**Datawarehouse Concepts**

**Data Warehouse:**

As per Bill Inmon "A warehouse is a Historical, subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process".

By Historical we mean, the data is continuously collected from sources and loaded in the warehouse. The previously loaded data is not deleted for long period of time. This results in building historical data in the warehouse.

By Subject Oriented we mean data grouped into a particular business area instead of the business as a whole.

By Integrated we mean, collecting and merging data from various sources. These sources could be disparate in nature.

By Time-variant we mean that all data in the data warehouse is identified with a particular time period.

By Non-volatile we mean, data that is loaded in the warehouse is based on business transactions in the past, hence it is not expected to change over time.

An implementation of an informational database used to store sharable data sourced from an operational database of record. It is typically a subject database that allows users to tap into a company’s vast store of operational data to track and respond to business trends and facilitate forecasting and planning efforts. It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is relevant to some moment in time. The data warehouse contains atomic data and lightly summarized data.

**Data Warehousing (DWH)**

Data warehousing is the vast field that includes the processes and methodologies involved in the creating, populating, querying and maintaining a data warehouse. The objective of DWH is to help users make better business decisions.

**Enterprise Data Warehouse (EDW)**

Enterprise Data Warehouse (EDW) is a central normalized repository containing multiple subject area data for an organization. EDW are generally used for feeding data to subject area specific data marts. We must consider building an EDW when

1. We have clarity on multiple business processes requirements.

2. Want a centralized architecture catering to single version of truth.

An EDW is a database that is subject-oriented, integrated, non-volatile (read-only) and time-variant (historical). It is usually used by financial analysts for historical trend analysis reporting, data mining and other activities that need historical data. An EDW keeps growing as you add more historical snapshots, daily, weekly or monthly. Because an EDW has historical data (and the ODS usually does not), some companies use the EDW as a hub for loading their data marts.

CIF stands for Corporate Information Factory in the context of data warehousing. CIF is an overall architecture that includes ODS, EDW, data marts, oper marts, exploration warehouse, Meta data repository, extract, transfer and load (ETL) processes, portals, OLAP and BI applications like dashboards and scorecards, and so on. Since the question was about data mart consolidation, the other option you have is to create a normalized persistent staging area (PSA) database, which is not accessible by end users but serves as a hub for loading your current and future data marts.

Enterprise Data Warehouse is a centralized warehouse which provides service for the entire enterprise. A data warehouse is by essence a large repository of historical and current transaction data of an organization. An Enterprise Data Warehouse is a specialized data warehouse which may have several interpretations.

Several terms used in information technology have been used by a so many different vendors, IT workers and marketing ad campaigns that has left many confused about what really the term Enterprise Data Warehouse means and what makes it different from a general data warehouse.

Enterprise Data Warehouse has emerged from the convergence of opportunity, capability, infrastructure and need for data which has exponentially increased during the last few years as technology has advanced too fast and Business Enterprises tried to do their best to catch up and be on the top of the industry competition.

**Enterprise data warehouses replace data marts**

Businesses are replacing departmental databases - known as data marts - with enterprise data warehouses designed to pool information across the company.

"Data marts are very expensive," said Gartner analyst Donald Feinberg. "If you have six departments each with its own data mart, you end up with six hardware systems and six database licenses. You also need people who can maintain each data mart."

Businesses often find they end up with the same information replicated in each data mart, he said. They no longer have a single master copy of the data and must spend more on storage.

"In an enterprise data warehouse, data quality is higher," said Feinberg. "A data warehouse project does not have to take a 'big bang' approach. You can start on a small project and design it around a data warehouse."

**Data Mart**

When we restrict the scope of a data warehouse to a sub set of business process or subject area we call it a Data Mart. Data Marts can be stand-alone entities, built from warehouses or be used to build warehouse.

1. Stand-Alone Data Marts – These are single subject areas data warehouses. Stand alone data marts could cause many stove-pipe/silo data marts in the future.
2. Data Marts fed from Enterprise Data Warehouses (Top-Down). This is the Top-Down strategy suggested by Bill Inmon. Here we build an enterprise data warehouse (EDW) hosting data for multiple subject areas. From the EDW we build subject area specific data marts.
3. Data Marts building Data Warehouses (Bottom-Up). This is the Bottom-Up strategy suggested by Ralph Kimball. Here we build subject area specific data marts first, keeping in mind the shared business attributes (called conformed dimensions). Using the conformed dimensions, we build a warehouse.
4. A subset of a data warehouse that focuses on one or more specific subject areas. The data is usually extracted from the data warehouse and further denormalised and indexed to support intense usage by targeted customers.

**Legacy Systems:**

Legacy system is the term given to a system having very large storage capabilities. Mainframes and AS400 are the two examples of legacy systems. Applications running on legacy systems are generally OLTP systems. Due to their large storage capabilities the business transactional data from legacy applications and other OLTP systems is stored on them and maintained for a long period of time. In the world of DWH, legacy systems are often used to extract past historical business data.

**Data Profiling:**

This is a very important activity mostly performed while collecting the requirements and the system study. Data Profiling is a process that familiarizes you with the data you will be loading into the warehouse/mart.

In the most basic form, data profiling process would read the source data and report on the

Type of data – Whether data is Numeric, Text, Date etc.

Data statistics – Minimum and maximum values, most occurred value.

Data Anomalies like Junk values, incorrect business codes, invalid characters, values outside a given range.

Using data profiling we can validate business requirements with respect to business codes and attributes present in the sources, define exceptional cases when we get incorrect or inconsistent data.

A lot of Data Profiling tools are available in the market. Some examples are Trillium Software and IBM Websphere Data Integrator’s Profile Stage.

In the simplest form of data profiling, we can even hand code (programming) data profile application specific to projects. Programming languages like PERL, Java, PLSQL and Shell Scripting can be very handy in creating these applications.