SYLLABUS FOR ST- 2
DATA WAREHOUSING CONCEPTS AND ETL TECHNOLOGIES

14-17	**Multidimensional Data Modelling:** Introduction to data modeling techniques, Ralph Kimball's Approach vs. W.H. Inman's Approach, Fact Table, Dimensional Table.
18-20	**Typical Dimensional Models:** Star Schema, Snowflake Schema, Fact constellation Schema
21-25	**Introduction to OLTO and OLAP:** Difference between OLTP and OLAP, OLAP Guidelines, categories, OLAP Tools, OLAP operations.
26-33	**The ETL Process:** Introduction and challenges in ETL process, Data Extraction – identification of data sources and extracting data, Data Transformation – tasks involved in data transformation, Data Loading – techniques of data loading

The ETL Process

Introduction

ETL functions (extract transform and load) take place in the data staging area, reshape the relevant data from the source systems into useful information to be stored in the data warehouse.

If source data is not extracted correctly cleansed and integrated in the proper format , query processing will not take place. Design and implementation of the automated ETL process often represents a major part of effort to develop data warehouse (70%).

Challenges in ETL

ETL functions are challenging because of the nature of the source systems. A lot of disparities in the source systems make the ETL functions a challenging task to accomplish.

Given below is the list of reasons for the types of difficulties in the ETL function

Source Systems are diverse and disparate

Source systems run on different platforms and have different operating systems installed.

Most of the operational systems do not preserve historical data which is critical for data warehouse

Quality of data cannot be guaranteed in older operational source system

Structures of the source system keep changing over time with the advent of new technology.

The prevalence of data inconsistency in the source system is a major challenge

Data in the source systems may be ambiguous or stored in cryptic form which is hard to handle

The data type, format and naming convention may be different during the ETL Process

ETL Process

- Determine the target data
- Determine all the data sources (Internal & External)
- Prepare Mapping for target data element from sources

- Plan for aggregate tables
- Determine data cleansing and transformation rules
- Determine data extraction rules

- Organize data staging area
- Write procedure for all data loads
- ETL for Dimension Tables
- ETL for Fact Tables

Select the source data/ Identification of Data sources

Source data is grouped into four categories

a) Production Data

b) Internal Data

c) External Data

d) Archived Data

Production Data – Comes from operational systems (Directly comes from the OLTP systems)

It is the main source of data in the warehouse

Internal Data – Internal to any organization (Employees)

Internal data is taken from private files and could include the data that is not stored on computer

(Customer profiles, personal spreadsheet) , Data extracted from individuals documents and private files

Archived data – Back up

An operational system is used to run the day to day business transactions and for this you need to keep only current information in the database. So periodically the old data is taken from systems and stored in archived files.

Stage 1- Recent data is archived to separate archival database that may still be online

Stage 2 – The older data is archived to flat files on disk storage

Stage 3- The oldest data is archived to tape cartridge and even kept off site.

External Data - Strategic Decisions – (Unstructured data)

The external data is mainly collected from business magazines, industry newsletters technology reports , reports generated by consultants, competitive analysis report , sales and marketing analysis report etc.

Extract Data from the source systems

The data extraction process has to deal with **multiple data sources**. In data warehouse environment one thing that is pretty sure is inconsistent , noisy and different formats.

Abinitio, Mapforce, Pentaho , Octopus, Benetl, Clover ETL.

Therefore the data which is extracted from the source systems is temporarily stored and prepared for loading into the data warehouse

The data extraction process performs the following functions

Placement trend for a particular university

a) Data has to be selected CSE, MBA,MCA, ME , ECE, Pharamacy

1) Identify the sources of data

2) Finalize the filters that will be applied to every individual source systems to extract the data

3) Produce automatic extract files from the operational systems

4) Generate intermediary files to store selected data to be merged later

5) Render automated job control services to create extract files

6) Reformat and standardize the input from departmental data files database and spreadsheets

7) Produce common application code for data extraction

8) Resolve inconsistencies for common data that will be extracted from multiple source systems.

Few examples of Inconsistencies while integrating and extracting the data

	Sales Voucher	Purchase Order	Inventory
Description	Customer Name IBM	C_Name International Business Machine	CNAME IBM

Encoding	Gender
	1 = Male
	2 = Female
		Gender
		X = Male
		Y = Female
			Gender
			M= Male
			F = Female

Units	Cable Length Centimeters	Cable Length Meters	Cable Length Inches
Coding	Key Character (10)	Key Integer	Key ‘99999999’

Extraction –

1) Static Extraction – The snapshot of the operational data

2) Incremental Extraction –

Techniques used for data extraction

Immediate

a) Capture through transaction Logs –

This tech makes use of transaction logs of the DBMS. The transaction logs are maintained for recovery from possible failures. As each business transaction adds, updates or deletes a row from a database table, the dbms immediately updates the log files.

Capture through transaction logs technique reads the transaction log and selects all committed transactions. Since logs are already maintained in all DBMS so no extra overhead is incurred in OLTP systems.

This tech is best if we are using DBMS but if we are maintaining the data in the flat files then this tech cannot be used.

b) Capture through Database triggers

The database triggers are stored procedures which gets fired when one or more event occur in a DBMS.

However maintaining trigger is also an additional overhead in OLTP systems

c) Capture in source applications

In this technique the source application is used to capture the data for the data warehouse. All relevant applications that write to a source files are modified to write all add updates and deletes to both the source and dbs tables

This tech works well for all types of source data (dbs, indexed files, flat files and all other files)

Deferred Data Extraction

a) Capture based on date and timestamp

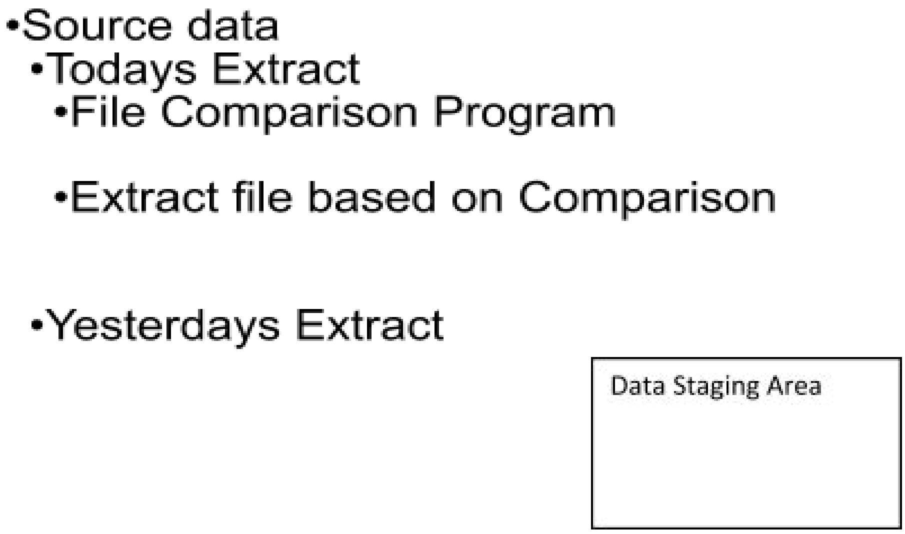
Every time a record is created in a database system it is marked with a timestamp that will be used for selecting the records for data extraction. The timestamp shows the date and time and unlike immediate data extraction this mechanism takes place at a later point of time and not while each record is created or updated.

Provided that the timestamps are created and maintained this tech works well with each types of file (Flat, Indexed, DBS)

b) Capture by comparing files

If none of the technique discussed above suits the data warehouse env or are not feasible for some or the other reason then this tech can be taken as the last option.

This technique is also known as snapshot differential technique because it compares two snapshots of the source data.



•Source data

•Todays Extract

•File Comparison Program

•Extract file based on Comparison

•Yesterdays Extract

Data Staging Area

Intermediate Stages of ETL

1) Data Mapping

It is the process of generating data element mapping between two distinct models.

It is the first process that is performed for a variety of data integration tasks which include

a) Data transformation between data source and data destination

b) Identification of data relationships

c) Discovery of hidden sensitive data

d) Consolidation of multiple databases into a single database

2) Data Staging

A data staging area can be defined as intermediate storage area that falls between the operational/transactional sources of data and the data warehouse (DW) or data mart (DM).

DATA STAGING AREA

Data staging area is the place where all the extracted data is temporarily stored and prepared for loading into the data warehouse. It is rightly compared to an assembly plant where

the extracted files are examined

business rules are reviewed

the data transformation functions are performed

data is stored and merged

inconsistencies are resolved

data is cleansed

Finally the data is processed and prepared for the enterprise wide view

At data staging are we perform three functions (Extraction, Transformation and Loading (ETL)

1) Cycle Initiation

2) Build Reference Data   Data Mapping

3) Extract Actual Data   Data Staging Area

4) Validate

5) Transform

6) Audit Reports

7) Publish (Load into target tables)

8) Archive

9) Clean up

Data Transformation

Before moving the data into the datawarehouse various transformations have to be performed. Since this data is coming from several disparate sources so therefore there is a strong need to transform the data according to a standardized format.

Tasks involved in data transformation

Format Revision

Decoding of Fields

Splitting of Fields

Merging of information

Character set conversion

Conversion of Units

Date and time of Conversions

Summarization

Key Restructuring

De-Duplication

Role of data Transformation Process

Map the input data from the source systems to warehouse

Clean the data

Remove duplicats

Perform splitting and merging of fields

Sort the records

De-normalize the extracted data