

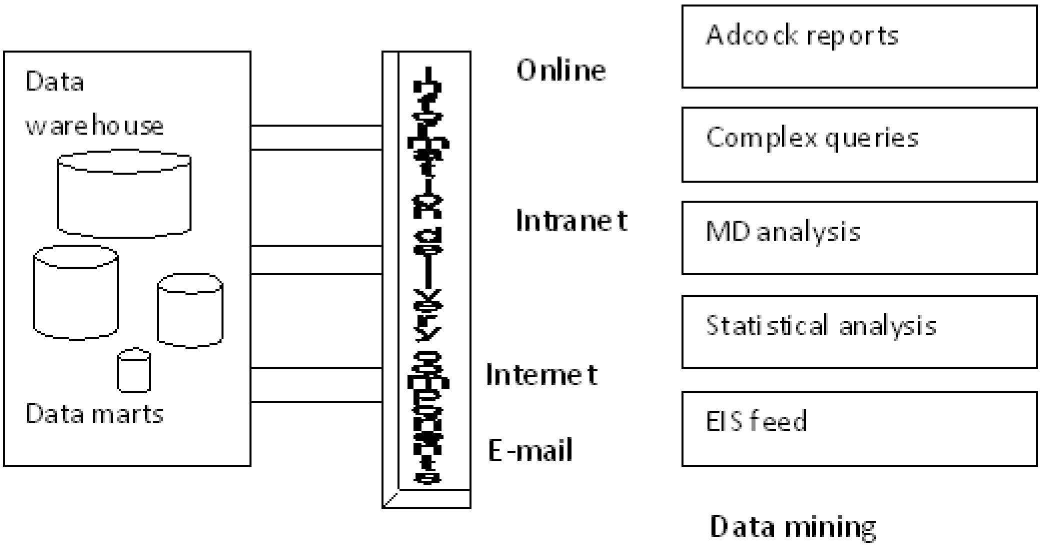


Figure: information delivery components

Ad hoc reports are predefined reports primarily meant for novice and casual users. Provision for complex queries multidimensional (MD) Analysis and statistical analysis cater to the needs of business analysis. Information fed into the executive information systems (EIS) is meant for senior executives and high level managers. Data mining applications helps to discover trends and patterns from the usage of your data.

1.10 MULTIDIMENSIONAL DATA MODEL

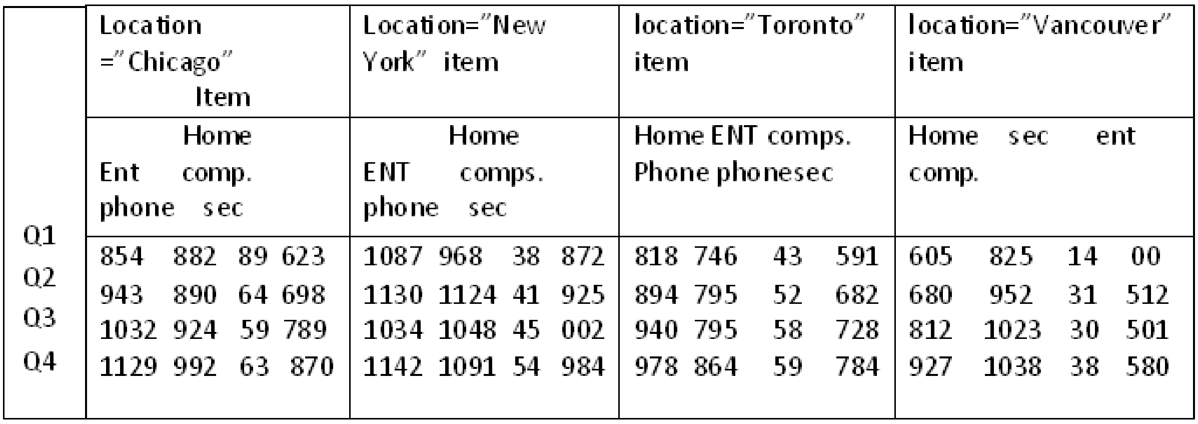
A data warehouse is based on multidimensional data model which views data in the form of data cube.

1.10.1 From tables and spreadsheets to data cubes

- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions.
- Dimensions are perspectives or entities with respect to which an organization wants to keep records such as time, item, branch, location etc.

  -Dimension table, such as item (item name, brand, type), or time (day, week, month, quarter, year) gives further descriptions about dimensions

  - Fact table contains measures (such as dollars _sold) and keys to each of the related dimension tables.
- In data warehousing literature, an n-D base cube is called base cuboids. The top most o-D cuboids, which hold the highest-level of summarization, called the apex cuboids. The lattice of cuboids forms a data cube.
- Table 1.1 A 3-D view of sales data warehouse, according to the dimensions time, item, and location. The measure displayed is dollars sold (in thousands).



	Location ="Chicago" Item	Location="New York" item	location="Toronto" item	location="Vancouver" item
	Home Ent comp. phone sec	Home ENT comps. phone sec	Home ENT comps. Phone phonesec	Home sec ent comp.
Q1	854 882 89 623	1087 968 38 872	818 746 43 591	605 825 14 00
Q2	943 890 64 698	1130 1124 41 925	894 795 52 682	680 952 31 512
Q3	1032 924 59 789	1034 1048 45 002	940 795 58 728	812 1023 30 501
Q4	1129 992 63 870	1142 1091 54 984	978 864 59 784	927 1038 38 580

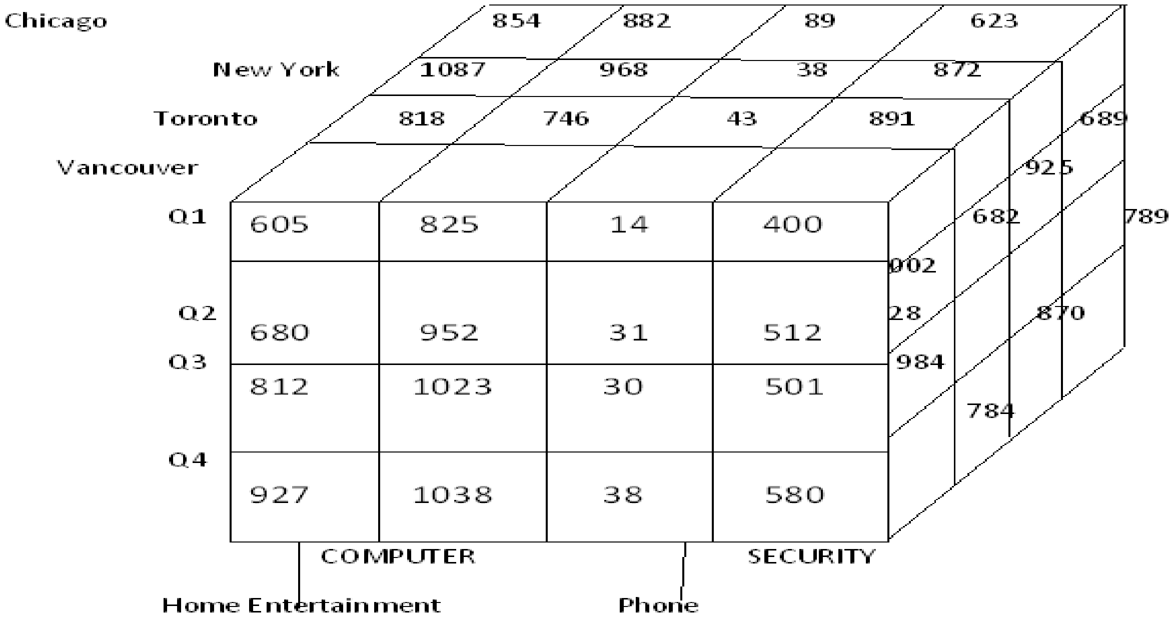
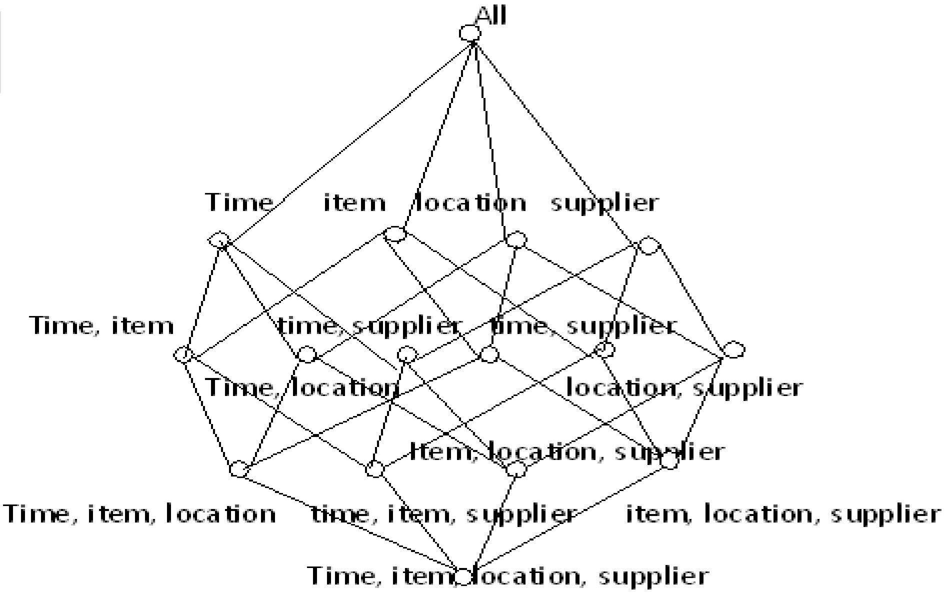


Figure 1.14 A 3-D Data cube representation of the data in table 2.1,

According to the dimensions time, item, and location.



1.10.2 Conceptual modeling of data warehouses

Star schema: A fact table in the middle connected to a set of dimensional tables

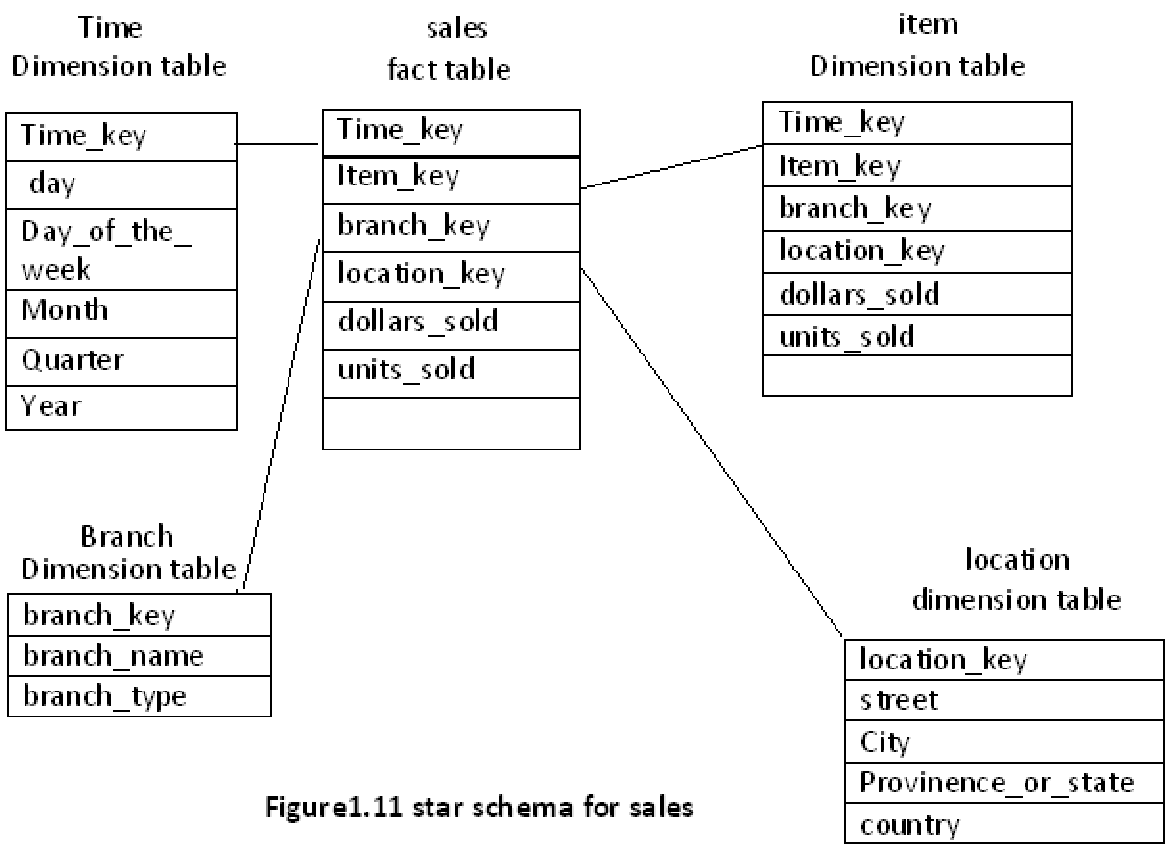


Figure1.11 star schema for sales

Snow flake schema: A refinement of a star schema where some dimensional hierarchy is normalized into set of smaller tables, forming a shape similar snowflake

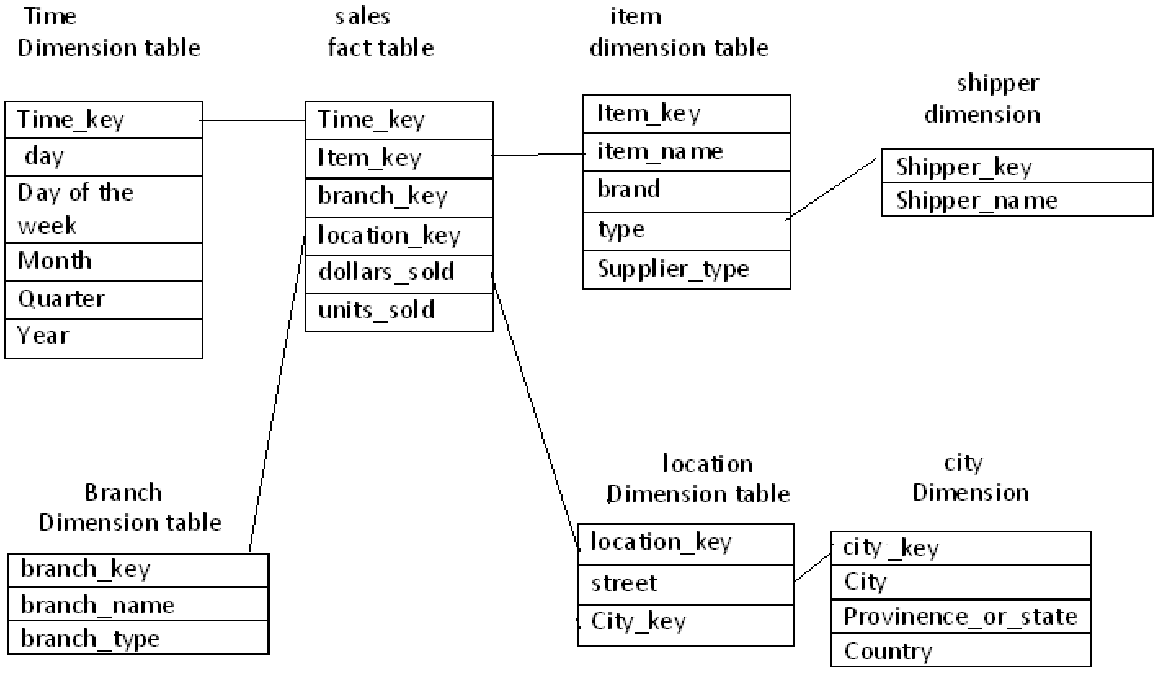


Figure 1.12 Snowflake schemas for sales

Fact constellations: Multiple fact tables share dimension table, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

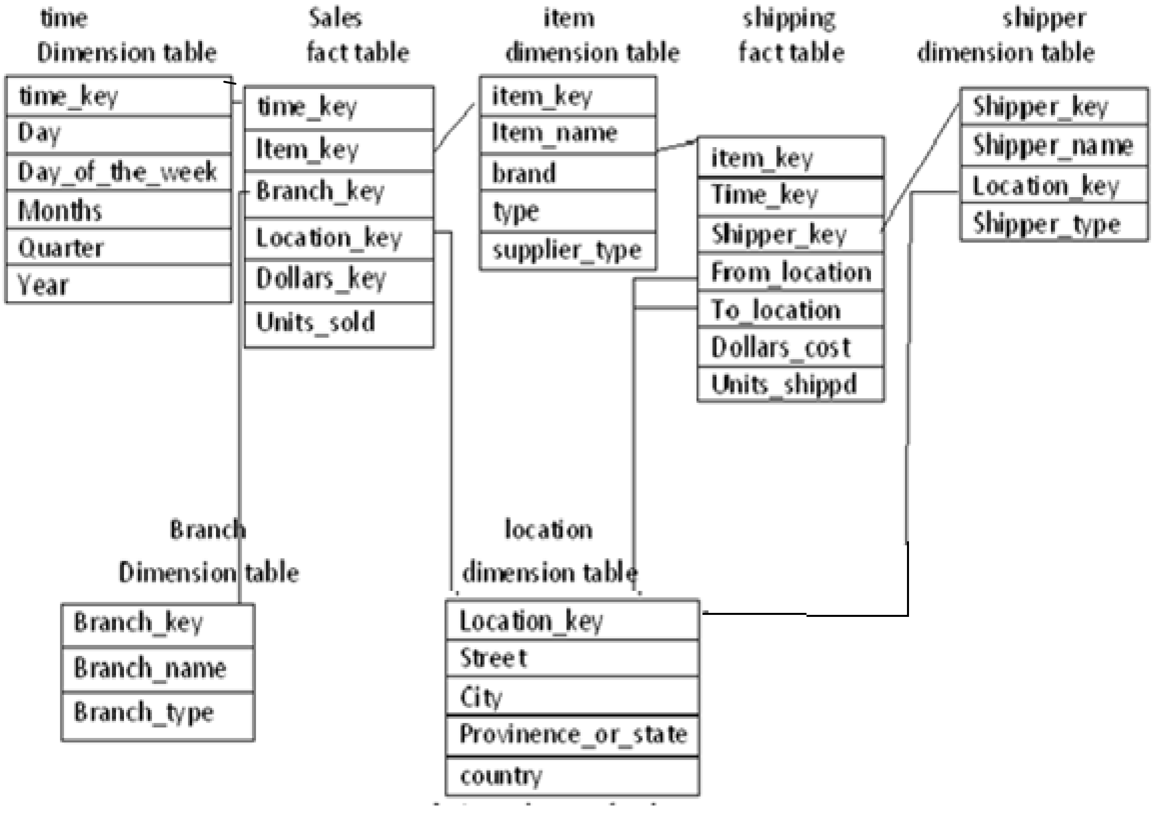


Figure:1.14 fact constellation schema of a data warehouse for sales and shipping

1.10.3 Measures: three categories

- Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all data without partitioning.
  E.g., count (), sum (), min (), max (),
- Algebraic: if it can be computed by an algebraic function with M argument (where M is a bounded integer), each of which obtained by applying a distributive aggregate function.
  E.g., avg (), min_N (), standard_deviation ().
- Holistic: if there is no constant bound on the storage size needed to describe a sub aggregate.
  E.g., median (), mode (), rank ().

1.10.4 A Concept Hierarchy

A concept hierarchy defines a sequence of mapping from a set of low-level concepts to higher-level, more general concepts.

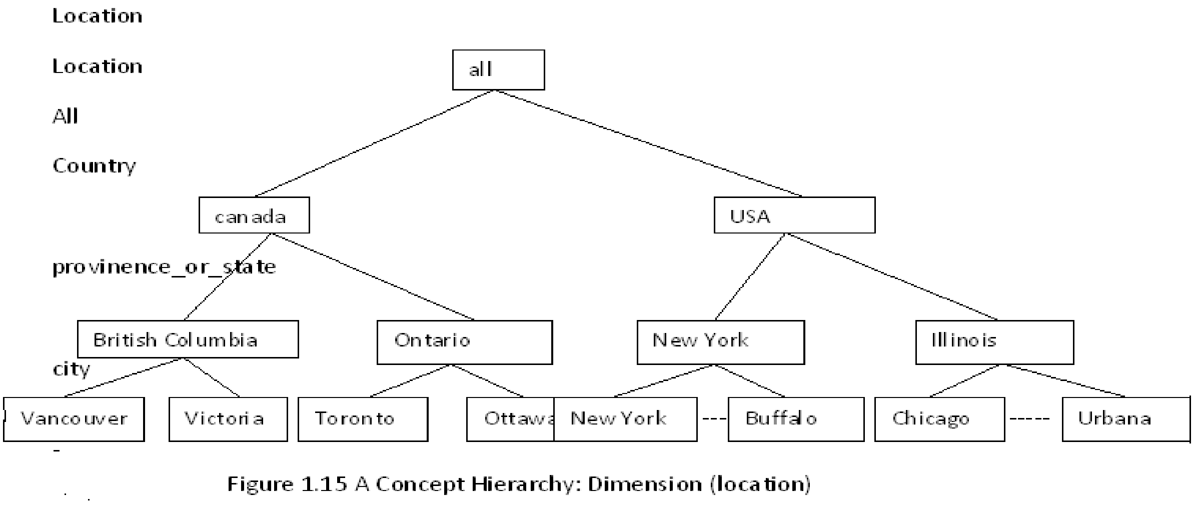


Figure 1.15 A Concept Hierarchy: Dimension (location)

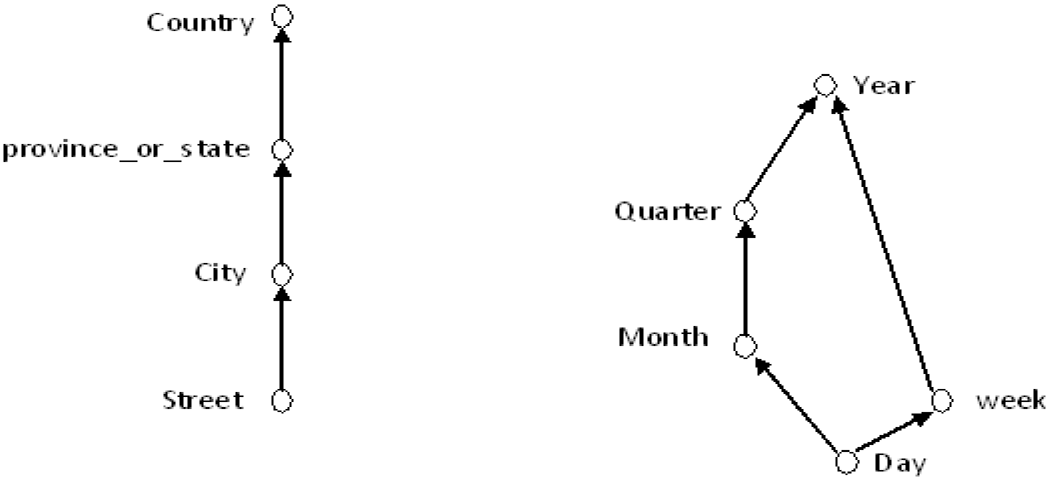


Figure 1.16 (a) A hierarchy for location; (b) A lattice for time.

1.10.5 OLAP Operations

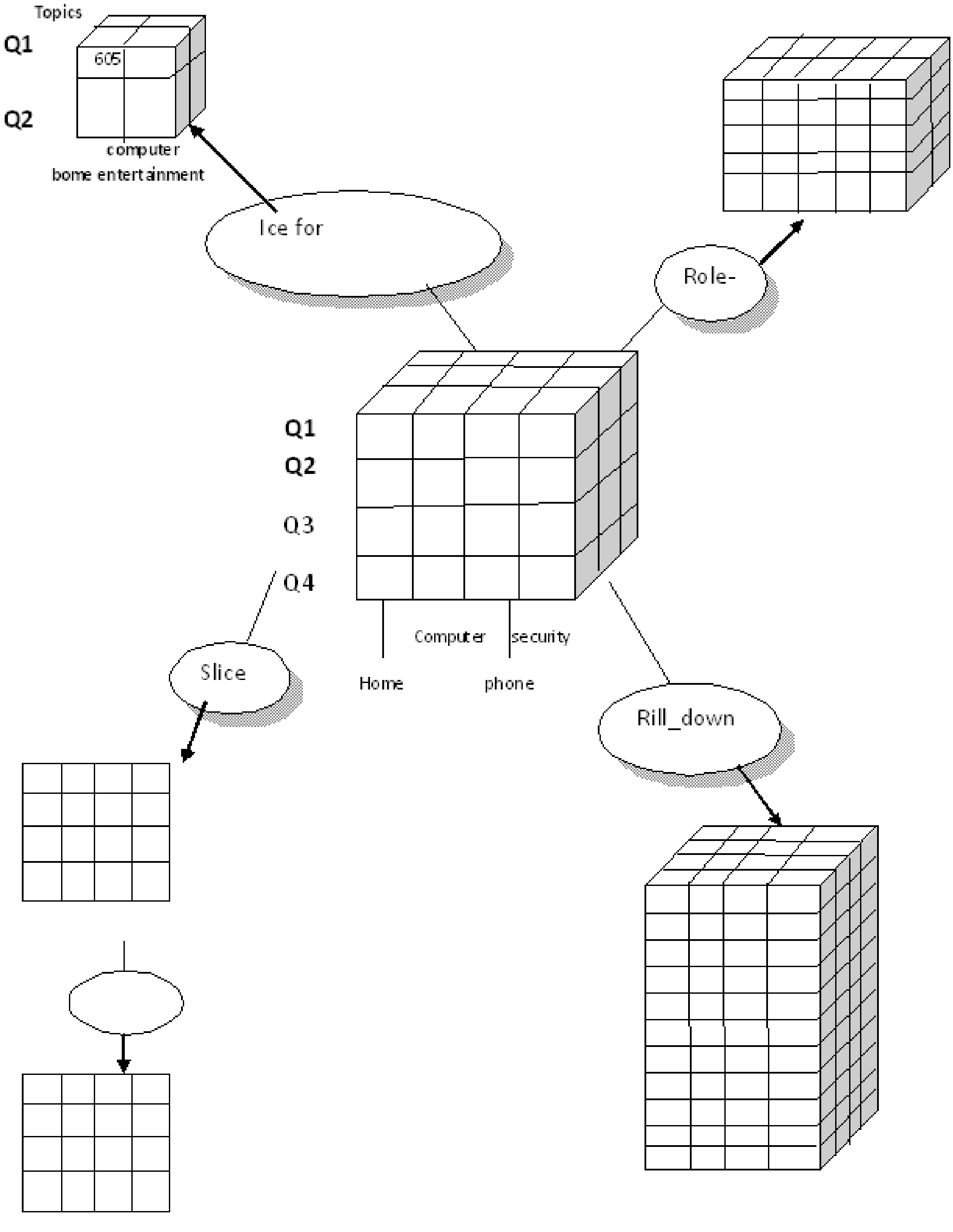


Figure 1.15 Examples of typical OLAP operations on multidimensional data.

- Roll up (rill-up): summarize data
  - By climbing up hierarchy or dimension reaction
- Drill down (roll down): reverse of roll-up
  - From higher level summary to lower level summary or detailed data, or introducing new dimensions.
- Slice and dice:
  - Project and select

- Pivoted(rotate):
  - Reorient the cube, visitation, 3D to series of 2D planes.
- Other operations
- Drill-within: It is switching from one classification to different one within the same dimension.
  - Drill across: involving (across) more than one fact table
  - Drill through: through bottom level of the cube to its back-end relational tables (using SQL)

1.10.6 A Star-Net Query Model

A star net model consists of radical 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.

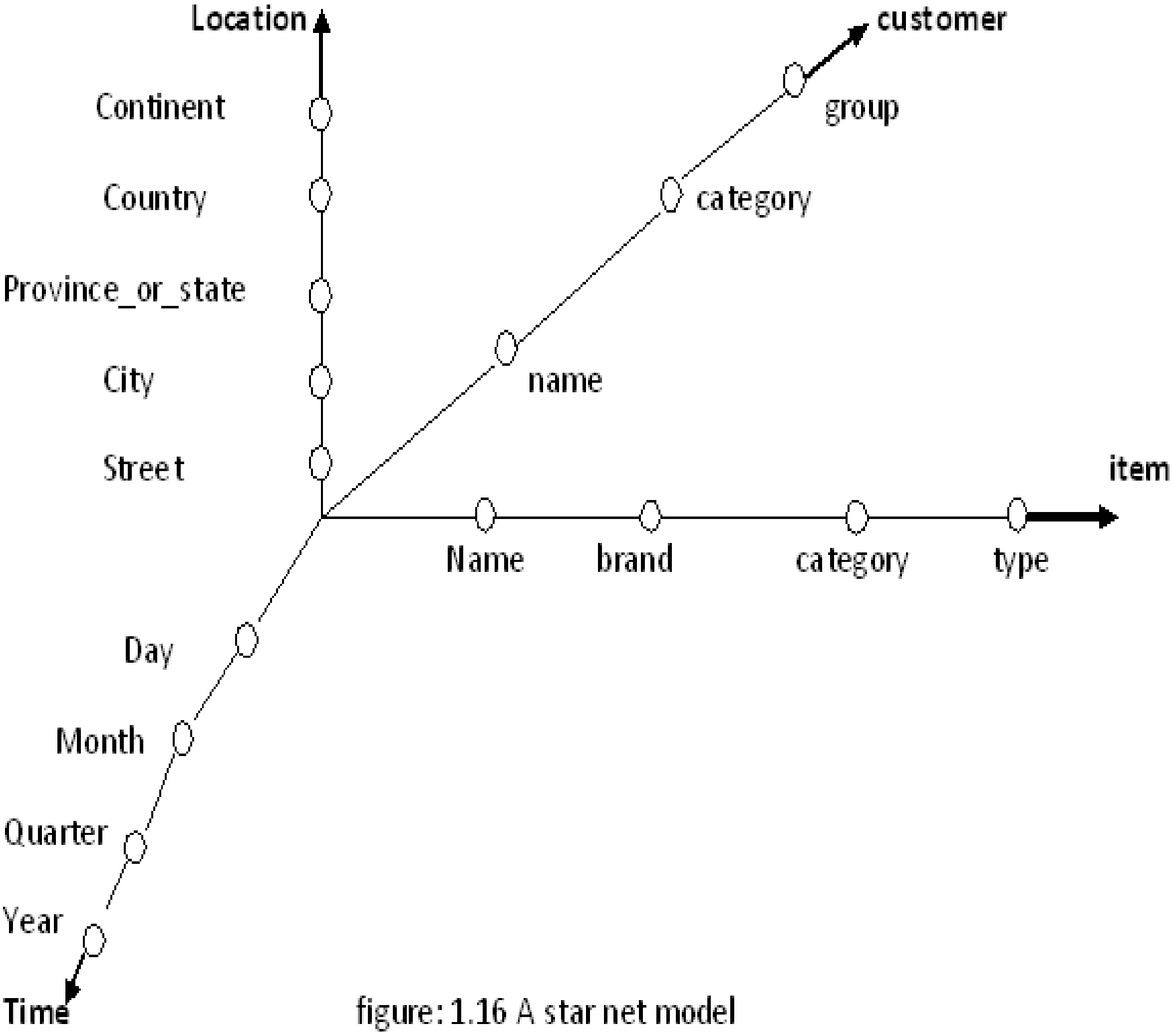


figure: 1.16 A star net model

1.11 Data warehouse Architecture

Steps for the Design and construction of Data warehouse are:

- Design of Data warehouse : A business analysis framework
- Data warehouse design process

Design of a data warehouse: a Business Analysis Framework

To design an effective data warehouse we need to understand and analyze business needs and construct a business framework. The construction of a large and complex information system can be viewed as the construction of a large and complex building, for which the owner, architect, a builder have different views. These views are combined to form a complex framework that represents the top-down, business-driven, or owner's perspective, as well as the bottom-up, builder-driven, or implementer's view of the information system.

Four views regarding the design of a data warehouse

- Top-down view
- Allows selection of the relevant information necessary for the data warehouse
- Data source view
  - Exposes the information being captured, stored, and managed by operational systems
- Data warehouse view
  - Consists of fact tables an dimensional tables
- Business query view
  - Sees the perspective of data in the warehouse from the view of end-user

Data warehouse Design process

- Top-down, Bottom-up approaches or a combination of both
- Top-down: Starts with overall Design an planning (mature)
- Bottom-up: Starts with experiments an prototypes(rapid)
  - From software engineering point of view
- Water fall: Structured an systemically analysis at each step before proceeding to the next
- Spiral: Rapid generation of increasingly functional systems, short turn around time, quick turn around
  - Typical data warehouse design process
- Choose a business process to model, e.g., orders, invoices, etc.
- Choose the grain (atomic level of data) of the business process
- Choose the dimensions that will apply to each fact table record.
- Dimension the measure that will populate each fact table record.

Multi-Tiered Architecture

Data warehouse often adopt a three-tier architecture

1. The bottom tier is a warehouse database server that us almost always 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 (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning an transformation. This tier also contains a metadata respiratory, which stores information about the data warehouse and its contents
2. The middle tier is an OLAP server that is typically implemented using either(1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implemented multidimensional data and operations.

3. The top tier is a front-end client layer, which contains query and reporting, analysis tools, and/or data mining tools (e.g. trend analysis, prediction, and so on).

Three Data Warehouse Models

- Enterprise warehouse

- Collects all of the information about subjects spanning in the enterprise the entire organization.

- Data Mart

- A subset of corporate-wide data that is of value to a specific groups, such as marketing
  Data mart.

- Virtual warehouse

- A set of views over operational database.
- Only some of the possible summary views may be materialized.

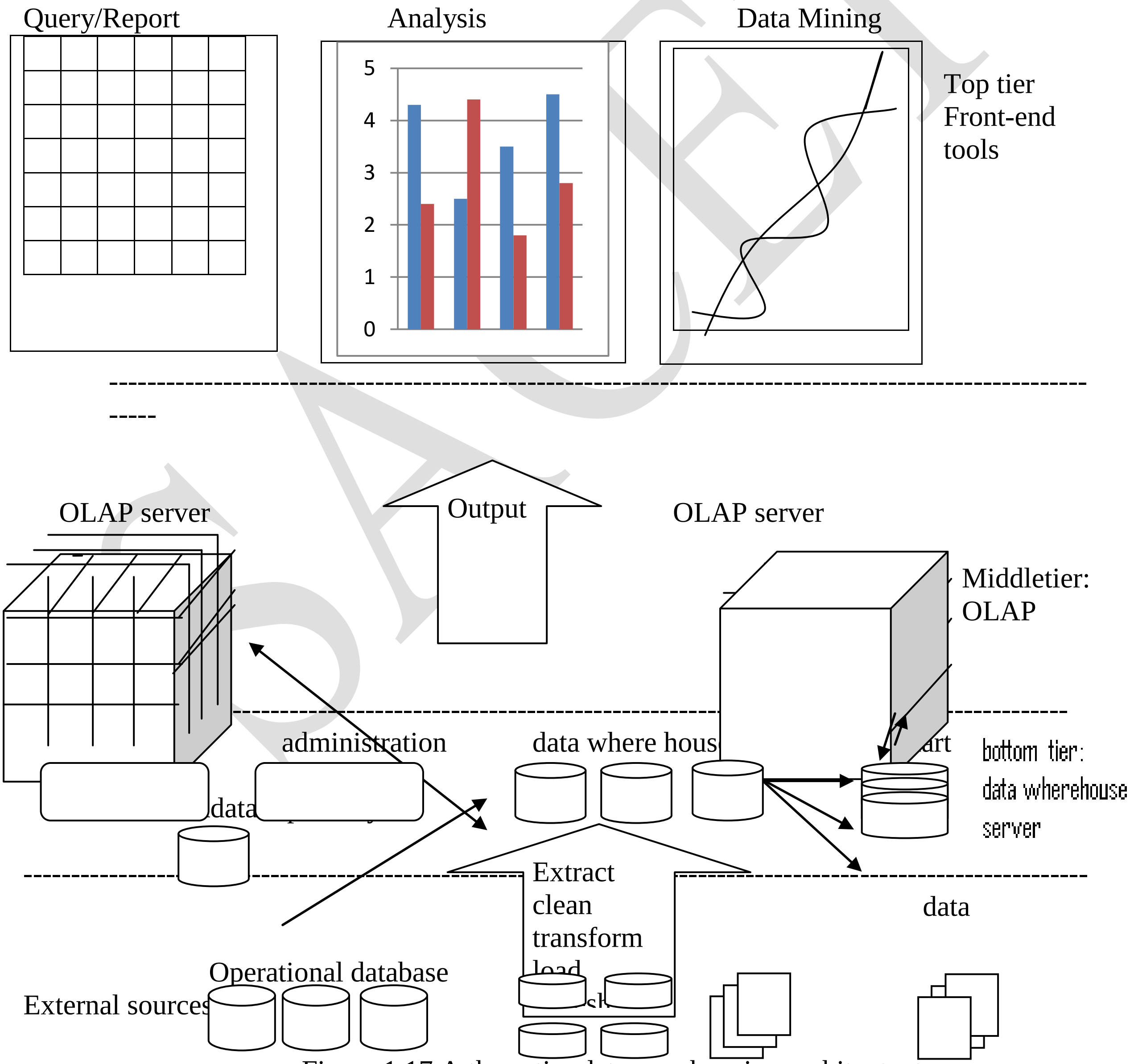


Figure 1.17 A three-tier data warehousing architucture

Three Data Warehouse Models

- Enterprise warehouse

- Collects all of the information about subjects spanning in the enterprise the entire organization.

- Data Mart

- A subset of corporate-wide data that is of value to a specific groups, such as marketing
  Data mart.

- Virtual warehouse

- A set of views over operational database.
- Only some of the possible summary views may be materialized.

Data warehouse Development: A Recommended Approach

A recommended method for the development of data warehouse systems is to implement the data warehouse in an increment and evolutionary manner. First, a high-level corporate data model is defined within is defined within a responsibly short period(such as one or two months) that provides a corporate-wide, consistent, intenerated view of data among the different subjects and potentials uses. Second, independent data mart can be implemented in parallel with the enterprise warehouse based on the same corporate data model such as above. Third, distributed data mart can be constructed to integrate different data marts via hub servers. Finally, a multitier data warehouse is connected to where the enterprise warehouse is the soul for custodian of all warehouse data, which is then distributed to the various dependent data marts.

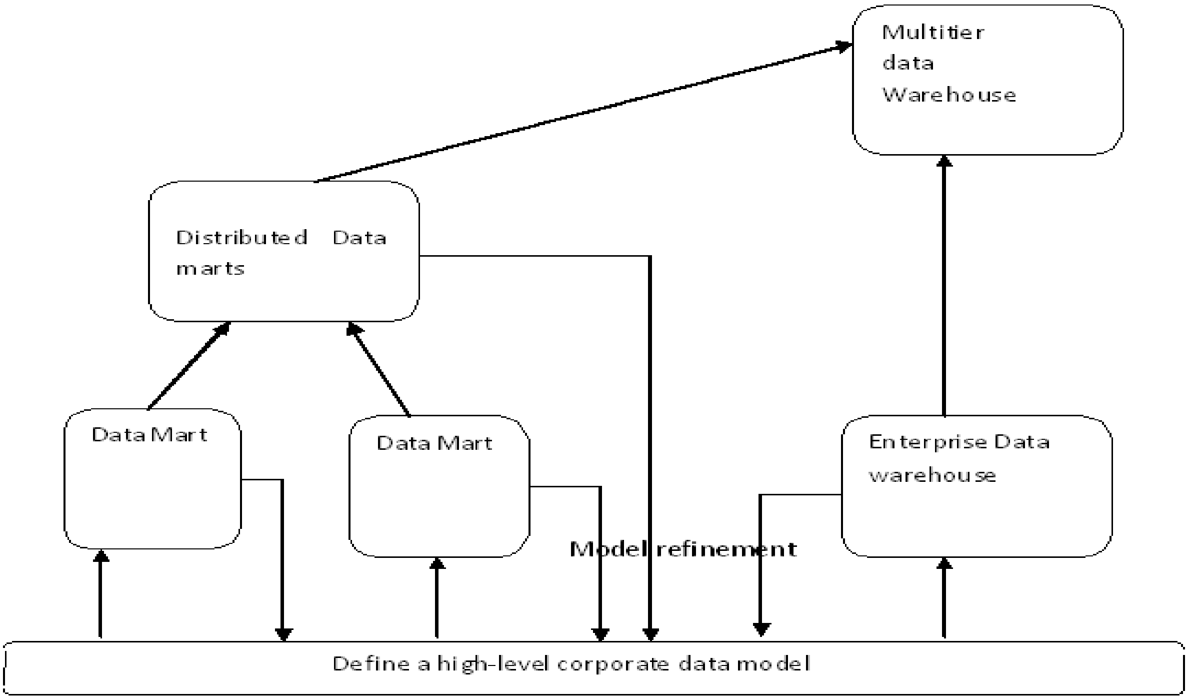


Figure 1.16 a recommended approach for data warehouse development.

OLAP Server Architectures

- Relational OLAP (ROLAP)

- Use relational or extended-relational DBMS to store and manage warehouse data

And OLAP middle ware to support missing pieces

  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Array-based multidimensional storage engine(sparse matrix techniques)
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP)
  - User flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers
  - Specialized support for SQL queries over star/snowflake schemas

1.12 DATA WAREHOUSE IMPLMENTATION

Data warehouse contain huge volumes of data. OLAP servers demand that decision support queries be answered in the order of seconds. Therefore, it is crucial for data warehouse systems to support highly efficient cube computation techniques, access methods, and query processing techniques.

Methods for the efficient Implementation of data warehouse systems are :

1.12.1 Efficient computations of data cubes

At the core of multidimensional data analysis is the efficient computation of aggregations across many sets of dimensions. In SQL terms, these aggregations are referred to as group by's. Each group by can be represented by a cuboids, where the set of group by's forms a lattice of cuboids defining a data cube.

The compute cube operator and its implementation

The compute cube operator aggregates overall subsets of the dimensions specified in the operation. This can require excessive storage space, especially for large numbers of dimensions.

An SQL query containing no group-by, such as "compute the sum of total sales," is a zero-dimensional operation. An SQL query containing one group-by, such as "compute the sum of sales, group by city," is a one-dimensional operation. A cube operator on n dimensions is equivalent to a collection of group by statements, one for each subject of then n dimensions. Therefore, the cube operator is the n-dimensional generalization of the group by operator.

Using DMQL, data cube is defined as:

Define cube sales cube [city, item, and year]: sum (sales in dollars)

For a cube with n dimensions and no hierarchies associated with each dimension, then the total number of cuboids is 2"

For an n-dimensional data cube associated with hierarchies, the total number of cuboids that can be generated (including the cuboids generated by climbing up the hierarchies along each dimension) is

$$\text{Total number of cuboids} = \prod_{I=1}^{n} (Li+1)$$

Where Li is the number of levels associated with dimension I

Partial materialization: Selected Computation of Cuboids

There are three choices for data cube materialization given base cuboids:

1. No materialization: Do not recompute any of the "nonbiased" cuboids. This leads to computing expensive multidimensional aggregates on the fly, which can be extremely slow.

2. Full materialization: Recomputed all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice typically requires huge amounts of memory space in order to store all of the recomputed cuboids.
3. Partial materialization: Selectivity compute a proper subset of the whole set of possible cuboids. Partial materialization represents an interesting trade-off between storage space and response time.

The partial materialization of cuboids or sub cubes should consider three factors:

1. Identify the subset of cuboids or sub cubes to materialize;
2. Exploit the materialized cuboids or sub cubes during query processing
3. Efficiency updates the materialized cuboids or sub cubes during load and refresh.

1.12.2 Index OLAP Data

To facilitate efficient data accessing, most data warehouse system support index structures and materialized views (using cuboids).

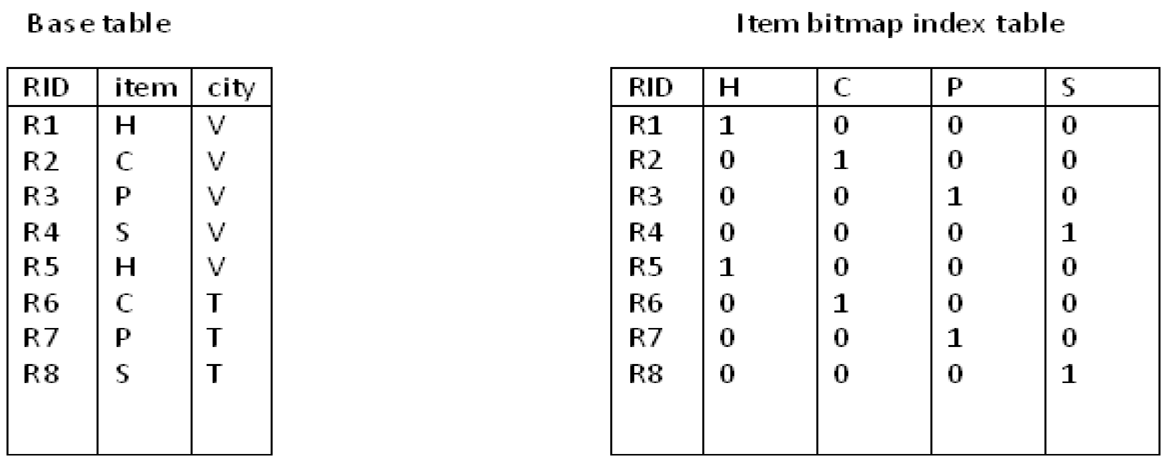
Type of indexing OLAP Data

The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes. The bitmap index is an alternative representation of the record ID (RID) list. In the bitmap index for a given attribute, there is a distinct bit vector $B_{y.}$ for each value v in the domain of the attribute. If the domain of the given attributes consists of n values, then n bits are needed for each entry in the bit image index (i.e. there are n bits vector). If the attributes has the values v for a given row in the data table, then the bit representing that values is set to 1 in the corresponding row of the bit image index. All other bits for that row are set to 0.

Advantages

- Useful for low cardinality domains.
- Reduction in apace a processing time.

Base (data) table containing the dimensions items and city, and its mapping to bitmap index table of the dimensions are given below.

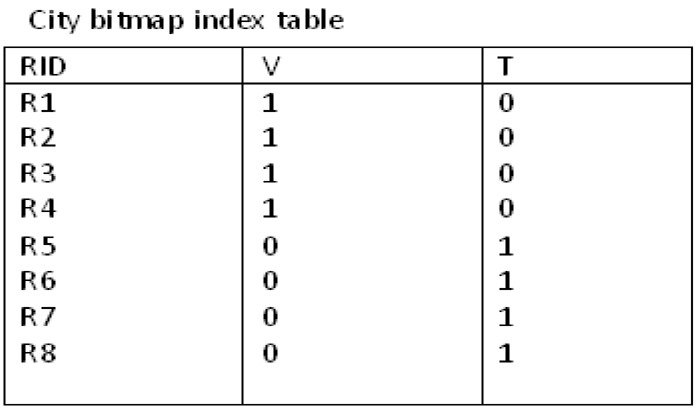


Base table

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

Item bitmap index table

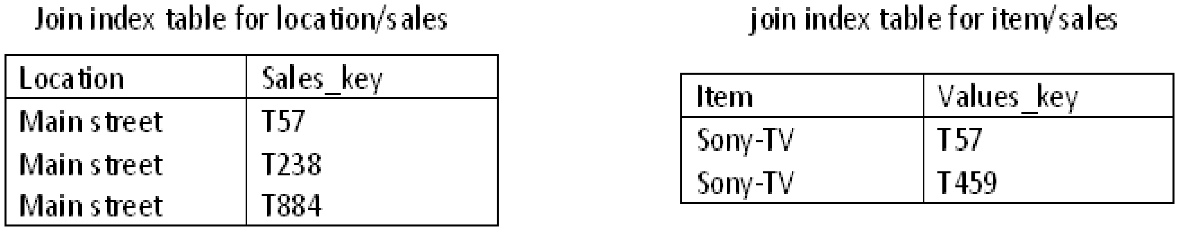
RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1



City bitmap index table

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

THE JOIN INDEXING method gained popularity from its use in relational data base query processing. Join indexing registers the joinable rows relation data base. For example, if two relation R(RID ,A) and S(B,SID) join on the attributes A and B, then the join index record contains the pair (RID,SID),where RID and SID are record identifiers from the R and S relations, respectively. Hence, the join index records can identify joinable topples without performing costly join operations. Join indexing is useful for maintaining the relationship between a foreign key and its matching primary keys, form the joinable relation.

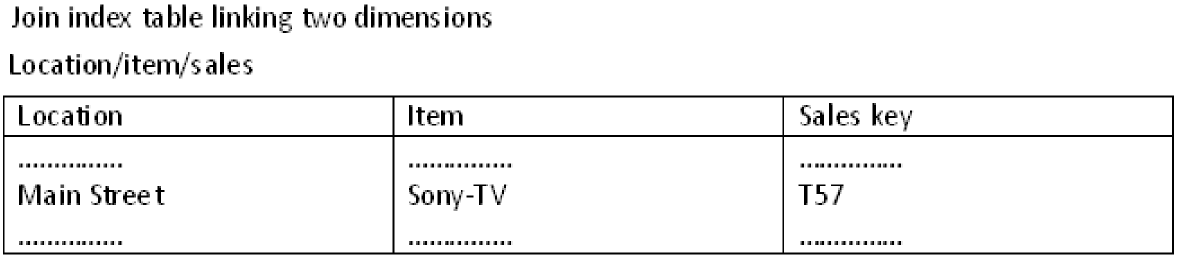


Join index table for location/sales

Location	Sales_key
Main street	T57
Main street	T238
Main street	T884

join index table for item/sales

Item	Values_key
Sony-TV	T57
Sony-TV	T459



Join index table linking two dimensions

Location/item/sales

Location	Item	Sales key
……………	……………	…………..
Main Street	Sony-TV	T57
……………	……………	…………..

Figure:1.19 join index table based on the linkages between the sales fact table

And dimensions table for location and item

1.12.3 Efficient processing of OLAP queries

The purpose of materializing cuboids and constructing OLAP index structure is to speed Up query processing in data cubes. Given materialized views, query processing should proceed as follows;

1. Determine which operation should be performed on the available: this involves transforming any selection, projection, roll-up(group-by), and drill –down operations specified in the query into corresponding SQL and/or OLAP operations.

1.12.4 Metadata Repository

- Meta data is the data defining warehouse objects .it has the following kinds.
- Describing of the structure of the warehouse.
- Schema, view ,dimensions ,hierarchies, derived data define ,data mart locations and contents
- Operational meta data
- Data linkage (history of migrated data and transformational path), currency of data (activ
- e, archived, or purged), monitoring informational (ware house usage statistics ,error retrials)
- The algorithms used for summarization.
- The mapping from operational environmental to the data ware house.
- Data related to system performance.
- Warehouse schema, view and derive data warehouse.
- Business data
- Business terms and definitions, ownership of data, charging policies.

1.12.5 Data ware house back-en tools and utilities

Data ware house systems use back-end tools and utilities to populate an refresh their data. These tools and utilities include the following functions:

Data extractions- gather data from multiple, heterogeneous, and external sources. Data Cleaning- Detects errors in the data and rectifies them when possible

Data transformation-converts data from legacy or host format to ware house format.

Data Loading –sorts, summaries, consolidates, computes views, checks integrity, and builds indices and partitions.

Refresh- propagates the updates from the data sources to the ware house

1.13 From data warehousing to data mining

1.13.1 Data ware house usage

Business executives use the data in data ware houses and data marts to perform data analysis and make strategic decisions. In many firms, data warehouses are used as an integral part of a *plan-execute-assess* “close-loop” fed back system for enterprise management. Initially, the data house is mainly used for generating reports an answering pre defined queries. Then, it is used to analyze summarized and detailed data, presented in the form of reports and charts. Later, the data ware house is used for performing multi-dimensional analysis and sophisticated slice-and –dice operations. finally, the data ware house may be employed for knowledge discovery and strategy decision making using data mining tools .in this context the tools for data where housing can be categorized in to access an retrieval tools , database reporting tools, data analysis tools and data mining tools.

There are three kinds of data warehouse applications:

- information processing support querying, basic statistical analysis, and reporting using crosstabs, tables, charts or graphs
- analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, pivoting
- data mining supports knowledge discovery by finding hidden patterns and associations, constricting analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Relation between information processing an on-line analytical processing

Online analytical (OLAP) and data mining are integral parts of any decision support process. However the two tools have a fundamental difference in terms of the direction of the query. Data mining is driven by data while OLAP is driven by the user or the user's intention to verify his or her queries. OLAP tools provide multi dimensional data analysis that is, they allow data to be broken down and summarized (such as by regional sales). For example, OLAP typically involves the summation of multiple databases into highly complex tables. Data mining, on the other hand, is about ratios, patterns, and influences in a data set. As such, data mining is division. This is not to say that both OLAP and data mining should not be used in conjunction to gain a powerful insight in to your company databases, customer information file, data marts and data warehouse. In fact, aggregate an inductive analysis can complement each other. For example, a data mining analysis can discover a significant relationship in a set of attributes. OLAP can then expand on this and generate a report detailing the impact of discovery.

1.13.2 from on-line Analytical Processing to on-line analytical mining

On-line analytical mining (OLAM) (also called OLAP mining) integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional data bases.

Advantages of OLAM:

- high quality
- quality of data in data where houses
- available information processing infrastructure surrounding data where houses
- OLAP-based exploratory data analysis
- On-line selection of data mining functions

Architecture for on-line data analytical mining

An OLAM server performs analytical mining in data cubes in a similar manner as an OLAP server performs on-line analytical processing. The OLAM and OLAP servers both accept user on-line analytical processing. The OLAM and OLAP servers both accept user on-line queries (and commands) via a graphical user interface API and work with the tube in the data analysis via a cube API. A metadata director is used to guide the access of data cube. The data cube can be constructed by accessing and /or integrating multiple databases via an MDDB API and /or by filtering a data warehouse via a database API that may support OLE DB or ODBC connections. Since an OLAM server may perform multiple data mining tasks, such as concept description, association , classification, prediction, clustering, time-series analysis, and so on, it usually consist of multiple integrated data mining modules and is more sophisticated than an OLAP server.

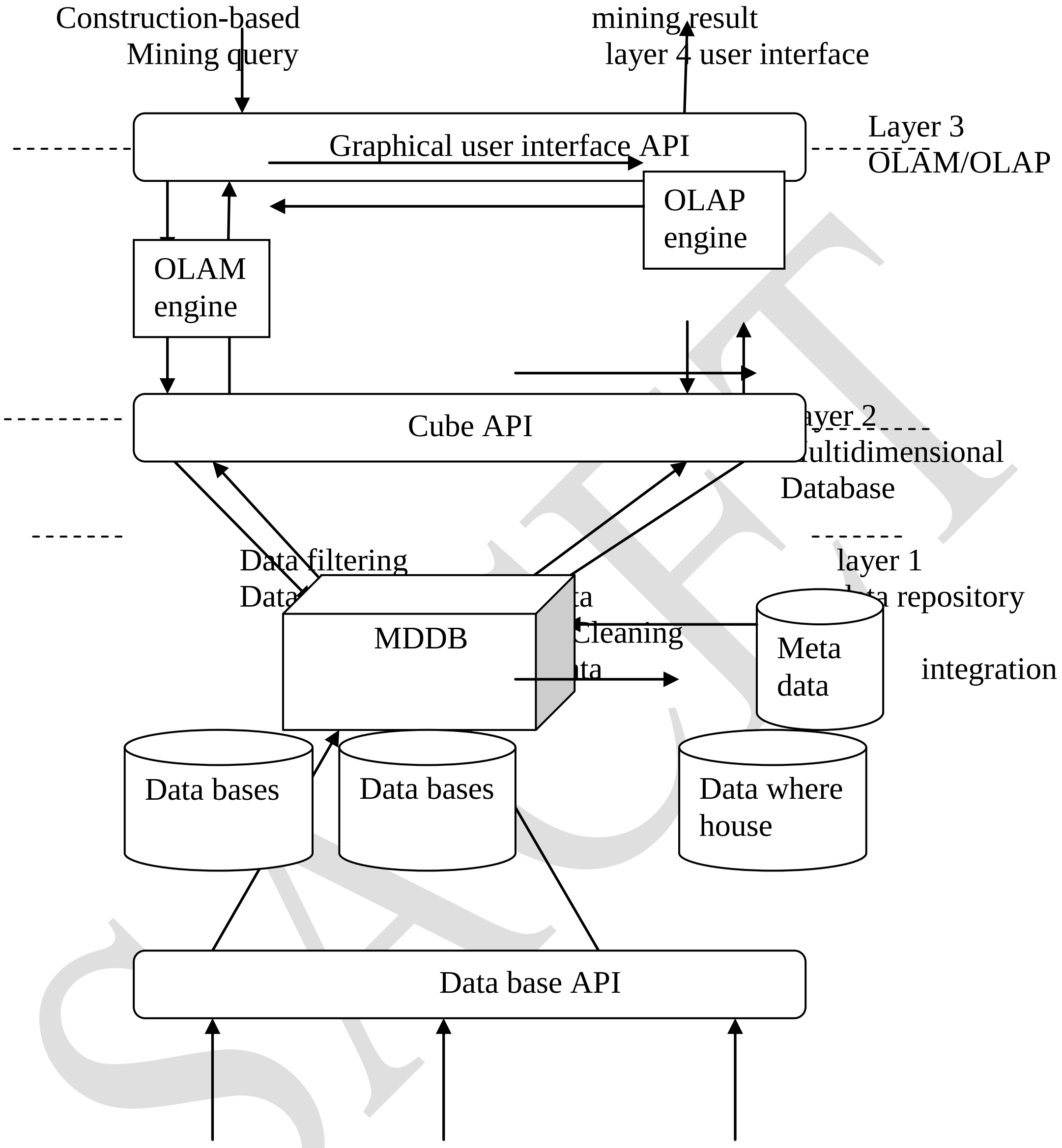


Figure 1.18 an integrated OLAM and OLAP architecture

1.14 ON-LINE ANLITICAL PROCESSING (OLAP)

1.14.1 Need for OLAP

On-line analytical processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

OLAP functionality is characterized by the dynamic multi-dimensional analysis of consolidated enterprise data supporting end user analytical and navigational activities including:

- Calculations and modeling applied across dimensions, through hierarchies and/or across members
- Trend analysis over squentional time periods
- Slicing subsets for on-screen viewing
- Drill-down to deep err levels of consolidation
- Reach-through to underlying detail datprisa
- Rotation to new dimensional comparisons in the viewing area

OLAP is implemented in a Multi-user client/server mode and offers consistently rapid response to queries, regardless of data base size and complexity. LAP helps the user synthesize enterprise information through comparative personalized viewing as well as through analysis projected data in various “what-if” data model scenarios. This is achieve through use of an OLAP sever.

OLAP SERVER

An OLAP server is a high-capacity, multi-user data maniplication engine specifically designed to support and operate on multy-dimsional data structures. A multy-dimentional structure is arranged so that every data item is located and accessed based on the intersection of the dimension members which define that item. The design of the server and the structure of the data are optimized for rapid ad-hoc information retrieval in any orientation, as well as for fast, flexible calculation and the transformation of raw data base on formulaic relationships. The OLAP server may either physically stage the processed multi-dimensional information to deliver consistent and rapid response times to end users, or it may populate its data structures in real-time from relational or other databases, or offer a choice of both. Given the current state of technology and the end user requirement for consistent and rapid response times, staging the multi-dimensional data in the OLAP server is often the performed method.

1.14.2 Categorization of OLAP tools

In the OLAP world, there are mainly two different types: multi-dimensional OLAP (MOLAP) and relational OLAP (ROLAP). Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.

MOLAP:

This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multi-dimensional cube. The storage is not in the relational database, but in proprietary formats.

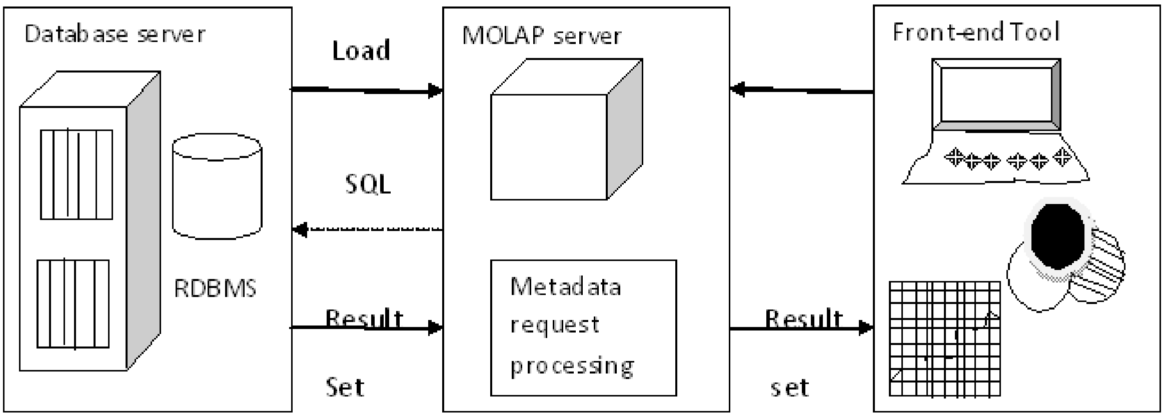


Figure 1.18 MOLAP architecture

Advantages:

- Excellent performance; MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
- Can perform complex calculations: all calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return
- quickly.

Disadvantages:

- Limited in the amount data it can handle: Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in cube itself.
- Requires additional investment: cube technologies are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.

ROLAP :

This methodology relies on manipulating the in the stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a “WHERE” clause in the SQL statement.

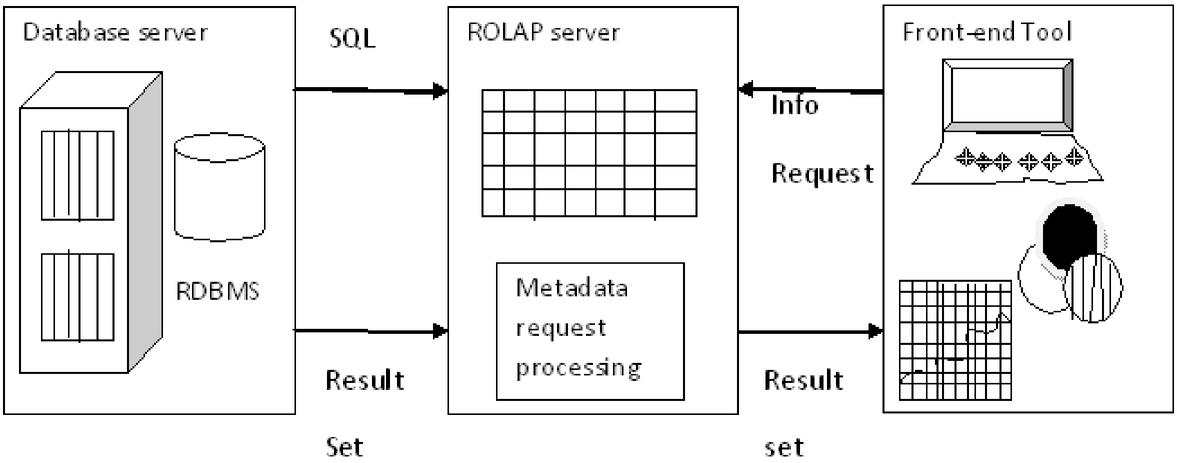


Figure 1.18 ROLAP architecture

Advantages:

- Can handle large amount of data: The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.
- Can leverage functionalities inherent in the relational database: often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages:

- Performance can be slow: because ROLAP technology mainly relies on generating SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
- Limited by SQL functionalities: because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs (for example, it is difficult to perform complex calculations using SQL),ROLAP technologies are therefore gravitationally limited by what SQL can do. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability into allow users to define their own functions.

HOLAP:

Hybrid online analytical processing (HOLAP) is a combination of relational OLAP (ROLAP) and multidimensional OLAP (usually referred to simply as OLAP). HOLAP was developed to combine the greater data capacity of ROLAP with the superior processing capability of OLAP.

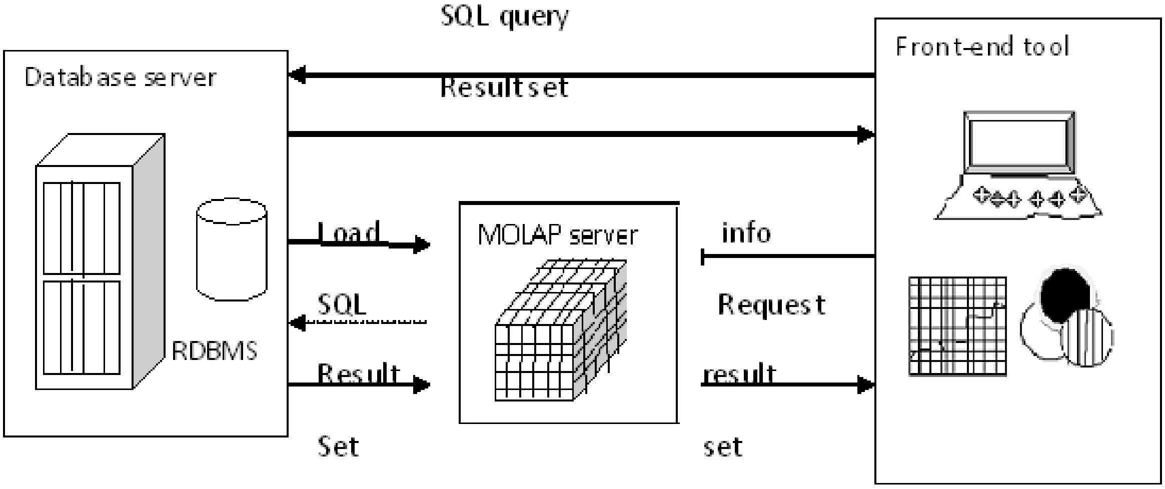


Figure 1.19 HOLAP architecture

HOLAP can use varying combinations of ROLAP and OLAP technology. Typically it stores data in a both a relational database (RDB) and multidimensional database (MDDB) and uses whichever one is best suited to the type of processing desired. The databases used to store data in the most functional way. For data-heavy processing, the data is more efficiently stored in a RDB, while for speculative processing; the data is more defectively stored in an MDDB.

HOLAP users can choose to store the results of queries to the MDDB to save the effort of looking for the same data over and over which saves time. Although this technique- called “materializing cells” – improves performance, it takes a tool on storage. The user as to strike a balance between performance and storage demand to get the most out of HOLAP nevertheless, because it offers the best features of OLAP and ROLAP, HOLAP is increasingly preferred.