1. You are the staging area expert on the project team for a large toy manufacturer. Discuss the four modes of applying data to the data warehouse. Select the modes you want to use for your data warehouse and explain the reasons for your selection.

Data may be applied in the following four different modes: load, append, destructive merge, and constructive merge. Working of each mode is explained below:

Load: If the target table to be loaded already exists and data exists in the table, the load process wipes out the existing data and applies the data from the incoming file. If the table is already empty before loading, the load process simply applies the data from the incoming file.

For a manufacturing company, full loads are needed by the toy manufacturer to initially load Master data of items in factory in a location, bill of Material by model of toy etc.

Append: Append can be treated as an extension of the load. If data already exists in the table, the append process unconditionally adds the incoming data, preserving the existing data in the target table. When an incoming record is a duplicate of an already existing record, you may define how to handle an incoming duplicate. The incoming record may be allowed to be added as a duplicate. In the other option, the incoming duplicate record may be rejected during the append process.

Operations can be appended to a particular routing from a factory situated in a particular location. Further resources needed to perform a routing operation can also be appended using this mode.

Destructive Merge: In this mode, you apply the incoming data to the target data. If the primary key of an incoming record matches with the key of an existing record, update the matching target record. If the incoming record is a new record without a match with any existing record, add the incoming record to the target table.

Count of Toys/items whose manufacturing is in progress and listed by toy/item. Such a staging requirement would need the counts of in progress items to be overridden with the new Count at this point in time.

Constructive Merge: This mode is slightly different from the destructive merge. If the primary key of an incoming record matches with the key of an existing record, leave the existing record, add the incoming record, and mark the added record as superseding the old record.

This mode is used where model/Item revisions need to have the latest revisions updated and marked for the same Item key.

2. Assume that you are the data quality expert on the data warehouse project team for a large financial institution with many legacy systems dating back to the 1970's. Review the types of data quality problems you are likely to have and make suggestions on how to deal with those

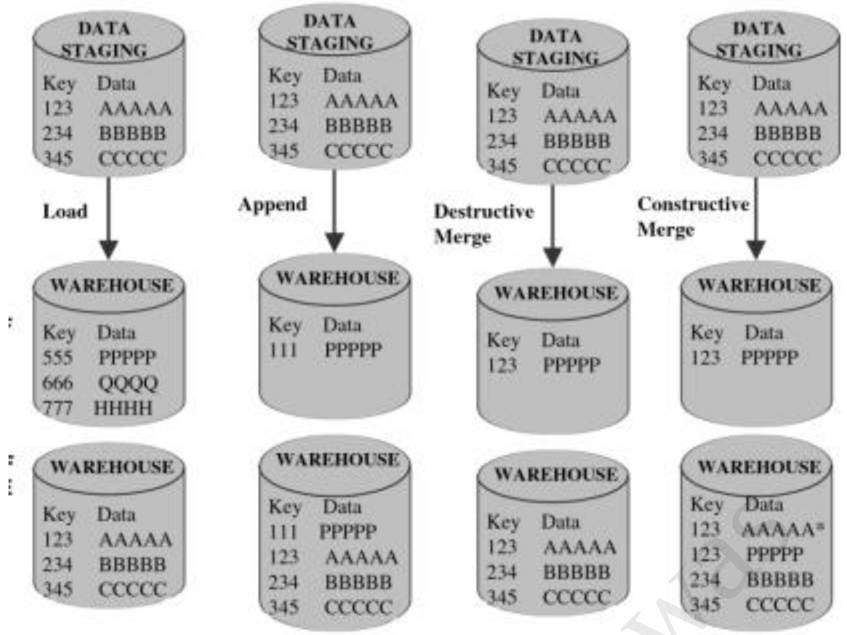

Fig: Modes of applying data

2. Assume that you are the data quality expert on the data warehouse project team for a large financial institution with many legacy systems dating back to the 1970's. Review the types of data quality problems you are likely to have and make suggestions on how to deal with those.

As a data quality expert, I might encounter the following types of data quality problems.

¤ Dummy values in source system fields:

¤ Absence of data in source system fields

¤ Multipurpose fields

¤ Cryptic data

¤ Contradicting data

¤ Improper use of name and address lines

¤ Violation of business rules

¤ Reused primary keys

¤ Non-unique identifiers

In spite of the enormous importance of data quality, many companies don't often pay much attention to data quality. The following suggestions can help deal with the types of data quality problems listed above. First, you may opt to let only clean data into your data warehouse. This means only data with a 100% quality can be loaded into the data warehouse. Data that is in any way polluted must be cleansed before it can be loaded. This is an ideal approach, but it takes a while to detect incorrect data and even longer to fix it. This approach is ideal from the point of view of data quality, but it will take a very long time before all data is cleaned up for data loading.

The second approach is a "clean as you go" method. In this method, you load all the data as is into the data warehouse and perform data cleansing operations in the data warehouse at a later time. Although you do not withhold data loads, the results of any query are suspect until the data gets cleansed. Depending on needs of the data warehouse, one can opt any of the two suggestions.

3. Compare the usage and value of information in the data warehouse with those in operational systems. Explain the major differences. Discuss and give examples.

The difference relates to two aspects of the information contained in these databases. First, they differ in the usage of the information. Next, they differ in the value of the information.
the contents and locate what they want.

Difference in usage: Users go to the data warehouse to find information on their own. The users formulate their own queries and run them. They format their own reports, run them, and receive the results. Some users may use predefined queries and preformatted reports but, by and large, the data warehouse is a place where the users are free to make up their own queries and reports. They move around the contents and perform their own analysis, viewing the data in ever so many different ways. Each time a user goes to the data warehouse, he or she may run different queries and different reports, not repeating the earlier queries or reports. The information delivery is interactive.

While because of efficiency considerations, operational systems are not designed to let users loose on the systems. The users may impact the performance of the system adversely with runaway queries. Another important point is that the users of operational systems do not exactly know the contents of the databases and metadata or data dictionary entries are typically unavailable to them. Interactive analysis, which forms the bedrock of information delivery in the data warehouse, is almost never present in an operational system.

Difference in value: The value of information from an operational system enables the users to monitor and control the current operations. On the other hand, information from the data warehouse gives the users the ability to analyze growth patterns in revenue, profitability, market penetration, and customer base. Based on such analysis, the users are able to make strategic decisions to keep the enterprise competitive and sound.

Since, information delivery from the data warehouse is markedly different from information delivery from operational systems, you should not try to apply the principles of information delivery from operational systems to the data warehouse.

4. As a senior analyst on the project team of a publishing company exploring the options for a data warehouse, make a case for OLAP. Describe the merits of OLAP and how it will be essential in your environment.

Analytical data processing for the Publishing company is key to analyzing subscription trends, performance of Distributors, publications, articles and subscriptions etc.

OLAP allows business users like Printing, Web publishing and other departments to slice and dice data at will. OLAP (or Online Analytical Processing) has been growing in popularity due to the increase in data volumes and the recognition of the business value of analytics. It uses database tables (fact and dimension tables) to enable multidimensional viewing, analysis and querying of large amounts of data. E.g. OLAP technology could provide the Publisher's management team with fast answers to complex queries on their operational data or enable them to analyze the Publishers company's historical data for trends and patterns.

Normally data in an organization is distributed in multiple data sources and are incompatible with each other. A retail example: Point-of-sales data and sales made via call-center or the Web are stored in different location and formats. It would a time consuming process for an executive to obtain OLAP reports such as - What are the most popular products purchased by customers between the ages 15 to 30?

Part of the OLAP implementation process involves extracting data from the various data repositories and making them compatible. Making data compatible involves ensuring that the meaning of the data in one repository matches all other repositories. An example of incompatible data: Customer ages can be stored as birth date for purchases made over the web and stored as age categories (i.e. between 15 and 30) for in store sales. The major OLAP vendors are Hyperion, Cognos, Business Objects,

5. Prepare an outline for a standards manual for your data warehouse. Consider all types of objects and their naming conventions. Indicate why standards are important. Produce a detailed table of contents.

Significance of standards: Standards in a data warehouse environment cover a wide range of objects, processes, and procedures. With regard to the physical model, the standard for naming the objects take on special significance. Standards provide a consistent means for communication. As we know, users are more directly involved in accessing information from a data warehouse than they are in an OLTP environment, effective communication must take place among the members of the project and standards help establish such effective communication**.**

Naming of Database Objects:

Components of Object Names**:** The name itself must be able to convey the meaning and description of the object. For example, look at the name of a column: customer_loan_balance. This naming convention immediately identifies the column as containing values of balance amounts.

Word Separators: Standardize the separators that are also called the delineators. Dashes (-) or underscores (_) are commonly used. If your DBMS has specific conventions or requirements, follow those conventions.

Names in Logical and Physical Models. Names for objects such as tables and attributes may include both the logical model versions and the physical model versions. You need naming standards for both versions. Analysts and logical model designers communicate with each other through the logical model names. When the users need to refer to the tables and columns for data retrieval, they are communicating at the level of the physical model. Therefore, you need to adapt the standards for the physical model for the users.

Naming of Files and Tables in the Staging Area.

Indicate the Process. Identify the process to which the file relates. If the file is the output from a transformation step, let the name of the file denote that. If the file is part of the daily incremental update, let that be clear from the name of the file.

Express the Purpose. Suppose you are setting up the scheduling of the weekly update to the product dimension table. You need to know the input load file for this purpose. If the name of the file indicates the purpose for which it was created, that will be a big help when you are setting up the update schedule. Develop standards for the staging area files to include the purpose of the file in the name. For example, sale_units_daily_stage, customer_daily_update, product_full_refresh

Standards for Physical Files. Your standards must include naming conventions for all types of files. These files are not restricted to data and index files for the data warehouse database. There are other files as well. Establish standards for the following:

Files holding source codes and scripts

Database files

Application documents

Saudi Telecom – Questions for Discussion

1. **Why do you think telecommunications companies are among the prime users of information visualization tools?**

Telecommunications companies typically have millions of customers, contacting them constantly for billing, payment, network usage, and support. All of this information has to be monitored in a timely fashion.

2. How did Saudi Telecom use information visualization?

STC used visualization to monitor information regarding network statistics, service analytics, and customer calls. Part of this requires prioritizing and contextualizing the data, identifying the relevant metrics, and viewing KPIs.

3. What were their challenges, the proposed solution, and the obtained results?

STC needed to identify the relevant metrics, properly visualize them, and provide them to the right people, often with time-sensitive information. But executives didn’t have the ability to see key performance indicators in a timely fashion. They would have to contact the technical teams to get status reports, which often came too late and to be of benefit for preventing problems before they happen. The solution was to contract with Dundas business intelligence consultants to refine the telecommunication dashboards and improve their functionality. This led to engagement on an enterprise-wide, mission-critical project to transform their data center and create a more proactive monitoring environment. Dundas’ information visualization tools allowed STC to see trends and correct issues before they became problems. This resulted in a decrease in the amount of service tickets by 55 percent.

.

.