**Unit II Data Warehousing and Data Mining**

Part-A:

Data Ware housing(DwH): - Definition, Characteristic, types, Date ware housing frame work, DwH 3 tier architecture, Alternative Architectures, Data ware housing Integration, Data ware housing- Development Approaches, Real time Data ware housing.

**PART-A**

1. **Data Warehousing(DwH): Definition:**

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

In simple terms, a **Data Warehouse (DwH)** is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities.

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

1. **Characteristics of Data Warehousing:**

The following are the main Characteristics of Data Warehousing: -

1. **Subject oriented:**

Data are organized by detailed subject, such as Sales, Products, or Customers, containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing, but why & how is performing?

A deep understanding will help in developing sales procedures that are defined within the bounds. It deals with all the subject matters that have warehouse.

**2) Integrated:**

Integration is closely related to subject orientation. Data warehouses must place data from different sources into a consistent format. To do so, they must deal with naming conflicts and discrepancies among units of measure. A data warehouse is presumed to be totally integrated.

**3) Time variant:**

A warehouse maintains historical data. The data do not necessarily provide current status (except in real-time systems). They detect trends, deviations and long-term relationships for forecasting and comparisons, leading to decision-making.

Time is the one important dimension that all data warehouses must support. Data for analysis from multiple sources contain multiple time points (ex: daily, weekly, monthly views).

**4) Non volatile:**

After data are entered into a data warehouse users cannot change or update the data. Obsolete data are discarded and changes are recorded as new data.

1. **Application Areas of Data Warehouse:**

****

1. **Types of Data warehousing:**

There are three main types of data warehousing:

1. **Data Marts:**

Whereas a data warehouse combines databases across an entire enterprise, a data mart is usually smaller and focuses on a particular subject or department. A data mart is a subset of a data warehouse, typically consisting of a single subject area (ex: Marketing, Operations).

1. **Operational Data Store:**

An Operational Data Store (ODS) provides a fairly recent form of customer information file (CIF). This type of database is often used as an interim staging area for a data warehouse. Unlike the static contents of a data warehouse, the contents of an ODS are updated throughout the course of business operations.

An ODS is used for short-term decisions involving mission-critical applications rather than for the medium and long-term decisions associated with an Enterprise Data Warehouses (EDW).

1. **Enterprise Data Warehouses (EDW):**

An Enterprise Data Warehouse (EDW) is a large-scale data warehouse that is used across the enterprise for decision support. The large-scale nature provides integration of data from many sources into a standard format for effective Business Intelligence and decision support applications.

EDW are used to provide data for many types of DSS, including CRM, Supply-Chain Management 9SCM), business performance management (BPM), business activity monitoring (BAM), product life cycle management, revenue management and sometimes even knowledge management system etc.

1. **Data warehousing frame work:**

The Data Warehousing Framework describes the relationships between the various components used in the process of building, using, and managing a data warehouse.

Two enabling technologies comprise the core Data Warehousing Framework:

1. **The integrated metadata repository:**

Metadata repository is a database created to store metadata. Metadata is information about the structures that contain the actual data. Metadata is often said to be "data about data", but this is misleading.

Data profiles are an example of actual "data about data". Metadata is one layer of abstraction removed from this – it is data about the structures that contain data. Metadata may describe the structure of any data, of any subject, stored in any format.

1. **The data transport layer (OLE DB):**

The transport layer is the layer in the open system interconnection (OSI) model responsible for end-to-end communication over a network. It provides logical communication between application processes running on different hosts within a layered architecture of protocols and other network components.

The transport layer is also responsible for the management of error correction, providing quality and reliability to the end user. This layer enables the host to send and receive error corrected data, packets or messages over a network and is the network component that allows multiplexing.

1. **DwH 3 tier architecture:**

****

****

****

1. **Alternative Architectures:**
2. Master-Slave Architecture
3. Client-Server Architecture
4. Peer-to-Peer (P2P) Architecture
5. eSharing Digital Objects using P2P
6. eTrading Digital Objects using P2P
7. 'The Cloud'
8. **Master-Slave Architecture**

**Star Topology**

****

**Multi-Personal Client-Server**

**The 'PC’ Era – Mid-Late 1970s Onwards**

****

1. **Client-Server Architecture**

**Mid-1980s Onwards**

****

1. **Peer-to-Peer (P2P) Architecture:**
* **Network Management**
* **Directory Management**
* **Object Management**
* **Process Management**

Server & Client

in

Workstation

Server & Client

in

Workstation

**Client-Server Architecture – Internet-Mediated
Mid-1990s Onwards**

****

**Key Developments Since the Mid-1990s**

* Workstation Capacity (now rivals Hosts)
* Workstation Diversity (vast, expanding)

desktops, laptops, smartcards, phones ==> smartphones,

Personal Digital Assistants (PDAs) ==> tablets etc.

* Broadband Connectivity (widespread)
* This enables dispersion and replication of devices capable of providing services Wireless Connectivity (increasingly widespread).
* This enables Mobility which means Devices change networks along with a change their IP-addresses.
1. **eSharing Files using Peer-to-Peer (P2P):**
2. **The Scope for eTrading in Music:**
3. **The Cloud:**

**Meaning of the Cloud:**

The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

1. Outsourcing
	* of IT Infrastructure
	* of IT platform, i.e. software environment
	* if Applications
2. 'Server Virtualisation'
3. Cloudsourcing

****

1. **Data ware housing Integration:**

****

1. **Data warehousing- Development Approaches:**

When it comes to designing a data warehouse for the business, the two most commonly discussed Development Approaches were introduced by Bill Inmon and Ralph Kimball.

**Bill Inmon’s enterprise data warehouse approach (the top-down design):**

A normalized data model is designed first. Then the dimensional data marts, which contain data required for specific business processes or specific departments are created from the data warehouse.

**Ralph Kimball’s dimensional design approach (the bottom-up design):**

The data marts facilitating reports and analysis are created first; these are then combined together to create a broad data warehouse.

**Inmon’s top-down approach:**

Inmon defines data warehouse as a [centralized repository](http://searchsqlserver.techtarget.com/definition/data-warehouse) for the entire enterprise. Data warehouse stores the ‘atomic’ data at the lowest level of detail. [Dimensional data marts](http://searchdatamanagement.techtarget.com/definition/dimension) are created only after the complete data warehouse has been created. Thus, data warehouse is at the center of the Corporate Information Factory (CIF), which provides a logical framework for [delivering business intelligence](http://searchcio.techtarget.in/tutorial/BI-tools-guide-for-managers).

Inmon defines the data warehouse in the following terms:

1. **Subject-oriented:**The data in the data warehouse is organized so that all the data elements relating to the same real-world event or object are linked together
2. **Time-variant:**The changes to the data in the database are tracked and recorded so that reports can be produced showing changes over time
3. **Non-volatile:** Data in the data warehouse is never over-written or deleted -- once committed, the data is static, read-only, and retained for future reporting
4. **Integrated:** The database contains data from most or all of an organization's operational applications, and that this data is made consistent.

**Kimball’s bottom-up approach:**

Keeping in mind the most important business aspects or departments, data marts are created first. These provide a thin view into the organizational data, and as and when required these can be combined into a larger data warehouse. Kimball defines data warehouse as “A copy of transaction data specifically structured for query and analysis”.

Dimensional modeling focuses on ease of end user accessibility and provides a high level of performance to the data warehouse.

**Essential Differences Between Inmon’s Kumball’s Approaches**

|  |  |  |
| --- | --- | --- |
| **Characteristic** | **Inmon** | **Kimbal** |
| Overall Approach | Top-Down | Bottom-Up |
| Architecture Structure | Enterprise-wide;  | Single business process |
| Complexity of the method | Quite complex | Fairly simple |
| Data orientation  | Fairly thorough | Fairly light |
| End-user accessibility | Low | High |
| Primary audience | IT Professionals | End users |
| Place in Organisation | Integral part of the Corporate Information factory | Transformer and retainer of operational data.  |
| Objective | Deliver a sound technical solution | Deliver a solution that makes it easy for end users |

**10) Real time Data warehousing:**

Real-time Data Warehousing, also known as Active Data Warehousing, is the process of loading and providing data via the Data Warehouse as they become available. It evolved from the Enterprise Data Warehouse (EDW) concept.

People throughout the organization who interact directly with customers and suppliers will be empowered with information-based decision making at their fingertips.

**END OF PART-A**