B.C.A. (Sem – VI)

B.C.A. - 603

Data Warehousing & Data Mining

Purushottam Singh

Unit:- 1

Data Warehousing & Mining

UNIT-1

Definition: - *A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.*

- **Subject oriented** -- data are organized around sales, products, etc.
- **Integrated** -- data are integrated to provide a comprehensive view.
- **Time variant** -- historical data are maintained.
- **Nonvolatile** -- data are not updated by users.

Data warehouse Usage:-

Information processing:-

- Supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs.

Analytical processing :-

- Multidimensional analysis of data warehouse data.
- Supports basic OLAP operations, slice-dice, drilling, pivoting.

Data mining:-

- Knowledge discovery from hidden patterns.
- Supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Datawarehouse Trends:-

Data Warehouse Appliances :

- There are many reasons why organizations consider buying an appliance, the main reason is simplicity.
- The appliance is delivered complete and installs rapidly.

By- Prof. Vishal Trivedi

- The vendor builds and certifies the configuration, balancing hardware, software and services for a predictable performance.
- If there are any problems, a single call to the appliance vendor is the first course of action.

Data Warehouse as a Service and Cloud :

- Data warehouse as a service comes in two "flavors" — software as a service and outsourced data warehouses.
- . Data warehouse in the cloud is primary an infrastructure design option as a data model.
- An integration strategy must be row wise and BI user access must be enabled and managed.
- Private clouds are a useful infrastructure design choice for some organizations in supporting their data warehouse and analysis.

Using an Open-Source DBMS to Deploy the Data Warehouse :

- Open-source DBMSs are still being used in both experimental and more formalized approaches.
- At this point, open-source warehouses are rare and usually smaller than traditional.
- Some solutions are optimized specifically for data warehousing.

In-memory technology :

- In-memory databases is to sub-second response requirements, do not share any common business space with data warehousing tools.
- Informatica is working closely and modify with other data warehouse vendors in these areas; for example, Greenplum, Teradata/Aster Data, HP/Vertica, etc.

- Data warehousing tools are marching towards in-memory in order to achieve higher performance.

No SQL

- Most commercial applications and solutions use a relational database under the cover for metadata/content store.
- Customers are highly sensitive to the application and database support availability of the solution.
- NoSQL databases being an open source , will have to support challenges before it can be part of any critical business system.

Difference between DBMS v/s DATAWAREHOUSE:-

- DBMS Used for Online Transactional Processing (OLTP).
- Data warehouse Used for Online Analytical Processing (OLAP).
- The tables and joins are complex since they are normalized (for RDMS).
- The Tables and joins are simple since they are de-normalized.
- Entity – Relational modeling techniques are used for RDMS database design.
- Data – Modeling techniques are used for the Data Warehouse design.
- Optimized for write operation.
- Optimized for read operations.
- Performance is low for analysis queries.
- High performance for analytical queries.

DATA MART:-

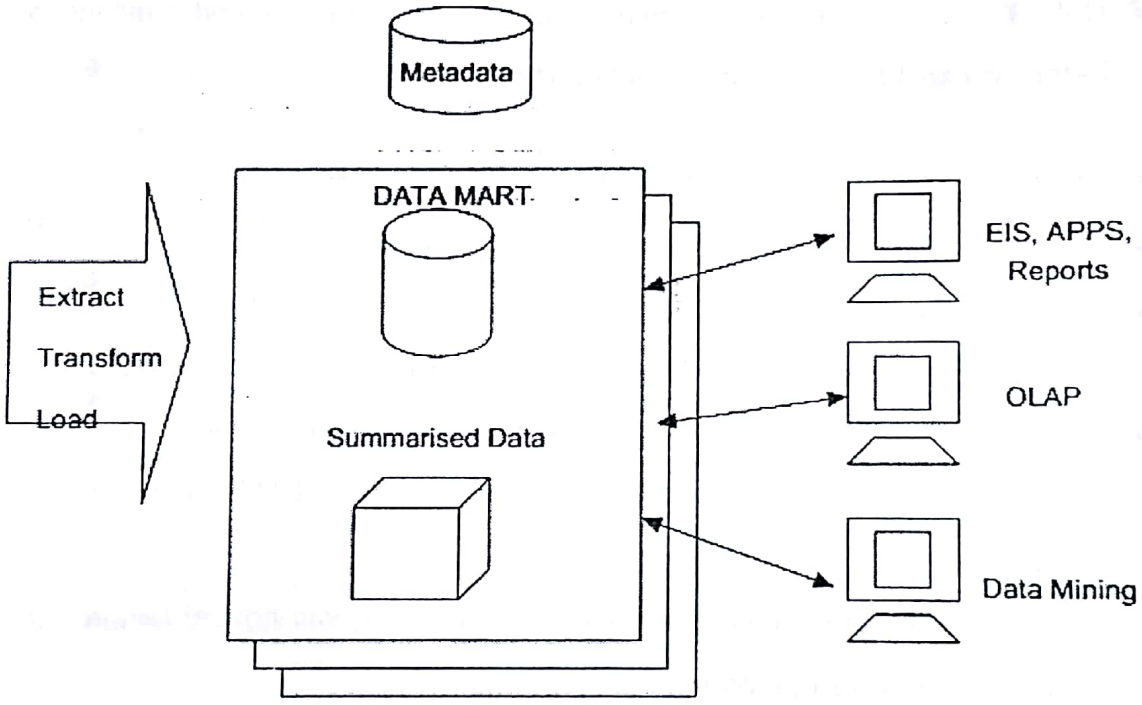
Data Warehousing & Mining

A subset of a data warehouse that supports the requirements of a particular department or business function.

- A data mart stores data for a limited number of subject areas, such as marketing and sales data. It is used to support specific applications.

- An independent data mart is created directly from source systems.
- A dependent data mart is populated from a data warehouse.

Figure of Data mart:-



per functional area

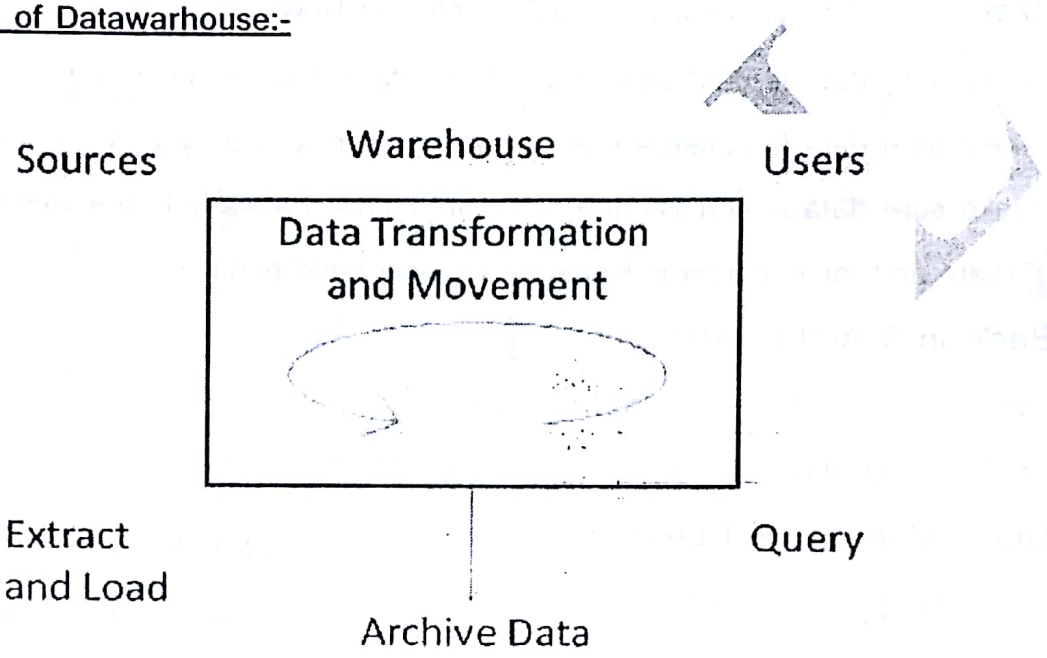
There are two kinds of Data Mart:-

1) Dependent
2) Independent

(1) Dependent: - A Dependent Data mart is one whose source is Data warehouse.

(2) Independent: - An Independent Data mart is one whose source is the legacy applications environment.

Process of Datawarhouse:-



Four types of process in Datawarehouse:-

1) Extract load and Process
2) Clean & Transformation
3) Back up & Archive process
4) Query management process

1) Extract load and process:-

- **Controlling the process:** determine when to start extracting the data, run transformation, consistency check & so on.
- **Example:** Retail sales analysis
- **When to initiate the extract :** Data should be in a consistent state.

- Same instance of time Ex. Telecom
- Loading the Data : Temporary data store. Clean up & consistency check.
- Example : current subscriber & current Event DB.

2) Clean & Transformation :-

- Clean & Transform the data in to a structure that speed up queries.
- Make sure data is consistent within itself. Ex: Row
- Make sure data is consistent with other data within the same source.
- Make sure data is consistent with data in other source system.
- Make sure data is consistent with the information already in the warehouse.
- Create and/or maintain indexes, views, and table partitions.

3) Back-up & Archive Process :-

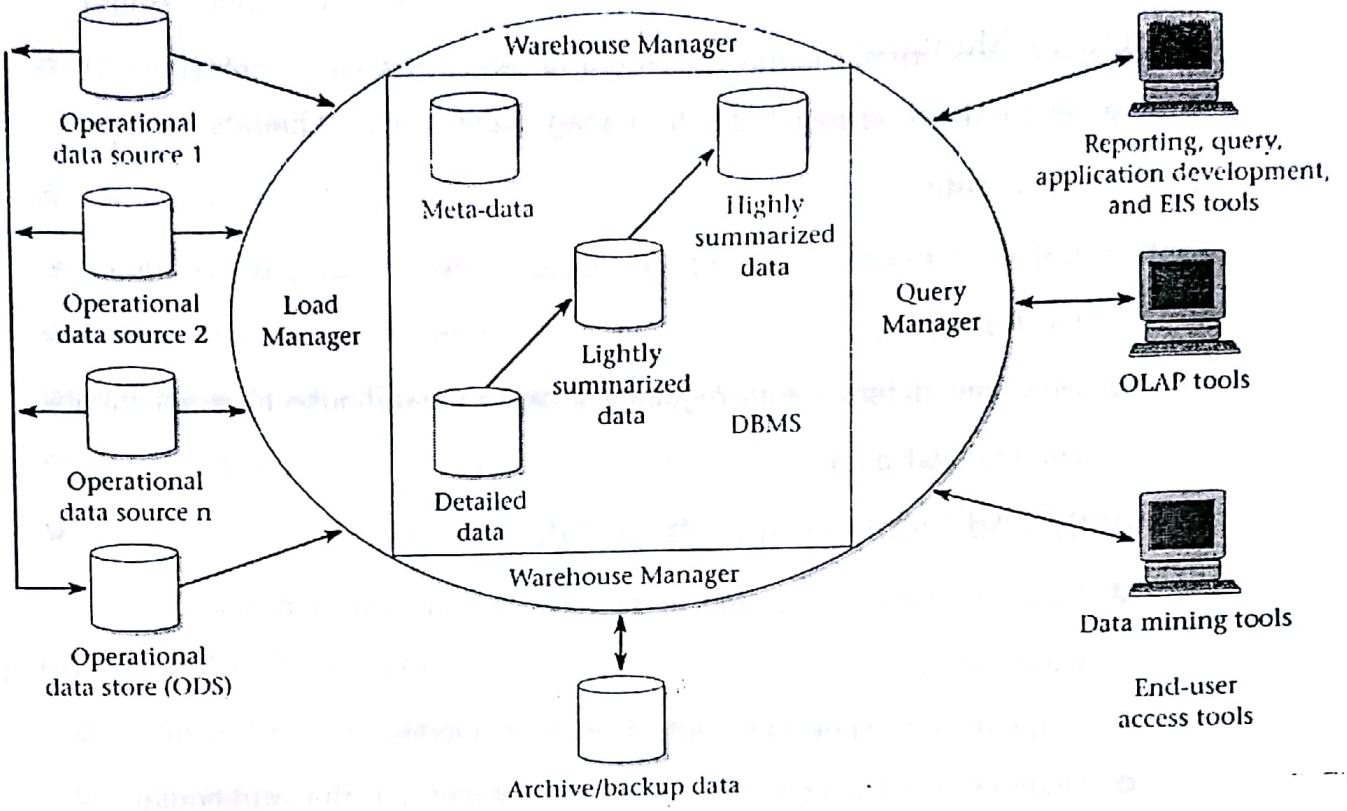
- Back-up regularly recover from loss/failure.
- In Archiving older data is removed from system.

4) Query Management process :-

- Directing query to most effective data source.

Data Warehousing & Mining

Architecture of Datawarehouse:-



- **Operational data:**
  - Supplied from mainframes, proprietary file systems, private workstations and servers, and external systems such as the Internet.
- **Operational data store (ODS):**
  - Repository of current and integrated operational data used for analysis.
  - Often structured and supplied with data in the same way as the data warehouse.
  - May act simply as a staging area for data to be moved into the warehouse.
- **Load Manager:**
  - Performs all operations associated with extraction and loading of data into warehouse.
- **Warehouse Manager**

- Performs all operations associated with management of data in the warehouse, such as merging data sources.

❖ Query Manager

- Performs all associated with management of user Queries.

❖ Detailed data :

- Not stored online but made available by summarizing data to the next level of detail.
- However, detailed data regularly added to warehouse to supplement summarized data.

❖ Lightly and highly summarized data:

- Predefined and generated by warehouse manager and stored in warehouse.
- Purpose is to speed up performance of queries.
- Updated continuously as new data is loaded into the warehouse.

❖ Meta-data (data about data):

- Used by all processes in the warehouse.

❖ End-user access tools:

- Principal purpose of data warehousing is to provide information to business users for strategic decision-making.
- Users interact with warehouse using end-user access tools.
- Warehouse must efficiently support *ad hoc* and routine analysis.
- Includes EIS, OLAP and data mining tools.

Characteristics of Data Warehouse:-

- Subject oriented: - Data are organized based on how the users refer to them.

Data Warehousing & Mining

- **Integrated:** - All inconsistencies regarding naming convention and value representations are removed.
- **Nonvolatile:** - Data are stored in read-only format and do not change over time.
- **Time variant:** - Data are not current but normally time series.
- **Summarized:**- Operational data are mapped into a decision-usable format
- **Large volume:** - Time series data sets are normally quite large.
- **Not normalized:** - DW data can be, and often are, redundant.
- **Metadata:** - Data about data are stored.
- **Data sources:** - Data come from internal and external unintegrated operational systems.

Need for Data Warehouse:-

- Industry has huge amount of operational data.
- Knowledge worker wants to turn this data into useful information.
- This information is used by them to support strategic decision making.
- It is a platform for consolidated historical data for analysis.
- It stores data of good quality so that knowledge worker can make correct decisions

- From business perspective

-It is latest marketing weapon

-Helps to keep customers by learning more about their needs.

-Valuable tool in today's competitive fast evolving world.

Basic Elements of Data Warehouse:-

(1) Source System:-

- Typically in any organization the data is stored in various databases, usually divided up by the systems.
- There may be data for marketing, sales, payroll, engineering, etc.

- These systems might be legacy/mainframe systems or relational database systems.

(2) Staging Area:-

- The data coming from various source systems is first kept in a staging area.
- The staging area is used to clean, transform, combine, de-duplicate, household, archive, and to prepare source data for use in data warehouse.
- This need not be based on relational terminology.
- Sometimes managers of the data are comfortable with normalized set of data.
- Staging area doesn't provide querying/presentation services.

(3) Presentation server:-

- Once the data is in staging area, it is cleansed, transformed and then sent to Data warehouse.
- You may or may not have ODS before transferring data to Data Warehouse.

(4) OLAP:-

- The data in Data Warehouse has to be easily manipulated in order to answer the business questions from management and other users.
- This is accomplished by connecting the data to fast and easy-to-use tools known as Online Analytical Processing (OLAP) tools.
- OLAP server, data is reorganized to meet the reporting and analysis
  - Requirements of the business, including:
  - Exception reporting
  - Ad-hoc analysis
  - Actual vs. budget reporting
  - Data mining

(5) DataMart:-

- Data mart is a logical subset of complete data warehouse.
- It is often viewed as the restriction of data warehouse to a single business process or to a group of related business processes targeted toward a particular business group.
- For example an organization may have a data mart for Sales or Inventory.

(6) Datawarehouse Tools:-

ETL Tools	BO Data Integrator, DMExpress, Informatica, IBM Data Stage, Microsoft SSIS, Oracle Warehouse Builder, LogiXML,
OLAP Server	IBM DB2 OLAP Server, Microsoft SQL Server OLAP Services, Oracle Express Server, Palo OLAP Server
OLAP Tools	IBM Cognos, MicroStrategy, Microsoft SSAS, Microsoft SSRS, Oracle Express Suite,SAP Business Objects
Data Warehouse	Informix, IBM DB2 UDB, Microsoft SQL Server,Oracle, Red Bricks, Teradata

User Requirements:-

- User requirements describe the tasks that the users must be able to accomplish with the help of the data warehouse system.
- User requirements must be collected from people who will actually use and work with the data warehouse system.
- The user requirements must align with the context and objectives established by the business requirements.

- User requirements on the modeling process and techniques applied for
- Data marts become even more important for data warehouses.
- User Requirements gathering is often incorporated in some way into studies of the business processes and information analysis activities in which end users are involved.
- User requirements are further investigated, and initial dimensional models are produced showing facts, measures, dimension keys, and dimension hierarchies.

Requirements Modeling:-

- Validated initial models are further developed into detailed dimensional models, showing all elements of the model and their properties.
- Detailed dimensional models can further be extended and optimized.
- Many techniques in this area should be thought of as advanced modeling techniques.
- Not every project requires all of them to be applied.
- The dimensional model usually tends to become complex and dense. This may cause problems for end users.
- Requirements modeling consists of several important activities that all are performed with the intent of producing a detailed conceptual model that represents at best the problem domain of the information analyst.

Temporal Data Modeling:-

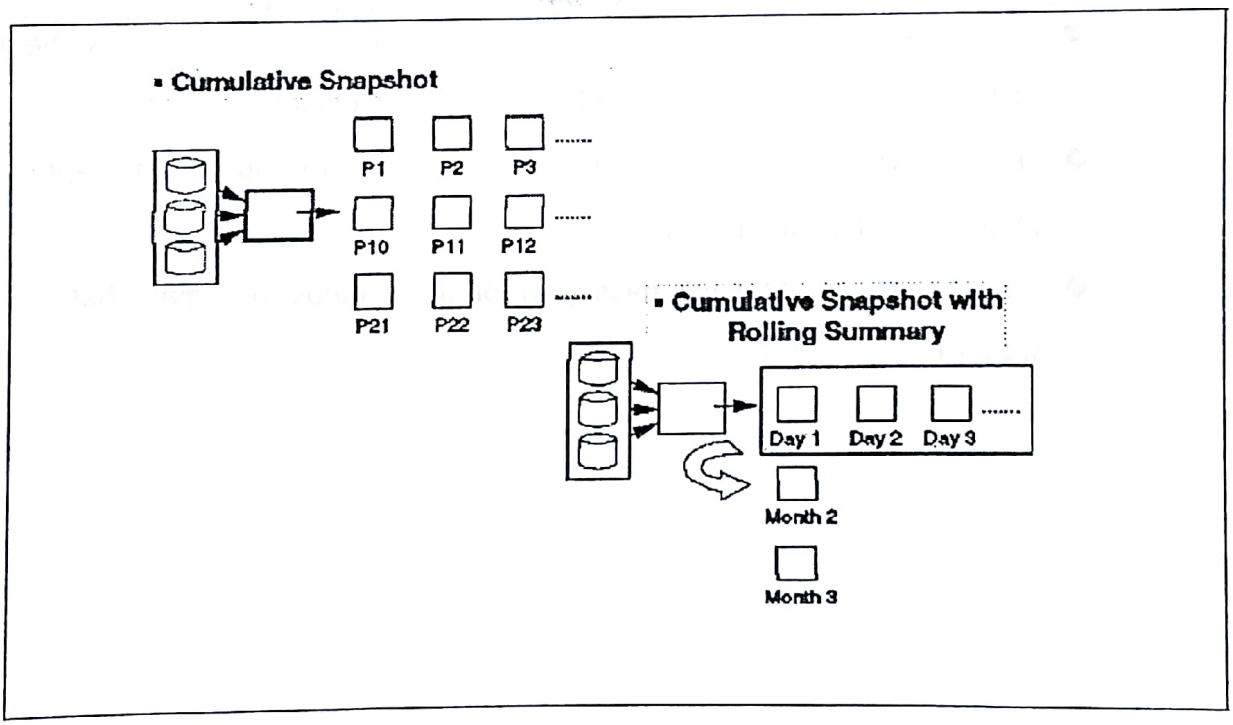
- Temporal data modeling consists of a collection of modeling techniques that are used to construct a temporal or historical data model.
- A temporal data model can loosely be defined as a data model that represents not only data items and their inherent structure but also

changes to the model and its content over time including, importantly, when these changes occurred or when they were valid.

- Temporal or historical data models distinguish themselves from traditional data models in that they incorporate one additional dimension in the model: the time dimension.
- Temporal modeling techniques from a general point of view, disregarding where the techniques are used in the process of data warehouse modeling.
- Temporal modeling can add substantial complexity to the modeling process and to the resulting data model.

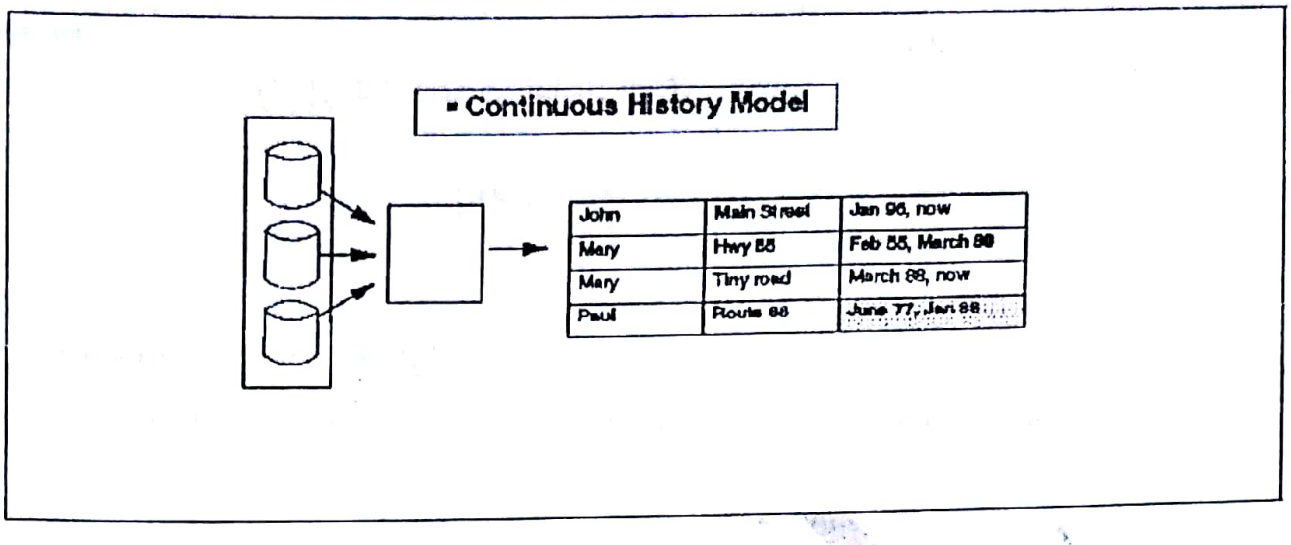
Temporal Data Modeling Style:-

- Temporal modeling styles or approaches have two of the most widely used modeling styles are cumulative snapshots and continuous history models.



Data Warehousing & Mining

- A database snapshot is a consistent view of the database, at a given point in time. For instance, the content of a database at the end of each day, week, or month represents a snapshot of the database at the end of each day, week, or month.
- Temporal modeling using a cumulative snapshot modeling style consists of collecting snapshots of a database or parts.
- It accumulating the snapshots in a single database, which then presents one form of historical dimension of the data in the database.
- If the snapshots are taken at the end of each day, the cumulative snapshot database will present a perception of history of the data.
- The technique of cumulative snapshots is often applied without considering a temporal modeling approach.
- It is a simple approach, for both end users and data modelers, but unfortunately, it has some serious drawbacks.
- Cumulative snapshots do tend to produce an overload of data in the resulting database.
- The other major drawback of cumulative snapshot modeling is the problem of information loss, which is inherent to the technique.
- Except when snapshotting transaction tables or tables that capture record changes in the database,
- Snapshots will always miss part of the change activities that take place within the database.



- Sometimes, the information loss problem can be reduced by taking snapshots more frequently.
- The continuous history model approach aims at producing a data model that can represent the full history of changes applied to data in the database.
- This approach leads to much more reliable solutions that do not suffer from the information loss problem associated with cumulative snapshots.
- In the remainder of this section, we explore techniques for temporal modeling using a continuous history modeling approach.