Lesson 3: Dimensional Data Warehouse

CONTENTS

Objectives

Introduction

3.1-Dimensional Model

3.2 Facts Table

3.2.1 Types of Measure

3.2.2 Types of Fact Table

3.3 Dimension Tables

3.4 Surrogate Keys and Alternative Table Structure

3.4.1 Advantages of Surrogate Keys

3.4.2 Disadvantages of Surrogate Keys

3.4.3 Alternative Tables used in Data Warehousing

3.5 Multidimensional OLAP

3.5.1 MOLAP

3.5.2 ROLAP

3.5.3 HOLAP

3.6 Summary

3.7 Keywords

3.8 Review Questions

3.9 Further Readings

Objectives

After studying this unit, you will be able to:

Describe about Dimensional Model

Construct Facts Table

Demonstrate Dimension Tables

Discuss about Surrogate Keys and Alternative Table Structure

Explain Multidimensional OLAP

Introduction

Dimensions are a common way of analysing data. Dimension model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data. Dimensional data modelling is the preferred modelling technique in a BI environment. Knowing the basics of data warehousing and dimensions helps you design a better data warehouse that fits your reporting

needs. This unit on data warehousing dimensions explains the importance of dimensions and dimension granularity and stresses the importance of flattening hierarchies—with the goal being to make data more accessible and useful to users. It also focuses on fact and dimension table.

3.1-Dimensional Model

Dimensional model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data. Since Business Intelligence reports are used in assessing the facts (aggregates) across various dimensions, dimensional data modelling prefer the modelling technique in a BI environment.

Facts are normally calculated data like dollars' worth or Sales or income. They correspond to the aim of a conclusion support analysis.

Dimensions define the axis of enquiry of a fact.

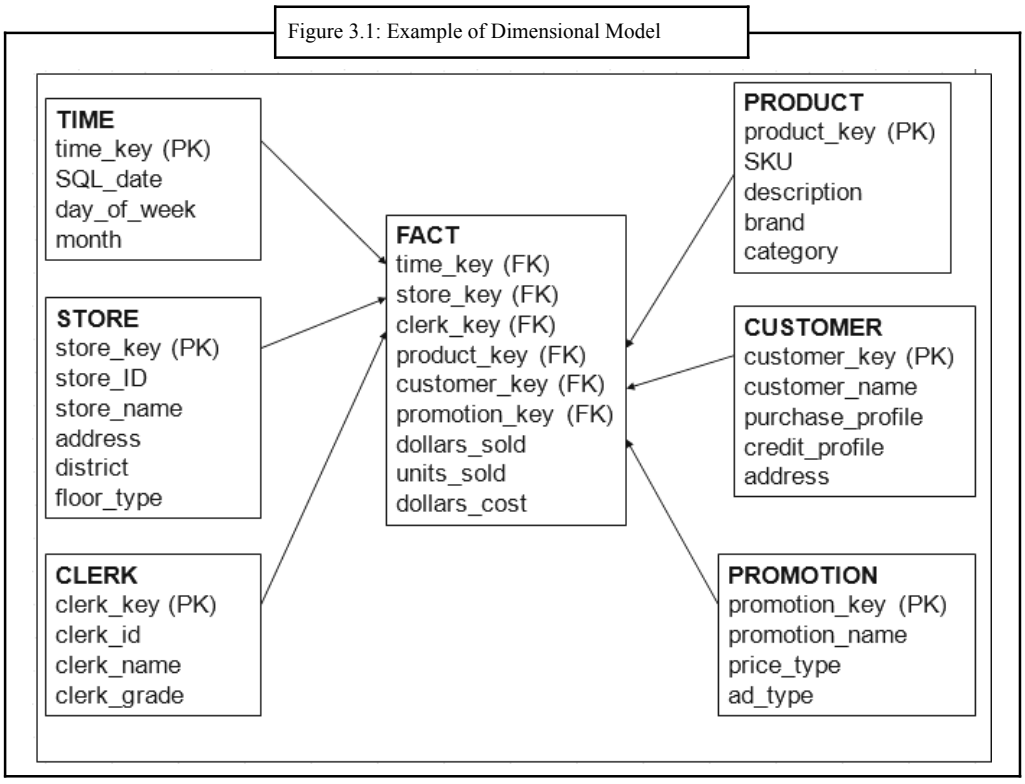
Example: For example, Product, Region and Time are the axes of enquiry of the Sales detail.

One such enquiry could be a scenario where the user might require to see the Sales (in dollars) for a specific item in a market over a specific time span of time. In this case, we are calculating the fact (Sales) over three dimensions (Product, Region and Time).

Thus we can say that dimensions give different views of the facts. They give structure to the otherwise unstructured facts.

It typically contains the attributes for the SQL answer set. Figure 3.1 shows an example of dimensional model.

Figure 3.1: Example of Dimensional Model



Notes

Self-Assessment

Fill in the blanks:

Dimensional model comprises of a ..... and numerous dimensional tables and is used for assessing summarized data.

2. ..... define the axis of enquiry of a fact.

3.2 Facts Table

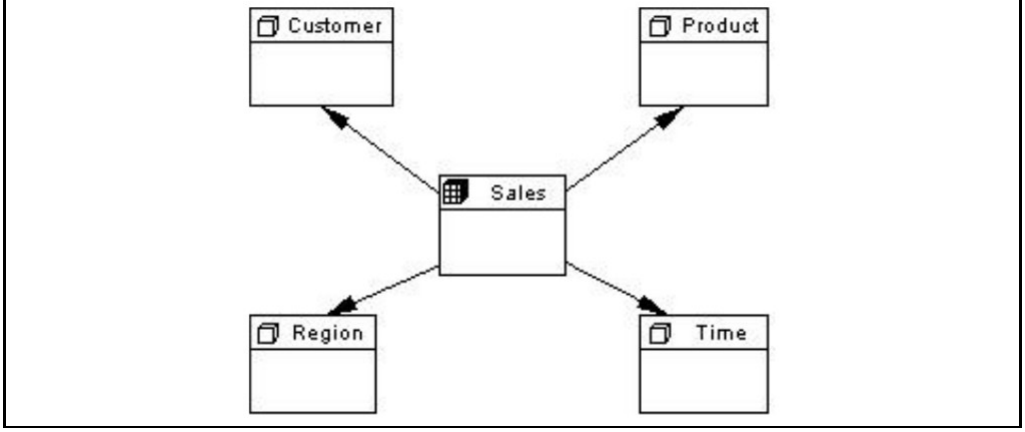
Fact table generally represent a process or reporting environment that is of value to the organization. It is important to determine the identity of the fact table and specify exactly what it represents. A fact table typically corresponds to an associative entity in the E-R model.

They must be listed in a logical fact table. Each measure has its own aggregation rules such as ADD, AVG, MIN or MAX. Aggregation rules define the way by which business would like to contrast standards of a measured value.

Caution Facts are the measurements associated with fact table records at fact table granularity.

The Figure 3.2 displays how Sales detail table is connected in a One-to-Many relationships with other dimension tables.

Figure 3.2: Sales Details Table One-to-many Relationship



3.2.1 Types of Measure

Various types of measure in a fact table are:

Additive - Measures that can be added across any dimensions are additive measure.

Semi Additive - Measures that can be added across only some dimensions are semi additive.

Non Additive - Measures that cannot be added across any dimension are non-additive.

3.2.2 Types of Fact Table

There are basically three types of fact tables:

Transactional: A transactional table is the most basic and fundamental type of fact table. The grain associated with a transactional fact table is usually specified as one row per line in a transaction, e.g., every line on a receipt represents a transaction.

Periodic Snapshots: It takes a picture of the moment, where the moment could be anything like performance summary of a salesman over the previous 3 months. A periodic snapshot table is dependent on the transactional table.

Accumulating Snapshots: In this type of fact table the activity of a process is shown such that it has a well-defined beginning and end.


Example: The processing of an order where an order moves through specific steps until it is completed.

As steps towards fulfilling the order are completed, the row which is associated with it is updated in the fact table. This type of table often has multiple date columns, each representing a complete step in the process. Therefore, it's important to have an entry in the date dimension that represents an unknown date, as many of the milestone completion time are unknown at the time the row is created.

Self-Assessment

Fill in the blanks:

A fact table typically corresponds to an associative entity in the .....

Measures that can be added across only some dimensions are .....

5. ..... take a picture of the moment, where the moment could be anything.

In ..... table often has multiple date columns, each representing a complete step in the process.

3.3 Dimension Tables

Dimension tables consist of attributes that describe fact records in the fact table. Some of these attributes provide descriptive information; others are used to specify how fact table data should be summarized to provide useful information to the person who is analyzing the information. Every dimension has a set of descriptive attributes. Dimension tables contain attributes that describe business entities.


Example: The Client dimension can contain attributes like C_No., Area, State, Country etc.

Did u know?

In a dimensional table, columns can be used to categorize the information into hierarchical levels.

Notes

For example, a dimension table for stores in the StandardMart sample database includes the following columns:

Table 3.1: Sample Dimension Table

Column	Description
store_country	Specifies the country or region in which the store is located. This is the country level of the hierarchy.
store_state	Specifies the state in which the store is located. This is the state level of the hierarchy.
store_city	Specifies the city or province in which the store is located. This is the city level of the hierarchy.
store_id	Specifies the individual store. This is the lowest level of the hierarchy. This field contains the primary key of the store dimension table and is used to join the dimension table to the fact table.
store_name	Specifies the name of the store. The values in this column are used to identify the store to users in a readable form.

Self Assessment

Fill in the blanks:

Dimension tables consist of attributes that describe ..... in the fact table.

..... contain attributes that describe business entities.

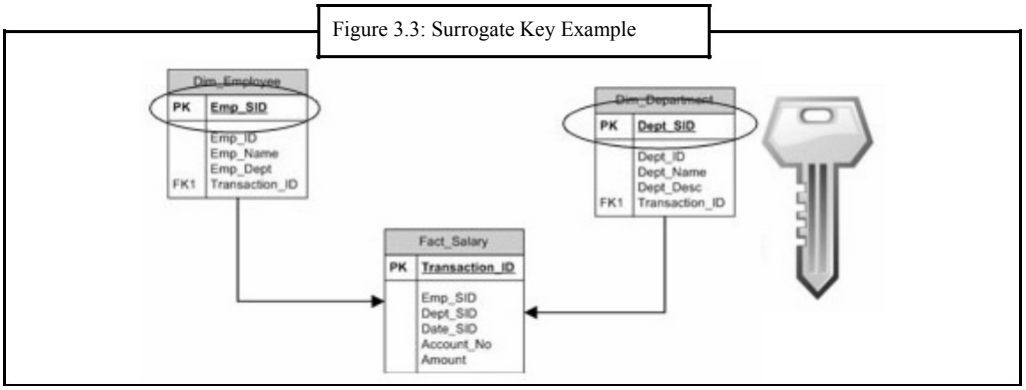
3.4 Surrogate Keys and Alternative Table Structure

A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database. The surrogate key is not derived from application data. Surrogate keys are keys that are maintained within the data warehouse instead of keys taken from source data systems.

Example: Say for the employee 'Emp12 the Business unit changes from B1 to B2. Now, if you use the natural primary key 'Emp12 for your employees within your data warehouse then everything would be allocated to Business unit 'B22 even what actually belongs to 'B1.'

If you use surrogate keys, you could create on the other day a new record for the Employee 'Emp12 in your Employee Dimension with a new surrogate key.

Figure 3.3: Surrogate Key Example



This way, in your fact table, you have your old data (i.e. before the day you added) with the SID of the Employee 'Emp12 >> 'B1.' All new data (i.e. after the day you added) would take the SID of the employee 'Emp12 >> 'B2.'

Notes

3.4.1 Advantages of Surrogate Keys

Immutability: Surrogate keys do not change while the row exists. Thus applications cannot misplace their reference in the database.

Change in Requirements: Attributes that uniquely recognize an entity might change over the time, which might lead to invalidation of the suitability of the compound keys.

Example: An employee's network username is chosen as a natural key. If it is merged with another company, new employees must be inserted. Now, some of the new user names may lead to conflict because their user names were developed independently.

In these cases, usually a new attribute should be added to the natural key (for example, an old_company column). In the case of a surrogate key, only the table that characterizes the surrogate key must be altered. But in the case of natural keys, all tables that use the natural key will have to change.

Performance: Surrogate keys tend to be a compact data type, such as a four-byte integer. This allows the database to query the single key column faster than it could multiple columns.

Uniformity: When every table has a uniform surrogate key, some tasks can be easily automated by composing the code in a table-independent way.

Validation: It is possible to design key-values that are in coordination with a well-known pattern which can be automatically verified.

Example: The keys that are intended to be used in some column of some table might be designed to "look differently from" those that are intended to be used in another column or table, thereby simplifying the detection of application errors in which the keys have been misplaced.

3.4.2 Disadvantages of Surrogate Keys

But surrogate keys also come with some disadvantages. The values of surrogate keys have no relationship with the real world meaning of the data held in a row. Therefore over usage of surrogate keys lead to the problem of disassociation and creates unnecessary ETL burden and performance degradation.

Query optimization also becomes difficult when one disassociates the surrogate key with the natural key. This is because when surrogate key takes the place of primary key, unique index is applied on that column. And any query based on natural key identifier leads to full table scan as that query cannot take the advantage of unique index on the surrogate key.

Referential Integrity: Referential integrity must be maintained between all dimension tables and the fact table. Each fact record contains foreign keys which are related to primary keys in the dimension tables.

Notes

Caution Every fact record must have a related record in every dimension table used with that particular fact table.

Shared Dimensions: To maintain consistency dimension tables that are shared are created. These tables are used by all components and data marts in the data warehouse.

3.4.3 Alternative Tables used in Data Warehousing

Auxiliary Table

This table is created with the SQL statements CREATE AUXILIARY TABLE and is used to hold the data for a column that is defined in a base table.

Base Table

The most common type of table is base table. You can create a base table with the SQL CREATE TABLE statement. All programs and users that refer to this type of table refer to the same description of the table and to the same instance of the table.

Clone Table

A table that is structurally identical to a base table is known as clone table. You can create a clone table by using an ALTER TABLE statement for the base table that includes an ADD CLONE clause.

Example: In the DB2 catalogue, SYSTABLESPACE.CLONE indicates that a clone table exists.

Empty Table

A table with zero rows is an empty table.

History Table

A history table is used by Database to store historical versions of rows from the associated system period temporal table.

Materialized Query Table

Materialized query tables are useful for complex queries that run on large amounts of data.

Notes They are commonly used in data warehousing and business intelligence applications.

Result Table

A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table. A result table is not an object that you can define using a CREATE statement.

Temporal Table

A temporal table is a table that records the period of time when a row is valid.

A table that is defined by the SQL statement CREATE GLOBAL TEMPORARY TABLE or DECLARE GLOBAL TEMPORARY TABLE is temporary table. It is used to hold data temporarily.

XML Table

It is a special table that holds only XML data. When you create a table with an XML column, database implicitly creates an XML table space and an XML table to store the XML data.

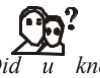
Self-Assessment

State whether the following statements are true or false:

- ✓ The surrogate key is derived from application data.
- ✓ Surrogate keys change while the row exists.
- ✓ In the case of natural keys, all tables that use the natural key will have to change.
- ✓ When every table has a uniform surrogate key, some tasks can be easily automated by composing the code in a table-independent way.
- ✓ The values of surrogate keys have relationship with the real world meaning of the data held in a row.
- ✓ The most common type of table is base table.
- ✓ A table with zero rows is an empty table.
- ✓ A temporal table is a table that records the period of time when a row is valid.
- ✓ XML table is used to hold data temporarily.

3.5 Multidimensional OLAP

OLAP stands for On-Line Analytical Processing. In computing, OLAP is an approach to answering Multi-Dimensional Analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also includes relational database, report writing and data mining. Depending on the underlying technology used, OLAP can be broadly divided into MOLAP and ROLAP.

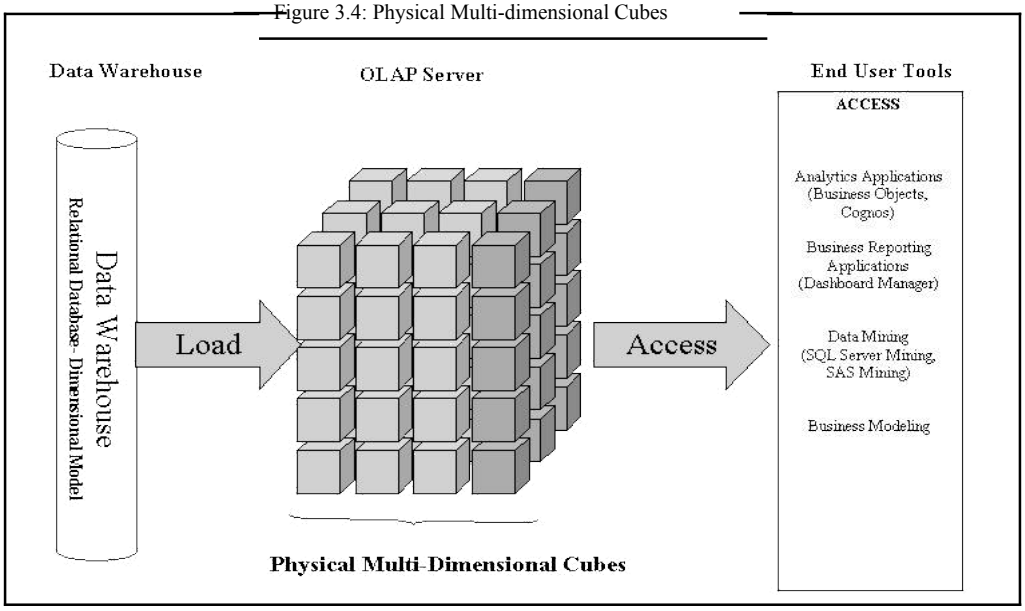


Did u know? In the OLAP world, there are mainly two different types: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP). Hybrid OLAP (HOLAP) is combination of MOLAP and ROLAP.

3.5.1 MOLAP

In MOLAP, data is stored in a multidimensional cube. It fulfils the requirements for an analytic application, where you require to access only summarized level of data. The storage is not in the relational database, but in proprietary formats. Figure 3.4 shows physical multi-dimensional cubes.

Figure 3.4: Physical Multi-dimensional Cubes



This method stores the data in multi-dimensional arrays which is different from the two dimensional relational structure.

Advantages:

MOLAP cubes are built for fast data retrieval and are thus optimal for slicing operations. MOLAP can perform complex calculations quickly.

Disadvantages:

MOLAP is limited in the amount of data it can handle because all the calculations are performed when the cube is built.

Cube technology generally do not already exist in the organization, therefore, to adopt MOLAP technology, chances are additional investments in the form of human and capital is needed.

3.5.2 ROLAP

This methodology depends on manipulating the data stored in the relational database. There are detail level values in relational data warehouse.

Advantages:

ROLAP can handle large amounts of data.

ROLAP can leverage functionalities inherent in the relational database as they sit on top of the relational database.

Disadvantages:

In ROLAP the performance can be slow. As it is known that ROLAP report is essentially a SQL query on the relational database, the query time can be long if the underlying data size is large thus the performance of same can be slow.

ROLAP can be limited by SQL functionalities. As, ROLAP technology mainly relies on SQL statements and SQL statements do not fit all needs (like it's not easy to do complex queries in SQL), thus what ROLAP can do is traditionally limited by what SQL can do.

Notes

3.5.3 HOLAP

HOLAP technologies combine the advantages of MOLAP and ROLAP. The first product to provide HOLAP storage was Holos but with time the technology also became available in other commercial products such as Microsoft Analysis Services (MAS), Oracle Database OLAP Option, MicroStrategy etc.

Task Compare and contrast the MOLAP and ROLAP.

Self-Assessment

Fill in the blanks:

18. OLAP stands for .....
19. OLAP can be broadly divided into ..... and .....
20. In MOLAP, data is stored in a .....
21. ..... can leverage functionalities inherent in the relational database as they sit on top of the relational database.
22. ..... technologies combine the advantages of MOLAP and ROLAP.

Case Study

Lolopop: Automated Data Warehouse

The essential concept of a data warehouse is to provide the ability to gather data into optimized databases without regard for the generating applications or platforms. Data warehousing can be formally defined as "the coordinated, architected, and periodic copying of data from various sources into an environment optimized for analytical and informational processing".

The Challenge

Meaningful analysis of data requires us to unite information from many sources in many forms, including: images; text; audio/video recordings; databases; forms, etc. The information sources may never have been intended to be used for data analysis purposes. These sources may have different formats, contain inaccurate or outdated information, be of low transcription quality, be mislabelled or be incompatible.

New sources of information may be needed periodically and some elements of information may be one time only artefacts.

A data warehouse system designed for analysis must be capable of assimilating these data elements from many disparate sources into a common form. Correctly labelling and describing search keys and transcribing data in a form for analysis is critical. Qualifying

Contd....

Notes

the accuracy of the data against its original source of authority is imperative. Any such system must also be able to: apply policy and procedure for comparing information from multiple sources to select the most accurate source for a data element; correct data elements as needed; and check inconsistencies amongst the data. It must accomplish this while maintaining a complete data history of every element before and after every change with attribution of the change to person, time and place. It must be possible to apply policy or procedure within specific periods of time by processing date or event data to assure comparability of data within a calendar or a processing time horizon. When data originates from a source where different policies and procedures are applied, it must be possible to reapply new policies and procedures. Where quality of transcription is low qualifying the data through verification or sampling against original source documents and media is required. Finally, it must be possible to recreate the exact state of all data at any date by processing time horizon or by event horizon.

The analytical system applied to a data warehouse must be applicable to all data and combinations of data. It must take into account whether sufficient data exists at the necessary quality level to make conclusions at the desired significance level. Where possible it must facilitate remediation of data from original primary source(s) of authority.

When new data is acquired from new sources, it must be possible to input and register the data automatically. Processing must be flexible enough to process these new sources according to their own unique requirements and yet consistently apply policy and procedure so that data from new sources is comparable to existing data.

When decisions are made to change the way data is processed, edited, or how policy and procedure is applied, it must be possible to exactly determine the point in time that this change was made. It must be possible to apply old policies and procedures for comparison to old analyses, and new policy and procedure for new analyses.

Defining Data Warehouse Issues

The Lolopop partners served as principals in a data warehouse effort with objectives that are shared by most users of data warehouses. During business analysis and requirements gathering phase, we found that high quality was cited as the number one objective. Many other objectives were actually quality objectives, as well. Based on our experiences, Lolopop defines the generalized objectives in order of importance as:

Quality information to Create data and/or combine with other data sources

In this case, only about one in eight events could be used for analysis across databases. Stakeholders said that reporting of the same data from the same incoming information varied wildly when re-reported at a later date or when it came from another organization’s analysis of the same data. Frequently the data in computer databases was demonstrably not contained in the original documents from which they were transcribed. Conflicting applications of policy and procedure by departments with different objectives, prejudices and perspectives were applied inconsistently without recording the changes or their sources, leaving the data for any given event a slave to who last interpreted it.

Timely response to requests for data

Here, the data was processed in time period batches. In some instances, it could take up to four years to finalize a data period. Organizations requiring data for analysis simply went to the reporting source and got their own copies for analysis, entirely bypassing the official data warehouse and analytical sources.

Contd....

Notes

Consistent relating of information

An issue as simple as a name — the information that could be used to connect data events to histories for individuals or other uniting objects — had no consistent method to standardize or simplify naming conventions. Another example, Geographical Information System (GIS) location information had an extravagant infrastructure that was constantly changing. This made comparisons of data from two different time periods extremely difficult.

Easy access to information

Often data warehouse technologies assume or demand a sophisticated understanding of relational databases and statistical analysis. This prevents ordinary stakeholders from using data effectively and with confidence. In some instances, the personnel responsible for analysis lack the professional and technical skills to develop effective solutions. This issue can stultify reporting to a few kinds of reports and variants that have been programmed over time, and reduces data selection for the analyses to kind of magic applied by clerical personnel responsible for generating reports.

Unleash management to formulate and uniformly apply policy and procedure

We found that management decisions and mandates could be hindered by an inability to effectively capture, store, retrieve and analyse data.

In this particular instance, no management controls existed to analyse: source of low quality; work rates; work effort to remediate (or even a concept of remediation); effectiveness of procedures; effectiveness of work effort; etc.

Remediation is a good case in point. Management experienced difficulty with the concept of remedying data transcription from past paper forms — even though the forms existed in images that could be automatically routed. The perception was that quantity of data, not quality, was the objective and that no one would ever attempt to fix data by verifying it or comparing it to original documents.

Manage incoming data from non-integrated sources

Data from multiple, unrelated sources requires a plan to convert electronic data, manage imaging and documents inputs, manage workflow and manage the analysis of data. In this case, every interface required manual intervention. Since there was no system awareness at the beginning of the capture process as to what was needed for analysis at the end, it was very difficult to make rapid and time effective changes to accommodate changing stakeholder needs.

Reproducible Reporting Results

We found that reporting of data was not reproducible and the reasons for differences in reporting were not retrievable, undermining confidence in the data, analysis and reporting. One may essentially summarize these objectives as quality challenges that require a basic systems engineering approach for resolution.

Questions:

What were the challenges of lolopop automated data warehouse?

What were the data warehouse issues?

Notes

3.6 Summary

Dimensional model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data.

Fact table generally represent a process or reporting environment that is of value to the organization.

A fact table typically corresponds to an associative entity in the E-R model.

Various types of measure in a fact table are: Additive, Semi Additive, Non-Additive.

There are basically three types of fact tables: Transactional, Periodic snapshots and accumulating snapshots.

Dimension tables consist of attributes that describe fact records in the fact table.

A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database.

Attributes that uniquely recognize an entity might change over the time, which might lead to invalidation of the suitability of the compound keys.

But surrogate keys also come with some disadvantages. The values of surrogate keys have no relationship with the real world meaning of the data held in a row.

Referential integrity must be maintained between all dimension tables and the fact table.

The most common type of table is base table. You can create a base table with the SQL CREATE TABLE statement.

A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table.

OLAP stands for On-Line Analytical Processing. In computing, OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly.

In MOLAP, data is stored in a multidimensional cube. It fulfils the requirements for an analytic application, where you require to access only summarized level of data.

HOLAP technologies combine the advantages of MOLAP and ROLAP.

3.7 Keywords

Accumulating Snapshots: In this type of fact table the activity of a process is shown such that it has a well-defined beginning and end.

Auxiliary Table: This table is created with the SQL statements CREATE AUXILIARY TABLE and is used to hold the data for a column that is defined in a base table.

Dimension Tables: Dimension tables consist of attributes that describe fact records in the fact table.

Dimensional Model: Dimensional Modelling (DM) is the name of a set of techniques and concepts used in data warehouse design. It is considered to be different from entity-relationship modelling (ER).

Empty Table: It is a table with zero rows is an empty table.

E-R Model: In software engineering, an Entity-relationship model (ER model) is a data model for describing a database in an abstract way.

Fact Table: Fact table generally represent a process or reporting environment that is of value to the organization.

HOLAP: HOLAP (Hybrid Online Analytical Processing) is a combination of ROLAP (Relational OLAP) and MOLAP (Multidimensional OLAP) which are other possible implementations of OLAP.

Multidimensional Online Analytical Processing (MOLAP): This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Result Table: A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table.

ROLAP: This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality.

Surrogate Key: A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database.

Temporal Table: A temporal table is a table that records the period of time when a row is valid.

Transactional Table: The grain associated with a transactional fact table is usually specified as one row per line in a transaction.

XML Table: It is a special table that holds only XML data.

3.8 Review Questions

- ✓ What is dimension?
- ✓ Explain the dimensional model.
- ✓ "Dimensions define the axis of enquiry of a fact." Elucidate.
- ✓ Give examples of dimensional model.
- ✓ What do you understand by fact table?
- ✓ Explain the types of measure and fact table.
- ✓ "Dimension tables consist of attributes that describe fact records in the fact table". Discuss.
- ✓ Define the concept of surrogate key. Also write down the advantages and disadvantages.
- ✓ What do you understand by alternative tables used in data warehousing?
- ✓ Briefly explain about multidimensional OLAP.

Answers: Self-Assessment

1. Fact table
2. Dimensions
3. E-R model
4. Semi additive
5. Periodic snapshots
6. Accumulating snapshots
7. Fact records
8. Dimension tables
9. False
10. False

Notes

11. True
12. True
13. False
14. True
15. True
16. True
17. False
18. On-Line Analytical Processing
19. MOLAP, ROLAP
20. Multidimensional cube
21. ROLAP
22. HOLAP