**The information flow mechanism**

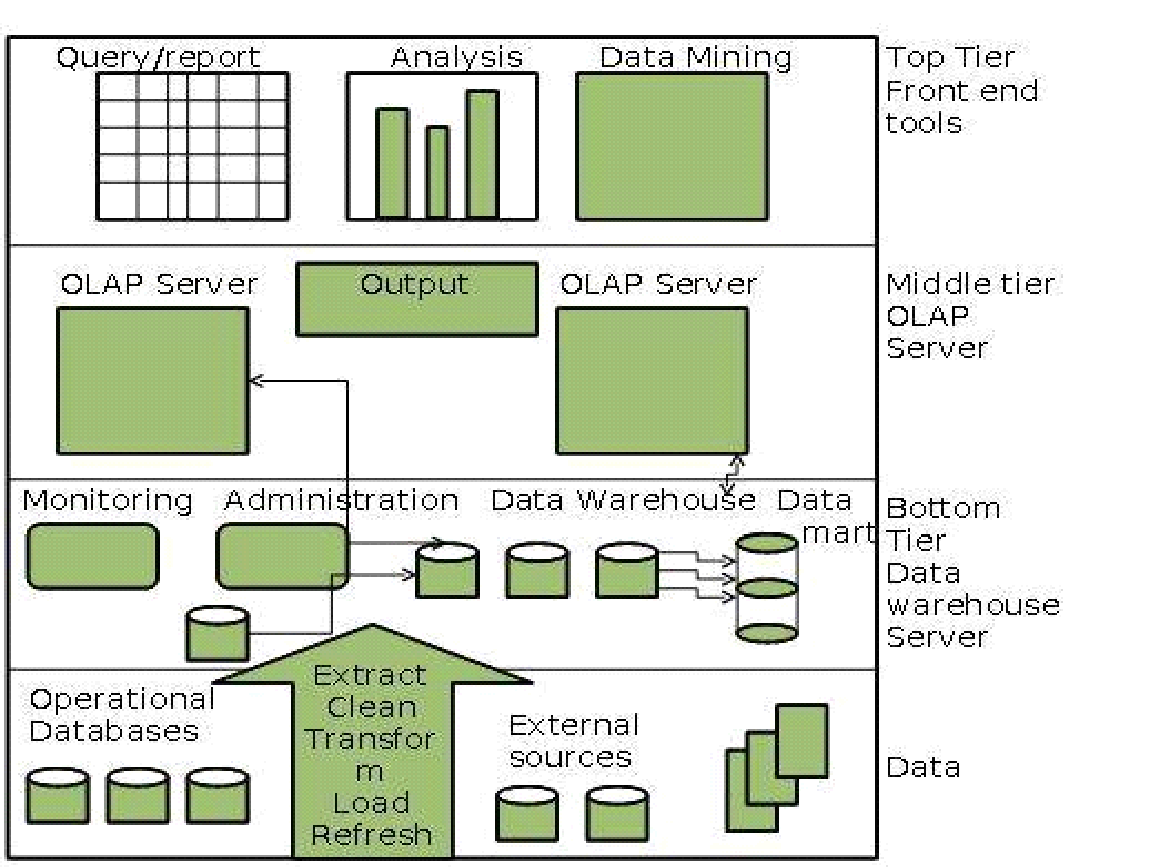
**1) Select the Source data**

**2) Extract the data from source systems**

3) Transform the extracted data

4) Load the transformed data into the data warehouse

5) Deliver the information to the end users



**Select the source data**

**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’ |

**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)**

**Extraction –**

**1) Static Extraction – The snapshot of the operational data**

**2) Incremental Extraction –**

**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).

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**