7. THE ARCHITECTURAL COMPONENTS

REVIEW QUESTIONS

Q.1.What is your understanding of data warehouse architecture? Describe in one or two paragraphs.

Ans: The structure that brings all the components of a data warehouse together is known as the architecture. For example, take the case of the architecture of a school building. The architecture of the building is not just the visual style. It includes the various classrooms, of fices, library, corridors, gymnasiums, doors, windows, roof, and a large number of other such components. When all of these components are brought and placed together, the structure that ties all of the components together is the architecture of the school building. If you can extend this comparison to a data warehouse, the various components of the data warehouse together form the architecture of the data warehouse.

In your data warehouse, architecture includes a number of factors. Primarily, it includes the integrated data that is the center piece. The architecture includes everything that is needed to prepare the data and store it. On the other hand, it also includes all the means for delivering information from your data warehouse. The architecture is further composed of the rules, procedures, and functions that enable your data warehouse to work and ful-fill the business requirements. Finally, the architecture is made up of the technology that empowers your data warehouse.

What is the general purpose of the data warehouse architecture? The architecture provides the overall framework for developing and deploying your data warehouse; it is a comprehensive blueprint. The architecture defines the standards, measurements, general design, and support techniques.

Q.2. What are the three major areas in the data warehouse? Is this a logical division? If so, why do you think so? Relate the architectural components to the three major areas.

Ans: As you already know, the three major areas in the data warehouse are: Data acquisition Data storage Information delivery

The following are the major building blocks of the data warehouse:

- Source data
- Data staging
- Data storage
- Information delivery
- Metadata Management and control.

Figure 7-1 groups these major architectural components into the three areas. In this chapter, we will study the architecture as it relates to these three areas. In each area, we will consider the supporting architectural components. Each of the components has definite functions and provides specific services.

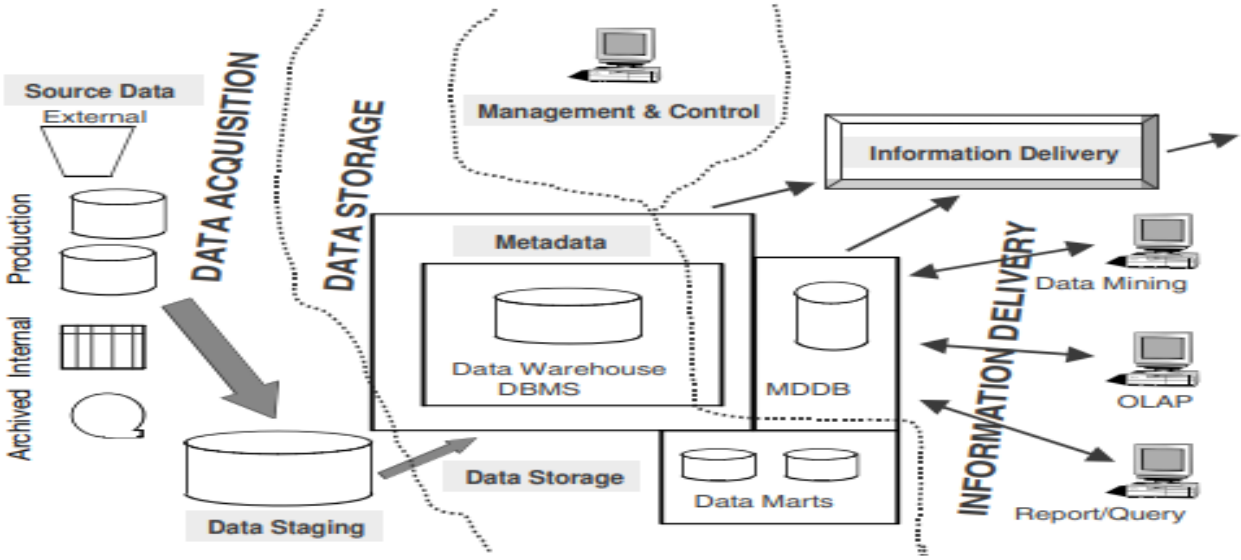


Figure 7-1 Architectural components in the three major areas.

Q.3.Name four distinguishing characteristics of data warehouse architecture. Describe each briefly.

Ans: Different Objectives and Scope

The architecture has to support the requirements for providing strategic information. Strategic information is markedly different from information obtained from operational systems. When you provide information from an operational application, the information content and quantity per user session is limited. As an example, at a particular time, the user is interested only in information about one customer and all the related orders. From a data warehouse, however, the user is interested in obtaining large result sets.

. Basically, the extent to which a decision support system is different from an operational system directly translates into just one essential principle: a data warehouse must have a different and more elaborate architecture

Data Content

The “read-only” data in the data warehouse sits in the middle as the primary component in the architecture. In an operational system, although the database is important, this importance does not measure up to that of a data warehouse data repository. Before data is brought into your data warehouse and stored as read-only data, a number of functions must be performed.

These exhaustive and critical functions do not compare with the data conversion that happens in an operational system. In your data warehouse, you keep data integrated from multiple sources. After extracting the data, which by itself is an elaborate process, you transform the data, cleanse it, and integrate it in a staging area. Only then you move the integrated data into the data warehouse repository as read-only data. Operational data is not “read-only” data.

Complex Analysis and Quick Response

Your data warehouse architecture must support complex analysis of the strategic information by the users. Information retrieval processes in an operational system dwindle in complexity when compared to the use of information from a data warehouse. Most of the online information retrieval during a session by a user is interactive analysis. A user does not run an isolated query, go away from the data warehouse, and come back much later for the next single query. A session by the user is continuous and lasts a long time because the user usually starts with a query at a high level, reviews the result set, initiates the next query looking at the data in a slightly different way, and so on. Your data warehouse architecture must, therefore, support variations for providing analysis. Users must be able to drill down, roll up, slice and dice data, and play with “what-if ” scenarios.

Users must have the capability to review the result sets in different output options. Users are no longer content with textual result sets or results displayed in tabular formats. Every result set in tabular format must be translated into graphical charts

Metadata-driven As the data moves from the source systems to the end-users as useful, strategic information, metadata surrounds the entire movement. The metadata component of the architecture holds data about every phase of the movement, and, in a true sense, makes the movement happen. In an operational system, there is no component that is equivalent to metadata in a data warehouse. The data dictionary of the DBMS of the operational system is just a faint shadow of the metadata in a data warehouse. So, in your data warehouse architecture, the metadata component interleaves with and connects the other components.

Q.4. Trace the flow of data through the data warehouse from beginning to end.

Ans: Please look at Figure 7-2. This figure shows the flow of data from beginning to end and also highlights the architectural components enabling the flow of data as the data moves along. Let us now follow the flow of the data. At each stop along the passage, let us identify the architectural components. Some of the architectural components govern the flow of data from beginning to end. The management and control module is one such component. This module touches every step along the data movement.

At the Data Source. Here the internal and external data sources form the source data architectural component. Source data governs the extraction of data for preparation and

storage in the data warehouse. The data staging architectural component governs the transformation, cleansing, and integration of data.

In the Data Warehouse Repository. The data storage architectural component includes the loading of data from the staging area and also storing the data in suitable formats for information delivery. The metadata architectural component is also a storage mechanism to contain data about the data at every point of the flow of data from beginning to end. At the User End. The information delivery architectural component includes dependent data marts, special multidimensional databases, and a full range of query and reporting facilities.

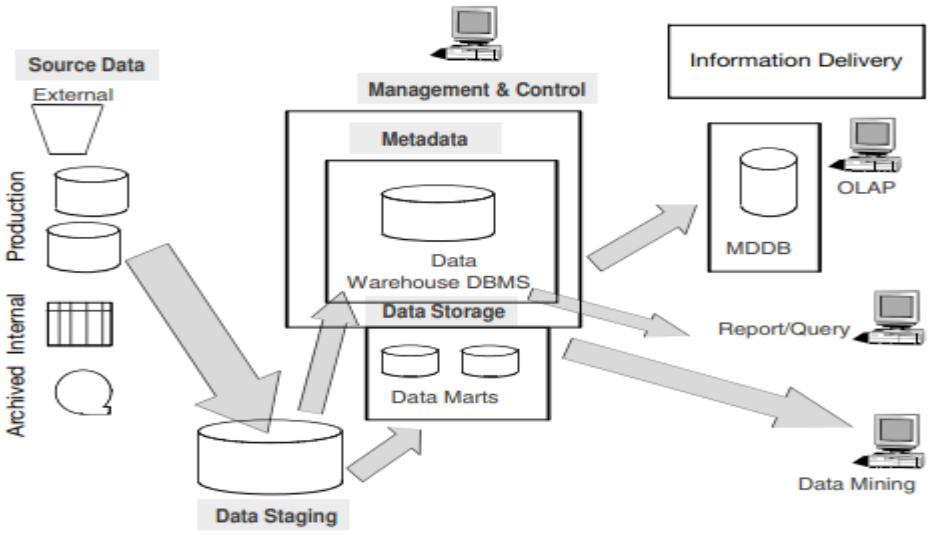


Figure 7-2 Architectural framework supporting the flow of data.

Q.5. For information delivery, what is the difference between top-down and bottom-up approaches to data warehouse implementation?

Ans: For information delivery, the data flow begins at the enterprise-wide data warehouse and the dependent data marts when the design is based on the top-down technique. When the design follows the bottom-up method, the data flow starts at the set of conformed data marts. Generally, data transformed into information flows to the user desktops during query sessions. Also, information printed on regular or ad hoc reports reaches the users. Sometimes, the result sets from individual queries or reports are held in proprietary data stores of the query or reporting tool vendors. The stored information may be put to faster repeated use. In many data warehouses, data also flows into specialized downstream decision support applications such as executive information systems (EIS) and data mining. The other more common flow of information is to proprietary multidimensional databases for OLAP.

Q.6. In which architectural component does OLAP fit in? What is the function of OLAP?

Ans: This area spans a broad spectrum of many different methods of making information available to users. For your users, the information delivery component is the data

warehouse. They do not come into contact with the other components directly. For the users, the strength of your data warehouse architecture is mainly concentrated in the robustness and flexibility of the information delivery component. The information delivery component makes it easy for the users to access the information either directly from the enterprise-wide data warehouse, from the dependent data marts, or from the set of conformed data marts. Most of the information access in a data warehouse is through online queries and interactive analysis sessions. Nevertheless, your data warehouse will also be producing regular and ad hoc reports. Almost all modern data warehouses provide for online analytical processing (OLAP). In this case, the primary data warehouse feeds data to proprietary multidimensional databases (MDDBs) where summarized data is kept as multidimensional cubes of information. The users perform complex multidimensional analysis using the information cubes in the MDDBs. Refer to Figure 7-6 for a summarized view of the technical architecture for information delivery.

Q.7. Define technical architecture of the data warehouse. How does it relate to the individual architectural components?

Ans: The technical architecture of a data warehouse is, therefore, the complete set of functions and services provided within its components. The technical architecture also includes the procedures and rules that are required to perform the functions and provide the services. The technical architecture also encompasses the data stores needed for each component to provide the services.

The architecture is not the set of tools needed to perform the functions and provide the services. When we refer to the data extraction function within one of the architectural components, we are simply mentioning the function itself and the various tasks associated with that function. Also, we are relating the data store for the staging area to the data extraction function because extracted data is moved to the staging area. Notice that there is no mention of any tools for performing the function. Where do the tools fit in? What are the tools for extracting the data? What are tools in relation to the architecture? Tools are the means to implement the architecture. That is why you must remember that architecture comes first and the tools follow.

When you establish the architecture for your data warehouse, you will prepare the architectural plan that will include all the components. The plan will also state in detail the extent and complexity of all the functions, services, procedures, and data stores related to each architectural component. The architectural plan will serve as the blueprint for the design and development. It will also serve as a master checklist for your tool selection.

Q.8.List five major functions and services in the data storage area.

Ans: List of Functions and Services Load data for full refreshes of data warehouse tables Perform incremental loads at regular prescribed intervals Support loading into multiple tables at the detailed and summarized levels Optimize the loading process Provide

automated job control services for loading the data warehouse Provide backup and recovery for the data warehouse database Provide security Monitor and fine-tune the database Periodically archive data from the database according to pre-set conditions

Q.9. What are the types of storage repositories in the data staging area?

Ans: This is the place where all the extracted data is put together and prepared for loading into the data warehouse. The staging area is like an assembly plant or a construction area. In this area, you examine each extracted file, review the business rules, perform the various data transformation functions, sort and merge data, resolve inconsistencies, and cleanse the data. When the data is finally prepared either for an enterprise wide data warehouse or one of the conformed data marts, the data temporarily resides in the staging area repository awaiting to be loaded into the data warehouse repository. In a large number of data warehouses, data in the staging area is kept in sequential or flat files. These flat files, however, contain the fully integrated and cleansed data in appropriate formats ready for loading. Typically, these files are in the formats that could be loaded by the utility tools of the data warehouse RDBMS. Now more and more staging area data repositories are becoming relational databases. The data in such staging areas are retained for longer periods. Although extracts for loading may be easily obtained from relational databases with proper indexes, creating and maintaining these relational databases involves overhead for index creation and data migration from the source systems. The staging area may contain data at the lowest grain to populate tables containing business measurements. It is also common for aggregated data to be kept in the staging area for loading. The other types of data kept in the staging area relate to business dimensions such as product, time, sales region, customer, and promotional schemes.

Q.10. List four major functions and services for information delivery. Describe each briefly.

Ans:

- Provide security to control information access
- Monitor user access to improve service and for future enhancements
- Allow users to browse data warehouse content
- Simplify access by hiding internal complexities of data storage from users
- Automatically reformat queries for optimal execution
- Enable queries to be aware of aggregate tables for faster results
- Govern queries and control runaway queries
- Provide self-service report generation for users, consisting of a variety of flexible options to create, schedule, and run reports
- Store result sets of queries and reports for future use
- Provide multiple levels of data granularity

- Provide event triggers to monitor data loading
- Make provision for the users to perform complex analysis through online analytical processing (OLAP)
- Enable data feeds to downstream, specialized decisions support systems such as EIS and data mining

EXERCISES

1. Indicate if true or false:

A. Data warehouse architecture is just an overall guideline. It is not a blueprint for the data warehouse.

Ans: False

B. In a data warehouse, the metadata component is unique, with no truly matching component in operational systems.

Ans: True

C. Normally, data flows from the data warehouse repository to the data staging area.

Ans: False

D. The management and control component does not relate to all operations in a data warehouse.

Ans: False

E. Technical architecture simply means the vendor tools.

Ans: False

F. SQL-based languages are used to extract data from hierarchical databases.

Ans: False

G. Sorts and merges of files are common in the staging area.

Ans: True

H. MDDBs are generally relational databases.

Ans: False

I. Sometimes, results of individual queries are held in temporary data stores for repeated use.

Ans: True

J. Downstream specialized applications are fed directly from the source data component.

Ans: False

2. You have been recently promoted to administrator for the data warehouse of a nationwide automobile insurance company. You are asked to prepare a checklist for selecting a proper vendor tool to help you with the data warehouse administration. Make a list of the functions in the management and control component of your data warehouse architecture. Use this list to derive the tool selection checklist.

Ans:
The following are the functions listed in the management and control component of the data warehouse architecture:
• Data Quality Checks.
• Managing and updating of metadata.
• Backup and recovery of data.
• Purging data.
• Data warehouse storage management.
• Information delivery function.
• Priority and security management.
• Monitoring updates of data from multiple sources.
• Distributing and sub-setting data.
• Auditing and reporting data warehouse usage and status.
• Data replication.
According to the information listed above, the best vendor tool will have to able to meet the majority of the most significant management and control components of the data warehouse function (Pasyeka & Pasyeka, 2016). The best vendor tool will, therefore, need to fulfil the above checklist. The list of the appropriate vendors that currently provide proper tools includes Amazon, IBM infosphere, Teradata and Oracle. For an organization with demanding data needs and an increasing amount of data to handle, the proper industry established data warehouse administration applications such as applications such as IBM Infosphere and Oracle. These prove to be the most reliable and operational data warehouse administrator tools that can be employed in an organization of dynamic nature.

3. As the senior analyst responsible for data staging, you are responsible for the design of the data staging area. If your data warehouse gets input from several legacy systems on multiple platforms, and also regular feeds from two external sources, how will you organize your data staging area? Describe the data repositories you will have for data staging.

Ans:

Data staging area is resting area for the data before being loaded into the data warehouse. In this area the data extracted from the source is examined, reviewed, and transformed to fit in the storage. Various operations like sorting, merging, resolving inconsistencies, cleansing of data, providing backup and recovery for staging area recovery, and preserving audit trail for maintaining data relations are done in this area. The data usually is kept into the flat files after being extracted from the source and ready to be loaded into the warehouse repository.

For organizing the data staging area that gets inputs from several legacy systems (mainframes, mini, UNIX) and feeds from external sources the single system infrastructure for the staging area will not work. A hybrid option needs to be considered in-order to perform the staging functions. A brief detail of the process and how staging area should be organized is provided below.

Once the data is extracted from multiple sources (legacy systems and external feeds) and before it reaches the staging area, initial reformatting and merging (into smaller number), reconciliation (record count with source), and cleansing (populating missing and default values) needs to be done. Now because we have the source platform as legacy systems therefore it is prudent to perform these operations on one of the source platform (say mainframes) only.

Because we are getting the data from legacy systems and as extracts (flat files) therefore it is prudent to perform the operations on one of the source platform (say mainframes) only. Various functions should be performed in the staging area:

Transformation and consolidation: The integration of data from various sources is done. In our case we are getting the data from multiple legacy platforms and our staging is done on one of the legacy systems. Other legacy system formats and extracts should be converted and integrated as per the staging area platform (say mainframes) format.

Validation and Quality check: A final quality audit of transformed data (against source) should be done so as to ensure the data integrity.

Creation of load image: Finally, images for individual files on the data warehouse repository should be created keeping in mind the platform of the storage. In our case I assume that the platform is one of the legacy systems (day mainframes), hence there will be no issue for data load.

For the store the repository can be product mater, employee master, stores, customer master, and offers etc.

4. You are the data warehouse architect for a leading national department store chain. The data warehouse has been up and running for nearly a year. Now the management has decided to provide the power users with OLAP facilities. How will you alter the

information delivery component of your data warehouse architecture? Make realistic assumptions and proceed.

Ans:

OLAP helps in reducing the load of the database and also improves the utilization rate of the data with analyzing the data.
Some of the ways in which the data can be shifted to OLAP is as follows:
The data gets collected by OLTP first which gets synchronized to OLAP as OLTP is short yet rapid as well as high frequency querying and hence feeds the data to the data warehouse whereas OLAP analyzes the data.

There are methods which are used when a synchronization takes place:
1) Full synchronization where the data from among the entire table are put to OLAP every time.
2) Increamental /synchronization where the procedure synchronizes the modified data for the OLAP table every time.

5. You recently joined as the data extraction specialist on the data warehouse project team developing a conformed data mart for a local but progressive pharmacy. Make a detailed list of functions and services for data extraction, data transformation, and data staging.

Ans:

Functions and Services. Please review the general list of functions and services given in this section. The list relates to the data acquisition area and covers the functions and services in three groups. This is a general list. It does not indicate the extent or complexity of each function or service. For the technical architecture of your data warehouse, you have to determine the content and complexity of each function or service.

List of Functions and Services

Data Extraction

Select data sources and determine the types of filters to be applied to individual sources

- Generate automatic extract files from operational systems using replication and other techniques
- Create intermediary files to store selected data to be merged later Transport extracted files from multiple platforms
- Provide automated job control services for creating extract files
- Reformat input from outside sources
- Reformat input from departmental data files, databases, and spreadsheets
- Generate common application code for data extraction
- Resolve inconsistencies for common data elements from multiple sources

Data Transformation

- Map input data to data for data warehouse repository
- Clean data, deduplicate, and merge/purge
- Denormalize extracted data structures as required by the dimensional model of the data warehouse
- Convert data types
- Calculate and derive attribute values
- Check for referential integrity
- Aggregate data as needed
- Resolve missing values
- Consolidate and integrate data

Data Staging

- Provide backup and recovery for staging area repositories
- Sort and merge files
- Create files as input to make changes to dimension tables
- If data staging storage is a relational database, create and populate database
- Preserve audit trail to relate each data item in the data warehouse to input source
- Resolve and create primary and foreign keys for load tables
- Consolidate datasets and create flat files for loading through DBMS utilities
- If staging area storage is a relational database, extract load files

8. INFRASTRUCTURE AS THE FOUNDATION FOR DATA WAREHOUSING

REVIEW QUESTIONS

Q.1. What is the composition of the operational infrastructure of the data warehouse? Why is operational infrastructure equally as important as the physical infrastructure?

Ans: To understand operational infrastructure, let us once again take the example of data staging. One part of foundational infrastructure refers to the computing hardware and the related software. You need the hardware and software to perform the data staging functions and render the appropriate services. You need software tools to perform data transformations. You need software to create the output files. You need disk hardware to place the data in the staging area files. But what about the people involved in performing these functions? What about the business rules and procedures for the data transformations? What about the management software to monitor and administer the data transformation tasks? Operational infrastructure to support each architectural component consists of People Procedures Training Management software These are not the people and procedures needed for developing the data warehouse. These are the ones needed to keep the data warehouse going. These elements are as essential as the hardware and software that keep the data warehouse running. They support the management of the data warehouse and maintain its efficiency. Data warehouse developers pay a lot of attention to the hardware and system software elements of the infrastructure. It is right to do so. But operational infrastructure is often neglected. Even though you may have the right hardware and software, your data warehouse needs the operational infrastructure in place for proper functioning. Without appropriate operational infrastructure, your data warehouse is likely to just limp along and cease to be effective. Pay attention to the details of your operational infrastructure.

Q.2. List the major components of the physical infrastructure. Write two or three sentences to describe each component.

Ans: Figure 8-2 highlights the major elements of physical infrastructure. What do you see in the diagram? As you know, every system, including your data warehouse, must have an overall platform on which to reside. Essentially, the platform consists of the basic hardware components, the operating system with its utility software, the network, and the network software. Along with the overall platform is the set of tools that run on the selected platform to perform the various functions and services of individual architectural components. We

will examine the elements of physical infrastructure in the next few sections. Decisions about the hardware top the list of decisions you have to make about the infrastructure of your data warehouse. Hardware decisions are not easy. You have to consider many factors. You have to ensure that the selected hardware will support the entire data warehouse architecture. Perhaps we can go back to our mainframe days and get some helpful hints. As newer models of the corporate mainframes were announced and as we ran out of steam on the current configuration, we stuck to two principles. First, we leveraged as much of the existing physical infrastructure as possible. Next, we kept the infrastructure as modular as possible. When needs arose and when newer versions became available at cheaper prices, we unplugged an existing component and plugged in the replacement

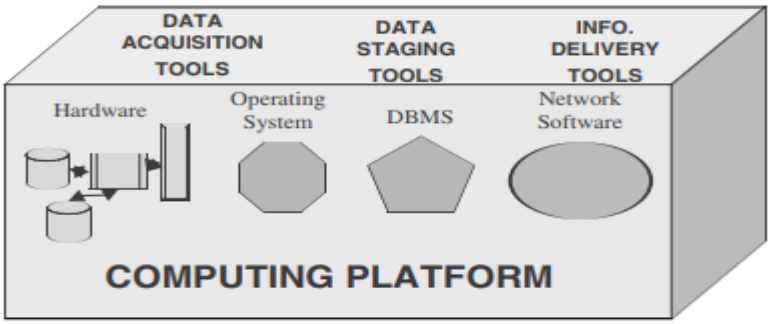


Figure 8-2 Physical infrastructure.

Q.3. Briefly describe any six criteria you will use for selecting the operating system for your data warehouse.

Ans: Scalability. When your data warehouse grows in terms of the number of users, the number of queries, and the complexity of the queries, ensure that your selected hardware could be scaled up. Support. Vendor support is crucial for hardware maintenance. Make sure that the support from the hardware vendor is at the highest possible level.

Vendor Stability. Check on the stability and staying power of the vendor. Next let us quickly consider a few general criteria for the selection of the operating system. First of all, the operating system must be compatible with the hardware. A list of criteria follows.

Scalability. Again, scalability is first on the list because this is one common feature of every data warehouse. Data warehouses grow, and they grow very fast. Along with the hardware and database software, the operating system must be able to support the increase in the number of users and applications.

Security. When multiple client workstations access the server, the operating system must be able to protect each client and associated resources. The operating system must provide each client with a secure environment. Reliability. The operating system must be able to protect the environment from application malfunctions.

Availability. This is a corollary to reliability. The computing environment must continue to be available after abnormal application terminations

Q.4. What are the platform options for the staging area? Compare the options and mention the advantages and disadvantages.

Ans: In One of Legacy Platforms. If most of your legacy data sources are on the same platform and if extra capacity is readily available, then consider keeping your data staging area in that legacy platform. In this option, you will save time and effort in moving the data across platforms to the staging area.

On the Data Storage Platform. This is the platform on which the data warehouse DBMS runs and the database exists. When you keep your data staging area on this platform, you will realize all the advantages for applying the load images to the database. You may even be able to eliminate a few intermediary substeps and apply data directly to the database from some of the consolidated files in the staging area.

On a Separate Optimal Platform. You may review your data source platforms, examine the data warehouse storage platform, and then decide that none of these platforms are really suitable for your staging area. It is likely that your environment needs complex data transformations. It is possible that you need to work through your data thoroughly to cleanse and prepare it for your data warehouse.

Here are some distinct advantages of a separate platform for data staging: You can optimize the separate platform for complex data transformations and data cleansing. What do we mean by this? You can gear up the neutral platform with all the necessary tools for data transformation, data cleansing, and data formatting. While the extracted data is being transformed and cleansed in the data staging area, you need to keep the entire data content and ensure that nothing is lost on the way. You may want to think of some tracking file or table to contain tracking entries. A separate environment is most conducive for managing the movement of data. We talked about the possibility of having specialized tools to manipulate the data in the staging area. If you have a separate computing environment for the staging area, you could easily have people specifically trained on these tools running the separate computing equipment.

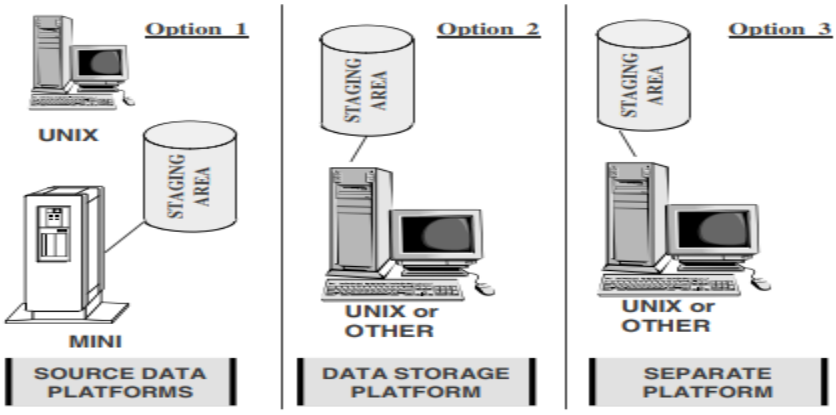


Figure 8-5 Platform options for the staging area.

Q.5. What are the four common methods for data movement within the data warehouse? Explain any two of these methods.

Ans: Shared Disk. This method goes back to the mainframe days. Applications running in different partitions or regions were allowed to share data by placing the common data on a shared disk. You may adapt this method to pass data from one step to another for data acquisition in your data warehouse. You have to designate a disk storage area and set it up so that each of the two platforms recognizes the disk storage area as its own.

Mass Data Transmission. In this case, transmission of data across platforms takes place through data ports. Data ports are simply inter platform devices that enable massive quantities of data to be transported from one platform to the other. Each platform must be configured to handle the transfers through the ports. This option calls for special hard ware, software, and network components. There must also be sufficient network bandwidth to carry high data volumes.

Real-Time Connection. In this option, two platforms establish connection in real time so that a program running on one platform may use the resources of the other platform. A program on one platform can write to the disk storage on the other. Also, jobs running on one platform can schedule jobs and events on the other. With the widespread adoption of TCP/IP, this option is very viable for your data warehouse.

Manual Methods. Perhaps these are the options of last resort. Nevertheless, these options are straightforward and simple. A program on one platform writes to an external medium such as tape or disk. Another program on the receiving platform reads the data from the external medium.

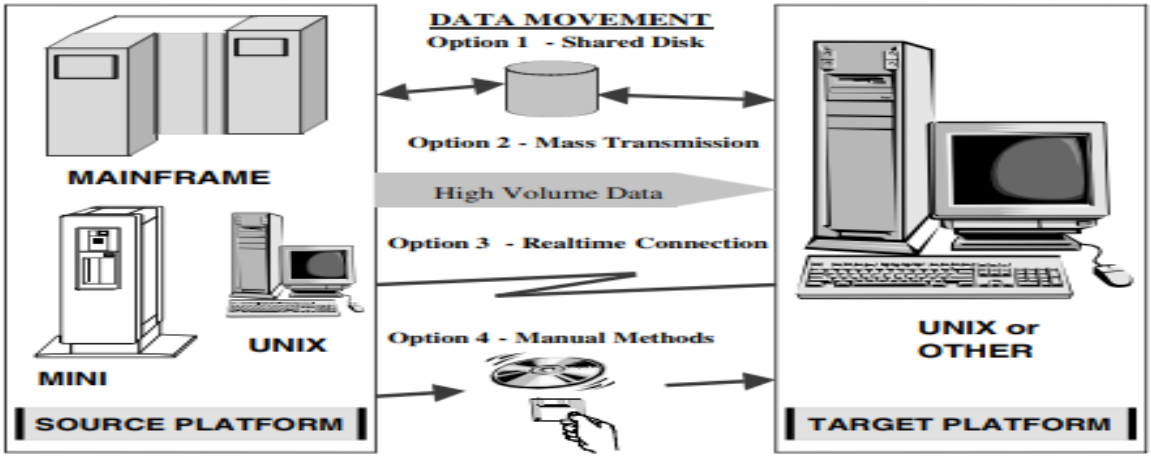


Figure 8-6 Data movement options.

Q.6. Write two brief paragraphs on the considerations for client workstations.

Ans: When you are ready to consider the configurations for the workstation machines, you will quickly come to realize that you need to cater to a variety of user types. We are only considering the needs at the workstation with regard to information delivery from the data warehouse. A casual user is perhaps satisfied with a machine that can run a Web browser to access HTML reports. A serious analyst, on the other hand, needs a larger and more

powerful workstation machine. The other types of users between these two extremes need a variety of services. Do you then come up with a unique configuration for each user? That will not be practical. It is better to determine a minimum configuration on an appropriate platform that would support a standard set of information delivery tools in your data warehouse. Apply this configuration for most of your users. Here and there, add a few more functions as necessary. For the power users, select another configuration that would support tools for complex analysis. Generally, this configuration for power users also supports OLAP. The factors for consideration when selecting the configurations for your users' workstations are similar to the ones for any operating environment. However, the main consideration for workstations accessing the data warehouse is the support for the selected set of tools. This is the primary reason for the preference of one platform over another. Use this checklist while considering workstations: Workstation operating system Processing power Memory Disk storage Network and data transport Tool support

Q.7. What are the four parallel server hardware options? List the features, benefits, and limitations of any one of these options.

Ans: The four parallel server hardware options are:

- SMP (Symmetric Multiprocessing).
- Clusters.
- MPP (Massively Parallel Processing).

ccNUMA or NUMA (Cache-coherent Nonuniform Memory Architecture).

SMP (Symmetric Multiprocessing). Refer to Figure 8-11. Features: This is a shared-everything architecture, the simplest parallel processing machine. Each processor has full access to the shared memory through a common bus. Communication between processors occurs through common memory. Disk controllers are accessible to all processors. Benefits: This is a proven technology that has been used since the early 1970s. Provides high concurrency. You can run many concurrent queries. Balances workload very well. Gives scalable performance. Simply add more processors to the system bus. Being a simple design, you can administer the server easily

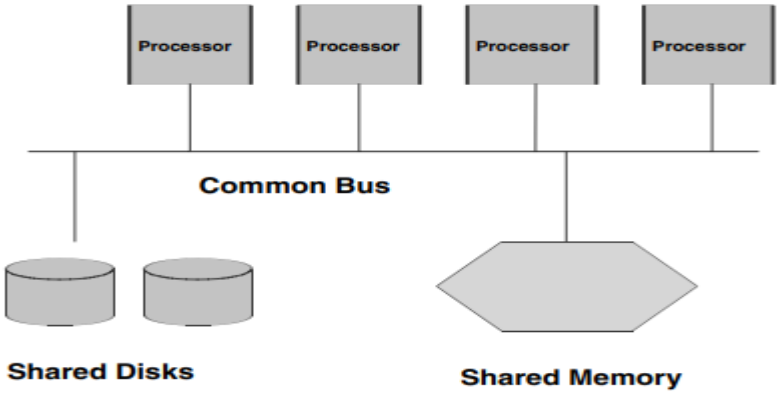


Figure 8-11 Server hardware option: SMP.

Limitations: Available memory may be limited. May be limited by bandwidth for processor-to-processor communication, I/O, and bus communication. Availability is limited; like a single computer with many processors. You may consider this option if the size of your data warehouse is expected to be around a two or three hundred gigabytes and concurrency requirements are reasonable.

Q.8. How have the RDBMS vendors enhanced their products for data warehousing? Describe briefly in one or two paragraphs.

Ans: Examine the features of the leading commercial RDBMSs. As data warehousing becomes more prevalent, you would expect to see data warehouse features being included in the software products. That is exactly what the database vendors are doing. Data-warehouserelated add-ons are becoming part of the database offerings. The database software that started out for use in operational OLTP systems is being enhanced to cater to decision support systems. DBMSs have also been scaled up to support very large databases. Some RDBMS products now include support for the data acquisition area of the data warehouse. Mass loading and retrieval of data from other database systems have become easier. Some vendors have paid special attention to the data transformation function. Replication features have been reinforced to assist in bulk refreshes and incremental loading of the data warehouse. Bit-mapped indexes could be very effective in a data warehouse environment to index on fields that have a smaller number of distinct values. For example, in a database table containing geographic regions, the number of distinct region codes is few. But frequently, queries involve selection by regions. In this case, retrieval by a bit-mapped index on the region code values can be very fast. Vendors have strengthened this type of indexing. We will discuss bit-mapped indexing further in Chapter 18. Apart from these enhancements, the more important ones relate to load balancing and query performance. These two features are critical in a data warehouse. Your data warehouse is query-centric. Everything that can be done to improve query performance is most desirable. The DBMS vendors are providing parallel processing features to improve query performance.

Q.9.What is intraquery parallelization by the DBMS? What are the three methods?

Ans: Intraquery Parallelization. We will use Figure 8-15 for our discussion of intraquery parallelization, so please take a quick look and follow along. This will greatly help you in matching up your choice of server hardware with your selection of RDBMS. Let us say a query from one of your users consists of an index read, a data read, a data join, and a data sort from the data warehouse database. A serial processing DBMS will process this query in the sequence of these base operations and produce the result set. However, while this query is executing on one processor in the SMP system, other queries can execute in parallel. This method is the interquery parallelization discussed above. The first group of operations in Figure 8-15 illustrates this method of execution. Using the intraquery parallelization technique, the DBMS splits the query into the lower-level operations of index read, data read, data join, and data sort. Then each one of these basic operations is executed in parallel on a single processor. The final result set is the consolidation of the intermediary results. Let us review three ways a DBMS can provide intraquery parallelization, that is, parallelization of parts of the operations within the same query itself.

Horizontal Parallelism. The data is partitioned across multiple disks. Parallel processing occurs within each single task in the query, for example, data read, which is performed on multiple processors concurrently on different sets of data to be read from multiple disks. After the first task is completed from all of the relevant parts of the partitioned data, the next task of that query is carried out, and then the next one after that task, and so on. The problem with this approach is the wait until all the needed data is read. Look at Case A in Figure 8-15.

Vertical Parallelism. This kind of parallelism occurs among different tasks, not just a single task in a query as in the case of horizontal parallelism. All component query operations are executed in parallel, but in a pipelined manner. This assumes that the RDBMS has the capability to decompose the query into subtasks; each subtask has all the operations of index read, data read, join, and sort. Then each subtask executes on the data in serial fashion. In this approach, the database records are ideally processed by one step and immediately given to the next step for processing, thus avoiding wait times. Of course, in this method, the DBMS must possess a very high level of sophistication in decomposing tasks. Now, please look at Case B in Figure 8-15.

Hybrid Method. In this method, the query decomposer partitions the query both horizontally and vertically. Naturally, this approach produces the best results. You will realize the greatest utilization of resources, optimal performance, and high scalability. Case C in Figure 8-15 illustrates this method.

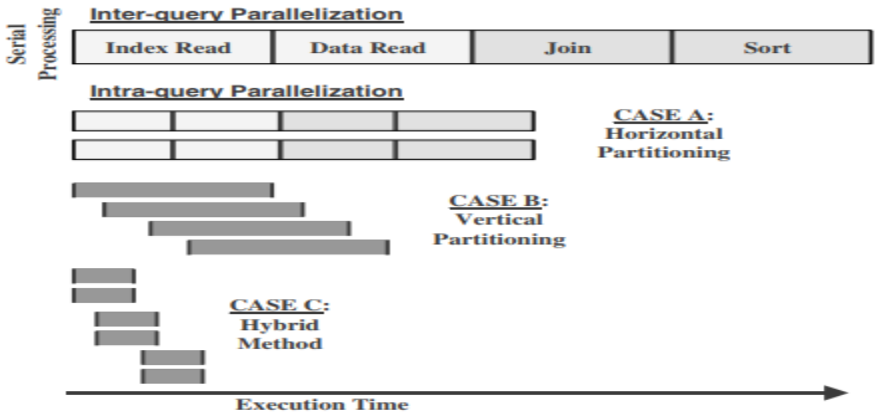


Figure 8-15 Intraquery parallelization by DBMS.

Q.10. List any six types of software tools used in the data warehouse. Pick any three types from your list and describe the features and the purposes.

Ans:

- Data Loading
- Data Quality
- Queries and Reports
- Online Analytical Processing (OLAP)
- Alert Systems
- Middleware and Connectivity
- Data Warehouse Management
- Data Transformation
- Data Extraction
- Data Modeling

Data Transformation Transform extracted data into appropriate formats and data structures. Provide default values as specified. Major features include field splitting, consolidation, standardization, and deduplication. Data Loading Load transformed and consolidated data in the form of load images into the data warehouse repository. Some loaders generate primary keys for the tables being loaded. For load images available on the same RDBMS engine as the data warehouse, precoded procedures stored on the database itself may be used for loading. Data Quality Assist in locating and correcting data errors. May be used on the data in the staging area or on the source systems directly. Help resolve data inconsistencies in load images.

EXERCISES

1. Match the columns:

1. operational infrastructure	**F. people, procedures, training**
2. preemptive multitasking	**D. operating system feature**
3. shared disk	**J. data movement option**
4. MPP	**A. shared-nothing architecture**
5. SMP	**G. easy administration**
6. interquery parallelization	**E. vertical parallelism**
7. intraquery parallelization	**B. provides high concurrency**
8. NUMA	**C. single memory address space**
9. UNIX-based system	**H. choice data warehouse platform**
10. data staging area	**I. optimize for data transformation**

2. In your company, all the source systems reside on a single UNIX-based platform, except one legacy system on a mainframe computer. Analyze the platform options for your data warehouse. Would you consider the single-platform option? If so, why? If not, why not?

Ans:

This is the most straightforward and simplest option for implementing the data warehouse architecture. In this option, all functions from the backend data extraction to the front-end query processing are performed on a single computing platform. This was perhaps the earliest approach, when developers were implementing data warehouses on existing mainframes, minicomputers, or a single UNIX-based server. Because all operations in the data acquisition, data storage, and information delivery areas take place on the same platform, this option hardly ever encounters any compatibility or interface problems. The data flows smoothly from beginning to end without any platform-to-platform conversions. No middleware is needed. All tools work in a single computing environment. In many companies, legacy systems are still running on mainframes or minis. Some of these companies have migrated to UNIX-based servers and others have moved over to ERP systems in client/server environments as part of the transition to address the Y2K challenge. In any case, most legacy systems still reside on mainframes, minis, or UNIX based servers. What is the relationship of the legacy systems to the data warehouse? Remember, the legacy systems contribute the major part of the data warehouse data. If these companies wish to adopt a single-platform solution, that platform of choice has to be a mainframe, mini, or a UNIX-based server. If the situation in your company warrants serious consideration of the

single-platform option, then analyze the implications before making a decision. The single-platform solution appears to be an ideal option. If so, why are not many companies adopting this option now? Let us examine the reasons.

Legacy Platform Stretched to Capacity.

In many companies, the existing legacy computing environment may have been around for a couple of decades and already fully stretched to capacity. The environment may be at a point where it can no longer be upgraded further to accommodate your data warehouse.

Nonavailability of Tools. Software tools form a large part of the data warehouse infrastructure. You will clearly grasp this fact from the last few subsections of this chapter. Most of the tools provided by the numerous data warehouse vendors do not support the mainframe or minicomputer environment. Without the appropriate tools in the infrastructure, your data warehouse will fall apart.

Multiple Legacy Platforms. Although we have surmised that the legacy mainframe or minicomputer environment may be extended to include data warehousing, the practical fact points to a different situation. In most corporations, a combination of a few mainframe systems, an assortment of minicomputer applications, and a smattering of the newer PC-based systems exist side by side. The path most companies have taken is from mainframes to minis and then to PCs. Figure 8-3 highlights the typical configuration. If your corporation is one of the typical enterprises, what can you do about a singleplatform solution? Not much. With such a conglomeration of disparate platforms, a single-platform option having your data warehouse alongside all the other applications is just not tenable.

3. You are the manager for the data warehouse project of a nationwide car rental company. Your data warehouse is expected to start out in the 500 GB range. Examine the options for server hardware and write a justification for choosing one.

Ans:
As the manager of the data warehouse I have to make sure that the server has to be scalable and extensible. For this to be achieved i have to consider the following for server hardware; RAM, version of the data version of the SQL server, the size of the server and processor. The data within the SQL Server will be loaded and processed into company server's RAM before processing alongside it, for the process to be faster and reliable then RAM capacity has to be considered depending on the SQL server. It is vital to note down which version of SQL Server that the company will be using when defining the server's hardware concerning RAM supported by the server for better performance of the company.
When carrying out the sizing the server, it is ideally the car rental company wants something that is going to last for long. Depending on the size of the database the company will have to purchase a server with a hard drive that will accommodate 500GB of data. Car Rental Company has to also recommend 8 to 16 cores of the processor for better performance of the server

4. As the administrator of the proposed data warehouse for a hotel chain with a leading presence in ten eastern states, write a proposal describing the criteria you will use to select the RDBMS for your data warehouse. Make your assumptions clear.

Ans:

Selection of the DBMS

Our discussions of the server hardware and the DBMS parallel processing options must have convinced you that selection of the DBMS is most crucial. You must choose the server hardware with the appropriate parallel architecture. Your choice of the DBMS must match with the selected server hardware. These are critical decisions for your data warehouse.

While discussing how business requirements drive the design and development of the data warehouse in Chapter 6, we briefly mentioned how requirements influence the Inter-query Parallelization Index Read Data Read Join Sort Intra-query Parallelization Execution Time Serial Processing CASE A: Horizontal Partitioning CASE B: Vertical Partitioning CASE C: Hybrid Method Figure 8-15 Intra query parallelization by DBMS. Selection of the DBMS. Apart from the criteria that the selected DBMS must have load balancing and parallel processing options, the other key features listed below must be considered when selecting the DBMS for your data warehouse.

Query governor—to anticipate and abort runaway queries

Query optimizer—to parse and optimize user queries

Query management—to balance the execution of different types of queries

Load utility—for high-performance data loading, recovery, and restart

Metadata management—with an active data catalog or dictionary

Scalability—in terms of both number of users and data volumes

Extensibility—having hybrid extensions to OLAP databases

Portability—across platforms

Query tool APIs—for tools from leading vendors

Administration—providing support for all DBA functions

5. You are the Senior Analyst responsible for the tools in the data warehouse of a large local bank with branches in only one state. Make a list of the types of tools you will provide for use in your data warehouse. Include tools for developers and users. Describe the features you will be looking for in each tool type.

Ans:

The types of software tools for your data warehouse. As mentioned earlier, more details will be added in the later chapters. These chapters will also elaborate on individual tool types. In the following subsections, we mention the basic purposes and features of the type of tool indicated by the title of each subsection.

Data Modeling

- Enable developers to create and maintain data models for the source systems and the data warehouse target databases. If necessary, data models may be created for the staging area.
- Provide forward engineering capabilities to generate the database schema.
- Provide reverse engineering capabilities to generate the data model from the data dictionary entries of existing source databases.
- Provide dimensional modeling capabilities to data designers for creating STAR schemas.

Data Extraction

- Two primary extraction methods are available: bulk extraction for full refreshes and change-based replication for incremental loads.
- Tool choices depend on the following factors: source system platforms and databases, and available built-in extraction and duplication facilities in the source systems.

Data Transformation

- Transform extracted data into appropriate formats and data structures.
- Provide default values as specified.
- Major features include field splitting, consolidation, standardization, and deduplication.

Data Loading

- Load transformed and consolidated data in the form of load images into the data warehouse repository.
- Some loaders generate primary keys for the tables being loaded.
- For load images available on the same RDBMS engine as the data warehouse, precoded procedures stored on the database itself may be used for loading.

Data Quality

- Assist in locating and correcting data errors.

- May be used on the data in the staging area or on the source systems directly.
- Help resolve data inconsistencies in load images.

Queries and Reports

- Allow users to produce canned, graphic-intensive, sophisticated reports.
- Help users to formulate and run queries.
- Two main classifications are report writers, report servers.

Online Analytical Processing (OLAP)

- Allow users to run complex dimensional queries.
- Enable users to generate canned queries.
- Two categories of online analytical processing are multidimensional online analytical processing (MOLAP) and relational online analytical processing (ROLAP). MOLAP works with proprietary multidimensional databases that receive data feeds from the main data warehouse. ROLAP provides online analytical processing capabilities from the relational database of the data warehouse itself.

Alert Systems

- Highlight and get user's attention based on defined exceptions.
- Provide alerts from the data warehouse database to support strategic decisions.
- Three basic alert types are: from individual source systems, from integrated enterprise-wide data warehouses, and from individual data marts.

Middleware and Connectivity

- Transparent access to source systems in heterogeneous environments.
- Transparent access to databases of different types on multiple platforms.
- Tools are moderately expensive but prove to be invaluable for providing interoperability among the various data warehouse components.

Data Warehouse Management

- Assist data warehouse administrators in day-to-day management.
- Some tools focus on the load process and track load histories.
- Other tools track types and number of user queries.

9. THE SIGNIFICANT ROLE OF METADATA

REVIEW QUESTIONS

Q.1. Why do you think metadata is important in a data warehouse environment? Give a general explanation in one or two paragraphs.

Ans: Metadata in a data warehouse contains the answers to questions about the data in the data warehouse. You keep the answers in a place called the metadata repository. Even if you ask just a few of data warehousing practitioners or if you read just a few of the books on data warehousing, you will receive seemingly different definitions for metadata. Here is a sample list of definitions: Data about the data Table of contents for the data Catalog for the data Data warehouse atlas Data warehouse roadmap Data warehouse directory Glue that holds the data warehouse contents together Tongs to handle the data The nerve center

So, what exactly is metadata? Which one of these definitions comes closest to the truth? Let us take a specific example. Assume your user wants to know about the table or entity called Customer in your data warehouse before running any queries on the customer data. What is the information content about Customer in your metadata repository? Let us review the metadata element for the Customer entity as shown in Figure 9-1. What do you see in the figure? The metadata element describes the entity called Customer residing the data warehouse. It is not just a description. It tells you more. It gives more than the explanation of the semantics and the syntax. Metatada describes all the pertinent aspects of the data in the data warehouse fully and precisely. Pertinent to whom? Pertinent primarily to the users and also to you as developer and part of the project team.

Entity Name:	Customer
Alias Names:	Account, Client

Definition:	A person or an organization that purchases goods or services from the company.
Remarks:	Customer entity includes regular, current, and past customers.
Source Systems:	Finished Goods Orders, Maintenance Contracts, Online Sales.

Create Date:	January 15, 1999
Last Update Date:	January 21, 2001
Update Cycle:	Weekly
Last Full Refresh Date:	December 29, 2000
Full Refresh Cycle:	Every six months
Data Quality Reviewed:	January 25, 2001
Last Deduplication:	January 10, 2001
Planned Archival:	Every six months
Responsible User:	Jane Brown

Figure 9-1 Metadata element for *Customer* entity.

Q.2. Explain how metadata is critical for data warehouse development and administration.

Ans: For Using the Data Warehouse. There is one big difference between a data warehouse and any operational system such as an order processing application. The difference is in the usage—the information access. In an order processing application, how do your users get information? You provide them with GUI screens and predefined reports. They get information about pending or back orders through the relevant screens. They get information about the total orders for the day from specific daily reports. You created the screens and you formatted the reports for the users. Of course, these were designed based on specifications from the users. Nevertheless, the users themselves do not create the screen formats or lay out the reports every time they need information. In marked contrast, users themselves retrieve information from the data warehouse. By and large, users themselves create ad hoc queries and run these against the data warehouse. They format their own reports. Because of this major difference, before they can create and run their queries, users need to know about the data in the data warehouse. They need metadata. In our operational systems, however, we do not really have any easy and flexible methods for knowing the nature of the contents of the database. In fact, there is no great need for user-friendly interfaces to the database contents. The data dictionary or catalog is meant for IT uses only. The situation for a data warehouse is totally different. Your data warehouse users need to receive maximum value from your data warehouse. They need sophisticated methods for browsing and examining the contents of the data warehouse. They need to know the meanings of the data items. You have to prevent them from drawing wrong conclusions from their analysis through their ignorance about the exact meanings. Earlier data mart implementations were limited in scope to probably one subject area. Mostly, those data marts were used by small groups of users in single departments. The users of those data marts were able to get by with scanty metadata. Today's data warehouses are

much wider in scope and larger in size. Without adequate metadata support, users of these larger data warehouses are totally handicapped.

For Building the Data Warehouse. Let us say you are the data extraction and transformation expert on the project team. You know data extraction methods very well. You can work with data extraction tools. You understand the general data transformation techniques. But, in order to apply your expertise, first you must know the source systems and their data structures. You need to know the structures and the data content in the data warehouse. Then you need to determine the mappings and the data transformations. So far, to perform your tasks in building the data extraction and data transformation component of the data warehouse, you need metadata about the source systems, source-to-target mappings, and data transformation rules. Try to wear a different hat. You are now the DBA for the data warehouse database. You are responsible for the physical design of the database and for doing the initial loading. You are also responsible for periodic incremental loads. There are more responsibilities for you. Even ignoring all the other responsibilities for a moment, in order to perform just the tasks of physical design and loading, you need metadata about a number of things. You need the layouts in the staging area. You need metadata about the logical structure of the data warehouse database. You need metadata about the data refresh and load cycles. This is just the bare minimum information you need. If you consider every activity and every task for building the data warehouse, you will come to realize that metadata is an overall compelling necessity and a very significant component in your data warehouse. Metadata is absolutely essential for building your data warehouse.

For Administering the Data Warehouse. Because of the complexities and enormous sizes of modern data warehouses, it is impossible to administer the data warehouse without substantial metadata. Figure 9-2 lists a series of questions relating to data warehouse administration. Please go through each question on the list carefully. You cannot administer your data warehouse without answers to these questions. Your data warehouse metadata must address these issues.

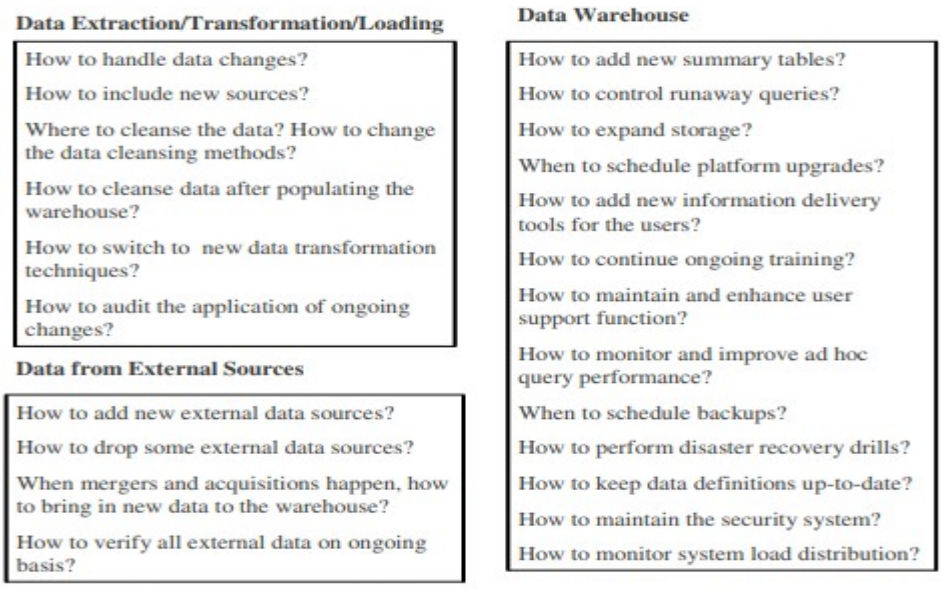


Figure 9-2 Data warehouse administration: questions and issues.

Q.3. Examine the concept that metadata is like a nerve center. Describe how the concept applies to the data warehouse environment.

Ans: Various processes during the building and administering of the data warehouse generate parts of the data warehouse metadata. Parts of metadata generated by one process are used by another. In the data warehouse, metadata assumes a key position and enables communication among various processes. It acts like a nerve center in the data warehouse. Figure 9-4 shows the location of metadata within the data warehouse. Use this figure to determine the metadata components that apply to your data warehouse environment. By examining each metadata component closely, you will also perceive that the individual parts of the metadata are needed by two groups of people: (1) end-users, and (2) IT (developers and administrators).

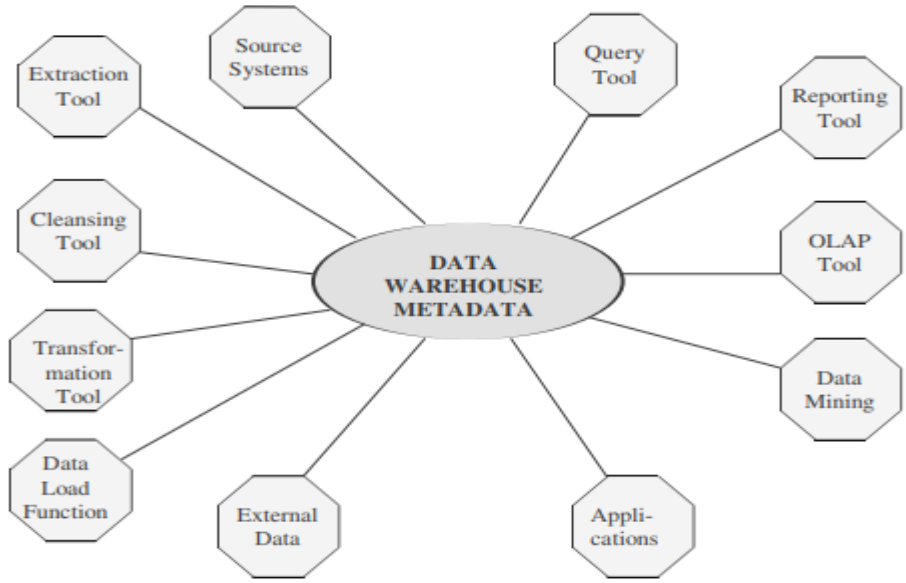


Figure 9-4 Metadata acts as a nerve center.

Q.4. List and describe three major reasons why metadata is vital for end-users

Ans: The Marketing VP of your company has asked this business analyst to do a thor ough analysis of a problem that recently surfaced. Because of the enormous sales potential in the Midwest and Northeast regions, your company has opened five new stores in each region. Although overall countrywide sales increased nicely for two months following the opening of the stores, after that the sales went back to the prior levels and remained flat. The Marketing VP wants to know why, so that she can take appropriate action. As a user, the business analyst expects to find answers from the new data warehouse, but he does not know the details about the data in the data warehouse. Specifically, he does not know the answers to the following questions: Are the sale units and dollars stored by individual transactions or as summary totals, by product, for each day in each store? Can sales be analyzed by product, promotion, store, and month? Can current month sales be compared to sales in the same month last year? Can sales be compared to targets? How is profit margin calculated? What are the business rules? What is the definition of a sales region? Which districts are included in each of the two regions being analyzed? Where did the sales come from? From which source systems? How old are the sales numbers? How often do these numbers get updated? If the analyst is not sure of the nature of the data, he is likely to interpret the results of the analysis incorrectly. It is possible that the new stores are cannibalizing sales from their own existing stores and that is why the overall sales remain flat. But the analyst may not find the right reasons because of misinterpretation of the results. The analysis will be more effective if you provide adequate metadata to help as a powerful roadmap of the data. If there is sufficient and proper metadata, the analyst does not have to get assistance from IT every time he needs to run an analysis. Easily accessible metadata is crucial for end-users. Let us take the analogy of an industrial warehouse storing items of merchandise sold through catalog. The customer refers to the catalog to find the

merchandise to be ordered. The customer uses the item number in the catalog to place the order. Also, the catalog indicates the color, size, and shape of the merchandise item. The customer calculates the total amount to be paid from the price details in the catalog. In short, the catalog covers all the items in the industrial warehouse, describes the items, and facilitates the placing of the order. In a similar way, the user of your data warehouse is like the customer. A query for information from the user is like an order for items of merchandise in the industrial warehouse. Just as the customer needs the catalog to place an order, so does your user need metadata to run a query on your data warehouse. Figure 9-5 summarizes the vital need of metadata for end-users. The figure shows the types of information metadata provides to the end-users and the purposes for which they need these types of information.

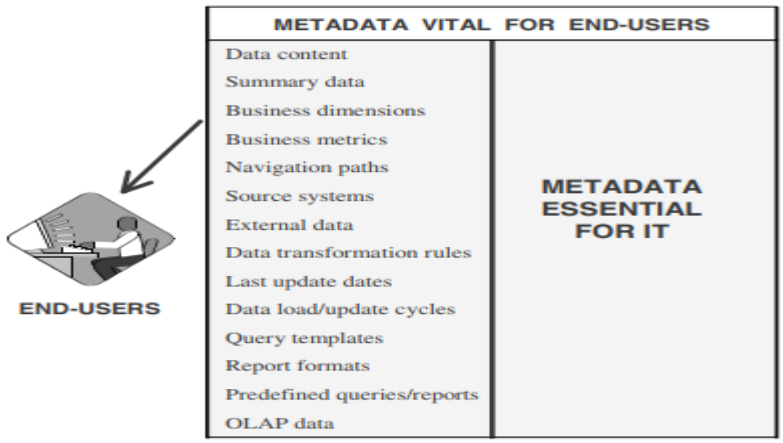


Figure 9-5 Metadata vital for end-users.

Q.5. Why is metadata essential for IT? List six processes in which metadata is significant for IT and explain why.

Ans: Development and deployment of your data warehouse is a joint effort between your IT staff and your user representatives. Nevertheless, because of the technical issues, IT is pri marily responsible for the design and ongoing administration of the data warehouse. For performing the responsibilities for design and administration, IT must have access to proper metadata. Throughout the entire development process, metadata is essential for IT. Beginning with the data extraction and ending with information delivery, metadata is crucial for IT. As the development process moves through data extraction, data transformation, data integration, data cleansing, data staging, data storage, query and report design, design for OLAP, and other front-end systems, metadata is critical for IT to perform their development activities. Here is a summary list of processes in which metadata is significant for IT: Data extraction from sources Data transformation Data scrubbing Data aggregation and summarization Data staging Data refreshment Database design Query and report design Figure 9-6 summarizes the essential need for metadata for IT. The figure shows the types of information metadata provides IT staff and the purposes for which they need these types of information.

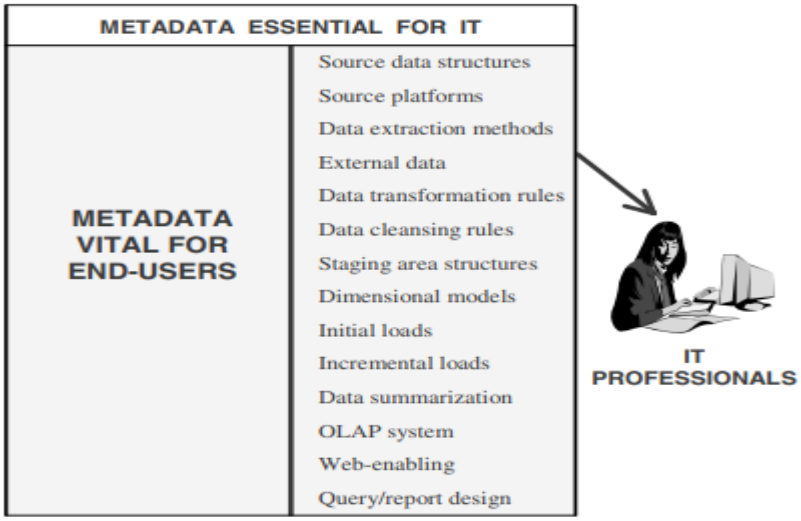


Figure 9-6 Metadata essential for IT.

Q.6. Pick three processes in which metadata assists in the automation of these processes. Show how metadata plays an active role in these processes.

Ans: Maintaining metadata is no longer a form of glorified documentation. Traditionally, metadata has been created and maintained as documentation about the data for each process. Now metadata is assuming a new active role. Let us see how this is happening. As you know, tools perform major functions in a data warehouse environment. For example, tools enable the extraction of data from designated sources. When you provide the mapping algorithms, data transformation tools transform data elements to suit the target data structures. You may specify valid values for data elements and the data quality tools will use these values to ensure the integrity and validity of data. At the front end, tools empower the users to browse the data content and gain access to the data warehouse. These tools generally fall into two categories: development tools for IT professionals, and information access tools for end-users. When you, as a developer, use a tool for design and development, in that process, the tool lets you to create and record a part of the data warehouse metadata. When you use another tool to perform another process in the design and development, this tool uses the metadata created by the first tool. When your end-user uses a query tool for information access at the front end, that query tool uses metadata created by some of the back-end tools. What exactly is happening here with metadata? Metadata is no longer passive documentation. Metadata takes part in the process. It aids in the automation of data warehouse processes. Let us consider the back-end processes beginning with the defining of the data sources. As the data movement takes place from the data sources to the data warehouse database through the data staging area, several processes occur. In a typical data warehouse, appro priate tools assist in these processes. Each tool records its own metadata as data movement takes place. The metadata recorded by one tool drives one or more processes that follow. This is how metadata assumes an active role and assists in the automation of data warehouse processes. Here is a list of back-end processes shown in the order in which they generally occur: 1. Source data structure

definition 2. Data extraction 3. Initial reformatting/merging 4. Preliminary data cleansing 5. Data transformation and consolidation 6. Validation and quality check 7. Data warehouse structure definition 8. Load image creation Figure 9-7 shows each of these eight processes. The figure also indicates the metadata recorded by each process. Further, the figure points out how each process is able to use the metadata recorded in the earlier processes. Metadata is important in a data warehouse because it drives the processes. However, our discussion above leads to the realization that each tool may record metadata in its own proprietary format. Again, the metadata recorded by each tool may reside on the platform where the corresponding process runs.

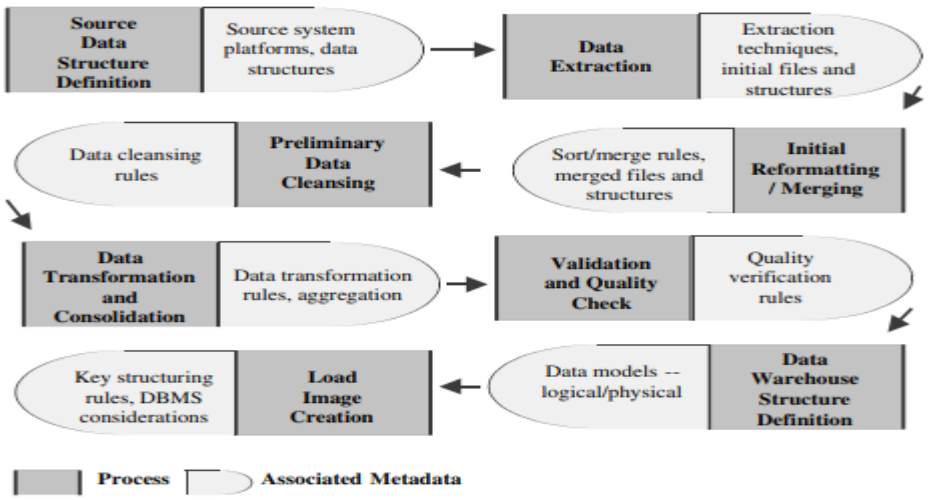


Figure 9-7 Metadata drives data warehouse processes.

Q.7. What is meant by establishing the context of information? Briefly explain with an example how metadata establishes the context of information in a data warehouse.

Ans: Imagine this scenario. One of your users wants to run a query to retrieve sales data for three products during the first seven days of April in the Southern Region. This user composes the query as follows:

Product = Widget-1 or Widget-2 or Widget-3

Region = 'SOUTH'

Period = 04-01-2000 to 04-07-2000

The result comes back:

	Sale Units	Amount
Widget-1—	25,355	253,550
Widget-2—	16,978	254,670
Widget-3—	7,994	271,796

Let us examine the query and the results. In the specification for region, which territories does region “SOUTH” include? Are these the territories your user is interested in? What is the context of the data item “SOUTH” in your data warehouse? Next, does the data item 04-01-2000 denote April 1, 2000 or January 4, 2000? What is the convention used for dates in your data warehouse?

Look at the result set. Are the numbers shown as sale units given in physical units of the products, or in some measure such as pounds or kilograms? What about the amounts shown in the result set? Are these amounts in dollars or in some other currency? This is a pertinent question if your user is accessing your data warehouse from Europe.

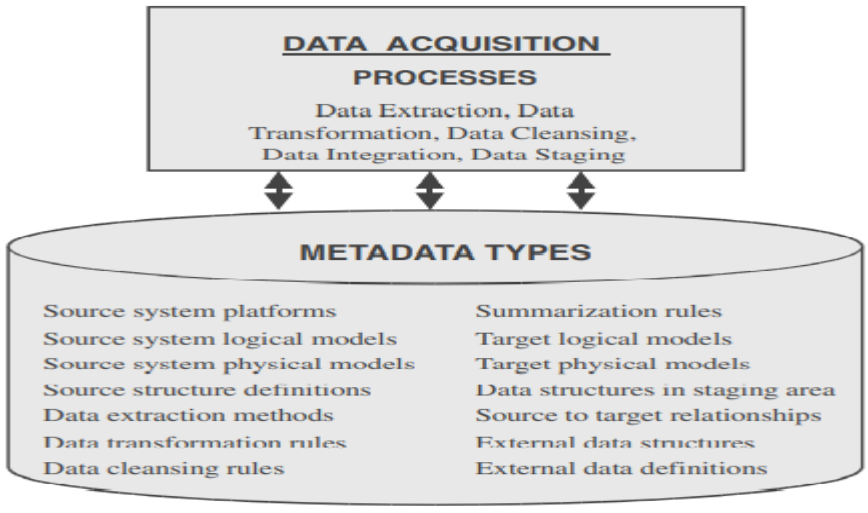
For the dates stored in your data warehouse, if the first two digits of the date format indicate the month and the next two digits denote the date, then 04-01-2000 means April 1, 2000. Only in this context is the interpretation correct. Similarly, context is important for the interpretation of the other data elements.

How can your user find out what exactly each data element in the query is and what the result set means? The answer is metadata. Metadata gives your user the meaning of each data element. Metadata establishes the context for the data elements. Data warehouse users, developers, and administrators interpret each data element in the context established and recorded in metadata

Q.8. List four metadata types used in each of the three areas of data acquisition, data storage, and information delivery.

Ans: Data Acquisition

For metadata types recorded and used in the data acquisition area, please refer to Figure. This figure summarizes the metadata types and the relevant data warehouse processes.

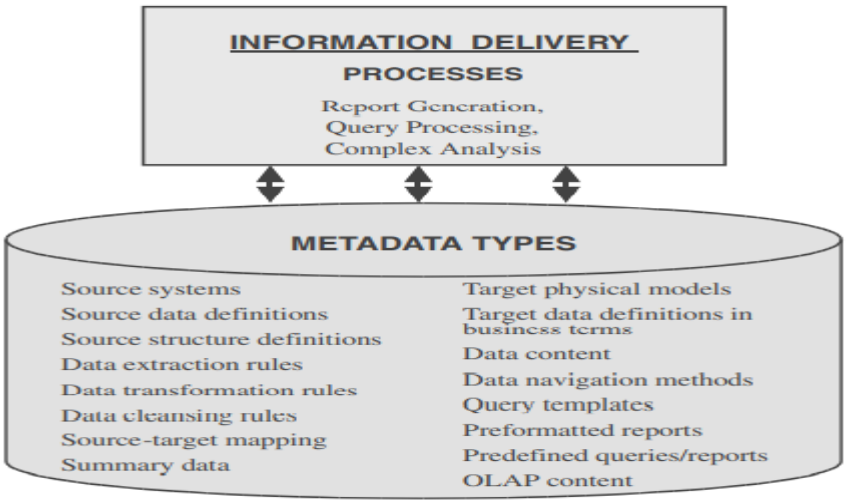


Data Storage

For metadata types recorded and used in the data storage area, please refer to Figure. This figure summarizes the metadata types and the relevant data warehouse processes.

Information Delivery

For metadata types recorded and used in the information delivery area, see Figure. This figure summarizes the metadata types and the relevant data warehouse processes.



Q.9. List any ten examples of business metadata.

Ans:

- Connectivity procedures
- Security and access privileges
- The overall structure of data in business terms
- Source systems
- Source-to-target mappings
- Data transformation business rules
- Summarization and derivations
- Table names and business definitions
- Attribute names and business definitions
- Data ownership
- Query and reporting tools
- Predefined queries
- Predefined reports
- Report distribution information
- Common information access routes

- Rules for analysis using OLAP
- Currency of OLAP data
- Data warehouse refresh schedule

Q.10. List four major requirements that metadata must satisfy. Describe each of these four requirements.

Ans:

Capturing and Storing Data. The data dictionary in an operational system stores the structure and business rules as they are at the current time. For operational systems, it is not necessary to keep the history of the data dictionary entries. However, the history of the data in your data warehouse spans several years, typically five to ten in most data warehouses. During this time, changes do occur in the source systems, data extraction methods, data transformation algorithms, and in the structure and content of the data warehouse database itself. Metadata in a data warehouse environment must, therefore, keep track of the revisions. As such, metadata management must provide means for capturing and storing metadata with proper versioning to indicate its time-variant feature.

Variety of Metadata Sources. Metadata for a data warehouse never comes from a single source. CASE tools, the source operational systems, data extraction tools, data transformation tools, the data dictionary definitions, and other sources all contribute to the data warehouse metadata. Metadata management, therefore, must be open enough to capture metadata from a large variety of sources.

Metadata Integration. We have looked at elements of business and technical metadata. You must be able to integrate and merge all these elements in a unified manner for them to be meaningful to your end-users. Metadata from the data models of the source systems must be integrated with metadata from the data models of the data warehouse databases. The integration must continue further to the front-end tools used by the endusers. All these are difficult propositions and very challenging.

Metadata Standardization. If your data extraction tool and the data transformation tool represent data structures, then both tools must record the metadata about the data structures in the same standard way. The same metadata in different metadata stores of different tools must be represented in the same manner.

Rippling Through of Revisions. Revisions will occur in metadata as data or business rules change. As the metadata revisions are tracked in one data warehouse process, the revisions must ripple throughout the data warehouse to the other processes.

Keeping Metadata Synchronized. Metadata about data structures, data elements, events, rules, and so on must be kept synchronized at all times throughout the data warehouse.

Metadata Exchange. While your end-users are using the front-end tools for information access, they must be able to view the metadata recorded by back-end tools like the data transformation tool. Free and easy exchange of metadata from one tool to another must be possible Support for End-Users. Metadata management must provide simple graphical and

tabular presentations to end-users, making it easy for them to browse through the metadata and understand the data in the data warehouse purely from a business perspective.

The requirements listed are very valid for metadata management. Integration and standardization of metadata are great challenges. Nevertheless, before addressing these issues, you need to know the usual sources of metadata. The general list of metadata sources will help you establish a metadata management initiative for your data warehouse.

EXERCISES

1. Indicate if true or false:

A. The importance of metadata is the same in a data warehouse as it is in an operational system.

Ans: False

B. Metadata is needed by IT for data warehouse administration.

Ans: True

C. Technical metadata is usually less structured than business metadata.

Ans: False

D. Maintaining metadata in a modern data warehouse is just for documentation.

Ans: False

E. Metadata provides information on predefined queries.

Ans: True

F. Business metadata comes from sources more varied than those for technical metadata.

Ans: True

G. Technical metadata is shared between business users and IT staff.

Ans: False

H. A metadata repository is like a general purpose directory tool.

Ans: True

I. Metadata standards facilitate metadata interchange among tools.

Ans: True

J. Business metadata is only for business users; business metadata cannot be understood or used by IT staff.

Ans: False

2. As the project manager for the development of the data warehouse for a domestic soft drinks manufacturer, your assignment is to write a proposal for providing meta-data. Consider the options and come up with what you think is needed and how you plan to implement a metadata strategy.

Ans:

In simple terms, metadata is \"data about data,\" and if managed properly, it is generated whenever data is created, acquired, added to, deleted from, or updated in any data store and data system in scope of the enterprise data architecture.
Metadata provides a number of very important benefits to the enterprise, including:
There are three broad categories of metadata:
All these types of metadata have to be persistent and available in order to provide necessary and timely information to manage often heterogeneous and complex data environments such as those represented by various Data Hub architectures. A metadata management facility that enables collection, storage, maintenance, and dissemination of metadata information is called ametadata repository.
Topologically, metadata repository architecture defines one of the following three styles:
The centralized architecture is the traditional approach to building a metadata repository. It offers efficient access to information, adaptability to additional data stores, scalability to capture additional metadata, and high performance. However, like any other centralized architecture, centralized metadata repository is a single point of failure. It requires continuous synchronization with the participants of the data environment, may become a performance bottleneck, and may negatively affect quality of metadata. Indeed, the need to copy information from various applications and data stores into the central repository may compromise data quality if the proper data validation procedures are not a part of the data acquisition process.

3. As the data warehouse administrator, describe all the types of metadata you would need for performing your job. Explain how these types would assist you.

Ans:
In a data warehouse metadata acts as a nervous system of the organization. It is basically data about data (or information about data), for example, catalog for data warehouse, Data warehouse repository and data warehouse road map etc. Metadata is necessary for building and administrating the data warehouse.
As a DBA, currently I am responsible for physical designing of the database and for doing the initial loading and also responsible for periodic incremental loads.

For physical designing I need metadata layout in staging area. For defining the logical structure of the data warehouse, for data refresh and load cycles. In our warehouse we have defined the few ways in which metadata is classified e.g.

1. Administrative/End User/Optimization
2. Development Usage
3. In the Data Mart/at the workstations
4. Technical Business
5. Backroom/Front room
6. Internal/External

Our data warehouse we have three main function areas

1. Data acquisition: For data acquisition we have processes like data extraction, data transformation, data cleansing etc. The metadata will use for data acquisition are - data extraction method, data cleansing rule, and data transformation types.
2. Data Storage: Data storage process relate to data loading, data achieving, and data management. We will use Source system, source data definition, data extraction method, data cleansing rule, data transformation type's method for data storage
3. Information delivery: Information delivery process includes report generation, query processing, and complex analysis. We will use Source system, source data definition, data extraction method, data cleansing rule, data transformation type's method for information delivery.

Metadata can be further categorized as:

1. Business Metadata: Business metadata focuses on providing the support for end-user at their workstation. It makes easy for the user to understand what data is available in warehouse. It also helps end-user to check predefined queries, predefined data, predefined reports, common information access routes, and data warehouse refresh schedule. Business metadata is primarily benefits the managers, end-users, regular users, power-users, casual users and senior managers.
2. Technical Metadata: Technical metadata is basically for it staff responsible for development and administration of data warehouse. The metadata is required for initial development of data warehouse and technical metadata is essential for ongoing growth and maintenance of data warehouse. Additionally, technical metadata is required for continuous administering of production data warehouse.

4. You are responsible for training the data warehouse end-users. Write a short procedure for your casual end-users to use the business metadata and run queries. Describe the procedure in user terms without using the word metadata.

Ans:

These tools generally fall into two categories: development tools for IT professionals, and information access tools for end-users. When you, as a developer, use a tool for design and development, in that process, the tool lets you to create and record a part of the data warehouse metadata. When you use another tool to perform another process in the design and development, this tool uses the metadata created by the first tool. When your end-user uses a query tool for information access at the front end, that query tool uses metadata created by some of the back-end tools. What exactly is happening here with metadata? Metadata is no longer passive documentation. Metadata takes part in the process. It aids in

the automation of data warehouse processes. Let us consider the back-end processes beginning with the defining of the data sources. As the data movement takes place from the data sources to the data warehouse database through the data staging area, several processes occur. In a typical data warehouse, appropriate tools assist in these processes. Each tool records its own metadata as data movement takes place. The metadata recorded by one tool drives one or more processes that follow. This is how metadata assumes an active role and assists in the automation of data warehouse processes.
Here is a list of back-end processes shown in the order in which they generally occur:

1. Source data structure definition
2. Data extraction
3. Initial reformatting/merging
4. Preliminary data cleansing
5. Data transformation and consolidation
6. Validation and quality check
7. Data warehouse structure definition
8. Load image creation

One of your users wants to run a query to retrieve sales data for three products during the first seven days of April in the Southern Region.

This user composes the query as follows:

Product = Widget-1 or Widget-2 or Widget-3

Region = 'SOUTH'

Period = 04-01-2000 to 04-07-2000

The result comes back:

	Sale Units	Amount
Widget-1—	25,355	253,550
Widget-2—	16,978	254,670
Widget-3—	7,994	271,796

Let us examine the query and the results.

In the specification for region, which territories does region "SOUTH" include?

Are these the territories your user is interested in?

What is the context of the data item "SOUTH" in your data warehouse?

Next, does the data item 04-01-2000 denote April 1, 2000 or January 4, 2000?

What is the convention used for dates in your data warehouse?

Look at the result set. Are the numbers shown as sale units given in physical units of the products, or in some measure such as pounds or kilograms? What about the amounts shown

in the result set? Are these amounts in dollars or in some other currency? This is a pertinent question if your user is accessing your data warehouse from Europe. For the dates stored in your data warehouse, if the first two digits of the date format indicate the month and the next two digits denote the date, then 04-01-2000 means April 1, 2000. Only in this context is the interpretation correct. Similarly, context is important for the interpretation of the other data elements. How can your user find out what exactly each data element in the query is and what the result set means? The answer is metadata. Metadata gives your user the meaning of each data element. Metadata establishes the context for the data elements. Data warehouse users, developers, and administrators interpret each data element in the context established and recorded in metadata

5. As the data acquisition specialist, what types of metadata can help you? Choose one of the data acquisition processes and explain the role of metadata in that process.

Ans:
Metadata is "data about data," and if managed properly, it is generated whenever data is created, acquired, added to, deleted from, or updated in any data store and data system in the scope of the enterprise data architecture.
The types of metadata that are helpful to the data acquisition specialist include but are not limited to:
•External data Structure
•Data extraction methods
•Source System Platforms
•Target Physical models
•External data Definitions
•Data structures in staging area
Data acquisition is the process of extracting data from the data sources, moving all the extracted data to the staging area, and preparing the data for loading into the data warehouse
repository. It includes data extraction, data transformation, data cleansing, data integration, and data staging. Metadata assists users in understanding the origin of the data: which source
systems it came from and which transformations were applied before it was made available in the data warehouse
Roles of metadata in data transformation and cleansing are listed below:
•Specifications for mapping extracted files to data staging files
•Conversion rules for individual files
•Default values for fields with missing values
•Business rules for validity checking
•Sorting and resequencing arrangements Audit trail for the movement from data extraction to data staging